UNIVERSIDAD
DE GRANADA

PROGRAMA DE DOCTORADO EN CIENCIAS SOCIALES

GENERACIÓN DE HERRAMIENTAS DE EVALUACIÓN BIBLIOMÉTRICA A PARTIR DE GOOGLE SCHOLAR

CREATION OF BIBLIOMETRIC TOOLS FOR EVALUATION BASED ON DATA FROM GOOGLE SCHOLAR

ALBERTO MARTÍN MARTÍN

Director: Emilio Delgado López-Cózar

TESIS DOCTORAL

GRANADA, 2019

El doctorando / *The doctoral candidate* **Alberto Martín Martín** y el director de tesis / *and the thesis supervisor* **Emilio Delgado López-Cózar**

Garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección del director de tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Guarantee, by signing this doctoral thesis, that the work has been done by the doctoral candidate under the direction of the thesis supervisor and, as far as our knowledge reaches, in the performance of the work, the rights of other authors to be cited (when their results or publications have been used) have been respected.

Lugar y fecha / *Place and date:*

Granada, …

Director de la tesis / *Thesis supervisor* Doctorando / *Doctoral candidate*

Firma / *Signed* Firma / *Signed*

TABLE OF CONTENTS

List of acronyms

AHSS	Arts, Humanities, and Social Sciences (fields)
GS	Google Scholar (search engine)
GSC	Google Scholar Citations (author profiles service)
GSCP	Google Scholar’s Classic Papers (service that displays highly cited articles)
GSM	Google Scholar Metrics (journal ranking service)
JSM	Journal Scholar Metrics (web application)
MADAP	Multifaceted Analysis of Disciplines through Academic Profiles
OA	Open Access
PoP	Publish or Perish (software)
RG	ResearchGate
SERP	Search Engine Results Page
STEM	Science, Technilogy, Engineering, and Mathematics (fields)
WoS	Web of Science

List of tables

INTRODUCTION

Introducción

CHAPTER 1

CAPÍTULO 1

CHAPTER 2

CHAPTER 3

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 10

List of figures

INTRODUCTION

INTRODUCCIÓN

CHAPTER 2

CHAPTER 3

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 8

CAPÍTULO 8

CHAPTER 9

CHAPTER 10

CHAPTER 11

CHAPTER 12

CHAPTER 13

CHAPTER 14

CHAPTER 15

CHAPTER 16

Agradecimientos

Realizar esta tesis ha requerido, como supongo suele ocurrir en estos casos, un gran esfuerzo y dedicación, en este caso durante casi cinco años. Pero este esfuerzo no ha sido solo mío, pues gracias a las personas de mi entorno personal y profesional, y a la obtención de un contrato predoctoral, he tenido la suerte de poder trabajar en esta tesis bajo unas condiciones verdaderamente privilegiadas. En estos agradecimientos espero transmitir como, realmente, si no fuera por estas personas no habría ninguna posibilidad de que yo estuviera hoy escribiendo estas líneas.

En primer lugar, soy un privilegiado por tener los padres que tengo. Esta tesis ha sido posible gracias a la educación que me han dado desde que nací, en la que me inculcaron la importancia de tener una buena formación. Mis padres siempre me han puesto en bandeja todas las oportunidades de formación que he necesitado (y más). Por poner varios ejemplos, fue mi madre la que decidió ponerme profesores de apoyo de inglés, ya que la formación en el colegio no era suficiente. Gracias a esta inversión continuada durante mis años en la educación secundaria pude desarrollar una habilidad que ha sido crucial en este periodo, pues me ha permitido acceder a recursos con los que formarme autónomamente, así como ser autosuficiente en la escritura de artículos (en lo que respecta al idioma) durante el desarrollo de esta tesis. Fue mi padre quien, cuando yo estaba preparándome la selectividad aun sin tener idea de lo que quería estudiar, se compró un libro que describía todas las carreras universitarias que se podían estudiar en España, se lo leyó, y me dio a conocer la diplomatura en biblioteconomía y documentación. Él sabía que a los 15 años yo ya había diseñado por mi cuenta varias bases de datos para catalogar mis libros y cómics en Microsoft Access, así que era posible que esa carrera me pudiera interesar (acertó de lleno). Han sido ellos dos los que me han apoyado especialmente en este periodo, para que lo único de lo que yo tuviera que preocuparme fuera de hacer mi trabajo lo mejor posible. Gracias mamá, gracias papá.

Empezando con el entorno profesional, he tenido el privilegio de tener un director de tesis de primer nivel cuyo nivel de compromiso ha ido años luz más allá de lo que cualquier doctorando podría esperar. Es imposible trabajar con Emilio y no contagiarse de su pasión por descubrir y aprender cosas nuevas. Él ha sido sin duda el motor de nuestra pequeña cuadrilla durante estos años, tanto por el incesante manantial de ideas que es su mente, como por su impresionante capacidad de trabajo. Las horas de conversación acumuladas en este periodo hablando sobre Google Scholar y otras bases de datos, evaluación científica, acceso abierto… se cuentan por cientos (y no exagero en absoluto). Siempre ha sido mi revisor más exigente, requiriendo un alto nivel de precisión y meticulosidad en cada tarea (desde diseñar e implementar un análisis, hasta para publicar un tweet). Para cumplir con sus expectativas, los límites de mis conocimientos y habilidades se han tenido que ampliar considerablemente. Pero lo que más le agradezco es que desde el primer momento y hasta ahora, todas sus acciones han ido encaminadas a buscar lo mejor para el presente y futuro de los que trabajamos con él. En ocasiones incluso a expensas de su salud personal, lo que nunca debería ocurrir. Es por esto que el título de director de tesis en este caso se queda desmesuradamente corto, y es más adecuado decir que Emilio ha sido para mí un verdadero maestro. Gracias, Emilio.

No solo he tenido suerte con el director, sino que también he tenido el privilegio de poder trabajar con grandes profesionales. A Enrique Orduña le quiero dar las gracias por servir en muchas ocasiones como codirector de esta tesis, un título que le correspondería sin lugar a dudas por todos los consejos y ayuda que me ha prestado a lo largo de estos años. Sin embargo, Enrique asumió esta carga sin reclamar el cargo, y no solo eso, sino que además hizo un trabajo espléndido para asegurar que nuestros trabajos llegaran a buen puerto. Esto es algo que no creo que llegue a poder agradecerle nunca del todo. Sin su ayuda, esta tesis sería muy diferente. También quiero dar las gracias a Juanma Ayllón, que nos acompañó durante gran parte de este camino resolviendo en muchas ocasiones los problemas de calidad y normalización de datos que Google Scholar nos presentaba continuamente. Agradezco mucho poder haber trabajado codo con codo con vosotros dos y con Emilio durante este tiempo, y me siento orgulloso de los trabajos que se han derivado de esta colaboración y que ahora forman parte de esta tesis.

Quiero agradecer también su inestimable ayuda a todos los colegas internacionales que también me han acompañado en diversas partes del camino. Gracias a Anne-Wil Harzing por su colaboración en uno de los trabajos que forman esta tesis, y por las conversaciones sobre Google Scholar que, junto con Emilio, hemos mantenido a lo largo de los años por correo electrónico. Gracias también a Mike Thelwall por aceptar que realizara una estancia de tres meses en 2016 en la University of Wolverhampton (Reino Unido), y por su ayuda durante la estancia, que en 2018 finalmente culminó en la publicación de otro de los artículos que forman esta tesis. Gracias a Ludo Waltman y Nees Jan van Eck del Centre for Science and Technology Studies (CWTS) de Leiden University (Países Bajos), que aceptaron que realizara otra estancia de tres meses en su centro en 2017, y que me ayudaron a encontrar una manera para aplicar sus algoritmos y programas a datos de Google Scholar. Aunque este trabajo todavía no haya tenido la oportunidad de manifestarse en resultados concretos, estoy impaciente por continuar trabajando en ello una vez defienda la tesis. Finalmente, gracias a Rodrigo Costas y Thed van Leeuwen, también del CWTS, por acompañarme en mi primera aproximación al tema del acceso abierto a la comunicación científica. Me alegra mucho que aquel proyecto del que hablamos por primera vez en el ISSI de 2015 en Estambul finalmente tomara forma como uno más de los artículos de esta tesis. De Rodrigo, además, no olvidaré la agradable acogida que me ofreció a mi paso por Leiden (sobre todo las cenas en el Shabu Shabu).

Esta tesis tampoco hubiera sido posible sin la financiación del Ministerio de Educación, Cultura y Deportes (MECD) a través del programa de Formación de Profesorado Universitario. Obtener un contrato predoctoral de este programa me permitió no tener que preocuparme del aspecto económico durante cuatro años, y así poder concentrarme a diario en la tesis como mi trabajo principal. El MECD, además, también cubrió los gastos de mi estancia en Leiden. Agradezco asimismo a la Universidad de Granada y al CEI BioTic Granada el haber financiado los gastos de mi estancia en Wolverhampton a través del programa propio del vicerrectorado de investigación, y al Departamento de Información y Comunicación de la Universidad de Granada el haber financiado los gastos de inscripción y asistencia a los congresos internacionales STI de 2017 y 2018.

Quiero acordarme también de los compañeros y compañeras que están trabajando en sus tesis en condiciones no tan favorables como han sido las mías: gente a la que se le negó el contrato predoctoral en España y ahora están contribuyendo muy satisfactoriamente a la ciencia en otros países, o gente que decidió quedarse en España y busca tiempo para hacer la tesis una vez acaba su jornada laboral. También de gente que compagina la tesis con cargas familiares o la saca adelante a pesar de sufrir problemas de salud, y de gente que no recibe suficiente apoyo de su director o directora. En definitiva, de toda la gente que tiene que superar obstáculos adicionales a los que ya de por sí presenta un proyecto de la envergadura de una tesis doctoral. Hay mucho camino por recorrer si queremos que buenas experiencias como la que yo he podido disfrutar se conviertan en la norma, en vez de ser la excepción.

Durante la realización de esta tesis no solo he tenido suerte en el aspecto académico/profesional, sino que de una manera tan inesperada como bienvenida, a raíz de esta tesis también ha habido cambios importantes en mi vida personal. Hoy en día ya no es tan raro conocer a gente gracias a aplicaciones, pero quizás es más inusual cuando esta aplicación es… Mendeley. Hace ya casi tres años que conocí a JingXuan, mi pareja, en aquel curso de gestión de referencias bibliográficas que impartí junto a Emilio a un grupo de estudiantes de doctorado. Desde entonces, JingXuan se ha convertido en mi mejor amiga y compañera de viaje, y ya no puedo imaginar cómo sería mi vida sin ella. Conocedora de las dificultades que entraña una tesis, ha sido mi principal apoyo en los momentos más difíciles de la segunda mitad de este periodo, la que más ánimos me ha dado para continuar cuando estaba bloqueado, y la primera que se alegraba cuando le contaba buenas noticias. Por todo esto y mucho más, muchas gracias, JingXuan.

Por último, me gustaría dedicarle esta tesis también a mi hermano menor, David. Todo lo que vale la pena en esta vida requiere esfuerzo, pero la dedicación siempre se ve recompensada al final. Tú tienes la capacidad de hacer cualquier cosa que te propongas en tu vida, y tu familia además va a estar a tu lado para ayudarte en lo que haga falta.

Acknowledgements

Working on this thesis has required, as I suppose is usually the case, great effort and dedication, in this case during almost five years. However, I have not been the only one to put effort into it. Thanks to the people in my personal and professional environment, as well as to the obtention of a doctoral grant, I have been able to work in this thesis under extremely privileged circumstances. In these acknowledgements my intention is to try to get across how without any doubt, if it were not for these people, I would not be writing these lines today.

First and foremost, I want to thank my parents. This thesis has been possible thanks to the education I have received from them since I was born. They instilled in me the importance of a good education, and provided me with all the opportunities to learn that I could ever need (and more). A few examples: when it became clear that the English class in school wouldn't be enough to learn English adequately, my mother decided to invest in private teachers throughout all my years in secondary education. The skills I learned have proved to be critical during this period, because they have allowed me to access resources with which to learn independently, as well as to be relatively self-sufficient in the process of writing this thesis. For his part, my father was the one who first pointed me to the Library and Information Sciences undergraduate program at the university, at a time when I had no idea what to do after high school. He bought a book that described all undergraduate programs available in Spain, and suggested that the LIS program might be a good fit for me. He knew I had spent a considerable amount of my teen years designing databases for my book and comics collection. My parents have supported me especially during this period, so that the only thing I had to worry about was how to do my work as best as possible. Thank you Mum, thank you Dad.

In the professional environment, I have had the privilege of working with a first-class thesis advisor whose level of commitment to this thesis has gone far and above what any PhD candidate could reasonably expect. It is impossible to work with Emilio and not be swayed by his passion to discover and learn new things. He has been without a doubt the main engine of our small team during these years, both because of the unceasing stream of ideas that is his mind, and because of his amazing work capacity. It is no exaggeration to say that the accumulated hours of conversation talking about Google Scholar, other databases, scientific evaluation, open access… are in the high hundreds. He has always been my most critical reviewer, and has required a high level of precision and meticulousness in every task (from designing and implementing an analysis, to publishing a tweet). To meet his expectations, the limits of my knowledge and abilities have had to expand considerably. However, what I value most is that from the first moment and until now, all his actions have been driven by will to achieve what is best for the present and future of the people who work with him. In some cases, even at the expense of his personal health, which is a line that should never be crossed. This is why the title of thesis advisor is in this case evidently insufficient, and it is more appropriate to say that Emilio has been a true mentor. Thank you, Emilio.

I have also been lucky to be able to work with great professionals. I want to thank Enrique Orduña for fulfilling in many occasions the role of co-advisor of this thesis, a title that he undoubtedly deserves after all the advice and help that he has provided me over these years. However, Enrique took up this workload without ever claiming the position, and not only that, he also did a splendid job to make sure that our studies reached a good destination in the end. This is something for which I will be always grateful to him. Without his help, this thesis would be very different. I would also like to thank Juanma Ayllón, who accompanied us during a large fraction of this time, helping us with the issues of quality and normalization of GS data. I am very grateful for having been able to work with you both and with Emilio during this time, and I am proud of the works that have resulted from this collaboration and now are part of this thesis.

I also want to thank their invaluable help to all the international colleagues who have accompanied me in various parts of this path. I would like to thank Anne-Wil Harzing, for her collaboration in one of the studies that are part of this thesis, and for the conversations about Google Scholar that, together with Emilio, we have maintained over e-mail. I would like to thank Mike Thelwall for accepting my proposal to do a three-month research stay in 2016 at the University of Wolverhampton, and for his help during the stay, which in 2018 resulted in the publication of one of the studies that are part of this thesis. I would like to thank Ludo Waltman and Nees Jan van Eck from the Centre for Science and Technology Studies (CWTS) at Leiden University, who accepted my proposal to do a three-month research stay at CWTS in 2017, and who helped me to find ways to apply their algorithms and software to data from GS. Although this work has not still materialized in concrete results, I am impatient about continuing working on it after completing the thesis. Lastly, I would like to thank Rodrigo Costas and Thed van Leeuwen, also from CWTS, for helping me in my first formal incursion to the topic of Open Access to scientific communication. I am very happy that the project we first discussed at the 2015 ISSI in Istanbul finally could take shape as one of the articles of this thesis. To Rodrigo, I am also grateful for the warm welcome that he offered me while I was in Leiden (especially the dinners at Shabu Shabu).

This thesis could not have been possible without the funding provided by the Ministerio de Educación, Cultura y Deportes (MECD) through the Formación de Profesorado Universatio program (FPU). Obtaining a four-year contract meant that I did not have to worry about the economic aspect during this time, and so I could focus on my thesis as my primary job. Furthermore, the MECD also covered the expenses of my research stay in Leiden. I would also like to thank Universidad de Granada and CEI BioTic Granada for covering the expenses of my research stay at the University of Wolverhampton. Lastly, I would also like to thank the Department of Information and Communication at Universidad de Granada for covering the expenses to attend the STI international conference in 2017 and 2018.

I would also like to bring attention to the fact that there are PhD candidates that are working in their thesis in conditions that are not as good as the ones I have been able to enjoy: people who were denied a grant in Spain, but are currently contributing to Science in other countries very satisfactorily, or people who decided to stay in Spain and struggle to find time to work on their thesis after their full-time jobs, people who have to balance their thesis with family life or health issues, and last but not least, people who do not receive the necessary support from their advisors. In short, I want to remember everyone who realise they have to overcome other obstacles in addition to those that a project such as a thesis naturally presents on its own. There is a long way to go if we want that experiences such as the one I have been able to enjoy become the norm, instead of being an exception.

During this thesis I have not only been lucky in the academic/professional aspects. In a way that was as unexpected as it was welcome, activities related to this thesis have also provoked important changes in my personal life. Nowadays it is not so rare to meet people thanks to applications, but perhaps it is more unusual if this application is… Mendeley. It has been almost three years since I met JingXuan, my girlfriend, in a course that Emilio and I taught to help other doctoral students learn to manage references. Since then, JingXuan has become my best friend and life companion, and I cannot imagine my life without her. Being aware of the difficulties of working on a thesis, she has been my main support in the difficult moments of the second part of this period, the one who encouraged me to continue on it when I was blocked, and the first one to celebrate it when there were good news. For all this and much more, thank you, JingXuan.

Lastly, I would like to dedicate this thesis to my younger brother, David. Everything that is worth doing requires an effort, but this dedication is always rewarded in the end. You have the capacity to do anything you want in your life, and your family is going to support you.

Introduction

Description of Google Scholar

Google Scholar (GS) is a freely-accessible academic search engine that indexes academic literature from a wide range of disciplines, document types, and languages. Its main goal is to help users find relevant resources of an academic nature. It also provides other additional services, such as quick generation of bibliographic references in various styles and formats, access to the full text of the academic documents (when they are freely accessible on the web, or when the user has access to these documents via institutional subscriptions), and identification of the number of times a document has been cited in other documents. Unlike the general Google search engine and other products developed by Google, GS has never been monetized. The platform has never displayed any advertisement whatsoever nor, according to its founder, has it ever collected usage data from users that could be reused in a commercial way (Acharya, 2015). Thus, GS can be considered to be a not-for-profit project subsidized by a for-profit company.

GS was developed by Google engineers Anurag Acharya and Alex Verstak while on sabbatical from the company. Both of them had spent some time at the academia, and knew the difficulties researchers face when they need to locate scientific information, especially in developing countries. Acharya and Verstak presented the idea to Larry Page, and on the 18th of November, 2004, GS was launched. It was in this manner that Google, a company whose initial success had stemmed from an algorithm that had been strongly influenced by the scientific literature on citation analysis (Brin & Page, 1998), embarked in a project that applied similar ideas to academic documents.

GS was innovative in several ways. First and foremost, unlike Web of Science (WoS) and Scopus, which had a selective approach to document indexing (they only index documents published in certain venues), GS chose an inclusive approach, indexing any seemingly academic document that its crawlers could find on the web. In addition to that, although most of the other academic databases only handled metadata (authors, title, abstract, keywords… of documents), GS made a point of indexing the full text of the documents whenever possible (when the full text was freely accessible, or when agreements with publishers allowed it). In this way, user queries were not only confronted with a selective index of metadata, but an all-encompassing index of both metadata and full text. For a more detailed description of GS's general functioning, we refer to Delgado López-Cózar, Orduna-Malea, Martín-Martín, & Ayllón (2017) and Orduña-Malea, Martín-Martín, Ayllón, & Delgado López-Cózar (2016, Chapter 4). Another thing that made GS different from most other academic databases was its simple search interface. While most databases presented an advanced search panel in which users had to carefully design every aspect of the query (fields, operators, filtering options…) GS replicated the simplicity of the Google search engine: a search box, complemented with very limited advanced query and filtering options. For a more detailed description of GS's search functionality we refer to Orduña-Malea et al. (2016, Chapter 6).

Accessing the full text of the documents published by publishers of all stripes (commercial, non-commercial, university presses, scientific societies…) and other channels of communication such as repositories or researchers' personal websites also enabled GS to build its own citation graph by processing the references at the end of each document and matching them to documents already identified in their index. The citation graph was key to implement GS's relevance ranking algorithm. Meanwhile, most other databases only permitted simple document sorting (by author, by title, by date of publication). Of the remaining major multidisciplinary academic databases available at the time, only the subscription-based database WoS, and the recently created Scopus (also subscription-based) had the option of sorting documents by number of citations. For many years, GS was the only freely-accessible multidisciplinary academic database with an extensive coverage that displayed citation data.

Its success as a search engine was immediate, despite having been launched just a few weeks after Elsevier's Scopus (3rd of November, 2014) (Elsevier, 2004). Before the year was out, both *Science* (Science, 2004) and *Nature* (Butler, 2004) had already reported on the arrival of this new search engine. According to Giles (2005), just one year after its launch GS already directed more traffic to *Nature* than any other multidisciplinary academic search engine. Similarly, Giustini (2005) found that Google and GS were the sites that directed more traffic to the *British Medical Journal*, followed at a distance by Yahoo and Pubmed-Medline. Wang & Howard (2012) found that between 2006 and 2011 the use of GS by users at San Francisco State University grew ten-fold, becoming "the top-ranked SFX source for requests in 2011", and favored above the university's search tool.

Numerous surveys also have pointed in the direction that GS has become many researchers' tool of choice to search academic information. van Noorden (2014) reported that according to the results of a survey to over 3,000 scientists and engineers, GS was the most visited site. Mussell & Croft (2013) and Nicholas et al. (2017) found similar results when studying students and early career researchers, respectively. For their part, Bosman & Kramer (2016) found that GS was the most used tool to search literature by far (over twice as many respondents as the following option). Their results also show that it was the most used tool to get alerts/recommendations. In 2015, WoS acknowledged that "Google Scholar is increasingly the starting spot for researchers" (Clarivate Analytics, 2015) and for this reason entered into an agreement with GS so that WoS citation counts (and a link to access WoS) would appear embedded in the results pages of GS (only for users with a subscription to WoS). Despite its continued growth and success, it was reported that by 2014 only nine people worked in Google Scholar (Van Noorden, 2014b).

In addition to the academic search engine, in 2011 GS launched its author profile service Google Scholar Citations (GSC)[1]. For a few months, this service could only be used by a limited number of beta users, but on the 16th of November of the same year it became available to all users. With this service, GS added some important features to its portfolio: by leveraging the vast document coverage of the search engine, users could now easily create a personal profile that listed all their publications. Citation counts were also visible, and several bibliometric indicators were automatically computed from the list of publications: total number of citations received by the author, h-index (Hirsch, 2005), and i10-index (an indicator invented by GS, defined as the number of documents in the profile with at least 10 citations). Each of the three indicators is computed twice, in one version they do not apply any limitation to the citation window (the span of years used in the calculation), while in the other version they use a citation window of 5 years (the last 5 complete years at any given time). The service gives total freedom to the user to add, edit, merge, and remove documents from the profile, with no external controls whatsoever. Users can set up alerts to be notified when their profiles or other authors' profiles are updated (in terms of both publications and citations). Bosman & Kramer (2016) found that GSC is the second most widely used tool to create author profiles (closely following ResearchGate). For a more detailed description of the service, we refer to Orduña-Malea, Martín-Martín, Ayllón, & Delgado López-Cózar (2016, Chapter 9).

Roughly a year later, in April of 2012[2], GS surprised the scientific community again by launching Google Scholar Metrics for Publications (GSM), a ranking of publication venues (scientific journals, but also some conference proceedings and repositories). On a first level, publication venues are classified by language. For languages other than English, the top 100 publication venues with a higher h5-index (h-index of documents published in the last 5 complete years) are displayed in the language ranking. Only for English-language venues a journal-level subject categorization is provided. The subject categorization has two levels: a first level with 8 broad areas (Business, Economics & Management; Chemical & Material Sciences; Engineering & Computer Science; Health & Medical Sciences; Humanities, Literature & Arts; Life Sciences & Earth Sciences; Physics & Mathematics; Social Sciences), and a second level with approximately 250 categories (it changes slightly from year to year). Within each broad category or subcategory, the top 20 publication venues with a highest h5-index are displayed. Besides the publication venues displayed in the rankings, many more can be found by using the available search tool. For each publication venue, the list

[1] https://scholar.googleblog.com/2011/07/google-scholar-citations.html

[2] https://scholar.googleblog.com/2012/04/google-scholar-metrics-for-publications.html

of documents above the h5-index threshold, as well as the list of citing documents, can be visualized within the platform. Its inclusion criteria state that in order for a publication venue to be included, it must have published at least 100 documents in the last 5-year period, and must have received at least one citation. GSM has been released once a year since 2012. For a more detailed description of the service we refer to Orduña-Malea et al. (2016, Chapter 8).

In 2017 GS introduced an additional (but apparently short-lived) (Orduna-Malea, Martín-Martín, & Delgado López-Cózar, 2018) service to its portfolio: Classic Papers: Articles that have stood the test of time[3] (GSCP). This service was launched on the 14th of June, 2017, shortly before the 2017 edition of GSM. However, while the 2018 edition of GSM was released on August of 2018, no new edition of GSCP has been released. Additionally, although the service is still accessible[4], the link to access it has been removed from the Metrics section of GS, making it virtually inaccessible to anyone who does not already know about its existence. This service provides lists of the top ten most highly-cited English-language journal articles published in 2006 (to provide a citation window of 10 years), by subject categories. The structure of the categories is the same as the one provided in GSM (two levels: eight broad categories, and 252 subcategories), but unlike in GSM, the classification was applied at the article level (articles published in multidisciplinary journals are classified in the appropriate specific category). In total, 2,515 documents can be visualized. This is because the service set a minimum threshold of 20 citations per document, and in one category (French Studies) only five documents could be found that met that criterion.

The widespread use of GS and its spin-off services GSC and GSM spurred the publication of a large number of studies that analysed its characteristics, in many cases by making comparisons with other academic databases. In the course of our work, my colleagues and I have identified almost 300 studies that analysed GS or used data from GS in some way[5]. Figure 1 shows the distribution of studies by year of publication. The peak publication year was 2014 (45 studies), the year when GS celebrated its tenth anniversary.

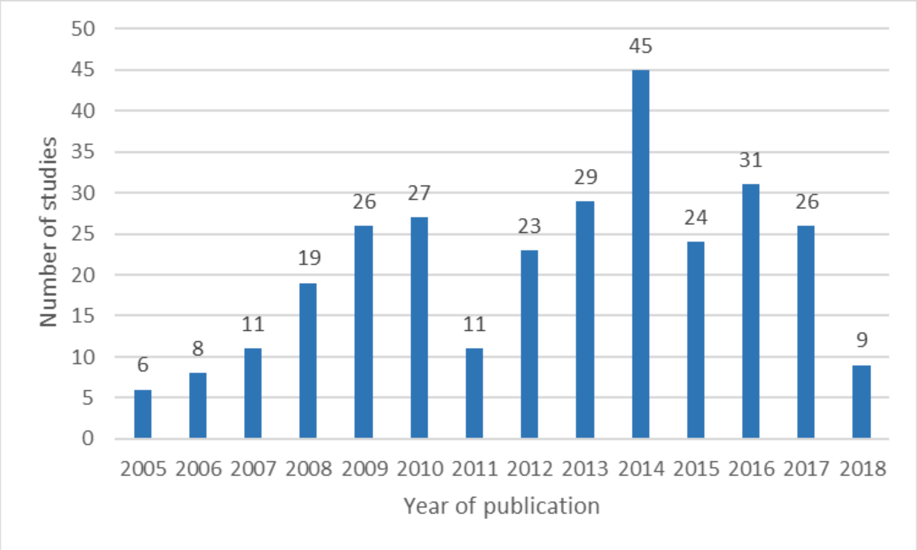


Figure 1. Distribution of studies concerning Google Scholar, by year of publication

Studies that analyse GS generally fall into two groups. First, there are those which are concerned with its capabilities as a tool to search scientific literature. Second, there are those which are interested in GS as a source of bibliographic and citation data that can be used to carry out bibliometric analyses. In this work we

[3] https://scholar.googleblog.com/2017/06/classic-papers-articles-that-have-stood.html
[4] https://scholar.google.com/citations?view_op=list_classic_articles&hl=en&by=2006
[5] http://googlescholardigest.blogspot.com/p/bibliography.html

are mainly concerned with the latter. For a review of studies about GS as a search tool, we refer to Orduña-Malea et al. (2016, Chapter 2).

Although GS is freely accessible and displays citation data, it has never facilitated a suitable method (neither paid nor free) to export data in bulk in order to carry out bibliometric analyses. Unlike Clarivate Analytics and Elsevier, which in addition to offering a research discovery solution (Web of Science, Scopus), they also sell licenses to their data as well as products that facilitate bibliometric analyses (InCites, SciVal), the main goal of GS has always been to facilitate content discovery. Citation counts are considered useful to the extent that they facilitate the achievement of this goal. Even the spin-off products GSC and GSM, which may be considered to have a more bibliometric nature to them, are useful to find people who are working in any given topic and stay up to date with their new work (GSC), or to identify the journals and articles with a higher impact in any given language or field (GSM). Additionally, the agreements that GS made with some publishers in order to access their content preclude any form of automated or bulk access to the data (Van Noorden, 2014b). GS asks users not to use automated methods to extract data from the search engine[6]. Furthermore, it has enforced this policy by putting in place a very strict CAPTCHA system: if users make too many queries too quickly (the threshold has changed over time, usually to become even more strict), they are asked to solve a CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart). GS uses the reCAPTCHA technology, which was acquired by Google in 2009[7]. This test used to consist in users having to fill in a text box with the same letters that appeared in a warped image, but in the last few years it was updated, and now users are asked to select among a series of images those that meet some specific criterion (images that contain cars, traffic signs, roads, bridges, sushi…).

Despite these difficulties, a number of applications have been developed to extract data from GS automatically for diverse purposes (Martín-Martín & Delgado López-Cózar, 2018). Without a doubt, the most widely known of these applications is *Publish or Perish* (PoP), developed by Anne-Wil Harzing (Harzing, 2007). The availability of GS in combination with PoP facilitated the development of bibliometric studies that used data from GS, thus contributing to the democratization of citation analysis (Harzing & Mijnhardt, 2015; Harzing & van der Wal, 2008), which up to 2004 had been a type of analysis that only those with access to WoS could perform.

Research on Google Scholar as a source of bibliographic and citation data

During the first years of life of GS, Verstak and Acharya spent considerable effort getting commercial publishers to agree to let GS's crawlers to index their websites. They understood that an academic database only makes sense if it contains the documents that users need. The goal was that, even if the documents couldn't be freely accessed by all users, users should at least be aware that the documents exist.

Because GS does not provide a public list of the sources or documents it indexes, it has always been difficult to know the exact agreements it has reached with each publisher. Despite this, in 2005 it already covered many important publishers, such as: Association of Computer Machinery (ACM), Blackwell, Institute of Electrical and Electronic Engineers (IEEE), Ingenta, Nature Publishing Group, Springer, Wiley, as well as other repositories such as the NASA Astrophysics Data System, arXiv, or PubMed, although not all of them exhaustively (Notess, 2005). For example, although PubMed was already indexed in 2005, Vine (2005) discovers that documents in PubMed appear in GS with a delay of one year, which of course reduced the usefulness of the tool to search the latest publications.

[6] https://scholar.google.com/intl/en/scholar/help.html#export

[7] https://googleblog.blogspot.com/2009/09/teaching-computers-to-read-google.html

Moreover, negotiations with commercial publishers to get their content indexed in GS were sometimes complicated. According to Butler (2004), at the time Elsevier had "declined to allow Google to index its text". Vine (2006) reports that the full texts of papers published by Elsevier and the American Chemical Society were still not being indexed by GS. It isn't until 2007 when Elsevier journals began to be indexed through the platform ScienceDirect (Brantley, 2007).

Perhaps the most famous case of tension between a major commercial publisher and GS was that of the American Chemical Society (ACS), who sued Google over the use of the term "Scholar" in its new academic search engine, because they considered that "Google's use of the word scholar infringes on ACS's SciFinder Scholar [a subscription-based search application] and Scholar trademarks and constitutes unfair competition" (Mehta, 2004). The lawsuit was settled out in 2006, each side agreeing to pay its own attorney fees (McCullagh, 2006). In part because the ACS didn't allow GS to index its papers, the document coverage of the field of Chemistry in GS was limited for a number of years, as evidenced by the results of Bornmann et al. (2009). However, this gap in coverage had been corrected by 2013 (Harzing, 2013b, 2013a).

The inclusion of JSTOR was another milestone in the history of GS. Although JSTOR was one of the largest subscription-based platforms in terms of the number of journal articles offered, non-subscribers did not have access to basic bibliographic information such as the abstract of the article, in part because most of their library consisted of scanned articles. GS convinced JSTOR to at least display the first page of each article (the page that usually contains the basic bibliographic information and the abstract) to non-subscribers (Levy, 2014). This enabled GS users to make better decisions regarding which documents could be relevant for them.

From early on, GS received a mix of criticism and favorable reviews. In December of 2004, Wentz (2004) was one of the first to state that "GS citation figures are unreliable at best, a waste of time at worst", and that "Google Scholar should withdraw the 'cited by' feature from its Beta version and probably not offer it in the final version", based on the examination of several papers. In 2005, Peter Jacsò carried out a number of small scale comparison between data from WoS, Scopus, and GS (Jacsó, 2005a). After analysing the coverage in these three sources for a small number of articles published in *Current Science*, he found that many of them were not covered by GS, and those that were covered had lower citation counts than in WoS and Scopus. In another study (Jacsó, 2005b), he found similar results: for a sample of articles from the *Asian Pacific Journal of Allergy and Immunology*, WoS found a significantly higher number of citations (1,355) than GS (595). Nevertheless, Jacsó also recognized the merit of the platform as an academic search engine (Jacsó, 2005c).

Pauly & Stergiou (2005) found that while GS found less citations than WoS for older documents (published before 1989) in a sample of 114 papers in several scientific areas, it found more citations than WoS when it came to more recent documents. Additionally, correlations of citation counts were very strong (.84-.99). Bar-Ilan (2006) analysed the document coverage for the scientific production of a mathematician and computer scientist in WoS, GS, and CiteSeer, finding that GS found slightly more citations than WoS, and these two found significantly more citations than CiteSeer. Walters (2007) compared the coverage of several databases and finds that, for a sample of 155 pre-selected articles in the topic of later-life migration, GS is the tool that finds more of them (93%).

Jacsó (2006) countered that studies that were enthusiastic of GS's coverage were shallow, because they did not consider that GS also covered non-journal document types such as conference papers, books, chapters, and dissertations, while other databases like WoS did not. Jacsó also questioned the validity of the data in GS because of its errors of "artificial unintelligence", such as creating incorrect citation matches that inflated citation counts, or returning over 40,000 hits for a search of the author "I Introduction".

Errors like the one described above were very common in the first years of GS, when many of the documents it indexed did not have structured and standard metadata, and GS had to guess the metadata from the text of the pdfs, each with its own particular layout, or even worse, created from scanned images. In some cases, the section "1. Introduction" was erroneously taken as the author (or sometimes the title) of the document by GS's parsers, giving way to the error Jacsó found. Harzing & van der Wal (2008) replicated this search and found that the hit count had decreased to 956, and Orduña-Malea et al. (2016) did the same in 2015 and found that just 5 results were returned by the query, confirming that GS gradually solved this problem over time.

Bakkalbasi, Bauer, Glover, & Wang (2006) published the first study that performed a citation-by-citation coverage comparison between WoS, Scopus, and GS. According to their results, GS only found 53% of all possible citations (WoS found 70% and Scopus 76%), and 13% of the citations could only be found by GS (while in WoS 28% of the citations were unique, and in Scopus 31%). This study would in time be followed by many others of a similar nature (Bar-Ilan, 2010; de Winter, Zadpoor, & Dodou, 2014; Jacimovic, Petrovic, & Zivkovic, 2010; Kousha & Thelwall, 2008; Lasda Bergman, 2012; Meho & Yang, 2007; Moed, Bar-Ilan, & Halevi, 2016; Sember, Utrobicić, & Petrak, 2010; Yang & Meho, 2007). We refer to Martín-Martín, Orduna-Malea, Thelwall, & Delgado López-Cózar (2018) (Chapter 7 of this thesis) and Delgado López-Cózar, Orduna-Malea, & Martín-Martín (2019) for a more in-depth revision of this topic.

Christianson (2007) was the first who leveraged GS's feature of pointing users to freely available versions of documents to analyze the degree of Open Access of publications in the field of ecology. She found that 38% of the articles were freely available. Many other articles that performed similar analyses followed this one (Abad-García, González-Teruel, & González-Llinares, 2018; Jamali & Nabavi, 2015; Khabsa & Giles, 2014; Laakso & Lindman, 2016; Laakso & Polonioli, 2018; Martín-Martín, Orduna-Malea, Ayllón, & Delgado López-Cózar, 2016; Mikki, Ruwehy, Gjesdal, & Zygmuntowska, 2018; Norris, Oppenheim, & Rowland, 2008; Pitol & De Groote, 2014; Teplitzky, 2017). For a more in-depth revision of this topic, we refer to Martín-Martín, Costas, Van Leeuwen, & Delgado López-Cózar (2018) (Chapter 16 of this thesis).

Jacsó (2008) acknowledged that GS had improved its coverage in terms of journals, books, and other document types from all parts of the geography and all languages, but at the same time he criticized the inconsistent hit counts returned by a number of queries and related variants, and metadata interpretation errors made by the automatic parser. His criticism of GS was further elaborated in Jacsó (2008b, 2009, 2010, 2012a, 2012b).

Harzing & van der Wal (2008) opened the way to the use of Google Scholar as a source of data for citation analysis with the creation of the *Publish or Perish* software. They stated that GS could be of use for citation analyses that involve fields which are not well covered by other citation databases. From that moment, evidence started to accumulate suggesting that the coverage of GS is usually more comprehensive than in most of the selective databases (especially in fields like the Humanities and Social Sciences) and the citation counts they provided higher than those provided by other databases (Amara & Landry, 2012; Bar-Ilan, 2010; Cabezas-Clavijo & Delgado-López-Cózar, 2013; Chen, 2013; de Winter et al., 2014; Delgado-López-Cózar & Repiso-Caballero, 2013; Franceschet, 2009; García-Pérez, 2010; A.-W. Harzing, 2013, 2014; Hodge & Lacasse, 2011; Howland, Howell, Wright, & Dickson, 2009; Jacimovic et al., 2010; Jacobs, 2009; Kousha & Thelwall, 2008; Kulkarni, Aziz, Shams, & Busse, 2009; Lasda Bergman, 2012; Martell & Martell, 2009; Mikki, 2009; Minasny, Hartemink, McBratney, & Jang, 2013; Mingers & Lipitakis, 2010; Ocholla & Onyancha, 2009; Rosenstreich & Wooliscroft, 2009; Sember et al., 2010; Zarifmahmoudi, Kianifar, & Sadeghi, 2013; Zarifmahmoudi & Sadeghi, 2012). There were also studies that reported that the opposite was true in certain fields (Adriaanse, Rensleigh, & Rensleigh, 2011; Bornmann et al., 2009; Adriaanse & Rensleigh, 2013). Additionally, Aguillo (2011) provided one of the first estimations of the size of GS (86 million records, from data collected in 2010) and considered that GS lacked "the quality control needed for its use as a bibliometric tool", and Delgado López-Cózar, Robinson-García, & Torres-Salinas (2014) brought attention to the fact that citation counts in GS can be easily gamed.

2014 was the year when GS celebrated its tenth anniversary, and for this occasion the team behind GS published two studies on the effect that the web in general, and GS in particular had had on scholarly communication. In Verstak et al. (2014) and Acharya et al. (2014), the GS team provided evidence that the percentage of citations to old documents, as well as the fraction of highly-cited articles published in non-elite journals, were increasing over time. They posited that this was possible because in a web environment, finding and reading relevant older articles, or relevant articles published in non-elite journals is about as easy as finding and reading recent articles, or articles published in elite journals. This effect was later confirmed using a sample of data from WoS (Martín-Martín, Orduna-Malea, Ayllón, & Delgado-López-Cózar, 2016). In 2014 several estimations of the size of GS were published: according to Khabsa & Giles (2014), who analyzed only English-language documents, its size at that time was nearly 100 million documents. Ortega (2014), using data from 2012, reached a similar figure: 95 million. Shortly after, another study that used different methods estimated its size in approximately 160 million documents (Orduña-Malea, Ayllón, Martín-Martín, & Delgado-López-Cózar, 2014). This study was replicated in 2017 (Delgado López-Cózar et al., 2019), finding that the number of articles in GS had increased to approximately 200 million (331 million if cited references and patents were included). In 2018, it was replicated again (Gusenbauer, 2018), finding that the record count had yet increased to 389 million.

For a more in-depth review of all the studies on GS as a source of bibliographic and citation data we refer to (Delgado López-Cózar et al., 2019; Halevi, Moed, & Bar-Ilan, 2017; Orduna-Malea, Ayllón, Martín-Martín, & Delgado López-Cózar, 2015).

Objectives

2014 was also the year in which I initiated my doctoral training. Given the evidence available up to that point that GS could be a useful source of bibliographic and citation data, especially in the fields where other citation indexes like WoS and Scopus were known to have poor coverage, such as the Humanities and Social Sciences, the general goal of this thesis was the following: to explore whether it is feasible and sustainable to re-use data available in GS to generate data products or tools of a bibliometric nature that provide functionalities that GS does not provide. To accomplish this goal, I joined my thesis advisor Emilio Delgado López-Cózar in his study of GS as a source of data for bibliometric analyses.

To accomplish the general goal of the thesis, we followed two approaches that have ran side by side. First, we endeavored to carry out studies that analysed the general characteristics of Google Scholar as a source of data: its strengths and weaknesses related to a number of aspects (detailed in the research questions below). When possible, these results were benchmarked against the subscription citation databases WoS and/or Scopus. Some of the studies in this category are a part of this thesis, while others, in which I performed in a supporting role, are not part of this thesis. Nevertheless, these are also listed below to provide a complete overview of the work that we have carried out over the last five years.

The second of the main objectives of this thesis was to test the knowledge obtained in the previous studies in practical real-life situations. These projects took the form of tailored web applications built for a variety of purposes, and open to everyone. The applications display data extracted from Google Scholar (and sometimes also other services) in ways that the native GS, GSC and GSM interfaces do not, thus expanding the range of ways in which users can interact with this information. The publications that describe these web applications, as well as the publications that analyse the data included in the web applications, are included in this thesis.

It is also important to note that during this thesis we decided to follow a publishing strategy aligned with the principles of Open Science. As a result, we first published the results and data used in our exploratory analyses as working papers that we deposited in preprint servers. These working papers were subsequently refined into one or several articles that were formally published in peer-reviewed journals. However, not all the content presented in the working papers made it to the journal articles. For this reason, in this thesis we include both the original working papers, and the journal articles.

Lastly, this thesis also contains several chapters describing works in progress that have not yet appeared published in any form: one of this chapters is description of a plan to create a web application that displays exhaustive and detailed bibliographic and bibliometric data about researchers working in Spain who have a Google Scholar Citations profile (and their publications). The second unpublished chapter describes the features of a web application to display data on OA status of publications at various levels of aggregation. This application has already been implemented, but it has not been made public.

The specific GS-related topics my colleagues and I have covered during the last five years are listed below, and Table 1 provides the list of documents that address each of these topics. The table also groups together the initial working papers with the journal articles that were subsequently published, and specifies whether the document is included in the thesis (in which case, the chapter number is provided).

- Description of GS as a platform: documents that describe the general functioning of GS, and the features of the search engine and its main spin-off services (GSM, GSC, GSCP).
- Size and coverage of GS: documents in which we analyse GS's database, at the level of documents, journals, or authors. In some cases, results are benchmarked with other databases.
- Errors / limitations of GS: documents in which we describe the various types of errors and limitations that we have found while trying to use GS data for bibliometric purposes.
- Indicators in GS: studies that analyse the bibliometric indicators provided by GS, sometimes comparing them to indicators provided by other databases.
- Open Access: studies that analyse the suitability of GS as a tool to find freely available versions of documents in the Web.
- Unofficial web applications: documents that describe the web applications based on data from GS that have been developed for this thesis.

Table 1. Google Scholar-related studies that my colleagues and I have carried out over the last five years (2014-2018)

Doc. type	Included in this thesis	Reference (sorted by date of publication except when related to previous working paper)	Topics covered					
			Description of GS as platform	Size & coverage of GS	Errors / limitations of GS	Indicators in GS	Open Access	Unofficial web applications
Working paper	No	Martín-Martín, A., Ayllón, J. M., Orduña-Malea, E., & Delgado-López-Cózar, E. (2014). **Google Scholar Metrics 2014: a low cost bibliometric tool** (EC3 Working Papers No. 17). Retrieved from http://arxiv.org/abs/1407.2827	X					
Working paper	No	Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado-López-Cózar, E. (2014). **About the size of Google Scholar: playing the numbers** (EC3 Working Papers No. 18). Retrieved from http://arxiv.org/abs/1407.6239		X				
Journal article	No	↳ Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). **Methods for estimating the size of Google Scholar**. Scientometrics, 104(3), 931–949.		X				
Working paper	**Ch. 2**	Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2014). **Does Google Scholar contain all highly cited documents (1950-2013)?** (EC3 Working Papers No. 19). Retrieved from http://arxiv.org/abs/1410.8464	X	X	X	X	X	
Letter to the editor	**Ch. 4**	↳ Martín-Martín, A., Ayllón, J. M., Delgado López-Cózar, E., & Orduna-Malea, E. (2015). **Nature 's top 100 Re-revisited**. Journal of the Association for Information Science and Technology, 66(12), 2714–2714.			X			
Journal article	**Ch. 3**	↳ Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). **A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013)**. Revista Española de Documentacion Cientifica, 39(4), e149.		X			X	
Journal article	**Ch. 5**	↳ Martin-Martin, A., Orduna-Malea, E., Harzing, A.-W., & Delgado López-Cózar, E. (2017). **Can we use Google Scholar to identify highly-cited documents?** Journal of Informetrics, 11(1), 152–163.	X					
Book	No	Orduña-Malea, E., Martín-Martín, A., Ayllón, J. M., & Delgado López-Cózar, E. (2016). **La revolución Google Scholar: Destapando la caja de Pandora académica.** Granada: Universidad de Granada y Unión de Editoriales Universitarias Españolas.	X	X	X	X	X	X
Working paper	**Ch. 10**	Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2016). **The counting house, measuring those who count: Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in GSC (Google Scholar Citations), ResearcherID, ResearchGate, Mendeley, & Twitter** (EC3 Working Papers No. 21). Retrieved from https://arxiv.org/abs/1602.02412		X	X	X		
Journal article	**Ch. 12**	↳ Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). **A novel method for depicting academic disciplines through Google Scholar Citations: The case of Bibliometrics**. Scientometrics, 114(3), 1251–1273.		X		X		
Journal article	**Ch. 13**	↳ Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). **Author-level metrics in the new academic profile platforms: The online behaviour of the Bibliometrics community**. Journal of Informetrics, 12(2), 494–509.				X		
Journal article	No	Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2017). **Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors**. Revista Española de Documentación Científica, 40(4), e185.			X			

Doc. type	Included in this thesis	Reference (sorted by date of publication except when related to previous working paper)	Topics covered					
			Description of GS as platform	Size & coverage of GS	Errors / limitations of GS	Indicators in GS	Open Access	Unofficial web applications
Journal article	No	Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2017). **The lost academic home: institutional affiliation links in Google Scholar Citations**. Online Information Review, 41(6), 762–781.	X					
Book chapter	No	Delgado López-Cózar, E., Orduna-Malea, E., Martín-Martín, A., & Ayllón, J. M. (2017). **Google Scholar : The Big Data Bibliographic Tool**. In F. J. Cantu-Ortiz (Ed.), Research analytics : boosting university productivity and competitiveness through scientometrics (pp. 59–80). Boca Raton, FL: CRC Press.	X	X				X
Conf. paper	**Ch. 9**	Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017). **Journal Scholar Metrics: building an Arts, Humanities, and Social Sciences journal ranking with Google Scholar data**. In 22nd International Conference on Science, Technology & Innovation Indicators (STI). Paris.						X
Conf. paper	**Ch. 11**	Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017). **Scholar Mirrors: Integrating evidence of impact from multiple sources into one platform to expedite researcher evaluation**. In 22nd International Conference on Science, Technology & Innovation Indicators (STI). Paris.						X
Journal article	No	Delgado López-Cózar, E., & Martín-Martín, A. (2018). **Apagón digital de la producción científica española en Google Scholar [Digital blackout of Spanish scientific production in Google Scholar]**. Anuario ThinkEPI, 12, 265–276.			X			
Journal article	**Ch. 16**	Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). **Evidence of open access of scientific publications in Google Scholar: A large-scale analysis**. Journal of Informetrics, 12(3), 819–841.		X			X	
Journal article	**Ch. 6**	Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). **Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison**. Scientometrics, 116(3), 2175–2188.		X		X		
Journal article	**Ch. 7**	Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). **Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories**. Journal of Informetrics, 12(4), 1160–1177.		X		X		
Conf. paper	No	Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2018). **Classic papers: using Google Scholar to detect the highly-cited documents**. In 23rd International Conference on Science and Technology Indicators (pp. 1298–1307). Leiden.	X					
Book chapter	No	Delgado López-Cózar, E., Orduna-Malea, E., & Martín-Martín, A. (2019). **Google Scholar as a data source for research assessment**. In W. Glaenzel, H. Moed, U. Schmoch, & M. Thelwall (Eds.), Springer Handbook of Science and Technology Indicators. Springer.	X	X	X	X		
Not published	**Ch. 14**	Web application that displays exhaustive and detailed bibliographic and bibliometric data about researchers working in Spain who have a Google Scholar Citations profile (and their publications)						X
Not published	**Ch. 15**	Description of a web application that presents data on Open Access of scientific publications at various levels of aggregation, based on data from Google Scholar						X

References

Abad-García, M.-F., González-Teruel, A., & González-Llinares, J. (2018). Effectiveness of OpenAIRE, BASE, Recolecta, and Google Scholar at finding spanish articles in repositories. *Journal of the Association for Information Science and Technology*, *69*(4), 619–622. https://doi.org/10.1002/asi.23975

Acharya, A. (2015, September 21). What happens when your library is worldwide and all articles are easy to find? Retrieved from https://youtu.be/S-f9MjQjLsk

Acharya, A., Verstak, A., Suzuki, H., Henderson, S., Iakhiaev, M., Lin, C. C. Y., & Shetty, N. (2014). Rise of the Rest: The Growing Impact of Non-Elite Journals. Retrieved from http://arxiv.org/abs/1410.2217

Adriaanse, L. S., Rensleigh, C., & Rensleigh, C. (2011). Comparing Web of Science, Scopus and Google Scholar from an Environmental Sciences perspective. *South African Journal of Libraries and Information Science*, *77*(2). https://doi.org/10.7553/77-2-58

Aguillo, I. F. (2011, December 21). Is Google Scholar useful for bibliometrics? A webometric analysis. https://doi.org/10.1007/s11192-011-0582-8

Amara, N., & Landry, R. (2012). Counting citations in the field of business and management: why use Google Scholar rather than the Web of Science. *Scientometrics*, *93*(3), 553–581. https://doi.org/10.1007/s11192-012-0729-2

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, *3*(1), 7. https://doi.org/10.1186/1742-5581-3-7

Bar-Ilan, J. (2006). An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. *Information Processing & Management*, *42*(6), 1553–1566. https://doi.org/10.1016/j.ipm.2006.03.019

Bar-Ilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics*, *82*(3), 495–506. https://doi.org/10.1007/s11192-010-0185-9

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A., & Daniel, H.-D. (2009). Convergent validity of bibliometric Google Scholar data in the field of chemistry—Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, S. *Journal of Informetrics*, *3*(1), 27–35. https://doi.org/10.1016/j.joi.2008.11.001

Bosman, J., & Kramer, B. (2016). Innovations in scholarly communication - data of the global 2015-2016 survey. https://doi.org/10.5281/ZENODO.49583

Brantley, P. (2007). Science Direct-ly into Google. *TOC: Tools of Change for Publishing*. Retrieved from http://toc.oreilly.com/2007/07/science-directly-into-google.html

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Seventh International World-Wide Web Conference (WWW 1998).* Retrieved from http://ilpubs.stanford.edu:8090/361/

Butler, D. (2004). Science searches shift up a gear as Google starts Scholar engine. *Nature*, *432*(7016), 423–423. https://doi.org/10.1038/432423a

Cabezas-Clavijo, A., & Delgado-López-Cózar, E. (2013). Google Scholar and the h-index in biomedicine: The popularization of bibliometric assessment. *Medicina Intensiva (English Edition)*, *37*(5), 343–354. https://doi.org/10.1016/j.medine.2013.05.002

Chen, X. (2013). Google Scholar's Dramatic Coverage Improvement Five Years after Debut a. *Serials*

Review, *36*(4), 221–226. https://doi.org/10.1080/00987913.2010.10765321

Christianson, M. (2007). Ecology Articles in Google Scholar: Levels of Access to Articles in Core Journals. *Issues in Science and Technology Librarianship*. https://doi.org/10.5062/F4MS3QPD

Clarivate Analytics. (2015). Web of Science & Google Scholar collaboration. Retrieved June 5, 2018, from http://wokinfo.com/googlescholar/

de Winter, J. C. F., Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics*, *98*(2), 1547–1565. https://doi.org/10.1007/s11192-013-1089-2

Delgado-López-Cózar, E., & Repiso-Caballero, R. (2013). The Impact of Scientific Journals of Communication: Comparing Google Scholar Metrics, Web of Science and Scopus. *Comunicar*, *21*(41), 45–52. https://doi.org/10.3916/C41-2013-04

Delgado López-Cózar, E., Orduna-Malea, E., & Martín-Martín, A. (2019). Google Scholar as a data source for research assessment. In W. Glaenzel, H. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*. Springer.

Delgado López-Cózar, E., Orduna-Malea, E., Martín-Martín, A., & Ayllón, J. M. (2017). Google Scholar : The Big Data Bibliographic Tool. In F. J. Cantu-Ortiz (Ed.), *Research analytics : boosting university productivity and competitiveness through scientometrics* (pp. 59–80). Boca Raton, FL: CRC Press.

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446–454. https://doi.org/10.1002/asi.23056

Elsevier. (2004). Scopus comes of age. Retrieved November 19, 2018, from https://www.elsevier.com/about/press-releases/science-and-technology/scopus-comes-of-age

Franceschet, M. (2009). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, *83*(1), 243–258. https://doi.org/10.1007/s11192-009-0021-2

García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in Psychology. *Journal of the American Society for Information Science and Technology*, *61*(10), 2070–2085. https://doi.org/10.1002/asi.21372

Giles, J. (2005). Start your engines. *Nature*, *438*(7068), 554–555. https://doi.org/10.1038/438554a

Giustini, D. (2005). How Google is changing medicine. *BMJ (Clinical Research Ed.)*, *331*(7531), 1487–1488. https://doi.org/10.1136/bmj.331.7531.1487

Gusenbauer, M. (2018). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 1–38. https://doi.org/10.1007/s11192-018-2958-5

Halevi, G., Moed, H., & Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the Literature. *Journal of Informetrics*, *11*(3), 823–834. https://doi.org/10.1016/J.JOI.2017.06.005

Harzing, A.-W. (2013). A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. *Scientometrics*, *94*(3), 1057–1075. https://doi.org/10.1007/s11192-012-0777-7

Harzing, A.-W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics*, *98*(1), 565–575. https://doi.org/10.1007/s11192-013-0975-y

Harzing, A.-W., & Mijnhardt, W. (2015). Proof over promise: towards a more inclusive ranking of Dutch academics in Economics &amp; Business. *Scientometrics*, *102*(1), 727–749. https://doi.org/10.1007/s11192-014-1370-z

Harzing, A. W. (2007). Publish or Perish. Retrieved from http://www.harzing.com/pop.htm

Harzing, A. W. K., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, *8*(1), 61–73. https://doi.org/10.3354/esep00076

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hodge, D. R., & Lacasse, J. R. (2011). Ranking disciplinary journals with the Google Scholar h-index: A new tool for construction cases for tenure, promotion, and other professional decisions. *Journal of Social Work Education*, *47*(3), 579–596. https://doi.org/10.5175/JSWE.2011.201000024

Howland, J. L., Howell, S., Wright, T. C., & Dickson, C. (2009). Google Scholar and the Continuing Education Literature. *The Journal of Continuing Higher Education*, *57*(1), 35–39. https://doi.org/10.1080/07377360902806890

Jacimovic, J., Petrovic, R., & Zivkovic, S. (2010). A citation analysis of Serbian Dental Journal using Web of Science, Scopus and Google Scholar. *Stomatoloski Glasnik Srbije*, *57*(4), 201–211. https://doi.org/10.2298/SGS1004201J

Jacobs, J. A. (2009). Where Credit Is Due: Assessing the Visibility of Articles Published in Gender & Society with Google Scholar. *Gender & Society*, *23*(6), 817–832. https://doi.org/10.1177/0891243209351029

Jacsó, P. (2005a). As we may search — Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*. Current Science Association. https://doi.org/10.2307/24110924

Jacsó, P. (2005b). Comparison and Analysis of the Citedness Scores in Web of Science and Google Scholar. In *International Conference on Asian Digital Libraries* (pp. 360–369). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11599517_41

Jacsó, P. (2005c). Google Scholar: the pros and the cons. *Online Information Review*, *29*(2), 208–214. https://doi.org/10.1108/14684520510598066

Jacsó, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review*, *30*(3), 297–309. https://doi.org/10.1108/14684520610675816

Jacsó, P. (2008a). Google Scholar revisited. *Online Information Review*, *32*(1), 102–114. https://doi.org/10.1108/14684520810866010

Jacsó, P. (2008b). The pros and cons of computing the h-index using Google Scholar. *Online Information Review*, *32*(3), 437–452. https://doi.org/10.1108/14684520810889718

Jacsó, P. (2009). Calculating the h-index and other bibliometric and scientometric indicators from Google Scholar with the Publish or Perish software. *Online Information Review*, *33*(6), 1189–1200. https://doi.org/10.1108/14684520911011070

Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, *34*(1), 175–191. https://doi.org/10.1108/14684521011024191

Jacsó, P. (2012a). Google Scholar Author Citation Tracker: is it too little, too late? *Online Information Review*, *36*(1), 126–141. https://doi.org/10.1108/14684521211209581

Jacsó, P. (2012b). Google Scholar Metrics for Publications. *Online Information Review*, *36*(4), 604–619. https://doi.org/10.1108/14684521211254121

Jamali, H. R., & Nabavi, M. (2015). Open access and sources of full-text articles in Google Scholar in different subject fields. *Scientometrics*, *105*(3), 1635–1651. https://doi.org/10.1007/s11192-015-1642-2

Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. *PloS One*, *9*(5), e93949. https://doi.org/10.1371/journal.pone.0093949

Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, *74*(2), 273–294. https://doi.org/10.1007/s11192-008-0217-x

Kulkarni, A. V, Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *JAMA*, *302*(10), 1092–1096. https://doi.org/10.1001/jama.2009.1307

Laakso, M., & Lindman, J. (2016). Journal copyright restrictions and actual open access availability: a study of articles published in eight top information systems journals (2010–2014). *Scientometrics*, *109*(2), 1167–1189. https://doi.org/10.1007/s11192-016-2078-z

Laakso, M., & Polonioli, A. (2018). Open access in ethics research: an analysis of open access availability and author self-archiving behaviour in light of journal copyright restrictions. *Scientometrics*, 1–27. https://doi.org/10.1007/s11192-018-2751-5

Lasda Bergman, E. M. (2012). Finding Citations to Social Work Literature: The Relative Benefits of Using Web of Science, Scopus, or Google Scholar. *The Journal of Academic Librarianship*, *38*(6), 370–379. https://doi.org/10.1016/j.acalib.2012.08.002

Levy, S. (2014). The Gentleman who Made Scholar. *Wired*. Retrieved from https://www.wired.com/2014/10/the-gentleman-who-made-scholar

Martell, C., & Martell, C. (2009). A Citation Analysis of College & Research Libraries Comparing Yahoo, Google, Google Scholar, and ISI Web of Knowledge with Implications for Promotion and Tenure. *College & Research Libraries*, *70*(5), 460–473. https://doi.org/10.5860/0700460

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, *12*(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012

Martín-Martín, A., & Delgado López-Cózar, E. (2018). Google Scholar's citation graph: comprehensive, global… and inaccessible. *Open Citations Seminar Organised by Uppsala Universitet*. Retrieved from http://hdl.handle.net/10481/51153

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2016). Back to the past : on the shoulders of an academic search engine giant. *Scientometrics*, *107*(3), 1477–1487. https://doi.org/10.1007/s11192-016-1917-2

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentacion Cientifica*, *39*(4), e149. https://doi.org/10.3989/redc.2016.4.1405

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

McCullagh, D. (2006, July 20). Google Scholar trademark case ends. *ZDNet*. Retrieved from https://www.zdnet.com/article/google-scholar-trademark-case-ends/

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105–2125. https://doi.org/10.1002/asi.20677

Mehta, A. (2004, December 10). ACS Takes Legal Action Against Google. *Chemical & Engineering News*. Retrieved from http://pubs.acs.org/cen/news/8250/8250acs.html

Mikki, S. (2009). Comparing Google Scholar and ISI Web of Science for Earth Sciences. *Scientometrics*, *82*(2), 321–331. https://doi.org/10.1007/s11192-009-0038-6

Mikki, S., Ruwehy, H. A. Al, Gjesdal, Ø. L., & Zygmuntowska, M. (2018). Filter bubbles in interdisciplinary research: a case study on climate and society. *Library Hi Tech*, LHT-03-2017-0052. https://doi.org/10.1108/LHT-03-2017-0052

Minasny, B., Hartemink, A. E., McBratney, A., & Jang, H.-J. (2013). Citations and the h index of soil researchers and journals in the Web of Science, Scopus, and Google Scholar. *PeerJ*, *1*, e183. https://doi.org/10.7717/peerj.183

Mingers, J., & Lipitakis, E. A. E. C. G. (2010). Counting the citations: a comparison of Web of Science and Google Scholar in the field of business and management. *Scientometrics*, *85*(2), 613–625. https://doi.org/10.1007/s11192-010-0270-0

Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics*, *10*(2), 533–551. https://doi.org/10.1016/j.joi.2016.04.017

Mussell, J., & Croft, R. (2013). Discovery Layers and the Distance Student: Online Search Habits of Students. *Journal of Library & Information Services in Distance Learning*, *7*(1–2), 18–39. https://doi.org/10.1080/1533290X.2012.705561

Nicholas, D., Boukacem-Zeghmouri, C., Rodríguez-Bravo, B., Xu, J., Watkinson, A., Abrizah, A., … Świgoń, M. (2017). Where and how early career researchers find scholarly information. *Learned Publishing*, *30*(1), 19–29. https://doi.org/10.1002/leap.1087

Norris, M., Oppenheim, C., & Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the American Society for Information Science and Technology*, *59*(12), 1963–1972. https://doi.org/10.1002/asi.20898

Notess, G. R. (2005). Scholarly Web Searching: Google Scholar and Scirus. *Information Today*. Retrieved from http://www.infotoday.com/online/jul05/onthenet.shtml

Ocholla, D., & Onyancha, O. B. (2009). Assessing researchers' performance in developing countries : is Google Scholar an alternative? *Mousaion*, *27*(1), 43–64. Retrieved from http://uir.unisa.ac.za/handle/10500/5269

Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado-López-Cózar, E. (2014). *About the size of Google Scholar: playing the numbers* (EC3 Working Papers No. 18). Retrieved from http://arxiv.org/abs/1407.6239

Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, *104*(3), 931–949. https://doi.org/10.1007/s11192-015-1614-6

Orduña-Malea, E., Martín-Martín, A., Ayllón, J. M., & Delgado López-Cózar, E. (2016). *La revolución Google Scholar : Destapando la caja de Pandora académica*. Granada: Universidad de Granada y Unión de Editoriales Universitarias Españolas.

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2018). Classic papers: using Google Scholar to detect the highly-cited documents. In *23rd International Conference on Science and Technology Indicators* (pp. 1298–1307). Leiden. https://doi.org/10.31235/osf.io/zkh7p

Ortega, J. L. (2014). *Academic search engines : a quantitative outlook*. Chandos Publishing.

Pauly, D., & Stergiou, K. (2005). Equivalence of results from two citation analyses: Thomson ISI's Citation Index and Google's Scholar service. *Ethics in Science and Environmental Politics*, *9*, 33–35. https://doi.org/10.3354/esep005033

Pitol, S. P., & De Groote, S. L. (2014). Google Scholar versions: do more versions of an article mean greater impact? *Library Hi Tech*, *32*(4), 594–611. https://doi.org/10.1108/LHT-05-2014-0039

Rosenstreich, D., & Wooliscroft, B. (2009). Measuring the impact of accounting journals using Google Scholar and the g-index. *The British Accounting Review*, *41*(4), 227–239. https://doi.org/10.1016/J.BAR.2009.10.002

S. Adriaanse, L., & Rensleigh, C. (2013). Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. *The Electronic Library*, *31*(6), 727–744. https://doi.org/10.1108/EL-12-2011-0174

Science. (2004). NET NEWS: A Google for Academia. *Science*, *306*(5702), 1661.3-1661. https://doi.org/10.1126/science.306.5702.1661c

Sember, M., Utrobicić, A., & Petrak, J. (2010). Croatian Medical Journal citation score in Web of Science, Scopus, and Google Scholar. *Croatian Medical Journal*, *51*(2), 99–103. https://doi.org/10.3325/CMJ.2010.51.99

Teplitzky, S. (2017). Open Data, [Open] Access: Linking Data Sharing and Article Sharing in the Earth Sciences. *Journal of Librarianship and Scholarly Communication*, *5*(General Issue), eP2150. https://doi.org/10.7710/2162-3309.2150

Van Noorden, R. (2014a). Online collaboration: Scientists and the social network. *Nature*, *512*(7513), 126–129. https://doi.org/10.1038/512126a

Van Noorden, R. (2014b, November 7). Google Scholar pioneer on search engine's future. *Nature*. https://doi.org/10.1038/nature.2014.16269

Verstak, A., Acharya, A., Suzuki, H., Henderson, S., Iakhiaev, M., Lin, C. C. Y., & Shetty, N. (2014). *On the Shoulders of Giants: The Growing Impact of Older Articles*. Retrieved from http://arxiv.org/abs/1411.0275

Vine, R. (2005). Google Scholar is a full year late indexing Pubmed content. Retrieved from http://web.archive.org/web/20060716085124/
http://www.workingfaster.com/sitelines/archives/2005_02.html

Vine, R. (2006). Google Scholar. *Journal of the Medical Library Association*, *94*(1), 97. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1324783/

Walters, W. H. (2007). Google Scholar coverage of a multidisciplinary field. *Information Processing & Management*, *43*(4), 1121–1132. https://doi.org/10.1016/J.IPM.2006.08.006

Wang, Y., & Howard, P. (2012). Google Scholar Usage: An Academic Library's Experience. *Journal of Web Librarianship*, *6*(2), 94–108. https://doi.org/10.1080/19322909.2012.672067

Wentz, R. (2004). WoS versus Google Scholar: Cited by...: Correction. *Medical Libraries Discussion List*. Retrieved from https://web.archive.org/web/20070630064521/http://listserv.acsu.buffalo.edu/cgi-bin/wa?A2=ind0412B&L=medlib-l&P=R5842&I=-3&m=95812

Yang, K., & Meho, L. I. (2007). Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–15. https://doi.org/10.1002/meet.14504301185

Zarifmahmoudi, L., Kianifar, H. R., & Sadeghi, R. (2013). Citation Analysis of Iranian Journal of Basic Medical Sciences in ISI Web of Knowledge, Scopus, and Google Scholar. *Iranian Journal of Basic Medical Sciences*, *16*(10), 1027–1030. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3874088&tool=pmcentrez&rendertype=abstract

Zarifmahmoudi, L., & Sadeghi, R. (2012). Comparison of ISI web of knowledge, SCOPUS, and Google scholar h-indices of Iranian nuclear medicine scientists. *Iranian Journal of Nuclear Medicine*, *20*(1), 1–4.

Introducción

Descripción de Google Scholar

Google Scholar (GS) es un motor de búsqueda académico gratuito que indiza literatura académica de un amplio rango de disciplinas, tipos documentales, e idiomas. Su objetivo principal es ayudar a los usuarios a encontrar información académica relevante. También proporciona otros servicios adicionales, como la generación automática de referencias bibliográficas en varios estilos y formatos, acceso al texto completo de los documentos académicos (cuando están libremente accesibles en la web, o cuando se detecta que el usuario tiene acceso a estos documentos a través de las subscripciones institucionales), así como el cálculo del número de veces que un documento ha sido citado en otros documentos. A diferencia del motor de búsqueda general de Google y de otros productos desarrollados por la misma compañía, GS nunca ha sido monetizado. La plataforma nunca ha mostrado publicidad y, de acuerdo con su fundador, tampoco ha recogido datos de usuarios individuales que pudieran ser reutilizado de manera comercial (Acharya, 2015). Por tanto, GS puede ser considerado como un proyecto sin ánimo de lucro subvencionado por una empresa con ánimo de lucro.

GS fue desarrollado por los ingenieros de Google Anurag Acharya y Alex Verstak durante un periodo sabático. Ambos habían pasado algún tiempo en el mundo académico, y conocían las dificultados que afrontan los investigadores cuando necesitan localizar información científica, especialmente en países en desarrollo. Acharya y Verstak presentaron la idea a Larry Page, y GS fue lanzado el 18 de noviembre de 2004. De esta manera Google, una empresa cuyo éxito inicial se había derivado de un algoritmo inspirado en la literatura científica del análisis de citas (Brin & Page, 1998), se embarcó en un proyecto que aplicaba ideas similares a los documentos académicos.

GS fue innovador en varios sentidos. En primer lugar, a contrario que la Web of Science (WoS) y Scopus, que tomaban un enfoque selectivo hacia la indización de documentos (solo indizan documentos publicados en ciertas fuentes seleccionadas), GS tomó un enfoque inclusivo, indizando cualquier documento con aspecto académico que sus rastreadores automáticos pueden encontrar en la web. Además de eso, aunque la mayoría de las demás bases de datos académicas solo trabajaban con metadatos (información sobre los autores, título, resumen, palabras clave… de los documentos), GS se desmarcó al indizar el texto completo de los documentos siempre que fuese posible (cuando el texto completo está accesible gratuitamente, o cuando GS llegaba a acuerdos con editoriales para indizar sus contenidos de pago). De esta manera, las consultas de los usuarios no solo se confrontaban con un índice selectivo de metadatos, sino con un índice exhaustivo de metadatos y textos completos. Una descripción más detallada del funcionamiento general de GS puede encontrarse en Delgado López-Cózar, Orduna-Malea, Martín-Martín, & Ayllón (2017) y en Orduña-Malea, Martín-Martín, Ayllón, & Delgado López-Cózar (2016, Chapter 4). Otro aspecto que diferenciaba a GS de la mayoría de las otras bases de datos académicas era su simple interfaz de búsqueda. Mientras que la mayoría de las bases de datos utilizaban avanzadas interfaces donde los usuarios tienen que diseñar cuidadosamente cada aspecto de la consulta (campos, operadores, opciones de filtrado…) GS replicó la simplicidad del buscador general de Google: una caja de búsqueda, complementada por unas opciones de búsqueda avanzada y filtrado muy limitadas. Una descripción más detallada de las opciones de búsqueda en GS se puede encontrar en Orduña-Malea et al. (2016, Chapter 6).

El poder acceder al texto completo de los documentos publicados por editoriales de todo tipo (comerciales, no comerciales, universitarias, de sociedades científicas…) así como a los documentos disponibles en otros canales de comunicación como repositorios y las páginas web personales de los investigadores, permitió a GS construir su propia red de citas (o grafo de citas), mediante el procesamiento de las referencias que se encuentran al final de cada documento y su matching (emparejamiento) con los documentos ya contenidos en su índice. Este grafo de citas fue clave para poder implementar su algoritmo de ordenación de resultados por relevancia. Mientras tanto, la mayoría de las otras bases de datos solo permitían ordenar los documentos mostrados en una búsqueda de maneras más simples (por apellido del autor, por título del documento, por fecha de publicación). De las bases de datos académicas multidisciplinares, solo WoS y Scopus (ambas solo accesibles mediante suscripción) tenían la opción de ordenar documentos por el número de citas. Durante muchos años, GS ha sido la única base de datos multidisciplinar con una cobertura extensa y que proporcionaba datos de citas de manera gratuita.

Su éxito como motor de búsqueda fue inmediato, a pesar de ser lanzado solo unas pocas semanas después que la base de datos Elsevier de Scopus (3 de noviembre de 2004) (Elsevier, 2004). Antes de que acabara el año, tanto *Science* (Science, 2004) como *Nature* (Butler, 2004) ya habían informado de la llegada de este nuevo motor de búsqueda. De acuerdo con Giles (2005), solo un año después de su lanzamiento GS ya dirigía más tráfico a *Nature* que cualquier otro motor de búsqueda académico multidisciplinar. De igual manera, Giustini (2005) inform que Google y GS eran los sitios que dirigían más tráfico al *British Medical Journal*, seguidos de lejos por Yahoo y Pubmed-Medline. Wang & Howard (2012) encontraron que entre 2006 y 2011 el uso de GS por los usuarios de la San Francisco State University multiplicó por 10, convirtiéndose en la primera fuente desde la que le llegaban peticiones a la biblioteca en 2011, por encima de la propia herramienta de búsqueda de la universidad.

Numerosas encuestas muestran resultados en el mismo sentido: GS se ha convertido en la herramienta más usada por muchos investigadores para buscar información académica. Van Noorden (2014) informaba de que de acuerdo a los resultados de una encuesta completada por más de 3,000 científicos e ingenieros, GS era el sitio más visitado. Mussell & Croft (2013) y Nicholas et al. (2017) encontraron resultados similares cuando estudiaron los hábitos de búsqueda de estudiantes e investigadores jóvenes, respectivamente. Por su parte, Bosman & Kramer (2016) encontraron que GS era de lejos la herramienta más usada para buscar literatura científica (con más del doble de encuestados seleccionando esta herramienta, respecto a la siguiente opción). Sus datos también muestran que GS también era la herramienta más usada para crear alertas informativas. En 2015, WoS reconoció que “Google Scholar es cada vez frecuentemente el punto de partida para los investigadores” (Clarivate Analytics, 2015) y por esta razón entró en un acuerdo con GS para que el número de citas según WoS (así como un enlace a la plataforma) apareciera embebido en las páginas de resultados de GS (solo para aquellos usuarios con suscripción a WoS). A pesar de su continuo crecimiento y éxito, en 2014 se informó de que solo nueve personas trabajaban directamente en Google Scholar (Van Noorden, 2014b).

Además del buscador académico, en 2011 GS lanzó un servicio de perfiles de autor llamado Google Scholar Citations (GSC)[8]. Durante unos meses, este servicio solo pudo ser usado por un número limitado de usuarios beta, pero el 16 de noviembre del mismo año fue abierto a todos los usuarios. Con este servicio, GS añadió algunas funciones importantes a su porfolio: aprovechando la extensa cobertura de documentos, a partir de ese momento los usuarios pudieron crear perfiles personales que listaran todas sus publicaciones fácilmente. El número de citas de cada documento también era visible, y varios indicadores bibliométricos a nivel de autor se calculaban automáticamente: total de citas recibidas por un autor, índice h (Hirsch, 2005), e índice i10 (un indicador inventado por GS, que se define como el número de documentos en el perfil con al menos 10 citas). Cada uno de los tres indicadores se calcula dos veces: en una versión no se aplica ninguna restricción en la ventana de citación, mientras que en otra versión se usa una ventana de citación de cinco años (los últimos cinco años completos). El servicio de perfiles da libertad completa al usuario para añadir, editar, unir, y eliminar documentos del perfil, sin ningún control externo. Los usuarios pueden configurar alertas para ser notificados cuando sus propios perfiles, o los de otros autores, se actualicen (ya sea por la adición de nuevos documentos, o por la identificación de nuevas citas). Según Bosman & Kramer (2016), GSC era la segunda herramienta de perfiles de autor más usada (ligeramente

[8] https://scholar.googleblog.com/2011/07/google-scholar-citations.html

por detrás de ResearchGate). Una descripción más detallada de GSC se puede encontrar en Orduña-Malea, Martín-Martín, Ayllón, & Delgado López-Cózar (2016, Chapter 9).

Aproximadamente un año después, en abril de 2012[9], GS sorprendió de nuevo a la comunidad científica al lanzar Google Scholar Metrics for Publications (GSM), un ranking de revistas científicas (aunque también se incluyen algunas conferencias y repositorios). En un primer nivel, las revistas están clasificadas según su idioma de publicación. Para todos los idiomas excepto el inglés, se muestra el top 100 de revistas con un mayor índice h5 (índice h de los documentos publicados en los últimos cinco años). Para las revistas, conferencias y repositorios donde se publica en inglés, el producto ofrece una clasificación de fuentes de acuerdo a su temática. La clasificación temática tiene dos niveles: un primer nivel con 8 áreas generales (Business, Economics & Management; Chemical & Material Sciences; Engineering & Computer Science; Health & Medical Sciences; Humanities, Literature & Arts; Life Sciences & Earth Sciences; Physics & Mathematics; Social Sciences), y un segundo nivel con aproximadamente 250 subcategorías (cambia ligeramente cada año). Dentro de cada categoría general o subcategoría se muestran el top 20 de las fuentes con un mayor índice h5. Además de las fuentes que se encuentran en estos listados por idioma o por categoría, GSM ofrece información sobre otras muchas fuentes, que se pueden encontrar utilizando la herramienta de búsqueda. Para cada revista, congreso, o repositorio, GSM puede mostrar el listado de documentos que se encuentran por encima de la barrera del índice h5, así como los listados de documentos citantes. Sus criterios de inclusión establecen que para que una fuente sea incluida, debe haber publicado al menos 100 documentos en los últimos cinco años, y recibido al menos una cita. Nuevas ediciones de GSM se han publicado una vez al año desde 2012. Una descripción más detallada del servicio se puede encontrar en Orduña-Malea et al. (2016, Chapter 8).

En 2017 GS lanzó un servicio adicional, aunque aparentemente efímero (Orduna-Malea, Martín-Martín, & Delgado López-Cózar, 2018): Classic Papers: Articles that have stood the test of time[10] (GSCP). Este servicio se lanzó el 14 de junio de 2017, poco antes de la actualización de GSM de 2017. Sin embargo, aunque la edición de 2018 de GSM fue lanzada en agosto de 2018, no sucedió lo mismo con GSCP. Es más, aunque el servicio todavía está accesible[11], el enlace para acceder al mismo ha sido eliminado de la sección Metrics de GS, por lo que es virtualmente inaccesible a cualquiera que no conozca ya su existencia. Este servicio proporciona listas de los diez documentos más citados publicados en inglés en 2006 (proporcionando así una ventana de citación de diez años) a nivel de categorías temáticas. La estructura de las categorías es la misma que la que se utiliza en GSM (dos niveles: ocho categorías generales, y 252 subcategorías), pero al contrario que en GSM, la clasificación se realizó a nivel de los documentos (los artículos publicados en revistas multidisciplinares fueron clasificados en sus respectivas categorías). En total, en este servicio se pueden visualizar 2,515 documentos. La razón por la que no son 2,520 es que el servicio establece un mínimo de 20 citas por documento, y en la categoría French Studies solo se pudieron encontrar 5 documentos que cumplieran ese criterio.

El uso generalizado de GS y sus servicios spin-off GSC y GSM espoleó la publicación un gran número de estudios que analizaban sus características, en muchos casos a la vez que se hacían comparaciones con otras bases de datos académicas. En el curso de nuestro trabajo, mis compañeros y yo hemos identificado casi 300 estudios que analizan GS o usan datos de GS de alguna forma[12]. La Figura 1 muestra la distribución de estudios por año de publicación. El pico mayor se alcanzó en 2014 (45 estudios), el año en el que GS celebró su décimo aniversario.

[9] https://scholar.googleblog.com/2012/04/google-scholar-metrics-for-publications.html
[10] https://scholar.googleblog.com/2017/06/classic-papers-articles-that-have-stood.html
[11] https://scholar.google.com/citations?view_op=list_classic_articles&hl=en&by=2006
[12] http://googlescholardigest.blogspot.com/p/bibliography.html

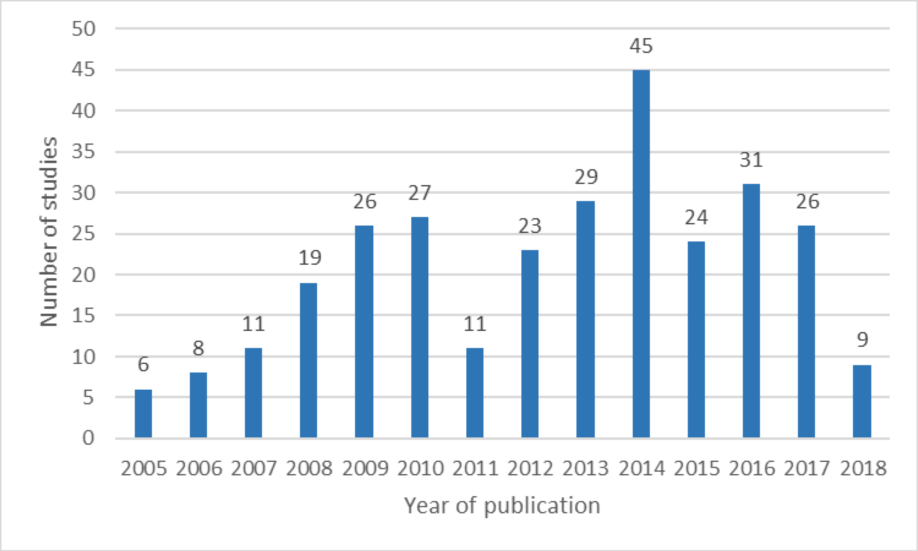


Figura 1. Distribución de estudios sobre Google Scholar, por año de publicación

Los estudios que analizan GS normalmente caen en uno de dos posibles grupos. Primero están aquellos que analizan GS como una herramienta para buscar literatura científica. Segundo, aquellos que están interesados en GS como una fuente de información bibliográfica y de citas que puede ser usada para llevar a cabo análisis bibliométricos. En esta tesis estamos principalmente interesados en el Segundo grupo. Una descripción extensa sobre los estudios que hay sobre GS como herramienta de búsqueda se puede encontrar en Orduña-Malea et al. (2016, Chapter 2).

Aunque GS es un servicio al que se puede acceder gratuitamente, en ningún momento se ha proporcionado una manera adecuada (ni gratuita, ni de pago) para exportar datos de manera masiva, lo cual sería necesario para llevar a cabo estudios bibliométrics. Al contrario que Clarivate Analytics y Elsevier, que además de ofrecer productos para que facilitan el descubrimiento de literatura científica (WoS, Scopus) también venden licencias para reutilizar sus datos con propósitos bibliométricos, y ofrecen productos especialment e diseñados para facilitar estos estudios (InCites, SciVal), el objetivo principal de GS siempre ha sido facilitar la búsqueda de información. GS genera un grafo de citas y proporciona esta información al usuario porque estos datos son considerados útiles para este objetivo principal. Incluso los productos spin-off GSC y GSM, que en principio parecen tener una naturaleza más bibliométrica, cumplen con la función de facilitar el descubrimiento de las personas más relevantes en un tema determinado, así como estar permanentemente informados de sus nuevos trabajos (GSC), o facilitan la identificación de las revistas y documentos más influyentes en un área (GSM). Se ha informado de que una de las barreras que impiden que GS proporcione acceso a los datos de manera masiva son los acuerdos a los que GS tuvo que llegar con las editoriales comerciales para que les dejaran acceder a sus datos (Van Noorden, 2014b). GS pide que sus usuarios no utilicen métodos automáticos para extraer datos de su buscador[13], y hace cumplir esta norma al implementar un sistema de CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart). GS usa la tecnología reCAPTCHA, que fue adquirida por Google en 2009[14]. Este test solía consistir en que los usuarios tenían que escribir en una caja de texto las mismas letras que aparecían en una imagen en la que aparecía un texto deformado. Sin embargo, en los últimos años el test se ha actualizado, y ahora se pide a los usuarios que seleccionen entre una serie de imágenes aquellas que cumplen con algún criterio específico (aquellas en las que aparecen coches, señales de tráfico, carreteras, puentes, sushi…).

[13] https://scholar.google.com/intl/en/scholar/help.html#export

[14] https://googleblog.blogspot.com/2009/09/teaching-computers-to-read-google.html

A pesar de estas dificultades, se desarrollaron una serie de aplicaciones no oficiales para extraer datos de GS con propósitos diversos (Martín-Martín & Delgado López-Cózar, 2018). Sin lugar a dudas, la más conocida de estas aplicaciones es *Publish or Perish* (PoP), desarrollada por Anne-Wil Harzing (Harzing, 2007). La aparición de GS, combinada con la disponibilidad y la facilidad de uso de PoP facilitaron el desarrollo de estudios bibliométricos que usaban datos de GS, contribuyendo así a la democratización del análisis de citas (Harzing & Mijnhardt, 2015; Harzing & van der Wal, 2008), que hasta 2004 había sido un tipo de análisis que solo aquellos con acceso a WoS podían realizar.

Investigaciones sobre Google Scholar como fuente de datos bibliográficos y de citas

Durante los primeros años de vida de GS, Verstak y Acharya dedicaron un esfuerzo considerable a conseguir que las editoriales comerciales accedieran a dejar que los rastreadores de GS indizaran sus páginas web. Ellos entendían que una base de datos académica solo tiene sentido si contiene los documentos que sus usuarios necesitan. El objetivo era que, incluso si no era posible acceder a un documento gratuitamente, los usuarios deberían al menos ser conscientes de que ese documento existe.

Como GS no proporciona una lista pública de las fuentes o documentos que indiza, siempre ha sido difícil saber exactamente a qué acuerdos ha llegado con cada editorial. A pesar de esto, in 2005 ya cubría muchas editorials importantes, como: Association of Computer Machinery (ACM), Blackwell, Institute of Electrical and Electronic Engineers (IEEE), Ingenta, Nature Publishing Group, Springer, Wiley, así como respositorios como NASA Astrophysics Data System, arXiv, o PubMed, aunque no todos ellos de manera exhaustiva (Notess, 2005). Por ejemplo, aunque PubMed ya estaba indizada en 2005, Vine (2005) descubrió que los documentos de PubMed aparecían en GS con un retraso de un año, lo que por supuesto reducía la utilidad de la herramienta para buscar publicaciones recientes.

Además de lo anterior, las negociaciones con editoriales comerciales para que éstas permitieran que GS indizara su contenido fueron a veces complicadas. De acuerdo con Butler (2004), en aquel momento Elsevier había "declinado permitir que Google indizara sus textos". Vine (2006) informó de que los textos completos de los artículos publicados por Elsevier y la American Chemical Society todavía no estaban siendo indizados en GS. No fue hasta 2007 cuando las revistas de Elsevier comenzaron a ser indizadas en GS a través de la plataforma ScienceDirect (Brantley, 2007).

Quizás el caso más famoso de tensión entre una gran editorial comercial y GS fue el de la American Chemical Society (ACS), que denunció a Google por el uso del término "Scholar" en su nuevo buscador académico, porque consideraba que "el uso de la palabra Scholar por Google infringe los derechos de la marca registrada de la ACS 'SciFinder Scholar' [una aplicación de pago para buscar información] y constituye competencia desleal" (Mehta, 2004). El caso se resolvió en 2006, cuando cada parte acordó pagar su parte de las costas (McCullagh, 2006). En parte porque la ACS no permitió en un principio que GS indizara sus artículos, la cobertura documental del campo de la Química en GS fue muy limitada durante un tiempo, como se evidenció en los resultados de Bornmann et al. (2009). Sin embargo, estas deficiencias ya habían sido subsanadas en 2013 (Harzing, 2013b, 2013a).

La inclusión de JSTOR fue otro hito en la historia de GS. Aunque JSTOR era una de las plataformas de suscripción más grandes en términos de número de artículos ofrecidos, los usuarios sin suscripción no tenían acceso a la información básica de los artículos tal como el resumen, en parte porque la mayoría de su biblioteca consistía en artículos escaneados. GS convenció a JSTOR para que al menos mostrara la primera página de cada artículo (la página que normalmente contiene la información bibliográfica básica del artículo y el resumen) a los usuarios sin suscripción (Levy, 2014). Esto permitió que los usuarios de GS pudieran tomar mejores decisiones acerca de qué documentos podrían ser relevantes para ellos.

Desde sus comienzos, GS recibió una mezcla de críticas y comentarios favorables. En diciembre de 2004, Wentz (2004) fue uno de los primeros en declarar que “los datos de citas de GS son poco fiables en el mejor de los casos, y una pérdida de tiempo en el peor”, y que “Google Scholar debería retirar la información de ‘cited by’ de su versión Beta, y probablemente no ofrecerla en la versión final”, todo esto basado en el análisis de varios artículos. En 2005, Peter Jacsò llevó a cabo una serie de comparaciones a pequeña escala entre WoS, Scopus y GS (Jacsó, 2005a). Después de analizar la cobertura de estas tres fuentes para un pequeño número de artículos publicados en *Current Science*, encontró que muchos de ellos no estaban cubiertos por GS, y que aquellos que lo estaban tenían unas cifras de citas inferiores a las proporcionadas por WoS y Scopus. En otro estudio (Jacsó, 2005b) encontró resultados similares para una muestra de artículos del *Asian Pacific Journal of Allergy and Immunology*: WoS proporcionaba cifras de citas significativamente superiores (1,455) que GS (595). Sin embargo, Jacsò también reconoció el mérito de la plataforma como buscador académico (Jacsó, 2005c).

Pauly & Stergiou (2005) encontraron que, aunque GS encontraba menos citas que WoS para documentos antiguos (publicados antes de 1989) en una muestra de 114 artículos de varias áreas científicas, sí encontraba más citas que WoS cuando se analizaban documentos recientes. Además, las correlaciones entre el número de citas eran muy altas (0.84-0.99). Bar-Ilan (2006) analizó la cobertura documental para producción científica de un matemático e informático en WoS, GS, y CiteSeer, e informó que GS encontraba ligeramente más citas que WoS, y que estas dos bases de datos encontraba un número mucho más alto de citas que CiteSeer. Walters (2007) comparó la cobertura de varias bases de datos y concluyó que, para una muestra de 155 artículos del tema de la migración, GS era la herramienta que encontraba un mayor número de ellos (93%).

Jacsó (2006) declaró que los estudios que eran entusiastas con la cobertura de GS eran superficiales, porque no consideraban que GS también cubría tipos documentales no provenientes de revistas, como artículos presentados en conferencias, libros, capítulos, y tesis, mientras que otras bases de datos como WoS no lo hacían. Jacsó también cuestionó la validez de los datos de GS debido a sus errores de “artificial unintelligence”, tales como la creación de emparejamientos incorrectos que incrementaban el número de citas, o devolver más de 40,000 resultados para una búsqueda del autor “I Introduction”.

Errores como los que se describen arriba eran muy comunes durante los primeros años de existencia de GS, cuando la mayoría de los documentos que indizaba no tenían asociados metadatos estructurados, y GS tenía que inferir los metadatos a partir del texto de los pdfs (cada uno con un formato particular, o en el peor de los casos, creados a partir de imágenes escaneadas). En algunos casos, los parsers de GS tomaban el encabezamiento “1. Introduction” erróneamente como el autor (otras veces como el título) del documento, dando lugar al error que Jacsó encontró. Harzing & van der Wal (2008) replicaron esta búsqueda y encontrar que el número de registros había disminuido a 956. Orduña-Malea et al. (2016) hicieron lo mismo en 2015 y encontraron que esta búsqueda solo devolvía cinco registros, confirmando que GS resolvió este problema gradualmente con el tiempo.

Bakkalbasi, Bauer, Glover, & Wang (2006) publicaron el primer estudio que llevó a cabo una comparación de la cobertura cita-por-cita entre WoS, Scopus, y GS. De acuerdo a sus resultados, GS solo encontró el 53% de todas las citas posibles (WoS encontró el 70% y Scopus el 76%). El 13% de las citas solo podían ser encontradas por GS (mientras que WoS tenía un 28% de citas únicas, y Scopus un 31%). A este estudio le siguieron con el tiempo muchos otros de una naturaleza similar (Bar-Ilan, 2010; de Winter, Zadpoor, & Dodou, 2014; Jacimovic, Petrovic, & Zivkovic, 2010; Kousha & Thelwall, 2008; Lasda Bergman, 2012; Meho & Yang, 2007; Moed, Bar-Ilan, & Halevi, 2016; Sember, Utrobicić, & Petrak, 2010; Yang & Meho, 2007). Martín-Martín, Orduna-Malea, Thelwall, & Delgado López-Cózar (2018) (capítulo 7 de esta tesis) proporcionan una revisión más detallada sobre este tema.

Christianson (2007) fue el primero que aprovechó la capacidad de GS de dirigir a sus usuarios a versiones gratuitas de los documentos para analizar el grado de Acceso Abierto a las publicaciones en el campo de la ecología. Ella encontró que el 38% de los artículos estaban disponibles gratuitamente. Muchos otros artículos realizaron análisis similares a este (Abad-García, González-Teruel, & González-Llinares, 2018;

Jamali & Nabavi, 2015; Khabsa & Giles, 2014; Laakso & Lindman, 2016; Laakso & Polonioli, 2018; Martín-Martín, Orduna-Malea, Ayllón, & Delgado López-Cózar, 2016; Mikki, Ruwehy, Gjesdal, & Zygmuntowska, 2018; Norris, Oppenheim, & Rowland, 2008; Pitol & De Groote, 2014; Teplitzky, 2017). Una revisión más detallada de este tema se puede encontrar en Martín-Martín, Costas, Van Leeuwen, & Delgado López-Cózar (2018) (capítulo 16 de esta tesis).

Jacsó (2008) reconoció que GS había mejorado su cobertura en términos de revistas, libros, y otros tipos documentales provenientes de todo el mundo y en muchos idiomas, pero a la vez criticó la inconsistencia de los números de resultados devueltos a una serie de consultas y sus variantes, así como los errores de interpretación de metadatos cometidos por el parser automático. Su crítica de GS continuó en Jacsó Jacsó (2008b, 2009, 2010, 2012a, 2012b).

Harzing & van der Wal (2008) abrieron la puerta al uso de Google Scholar como una fuente de datos para el análisis de citas con la creación del software *Publish or Perish.* Ellos declaraban que GS podía ser usado para análisis de citas en campos que no estuvieran bien cubiertos por los otros índices de citas. Desde ese momento, se empezaron a acumular evidencias que sugerían que la cobertura de GS era normalmente más extensa que la de las bases de datos selectivas (especialmente en las áreas de Humanidades y Ciencias Sociales), y que los números de citas proporcionados por GS eran normalmente superiores que los proporcionados por las otras bases de datos (Amara & Landry, 2012; Bar-Ilan, 2010; Cabezas-Clavijo & Delgado-López-Cózar, 2013; Chen, 2013; de Winter et al., 2014; Delgado-López-Cózar & Repiso-Caballero, 2013; Franceschet, 2009; García-Pérez, 2010; A.-W. Harzing, 2013, 2014; Hodge & Lacasse, 2011; Howland, Howell, Wright, & Dickson, 2009; Jacimovic et al., 2010; Jacobs, 2009; Kousha & Thelwall, 2008; Kulkarni, Aziz, Shams, & Busse, 2009; Lasda Bergman, 2012; Martell & Martell, 2009; Mikki, 2009; Minasny, Hartemink, McBratney, & Jang, 2013; Mingers & Lipitakis, 2010; Ocholla & Onyancha, 2009; Rosenstreich & Wooliscroft, 2009; Sember et al., 2010; Zarifmahmoudi, Kianifar, & Sadeghi, 2013; Zarifmahmoudi & Sadeghi, 2012). También hubo estudios que informaron de resultados contrarios a esta tendencia en algunos campos (Adriaanse, Rensleigh, & Rensleigh, 2011; Bornmann et al., 2009; Adriaanse & Rensleigh, 2013). Aguillo (2011) proporcionó una de las primeras estimaciones del tamaño de GS (86 millones de registros, con datos de 2010) y consideró que a GS le faltaba “el control de calidad necesario para su uso como una herramienta bibliométrica”. Delgado López-Cózar, Robinson-García, & Torres-Salinas (2014) avisaron de la facilidad con la que los conteos de citas en GS podían ser manipulados por cualquier persona.

2014 fue el año en el que GS celebró su décimo aniversario, y por esta ocasión el equipo detrás de GS publicó dos estudios que analizaban el efecto que la web en general, y GS en particular habían tenido sobre la comunicación científica. En Verstak et al. (2014) y Acharya et al. (2014), el equipo de GS proporcionó evidencias de que el porcentaje de citas a documentos antiguos, así como el porcentaje de documentos altamente citados publicados en revistas fuera de la élite, estaban aumentando con el tiempo. Ellos proponían que esto era posible gracias a que en el entorno web, encontrar y leer documentos más antiguos, o documentos publicados en revistas que no son de la élite es tan fácil como encontrar y leer documentos recientes, o documentos de las revistas de élite. Este efecto fue más tarde confirmado usando una muestra de datos de WoS (Martín-Martín, Orduna-Malea, Ayllón, & Delgado-López-Cózar, 2016). En 2014, además, se publicaron varias estimaciones del tamaño de GS: según Khabsa & Giles (2014), que analizaron solo documentos en inglés, su tamaño en aquel momento estaba cerca de los 100 millones de documentos. Ortega (2014), usando datos de 2012, llegó a una cifra similar: 95 millones. Poco después, otro estudio que utilizaba métodos diferentes estimó el tamaño en aproximadamente 160 millones de documentos (Orduña-Malea, Ayllón, Martín-Martín, & Delgado-López-Cózar, 2014). Este último estudio fue replicado en 2017 (Delgado López-Cózar et al., 2019), encontrando que el número de documentos en GS se había incrementado hasta aproximadamente 200 millones (331 millones si se incluían las referencias citadas y las patentes). En 2018, fue replicado de nuevo (Gusenbauer, 2018), encontrando que la cifra había aumentado a 389 millones.

Descripciones más detalladas de todos los estudios que analizan GS como una fuente de datos bibliográficos y de citas pueden encontrarse en Delgado López-Cózar et al., 2019; Halevi, Moed, & Bar-Ilan, 2017; Orduna-Malea, Ayllón, Martín-Martín, & Delgado López-Cózar, 2015.

Objetivos

2014 fue también el año en el que inicié mi formación como doctorando. Según la evidencia disponible hasta el momento, GS parecía ser una fuente de datos bibliográficos y de citas que podría ser útil, especialmente en los campos donde la cobertura de los otros índices de citas como WoS y Scopus era patentemente insuficiente, como en las Humanidades y las Ciencias Sociales. Debido a esto el objetivo general de esta tesis fue el siguiente: explorar si las posibilidades para reutilizar datos de GS para generar, de manera sostenible, productos y herramientas de naturaleza bibliométrica que proporcionaran funcionalidades que GS no ofrece por sí mismo. Para llevar a cabo este trabajo, me integré en la línea de investigación de mi director Emilio Delgado López-Cózar sobre el estudio de GS como una fuente de datos para análisis bibliométricos.

Para conseguir el objetivo general de la tesis, se siguieron dos enfoques que han sido ejecutados en paralelo. En primer lugar, llevamos a cabo estudios que analizaban las características generales de GS como una fuente de datos: sus fortalezas y debilidades respecto a varios aspectos (ver lista y Tabla 1 más abajo). Cuando era posible, estos resultados eran comparados con aquellos ofrecidos por las bases de datos WoS y/o Scopus. Algunos de los estudios que nuestro grupo ha realizado en esta categoría forman parte de esta tesis, mientras otros, en los que realicé un labor de apoyo, no son parte de esta tesis. Sin embargo, estos también están listados más abajo para proporcionar una visión general del trabajo que nuestro grupo ha realizado a lo largo de los últimos cinco años.

El segundo de los objetivos principales de esta tesis ha sido poner a prueba el conocimiento obtenido en los estudios anteriores en situaciones prácticas de la vida real. Estos proyectos tomaron la forma de aplicaciones web personalizadas, construidas para varios propósitos, y que fueron puestas a disposición de todo el mundo. Estas aplicaciones muestran datos extraídos de Google Scholar (y en algunas ocasiones también otras fuentes) de maneras que las interfaces nativas de GS, GSC y GSM no ofrecen, por tanto expandiendo el rango de maneras en las que los usuarios pueden interactuar con esta información. Las publicaciones que describen estas aplicaciones web, así como las publicaciones que analizan los datos incluidos en las aplicaciones, forman parte de esta tesis.

También es importante decir que durante la realización de esta tesis decidimos seguir una estrategia de publicación alineada con los principios de la Ciencia Abierto. De esta manera, primero publicamos los resultados y datos usados en nuestros análisis exploratorios como working papers que depositamos en servidores de preprints. Estos working papers fueron seguidamente refinados para crear uno o más artículos que fueron más tarde publicados en revistas revisadas por pares. Sin embargo, no todo el contenido presentado en los working papers llegó a aparecer en los artículos. Por esta razón, en esta tesis se incluyen tanto los working papers originales, como los artículos de revista.

Finalmente, esta tesis también contiene varios capítulos en los que se describen proyectos aún no finalizados que aún no han aparecido publicados de ninguna forma: uno de estos capítulos describe un plan para crear una aplicación web que muestra información bibliográfica y bibliométrica exhaustiva y detallada sobre investigadores que trabajan en España y tienen un perfil de GSC público, así como de las publicaciones de estos investigadores. El segundo capítulo inédito describe las funciones de una aplicación web que muestra datos sobre Acceso Abierto a publicaciones en varios niveles de agregación. Esta aplicación ya ha sido implementada, pero todavía no se ha hecho pública.

Los temas específicos relacionados con GS que mis compañeros y yo hemos trabajado durante estos cinco años están listados más abajo, y la Tabla 1 proporciona una lista de los documentos que tratan cada uno de estos temas. La tabla también agrupa bajo cada working paper los artículos que se derivaron de él, y especifica si el documento forma parte de esta tesis (y en caso afirmativo, se proporciona el capítulo de la tesis correspondiente).

- Descripción de GS como plataforma: documentos que describen el funcionamiento general de GS, las funcionalidades del buscador y sus servicios spin-off (GSC, GSM, GSCP).
- Tamaño y cobertura de GS: documentos en los que se analiza la base de datos de GS, al nivel de documentos, revistas, o autores. En algunos casos, los resultados se comparan con otras bases de datos.
- Errores / limitaciones de GS: documentos en los que se describen los varios tipos de errores y limitaciones que hemos encontrado mientras intentábamos usar GS para propósitos bibliométricos.
- Indicadores en GS: estudios que analizan los indicadores bibliométricos proporcionados por GS, algunas veces comparándolos con indicadores proporcionados por otras bases de datos.
- Acceso Abierto: estudios que analizan la idoneidad de GS como una herramienta para encontrar versiones gratuitas de documentos en la Web.
- Aplicaciones web no oficiales: documentos que describen aplicaciones web desarrolladas para esta tesis, basadas en datos extraídos de GS.

Tabla 1. Estudios relacionados con GS que mis compañeros y yo hemos llevado a cabo durante 2014-2018

Tipo doc.	**Incluido en esta tesis**	**Referencia (ordenado por fecha de publicación excepto cuando está relacionado con un working paper anterior)**	**Temas tratados**					
			Descripción GS como plataforma	**Tamaño y cobert GS**	**Errores / limitac. de GS**	**Indicador es en GS**	**Acceso Abierto**	**Aplicaciones web no oficiales**
Working paper	No	Martín-Martín, A., Ayllón, J. M., Orduña-Malea, E., & Delgado-López-Cózar, E. (2014). **Google Scholar Metrics 2014: a low cost bibliometric tool** (EC3 Working Papers No. 17). Retrieved from http://arxiv.org/abs/1407.2827	**X**					
Working paper	No	Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado-López-Cózar, E. (2014). **About the size of Google Scholar: playing the numbers** (EC3 Working Papers No. 18). Retrieved from http://arxiv.org/abs/1407.6239		**X**				
Artículo revista	No	↳ Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). **Methods for estimating the size of Google Scholar**. Scientometrics, 104(3), 931–949.		**X**				
Working paper	**Ch. 2**	Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2014). **Does Google Scholar contain all highly cited documents (1950-2013)?** (EC3 Working Papers No. 19). Retrieved from http://arxiv.org/abs/1410.8464	**X**	**X**	**X**	**X**	**X**	
Carta al editor	**Ch. 4**	↳ Martín-Martín, A., Ayllón, J. M., Delgado López-Cózar, E., & Orduna-Malea, E. (2015). **Nature 's top 100 Re-revisited**. Journal of the Association for Information Science and Technology, 66(12), 2714–2714.			**X**			
Artículo revista	**Ch. 3**	↳ Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). **A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013)**. Revista Española de Documentacion Cientifica, 39(4), e149.		**X**			**X**	
Artículo revista	**Ch. 5**	↳ Martin-Martin, A., Orduna-Malea, E., Harzing, A.-W., & Delgado López-Cózar, E. (2017). **Can we use Google Scholar to identify highly-cited documents?** Journal of Informetrics, 11(1), 152–163.	**X**					
Libro	No	Orduña-Malea, E., Martín-Martín, A., Ayllón, J. M., & Delgado López-Cózar, E. (2016). **La revolución Google Scholar: Destapando la caja de Pandora académica.** Granada: Universidad de Granada y Unión de Editoriales Universitarias Españolas.	**X**	**X**	**X**	**X**	**X**	**X**
Working paper	**Ch. 10**	Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2016). **The counting house, measuring those who count: Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in GSC (Google Scholar Citations), ResearcherID, ResearchGate, Mendeley, & Twitter** (EC3 Working Papers No. 21). Retrieved from https://arxiv.org/abs/1602.02412		**X**	**X**	**X**		
Artículo revista	**Ch. 12**	↳ Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). **A novel method for depicting academic disciplines through Google Scholar Citations: The case of Bibliometrics**. Scientometrics, 114(3), 1251–1273.		**X**		**X**		
Artículo revista	**Ch. 13**	↳ Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). **Author-level metrics in the new academic profile platforms: The online behaviour of the Bibliometrics community**. Journal of Informetrics, 12(2), 494–509.				**X**		
Artículo revista	No	Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2017). **Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors**. Revista Española de Documentación Científica, 40(4), e185.			**X**			

Tipo doc.	Incluido en esta tesis	Referencia (ordenado por fecha de publicación excepto cuando está relacionado con un working paper anterior)	Temas tratados					
			Descripción GS como plataforma	Tamaño y cobert GS	Errores / limitac. de GS	Indicador es en GS	Acceso Abierto	Aplicaciones web no oficiales
Artículo revista	No	Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2017). **The lost academic home: institutional affiliation links in Google Scholar Citations**. Online Information Review, 41(6), 762–781.	X					
Capítulo de libro	No	Delgado López-Cózar, E., Orduna-Malea, E., Martín-Martín, A., & Ayllón, J. M. (2017). **Google Scholar : The Big Data Bibliographic Tool**. In F. J. Cantu-Ortiz (Ed.), Research analytics : boosting university productivity and competitiveness through scientometrics (pp. 59–80). Boca Raton, FL: CRC Press.	X	X				X
Artículo congr.	**Ch. 9**	Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017). **Journal Scholar Metrics: building an Arts, Humanities, and Social Sciences journal ranking with Google Scholar data**. In 22nd International Conference on Science, Technology & Innovation Indicators (STI). Paris.						X
Artículo congr.	**Ch. 11**	Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017). **Scholar Mirrors: Integrating evidence of impact from multiple sources into one platform to expedite researcher evaluation**. In 22nd International Conference on Science, Technology & Innovation Indicators (STI). Paris.						X
Artículo revista	No	Delgado López-Cózar, E., & Martín-Martín, A. (2018). **Apagón digital de la producción científica española en Google Scholar [Digital blackout of Spanish scientific production in Google Scholar]**. Anuario ThinkEPI, 12, 265–276.			X			
Artículo revista	**Ch. 16**	Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). **Evidence of open access of scientific publications in Google Scholar: A large-scale analysis**. Journal of Informetrics, 12(3), 819–841.		X			X	
Artículo revista	**Ch. 6**	Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). **Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison**. Scientometrics, 116(3), 2175–2188.		X		X		
Artículo revista	**Ch. 7**	Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). **Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories**. Journal of Informetrics, 12(4), 1160–1177.		X		X		
Artículo congr.	No	Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2018). **Classic papers: using Google Scholar to detect the highly-cited documents**. In 23rd International Conference on Science and Technology Indicators (pp. 1298–1307). Leiden.	X					
Capítulo libro	No	Delgado López-Cózar, E., Orduna-Malea, E., & Martín-Martín, A. (2019). **Google Scholar as a data source for research assessment**. In W. Glaenzel, H. Moed, U. Schmoch, & M. Thelwall (Eds.), Springer Handbook of Science and Technology Indicators. Springer.	X	X	X	X		
No publicado	**Ch. 14**	Web application that displays exhaustive and detailed bibliographic and bibliometric data about researchers working in Spain who have a Google Scholar Citations profile (and their publications)						X
No publicado	**Ch. 15**	Description of a web application that presents data on Open Access of scientific publications at various levels of aggregation, based on data from Google Scholar						X

Referencias

Abad-García, M.-F., González-Teruel, A., & González-Llinares, J. (2018). Effectiveness of OpenAIRE, BASE, Recolecta, and Google Scholar at finding spanish articles in repositories. *Journal of the Association for Information Science and Technology*, *69*(4), 619–622. https://doi.org/10.1002/asi.23975

Acharya, A. (2015, September 21). What happens when your library is worldwide and all articles are easy to find? Retrieved from https://youtu.be/S-f9MjQjLsk

Acharya, A., Verstak, A., Suzuki, H., Henderson, S., Iakhiaev, M., Lin, C. C. Y., & Shetty, N. (2014). Rise of the Rest: The Growing Impact of Non-Elite Journals. Retrieved from http://arxiv.org/abs/1410.2217

Adriaanse, L. S., Rensleigh, C., & Rensleigh, C. (2011). Comparing Web of Science, Scopus and Google Scholar from an Environmental Sciences perspective. *South African Journal of Libraries and Information Science*, *77*(2). https://doi.org/10.7553/77-2-58

Aguillo, I. F. (2011, December 21). Is Google Scholar useful for bibliometrics? A webometric analysis. https://doi.org/10.1007/s11192-011-0582-8

Amara, N., & Landry, R. (2012). Counting citations in the field of business and management: why use Google Scholar rather than the Web of Science. *Scientometrics*, *93*(3), 553–581. https://doi.org/10.1007/s11192-012-0729-2

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, *3*(1), 7. https://doi.org/10.1186/1742-5581-3-7

Bar-Ilan, J. (2006). An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. *Information Processing & Management*, *42*(6), 1553–1566. https://doi.org/10.1016/j.ipm.2006.03.019

Bar-Ilan, J. (2010). Citations to the “Introduction to informetrics” indexed by WOS, Scopus and Google Scholar. *Scientometrics*, *82*(3), 495–506. https://doi.org/10.1007/s11192-010-0185-9

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A., & Daniel, H.-D. (2009). Convergent validity of bibliometric Google Scholar data in the field of chemistry—Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, S. *Journal of Informetrics*, *3*(1), 27–35. https://doi.org/10.1016/j.joi.2008.11.001

Bosman, J., & Kramer, B. (2016). Innovations in scholarly communication - data of the global 2015-2016 survey. https://doi.org/10.5281/ZENODO.49583

Brantley, P. (2007). Science Direct-ly into Google. *TOC: Tools of Change for Publishing*. Retrieved from http://toc.oreilly.com/2007/07/science-directly-into-google.html

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Seventh International World-Wide Web Conference (WWW 1998)*. Retrieved from http://ilpubs.stanford.edu:8090/361/

Butler, D. (2004). Science searches shift up a gear as Google starts Scholar engine. *Nature*, *432*(7016), 423–423. https://doi.org/10.1038/432423a

Cabezas-Clavijo, A., & Delgado-López-Cózar, E. (2013). Google Scholar and the h-index in biomedicine: The popularization of bibliometric assessment. *Medicina Intensiva (English Edition)*, *37*(5), 343–354. https://doi.org/10.1016/j.medine.2013.05.002

Chen, X. (2013). Google Scholar’s Dramatic Coverage Improvement Five Years after Debut a. *Serials Review*, *36*(4), 221–226. https://doi.org/10.1080/00987913.2010.10765321

Christianson, M. (2007). Ecology Articles in Google Scholar: Levels of Access to Articles in Core Journals. *Issues in Science and Technology Librarianship*. https://doi.org/10.5062/F4MS3QPD

Clarivate Analytics. (2015). Web of Science & Google Scholar collaboration. Retrieved June 5, 2018, from http://wokinfo.com/googlescholar/

de Winter, J. C. F., Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics*, *98*(2), 1547–1565. https://doi.org/10.1007/s11192-013-1089-2

Delgado-López-Cózar, E., & Repiso-Caballero, R. (2013). The Impact of Scientific Journals of Communication: Comparing Google Scholar Metrics, Web of Science and Scopus. *Comunicar*, *21*(41), 45–52. https://doi.org/10.3916/C41-2013-04

Delgado López-Cózar, E., Orduna-Malea, E., & Martín-Martín, A. (2019). Google Scholar as a data source for research assessment. In W. Glaenzel, H. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*. Springer.

Delgado López-Cózar, E., Orduna-Malea, E., Martín-Martín, A., & Ayllón, J. M. (2017). Google Scholar : The Big Data Bibliographic Tool. In F. J. Cantu-Ortiz (Ed.), *Research analytics : boosting university productivity and competitiveness through scientometrics* (pp. 59–80). Boca Raton, FL: CRC Press.

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446–454. https://doi.org/10.1002/asi.23056

Elsevier. (2004). Scopus comes of age. Retrieved November 19, 2018, from https://www.elsevier.com/about/press-releases/science-and-technology/scopus-comes-of-age

Franceschet, M. (2009). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, *83*(1), 243–258. https://doi.org/10.1007/s11192-009-0021-2

García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in Psychology. *Journal of the American Society for Information Science and Technology*, *61*(10), 2070–2085. https://doi.org/10.1002/asi.21372

Giles, J. (2005). Start your engines. *Nature*, *438*(7068), 554–555. https://doi.org/10.1038/438554a

Giustini, D. (2005). How Google is changing medicine. *BMJ (Clinical Research Ed.)*, *331*(7531), 1487–1488. https://doi.org/10.1136/bmj.331.7531.1487

Gusenbauer, M. (2018). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 1–38. https://doi.org/10.1007/s11192-018-2958-5

Halevi, G., Moed, H., & Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the Literature. *Journal of Informetrics*, *11*(3), 823–834. https://doi.org/10.1016/J.JOI.2017.06.005

Harzing, A.-W. (2013). A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. *Scientometrics*, *94*(3), 1057–1075. https://doi.org/10.1007/s11192-012-0777-7

Harzing, A.-W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics*, *98*(1), 565–575. https://doi.org/10.1007/s11192-013-0975-y

Harzing, A.-W., & Mijnhardt, W. (2015). Proof over promise: towards a more inclusive ranking of Dutch academics in Economics &amp; Business. *Scientometrics*, *102*(1), 727–749.

https://doi.org/10.1007/s11192-014-1370-z

Harzing, A. W. (2007). Publish or Perish. Retrieved from http://www.harzing.com/pop.htm

Harzing, A. W. K., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, *8*(1), 61–73. https://doi.org/10.3354/esep00076

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hodge, D. R., & Lacasse, J. R. (2011). Ranking disciplinary journals with the Google Scholar h-index: A new tool for construction cases for tenure, promotion, and other professional decisions. *Journal of Social Work Education*, *47*(3), 579–596. https://doi.org/10.5175/JSWE.2011.201000024

Howland, J. L., Howell, S., Wright, T. C., & Dickson, C. (2009). Google Scholar and the Continuing Education Literature. *The Journal of Continuing Higher Education*, *57*(1), 35–39. https://doi.org/10.1080/07377360902806890

Jacimovic, J., Petrovic, R., & Zivkovic, S. (2010). A citation analysis of Serbian Dental Journal using Web of Science, Scopus and Google Scholar. *Stomatoloski Glasnik Srbije*, *57*(4), 201–211. https://doi.org/10.2298/SGS1004201J

Jacobs, J. A. (2009). Where Credit Is Due: Assessing the Visibility of Articles Published in Gender & Society with Google Scholar. *Gender & Society*, *23*(6), 817–832. https://doi.org/10.1177/0891243209351029

Jacsó, P. (2005a). As we may search — Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*. Current Science Association. https://doi.org/10.2307/24110924

Jacsó, P. (2005b). Comparison and Analysis of the Citedness Scores in Web of Science and Google Scholar. In *International Conference on Asian Digital Libraries* (pp. 360–369). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11599517_41

Jacsó, P. (2005c). Google Scholar: the pros and the cons. *Online Information Review*, *29*(2), 208–214. https://doi.org/10.1108/14684520510598066

Jacsó, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review*, *30*(3), 297–309. https://doi.org/10.1108/14684520610675816

Jacsó, P. (2008a). Google Scholar revisited. *Online Information Review*, *32*(1), 102–114. https://doi.org/10.1108/14684520810866010

Jacsó, P. (2008b). The pros and cons of computing the h-index using Google Scholar. *Online Information Review*, *32*(3), 437–452. https://doi.org/10.1108/14684520810889718

Jacsó, P. (2009). Calculating the h-index and other bibliometric and scientometric indicators from Google Scholar with the Publish or Perish software. *Online Information Review*, *33*(6), 1189–1200. https://doi.org/10.1108/14684520911011070

Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, *34*(1), 175–191. https://doi.org/10.1108/14684521011024191

Jacsó, P. (2012a). Google Scholar Author Citation Tracker: is it too little, too late? *Online Information Review*, *36*(1), 126–141. https://doi.org/10.1108/14684521211209581

Jacsó, P. (2012b). Google Scholar Metrics for Publications. *Online Information Review*, *36*(4), 604–619. https://doi.org/10.1108/14684521211254121

Jamali, H. R., & Nabavi, M. (2015). Open access and sources of full-text articles in Google Scholar in different subject fields. *Scientometrics*, *105*(3), 1635–1651. https://doi.org/10.1007/s11192-015-1642-2

Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. *PloS One*, *9*(5), e93949. https://doi.org/10.1371/journal.pone.0093949

Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, *74*(2), 273–294. https://doi.org/10.1007/s11192-008-0217-x

Kulkarni, A. V, Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *JAMA*, *302*(10), 1092–1096. https://doi.org/10.1001/jama.2009.1307

Laakso, M., & Lindman, J. (2016). Journal copyright restrictions and actual open access availability: a study of articles published in eight top information systems journals (2010–2014). *Scientometrics*, *109*(2), 1167–1189. https://doi.org/10.1007/s11192-016-2078-z

Laakso, M., & Polonioli, A. (2018). Open access in ethics research: an analysis of open access availability and author self-archiving behaviour in light of journal copyright restrictions. *Scientometrics*, 1–27. https://doi.org/10.1007/s11192-018-2751-5

Lasda Bergman, E. M. (2012). Finding Citations to Social Work Literature: The Relative Benefits of Using Web of Science, Scopus, or Google Scholar. *The Journal of Academic Librarianship*, *38*(6), 370–379. https://doi.org/10.1016/j.acalib.2012.08.002

Levy, S. (2014). The Gentleman who Made Scholar. *Wired*. Retrieved from https://www.wired.com/2014/10/the-gentleman-who-made-scholar

Martell, C., & Martell, C. (2009). A Citation Analysis of College & Research Libraries Comparing Yahoo, Google, Google Scholar, and ISI Web of Knowledge with Implications for Promotion and Tenure. *College & Research Libraries*, *70*(5), 460–473. https://doi.org/10.5860/0700460

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, *12*(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012

Martín-Martín, A., & Delgado López-Cózar, E. (2018). Google Scholar's citation graph: comprehensive, global… and inaccessible. *Open Citations Seminar Organised by Uppsala Universitet*. Retrieved from http://hdl.handle.net/10481/51153

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2016). Back to the past : on the shoulders of an academic search engine giant. *Scientometrics*, *107*(3), 1477–1487. https://doi.org/10.1007/s11192-016-1917-2

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentacion Cientifica*, *39*(4), e149. https://doi.org/10.3989/redc.2016.4.1405

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

McCullagh, D. (2006, July 20). Google Scholar trademark case ends. *ZDNet*. Retrieved from https://www.zdnet.com/article/google-scholar-trademark-case-ends/

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105–2125. https://doi.org/10.1002/asi.20677

Mehta, A. (2004, December 10). ACS Takes Legal Action Against Google. *Chemical & Engineering News*. Retrieved from http://pubs.acs.org/cen/news/8250/8250acs.html

Mikki, S. (2009). Comparing Google Scholar and ISI Web of Science for Earth Sciences. *Scientometrics*, *82*(2), 321–331. https://doi.org/10.1007/s11192-009-0038-6

Mikki, S., Ruwehy, H. A. Al, Gjesdal, Ø. L., & Zygmuntowska, M. (2018). Filter bubbles in interdisciplinary research: a case study on climate and society. *Library Hi Tech*, LHT-03-2017-0052. https://doi.org/10.1108/LHT-03-2017-0052

Minasny, B., Hartemink, A. E., McBratney, A., & Jang, H.-J. (2013). Citations and the h index of soil researchers and journals in the Web of Science, Scopus, and Google Scholar. *PeerJ*, *1*, e183. https://doi.org/10.7717/peerj.183

Mingers, J., & Lipitakis, E. A. E. C. G. (2010). Counting the citations: a comparison of Web of Science and Google Scholar in the field of business and management. *Scientometrics*, *85*(2), 613–625. https://doi.org/10.1007/s11192-010-0270-0

Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics*, *10*(2), 533–551. https://doi.org/10.1016/j.joi.2016.04.017

Mussell, J., & Croft, R. (2013). Discovery Layers and the Distance Student: Online Search Habits of Students. *Journal of Library & Information Services in Distance Learning*, *7*(1–2), 18–39. https://doi.org/10.1080/1533290X.2012.705561

Nicholas, D., Boukacem-Zeghmouri, C., Rodríguez-Bravo, B., Xu, J., Watkinson, A., Abrizah, A., … Świgoń, M. (2017). Where and how early career researchers find scholarly information. *Learned Publishing*, *30*(1), 19–29. https://doi.org/10.1002/leap.1087

Norris, M., Oppenheim, C., & Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the American Society for Information Science and Technology*, *59*(12), 1963–1972. https://doi.org/10.1002/asi.20898

Notess, G. R. (2005). Scholarly Web Searching: Google Scholar and Scirus. *Information Today*. Retrieved from http://www.infotoday.com/online/jul05/onthenet.shtml

Ocholla, D., & Onyancha, O. B. (2009). Assessing researchers' performance in developing countries : is Google Scholar an alternative? *Mousaion*, *27*(1), 43–64. Retrieved from http://uir.unisa.ac.za/handle/10500/5269

Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado-López-Cózar, E. (2014). *About the size of Google Scholar: playing the numbers* (EC3 Working Papers No. 18). Retrieved from http://arxiv.org/abs/1407.6239

Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, *104*(3), 931–949. https://doi.org/10.1007/s11192-015-1614-6

Orduña-Malea, E., Martín-Martín, A., Ayllón, J. M., & Delgado López-Cózar, E. (2016). *La revolución Google Scholar : Destapando la caja de Pandora académica*. Granada: Universidad de Granada y Unión de Editoriales Universitarias Españolas.

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2018). Classic papers: using Google Scholar to detect the highly-cited documents. In *23rd International Conference on Science and Technology Indicators* (pp. 1298–1307). Leiden. https://doi.org/10.31235/osf.io/zkh7p

Ortega, J. L. (2014). *Academic search engines : a quantitative outlook*. Chandos Publishing.

Pauly, D., & Stergiou, K. (2005). Equivalence of results from two citation analyses: Thomson ISI's Citation Index and Google's Scholar service. *Ethics in Science and Environmental Politics*, *9*, 33–35. https://doi.org/10.3354/esep005033

Pitol, S. P., & De Groote, S. L. (2014). Google Scholar versions: do more versions of an article mean greater impact? *Library Hi Tech*, *32*(4), 594–611. https://doi.org/10.1108/LHT-05-2014-0039

Rosenstreich, D., & Wooliscroft, B. (2009). Measuring the impact of accounting journals using Google

Scholar and the g-index. *The British Accounting Review*, *41*(4), 227–239. https://doi.org/10.1016/J.BAR.2009.10.002

S. Adriaanse, L., & Rensleigh, C. (2013). Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. *The Electronic Library*, *31*(6), 727–744. https://doi.org/10.1108/EL-12-2011-0174

Science. (2004). NET NEWS: A Google for Academia. *Science*, *306*(5702), 1661.3-1661. https://doi.org/10.1126/science.306.5702.1661c

Sember, M., Utrobicić, A., & Petrak, J. (2010). Croatian Medical Journal citation score in Web of Science, Scopus, and Google Scholar. *Croatian Medical Journal*, *51*(2), 99–103. https://doi.org/10.3325/CMJ.2010.51.99

Teplitzky, S. (2017). Open Data, [Open] Access: Linking Data Sharing and Article Sharing in the Earth Sciences. *Journal of Librarianship and Scholarly Communication*, *5*(General Issue), eP2150. https://doi.org/10.7710/2162-3309.2150

Van Noorden, R. (2014a). Online collaboration: Scientists and the social network. *Nature*, *512*(7513), 126–129. https://doi.org/10.1038/512126a

Van Noorden, R. (2014b, November 7). Google Scholar pioneer on search engine's future. *Nature*. https://doi.org/10.1038/nature.2014.16269

Verstak, A., Acharya, A., Suzuki, H., Henderson, S., Iakhiaev, M., Lin, C. C. Y., & Shetty, N. (2014). *On the Shoulders of Giants: The Growing Impact of Older Articles*. Retrieved from http://arxiv.org/abs/1411.0275

Vine, R. (2005). Google Scholar is a full year late indexing Pubmed content. Retrieved from http://web.archive.org/web/20060716085124/ http://www.workingfaster.com/sitelines/archives/2005_02.html

Vine, R. (2006). Google Scholar. *Journal of the Medical Library Association*, *94*(1), 97. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1324783/

Walters, W. H. (2007). Google Scholar coverage of a multidisciplinary field. *Information Processing & Management*, *43*(4), 1121–1132. https://doi.org/10.1016/J.IPM.2006.08.006

Wang, Y., & Howard, P. (2012). Google Scholar Usage: An Academic Library's Experience. *Journal of Web Librarianship*, *6*(2), 94–108. https://doi.org/10.1080/19322909.2012.672067

Wentz, R. (2004). WoS versus Google Scholar: Cited by...: Correction. *Medical Libraries Discussion List*. Retrieved from https://web.archive.org/web/20070630064521/http://listserv.acsu.buffalo.edu/cgi-bin/wa?A2=ind0412B&L=medlib-l&P=R5842&I=-3&m=95812

Yang, K., & Meho, L. I. (2007). Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–15. https://doi.org/10.1002/meet.14504301185

Zarifmahmoudi, L., Kianifar, H. R., & Sadeghi, R. (2013). Citation Analysis of Iranian Journal of Basic Medical Sciences in ISI Web of Knowledge, Scopus, and Google Scholar. *Iranian Journal of Basic Medical Sciences*, *16*(10), 1027–1030. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3874088&tool=pmcentrez&rendertype=abstract

Zarifmahmoudi, L., & Sadeghi, R. (2012). Comparison of ISI web of knowledge, SCOPUS, and Google scholar h-indices of Iranian nuclear medicine scientists. *Iranian Journal of Nuclear Medicine*, *20*(1), 1–4.

Results

Section 1. Understanding the characteristics of Google Scholar: size, document coverage, citation data, and errors

Chapter 1. Summary of results

The first step towards establishing whether Google Scholar (GS) is a useful source of data for bibliometric analyses is to know the characteristics of its document base. This is also necessary if we are to benchmark GS against other sources of data that can be used for similar purposes. Unfortunately, GS has never been transparent about its size, coverage, and list of sources from which it indexes documents. In other words, GS operates as a black box system, in which users (or other interested parties such as publishers, repository managers, etc.) can only provide input (queries), and observe the output returned by GS, without the possibility of knowing what occurs in between. The official documentation available in its help pages (Google Scholar, 2019) keeps aspects such as size and coverage intentionally vague: "Google Scholar includes scholarly articles from a wide variety of sources in all fields of research, all languages, all countries, and over all time periods". Reportedly, the team behind GS does not think these aspects are something users should be concerned about. In an interview published in 2014 on the occasion of GS's tenth anniversary (Van Noorden, 2014), GS's founding engineer Anurag Acharya declared that "the index size might be a concern here if it was too small. But we are clearly large enough".

In an attempt to pry open the black box that is GS, or in other words, to reduce the uncertainty behind these unknown aspects of GS, my colleagues and I have carried out a series of analyses. These analyses have centered around the size of its index, its document coverage, and its citation data. Along the way, we have encountered and documented the errors that can be found in the platform, as well as the technical limitations that we faced when we tried to analyse the data available in GS.

What follows is a summary of the objectives, methods, and results of all the studies my colleagues and I have carried out during the past five years. Not all of them can be considered, strictly speaking, a part of my thesis, because there are some in which I only had a supporting role. Those studies are, of course, not included as a part of this thesis. They are all, however, part of the same research line and therefore we believe it is best to provide the full context of our work in this summary, rather than present a partial account. This is the natural consequence of working within a team.

Size of Google Scholar

We first tried to estimate the size of GS's document index. Our first approaches (Orduña-Malea, Ayllón, Martín-Martín, & Delgado-López-Cózar, 2014; Orduna-Malea, Ayllón, Martín-Martín, & Delgado López-Cózar, 2015) relied on a number of different estimation methods: estimations from empirical data partially based on Khabsa & Giles (2014), and estimations from direct queries to the search engine, which relied on the approximate number of results found (hit count) declared by GS for any given query. The estimations from direct queries were further subdivided into "empty queries" and "absurd queries". "Empty queries" were queries that did not contain any topic keyword which could limit the scope of the results returned by GS. Instead, only publication years (a range of years, or single years) where defined. "Absurd queries" were queries that took advantage of advanced query options to force the search engine to return as many results as possible. These queries contained single letters or numbers (such as “a” or “1”) and additionally, used the “site:” and “NOT” operators (in GS, the “NOT” operator is denoted with the “-“ symbol) in conjunction with a non-existent web domain. Thus, an absurd query would look like “1 -site:ssstfsffsdfasdfsf.com”, which would mean “return all documents that contain the number 1, and which are NOT hosted in the domain ssstfsffsdfasdfsf.com”.

Depending on the method used, and the specific parameters (inclusion or exclusion of cited references and patents), the results from our 2014 study ranged from 80 million scholarly documents (empty, year-by-year query, excluding citing references and patents) to 176 million (absurd query using the range of years 1700-2013, including cited references and patents). This study was replicated in 2017 (Delgado López-Cózar, Orduna-Malea, & Martín-Martín, 2019), finding estimates that ranged from 184 million records (absurd query, excluding cited references and patents) to 331 million (absurd query, including cited references and patents).

Other studies can also provide insight into the size of Google Scholar. It is also possible to gauge the size of GS using a similar methodology to Khabsa & Giles (2014), who analysed the citations found by two indexes (GS and Microsoft Academic Search) to a sample of documents that were covered by both indexes. In this regard, throughout these years we have worked with a number of diverse samples which allow us to gauge the size of GS by comparing it to the data found in other citation indexes for which the size is known.

Our studies show that for virtually any collection of documents covered both by GS and WoS, GS is able to find a higher amount of citations to those documents. According to the samples available in Table 1, the ratio of GS citations to WoS citations ranges from 1.54 (for a sample of 2.26 million articles and reviews published in 2009 or 2014) to 2.9 (for an admittedly much smaller sample of 239 of the most cited articles published by Spanish LIS researchers). Although the samples in Table 1 differ much from each other, there seems to be a pattern indicating that citation counts in GS are about 50-60% higher than in WoS when most of the documents in the sample are from STEM fields and published in English, but much higher (150-200%) when the articles are from the Social Sciences and/or published in languages other than English. Results at the level of subject categories can be found in Martín-Martín, Orduna-Malea, Thelwall, & Delgado López-Cózar (2018) (Chapter 7 of this thesis). At this level of aggregation, GS is found to provide at least 30% more citations than WoS (in the field of Chemistry & Chemical Engineering), and as much as four times as many citations as WoS (in the field of Literature).

This phenomenon is even more noticeable when we turn to researchers themselves as an object of study. In this case, we are not studying the same document collections, but instead all the documents covered by each database published by specific researchers. As Table 1 shows, for a sample 196 Bibliometrics & Scientometrics researchers, GS found 3.24 times more citations to documents published by those authors than WoS did. What’s more, for a sample of 337 Spanish LIS researchers, GS found 9.25 times more citations to documents by those authors than WoS did.

Table 1. Ratio of Google Scholar citations to Web of Science citations in various samples

Document-level citation counts						
Source	**Data collection**	**Description**	**N docs**	**GS citations**	**WoS citations**	**Ratio GS/WoS**
(Martín-Martín, Orduna-Malea, Thelwall, et al., 2018)	April-May 2018	Citation counts of the documents that cite any of the 2,299 highly-cited documents in GS's *Classic Papers*	1.03 million	47.2 million	29 million	1.63 (1.3-4.0)
		Highly-cited articles in 252 subject categories and published in 2006. Extracted from GS's *Classic Papers* product	2,299	2.30 million	1.27 million	1.81
(Delgado López-Cózar et al., 2019)	February 2017	Highly cited documents in GS by language and year 1950-2016). WoS data extracted only from GS/WoS integration	69,261	80.8 million	44.9 million	1.80
(Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018a)	June-October 2016	Articles and reviews with a DOI covered by WoS, published in 2009 or 2014. WoS data extracted from web interface	2.26 million	42.6 million	27.6 million	1.54
(Delgado López-Cózar, Martín-Martín, Orduña-Malea, & Ayllón, 2015b)	July 2015	Top cited documents extracted from GSC profiles of Bibliometricians. WoS data extracted from GS/WoS integration and cited reference search (for non-sourced materials)	1,055	240,066	97,281	2.46
(Delgado López-Cózar, Lorenzo-Sar, Martín-Martín, & Ayllón, 2015)	July 2015	Top 150 most cited documents by 30 prominent bibliometric researchers. WoS data extracted from GS/WoS integration and cited reference search (for non-sourced materials)	150	85,729	34,000	2.52
(Delgado López-Cózar, Martín-Martín, Orduña-Malea, & Ayllón, 2015a)	February 2015	Top cited documents extracted from GSC profiles of Spanish Library & Information Science researchers. WoS data extracted only from GS/WoS integration	239	11,343	3,900	2.9
(Martín-Martín, Orduña-Malea, Ayllón, & Delgado-López-Cózar, 2014b)	May 2014	Highly cited documents in Google Scholar by years (1950-2013). WoS data extracted only from GS/WoS integration	32,679	58.5 million	35.2 million	1.66
Author-level citation counts						
Source	**Data collection**	**Description**	**N docs**	**GS citations**	**WoS citations**	**Ratio GS/WoS**
(Delgado López-Cózar, Martín-Martín, et al., 2015b)	July 2015	Total number of citations received by Bibliometrics and Scientometrics researchers. WoS data extracted from ResearcherID	196	379,978	117,096	3.24
(Delgado López-Cózar, Martín-Martín, et al., 2015a)	February 2015	Total number of citations received by Spanish Library & Information Science researchers. WoS data extracted only from GS/WoS integration	337	68,259	8,267	9.25

Confidence level: 95%; *p* values < 0.05

Coverage of Google Scholar

Once we had a general idea of the size of GS's document index, we proceeded to analyse the actual composition of the document index in terms of document types and languages. When possible, the results of these analyses were benchmarked against the data available in the other two most widely used citation indexes: WoS and Scopus.

In our first approach, given the search limitations inherent to GS (Delgado López-Cózar et al., 2019; Orduña-Malea, Martín-Martín, Ayllón, & Delgado López-Cózar, 2016), we limited our analysis to highly-cited documents. We carried out 64 keyword-free queries. Each query limited results to those published in a specific year, and we carried out queries for the years within the range 1950-2013. Because GS displays a maximum of 1,000 records per query, we were able to collect 64,000 records using this method (Martín-Martín, Orduña-Malea, Ayllón, & Delgado-López-Cózar, 2014a) (Chapter 2 of this thesis). By removing the topic limitations imposed by using keywords in a query, we expected to obtain the most highly-cited documents in each publication year. This is because citation counts had previously been found to be “the highest weighted factor in Google Scholar’s ranking algorithm” (Beel & Gipp, 2009).

As expected, the results returned by GS were apparently the most cited documents indexed by GS. This is supported by the high correlation observed between the actual positions that each document occupied in GS’s relevance ranking in the results pages, and the position that those documents would occupy in a ranking solely based on citation counts (Martin-Martin, Orduna-Malea, Harzing, & Delgado López-Cózar, 2017) (Chapter 5 of this thesis). Additionally, it turned out that around the same time our results were published as a working paper (Martín-Martín et al., 2014a) (Chapter 2 of this thesis), a similar ranking of the top 100 most highly-cited documents using data from GS that was published by the journal Nature (Van Noorden, Maher, & Nuzzo, 2014). This ranking had been provided to the authors of the Nature piece by the team behind GS. Although we could only compare the top 100 most cited documents in our sample with those in the Nature ranking, the two were very similar, which further supports our hypothesis that had indeed recovered a list of the most highly-cited documents in GS.

Despite the limited metadata provided by GS, we were able to determine the document type of 71% of the documents in the sample. Just over half (51%) of the documents in our sample were journal articles (Martín-Martín, Orduna-Malea, Ayllón, & Delgado López-Cózar, 2016) (Chapter 3 of this thesis). 18% of them were books or book chapters, and the remaining 2% for which a document type could be identified were conference communications, and other types of scholarly documents. For 29% of the documents in this sample, a document type could not be automatically identified. It is worth noting that within the top 25 most cited documents of all time according to GS, 14 of them were books. The vast majority (93%) of the documents in the sample of highly-cited documents were published in English, and 7% in other languages (Spanish: 2%; Portuguese, German, French, Russian, and Chinese: 1% each; other languages: less than 1%). Taking advantage of the integration between GS and WoS that is available for WoS subscribers (Clarivate Analytics, 2015), we could also determine that only 51% of the documents in our sample were covered by WoS.

In a more recent study (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018) (Chapter 6 of this thesis) we analysed the coverage highly-cited documents in GS at the level of subject categories, and checked for coverage of the same documents in WoS and Scopus. The objective of this study was to check whether the general approach to document indexing in each database (the inclusive approach of Google Scholar, and the exclusive approach of WoS and Scopus) could affect the computation of bibliometric indicators based on counts of highly-cited documents. The sample of documents were all articles in GS’s *Classic Papers* (GSCP), which displays the top 10 most cited articles published in 2006 in each of 252 subject categories. The results showed that, even within this highly select group, in some areas WoS and Scopus did not cover a significant portion of the documents that GS finds to be highly-cited. Of the documents in the area of Humanities, Literature & Arts, 28.2% were not covered by WoS, and 17.1% were not covered by Scopus. In Social Sciences, WoS did not cover 17.5% of the documents, and Scopus 8.6%. WoS also had a low coverage in Engineering & Computer Science (11.6% of the documents were not covered) mainly because of its low coverage of conference proceedings and in Business, Economics & Management (6% of the documents were not covered). In the rest of the areas, there were some missing documents, but the proportion was lower than 3%.

In a subsequent study, we analysed the citations to the subset of GSCP that were covered by all three databases (GS, WoS, and Scopus) (Martín-Martín, Orduna-Malea, Thelwall, et al., 2018) (Chapter 7 of this thesis). We extracted 2,301,997 citations from GS, 1,270,225 citations from WoS, and 1,515,436 citations from Scopus. After extracting those three datasets, we computed the overlap of citations in the three databases, overall and by subject categories. According to the results, GS was the platform with the most exhaustive coverage: it found 94% of all the citations available in any of the three sources. In comparison, WoS only found 52% of all citations, and Scopus 60%. Regarding the overlap of citations, GS found 95% of the citations that WoS found, and 92% of the citations that Scopus found. These proportions were different when the citations were disaggregated according to the subject category of the cited article.

Regarding the document types of these citations, results showed that the distribution of document types was significantly different between citations only found by GS (GS unique citations), and overlapping citations (citations found by GS and also by at least one of the other sources), as well as by subject area. Of GS unique citations, around half were from non-journal sources (48%-65% depending on the area), including theses, books, conference papers, and unpublished materials. Among overlapping citations, citations from non-journal sources were much less common (6%-17%). As regards the language of the citations, non-English citations were much more common among GS unique citations (19%-38%) than among overlapping citations (0%-3%).

Lastly, there is an additional study that provided interesting evidence on the coverage of GS as compared to WoS, even if this was not its main objective (Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018b) (Chapter 16 in this thesis). This study analysed whether GS is a useful source to find evidence of OA levels of scientific publications. The sample selected for analysis were all articles and reviews with a DOI published in 2009 and 2014, and covered by the three main indexes in WoS (Science Citation Index, Social Sciences Citation Index, and Arts & Humanities Citation Index). Of the 2,269,022 documents in the sample, 97.6% of them were successfully found in GS.

Comparing citation data in Google Scholar with other sources

One of GS's most important assets, at least from a bibliometric perspective, is its citation graph. As one of the most popular tools for research discovery, GS citation counts are widely consulted by researchers, and sometimes used in research evaluations. For this reason, we have paid special attention to how GS citation data compares to citation data in the other two most widely used citation databases: WoS and Scopus.

Depending on the sample of documents or authors (Table 2), Spearman correlations of citation counts between GS and WoS range from 0.63 to 0.99. The lowest correlations are found in samples of highly-cited documents, and samples of documents in the areas of Humanities and Social Sciences (especially if the documents are not published in English), while higher correlations are found in samples of documents that are not limited to highly cited documents, especially those that contain documents in the fields of STEM. In the most recent samples (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018; Martín-Martín, Orduna-Malea, Thelwall, et al., 2018) (Chapters 6 and 7 of this thesis) we also had the opportunity to compare GS citations to Scopus citations. According to the results, Spearman correlations between GS and Scopus are even higher than between GS and WoS (0.93-0.99).

Table 2. Spearman correlations of citation counts between Google Scholar and Web of Science, and Google Scholar and Scopus, in various samples

Document-level citation counts						
Source	**Date of data collection**	**Description**	**GS-WoS N docs**	**GS-WoS Spearman correlation***	**GS-Scopus N docs**	**GS-Scopus Spearman correlation**
(Martín-Martín, Orduna-Malea, Thelwall, et al., 2018)	April-May 2018	Documents that cite documents in GS's *Classic Papers*. WoS and Scopus data extracted from their respective web interfaces	1.03 million	0.94 (0.78-0.98)	1.2 million	0.96 (0.93-0.99)
(Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018)	June 2017	Top 10 most highly cited documents published in 2006 in each of 252 subject categories, displayed in GS's *Classic Papers*	2,305	0.88	2,421	0.92
(Delgado López-Cózar et al., 2019)	February 2017	Highly cited documents in GS by language and year 1950-2016). WoS data extracted only from GS/WoS integration	69,261	0.91		
(Martín-Martín, Costas, van Leeuwen, et al., 2018a)	June-October 2016	Articles and reviews with a DOI covered by WoS, published in 2009 or 2014. WoS data extracted from web interface	2.26 million	0.91		
(Delgado López-Cózar, Martín-Martín, et al., 2015b)	July 2015	Top cited documents extracted from GSC profiles of Bibliometricians. WoS data extracted from GS/WoS integration and cited reference search (for non-sourced materials)	1,055	0.76		
(Delgado López-Cózar, Lorenzo-Sar, et al., 2015)	July 2015	Top 150 most cited documents by 30 prominent bibliometric researchers. WoS data extracted from GS/WoS integration and cited reference search (for non-sourced materials)	150	0.80		
(Delgado López-Cózar, Martín-Martín, et al., 2015a)	February 2015	Top cited documents extracted from GSC profiles of Spanish Library & Information Science researchers. WoS data extracted only from GS/WoS integration	239	0.63		
(Martín-Martín et al., 2014b)	May 2014	Highly cited documents in Google Scholar by years (1950-2013). WoS data extracted only from GS/WoS integration	32,679	0.73		
Author-level citation counts						
Source	**Date of data collection**	**Description**	**GS-WoS N docs**	**GS-WoS Spearman correlation***	**GS-Scopus N docs**	**GS-Scopus Spearman correlation**
(Delgado López-Cózar, Martín-Martín, et al., 2015a)	February 2015	Total number of citations received by Spanish Library & Information Science researchers. WoS data extracted only from GS/WoS integration	337	0.79		
(Delgado López-Cózar, Martín-Martín, et al., 2015b)	July 2015	Total number of citations received by Bibliometrics and Scientometrics researchers. WoS data extracted from ResearcherID	196	0.91		

Confidence level: 95%; *p* values < 0.05

Errors in Google Scholar

Unlike other citation indexes and bibliographic databases in general, GS has a completely automated approach to document indexing. This allows GS to build a much more comprehensive document index than citation indexes such as WoS or Scopus, but it comes with a trade-off (Harzing, 2016): GS's algorithms sometimes make mistakes that wouldn't happen in a system were records are manually curated by humans. Some researchers consider these errors the reason why data from GS should not be used for bibliometric purposes (Jacsó, 2006, 2010).

It is important to note that errors should not be confused with limitations, and especially the limitations of GS as tool to carry out bibliometric analyses, which as was mentioned in the introduction, was never the intended use envisioned by GS's creators. Our definition of error is the following: a deviation from a feature or service that GS declares to offer. An overview of the limitations of GS, GSC, and GSM for bibliometric analyses can be found in Delgado López-Cózar et al. (2019).

According to the taxonomy in Orduna-Malea, Martín-Martín, & Delgado López-Cózar (2017), while searching in GS we can find several types of errors:

- Coverage errors: these can further be classified into
    - false positives: records that GS should not have included in its index, because they are outside the scope, of because they are not scholarly documents. Examples include records that point to document types that GS considers to be "not appropriate" for its index (Google Scholar, 2019), such as magazine articles, book reviews, editorials, and course syllabi. What's more, in some cases, even less related sources such as pages from online stores , or websites containing adult content (Wesley, 2016) can be found on GS. In most cases, these are the results of deliberate SPAM attempts, and they disappear after some time.
    - false negatives: documents that GS should have included in its index, but has not included. False negatives are sometimes caused by journal or repository websites not following GS's indexing guidelines. Although it is not easy to establish direct communication with the GS team to fix these problems, their guidelines have become increasingly detailed in this respect through the years (Google Scholar, 2019). Therefore, in these cases it can be argued that Google Scholar is not the one at fault. Nevertheless, there are other cases where GS's behaviour is more difficult to justify. One example of this is what came to be known as the "Google Scholar preprint bug" (Wilke, 2014). This describes the phenomenon that occurs when a preprint of an article is published online ahead of its publication on a journal. When the article is published in the journal, GS sometimes fails to index the version of the article published in the journal, keeping the preprint version as the only version.
- Parsing errors: once GS has identified and decided to index a document, it may fail to extract the metadata of the document correctly from the source. In the early stages of GS this was fairly common, because not many sources provided standardized metadata, and GS had to infer this information from the layout of the PDF. This led to many mistakes, such as taking an incorrect string of text as the title of the document (i.e. the name of the journal, or the copyright declaration), or as the authors ("I Introduction", "et al.") (Jacsó, 2006). Another recurring error was taking the ISSN of the journal as the publication year (Jacsó, 2008). GS can also make mistakes in parsing the list of cited references of a document. Parsing errors usually trigger matching errors.
- Document and citation matching errors: these matching errors can occur between two or more versions of the same documents available on the web, or between a source document and a cited reference (when GS is building its citations graph). Sometimes, matching errors are caused by parsing errors (GS decides that two documents are not the same because their metadata does not

match). These errors are common in the cases when a document (or at least its metadata) is published in several languages (GS is not able to detect they are the same document), and in classic monographs or reference works that are re-published in many editions (and languages).

  - Errors caused by incorrect matching of different versions of source documents: because GS indexes the entire academic web, it sometimes finds various versions of the same document on the web. Usually, GS is able to merge together the different versions of the same document that it finds by comparing their metadata, but the automated system to merge versions of the same document sometimes falters. When this happens, we may find:
    - false negatives: When a matching that should occur does not, the result is duplicate records. This error can trigger other errors, such as duplicate citations (when a document is cited by another document whose versions GS is not able to merge correctly), and scattered citations (citations to a document whose versions GS is not able to merge correctly are scattered among the various versions).
    - false positives: when a matching that should not occur does occur, the result is that two (or more) documents are incorrectly merged, which may have consequences on the discoverability of some (or all) of the documents, and on their citation indicators.
  - Errors in citation matching: GS also carries out citation matching in order to build its citation graph. In this task there can also be false positives (incorrectly assigned citations, because the citing document does not actually cite the cited document), and false negatives (missed citations, when GS fails to recognize a citation that has actually occurred). Moreover, when GS is not able to link a reference found within a document with one of the source documents in its index, it creates a [CITATION]-type record (the equivalent to a cited reference record in WoS). Therefore, an incorrect matching in this case would also create duplicate records (although one of them would be a [CITATION]-type record). Errors in citation matching usually occur when the cited reference is not entirely correct in the citing document, or when it is done in a way that GS does not recognise (for example, placing citations in a footnote instead of at the end of the document, which is common in Law).

Apart from the errors that can be found on GS, GS's profile service, GSC, and its journal ranking, GSM, also contain errors. In GSC for example, apart from the errors inherited from GS, there can be:

- duplicate profiles: when the own researcher, or other people, create more than one profile about an author, and makes it publicly available. GS does not automatically create profiles, so duplicates in this case are always caused by external human intervention.
- misattributed documents in a profile: sometimes, a profile lists documents where the author did not actually participate as a co-author. This usually happens when users do not change the default option to update profiles, which is to let GS update the profile automatically without the intervention of the user. This option works well for people with uncommon surnames, but it is especially ill-suited for people with common surnames. The creator of the profile may or may not be aware that the profile contains documents that should not appear there. This, of course, has consequences on the author-level indicators that GS automatically computes based on the documents included in the profile. The easiest way to avoid this problem is to change setting to "confirm updates". That way, profile creators receive an e-mail each time GS thinks a document should be added or modified to a profile, and they only have to confirm or discard the modification. However, the problem remains that many profile creators neglect or completely abandon the profiles after creating them, which leaves the door open to the existence of increasingly inaccurate GSC profiles.
- Incorrect merging of documents: because creators of profiles are free to manage their publications as they wish, some profiles might contain merged documents that are not actually the same.

- Deliberately manipulated documents and citations in profiles:

Over the course of our analyses, we have also encountered many of these errors. These can be found in Martín-Martín et al. (2014a) (Chapter 2 of this thesis), Martín-Martín, Ayllón, Delgado López-Cózar, & Orduna-Malea (2015) (Chapter 4 of this thesis), and Martín-Martín, Orduna-Malea, Ayllón, & Delgado-López-Cózar (2016) (Chapter 10 of this thesis). For a complete literature review on the topic of Google Scholar errors we refer to Orduna-Malea et al. (2017).

References

Beel, J., & Gipp, B. (2009). Google Scholar's ranking algorithm: The impact of citation counts (An empirical study). In *2009 Third International Conference on Research Challenges in Information Science* (pp. 439–446). IEEE. https://doi.org/10.1109/RCIS.2009.5089308

Clarivate Analytics. (2015). Web of Science & Google Scholar collaboration. Retrieved June 5, 2018, from http://wokinfo.com/googlescholar/

Delgado López-Cózar, E., Lorenzo-Sar, V., Martín-Martín, A., & Ayllón, J. M. (2015). Classic Scholars' Profiles: Bibliometrics & Scientometrics. Retrieved April 1, 2017, from http://www.classic-scholars-profiles.infoec3.es/bibliometrics

Delgado López-Cózar, E., Martín-Martín, A., Orduña-Malea, E., & Ayllón, J. M. (2015a). La Biblioteconomía y Documentación española según Google Scholar Citations. Retrieved April 1, 2017, from http://www.biblioteconomia-documentacion-española.infoec3.es

Delgado López-Cózar, E., Martín-Martín, A., Orduña-Malea, E., & Ayllón, J. M. (2015b). Scholar Mirrors: Bibliometrics, Scientometrics, Informetrics, Webometrics, and Altmetrics in Google Scholar Citations, ResearcherID, ResearchGate, Mendeley, and Twitter. Retrieved April 1, 2017, from http://www.scholar-mirrors.infoec3.es

Delgado López-Cózar, E., Orduna-Malea, E., & Martín-Martín, A. (2019). Google Scholar as a data source for research assessment. In W. Glaenzel, H. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*. Springer.

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446–454. https://doi.org/10.1002/asi.23056

Google Scholar. (2019). Google Scholar Help. Inclusion Guidelines for Webmasters. Retrieved February 15, 2019, from https://scholar.google.es/intl/en/scholar/inclusion.html#content

Harzing, A.-W. (2016). Sacrifice a little accuracy for a lot more comprehensive coverage. Retrieved from https://harzing.com/blog/2016/08/sacrifice-a-little-accuracy-for-a-lot-more-comprehensive-coverage

Jacsó, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review*, *30*(3), 297–309. https://doi.org/10.1108/14684520610675816

Jacsó, P. (2008). Google Scholar revisited. *Online Information Review*, *32*(1), 102–114. https://doi.org/10.1108/14684520810866010

Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, *34*(1), 175–191. https://doi.org/10.1108/14684521011024191

Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. *PloS One*, *9*(5), e93949. https://doi.org/10.1371/journal.pone.0093949

Martín-Martín, A., Ayllón, J. M., Delgado López-Cózar, E., & Orduna-Malea, E. (2015). Nature 's top 100

Re-revisited. *Journal of the Association for Information Science and Technology*, *66*(12), 2714–2714. https://doi.org/10.1002/asi.23570

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018a). Dataset: sources of free full text found by Google Scholar for documents in Web of Science published in 2009 and 2014 (raw and aggregated). https://doi.org/10.17605/OSF.IO/FSUJY

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018b). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, *12*(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2016). *The counting house, measuring those who count: Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in GSC (Google Scholar Citations), ResearcherID, ResearchGate, Mendeley, & Twitter* (EC3 Working Papers No. 21). Retrieved from https://arxiv.org/abs/1602.02412

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2014a). *Does Google Scholar contain all highly cited documents (1950-2013)?* (EC3 Working Papers No. 19). Retrieved from http://arxiv.org/abs/1410.8464

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2014b). Highly Cited Documents on Google Scholar (1950-2013). https://doi.org/10.6084/m9.figshare.1224314

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentacion Cientifica*, *39*(4), e149. https://doi.org/10.3989/redc.2016.4.1405

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. *Scientometrics*, *116*(3), 2175–2188. https://doi.org/10.1007/s11192-018-2820-9

Martin-Martin, A., Orduna-Malea, E., Harzing, A.-W., & Delgado López-Cózar, E. (2017). Can we use Google Scholar to identify highly-cited documents? *Journal of Informetrics*, *11*(1), 152–163. https://doi.org/10.1016/j.joi.2016.11.008

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado-López-Cózar, E. (2014). *About the size of Google Scholar: playing the numbers* (EC3 Working Papers No. 18). Retrieved from http://arxiv.org/abs/1407.6239

Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, *104*(3), 931–949. https://doi.org/10.1007/s11192-015-1614-6

Orduña-Malea, E., Martín-Martín, A., Ayllón, J. M., & Delgado López-Cózar, E. (2016). *La revolución Google Scholar : Destapando la caja de Pandora académica*. Granada: Universidad de Granada y Unión de Editoriales Universitarias Españolas.

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2017). Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors. *Revista Española de Documentación Científica*, *40*(4), e185. https://doi.org/10.3989/redc.2017.4.1500

Van Noorden, R. (2014, November 7). Google Scholar pioneer on search engine’s future. *Nature*. https://doi.org/10.1038/nature.2014.16269

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature*, *514*(7524), 550–553. https://doi.org/10.1038/514550a

Wesley, B. (2016). Google Scholar is Filled with Junk. Retrieved February 17, 2019, from http://ipensatori.com/2016/06/27/google-scholar-is-filled-with-junk/

Wilke, C. (2014, November 1). The Google Scholar preprint bug. *The Serial Mentor [Blog]*. Retrieved from https://serialmentor.com/blog/2014/11/1/the-google-scholar-preprint-bug

Capítulo 1. Resumen de resultados

El primer paso para determinar si Google Scholar (GS) es una fuente de datos útil para llevar a cabo estudios bibliométricos es conocer las características de su base documental. Conocer esto también es necesario para comparar GS con otras fuentes de datos que pueden ser usadas con propósitos similares. Desafortunadamente, GS nunca ha sido transparente acerca de su tamaño, cobertura, y lista de fuentes desde las que indiza documentos. En otras palabras, GS funciona como un sistema tipo caja negra, en el que los usuarios (u otros actores interesados como editoriales, gestores de repositorios, etc.) solo pueden proporcionar inputs (consultas), y observar los outputs devueltos por GS, sin saber nunca exactamente lo que ocurre entre medias. La documentación oficial disponible en las páginas de ayuda (Google Scholar, 2019) mantiene intencionalmente vagos algunos aspectos tales como el tamaño y la cobertura: “Google Scholar cubre artículos académicos de una amplia variedad de fuentes en todos los campos de investigación, todos los idiomas, todos los países, y todos los periodos temporales”. Al parecer, el equipo tras GS piensa que estos aspectos no son algo de lo que se tengan que preocupar los usuarios. En una entrevista publicada en 2014 en ocasión del décimo aniversario de GS (Van Noorden, 2014), el ingeniero fundador de GS Anurag Acharya declaró que “el tamaño del índice podría ser causa de preocupación si fuera demasiado pequeño. Pero claramente somos suficientemente grandes”.

Con el objetivo de intentar abrir la caja negra que es GS, o en otras palabras, para reducir la incertidumbre que hay detrás de los aspectos desconocidos de GS, mis compañeros y yo hemos llevado a cabo una serie de análisis. Estos análisis se han centrado en el tamaño del índice, la cobertura documental, y los datos de citas. Por el camino, hemos encontrado y documentado los errores que se pueden encontrar en la plataforma, así como las limitaciones técnicas que hemos tenido que afrontar para analizar los datos disponibles en GS.

Lo que sigue es un resumen de los objetivos, metodologías, y resultados de todos los estudios que mis compañeros y yo hemos llevado a cabo durante los últimos cinco años. No todos ellos pueden considerarse, estrictamente hablando, parte de esta tesis como tal, porque hay trabajos en los que solo tuve una función de apoyo. Esos estudios, por supuesto, no están incluidos en la tesis. Sin embargo, todos forman parte de la misma línea de investigación y por tanto pensamos que es importante proporcionar una visión completa de nuestro trabajo en este sumario, en vez de presentar una descripción parcial. Esto es el resultado natural de trabajar dentro de un grupo.

El tamaño de Google Scholar

En nuestros primeros análisis intentamos estimar el tamaño de la base documental de GS. Nuestros primeros análisis (Orduña-Malea, Ayllón, Martín-Martín, & Delgado-López-Cózar, 2014; Orduna-Malea, Ayllón, Martín-Martín, & Delgado López-Cózar, 2015) utilizaron varios métodos de estimación diferentes: estimaciones basadas en datos empíricos parcialmente basados en Khabsa & Giles (2014), y estimaciones basadas en consultas directas al buscador, y apoyadas en el número de resultados (hit count) declarado por GS para cada consulta. Las estimaciones basadas en consultas directas se subdividieron en lo que llamamos “consultas vacías” y “consultas absurdas”. Las “consultas vacías” eran consultas en las que no se utilizaba ningún término temático que pudiera limitar el alcance temático de los resultados devueltos por GS. Estas consultas solo contenían filtros por años de publicación (rangos de años, o años individuales). Las “consultas absurdas” eran consultas que se aprovechaban de los operadores avanzados para forzar al buscador a devolver el mayor número de resultados posible. Estas consultas contenían letras o números sueltos (como “a” o “1”) y además, usaban los comandos “site:” y “NOT” (en GS, el operador “NOT” se denota con el símbolo “-“), en conjunción con un dominio web no existente (la parte “absurda” de la consulta). De esta manera, una “consulta absurda” podría ser “1 -site:ssstfsffsdfasdfsf.com”, que significaría “devuelve todos los documentos que contiene el número 1, y que NO están alojados en el dominio ssstfsffsdfasdfsf.com”.

Dependiendo del método utilizado, y de los parámetros específicos (inclusión o exclusión de referencias citadas y patentes), los resultados de nuestro estudio de 2014 iban desde los 80 millones de documentos académicos (“consulta vacía”, año por año, excluyendo referencias citadas y patentes) a 176 millones (“consulta absurda”, usando el rango de años 1700-2013, e incluyendo referencias citadas y patentes). Este estudio se replicó en 2017 (Delgado López-Cózar, Orduna-Malea, & Martín-Martín, 2019), encontrando estimaciones que iban desde los 184 millones de registros (“consulta absurda”, excluyendo referencias citadas y patentes) hasta los 331 millones (“consulta absurda”, incluyendo referencias citadas y patentes).

Hay otros tipos de estudios que también pueden ayudar a identificar el tamaño de Google Scholar. También es posible analizar el tamaño de GS usando una metodología similar a Khabsa & Giles (2014), que analizaron las citas encontradas por dos fuentes (GS y Microsoft Academic Search) a una muestra de documentos que estaban cubiertos por los dos índices. En esta línea, a lo largo de los años hemos trabajado con una serie de muestras diversas que nos permiten conocer el tamaño de GS al compararlo con los datos disponibles en otros índices de citas cuyo tamaño sí es conocido.

Nuestros estudios muestran que para virtualmente cualquier colección de documentos indizados tanto en GS como WoS, GS es capaz de encontrar una cantidad superior de citas que WoS. De acuerdo con las muestras listadas en la Tabla 1, el ratio de citas GS/WoS se mueve entre 1,54 (para una muestra de 2,26 millones de artículos y revisiones bibliográficas publicadas en 2009 o 2014) y 2,9 (para una muestra pequeña de 239 de los documento más citados por investigadores españoles del área de Biblioteconomía y Documentación). Aunque las muestras que aparecen en la Tabla 1 son muy variadas, parece haber un patrón emergente: los números de citas en GS son un 50-60% mayores que los de WoS cuando los documentos de la muestra pertenecen principalmente a campos STEM y están publicados en inglés, mientras que cuando los artículos son de Ciencias Sociales y/o están publicados en idiomas diferentes al inglés, las diferencias son mucho más grandes (las citas en GS son un 150%-200% más altas). Martín-Martín, Orduna-Malea, Thelwall, & Delgado López-Cózar (2018) (Capítulo 7 de esta tesis) realiza un análisis sistemático a nivel de 252 categorías temáticas. A este nivel, se encuentra que GS proporciona al menos un 30% más de citas que WoS (en el campo de la Química y la Ingeniería Química), y hasta cuatro veces más citas que WoS en campos como la Literatura.

Este fenómeno es todavía más fácilmente apreciable cuando analizamos datos a nivel de autores. En este caso no se estudian las mismas colecciones de documentos, sino todos los documentos publicados por un autor que están disponibles en determinadas bases de datos. Como se muestra en la Tabla 1, para una muestra de 196 investigadores internacionales en Bibliometría y Cienciometría, GS encontró 3,24 veces más citas que WoS a los documentos publicados por estos autores. Para una muestra limitada a 337 investigadores españoles en Biblioteconomía y Documentación, GS encontró 9,25 veces más citas de las encontradas por WoS.

Tabla 1. Ratio de citas entre Google Scholar y Web of Science para una serie de muestras

Conteos de citas a nivel de documento						
Referencia	**Extracción de datos**	**Descripción**	**N docs**	**Citas GS**	**Citas WoS**	**Ratio GS/WoS**
(Martín-Martín, Orduna-Malea, Thelwall, et al., 2018)	abril-mayo 2018	Conteos de citas de documentos que citan cualquiera de los 2,299 documentos altamente citados en GS *Classic Papers*	1,03 millones	47,2 millones	29 millones	1,63 (1,3-4,0)
		Artículos altamente citados en 252 categorías y publicados en 2006. Extraído de GS *Classic Papers*	2,299	2,30 millones	1,27 millones	1,81
(Delgado López-Cózar et al., 2019)	febrero 2017	Documentos altamente citados en GS por idioma y año de publicación (1950-2016). Los datos de WoS están extraídos solamente de la integración GS/WoS	69.261	80,8 millones	44,9 millones	1,80
(Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018a)	junio-octubre 2016	Artículos y revisiones con DOI cubiertas por WoS y publicadas en 2009 o 2014. Los datos de WoS están extraídos de la interfaz web	2,26 millones	42,6 millones	27,6 millones	1,54
(Delgado López-Cózar, Martín-Martín, Orduña-Malea, & Ayllón, 2015b)	julio 2015	Documentos altamente citados extraídos de perfiles GSC de bibliómetras. Los datos de WoS están extraídos de la integración GS/WoS y de WoS cited reference search (para documentos no fuente)	1.055	240.066	97.281	2,46
(Delgado López-Cózar, Lorenzo-Sar, Martín-Martín, & Ayllón, 2015)	julio 2015	Top 150 de documentos más citados por 30 investigadores clásicos de la bibliometría. Los datos de WoS están extraídos de la integración GS/WoS y de WoS cited reference search (para documentos no fuente)	150	85.729	34.000	2,52
(Delgado López-Cózar, Martín-Martín, Orduña-Malea, & Ayllón, 2015a)	febrero 2015	Documentos altamente citados extraídos de perfiles GSC de investigadores españoles de Biblioteconomía y Documentación. Los datos de WoS están extraídos solamente de la integración GS/WoS	239	11.343	3.900	2,9
(Martín-Martín, Orduña-Malea, Ayllón, & Delgado-López-Cózar, 2014b)	mayo 2014	Documentos altamente citados en GS por años de publicación (1950-2013). Los datos WoS están extraídos solamente de la integración GS/WoS	32.679	58,5 millones	35,2 millones	1,66
Conteos de citas a nivel de autor						
Referencia	**Extracción de datos**	**Descripción**	**N docs**	**Citas GS**	**Citas WoS**	**Ratio GS/WoS**
(Delgado López-Cózar, Martín-Martín, et al., 2015b)	julio 2015	Total de citas recibidas por investigadores de Bibliometría y Cienciometría. Los datos de WoS están extraídos de ResearcherID	196	379.978	117.096	3,24
(Delgado López-Cózar, Martín-Martín, et al., 2015a)	febrero 2015	Total de citas recibidas por investigadores españoles de Biblioteconomía y Documentación. Los datos de WoS están extraídos solamente de la integración GS/WoS	337	68.259	8.267	9,25

Nivel de confianza: 95%; *p* values < 0.05

Cobertura de Google Scholar

Una vez nos hicimos una idea general del tamaño de la base documental de GS, continuamos con el análisis de la composición específica de documentos en términos de tipos documentales e idiomas. Cuando fue posible, los resultados de estos análisis se compararon con datos de los dós índices de citas más usados para estudios bibliométricos: WoS y Scopus.

En una primera aproximación, dadas las limitaciones de búsqueda inherentes a GS (Delgado López-Cózar et al., 2019; Orduña-Malea, Martín-Martín, Ayllón, & Delgado López-Cózar, 2016), limitamos nuestros análisis a documentos altamente citados. Llevamos a cabo 64 consultas libres de términos temáticos. Cada consulta tenía como objetivo recuperar los documentos más citados en un año específico, y llevamos a cabo consultas para cada año del rango 1950-2013. Como GS muestra un máximo de 1.000 registros por consulta, fuimos capaces de recuperar 64.000 registros usando este método (Martín-Martín, Orduña-Malea, Ayllón, & Delgado-López-Cózar, 2014a) (Capítulo 2 de esta tesis). Al eliminar las limitaciones temáticas que se introducen cuando se utilizan palabras clave en una consulta, esperábamos obtener los documentos más altamente citados de cada año de publicación, ya que conocíamos que el criterio que más peso tiene en el ranking de ordenación de resultados por relevancia de GS es el número de citas (Beel & Gipp, 2009).

Como se esperaba, los resultados devueltos por GS eran los documentos más altamente citados en GS. Esta afirmación se apoya en la alta correlación observada entre las posiciones que cada documento ocupaba en el ranking de relevancia de GS, y las posiciones que esos mismos documentos ocuparían en un ranking basado únicamente en el número de citas (Martin-Martin, Orduna-Malea, Harzing, & Delgado López-Cózar, 2017) (capítulo 5 de esta tesis). Además, coincidentalmente, en el momento que publicamos estos resultados en un working paper (Martín-Martín et al., 2014a) (capítulo 2 de esta tesis), un ranking similar del top 100 de los documentos más altamente citados según GS fue publicado por la revista *Nature* (Van Noorden, Maher, & Nuzzo, 2014). Este ranking había sido proporcionado a los autores de la pieza en *Nature* por el equipo que trabaja en GS. Aunque solo se podían comparar los 100 documentos más citados en nuestra muestra con aquellos que aparecían en el ranking de *Nature*, los dos eran bastante similares, lo cual confirma nuestra hipótesis de que efectivamente con nuestra metodología habíamos extraído los documentos más altamente citados en GS.

A pesar de los limitados metadatos proporcionados por GS, fuimos capaces de determinar la tipología documental del 71% de los documentos de la muestra. Un poco más de la mitad (51%) de los documentos en nuestra muestra eran artículos de revista (Martín-Martín, Orduna-Malea, Ayllón, & Delgado López-Cózar, 2016) (capítulo 3 de esta tesis). Al menos el 18% de los mismos eran libros o capítulos de libro, y el 2% restante para los cuales se pudo identificar un tipo documental eran comunicaciones a congresos, y otros tipos de documentos académicos. No se pudo identificar automáticamente la tipología del 29% de los documentos en la muestra. Es importante resaltar que dentro del top 25 de los documentos más citados de todos los tiempos según GS, 14 eran libros. La gran mayoría de los documentos (93%) estaban publicados en inglés, y el 7% en otros idiomas (español: 2%; portugués, alemán, francés, ruso, y chino: 1% cada uno; otros idiomas: menos de un 1%). Aprovechando la integración entre GS y WoS disponible para los usuarios con suscripción a WoS (Clarivate Analytics, 2015), también pudimos determinar que solo el 51% de los documentos de la muestra estaban disponibles en WoS.

En un estudio más reciente (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018) (capítulo 6 de esta tesis) analizamos la cobertura de documentos altamente citados en GS a nivel de categorías temáticas, y comprobamos la cobertura de los mismos documentos en WoS y Scopus. El objetivo de este estudio era comprobar si la políticas de indización de cada base de datos (inclusivas en Google Scholar, y exclusivas en WoS y Scopus) podrían afectar al cálculo de indicadores bibliométricos basados en el conteo de documentos altamente citados. La muestra de documentos fueron todos los documentos mostrados en GSCP, que muestra el top 10 de documentos más citados en 2006 en cada una de 252 categorías temáticas. Los resultados mostraban que, incluso dentro de este selecto grupo de documentos altamente citados, WoS y Scopus no cubrían una fracción significativa de los mismos. De los documentos en el área de Humanidades, Literatura y Artes, el 28,2% no estaban cubiertos por WoS, y el 17,1% no estaban cubiertos por Scopus. En Ciencias Sociales, WoS no cubría el 17,5$ de los documentos, y Scopus el 8,6%. WoS también tenía deficiencias de coberturas en las áreas de Ingeniería e Informática (no cubría el 11,6% de los documentos) principalmente por su baja cobertura de actas de congreso, así como en el área de Empresa, Economía, y Gestión (no cubría el 6% de los documentos). En el resto de áreas había menos documentos altamente citados que WoS y Scopus no cubrían (menos del 3%).

En un estudio posterior, indagamos con más profundidad en este mismo asunto al analizar todos los documentos citantes recogidos por GS, WoS, y Scopus a la muestra de documentos altamente citados usada en el estudio anterior (Martín-Martín, Orduna-Malea, Thelwall, et al., 2018) (capítulo 7 de esta tesis). Se extrajeron 2.301.997 citas de GS, 1.270.225 citas de WoS, y 1.515.436 citas de Scopus. Después se calculó el solapamiento de citas entre las tres bases de datos, en general y por categorías temáticas. De acuerdo con los resultados, GS era la plataforma con una cobertura más exhaustiva: encontró el 94% de todas las citas posibles encontradas por las tres fuentes. En comparación, WoS solo econtró el 52% de todas las citas, y Scopus el 60%. Atendiendo al solapamiento relativo, GS encontró el 95% de todas las citas que WoS era capaz de encontrar, y el 92% de las citas que Scopus encontró. Estas proporciones eran diferentes cuando se desagregaban de acuerdo a la categoría temática del artículo citado.

Sobre los tipos documentales de estos documentos citantes, los resultados mostraron que la distribución de tipos documentales entre citas que solo GS encontraba (citas únicas en GS) y las citas solapadas (citas encontradas por GS y también por alguna otra base de datos) eran significativamente diferentes. Entre las citas únicas de GS, alrededor de la mitad venían de fuentes que no eran revistas (48%-65% dependiendo del área temática) como tesis, libros, comunicaciones a congresos, y materiales no publicados formalmente. Entre las citas solapadas, los documentos que no están publicados en revistas eran mucho menos comunes (6%-17%). Respecto al idioma de los documentos citantes, entre las citas únicas de GS los documentos no publicados en inglés eran más frecuentes (19%-38%) que entre el grupo de citas solapadas (0%-3%).

Finalmente, hay un estudio adicional que proporciona una evidencia interesante sobre la cobertura de GS comparada con WoS, incluso aunque este no era su objetivo principal (Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018b) (capítulo 16 de esta tesis). Este estudio analiza si GS es una fuente útil para encontrar evidencia de niveles de Acceso Abierto a las publicaciones científicas. La muestra selecionada para el análisis fueron todos los artículos y revisiones con DOI publicados en 2009 o 2014, y cubiertos por los tres índices principales de WoS (Science Citation Index, Social Sciences Citation Index, y el Arts & Humanities Citation Index). De los 2.269.022 documentos en la muestra, el 97,6% fueron encontrados satisfactoriamente en GS.

Comparación de datos de citas en Google Scholar con otras fuentes

Una de las características más interesantes de GS, al menos desde la perspectiva bibliométrica, es su grafo de citas. Como una de las herramientas más populares para encontrar información, los datos de citas de GS son ampliamente consultados por los investigadores, y a veces se utilizan en procesos de evaluación. Por esta razón, hemos prestado especial atención a los datos de citas de GS y los hemos comparado a los datos proporcionados por los índices de citas más utilizados para estudios bibliométricos: WoS y Scopus.

Dependiendo de la muestra de documentos o autores estudiada (Tabla 2), las correlaciones Spearman de los conteos de citas proporcionados por GS y WoS varían entre 0,63 y 0,99. Las correlaciones más bajas se encuentran en muestras de documentos altamente citados, y muestras de documentos en las áreas de Humanidades y Ciencias Sociales (especialmente si los documentos no están publicados en inglés), mientras que las correlaciones más altas se encuentran en muestras de documentos que no están limitadas a documentos altamente citados, especialmente cuando incluyen documentos en las áreas STEM. En las muestras estudiadas más recientemente (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018; Martín-Martín, Orduna-Malea, Thelwall, et al., 2018) (Chapters X and X of this thesis) también tuvimos la oportunidad de comparar las citas de GS con las de Scopus. De acuerdo a los resultados, las correlaciones en estos casos son incluso superiores que entre GS y WoS (0,93-0,99).

Tabla 2. Correlaciones Spearman entre conteos de citas en Google Scholar y Web of Science, o entre Google Scholar y Scopus, en varias muestras

Conteos de citas a nivel de documentos

Referencia	Fecha de extración de datos	Descripción	GS-WoS N docs	GS-WoS correlación Spearman*	GS-Scopus N docs	GS-Scopus correlación Spearman
(Martín-Martín. Orduna-Malea. Thelwall. et al.. 2018)	abril-mayo 2018	Documentos que citan a documentos en GSCP. Los datos de WoS y Scopus se extrajeron de sus respectivas interfaces web	1,03 millones	0,94 (0,78-0,98)	1,2 millones	0,96 (0,93-0,99)
(Martín-Martín. Orduna-Malea. & Delgado López-Cózar. 2018)	junio 2017	Top 10 de los documentos más altamente citados en 2006 en cada una de 252 categorías temáticas mostradas en GSCP	2.305	0,88	2.421	0,92
(Delgado López-Cózar et al.. 2019)	febrero 2017	Documentos altamente citados en GS por idioma y año de publicación (1950-2016). Los datos de WoS se extrajeron solamente de la integración GS/WoS	69.261	0,91		
(Martín-Martín. Costas. van Leeuwen. et al.. 2018a)	junio-octubre 2016	Artículos y revisiones con DOI cubiertas por WoS. publicadas en 2009 o 2014. Datos WoS extraídos de interfaz web	2,26 millones	0,91		
(Delgado López-Cózar. Martín-Martín. et al.. 2015b)	julio 2015	Documentos altamente citados extraídos de perfiles GSC de bibliométrics. Datos WoS extraídos de integración GS/WoS y WoS cited reference search (para documentos no fuente)	1.055	0,76		
(Delgado López-Cózar. Lorenzo-Sar. et al.. 2015)	julio 2015	Top 150 de los documentos más citados publicados por 30 investigadores clásicos de la bibliometría. Datos WoS extraídos de integración GS/WoS y WoS cited reference search	150	0,80		
(Delgado López-Cózar. Martín-Martín. et al.. 2015a)	febrero 2015	Documentos altamente citados extraídos de perfiles GSC de investigadores españoles en Biblioteconomía y Documentación. Datos WoS estraídos solamente de integración GS/WoS	239	0,63		
(Martín-Martín et al.. 2014b)	mayo 2014	Documentos altamente citados en GS por año de publicación (1950-2013). Datos WoS extraídos solamente de integración GS/WoS	32.679	0,73		

Conteos de citas a nivel de autor

Referencia	Fecha de extración de datos	Descripción	GS-WoS N docs	GS-WoS correlación Spearman*	GS-Scopus N docs	GS-Scopus correlación Spearman
(Delgado López-Cózar. Martín-Martín. et al.. 2015a)	febrero 2015	Total de citas recibidas por investigadores españoles en Biblioteconomía y Documentación. Datos WoS extraídos de integración GS/WoS	337	0,79		
(Delgado López-Cózar. Martín-Martín. et al.. 2015b)	febrero 2015	Total de citas recibidas por investigadores en Bibliometría. Datos WoS extraídos de ResearcherID	196	0,91		

Nivel de confianza: 95%; *p* values < 0.05

Errores en Google Scholar

Al contrario que otros índices de citas y bases de datos bibliográficas en general, GS utiliza un enfoque de indización de documentos totalmente automatizado. Esto permite a GS construir una base documental mucho más extensa que las de otros índices de citas como WoS y Scopus, pero tiene sus desventajas (Harzing. 2016). Los algoritmos de GS en ocasiones cometen errores que no ocurrirían en un sistema donde los registros son manualmente "curados" por personas. Algunos investigadores consideran que estos errores son la razón por la que GS no debería ser usado con propósitos bibliométricos (Jacsó. 2006. 2010).

Es importante diferenciar entre errores y limitaciones, y especialmente las limitaciones para llevar a cabo análisis bibliométricos, que como se mencionó en la introducción, nunca fue uno de los usos que los creadores de GS pretendieron para su producto. Nuestra definición de error es la siguiente: una desviación de una característica o función que GS declara ofrecer. Una revisión de las limitaciones de GS, GSC, y GSM a la hora de llevar a cabo análisis bibliométricos puede encontrarse en Delgado López-Cózar et al. (2019).

De acuerdo con la taxonomía de errores presentada en Orduna-Malea. Martín-Martín. & Delgado López-Cózar (2017), en GS se pueden encontrar los siguientes tipos de errores:

- Errores de cobertura: estos pueden subclasificarse en
  - falsos positivos: registros que GS no debería haber incluido en su índice, porque no entran en el ámbito declarado por GS, o porque no son documentos académicos. Algunos ejemplos son registros que representan tipos documentales que GS considera "no apropiados" para su índice (Google Scholar. 2019), tales como artículos de revistas no académicas (magazines), reseñas de libros, editoriales, o guías de asignaturas. En algunos casos, GS llega a indizar fuentes que no tienen ninguna relación con materiales académicos, tales como páginas de tiendas online, o páginas con contenido adulto (Wesley. 2016). En la mayoría de los casos, estos registros son el resultado de SPAM intencionado, y suelen desaparecer tras algún tiempo.
  - falsos negativos: documentos que GS debería haber incluído en su índice, pero no lo ha hecho. Los falsos negativos en ocasiones son consecuencia de que las páginas web de revistas o repositorios no siguen las directrices de indización de GS. Aunque no es fácil establecer una comunicación directa con el equipo de GS para arreglar estos problemas, las directrices que se proporcionan se han actualizado con el tiempo para ser más detalladas (Google Scholar. 2019). Por tanto, en estos casos se podría discutir que GS no es el causante del problema. Sin embargo, hay casos en los que el comportamiento de GS es más difícil de justificar. Un ejemplo de esto es lo que ha venido a denominarse el "Google Scholar preprint bug" (Wilke. 2014). Este es el fenómeno que ocurre cuando el un preprint se publica online antes de que aparezca como artículo publicado en una revista. Cuando el artículo finalmente es publicado en la revista, GS algunas veces no muestra, al menos durante un largo tiempo (hasta que se produce una actualización general), la versión final publicada en la revista, mostrando solo la versión como preprint.
- Errores de parsing: una vez GS ha identificado y decidido indizar un documento, puede equivocarse en la extracción de los metadatos del documento. En los primeros años de GS esto era bastante común, porque la mayoría de las fuentes no proporcionaban metadatos estandarizados, y GS tenía que inferir esta información del propio texto completo en el PDF. Esto conducía a muchos errores, tales como tomar una cadena de texto incorrecta como si fuera el título del documento (por ejemplo, el título de la revista, o la declaración de copyright), o como el autor del documento ("I Introduction". "et al.") (Jacsó. 2006). Otro error recurrente era tomar el

ISSN de la revista (formado por dos cadenas de cuatro dígitos, separados por un guión) como el año de publicación (Jacsó. 2008). GS también puede cometer errores al procesar la lista de referencias citadas de un documento (sobre todo si los autores no citan correctamente la fuente). Los errores de parsing suelen terminar provocando errores en el emparejamiento (matching) de documentos.

- Errores en el emparejamiento (matching) de documentos y citas: estos errores pueden ocurrir entre dos o más versiones del mismo documento disponibles en la web, o entre un documento fuente y una referencia citada (cuando GS está construyendo so grafo de citas). Algunas veces, los errores de matching están causados por errores de parsing (GS decide que dos registros no representan en realidad al mismo documento, porque sus metadatos no coinciden). Estos errores son comunes en los casos en los que un documento (o al menos sus metadatos) se publican en varios idiomas (GS no es capaz de detectar que son el mismo documento) y en los casos de monografías u obras de referencia clásicas que son reeditadas en varias ediciones (o en varios idiomas).
    - Errores causados por un matching incorrecto de diferentes versiones de documentos fuente: como GS indiza la web académica al completo, algunas veces encuentra varias versiones de un mismo documento alojadas en diferentes sitios. Normalmente, GS es capaz de unir estas versiones al detectar que sus metadatos coinciden, pero este proceso a veces falla. Cuando esto ocurre, podemos encontrar:
        - falsos negativos: cuando se deberían emparejar dos documentos, y no se hace. El resultado son entradas duplicadas para un mismo documento. Este error puede provocar otros errores, como citas duplicadas (cuando un documento es citado por otro documento para el cual GS encuentra varias versiones que no es capaz de emparejar), y citas dispersas (las citas a un documento cuyas versiones GS no es capaz de emparejar se distribuyen entre los registros duplicados).
        - falsos positivos: cuando un emparejamiento que no debería hacerse, se hace. El resultado es dos (o más) documentos incorrectamente unidos, lo que podría tener consecuencias tanto para la facilidad de encontrar en el buscador alguno (o todos) los documentos involucrados, como para los indicadores de citas que GS calcula.
    - Errores en el matching de citas: GS también realiza matching entre sus documentos fuente y las referencias que aparecen en cada documento, para poder generar su grafo de citas. En esta tarea también puede haber falsos positivos (citas incorrectamente asignadas, porque el documento citante según GS realmente no cita al documento citado) y falsos negativos (citas perdidas, cuando GS no reconoce que una cita realmente ha ocurrido). Además, cuando GS no es capaz de enlazar una referencia citada con uno de sus documentos fuente, crea un registro tipo [CITA] (el equivalente a una referencia citada en WoS). Por tanto, un matching incorrecto en este caso también crearía registros duplicados (aunque uno de ellos sería un registro tipo [CITA]). Los errores de matching de citas ocurren normalmente cuando la referencia no está correctamente representada en el documento citante, o cuando se hace en un formato que GS no reconoce (por ejemplo, utilizando el sistema cita-nota, muy común en el área de derecho y ciencias jurídicas).

Además de los errores que se pueden encontrar en GS, el servicio de perfiles de GS (GSC), y su ranking de revistas (GSM) también contienen errores propios. En GSC, aparte de los errores que se heredan de GS, podemos encontrar:

- perfiles duplicados: cuando el propio investigador, u otras personas, crear más de un perfil sobre un mismo autor, y lo hacen público. GS no crea perfiles automáticamente, así que los perfiles duplicados siempre son causados por intervenciones externas.

- Documentos incorrectamente atribuidos en un perfil: algunas veces, un perfil lista documentos en los que el autor no participó como coautor. Esto ocurre normalmente cuando los usuarios utilizan la configuración defecto, que permite a GS actualizar automáticamente el perfil sin requerir la intervención del usuario. Esta opción funciona bien para personas con apellidos poco comunes, pero no es adecuada para personas con apellidos comunes. El creador del perfil puede ser consciente o no de que el perfil contiene documentos que no deberían aparecer. Esto, por supuesto, afecta a los indicadores a nivel de autor que GS calcula automáticamente a partir de los documentos que aparecen en el perfil. La manera más fácil de evitar este problema es cambiar la configuración de actualización a "confirmar cambios". De esta manera, los creadores de perfiles son notificados cuando GS encuentra un documento que piensa que debería añadirse al perfil, y el usuario puede confirmar o rechazar la actualización. Sin embargo, esta opción no es utilizada por muchos usuarios, que en ocasiones crean el perfil y no lo visitan regularmente o directamente lo dejan abandonado, contribuyendo a una representación inexacta de los méritos de los autores.
- Unión incorrecta de documentos: como los usuarios pueden gestionar sus publicaciones como quieren, algunos perfiles podrían contener documentos unidos aunque realmente no hubieran debido ser unidos
- Perfiles donde se pueden encontrar documentos y citas deliberadamente manipuladas.

En el curso de nuestros análisis hemos encontrado muchos de estos errores. Una descripción de los mismos puede encontrarse en Martín-Martín et al. (2014a) (capítulo 2 de esta tesis). Martín-Martín. Ayllón. Delgado López-Cózar. & Orduna-Malea (2015) (capítulo 4 de esta tesis). y Martín-Martín. Orduna-Malea. Ayllón. & Delgado-López-Cózar (2016) (capítulo 10 de esta tesis). Una revisión más extensa de los errores que GS comete puede encontrarse en Orduna-Malea et al. (2017).

Referencias

Beel. J.. & Gipp. B. (2009). Google Scholar's ranking algorithm: The impact of citation counts (An empirical study). In *2009 Third International Conference on Research Challenges in Information Science* (pp. 439–446). IEEE. https://doi.org/10.1109/RCIS.2009.5089308

Clarivate Analytics. (2015). Web of Science & Google Scholar collaboration. Retrieved June 5. 2018. from http://wokinfo.com/googlescholar/

Delgado López-Cózar. E.. Lorenzo-Sar. V.. Martín-Martín. A.. & Ayllón. J. M. (2015). Classic Scholars' Profiles: Bibliometrics & Scientometrics. Retrieved April 1. 2017. from http://www.classic-scholars-profiles.infoec3.es/bibliometrics

Delgado López-Cózar. E.. Martín-Martín. A.. Orduña-Malea. E.. & Ayllón. J. M. (2015a). La Biblioteconomía y Documentación española según Google Scholar Citations. Retrieved April 1. 2017. from http://www.biblioteconomia-documentacion-española.infoec3.es

Delgado López-Cózar. E.. Martín-Martín. A.. Orduña-Malea. E.. & Ayllón. J. M. (2015b). Scholar Mirrors: Bibliometrics. Scientometrics. Informetrics. Webometrics. and Altmetrics in Google Scholar Citations. ResearcherID. ResearchGate. Mendeley. and Twitter. Retrieved April 1. 2017. from http://www.scholar-mirrors.infoec3.es

Delgado López-Cózar. E.. Orduna-Malea. E.. & Martín-Martín. A. (2019). Google Scholar as a data source for research assessment. In W. Glaenzel. H. Moed. U. Schmoch. & M. Thelwall (Eds.). *Springer Handbook of Science and Technology Indicators*. Springer.

Delgado López-Cózar. E.. Robinson-García. N.. & Torres-Salinas. D. (2014). The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*. *65*(3). 446–454.

https://doi.org/10.1002/asi.23056

Google Scholar. (2019). Google Scholar Help. Inclusion Guidelines for Webmasters. Retrieved February 15. 2019. from https://scholar.google.es/intl/en/scholar/inclusion.html#content

Harzing. A.-W. (2016). Sacrifice a little accuracy for a lot more comprehensive coverage. Retrieved from https://harzing.com/blog/2016/08/sacrifice-a-little-accuracy-for-a-lot-more-comprehensive-coverage

Jacsó. P. (2006). Deflated. inflated and phantom citation counts. *Online Information Review*. *30*(3). 297–309. https://doi.org/10.1108/14684520610675816

Jacsó. P. (2008). Google Scholar revisited. *Online Information Review*. *32*(1). 102–114. https://doi.org/10.1108/14684520810866010

Jacsó. P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*. *34*(1). 175–191. https://doi.org/10.1108/14684521011024191

Khabsa. M.. & Giles. C. L. (2014). The number of scholarly documents on the public web. *PloS One*. *9*(5). e93949. https://doi.org/10.1371/journal.pone.0093949

Martín-Martín. A.. Ayllón. J. M.. Delgado López-Cózar. E.. & Orduna-Malea. E. (2015). Nature 's top 100 Re-revisited. *Journal of the Association for Information Science and Technology*. *66*(12). 2714–2714. https://doi.org/10.1002/asi.23570

Martín-Martín. A.. Costas. R.. van Leeuwen. T.. & Delgado López-Cózar. E. (2018a). Dataset: sources of free full text found by Google Scholar for documents in Web of Science published in 2009 and 2014 (raw and aggregated). https://doi.org/10.17605/OSF.IO/FSUJY

Martín-Martín. A.. Costas. R.. van Leeuwen. T.. & Delgado López-Cózar. E. (2018b). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*. *12*(3). 819–841. https://doi.org/10.1016/j.joi.2018.06.012

Martín-Martín. A.. Orduna-Malea. E.. Ayllón. J. M.. & Delgado-López-Cózar. E. (2016). *The counting house. measuring those who count: Presence of Bibliometrics. Scientometrics. Informetrics. Webometrics and Altmetrics in GSC (Google Scholar Citations). ResearcherID. ResearchGate. Mendeley. & Twitter* (EC3 Working Papers No. 21). Retrieved from https://arxiv.org/abs/1602.02412

Martín-Martín. A.. Orduña-Malea. E.. Ayllón. J. M.. & Delgado-López-Cózar. E. (2014a). *Does Google Scholar contain all highly cited documents (1950-2013)?* (EC3 Working Papers No. 19). Retrieved from http://arxiv.org/abs/1410.8464

Martín-Martín. A.. Orduña-Malea. E.. Ayllón. J. M.. & Delgado-López-Cózar. E. (2014b). Highly Cited Documents on Google Scholar (1950-2013). https://doi.org/10.6084/m9.figshare.1224314

Martín-Martín. A.. Orduna-Malea. E.. Ayllón. J. M.. & Delgado López-Cózar. E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentacion Cientifica*. *39*(4). e149. https://doi.org/10.3989/redc.2016.4.1405

Martín-Martín. A.. Orduna-Malea. E.. & Delgado López-Cózar. E. (2018). Coverage of highly-cited documents in Google Scholar. Web of Science. and Scopus: a multidisciplinary comparison. *Scientometrics*. *116*(3). 2175–2188. https://doi.org/10.1007/s11192-018-2820-9

Martin-Martin. A.. Orduna-Malea. E.. Harzing. A.-W.. & Delgado López-Cózar. E. (2017). Can we use Google Scholar to identify highly-cited documents? *Journal of Informetrics*. *11*(1). 152–163. https://doi.org/10.1016/j.joi.2016.11.008

Martín-Martín. A.. Orduna-Malea. E.. Thelwall. M.. & Delgado López-Cózar. E. (2018). Google Scholar. Web of Science. and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*. *12*(4). 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

Orduña-Malea. E.. Ayllón. J. M.. Martín-Martín. A.. & Delgado-López-Cózar. E. (2014). *About the size of*

Google Scholar: playing the numbers (EC3 Working Papers No. 18). Retrieved from http://arxiv.org/abs/1407.6239

Orduna-Malea. E.. Ayllón. J. M.. Martín-Martín. A.. & Delgado López-Cózar. E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*. *104*(3). 931–949. https://doi.org/10.1007/s11192-015-1614-6

Orduña-Malea. E.. Martín-Martín. A.. Ayllón. J. M.. & Delgado López-Cózar. E. (2016). *La revolución Google Scholar : Destapando la caja de Pandora académica*. Granada: Universidad de Granada y Unión de Editoriales Universitarias Españolas.

Orduna-Malea. E.. Martín-Martín. A.. & Delgado López-Cózar. E. (2017). Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors. *Revista Española de Documentación Científica*. *40*(4). e185. https://doi.org/10.3989/redc.2017.4.1500

Van Noorden. R. (2014. November 7). Google Scholar pioneer on search engine's future. *Nature*. https://doi.org/10.1038/nature.2014.16269

Van Noorden. R.. Maher. B.. & Nuzzo. R. (2014). The top 100 papers. *Nature*. *514*(7524). 550–553. https://doi.org/10.1038/514550a

Wesley. B. (2016). Google Scholar is Filled with Junk. Retrieved February 17. 2019. from http://ipensatori.com/2016/06/27/google-scholar-is-filled-with-junk/

Wilke. C. (2014. November 1). The Google Scholar preprint bug. *The Serial Mentor [Blog]*. Retrieved from https://serialmentor.com/blog/2014/11/1/the-google-scholar-preprint-bug

Chapter 2. Does Google Scholar contain all highly cited documents (1950-2013)?

Working paper. Cite as:

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2014). Does Google Scholar contain all highly cited documents (1950-2013)? (EC3 Working Papers No. 19). Retrieved from http://arxiv.org/abs/1410.8464

Abstract (English)

The study of highly cited documents on Google Scholar (GS) has never been addressed to date in a comprehensive manner. The objective of this work is to identify the set of highly cited documents in Google Scholar and define their core characteristics: their languages, their file format, or how many of them can be accessed free of charge. We will also try to answer some additional questions that hopefully shed some light about the use of GS as a tool for assessing scientific impact through citations.

The decalogue of research questions is shown below:

1. Which are the most cited documents in GS?
2. Which are the most cited document types in GS?
3. What languages are the most cited documents written in GS?
4. How many highly cited documents are freely accessible?
    a. What file types are the most commonly used to store these highly cited documents?
    b. Which are the main providers of these documents?
5. How many of the highly cited documents indexed by GS are also indexed by WoS?
6. Is there a correlation between the number of citations that these highly cited documents have received in GS and the number of citations they have received in WoS?
7. How many versions of these highly cited documents has GS detected?
8. Is there a correlation between the number of versions GS has detected for these documents, and the number citations they have received?
9. Is there a correlation between the number of versions GS has detected for these documents, and their position in the search engine result pages?
10. Is there some relation between the positions these documents occupy in the search engine result pages, and the number of citations they have received?

To answer these questions, a set of 64,000 documents indexed in Google Scholar has been collected, after performing 64 queries by year (from 1950 to 2013) using Google Scholar's advanced search, and collecting the maximum number of records that GS displays for any given query, which as we know is always 1,000. These 64,000 documents receive 122,245,865 citations in Google Scholar and 35,182,077 in Web of Science Core Collection.

Abstract (Spanish)

Hasta ahora nunca se había estudiado de manera exhaustiva a los documentos altamente citados según Google Scholar (GS). El objetivo de este trabajo es identificar el conjunto de los documentos altamente citados en Google Scholar y definir sus principales características: idioma, formato de archivo, a cuántos se puede acceder gratuitamente. También intentaremos responder algunas preguntas adicionales que quizás puedan arrojar un poco de luz sobre el uso de GS como una herramienta con la que evaluar el impacto de las publicaciones científicas a través de sus citas.

El decálogo de preguntas es el siguiente:

1. ¿Cuáles son los documentos más altamente citados en GS?
2. ¿Cuáles son los tipos documentales más citados en GS?
3. ¿En qué idioma están escritos los documentos más altamente citados en GS?
4. ¿A cuántos documentos altamente citados se puede acceder gratuitamente?
    a. ¿Qué tipos de archivo son los más comunes para almacenar estos documentos altamente citados?
    b. ¿Cuáles son las principales plataformas que proporcionan estos documentos?
5. ¿Cuántos de estos documentos altamente citados en GS están también indizados en WoS?
6. ¿Existe una correlación entre el número de citas que estos documentos han recibido en GS, y el número de citas que han recibido según WoS?
7. ¿Cuántas versiones de estos documentos altamente citados ha detectado GS?
8. ¿Existe una correlación entre el número de versiones que GS ha detectado de estos documentos, y el número de citas que han recibido?
9. ¿Existe una correlación entre el número de versiones que GS ha detectado de estos documentos, y la posición que tienen en la página de resultados del buscador?
10. ¿Existe una correlación entre las posiciones que estos documentos ocupan en la página de resultados del buscador, y el número de citas que han recibido?

Para responder a estas preguntas, se extrajo un listado de 64,000 documentos indizados en Google Scholar, tras la realización de 64 consultas avanzadas en Google Scholar en las que se establecieron los años entre el rango 1950-2013. Para cada consulta se extrajo el listado de los 1,000 resultados que GS muestra como máximo para cada consulta. Estos 64,000 documentos habían recibido 122,245,865 citas en Google Scholar y 35,182,077 citas en la colección principal de Web of Science.

1. Introduction

1.1. About this title

The reason behind the title of this work and its structure as questions is not simply a rhetorical device intended to attract the reader's attention. It is a genuine statement of intentions, since there is no absolute empirical certainty that our sample contains all the highly cited documents present in Google Scholar (GS) at the moment we collected the data. If GS provided a feature that allowed us to sort documents according to number of citations, as traditional bibliometric databases do (Web of Science and Scopus), we wouldn't harbor any doubts about this matter. Since this is not the case, we cannot be completely sure that when we make a query by year of publication in GS, it will show us the 1,000 most cited documents published during that range of years (as we know, 1,000 is the maximum number of results GS will display for any given query). In short, we are not entirely sure that the data we collected comprises only highly cited documents in GS, and therefore it is likely that some of these documents don't actually belong to the group of "upper crust" documents in GS for each of the years in the selected range (1950-2013).

Nevertheless, there is strong evidence suggesting that our sample contains a very large portion of the highly cited documents in GS:

Firstly, in its documentation, GS explicitly declares that the number of citations received by a document is one of the factors involved in the calculation of the position this document will occupy on the results page, although they don't specify the overall weight of this factor in the calculation. A high correlation between the position documents occupy in the search engine results page (SERP) when they are sorted by Google Scholar's default relevance criteria, and the position they occupy when they are sorted simply by their number of citations (See question 10, Figure 24) would confirm that citation count is indeed the factor that is given the highest weight in Google Scholar's ranking algorithm, and therefore it would be safe to presume that the first positions of a query will always be occupied by the most cited documents that satisfy said query.

Secondly, we can see other evidences that support the validity of our sample: in order to verify that the documents in our sample were in fact highly cited documents, we retrieved the top 1,000 most cited documents on the Web of Science Core Collection for each year in the range 1950-2013 (as of October the 30th 2014), and compared the two sets of documents for each year. The results showed that, on average, 81% of the documents in our sample from GS with a link to a WoS record were also present in the ranking of the top 1,000 most cited documents in WoS. With the WoS dataset, we could also learn how many highly cited documents in WoS were missing from our GS dataset. In this respect, the results show that the number of highly cited documents in WoS that are not present in our GS sample is insignificant. There are only 396 (1.3%) documents in our WoS sample that have received enough citations to be included among the 30,000 most cited documents in our GS sample, but that according to their document ID are not present in this sample. Likewise, if we consider the 40,000 most cited documents in our sample, this figure raises to 1,645 (4.1%). As we lower the citation threshold, this figure obviously increases (See Question 1). This result seems logical for two reasons:

a) factor ranking: citations are the main ranking factor but not the only one. Therefore, for documents with the highest number of citations, the position achieved clearly correlated with citations. In contrast, in the lower positions, where the number of citations is also lower, the effect of other ranking factors is more evident.

b) statistical noise: in the first positions, the differences between the documents in terms of citations are high, so the statistical error must be very large to obtain documents in wrong positions. However, as we approach the border cut (1,000 documents), the differences between the documents are smaller, and therefore small errors can result in significant changes in positions over the lower ranks (especially for positions in the margin 800-1,200).

Lastly, our own experience, gained through the daily observation of hundreds of searches. Usually, the relevance ranking used by GS is reduced to simply placing the highest cited documents in the first results pages, with very rare exceptions. This is something anyone can check just by doing a search in Google Scholar. We encourage researchers to experience this for themselves.

To sum up, in this work we analyse the 1,000 documents that GS retrieves for each one of 64 queries by year, from 1950 until 2013. Presumably, among them we should be able to find the most cited documents published in each of those years.

1.2. Citation Classics: Highly Cited Documents

The idea of identifying the most influential documents in science using the number of citations they generate in the scientific literature was introduced, like many other bibliometric tools, by Eugène Garfield. On January 3rd 1977, Garfield published an essay entitled "Introducing Citation Classics: the human side of scientific papers" (1977), which appeared in Current Contents. The candidates for Citation Classics were selected from a group of 500 most cited papers during the years 1961-1975. Many of these had been listed before in Current Contents. From 1977 to 1993, 400 Citation Classic Commentaries were published in Current Contents. The full texts of these mostly one-page articles are now available in an open access server at http://garfield.library.upenn.edu/classics.html.

From 2001, the Highly Cited Papers were integrated in a new product from Thomson Scientific: the Essential Science Indicators. Neither Scopus nor other databases have released alternatives to this product.

What we do have is an extensive scientific literature, published during the last few decades, on the matter of highly cited documents in different journals, subject areas, institutions or countries (Oppenheim & Renn 1978; Narin & Frame 1983; Plomp 1990; Glänzel & Czerwon 1992; Glänzel, & Schubert 1992a-b; Glänzel et al. 1995; Tijssen et al. 2002; Aksnes 2003; Aksnes & Sivertsen 2004; Kresge et al. 2005; Levitt & Thelwall 2009; Smith 2009; Persson 2010). Recently, the need of ranking any product of scientific activity according to its citation performance has caused the emergence of this kind of classifications (top 1%, 10%, 15%). The calculation of percentiles, previously proposed explicitly by Maltrás (2003), has recently been rediscovered by other authors (Bornmann 2010, Bornmann & Mutz 2011, Bornmann et al. 2011).

The appearance of Google Scholar opened up new possibilities in this field. Its birth at the end of 2004 signaled a revolution in the way scientific publications were searched, retrieved and accessed (Jacsó, 2005).

From the get-go, GS became not only a search engine for scientific and academic documents, but also for the citations these documents receive. Although it took five years to get over its "beta" stage, today we can say without a doubt that GS is not only the largest database of scientific, academic and technical information in the world (Orduña-Malea et al., 2014, Ortega 2014), but also the richest and most varied, since Google's crawlers systematically parse and process the whole academic web, not making distinctions based on subject areas, languages, or countries (Ortega 2014). Despite the limitations of its spiders and processing software, the lack of normalization processes and quality control filters, GS is an irreplaceable source of global scientific knowledge.

Studies about GS have been limited to: a) explain how it works, its features, limitations, errors, etc.; b) define its coverage and size; c) compare the number of citations received by documents of a given subject area in GS, to the citations they receive in other databases; and d) its growth and evolution over time. However, the study of highly cited documents regardless of their discipline or field has never been addressed in a comprehensive manner.

Therefore, the objective of this work is to identify the set of highly cited documents in GS and define their core characteristics: language, file format, and how many of them can be accessed to free of charge. We will also try to answer some additional questions that - hopefully - shed some light about the use of GS as a tool for assessing impact through citations.

In short, we intend to answer the following questions:

1.3. Research Questions

1. Which are the most cited documents in GS?
2. Which are the most cited document types in GS?
3. In what languages are the most cited documents in GS written?
4. How many highly cited documents are freely accessible?
    a. What file types are the most commonly used to store these highly cited documents?
    b. Which are the main providers of these documents?
5. How many of the highly cited documents indexed by GS are also indexed by WoS?
6. Is there a correlation between the number of citations that these highly cited documents have received in GS and the number of citations they have received in WoS?
7. How many versions of these highly cited documents has GS detected?

8. Is there a correlation between the number of versions GS has detected for these documents, and the number citations they have received?
9. Is there a correlation between the number of versions GS has detected for these documents, and their position in the search engine result pages?
10. Is there some relation between the positions these documents occupy in the search engine result pages, and the number of citations they have received?

2. Materials and Methods

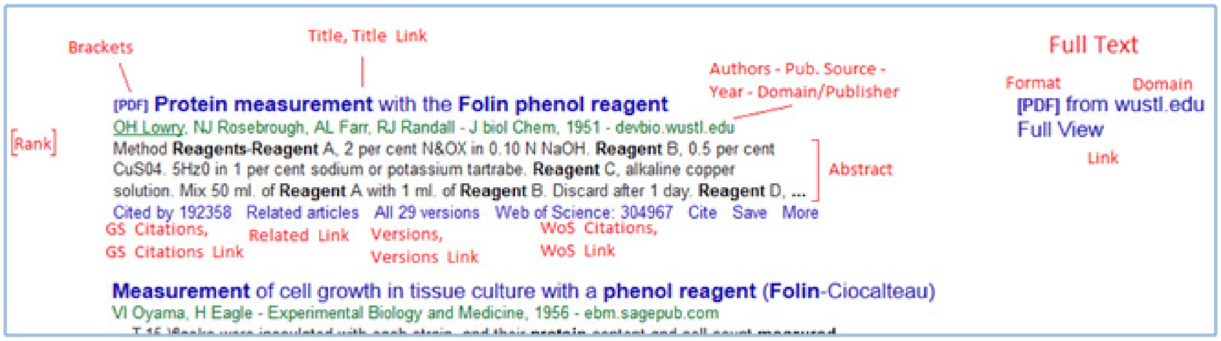
This longitudinal study describes a set of 64,000 documents indexed in Google Scholar, obtained after performing 64 queries by year (from 1950 to 2013) using Google Scholar's advanced search, and collecting the maximum number of records that GS displays for any given query, which as we know is always 1,000.

This process was carried out twice, with a few days between the first and the second download processes. In one case, it was done from a computer connected to our university's IP range (to obtain WoS data embedded in GS), and in the other case, from a computer with a normal Internet connection (to obtain data about open access links unadulterated by our university's subscriptions). Besides, this also worked as a reliability check, because we confirmed that the two datasets contained the same records. These processes took place on the 28th of May and 2nd of June, 2014.

We downloaded the source HTML code for each of the result pages in our queries, parsed them to extract all the relevant information, and saved it in spreadsheet, which is a format more appropriate for the analysis of data. The fields extracted were the following (Figure 1):

- **Publication year:** It is the year that was used in the query, and not that contained in the bibliographical description of the record retrieved.
- **Rank:** The position that each document occupies in the search engine results page of GS.
- **Full Text:** Only marked when GS found a freely accessible version of the document. Then, some additional fields were obtained:
    - **Domain:** The domain where GS has found a full text version of the document.
    - **Link:** Link to the full text of the document.
    - **Format:** File type of the full text version of the document.
- **Brackets:** Some records display text in square brackets before the title of the document. The most common occurrences are: "[BOOK]" (the record is a book) and "[CITATION]" (the record has only been found in the reference list of another document), "[PDF]" and "[HTML]" (to indicate that the document has been found in those formats).
- **Title:** Title of the document.
- **Title Link:** The URL pointing to where the record has been found (it is not a link to a freely accessible version of the full text, since the document may be behind a paywall).
- **Authors – Publication Source – Year – Domain/Publisher:** This field contains information about the authors, publication source, year of publication, and publisher of each document. However, not all this information is always displayed for all records, and it is usually cropped to fit one line.
    -
    - **Publication source:** Name of the source where the document has been published, and, sometimes, publication details (volume, issue, pages). This information is not always displayed, and when it is, it's not always complete.
    - **Year:** year when the document was published. This field has been proved to correspond with the field "Publication year", previously described.

  - **URL domain / Publisher:** Domain where this document has been found, or, sometimes, the name of its publisher (only for big publishers).
- **Abstract:** First lines of the abstract (it is also cropped to fit a fixed space).
- **GS Citations:** Number of citations the document has received according to GS.
- **Link to GS Citations:** URL pointing to the list of citing documents in Google Scholar.
- **Link to Related documents:** URL pointing to the list of related documents.
- **Versions:** Number of versions GS has found of the documents.
- **Link to Versions:** URL pointing to the list of versions GS has found of the same document.
- **Web of Science:** This data will only appear if the query is performed from a computer connected to an IP range with access to Thomson Reuters' Web of Science, and only for the documents that are indexed both in GS and WoS.
  - **WoS Citations:** Number of citations according to Web of Science.
  - **WoS accession number (UT)**: identification number of the document in Web of Science. This code allows us to accurately match a GS record with a WoS record.
  - **WoS Link:** URL pointing to the list of citing documents in Web of Science.


[PDF] **Protein measurement** with the **Folin phenol reagent**

OH Lowry, NJ Rosebrough, AL Farr, RJ Randall - J biol Chem, 1951 - devbio.wustl.edu

Method **Reagents-Reagent** A, 2 per cent N&OX in 0.10 N NaOH. **Reagent** B, 0.5 per cent CuS04. 5Hz0 in 1 per cent sodium or potassium tartrabe. **Reagent** C, alkaline copper solution. Mix 50 ml. of **Reagent** A with 1 ml. of **Reagent** B. Discard after 1 day. **Reagent** D, ...

Cited by 192358 Related articles All 29 versions Web of Science: 304967 Cite Save More

[PDF] from wustl.edu

Full View

Measurement of cell growth in tissue culture with a **phenol reagent** (**Folin**-Ciocalteau)

VI Oyama, H Eagle - Experimental Biology and Medicine, 1956 - ebm.sagepub.com

Figure 1. Fields extracted from Google Scholar's SERP

In addition to these fields, we added a few more in order to answer our questions related to: **type of the source publication**, and **language** of the document.

Given the difficulty of ascertaining the typologies of the documents indexed in Google Scholar (this information is not systematically provided by the search engine), we have devised three different strategies that, combined, have allowed us to know the type of a large portion of documents in our data set:

a) All documents where the field **Brackets** = **"**[BOOK]" have been considered as books (codified as "B").

b) For documents that were also indexed in WoS, GS data was merged with WoS data to obtain the document types. The correspondence is as follows:

  - Journal ("J"): "Article", "Letter", "Note", "Reviews".
  - Book ("B"): "Book", "Book Chapter".
  - Conference Proceedings ("C"): "Proceedings Papers".
  - Others ("O"): "Book Review", "Correction", "Correction, Addition", "Database Review", "Discussion", "Editorial Material", "Excerpt", "Meeting Abstract", "News Item", "Poetry", "Reprint", "Software Review".

c) Lastly, we analysed the **publication source** (where possible), searching for keywords that could indicate the type of the source publication:

- Journal (“J”): “Revista”, “Anuario”, “Cuadernos”, “Journal”, “Revue”, “Bulletin”, “Annuaire”, “Anales”, “Cahiers”,”Proceedings”[15].
- Conference Proceedings (“C”): “Proceedings”, “Congreso”, “Jornada”, “Seminar”, “Simposio”,”Congrès”, “Conference”, “symposi”, “meeting”.

Combining these three strategies, we identified the document type for 71% of the 64,000 documents in our sample. We couldn’t identify the document types for the remaining 29% because this would have required doing it manually for 18,590 documents, which would have taken an excessive amount of time. This information was saved in a new field called **Source Type**, and was codified as follows:

- B: Books or book chapters.
- J: Journal articles, reviews, letters and notes.
- C: Conference proceedings.
- O: Others (meeting abstracts, corrections, editorial material…).
- Unknown: we haven’t been able to assign a source type (29% of the sample).

As regards the language of the documents (GS doesn’t provide this information either), we used the language in which the title and abstract of the document were written, as well as WoS data (when available) as a basis for a new **Language** field.

In essence, we will show a sectional view (global results) as well as a longitudinal view (results by year, in order to detect potential changes) of this sample of documents.

The measures we have used to summarise the data are: absolute and relative frequencies of various aspects of the documents (questions 1-5), and the Pearson correlation (questions 6-10), with $p \leq 0.01$.

3. Results

The structure we have followed to present the results of each research question is as follows: first we describe the results we have obtained, and after that, under a separate heading called “Discussion & limitations”, we lay out and discuss possible inquiries and uncertainties raised by these findings.

Question 1. Which are the most cited documents in Google Scholar?

In Table 1 we present the top 25 most cited documents in Google Scholar. Additionally, Appendix A shows the top 1% most cited documents in our sample (a total of 640 documents).

These lists are a faithful reflection of the all-encompassing indexing policies of Google Scholar: the academic/scientific/technical world against the scientific world displayed in traditional citation-based databases. In this respect, we can state that GS offers an original and different vision as regards what the most influential documents in the academic/scientific world are, from the perspective of their citation count. This is caused by several reasons:

First, its coverage is not limited to seminal research works in the entire spectrum of scientific fields, but it also covers greatly influential works directed not only to researchers but also to people who are training to

[15] The word ”Proceedings” is used both for journals (i.e. “Proceedings of the National Academy of Sciences”) and for conference proceedings (i.e., “Proceedings of the 4th Conference…”). Initially, records containing this word in the “**Publication Source”** field were all considered as conference proceedings, but a manual check was carried out to reassign those that were really journal articles.

become researchers or practitioners in their respective fields. This is testified by the presence of statistical manuals (*Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables; Biostatistical Analysis; Statistical Power Analysis for the Behavioral Scienc*e), laboratory manuals (*Molecular cloning: a laboratory manual)*, manuals of research methodology (*Case study research: Design and methods*), and works that have become a *de facto* standard in professional practice (*Diagnostic and statistical manual of mental disorders, Numerical recipes: the art of scientific computing; Genetic algorithms in search, optimization, and machine learning*).

Second, a high proportion of the highly cited documents are books (a document type that is essential in the humanities and the social sciences as a vehicle for the communication of new results, and in the experimental sciences as a way to consolidate and disseminate knowledge). In fact, 62% of the top 1% most cited documents in our sample are books (see Appendix A). Moreover, books are the document type with a highest citation average: 2,700, against an average of 1,700 in journal articles, and 2,200 for conference proceedings. The importance of books and conference proceedings is therefore thoroughly proven.

Although the ranking is dominated by studies from the natural sciences, and within those, especially the life sciences, it also contains many works from the social sciences, especially from economics, psychology, sociology, education… and also from the Humanities (philosophy and history). For instance: *The structure of scientific revolutions; Diffusion of innovations;* and *Imagined communities. Reflections on the origin and spread of nationalism*).

Many of the works leading this ranking are clearly methodological in nature: they describe the steps of a certain procedure or how to handle basic tools to process and analyse all kinds of data. Precisely because they are essential to researchers, they reach such a high number of citations. This phenomenon is widely known in bibliometrics, where it has already been observed that works that deal with new data collecting and processing techniques or methodologies are more likely to receive a great number of citations.

Even though, as we comment before, GS presents a very different ranking of highly cited academic documents compared to the rankings offered by the traditional citation-based databases, in other aspects it presents a very similar portrait of the world of research to the one offered in traditional databases. This is so because the most cited scientific documents in GS match very closely with those that have been already identified as highly cited in the Web of Science (Garfield, 2005). This explains the high correlation found in the rankings of documents according to their number of citations in GS and WoS (See Question 6).

Therefore, it is not surprising that the most cited document according to GS is the already famous article written by Lowry, "*Protein measurement with the Folin phenol reagent"* published in 1951 in the *Journal of Biological Chemistry,* where he developed a new method to measure the concentration of a protein in a solution. The reasons for the success of this article were revealed by the author himself (Lowry, 1977), and in a short note published in the same journal on the occasion of its hundredth anniversary in 2005 (Kresge et al., 2005).[16]

We'll use this article as an example in the next section to comment some uncertainties and discuss the possible limitations of these results.

[16] See his profile on Google Scholar:

http://scholar.google.com/citations?user=YCS0XAcAAAAJ&hl=es

Table 1. Top 25 most cited documents in Google Scholar (1950-2013)

Document type	Bibliographic reference	1st ed. Pub. Year	GS Citations
J	LOWRY, O.H. et al., (1951). Protein measurement with the Folin phenol reagent.The Journal of biological chemistry, 193(1), 265-275.	1951	253671
J	LAEMMLI, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature, 227(5259), 680-685. DOI: 10.1038/227680a0	1970	221680
J	BRADFORD, M.M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein using the principle of protein dye binding. Analytical Biochemistry, 72, 248-254. DOI: 10.1006/abio.1976.9999	1976	185749
B	SAMBROOK, J., FRITSCH, E. F., & MANIATIS, T. (1982). Molecular cloning: a laboratory manual. New York, Cold Spring Harbor Laboratory Press.	1982	171004
B	AMERICAN PSYCHIATRIC ASSOCIATION. (1952). Diagnostic and statistical manual: mental disorders. Washington, American Psychiatric Assn., Mental Hospital Service.	1952	129473
B	PRESS, W. H. (1986). Numerical recipes: the art of scientific computing. Cambridge [Cambridgeshire], Cambridge University Press.	1986	108956
B	YIN, R. K. (1984). Case study research: design and methods. Beverly Hills, Calif, Sage Publications.	1984	82538
B	ABRAMOWITZ, M., & STEGUN, I. A. (1964). Handbook of mathematical functions: with formulas, graphs, and mathematical tables. Washington, Government printing office.	1964	80482
B	KUHN, T. S. (1962). The structure of scientific revolutions. Chicago, University of Chicago Press.	1962	70662
B	ZAR, J. H. (1974). Biostatistical analysis. Englewood Cliffs, Prentice Hall international.	1974	68267
J	SHANNON, C.E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423.	1948	66851
J	CHOMCZYNSKI, , & SACCHI, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Analytical Biochemistry, 162, 156-159. DOI: 10.1006/abio.1987.9999	1987	63871
J	SANGER F, NICKLEN S, & COULSON AR. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 74, 5463-7. DOI: 10.1073/pnas.74.12.5463	1977	63767
B	COHEN, J. (1969). Statistical power analysis for the behavioral sciences. New York, Academic Press.	1969	63766
B	GLASER, B. G., & STRAUSS, A. L. (1967). The discovery of grounded theory: strategies for qualitative research. New York, Aldine de Gruyter.	1967	61158
B	NUNNALLY, J. C. (1967). Psychometric Theory. New York , McGraw-Hill.	1967	60725
B	GOLDBERG, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Reading, Mass, Addison-Wesley Pub. Co.	1989	59764

Discussion & Limitations

How confident are we that the 64,000 documents that make up our sample really contain the most cited documents in GS?

Although there are certain evidences that suggest that we have been able to collect the vast majority of the most cited documents in GS between 1950 and 2013 (as of the 28th of May 2014), as we already explained at the beginning of this study (see Introduction), there are still some questions that should be cleared up.

To this end, first we have tried to find out if any of the documents in our sample aren't really highly cited documents, and second, if there are any highly cited documents that haven't been included in our sample. To do this, we have compared the 1,000 most cited papers in GS against the 1,000 most cited papers in WoS between 1950 and 2013 (Figure 2).

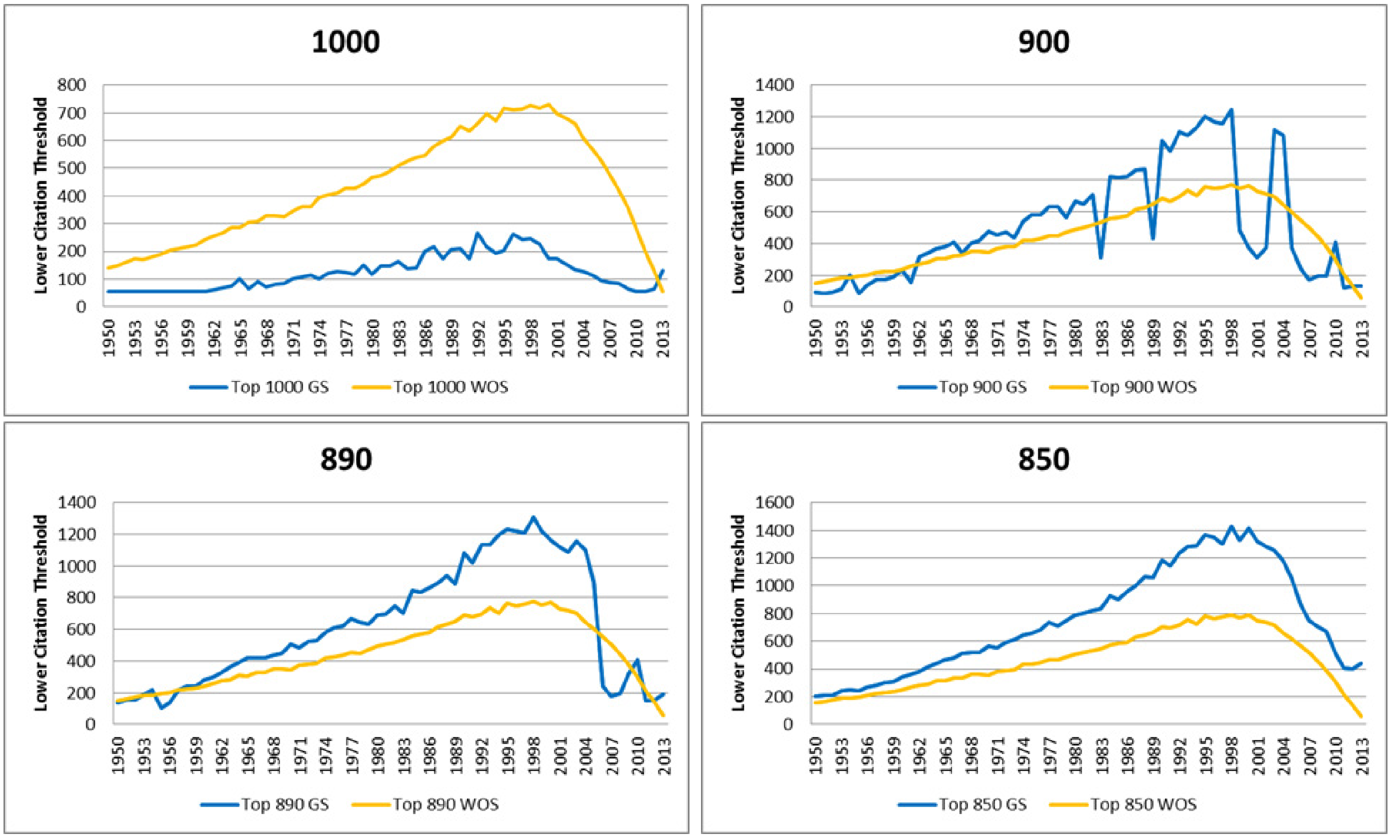


Figure 2. Minimum number of citations received by top cited (1,000, 900, 890, 850) documents in Google Scholar and WoS (1950-2013)

On the one hand, we have detected that the results displayed by GS to our queries become extremely erratic in terms of their citation count from about the 900th result onwards. This means that it is highly probable that approximately the last 100 documents for each year in our sample (a total of 6,400 documents) aren't actually highly cited documents, and therefore should be excluded from the sample.

In contrast, we also have checked that some documents in WoS with a number of citations that slightly exceed the threshold set by the 1,000 documents returned by GS, are not present in the first 1,000 results of the search engine.

Nonetheless, all these inconsistencies happen in the last 100 positions of each query for each year, whereas in the first 900 the consistency is high. To sum up, despite the various limitations described above, we can affirm that the majority of the documents in our sample are highly cited documents.

In order to be able to trust the results that our search strategy yielded, we must ask ourselves if the documents in our sample were really published in the year GS says they were published.

To answer this question we carried out two different tests. In the first place, we tested the internal consistency of the search engine. We checked if the results displayed by GS met the requirements of our query. We found that the year of publication of the documents according to GS matched the year we entered in our query in practically 100% of the cases. Only two records out of 64,000 displayed a different year to the one we typed in the search box.

Secondly, we tested the external consistency. For those documents that had been linked to a WoS record (32,680 out of 64,000), we compared the publication year according to GS to the one displayed in the WoS record. Since WoS is a controlled database with a minimum error rate as regards its bibliographic information, we have used it as a benchmark. The results showed that the publication years in GS and WoS matched in 96.7% of the cases (31,600 documents). Curiously enough, the years where we detected more mismatches were 2012 and 2013. Consequently, we must conclude that the error rate in the publication years is very low for this subset of the sample.

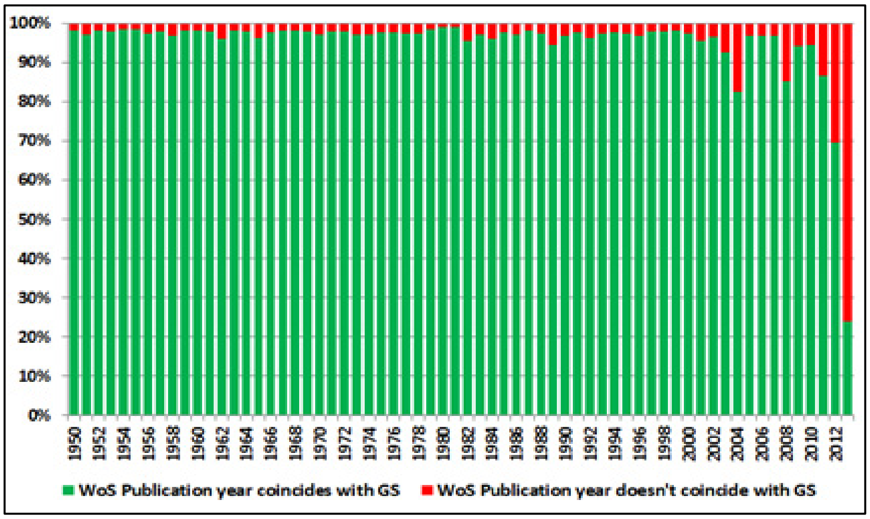


Figure 3. Publication year mismatches between journal articles in Google Scholar and Web of Science

However, we have observed that, in the case of books, Google lumps together all the different editions of a same book, and systematically selects the latest edition of the book as the primary version. As a result, GS takes the publication date of the last edition (and not the publication date of the first edition) as the publication date of the book. This decision, as understandable as it is from a search point of view (users will probably want to access the latest edition of a book), obviously affects our sample. In Figure 4, the frequency distributions for both the publication year of the top 600 most cited books in our sample according to Google Scholar, and the publication year of the 1st edition of these books are displayed.

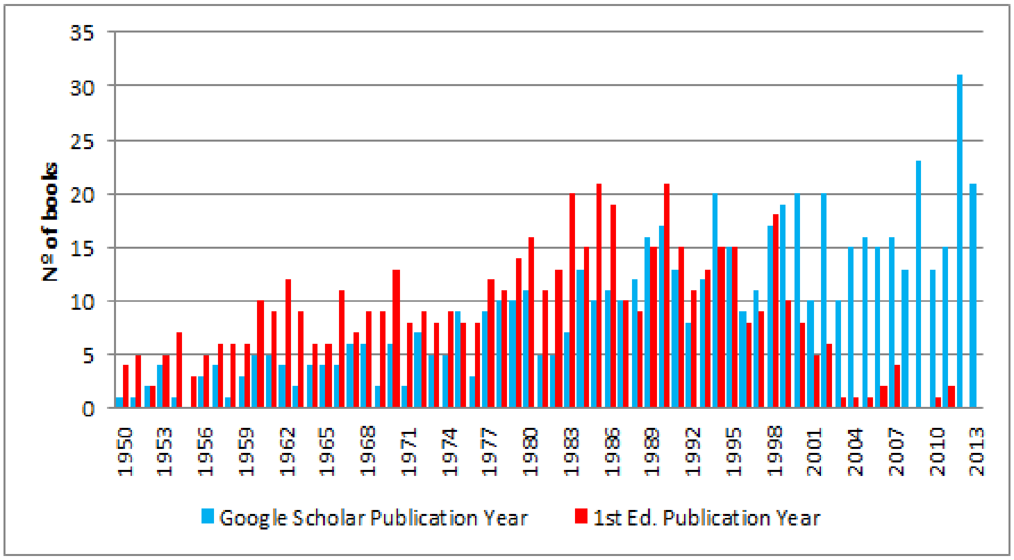
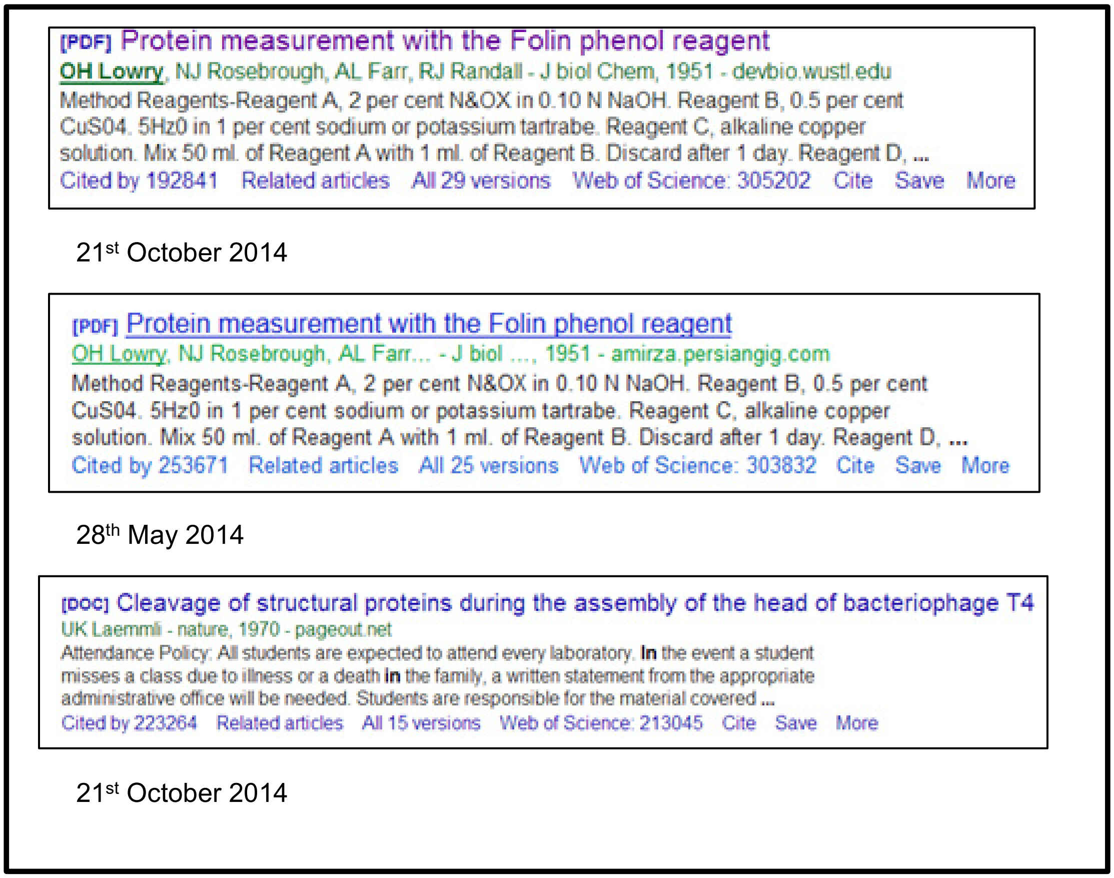


Figure 4. Differences between the publication year of the top 600 most cited books according to Google Scholar, and the publication year of the 1st edition of these books

In any case, it should be noted that this limitation doesn't affect the status of these books as highly cited documents, only the year of publication assigned to them[17]. Moreover, this fact may be the reason behind the higher number of books in the last five year of the sample (see Question 2).

When some time after collecting our sample, we checked again the number of citations to Lowry's article, we were taken by surprise by the result we found. As of the 21st of October, 2014, this study had 192,841 citations according to GS (Figure 5 top). However, on the 28th of May, 2014, when we collected our sample, this number was 253,671 (figure 5 middle). This means than within 5 months, Lowry's article has lost nothing less than 60,000 citations. Therefore, right now, it is not the highest cited article in GS, giving way to Laemli's work (Figure 5 bottom)

[17] With the exception, of the book *Mathematical theory of communication*, a special case study expanded and commented in Appendix B



[PDF] Protein measurement with the Folin phenol reagent
OH Lowry, NJ Rosebrough, AL Farr, RJ Randall - J biol Chem, 1951 - devbio.wustl.edu
Method Reagents-Reagent A, 2 per cent N&OX in 0.10 N NaOH. Reagent B, 0.5 per cent CuS04. 5Hz0 in 1 per cent sodium or potassium tartrabe. Reagent C, alkaline copper solution. Mix 50 ml. of Reagent A with 1 ml. of Reagent B. Discard after 1 day. Reagent D, ...
Cited by 192841 Related articles All 29 versions Web of Science: 305202 Cite Save More

21st October 2014

[PDF] Protein measurement with the Folin phenol reagent
OH Lowry, NJ Rosebrough, AL Farr... - J biol ..., 1951 - amirza.persiangig.com
Method Reagents-Reagent A, 2 per cent N&OX in 0.10 N NaOH. Reagent B, 0.5 per cent CuS04. 5Hz0 in 1 per cent sodium or potassium tartrabe. Reagent C, alkaline copper solution. Mix 50 ml. of Reagent A with 1 ml. of Reagent B. Discard after 1 day. Reagent D, ...
Cited by 253671 Related articles All 25 versions Web of Science: 303832 Cite Save More

28th May 2014

[DOC] Cleavage of structural proteins during the assembly of the head of bacteriophage T4
UK Laemmli - nature, 1970 - pageout.net
Attendance Policy: All students are expected to attend every laboratory. **In** the event a student misses a class due to illness or a death **in** the family, a written statement from the appropriate administrative office will be needed. Students are responsible for the material covered ...
Cited by 223264 Related articles All 15 versions Web of Science: 213045 Cite Save More

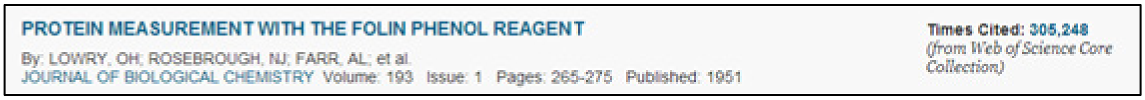
21st October 2014

Figure 5. Citation loss of the most cited document in Google Scholar and Web of Science (Lowry, 1951)

The debate is served...

How is it possible that the total number of citations of a document decreases over time? What are the reasons for these changes? Are the results offered by GS concerning citations stable and reliable, and consequently, the results concerning which the most cited documents are?

There is an explanation for this phenomenon, although it's difficult to justify that a document presents a lower number of citations in the present than the number it presented in the past. The behavior of this document in WoS is more logical, since in these months it has accumulated a few more citations: as of the end of May 2014, it had 303,832 citations, and on October the 21st, 2014, it had 305,202 according to GS (Figure 5 top), and 305,248 according to WoS (Figure 6 bottom). WoS data in GS is updated regularly but not in real time.



PROTEIN MEASUREMENT WITH THE FOLIN PHENOL REAGENT
By: LOWRY, OH; ROSEBROUGH, NJ; FARR, AL; et al.
JOURNAL OF BIOLOGICAL CHEMISTRY Volume: 193 Issue: 1 Pages: 265-275 Published: 1951

Times Cited: 305,248
(from Web of Science Core Collection)

Figure 6. Citation of the most cited document in Google Scholar and Web of Science (Lowry, 1951)

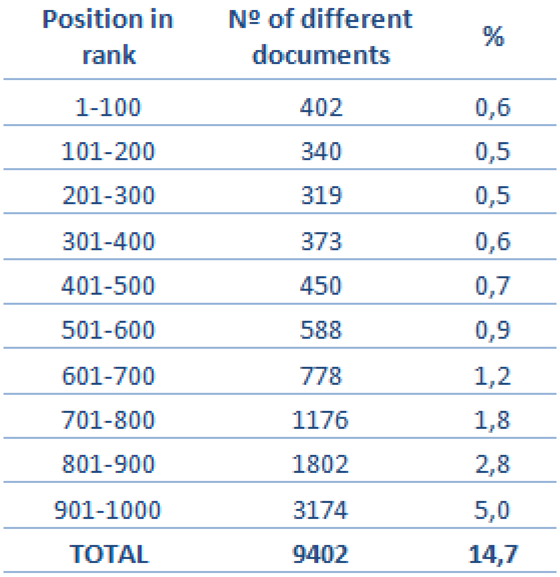
Why does this phenomenon occur in GS?

The answer is related to the dynamic nature of the Web: information is added and removed constantly, and therefore, GS always displays what is currently available on the Web. This is explained in Google Scholar's help pages[18], where they warn that “Google Scholar generally reflects the state of the web as it is currently visible to our search robots and to the majority of users”. Presumably, this drastic change in citations took place when GS made a major “re-crawling” of the documents in its database earlier this year (around the third week of June 2014 according to our data).

The consequences of this phenomenon in our study are self-evident: did we really collect the most cited documents?

To this end, we collected the entire sample again on the 4th of October, 2014, and compared the two samples to learn how many of the documents in our earlier sample are not present in the new sample (Table 2).

Table 2. Comparison of two samples of 64,000 highly cited documents (May and October, 2014)



Position in rank	Nº of different documents	%
1-100	402	0,6
101-200	340	0,5
201-300	319	0,5
301-400	373	0,6
401-500	450	0,7
501-600	588	0,9
601-700	778	1,2
701-800	1176	1,8
801-900	1802	2,8
901-1000	3174	5,0
TOTAL	**9402**	**14,7**

Only 14.7% of the 64,000 documents in the most recent sample were not also present in our earlier sample. Moreover, most of these new documents are placed in pretty low positions in Google Scholar’s ranking of results.

Are we sure that all versions of a same document (not only different editions or reprints, but also translations to other languages) have been successfully merged, and that all their respective citations have been added, removing any possible duplicates?

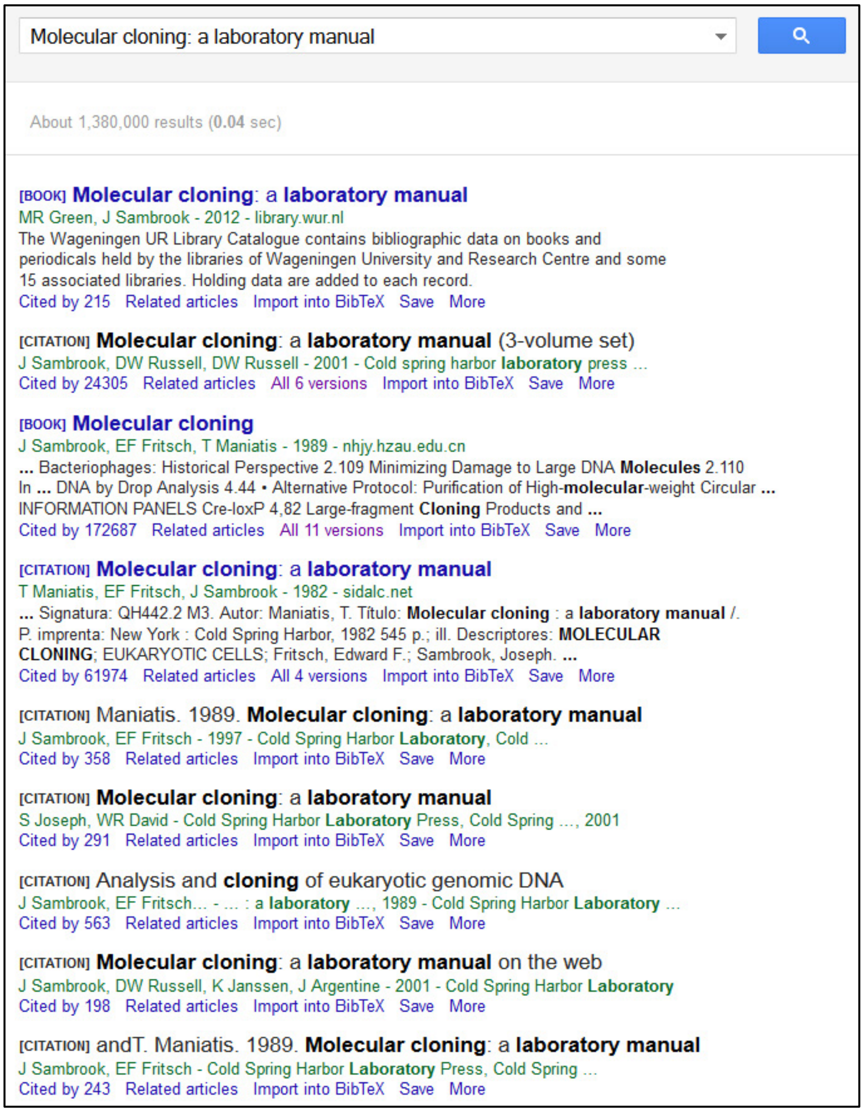
GS has declared that they do exactly this (Verstak & Acharya, 2013), but we don’t have empirical data to comment on the potential errors regarding this issue.

Nevertheless, it is not difficult to find obvious errors, like the case of the classic work in Molecular Biology “Molecular cloning: a laboratory manual” (Figure 7), where it is clear that there are still many different versions with a high number of citations that haven’t been merged. This, of course, is an exceptional case.

[18] **My citation counts have gone down. Help!**

http://scholar.google.com/intl/en/scholar/help.html#corrections [accessed on 24th October 2014]

Normally, documents will not present as many versions as this example (See Question 7; Table 7), nor as many citations.



Molecular cloning: a laboratory manual

About 1,380,000 results (**0.04** sec)

[BOOK] **Molecular cloning**: a **laboratory manual**
MR Green, J Sambrook - 2012 - library.wur.nl
The Wageningen UR Library Catalogue contains bibliographic data on books and periodicals held by the libraries of Wageningen University and Research Centre and some 15 associated libraries. Holding data are added to each record.
Cited by 215 Related articles Import into BibTeX Save More

[CITATION] **Molecular cloning**: a **laboratory manual** (3-volume set)
J Sambrook, DW Russell, DW Russell - 2001 - Cold spring harbor **laboratory** press ...
Cited by 24305 Related articles All 6 versions Import into BibTeX Save More

[BOOK] **Molecular cloning**
J Sambrook, EF Fritsch, T Maniatis - 1989 - nhjy.hzau.edu.cn
... Bacteriophages: Historical Perspective 2.109 Minimizing Damage to Large DNA **Molecules** 2.110 In **...** DNA by Drop Analysis 4.44 • Alternative Protocol: Purification of High-**molecular**-weight Circular **...** INFORMATION PANELS Cre-loxP 4,82 Large-fragment **Cloning** Products and **...**
Cited by 172687 Related articles All 11 versions Import into BibTeX Save More

[CITATION] **Molecular cloning**: a **laboratory manual**
T Maniatis, EF Fritsch, J Sambrook - 1982 - sidalc.net
... Signatura: QH442.2 M3. Autor: Maniatis, T. Título: **Molecular cloning** : a **laboratory manual** /. P. imprenta: New York : Cold Spring Harbor, 1982 545 p.; ill. Descriptores: **MOLECULAR CLONING**; EUKARYOTIC CELLS; Fritsch, Edward F.; Sambrook, Joseph. **...**
Cited by 61974 Related articles All 4 versions Import into BibTeX Save More

[CITATION] Maniatis. 1989. **Molecular cloning**: a **laboratory manual**
J Sambrook, EF Fritsch - 1997 - Cold Spring Harbor **Laboratory**, Cold ...
Cited by 358 Related articles Import into BibTeX Save More

[CITATION] **Molecular cloning**: a **laboratory manual**
S Joseph, WR David - Cold Spring Harbor **Laboratory** Press, Cold Spring ..., 2001
Cited by 291 Related articles Import into BibTeX Save More

[CITATION] Analysis and **cloning** of eukaryotic genomic DNA
J Sambrook, EF Fritsch... - ... : a **laboratory** ..., 1989 - Cold Spring Harbor **Laboratory** ...
Cited by 563 Related articles Import into BibTeX Save More

[CITATION] **Molecular cloning**: a **laboratory manual** on the web
J Sambrook, DW Russell, K Janssen, J Argentine - 2001 - Cold Spring Harbor **Laboratory**
Cited by 198 Related articles Import into BibTeX Save More

[CITATION] andT. Maniatis. 1989. **Molecular cloning**: a **laboratory manual**
J Sambrook, EF Fritsch - Cold Spring Harbor **Laboratory** Press, Cold Spring ...
Cited by 243 Related articles Import into BibTeX Save More

Figure 7. A few versions of Molecular cloning: a laboratory manual, by J. Sambrook et al. that Google hasn't merged

Lastly, a few well-known issues in bibliometrics (Garfield, 2005) should be kept in mind before proceeding to observe the ranking of the top 1% most cited documents in Google Scholar (see Appendix A). First, the citation windows: a document published in 1950 has had 64 years to receive citations, whereas a document published in 2013 has had only one year. Secondly, the different paces at which obsolescence takes place in the different scientific fields: generally, documents stop being cited at some point after their publication date. Thirdly, the exponential growth of production: as production volumes increase, the number of citations also increases.

Question 2. Which are the most cited document types in Google Scholar?

Document types and its evolution

The typologies of the documents in our sample are shown in Figure 8. As we stated in the methods section, we have been able to determine the typology of 45,410 documents in our sample (71%). The typologies of the remaining 29% are unknown.

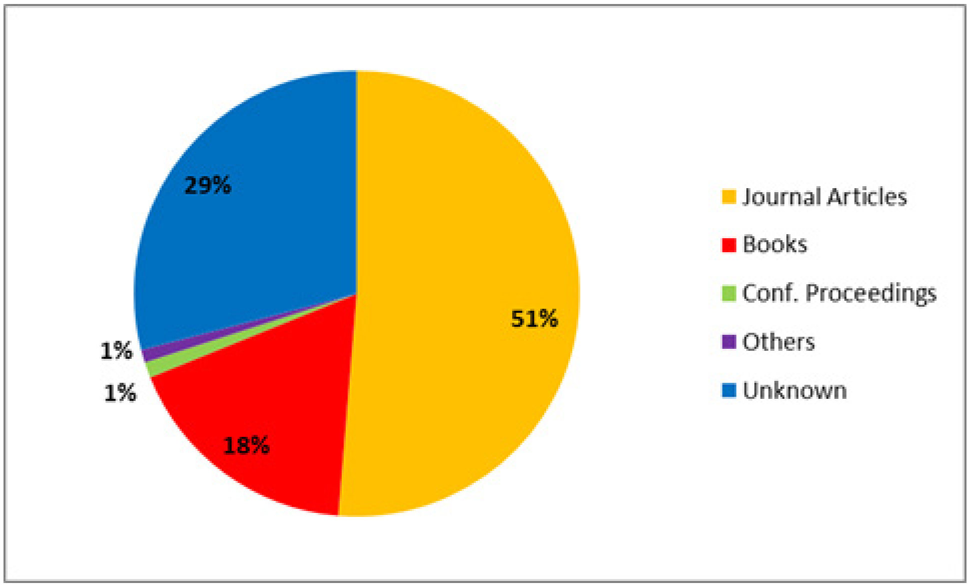


Figure 8. Document types of the highly cited documents in Google Scholar

There is a clear predominance of journal articles, which make up a much higher fraction of the total than books and book chapters. The presence of conference proceedings is almost non-existent. Admittedly, this distribution might have been different if we could have defined the document type of the remaining 29% of our sample.

Figure 9 presents this distribution from a longitudinal perspective, where we find the following three phenomena:

- A steady decrease over time in the number of documents with an unknown document type.
- A constant increase in the number of books, which become the most frequent document type in the last five years (2009-2013). As an example, in the 1,000 results for the year 2013, we only find 27 journal articles. What's the reason for this obvious overrepresentation of the book format over the rest of the formats in the last years? We believe this phenomenon has very much to do with the decision of using the most recent edition of a book (and therefore, the most recent publication date), as the primary version of the document (See Question 1, Figures 3-4). This causes, for example, that a classic book originally published in 1965, and reprinted over the years with its latest edition published in 2012, will be considered as having been published in 2012. Since Google Scholar only presents 1,000 results for any given query, and we only collected information about the primary versions of the documents, these books are overshadowing other publications that have really been published in these years.
- Conference proceedings play an insignificant role in this sample, although they achieve greater presence during the last decade of the twentieth century.

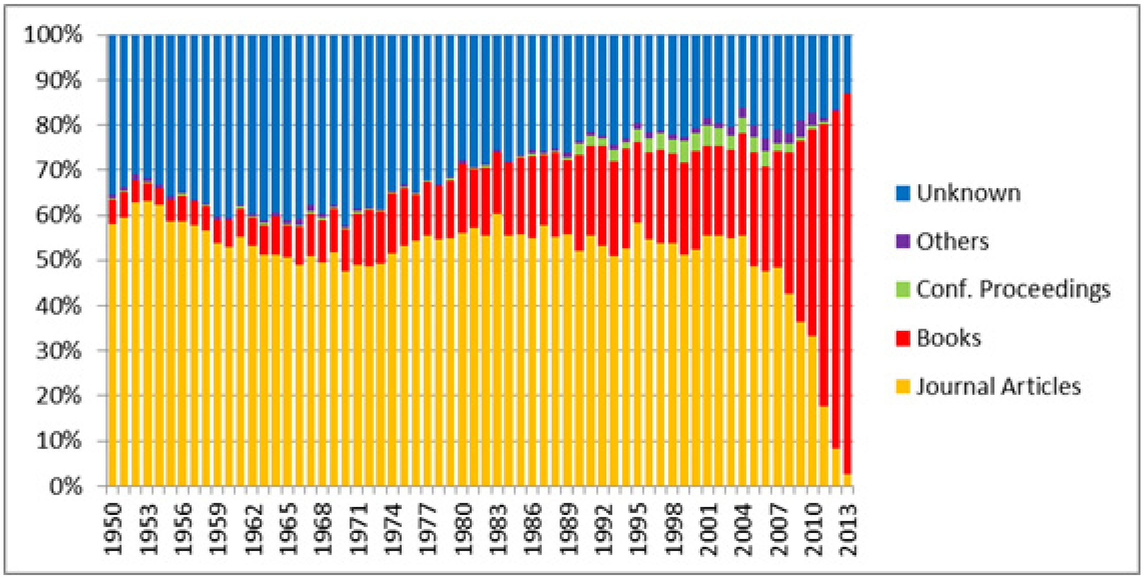
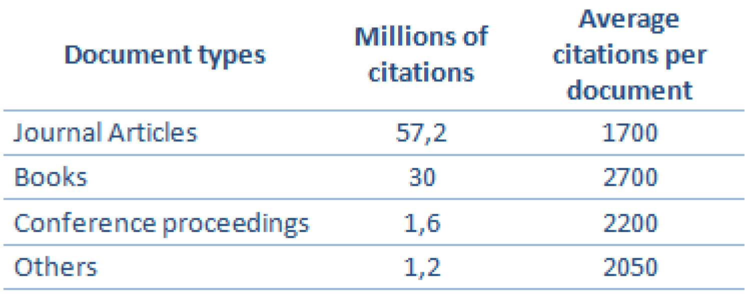


Figure 9. Document types of the highly cited documents in Google Scholar, broken down by years

Citations and document types

Books is the document type with a higher average citations per document (Table 3), followed by conference proceedings. Journal articles rank third in this list.

Table 3. Citations according to document types

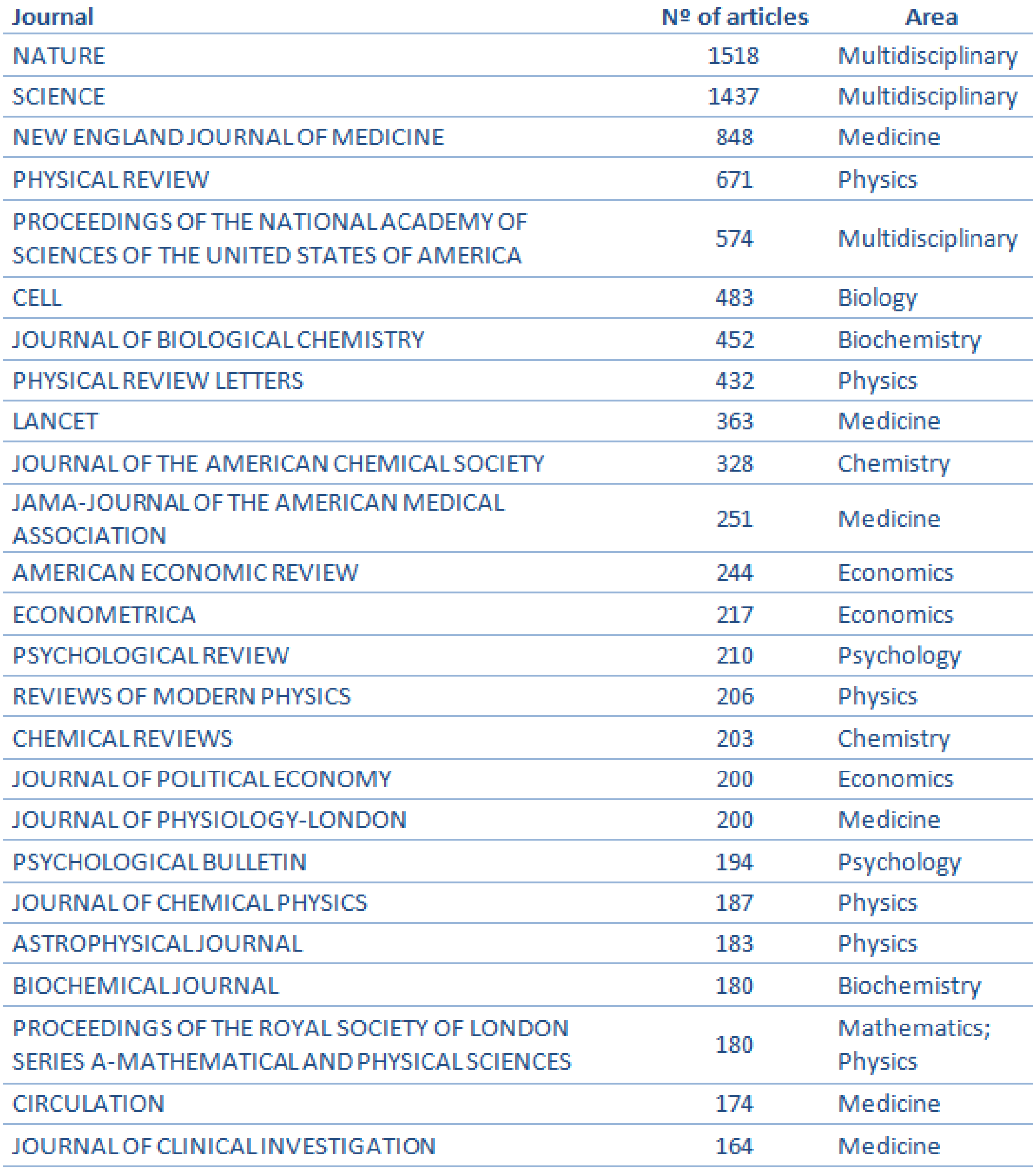


Document types	Millions of citations	Average citations per document
Journal Articles	57,2	1700
Books	30	2700
Conference proceedings	1,6	2200
Others	1,2	2050

Journals containing highly cited documents (1950 and 2013)

The articles contained in our sample have been published in a total of 3,131 different journals. In Table 4 we show the list of journals where the majority of articles are concentrated. As it could not be otherwise, multidisciplinary journals (Science and Nature) are the ones with the higher number of highly cited journals, followed by the major journals in the natural sciences (Physics and Chemistry). As regards the social sciences, only economics and psychology journals (American Economic Review, and Econometrica) are capable of reaching prominent positions.

Table 4. Top 25 Most frequent journals in the highly cited documents in Google Scholar



Journal	Nº of articles	Area
NATURE	1518	Multidisciplinary
SCIENCE	1437	Multidisciplinary
NEW ENGLAND JOURNAL OF MEDICINE	848	Medicine
PHYSICAL REVIEW	671	Physics
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	574	Multidisciplinary
CELL	483	Biology
JOURNAL OF BIOLOGICAL CHEMISTRY	452	Biochemistry
PHYSICAL REVIEW LETTERS	432	Physics
LANCET	363	Medicine
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY	328	Chemistry
JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION	251	Medicine
AMERICAN ECONOMIC REVIEW	244	Economics
ECONOMETRICA	217	Economics
PSYCHOLOGICAL REVIEW	210	Psychology
REVIEWS OF MODERN PHYSICS	206	Physics
CHEMICAL REVIEWS	203	Chemistry
JOURNAL OF POLITICAL ECONOMY	200	Economics
JOURNAL OF PHYSIOLOGY-LONDON	200	Medicine
PSYCHOLOGICAL BULLETIN	194	Psychology
JOURNAL OF CHEMICAL PHYSICS	187	Physics
ASTROPHYSICAL JOURNAL	183	Physics
BIOCHEMICAL JOURNAL	180	Biochemistry
PROCEEDINGS OF THE ROYAL SOCIETY OF LONDON SERIES A-MATHEMATICAL AND PHYSICAL SCIENCES	180	Mathematics; Physics
CIRCULATION	174	Medicine
JOURNAL OF CLINICAL INVESTIGATION	164	Medicine

Discussions & Limitations

Google Scholar does not provide document type information systematically for all its documents (only for books).

Because of this, we could not determine the document types of the entire data set, since this would have required a manual inspection of the remaining 18,590 documents. If we did this, our guess is that the fraction of books and book chapters would increase, since this is the typology that GS has more trouble identifying.

Would the weight of the book format be different over the years, had Google Scholar decided to take the first edition of books as their primary version?

Without a doubt, yes (see Question 1; Figure 4).

Question 3. In what languages are the most cited documents in Google Scholar written?

In Figure 10 we show the document distribution according to language. As we can see, English dominates over the rest of languages as the most widely used language for scientific communication, accounting for 92.5% of all the documents in our sample. The second and third places are occupied by Spanish and Portuguese respectively, but neither of them reach even 2% of the total.

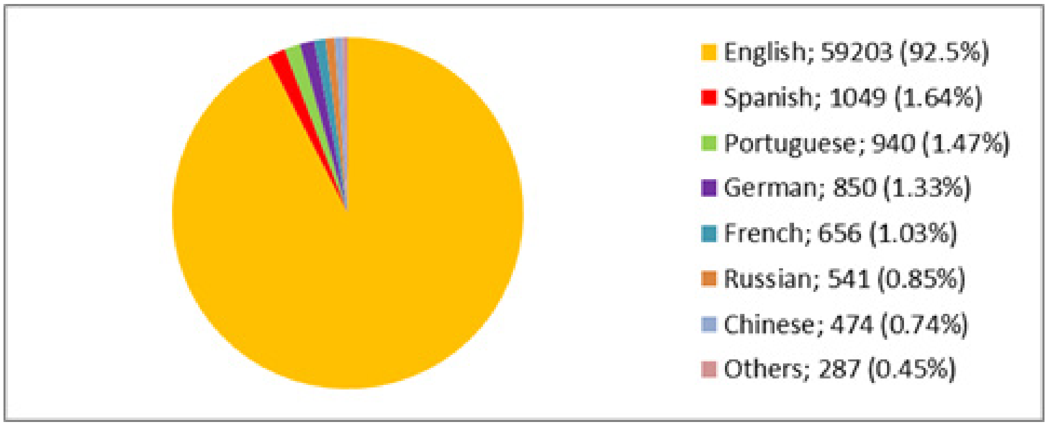


Figure 10. Distribution of languages used in the highly cited documents in GS

In Figure 11 we can observe the same data broken down by years. The results for the language variable are much more stable through the years than the ones found for the document types. In this case, the English language predominates in every year, with an oscillation between its maximum and minimum value of less than 10% (87% in 2013, and 95% in 1991).

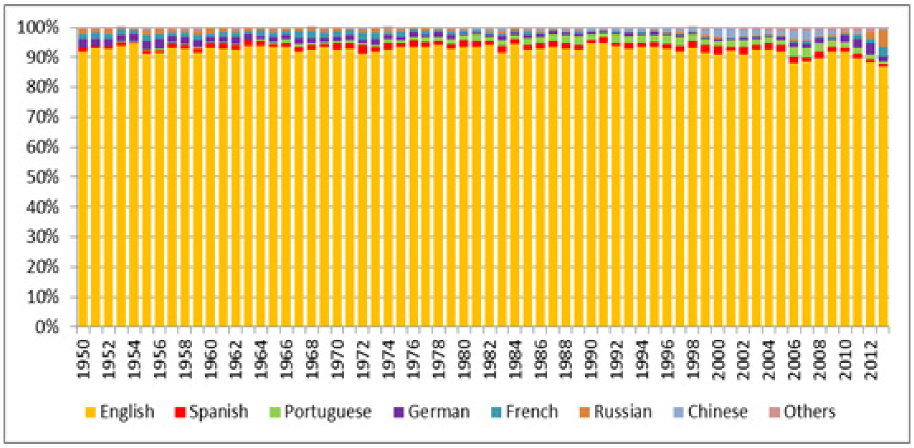


Figure 11. Distribution of languages in the highly cited documents in GS by years of publication

The "Others" category includes the following languages: Italian, Swedish, Indonesian, Finnish, Danish, Bulgarian, Polish, Norwegian, Turkish, Latin, Slovenian, Serbian, Dutch, Macedonian, Malayan, Japanese, Czech, Estonian, Slovak, Mongolian, Catalan, Croatian, Lithuanian, and Ukrainian.

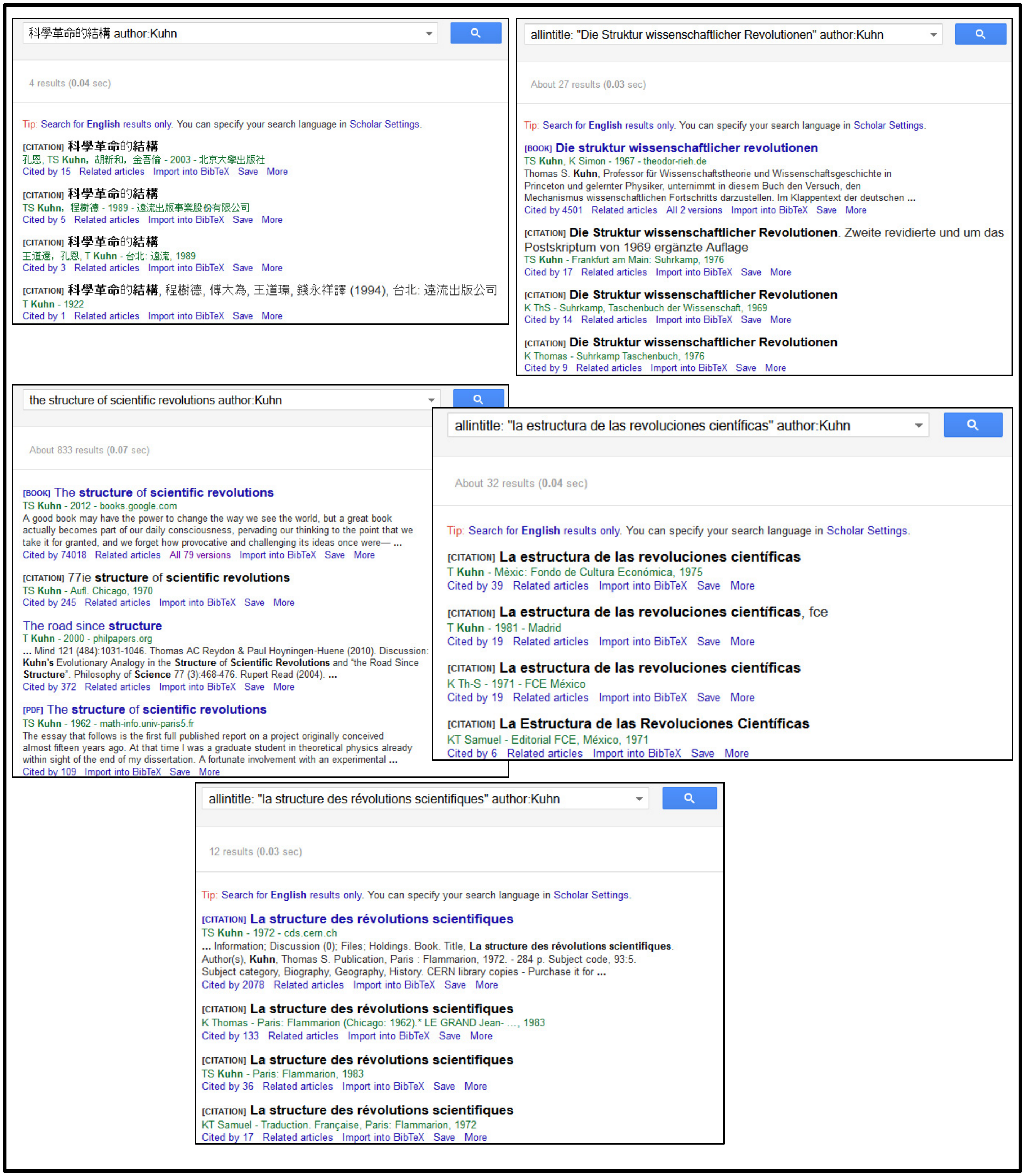
Discussions & Limitations

As with document types, Google Scholar does not provide information about the languages in which the documents it indexes are written.

Because of this, we developed a strategy to determine this information, using WoS data where possible (around 50% of the cases), and the title and abstract of the document in all the other cases. This approach, however, may have introduced an overrepresentation of the English language, since it is usual for a document written in a language other than English to provide its title and abstract in English as well, for the purpose of being indexed in international databases.

Additionally, our sample may contain records that are in fact translations of other documents (which may also be present in our sample).

As we pointed out in previous studies (Martín et al. 2014), Google Scholar usually fails to group together different translations of a same document. This is the case of journals that are published both in English and in other language, or books that are translated into various languages (see Figure 12). This issue has an immediate effect for the works affected by this problem: their citations are scattered across different records, and this could affect their status as highly cited documents.



科學革命的結構 author:Kuhn

4 results (0.04 sec)

Tip: Search for English results only. You can specify your search language in Scholar Settings.

[CITATION] 科學革命的結構
孔恩, TS Kuhn，胡新和，金吾倫 - 2003 - 北京大學出版社
Cited by 15 Related articles Import into BibTeX Save More

[CITATION] 科學革命的結構
TS Kuhn，程樹德 - 1989 - 遠流出版事業股份有限公司
Cited by 5 Related articles Import into BibTeX Save More

[CITATION] 科學革命的結構
王道還，孔恩, T Kuhn - 台北: 遠流, 1989
Cited by 3 Related articles Import into BibTeX Save More

[CITATION] 科學革命的結構, 程樹德, 傅大為, 王道環, 錢永祥譯 (1994), 台北: 遠流出版公司
T Kuhn - 1922
Cited by 1 Related articles Import into BibTeX Save More

allintitle: "Die Struktur wissenschaftlicher Revolutionen" author:Kuhn

About 27 results (0.03 sec)

Tip: Search for English results only. You can specify your search language in Scholar Settings.

[BOOK] Die struktur wissenschaftlicher revolutionen
TS Kuhn, K Simon - 1967 - theodor-rieh.de
Thomas S. Kuhn, Professor für Wissenschaftstheorie und Wissenschaftsgeschichte in Princeton und gelernter Physiker, unternimmt in diesem Buch den Versuch, den Mechanismus wissenschaftlichen Fortschritts darzustellen. Im Klappentext der deutschen ...
Cited by 4501 Related articles All 2 versions Import into BibTeX Save More

[CITATION] Die Struktur wissenschaftlicher Revolutionen. Zweite revidierte und um das Postskriptum von 1969 ergänzte Auflage
TS Kuhn - Frankfurt am Main: Suhrkamp, 1976
Cited by 17 Related articles Import into BibTeX Save More

[CITATION] Die Struktur wissenschaftlicher Revolutionen
K ThS - Suhrkamp, Taschenbuch der Wissenschaft, 1969
Cited by 14 Related articles Import into BibTeX Save More

[CITATION] Die Struktur wissenschaftlicher Revolutionen
K Thomas - Suhrkamp Taschenbuch, 1976
Cited by 9 Related articles Import into BibTeX Save More

the structure of scientific revolutions author:Kuhn

About 833 results (0.07 sec)

[BOOK] The structure of scientific revolutions
TS Kuhn - 2012 - books.google.com
A good book may have the power to change the way we see the world, but a great book actually becomes part of our daily consciousness, pervading our thinking to the point that we take it for granted, and we forget how provocative and challenging its ideas once were— ...
Cited by 74018 Related articles All 79 versions Import into BibTeX Save More

[CITATION] 77ie structure of scientific revolutions
TS Kuhn - Aufl. Chicago, 1970
Cited by 245 Related articles Import into BibTeX Save More

The road since structure
T Kuhn - 2000 - philpapers.org
... Mind 121 (484):1031-1046. Thomas AC Reydon & Paul Hoyningen-Huene (2010). Discussion: Kuhn's Evolutionary Analogy in the Structure of Scientific Revolutions and "the Road Since Structure". Philosophy of Science 77 (3):468-476. Rupert Read (2004). ...
Cited by 372 Related articles Import into BibTeX Save More

[PDF] The structure of scientific revolutions
TS Kuhn - 1962 - math-info.univ-paris5.fr
The essay that follows is the first full published report on a project originally conceived almost fifteen years ago. At that time I was a graduate student in theoretical physics already within sight of the end of my dissertation. A fortunate involvement with an experimental ...
Cited by 109 Import into BibTeX Save More

allintitle: "la estructura de las revoluciones científicas" author:Kuhn

About 32 results (0.04 sec)

Tip: Search for English results only. You can specify your search language in Scholar Settings.

[CITATION] La estructura de las revoluciones científicas
T Kuhn - Mèxic: Fondo de Cultura Económica, 1975
Cited by 39 Related articles Import into BibTeX Save More

[CITATION] La estructura de las revoluciones científicas, fce
T Kuhn - 1981 - Madrid
Cited by 19 Related articles Import into BibTeX Save More

[CITATION] La estructura de las revoluciones científicas
K Th-S - 1971 - FCE México
Cited by 19 Related articles Import into BibTeX Save More

[CITATION] La Estructura de las Revoluciones Científicas
KT Samuel - Editorial FCE, México, 1971
Cited by 6 Related articles Import into BibTeX Save More

allintitle: "la structure des révolutions scientifiques" author:Kuhn

12 results (0.03 sec)

Tip: Search for English results only. You can specify your search language in Scholar Settings.

[CITATION] La structure des révolutions scientifiques
TS Kuhn - 1972 - cds.cern.ch
... Information; Discussion (0); Files; Holdings. Book. Title, La structure des révolutions scientifiques. Author(s), Kuhn, Thomas S. Publication, Paris : Flammarion, 1972. - 284 p. Subject code, 93:5. Subject category, Biography, Geography, History. CERN library copies - Purchase it for ...
Cited by 2078 Related articles Import into BibTeX Save More

[CITATION] La structure des révolutions scientifiques
K Thomas - Paris: Flammarion (Chicago: 1962).* LE GRAND Jean- ..., 1983
Cited by 133 Related articles Import into BibTeX Save More

[CITATION] La structure des révolutions scientifiques
TS Kuhn - Paris: Flammarion, 1983
Cited by 36 Related articles Import into BibTeX Save More

[CITATION] La structure des révolutions scientifiques
KT Samuel - Traduction. Française, Paris: Flammarion, 1972
Cited by 17 Related articles Import into BibTeX Save More

Figure 12. Example of language versions (Chinese, German, English, Spanish, French) of The structure of scientific revolutions, by Kuhn

Question 4. How many highly cited documents are freely accessible?

The percentage of documents for which Google Scholar provides a freely accessible full text link can be observed in Figure 13. Over 40% of the documents in our sample provided a full text link, and these links are mostly concentrated in the last two decades. The lower rate of records with an open access link in the last four years might be explained by journal's and publisher's embargo policies. Additionally, the high percentage of books in the last 5 years of the sample may influence as well.

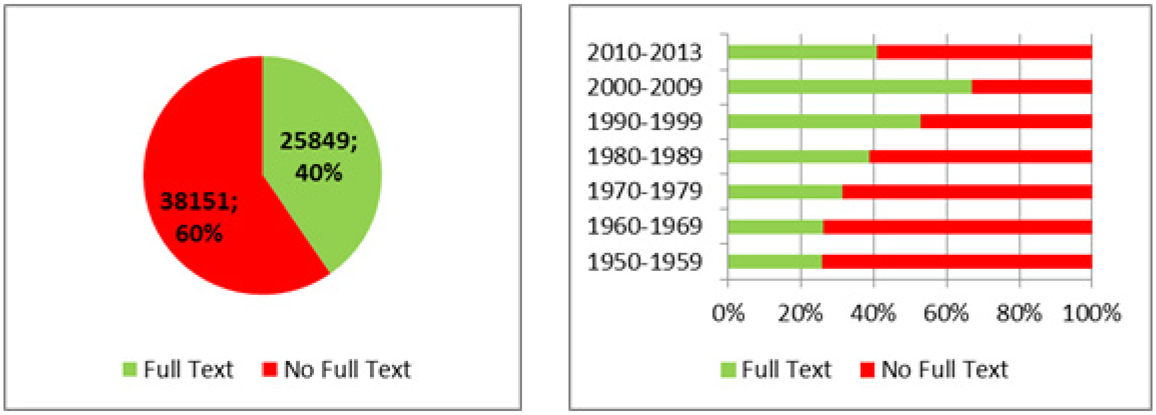


Figure 13. Percentage of freely accessible highly cited documents in Google Scholar. Global results for the 1950-2013 period (left), and broken down by decades (right)

These results are consistent with those published by Archambault et al. in 2013, (since they also found that over 40% of the articles from their sample were freely accessible from Google Scholar), and much higher than the results obtained by Khabsa and Giles (2014), and Björk et al. (2010), who found only a 24% and 20.4% of open access documents respectively.

What file types are the most commonly used to store these highly cited documents?

Full text links point to documents in a variety of formats. The most common one is the PDF format, followed by the HTML format. Figure 14 presents the distribution of these formats for all the documents that provide a Full Text Link. These results confirm the data previously identified, among others, by Aguillo, Ortega, Fernández & Utrilla (2010) and Orduña-Malea, Serrano-Cobos & Lloret-Romero, N. (2009).

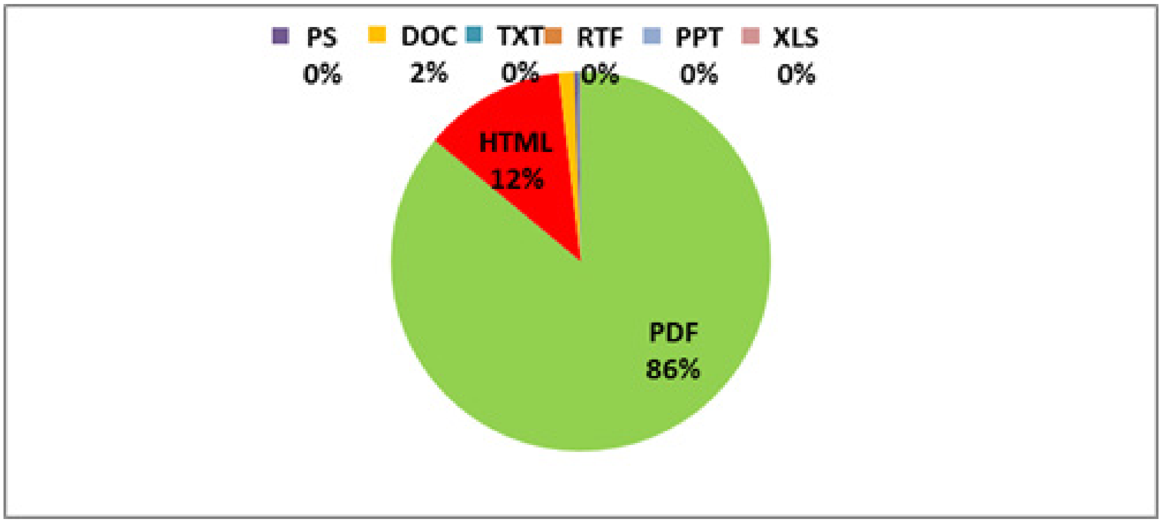


Figure 14. File Formats of the highly cited documents in Google Scholar freely accessible (1950-2013)

Figure 15 shows the same data broken down by years. We can see that the predominance of the PDF format is present throughout the entire range of years. However, it is also noteworthy that the HTML format has started gaining more presence for documents published in the last 25 years, with a peak of almost 20% of the share in 2010.

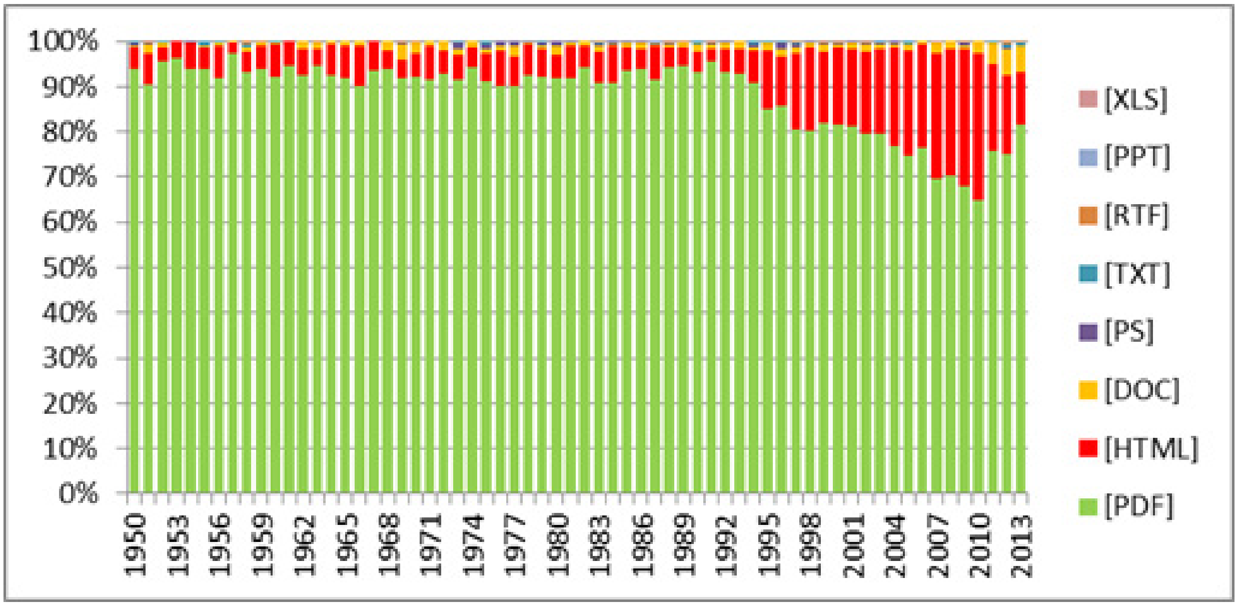
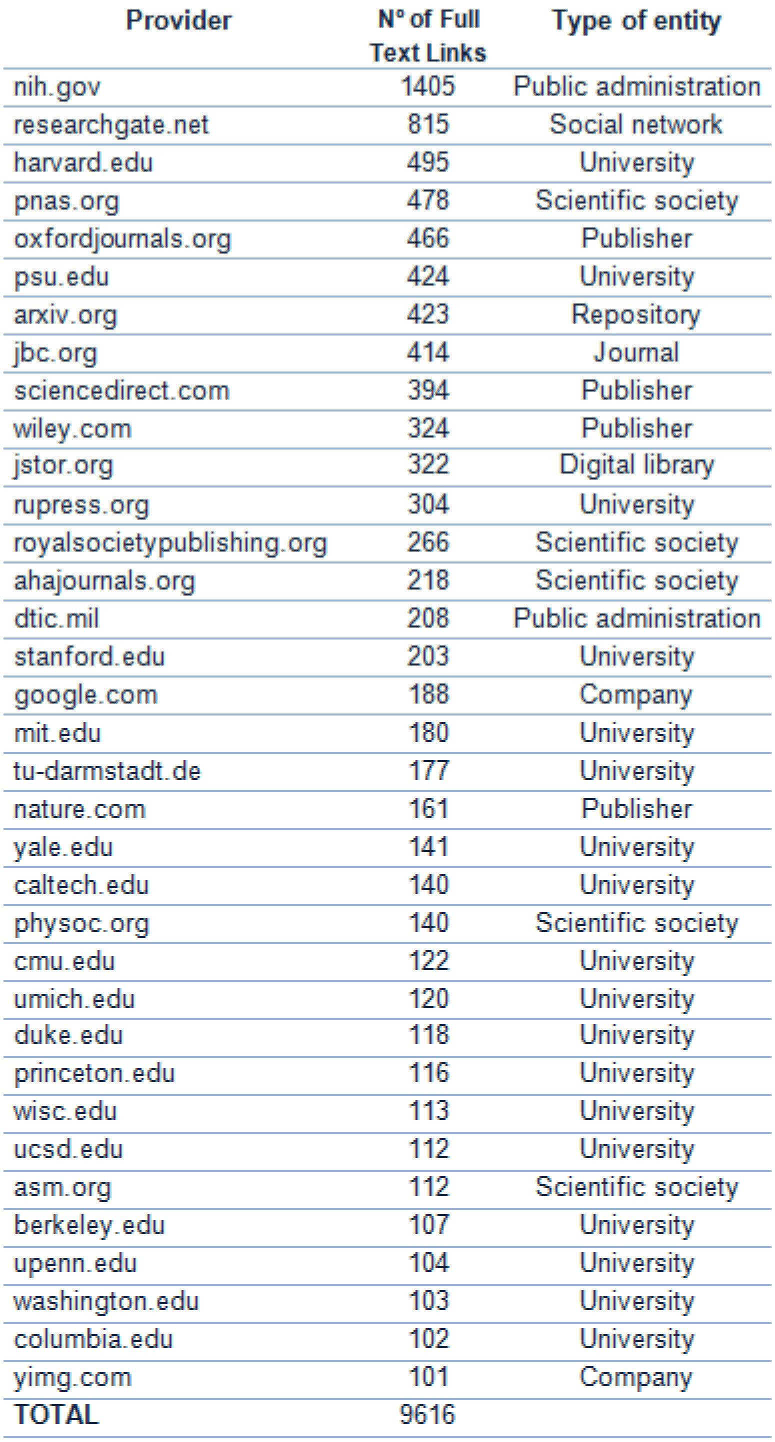


Figure 15. File Formats of the highly cited documents in Google Scholar that are freely accessible, broken down by years (1950-2013)

Which are the main providers of these documents?

We have found a total of 5,715 different providers of Full Text Links in our sample. However, a group of 35 providers account for more than a third of all the links. Table 5 shows these main providers.

Table 5. Full Text providers

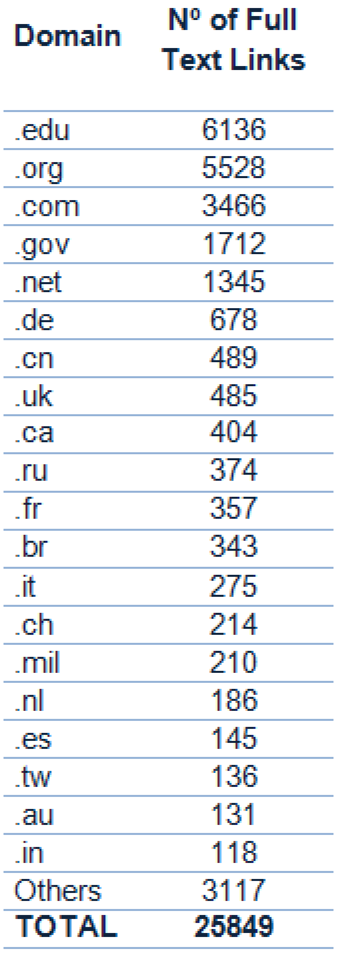


Provider	Nº of Full Text Links	Type of entity
nih.gov	1405	Public administration
researchgate.net	815	Social network
harvard.edu	495	University
pnas.org	478	Scientific society
oxfordjournals.org	466	Publisher
psu.edu	424	University
arxiv.org	423	Repository
jbc.org	414	Journal
sciencedirect.com	394	Publisher
wiley.com	324	Publisher
jstor.org	322	Digital library
rupress.org	304	University
royalsocietypublishing.org	266	Scientific society
ahajournals.org	218	Scientific society
dtic.mil	208	Public administration
stanford.edu	203	University
google.com	188	Company
mit.edu	180	University
tu-darmstadt.de	177	University
nature.com	161	Publisher
yale.edu	141	University
caltech.edu	140	University
physoc.org	140	Scientific society
cmu.edu	122	University
umich.edu	120	University
duke.edu	118	University
princeton.edu	116	University
wisc.edu	113	University
ucsd.edu	112	University
asm.org	112	Scientific society
berkeley.edu	107	University
upenn.edu	104	University
washington.edu	103	University
columbia.edu	102	University
yimg.com	101	Company
TOTAL	9616	

If we analyse the top-level domains of these links, the most frequent are academic institutions (.edu) and organizations (.org). Moreover, the number of links provided by academic institutions is probably higher than 6,136, because there are many universities that use national top-level domains instead of .edu. Table 6 shows the 20 most frequent top-level domains.

This means that GS feeds highly cited documents mainly, at least as far as our sample is concerned, from universities (institutional repositories) and public organizations (working papers, grey literature), and not from commercial publishers. Of special note is the role of the scientific social network ResearchGate, where researchers often upload their publications.

Table 6. Main top-level domains contributing Full Text links in Google Scholar



Domain	Nº of Full Text Links
.edu	6136
.org	5528
.com	3466
.gov	1712
.net	1345
.de	678
.cn	489
.uk	485
.ca	404
.ru	374
.fr	357
.br	343
.it	275
.ch	214
.mil	210
.nl	186
.es	145
.tw	136
.au	131
.in	118
Others	3117
TOTAL	**25849**

Discussions & Limitations

Do these links really point to full text versions of the documents?

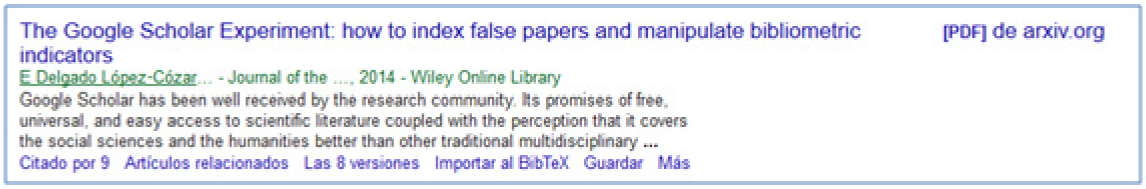
More rigorous analyses should be carried out in order to determine if there are false positives among these links. For example, a freely accessible PDF document containing a review of a book, or just the cover and the table of contents of a book could be mistaken for the book itself.

Moreover, the dynamic nature of the web means that a link that was accessible some time ago may no longer be available. How often does Google Scholar checks that these links are still functioning properly?

Our analysis deals only with the full text link provided for the version of the document GS considers as the primary version.

However, when the primary version of a document is not freely accessible, GS points the user to any other free version if available. Figure 16 is an example of a case where the primary version is the publisher's edition of a journal article, but the Full Text link is a preprint from arXiv).

Figure 16. Primary version, Publisher and Full Text provider



The Google Scholar Experiment: how to index false papers and manipulate bibliometric indicators [PDF] de arxiv.org

E Delgado López-Cózar... - Journal of the ..., 2014 - Wiley Online Library

Google Scholar has been well received by the research community. Its promises of free, universal, and easy access to scientific literature coupled with the perception that it covers the social sciences and the humanities better than other traditional multidisciplinary ...

Citado por 9 Artículos relacionados Las 8 versiones Importar al BibTeX Guardar Más

For documents with more than one version, there may be more than one full text version of the document.

These versions may be hosted in other domains. Again, we want to stress that we only study the Full Text Links displayed for the primary versions of the documents.

Question 5. How many of the highly cited documents indexed by GS are also indexed by WoS?

Almost half of the highly cited documents according to Google Scholar are not indexed on the Web of Science (Figure 17).

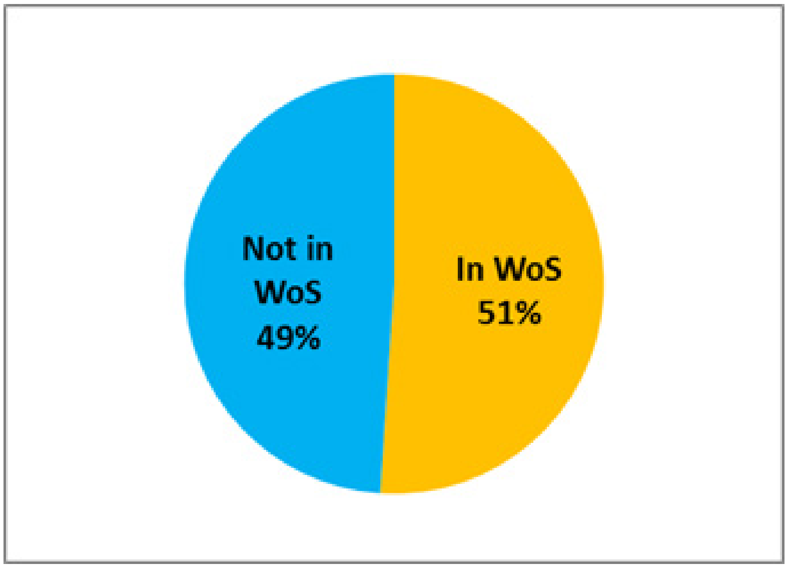


Figure 17. Percentage of highly cited documents in Google Scholar that are also indexed in the Web of Science (1950-2013)

This is extremely relevant, although the following issues should be taken into consideration:

- The different natures of GS and WoS as databases: GS covers academic documents (scientific, technical, educational…) published by all kinds of different sources and in all sorts of communication channels (books, theses, reports…), whereas the coverage in Web of Science Core Collection is oriented towards a more limited range of academic publications, i.e. journal articles and conference communications. This would confirm our hypothesis that GS measures a different kind of impact than the one measured by scientific databases: the academic impact.
- If we want to identify the most influential documents in the academic-scientific sphere, we must use GS.

- GS also identifies the most relevant scientific documents with a fair amount of reliability.

Furthermore, no significant differences are appreciated between 1950 and 2003 (Figure 18). However, the last decade suffers the consequences of the phenomenon we encountered in question 2: the overrepresentation of books in the last years caused by Google Scholar’s policy of taking the latest edition of books as their primary version.

Since Web of Science’s coverage of books is still very limited, it is not surprising that the reduction in the percentage of documents indexed in WoS in the last years closely matches the reduction in the number journal articles during the same years (Figure 9).

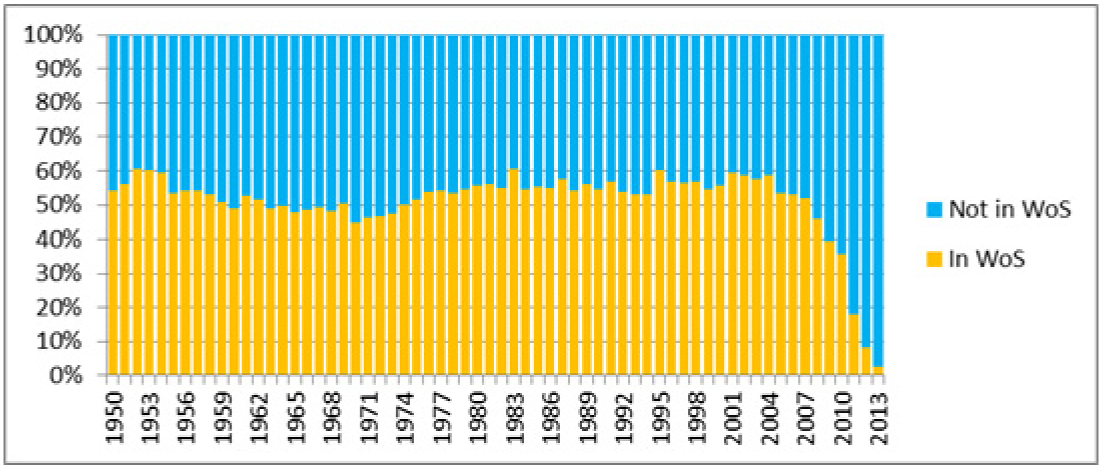


Figure 18. Percentage of highly cited documents in Google Scholar that are also indexed in the Web of Science, broken down by decades (1950-2013)

Discussions & Limitations

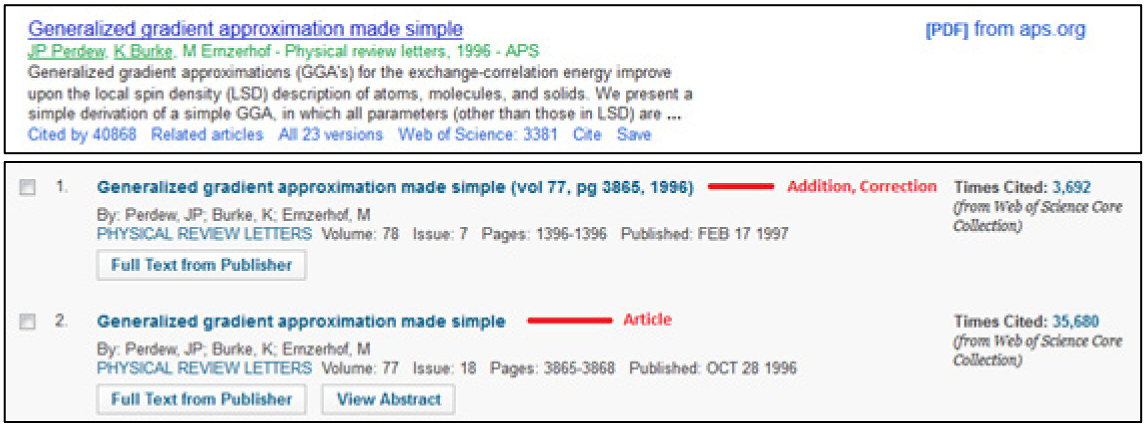
Is the GS-WoS connection correctly implemented?

A more in-depth study should be carried out to determine potential flaws in the matching of documents and the frequency with which they occur:

- False positives: a document in GS matched to a document in WoS even if they’re not really the same documents. For example, a book in GS might be matched to a review of that book indexed in WoS. This is the case of the book “The discovery of grounded theory: Strategies for qualitative research”, which was previously presented in Table 1.
- False negatives: documents indexed both in GS and WoS for which a connection hasn’t been established.

As a first approximation, we have selected the 398 most cited WoS documents between 1950 and 2013 that, according to their WoS ID (accession number), weren’t present in our GS sample. We have searched the titles of these documents on Google Scholar and found that 382 (96%) were in fact indexed in Google Scholar, and 300 of them were also connected to a different WoS record.

Therefore, these mistakes arise from incorrect connections between Google Scholar and Web of Science records, caused by the existence of various records with the same name in WoS. For example, a case where a document in Google Scholar has been connected to the Correction of an article in WoS, and not to the article itself is shown in Figure 19.



Generalized gradient approximation made simple [PDF] from aps.org
JP Perdew, K Burke, M Ernzerhof - Physical review letters, 1996 - APS
Generalized gradient approximations (GGA's) for the exchange-correlation energy improve upon the local spin density (LSD) description of atoms, molecules, and solids. We present a simple derivation of a simple GGA, in which all parameters (other than those in LSD) are ...
Cited by 40868 Related articles All 23 versions Web of Science: 3381 Cite Save

□ 1. Generalized gradient approximation made simple (vol 77, pg 3865, 1996) Addition, Correction Times Cited: 3,692 (from Web of Science Core Collection)
By: Perdew, JP; Burke, K; Ernzerhof, M
PHYSICAL REVIEW LETTERS Volume: 78 Issue: 7 Pages: 1396-1396 Published: FEB 17 1997
Full Text from Publisher

□ 2. Generalized gradient approximation made simple Article Times Cited: 35,680 (from Web of Science Core Collection)
By: Perdew, JP; Burke, K; Ernzerhof, M
PHYSICAL REVIEW LETTERS Volume: 77 Issue: 18 Pages: 3865-3868 Published: OCT 28 1996
Full Text from Publisher View Abstract

Figure 19. Incorrect connections between Google Scholar and Web of Science records

Is it possible that some highly cited articles according to the Web of Science are not indexed on Google Scholar?

As noted earlier in question 1, this may have happened in a very few cases, but not among the very highly cited (30,000 most cited documents in our sample).

The overrepresentation of books in the last decade

Again, this is one of the flaws in our sample, since it has caused that many journal articles published in those last years of the sample (2003-2013) and that have received many citations, are being left out in favor of books that were first published many years ago.

Question 6. Is there a correlation between the number of citations that these highly cited documents have received in GS and the number of citations they have received in WoS?

We have calculated Pearson's correlation coefficient for the number of citations that documents have received according to Google Scholar and the Web of Science, by year. The average correlation is 0.8 (calculated only for documents that are in both sources, which are 32,680). Figure 20 shows the Pearson correlation coefficient for each of the years in our sample.

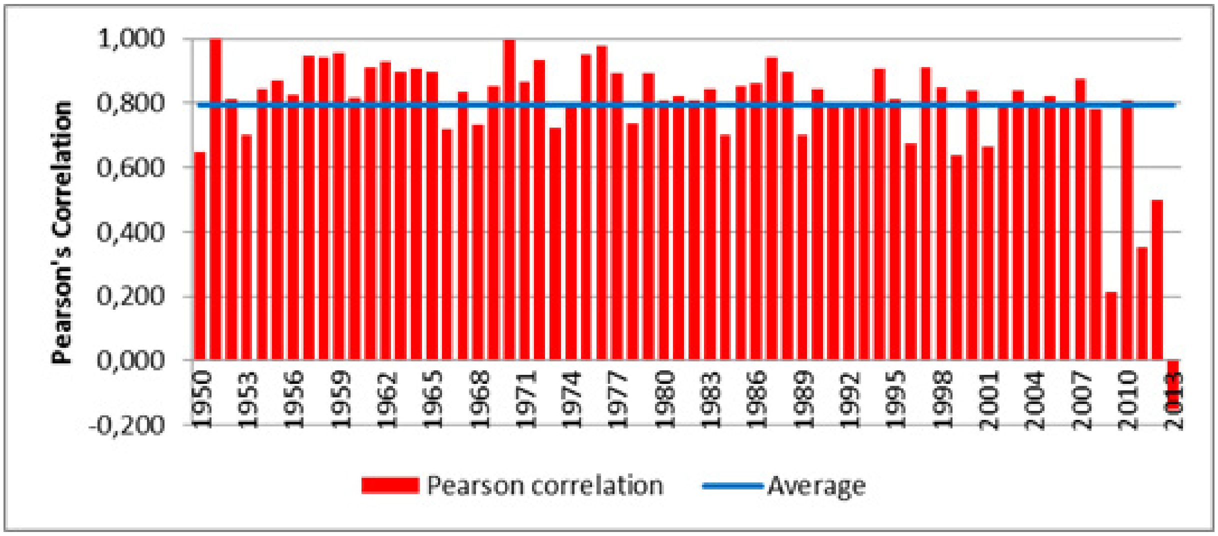


Figure 20. Pearson correlation coefficient between Google Scholar and Web of Science citations (1950-2013)

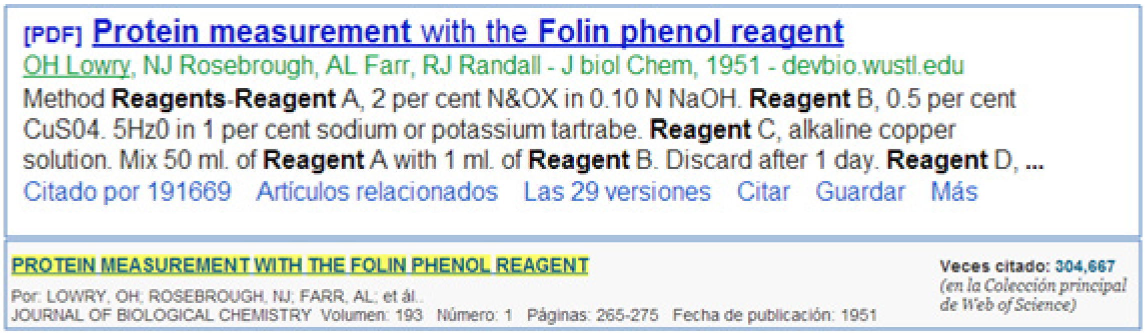
This finding is consistent with the results found in many previous studies (Sanderson 2008; Kousha, & Thelwall 2008; Meho & Rogers 2008; Franceschet, 2010; Delgado López-Cózar & Cabezas 2013; Delgado López-Cózar & Repiso 2013), who also found a high correlation among the journal indicators published by Google Scholar/Google Scholar Metrics and the Web of Science/Journal Citation Reports. However, none of these studies had analysed a sample as large as this one (32,680 documents).

It is common among the studies that compare Google Scholar and the Web of Science to quantify the number of citations they have been able to find for the documents they index. In our sample, 91.6% of the documents have received more citations in GS than in WoS. Only 3,079 documents (9.4%) have more citations according to WoS than in GS. Furthermore, the average number of citations per document in GS is 1,790, and 1,080 in WoS, which means that on average, GS has 70% more citations per document than WoS.

Discussions & Limitations

1. As in question 5, the quality of the matching between GS and WoS plays an important part.
2. The instability of Google Scholar's indicators is also an important factor and should be further analysed.

As an example, Lowry's classic article had 253,671 citations at the end of May, 2014, when we collected the data (see Table 1), but on August the 5th the count had went down to 191,669 (Figure 21). WoS data seems to be much more stable, but it also went down from 304,893 citations in May, to 304,667 in August (See also Question 1, Figure 5).



[PDF] Protein measurement with the Folin phenol reagent
OH Lowry, NJ Rosebrough, AL Farr, RJ Randall - J biol Chem, 1951 - devbio.wustl.edu
Method **Reagents**-**Reagent** A, 2 per cent N&OX in 0.10 N NaOH. **Reagent** B, 0.5 per cent CuS04. 5Hz0 in 1 per cent sodium or potassium tartrabe. **Reagent** C, alkaline copper solution. Mix 50 ml. of **Reagent** A with 1 ml. of **Reagent** B. Discard after 1 day. **Reagent** D, ...
Citado por 191669 Artículos relacionados Las 29 versiones Citar Guardar Más

PROTEIN MEASUREMENT WITH THE FOLIN PHENOL REAGENT
Por: LOWRY, OH; ROSEBROUGH, NJ; FARR, AL; et ál..
JOURNAL OF BIOLOGICAL CHEMISTRY Volumen: 193 Número: 1 Páginas: 265-275 Fecha de publicación: 1951
Veces citado: 304,667
(en la Colección principal de Web of Science)

Figure 21. The most cited scientific article in history, according to Google Scholar (top), and WoS (bottom). Screen capture from 7th of August, 2014

Question 7. How many versions of these highly cited documents has GS detected?

One of the most interesting features of Google Scholar as an academic search engine is its ability to identify and connect all the different instances of the same document that have been deposited across the Web. We should bear in mind that a document can be stored in various locations: the journal publisher's webpage (Cell), databases (Pubmed), aggregators (Ingenta), library catalogues (Dialnet), subject or institutional repositories, and authors' personal or institutional web pages. Moreover, documents might go through various versions and revisions, and they can be cited in different forms. Google acknowledges this reality and tries to find a solution.

Excerpt from Verstak, AA and Acharya, A (2013). Identifying multiple versions of documents. U.S. Patent No. 8,589,784. Washington, DC: U.S. Patent and Trademark Office:

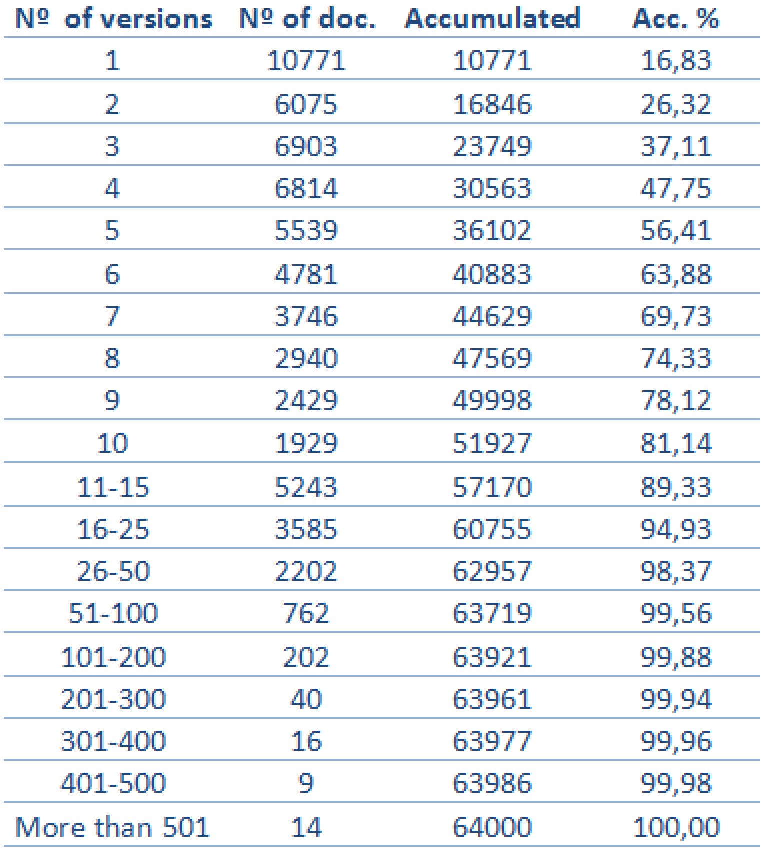
> "[…] it is typical that a particular document or portion thereof, appears in a number of different versions or forms in various online repositories. This generally results in multiple versions of a document being included in the search results for any given query. Because the inclusion of different versions of the same document does not provide additional useful information, this increase in the number of the search results does not benefit users. Also, search results including different versions of the same document may crowd out diverse contents that should be included. These problems have seriously affected the quality of a search result provided by a search engine.

Another problem arises in systems in which there are multiple versions of documents present. Documents in a document collection will have a number of citations to it by other documents. This is particularly the case for academic documents, legal documents, and the like. The number of citations (citation count) to a document is often reflective of the importance, significance, or quality of the document. Where there are different versions of a document present in a repository, each with its own citation count, a user does not have an accurate assessment of the actual significance, importance or quality of the document based on the individual citation counts.

For these reasons, it would be desirable to identify documents that are different versions of the same document in a document collection. It would also be desirable to manage these documents in an efficient manner such that the search engine can furnish the most appropriate and reliable search result."

83% of the documents in our sample have more than one version, whereas 40% have 6 or more versions, 19% have 10 or more versions, and 200 documents have more than 100 versions (0.1%). The distribution of documents according their number of versions can be observed in Table 7:

Table 7. Distribution of documents according to their number of versions



Nº of versions	Nº of doc.	Accumulated	Acc. %
1	10771	10771	16,83
2	6075	16846	26,32
3	6903	23749	37,11
4	6814	30563	47,75
5	5539	36102	56,41
6	4781	40883	63,88
7	3746	44629	69,73
8	2940	47569	74,33
9	2429	49998	78,12
10	1929	51927	81,14
11-15	5243	57170	89,33
16-25	3585	60755	94,93
26-50	2202	62957	98,37
51-100	762	63719	99,56
101-200	202	63921	99,88
201-300	40	63961	99,94
301-400	16	63977	99,96
401-500	9	63986	99,98
More than 501	14	64000	100,00

Discussions & Limitations:

Does GS correctly identify all versions of a same document? Does it make mistakes, like linking a document with a different document (i.e., a review of that document, or a citation found in the list of references of another document), or failing to connect two records that refer to the same document? How frequently does it make these mistakes?

In order to successfully answer these questions, we would need to analyse a sample of documents and study all their versions individually. While we carry out this study, we present, by way of an example, an illustrative example in Appendix B.

Question 8. Is there a correlation between the number of versions GS has detected for these documents, and the number citations they have received?

Using Pearson's correlation coefficient, we have been able to determine that there is no correlation whatsoever between the number of citations of a document in Google Scholar and its number of versions (r = 0.2**). Calculating it by year of publication yields similar results (Figure 22).

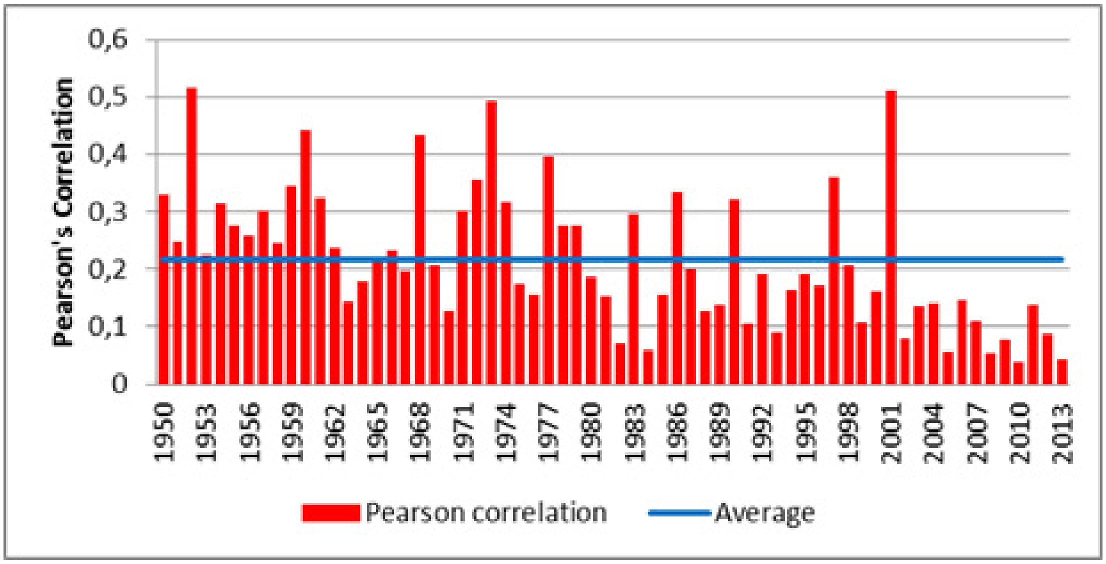


Figure 22. Pearson's correlation between the nº of citations and nº of versions in Google Scholar documents (64,000 most cited documents in Google Scholar; 1950-2013)

Question 9. Is there a correlation between the number of versions Google Scholar has detected for these documents, and their position in the result pages?

Using Pearson's correlation coefficient, we also have determined that there is no correlation whatsoever between the number of versions of a document in Google Scholar and the position it occupies in the search engine results page (Figure 23). The average correlation for the results we collected from 64 queries is r = -0.2**.

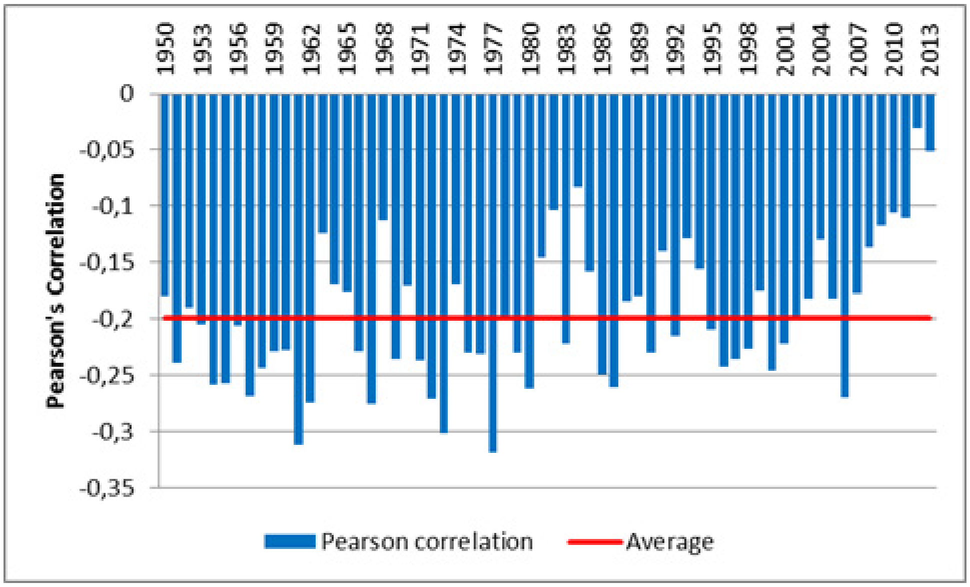


Figure 23. Pearson's correlation between the number of versions of the documents in Google Scholar and their rank in the SERP

Question 10. Is there some relation between the positions these documents occupy in the search engine result pages, and the number of citations they have received?

After calculating the Pearson correlation for each of the years in our queries, we obtained an average r = 0.9** (Figure 24). These results confirm that the most important factor in the calculation of the position a document will occupy in Google Scholar's SERP is its citation count, confirming the statement of Google Scholar in this regard.

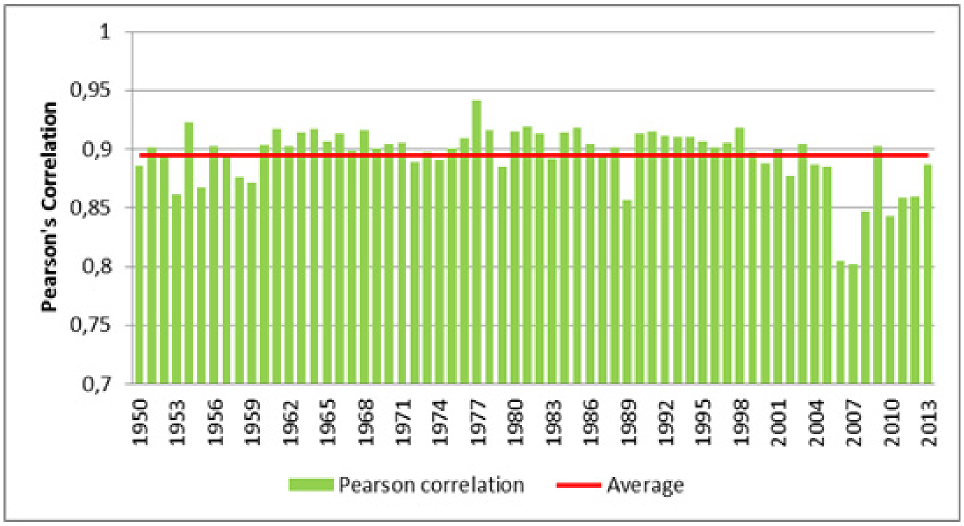


Figure 24. Pearson correlation between the number of citations of documents in Google Scholar and the position they occupy in the Search Engine Result Page

Moreover, according to the scatterplot in Figure 25, the correlation is almost perfect until we reach the last 100 results of the queries, but then the correlation becomes much more tenuous. If we calculate the Pearson correlation for the first 900 and the last 100 results of each query separately, the average correlation for all years is 0.97** and 0.61** respectively. Clearly, the problem is restricted to the tail of the distribution.

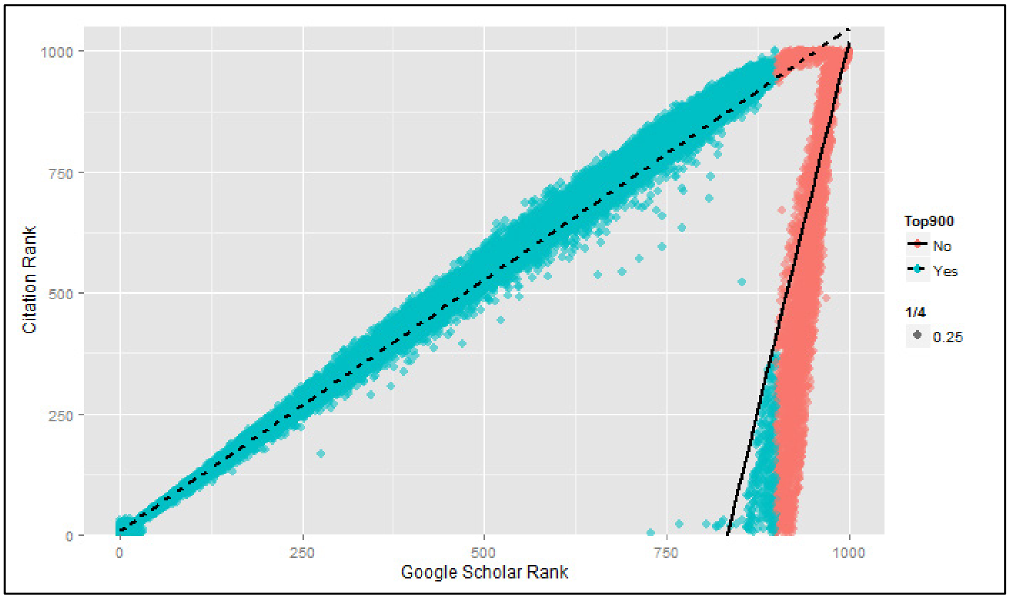


Figure 25. Relationship between the number of citations of documents in Google Scholar and the position they occupy in the Search Engine Result Page

4. Conclusions

As we’ve seen, the analysis of GS provides a very different vision to the question of which are the most influential academic, scientific and technical documents for the scientific, professional and educational community. This fact can be explained by Google Scholar’s own nature:

- Google Scholar’s crawlers sweep the entire academic web: the most well-known scholarly publishers (such as Elsevier, Springer, Sage, Willey, Taylor & Francis, IEEE, ACS, ACM, Macmillan, Wiley, Oxford University Press); their digital hosts/facilitators (such as HighWire Press, MetaPress, Ingenta); societies and other scholarly organizations (such as the American Physical Society, American Chemical Society, ACM), government agencies (National Institute of Health, National Oceanic and Atmospheric Administration, U.S. Geological Survey), databases (Pubmed, ERIC), disciplinary repositories (such as arXiv.org, Astrophysics Data System, RePEc, SSRN, CiteBase), institutional repositories from universities or research centers, library catalogs (Dialnet), as well as personal web pages from researchers, professors, research groups, departments, faculties… hosted inside the servers of the university or research center they belong to.
- While traditional citation-based databases deal with the strictly scientific world (mainly journal articles, conference communications, and some books), Google Scholar’s aim is to index all kinds of scientific documents (scientific and professional journals, conferences, books, working papers, reports…), as well as educational documents (master’s and doctoral theses, teaching materials…), and technical and professional documents (reports, patents, american case laws, annuals…) circulating in the Web.
- It covers documents written in all languages and from all countries.

In conclusion, thanks to the wide and varied sources from which GS feeds, we are able to measure not only scientific impact, but also educational and professional impact in the broadest sense of the term (Kousha and Thelwall, 2008).

At the same time, as regards strict scientific impact, the analysis of GS data provides very similar results to the results obtained from traditional citation-based databases, with the advantage of being able to retrieve a larger and more varied number of citations, since they come from a wider range of document types, different geographical environments, and languages different to English.

The profile of the average highly cited document is: a book or journal article written in English and available online in PDF format.

The rest of the findings of this study can be summarised as follows:

- 40% of the highly cited documents in GS are freely accessible, mostly from educational institutions (mainly universities), and other non-profit organizations. The availability of these documents is essential for GS as a search engine.
- Almost half of these highly cited documents are not indexed in Web of Science, which for many years has has been considered the most prestigious scientific information database.
- There is a high correlation (r = 0.8) between the number of citations of these documents in GS and their citations in WoS.
- GS has detected more than one version for the 83.17% of the documents in our sample.
- There is no correlation between the number of versions GS has detected, and the number citations they have received.
- There is no correlation between the number of versions GS has detected for these documents, and their position in the result pages (SERPs).
- There is a high correlation (r = 0.9) between the positions these documents occupy in the result pages and the number of citations they have received, at least in queries that only use the filtering option to select the documents published in a given year.

Funding acknowledgements

Alberto Martín-Martín enjoys a four-year doctoral fellowship (FPU2013/05863) granted by the Spanish Ministry of Education, Culture and Sports. Juan Manuel Ayllón enjoys a four-year doctoral fellowship (BES-2012-054980) granted by the Spanish Ministry of Economy and Competitiveness. This study has also been funded under project HAR2011-30383-C02-02 from Dirección General de Investigación y Gestión del Plan Nacional de I+D+I (Ministry of Economy and Competitiveness), of which Emilio Delgado López-Cózar is the principal investigator, and project APOSTD/2013/002 from the Regional Ministry of Education, Culture and Sport (Generalitat Valenciana, Spain), awarded to Enrique Orduña-Malea.

References

Aguillo, I. F., Ortega, J. L., Fernández, M., & Utrilla, A. M. (2010). Indicators for a webometric ranking of open access repositories. Scientometrics, 82(3), 477-486.

Aksnes, D. W. (2003). Characteristics of highly cited papers. Research Evaluation,12(3), 159-170.

Aksnes, D. W., & Sivertsen, G. (2004). The effect of highly cited papers on national citation indicators. Scientometrics, 59(2), 213-224.

Bornmann, L. (2010). Towards an ideal method of measuring research performance: Some comments to the Opthof and Leydesdorff (2010) paper. Journal of Informetrics, 4(3), 441–443

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: the avoidance of citation (ratio) averages in field-normalization. Journal of Informetrics, 5(1), 228-230.

Bornmann, L., de Moya-Anegón, F., & Leydesdorff, L. (2011). The new excellence indicator in the World Report of the SCImago Institutions Rankings 2011. arXiv preprint arXiv:1110.2305.

Delgado López-Cózar, E. & Repiso, R., (2013). The Impact of Scientific Journals of Communication: Comparing Google Scholar Metrics, Web of Science and Scopus. *Comunicar*, *21*(41), 45-52.

Delgado-López-Cózar, E., Cabezas-Clavijo, Á. (2013). Ranking journals: could Google Scholar Metrics be an alternative to Journal Citation Reports and Scimago Journal Rank?. Learned Publishing, 26(2), 101-114. DOI: http://dx.doi.org/10.1087/20130206

Franceschet, M. (2010). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar.*Scientometrics* 83.1: 243-258.

Garfield, E. (2005). The Agony and the Ecstasy—The History and Meaning of the Journal Impact Factor. International Congress on Peer Review And Biomedical Publication. Chicago, September 16, 2005. http://www.garfield.library.upenn.edu/papers/jifchicago2005.pdf

Glänzel, W., & Czerwon, H. J. (1992a). What are highly cited publications? A method applied to German scientific papers, 1980–1989. Research Evaluation, 2(3), 135-141.

Glänzel, W., & Schubert, A. (1992b). Some facts and figures on highly cited papers in the sciences, 1981–1985. Scientometrics, 25(3), 373-380.

Glänzel, W., Rinia, E. J., & Brocken, M. G. (1995). A bibliometric study of highly cited European physics papers in the 80s. Research Evaluation, 5(2), 113-122.

Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon.Scientometrics, 58(3), 571-586.

Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics, 74*(2), 273–294.

Kresge, N., Simoni, R. D., & Hill, R. L. (2005). The most highly cited paper in publishing history: Protein determination by Oliver H. Lowry. Journal of Biological Chemistry, 280(28), e25-e25. http://www.jbc.org/content/280/28/e25.full.pdf

Levitt, J. M., & Thelwall, M. (2009). The most highly cited Library and Information Science articles: Interdisciplinarity, first authors and citation patterns.Scientometrics, 78(1), 45-67.

Lowry, OH. (1977). Commentary by Lowry, OH on ”Protein measurement with folin phenol reagent,” Current Contents/Life Sciences (1):7 (January 3, 1977). http://garfield.library.upenn.edu/classics1977/A1977DM02300001.pdf

Maltrás Barba, B. (2003). Los indicadores bibliométricos: fundamentos y aplicación al análisis de la ciencia. Gijón: Trea.

Martín-Martín, A.; Ayllón, J.M.; Orduña-Malea, E.; Delgado López-Cózar, E. (2014). Google Scholar Metrics 2014: a low cost bibliometric tool. EC3 Working Papers, 17: 8 July 2014. http://arxiv.org/ftp/arxiv/papers/1407/1407.2827.pdf

Meho, L. I., & Rogers, Y. (2008). Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison between Scopus and Web of Science. *Journal of the American Society for Information Science and Technology, 59*(11), 1711–1726.

Narin, F., Frame, J. D., & Carpenter, M. P. (1983). Highly cited Soviet papers: An exploratory investigation. Social Studies of Science, 13(2), 307-319.

Oppenheim, C., & Renn, S. P. (1978). Highly cited old papers and the reasons why they continue to be cited. Journal of the American Society for Information Science,29(5), 225-231.

Ortega, JL. (2014). Academic Search Engines: A Quantitative Outlook. Elsevier, Chandos Information Professional Series

Orduña-Malea, E., Ayllón, J.M., Martín-Martín, A., Delgado López-Cózar, E. (2014). About the size of Google Scholar: playing the numbers.Granada: EC3 Working Papers, 18: 24 July 2014. http://arxiv.org/pdf/1407.6239

Orduña-Malea, E., Serrano-Cobos, J., & Lloret-Romero, N. (2009). Las universidades públicas españolas en Google Scholar: presencia y evolución de su publicación académica web. El profesional de la información, 18(5), 493-500.

Persson, O. (2010). Are highly cited papers more international?. Scientometrics,83(2), 397-401.

Plomp, R. (1990). The significance of the number of highly cited papers as an indicator of scientific prolificacy. Scientometrics, 19(3), 185-197.

Sanderson, M. (2008). Revisiting h measured on UK LIS academics. *Journal of the American Society for Information Science and Technology, 59*(7), 1184–1190.

Smith, D. R. (2009). Highly cited articles in environmental and occupational health, 1919–1960. Archives of environmental & occupational health, 64(sup1), 32-42.

Tijssen, R. J., Visser, M. S., & Van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: are highly cited research papers an appropriate frame of reference? Scientometrics, 54(3), 381-397.

Appendix A. List of the top 1% most-cited documents in our sample

The original publication of this working paper included the list of the top 1% most-cited documents as a table within the text document. We now refer to the supplemental data file where this list can be found:

Martín-Martín, A., Orduña-Malea, E., Manuel Ayllón, J., & Delgado-López-Cózar, E. (2014, October 31). Highly Cited Documents on Google Scholar (1950-2013) (Version 2). figshare. https://doi.org/10.6084/m9.figshare.1224314.v2

Appendix B. A case study: *The Mathematical Theory of Communication* in Google Scholar

This work, because of its bibliographic and bibliometric complexity, collects and illustrates the problems posed by this working paper on the treatment of highly cited documents. Therefore, it has been taken as a special case study, to develop it further.

Complexity

"A mathematical theory of communication" constitutes an article by Claude Shannon in 1948 in the Bell System Technical Journal and that was divided in two parts published separately.

Later, in 1949, this work is expanded and reedited in book form, published by the University of Illinois Press. On this occasion, is published co-authored by Claude Shannon and Warren Weaver, and the title varies imperceptibly: "The mathematical theory of communication".

Problematic

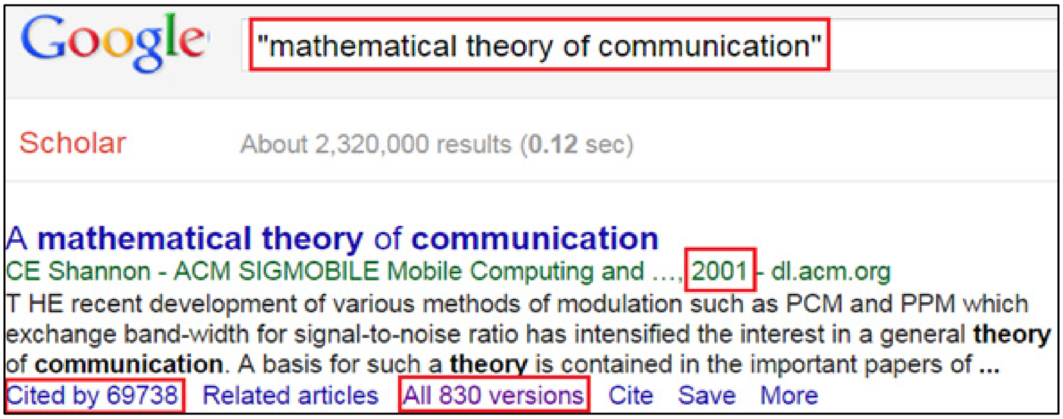
Despite being two articles published in 1948 and a book published in 1949, this work appears in the results of our analysis, which we remind that is limited to the period 1950-2013. So this raises a key question: Why this document appears in our sample?

Additionally, the fact that it is composed of two distinct works (article and book), both before 1950, generating different editions and different citations, raises a number of additional issues, which affect the functioning of the versions in Google Scholar as well as a number of additional issues raised in this working paper, for example:

Has GS identified all editions of the same document? Were successfully linked all editions of the same document? Were all citations received by each edition successfully merged? Was each citation successfully linked to each of the different editions?

Bibliographic search

In Figure B1 we show the query search for the work in Google Scholar by identifying the result with a higher number of versions.



"mathematical theory of communication"

About 2,320,000 results (**0.12** sec)

A **mathematical theory** of **communication**
CE Shannon - ACM SIGMOBILE Mobile Computing and ..., 2001 - dl.acm.org
T HE recent development of various methods of modulation such as PCM and PPM which exchange band-width for signal-to-noise ratio has intensified the interest in a general **theory** of **communication**. A basis for such a **theory** is contained in the important papers of **...**

Figure B1. Principal version of The mathematical theory of communication in Google Scholar

Even before trying to study the number of versions gathered, this point raises a fundamental issue: has Google Scholar merged all versions of this work?

To do this, we proceeded to refine this query (adding the search command author:Shannon), obtaining a total of 230 results, which have been analyzed to see which of them should not constitute a version (the raw data of this analysis is available in the complementary material).

Of the 230 results, 71.7% (165) are records that correspond to versions of the work, while the rest are not true versions, and they appear in the results of this query because they comment or review Shannon’s work.

Of the 165 records, the first one includes a larger number of versions (shown in Figure B1). However, in the remaining 164 records, there are 3,714 potential citations (without eliminating possible duplicates).

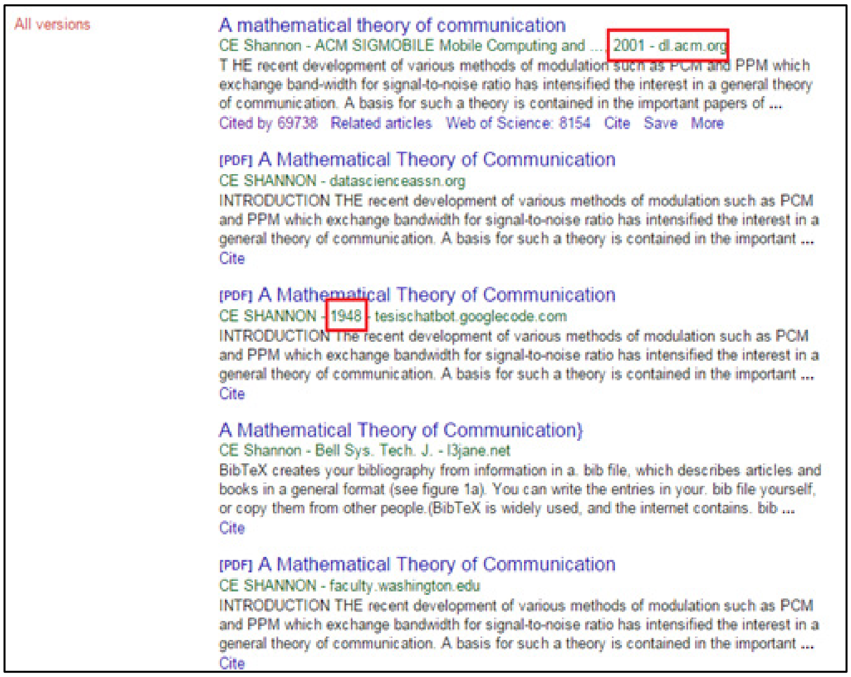
Number of versions for the main record

On the one hand, we can observe the large amount of retrieved versions (830) and, on the other, that among these versions we can find both versions of the article, and versions of the book, although the latter are a minority.

Otherwise, a manual analysis gives us only 763, not the 830 displayed in Figure B1. That is, the figure shown is an approximation of the number of recovered versions. This effect has already been seen in the Hit Count Estimates to general queries (Orduna-Malea et al, 2014).

Edition

However, the biggest problem we found was the year of publication (2001). This is taken from the primary version, which corresponds with the last identified reprint of the book, although the dates for the rest of the versions themselves are properly identified (Figure B2).



A mathematical theory of communication
CE Shannon - ACM SIGMOBILE Mobile Computing and ..., 2001 - dl.acm.org
T HE recent development of various methods of modulation such as PCM and PPM which exchange band-width for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of ...
Cited by 69738 Related articles Web of Science: 8154 Cite Save More

[PDF] A Mathematical Theory of Communication
CE SHANNON - datascienceassn.org
INTRODUCTION THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important ...
Cite

[PDF] A Mathematical Theory of Communication
CE SHANNON - 1948 - tesischatbot.googlecode.com
INTRODUCTION The recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important ...
Cite

A Mathematical Theory of Communication}
CE Shannon - Bell Sys. Tech. J. - l3jane.net
BibTeX creates your bibliography from information in a. bib file, which describes articles and books in a general format (see figure 1a). You can write the entries in your. bib file yourself, or copy them from other people.(BibTeX is widely used, and the internet contains. bib ...
Cite

[PDF] A Mathematical Theory of Communication
CE SHANNON - faculty.washington.edu
INTRODUCTION THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important ...
Cite

Figure B2. Versions grouped in Google Scholar about The mathematical theory of communication

Therefore, as the primary version gives the publication year, this causes it to appear in the results of our sample, which should have been limited to the period 1950-2013.

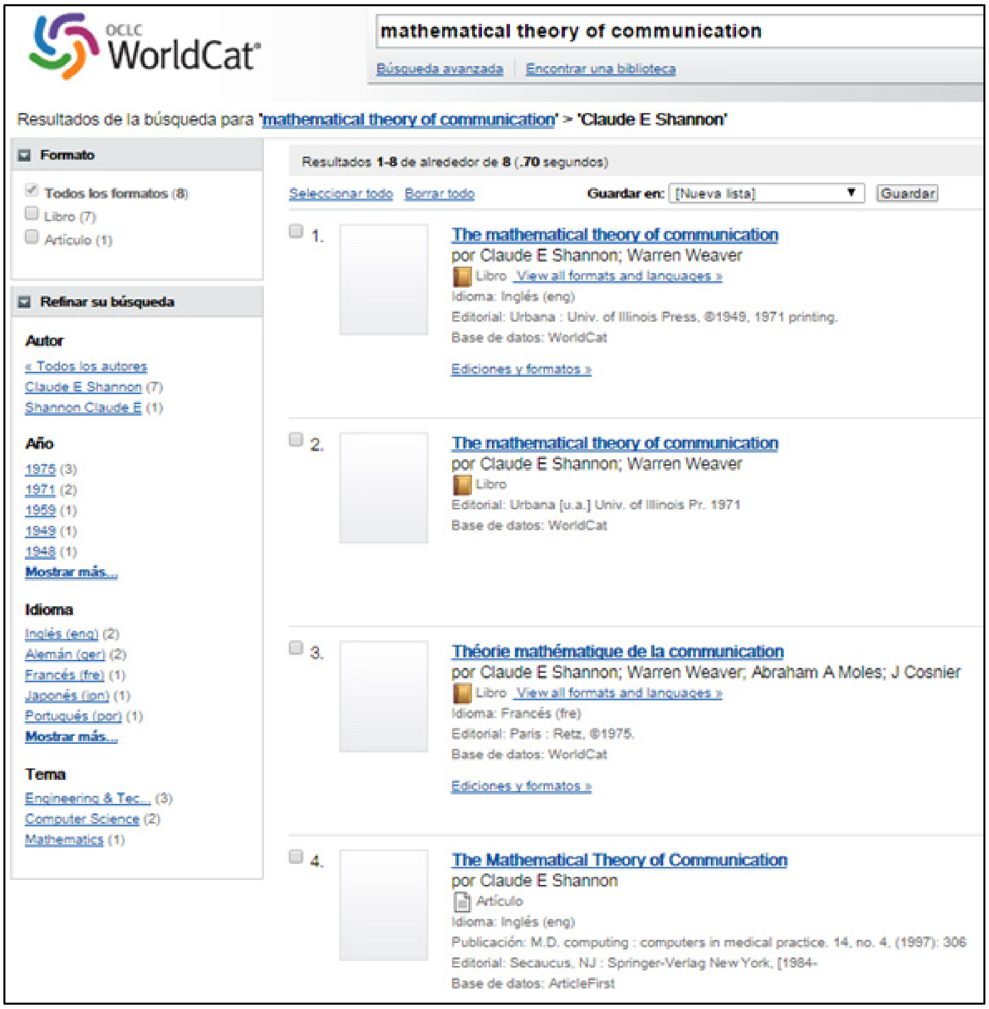
Given the magnitude and global impact of the Shannon and Weaver work, the number of reprints and editions (in different languages) are very high, and we should also add some book reviews (as sometimes are taken as versions).

These other inquiries are answered in the remaining questions raised in the working paper, in a more detailed and comprehensive way, but especially detectable in this case study.

The decision of taking the publication date from the last available release is understandable from the point of view of the search engine service, although it limits its potential as an object of bibliometric analysis, and therefore should be considered. Nonetheless, it is likely that the number of highly cited works that are affected by this issue relatively low.

As a counterpoint to the model taken by Google Scholar to group different versions (Figure B2), an example from the library world (Figure B3) is offered, using a catalog where authority control is used (Worldcat).

As can be seen, the system recovers, after querying for the title "mathematical theory of communication" and author (“Shannon”), different versions of the book (in different languages), as well as the original article.



mathematical theory of communication

Resultados de la búsqueda para 'mathematical theory of communication' > 'Claude E Shannon'

Formato

☑ Todos los formatos (8)
□ Libro (7)
□ Artículo (1)

Refinar su búsqueda

Autor

« Todos los autores
Claude E Shannon (7)
Shannon Claude E (1)

Año

1975 (3)
1971 (2)
1959 (1)
1949 (1)
1948 (1)
Mostrar más...

Idioma

Inglés (eng) (2)
Alemán (ger) (2)
Francés (fre) (1)
Japonés (jpn) (1)
Portugués (por) (1)
Mostrar más...

Tema

Engineering & Tec... (3)
Computer Science (2)
Mathematics (1)

Resultados **1-8** de alrededor de **8** (.**70** segundos)

Seleccionar todo Borrar todo **Guardar en:** [Nueva lista] ▼ Guardar

□ 1. The mathematical theory of communication
por Claude E Shannon; Warren Weaver
Libro View all formats and languages »
Idioma: Inglés (eng)
Editorial: Urbana : Univ. of Illinois Press, ©1949, 1971 printing.
Base de datos: WorldCat
Ediciones y formatos »

□ 2. The mathematical theory of communication
por Claude E Shannon; Warren Weaver
Libro
Editorial: Urbana [u.a.] Univ. of Illinois Pr. 1971
Base de datos: WorldCat

□ 3. Théorie mathématique de la communication
por Claude E Shannon; Warren Weaver; Abraham A Moles; J Cosnier
Libro View all formats and languages »
Idioma: Francés (fre)
Editorial: Paris : Retz, ©1975.
Base de datos: WorldCat
Ediciones y formatos »

□ 4. The Mathematical Theory of Communication
por Claude E Shannon
Artículo
Idioma: Inglés (eng)
Publicación: M.D. computing : computers in medical practice. 14, no. 4, (1997): 306
Editorial: Secaucus, NJ : Springer-Verlag New York, [1984-
Base de datos: ArticleFirst

Figure B3. The mathematical theory of communication in Worldcat

Analysis

To further analyze this case, we have proceeded to download the 830 versions of the book "The Mathematical Theory of Communication", belonging to the primary version in Google Scholar, in order to understand and describe it.

The raw data for this analysis is available in the supplementary material, in a spreadsheet file. For each version, we have considered the same parameters that we have used for the 64,000 highly cited documents, fully described in the Introduction of the Working Paper.

Additionally, for each version, we manually checked if it was done correctly or not. And in those cases where the connection is unsuccessful, we have classified the different errors into categories, further detailing the reason for the error where needed.

Of the total 830 versions, Google Scholar has really returned only 763, of which 602 (78.9%) are working properly. In the remaining versions (161), the following problems occur:

- False positive: when a document has been identified as a version of another document, but actually it is not.

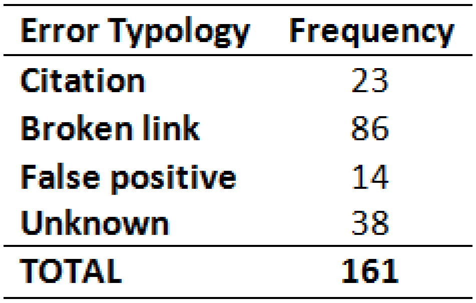
- Citation: false positive specific case, when the identified version is a citation rather than a document.

- Broken Link: If the link is not working properly.

- Unknown: when we have not been able to verify if the version was correct. This has occurred mainly in cases in which the files were available in PS (PostScript) file format.

In Table B1 we summarize all data about errors in the 830 different versions grouped.

Table B1. Types of errors in the different versions of a document



Error Typology	Frequency
Citation	23
Broken link	86
False positive	14
Unknown	38
TOTAL	**161**

Appendix C. Frequency table: number of highly-cited documents in our sample published in WoS-covered journals

The original publication of this working paper included the frequency table of the number of highly-cited documents published in WoS-covered journals as a table within the text document. We now refer to the supplemental data file where this table can be found:

Martín-Martín, A., Orduña-Malea, E., Manuel Ayllón, J., & Delgado-López-Cózar, E. (2014, October 31). Highly Cited Documents on Google Scholar (1950-2013) (Version 2). figshare. https://doi.org/10.6084/m9.figshare.1224314.v2

Chapter 3. A two-sided academic landscape: portrait of highly-cited documents in Google Scholar (1950-2013)

Journal Article. Cite as:

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). Revista Española de Documentacion Cientifica, 39(4), e149. https://doi.org/10.3989/redc.2016.4.1405

Abstract (English)

Despite its well-known limitations, the wide coverage of Google Scholar has various advantages when used as a tool to collect highly-cited documents. The main objective of this paper is to identify the set of highly-cited documents in Google Scholar and to define their core characteristics (document types, language, free availability, source providers, and number of versions). To do this, a longitudinal analysis was carried out by performing 64 keyword-free year queries, from 1950 to 2013 (one query per year). All available records (a maximum of 1,000 per query) were collected, obtaining a set of 64,000 records of which 40% provided a free full-text link. According to the results, the average highly-cited document is a journal article (72.3% of the documents for which a document type could be ascertained) or a book (62% of the top 1% most cited documents of the sample), written in English (92.5% of all documents) and available online in PDF format (86.0% of all documents). The research concludes that Google Scholar data offer an original and different vision of the most influential academic documents (measured from the perspective of their citation count). Moreover, these data enable the measurement of impact that stems from not only the scientific side of the academic landscape, but also from the educational side (doctoral dissertations, handbooks) and from the professional side (working papers, technical reports, patents), the last two being areas that haven't been explored as much as the first one.

Abstract (Spanish)

A pesar de sus conocidas limitaciones, la amplia cobertura de Google Scholar posee diversas ventajas a la hora de ser utilizada como herramienta para recopilar documentos altamente citados. El principal objetivo de este trabajo es identificar el conjunto de documentos altamente citados en Google Scholar y definir sus características nucleares (tipología documental, idioma, disponibilidad en abierto, fuentes y número de versiones). Para ello, se ha llevado a cabo un análisis longitudinal a partir de la ejecución de 64 consultas (incluyendo el año y excluyendo palabras clave incluida), desde 1950 hasta 2013 (una consulta por año). Los registros obtenidos (1.000 por consulta máximo) fueron recogidos, obteniendo una muestra de 64.000 registros (el 40% de los cuales proporcionaban un enlace al texto completo). Teniendo en cuenta los resultados obtenidos, el documento altamente citado "promedio" es un artículo de revista (considerando únicamente los documentos en los que se pudo determinar su tipología, que corresponden con el 72.3%) o libro (constituyen el 62% del top 1% de los documentos más citados de la muestra), escrito en inglés (92.5%) y disponible online en PDF (86% de la muestra). Se concluye que Google Scholar ofrece una visión original y diferente de los documentos académicos más influyentes (medidos desde la perspectiva de la contabilización de citas). Además, los datos obtenidos permiten la medición no sólo desde el punto de vista científico del panorama académico, sino además desde el lado educacional (tesis, manuales) y profesional (working papers, informes técnicos, patentes), áreas estas últimas menos exploradas.

1. Introduction

The idea of identifying the most influential scientific documents using the number of times they are cited in the scientific literature was introduced, like many other bibliometric procedures, by Garfield (1977). The candidates for "Citation classics" were selected from a group of the 500 most cited papers during the years 1961-1975 (http://garfield.library.upenn.edu/classics.html). From 2001, the highly cited papers were integrated in a new product: The Essential Science Indicators. Nevertheless, no other bibliographic database has released alternatives to this product.

However, we do have an extensive scientific literature on the matter of highly-cited documents in different journals, subject areas, institutions or countries (Oppenheim & Renn, 1978; Narin et al., 1983; Plomp, 1990; Glänzel & Czerwon, 1992; Glänzel & Schubert, 1992; Glänzel et al., 1995; Tijssen et al., 2002; Aksnes, 2003; Aksnes & Sivertsen, 2004; Kresge et al., 2005; Levitt & Thelwall, 2009; Smith, 2009; Persson, 2010).

Recently, the interest in these lists has returned with the development of rankings based on the concept of excellence through the calculation of percentiles, first proposed by Maltrás (2003) and recently popularized by other authors (Bornmann, 2010; Bornmann & Mutz, 2011; Bornmann et al., 2011).

To celebrate the fiftieth anniversary of the Science Citation Index, the journal *Nature* asked *Thomson Reuters* for the list of the top 100 most highly-cited papers of all time (Van Noorden et al., 2014). In this list, the classic "Protein measurement with the folin phenol reagent", by Lowry et al. (1951), achieves the first position, a place it has historically occupied (Garfield, 2005; Kresge et al., 2005). Although the authors explore the most-cited research of all time using data from the Web of Science Core Collection (WoScc), they also provide an alternative ranking using data from Google Scholar (GS). This alternative ranking is only available in the online version of that article as supplementary material.[1]

The appearance of Google Scholar at the end of 2004 signalled a revolution in the way scientific publications were searched, retrieved and accessed (Jacsó, 2005), becoming not only a search engine for academic documents, but also for the citations these documents receive (Ortega, 2014).

Google Scholar's crawlers systematically parse and analyse the entire academic web, not making distinctions based on subject areas, languages, or countries. This enables the calculation of impact metrics for a broader collection of documents, not only articles published in elite journals that are included in expensive citation indexes. Disciplines inside the Social Sciences and the Humanities, which use other channels of scientific communication apart from journal articles (such as doctoral theses, books, book chapters, working papers, and conference proceedings) could benefit from using this much broader source of scientific publication data (Meho & Yang, 2007; Harzing & Van der Wal, 2008; Bar-Ilan, 2010; Kousha et al., 2011; Kousha & Thelwall, 2008).

Its wide coverage and evolution (Aguillo, 2012; Khabsa & Giles, 2014; Ortega, 2014; Winter et al., 2014; Orduna-Malea et al., 2015) as well as its empirically tested capacity to obtain unique citations (citations that can only be found in Google Scholar) (Yang & Meho, 2006; Meho & Yang, 2007; Kousha & Thelwall, 2008; Bar-Ilan, 2010; Kousha et al., 2011; Harzing, 2013; Harzing, 2014; Orduna-Malea & Delgado López-Cózar, 2014), make of Google Scholar an exceptional source to collect highly-cited documents.

One issue that should be taken into account when using bibliographic and bibliometric data provided by Google Scholar is that the data may present errors. These errors have been already studied and classified (Jacsó, 2005; 2006; Bar-Ilan 2010; Jacsó 2008a; 2008b; 2012). Although the quality of the data has improved significantly over the years (Google Scholar is now over 11 years old), some of these errors still persist, especially those related to the detection of duplicate documents, and the correct allocation of citations (Martín-Martín et al., 2015; Orduna-Malea et al., 2015). Thus, Google Scholar data usually requires some cleaning before it is suitable for analysis. Failing to observe this measure might lead to unreliable results. This is the case of Nature's ranking of highly cited documents according to Google Scholar (Van Noorden et al., 2014), which presents various irregularities (Martín-Martín et al., 2015).

In spite of these shortcomings, Google Scholar is capable not only of identifying the most-cited papers, but also of providing a view of a broader academic landscape (including books, heavily cited in certain fields, and traditionally discriminated against).

It is important to note that Nature's ranking was drawn from the data that the Google Scholar's team provided directly to the authors. It would be necessary therefore to ascertain whether such listings could be obtained by an independent user through the use of specific search queries. This task has been carried out successfully (see supplementary material), demonstrating the soundness of Google Scholar for retrieving highly-cited documents, and providing an opportunity for the execution of studies describing the key bibliographic aspects of these highly-cited items. The unique coverage policy of Google Scholar (virtually no language, country, subject area, or document type restrictions) may provide interesting insights to the bibliometric community for understanding the characteristics of highly-cited documents.

Although some of the bibliographic properties of the documents indexed in Google Scholar (such as its sources or document types) have been previously treated in the existing literature, these works have never focused on samples made up entirely of highly-cited documents. Aguillo (2012) and Ortega (2014) performed two separate general analyses of the search engine (without considering the number of citations received by documents), while Jamali and Nabavi (2015) studied a sample of 8310 documents in different disciplinary fields (the 277 subcategories offered by Scopus), and limited to the period 2004-2014. In fact, the use of keyword queries prevented the authors from isolating highly-cited papers, since those queries were affected by Google Scholar's academic search engine optimization practices (Beel et al., 2010). This issue is circumvented in this work by means of using keyword-free year queries.

Therefore, the main objective of this paper is to identify the set of highly-cited documents in Google Scholar and define their core characteristics, in order to give an answer to the following research questions:

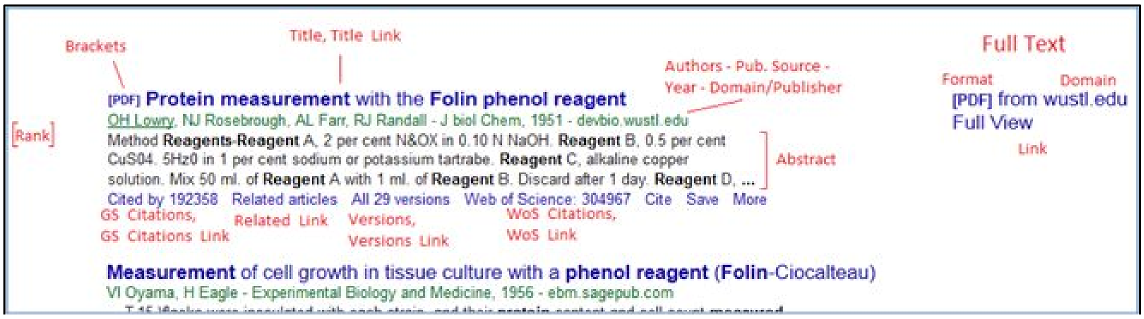
- Which are the most cited documents in Google Scholar?
- Which is the most frequent document type for these highly-cited documents?
- In what languages are the most cited documents written?
- How many highly-cited documents are freely accessible?
- What are the most common file formats to store these highly cited documents?
- Which are the main providers of these highly-cited full text documents?
- How many versions has Google Scholar found of these highly-cited documents?

2. Methods

In order to isolate a sample of highly-cited documents, we performed a series of keyword-free year queries (only the year field in Google Scholar's advanced search was used). By doing this, the results of the queries weren't limited to a specific topic.

A longitudinal analysis was carried out by performing 64 keyword-free year queries from 1950 to 2013 (one query per year). All the records displayed (a maximum of 1,000 per query) were extracted, obtaining a final set of 64,000 records.

This process was carried out twice (on the 28th of May, and on the 2nd of June, 2014). In the first case, it was performed from a computer with access to the Web of Science, in order to obtain WoS data embedded in Google Scholar (http://wokinfo.com/googlescholar). In the second case, the data extraction was made from a computer with a normal Internet connection, because we wanted to collect data about free full-text links that couldn't have been unadulterated by our university's subscriptions. This process doubled as a reliability check, because we confirmed that the two datasets contained the same records. After this, the HTML source code for each of the search engine result pages of every query was parsed, extracting all the bibliographic information available for every record (Fig. 1).



[PDF] Protein measurement with the Folin phenol reagent
OH Lowry, NJ Rosebrough, AL Farr, RJ Randall - J biol Chem, 1951 - devbio.wustl.edu
Method Reagents-Reagent A, 2 per cent N&OX in 0.10 N NaOH. Reagent B, 0.5 per cent CuS04. 5Hz0 in 1 per cent sodium or potassium tartrabe. Reagent C, alkaline copper solution. Mix 50 ml. of Reagent A with 1 ml. of Reagent B. Discard after 1 day. Reagent D, ...
Cited by 192358 Related articles All 29 versions Web of Science: 304967 Cite Save More

[PDF] from wustl.edu
Full View

Measurement of cell growth in tissue culture with a phenol reagent (Folin-Ciocalteau)
VI Oyama, H Eagle - Experimental Biology and Medicine, 1956 - ebm.sagepub.com

Figure 1. Fields extracted from each Google Scholar record in the search engine results page

The main fields extracted were the following: author name(s), publication source, year of publication, GS citations, and number of versions.

The full-text fields are available only when Google Scholar finds at least one freely accessible version among all the versions identified of a same document. In the cases where more than one free version is found, Google Scholar selects one of them and displays it right next to the bibliographic information of the primary version of the document. This study analyses only those selected full-text links, not all the full-text links that may be found when clicking the “View all X versions” link of a Google Scholar result. For each document with full-text data, the following fields were extracted: domain (the web domain where GS has found a free full-text version of the document), link (hyperlink to the full-text of the document), and format (file type of the full-text version of the document).

In addition to these fields, information about the document type and the language of the document, which are not systematically provided by Google Scholar, were assigned to each record as well.

Regarding the document types, some records display a text in square brackets before the title of the document (for example “[BOOK]”). Regrettably, this text is not always offered and in some cases the information does not refer to document types but to file types (for example “[PDF]” or “[HTML]”) or it is used to mark some special records (such as “[CITATION]”, references to a document that have been found cited in the reference list of a document indexed in Google Scholar, but are not linked to any web source).

Given the difficulty of ascertaining the typologies of the documents indexed in Google Scholar, we devised three different strategies that, combined, allowed us to some extent to define the typology of the documents in the data set:

a) All documents where the field brackets = “[BOOK]” were considered as books (codified as “B”).
b) For documents that were also indexed in WoS, Google Scholar data was merged with WoS data to obtain the document types. The correspondence is as follows:
    - Journal (“J”): “Article”, “Letter”, “Note”, “Reviews”.
    - Book (“B”): “Book”, “Book Chapter”.
    - Conference proceedings (“C”): “Proceedings Papers”.
    - Others (“O”): “Book Review”, “Correction”, “Correction, Addition”, “Database Review”, “Discussion”, “Editorial Material”, “Excerpt”, “Meeting Abstract”, “News Item”, “Poetry”, “Reprint”, “Software Review”.
c) Lastly, we analyzed the publication source (where possible), searching for keywords (in different languages) that could indicate the type of the source publication, searching the following terms:
    - Journal (“J”): “Revista”, “Anuario”, “Cuadernos”, “Journal”, “Revue”, “Bulletin”, “Annuaire”, “Anales”, “Cahiers”, “Proceedings”.
    - Conference Proceedings (“C”): “Proceedings”, “Congreso”, “Jornada”, “Seminar”, “Simposio”, “Congrès”, “Conference”, “symposi”, “meeting”.

Since the word “Proceedings” is used both for journals (i.e. “Proceedings of the National Academy of Sciences”) and for conference proceedings (i.e., “Proceedings of the 4th Conference…”), records containing this word in the publication source field were all considered initially as conference proceedings, but a manual check was carried out to reassign those that were really journal articles.

With respect to the language of the documents (GS doesn’t provide this information either), we manually checked the language in which the title and abstract of the document were written as well as WoS data (when available), as a basis to fill the language field.

Lastly, all the data was saved to a spreadsheet so it could be statistically analyzed. Pearson and Spearman correlations (α=0.01) were calculated with the XLstat statistical suite in order to find the connection between versions and citations.

3. Results

The most cited documents in Google Scholar

The top 25 most cited documents in GS (1950-2013) are listed in Table 1. In the case of books, the year of publication is the year of publication of the first edition. The top 1% most cited documents in our sample (640 documents) are provided in the supplementary material.[1]

Table 1. Top 25 most-cited documents in Google Scholar (1950-2013)

R	DOCUMENT (Author, Title, Publisher)	YEAR (1ST ED.)	CITATIONS	TYPE
1	LOWRY, O.H. et al. Protein measurement with the Folin phenol reagent. *The Journal of biological chemistry*.	1951	253,671	J
2	LAEMMLI, U.K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*.	1970	221,680	J
3	BRADFORD, M.M. A rapid and sensitive method for the quantitation of microgram quantities of protein using the principle of protein Dye binding. *Analytical Biochemistry*.	1976	185,749	J
4	SAMBROOK, J., FRITSCH, E. F., & MANIATIS, T. Molecular cloning: a laboratory manual. New York, Cold Spring Harbor Laboratory Press.	1982	171,004	B
5	AMERICAN PSYCHIATRIC ASSOCIATION. Diagnostic and statistical manual: mental disorders. Washington, American Psychiatric Assn.	1952	129,473	B
6	PRESS, W. H. Numerical recipes: the art of scientific computing. Cambridge: Cambridge University Press.	1986	108,956	B
7	YIN, R. K. Case study research: design and methods. Beverly Hills (CA): Sage Publications.	1984	82,538	B
8	ABRAMOWITZ, M., & STEGUN, I. A. Handbook of mathematical functions: with formulas, graphs, and mathematical tables. Washington, Government printing office.	1964	80,482	B
9	KUHN, T. S. The structure of scientific revolutions. Chicago, University of Chicago Press.	1962	70,662	B
10	ZAR, J. H. Biostatistical analysis. Englewood Cliffs: Prentice Hall international.	1974	68,267	B
11	SHANNON, C.E. A mathematical theory of communication. *The Bell System Technical Journal*.	*1948	66,851	J
12	CHOMCZYNSKI & SACCHI, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry*	1987	63,871	J
13	SANGER F, NICKLEN S, & COULSON AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*.	1977	63,767	J
14	COHEN, J. Statistical power analysis for the behavioral sciences. New York: Academic Press.	1969	63,766	B
15	GLASER, B. G., & STRAUSS, A. L. The discovery of grounded theory: strategies for qualitative research. New York: Aldine de Gruyter.	1967	61,158	B
16	NUNNALLY, J. C. Psychometric Theory. New York: McGraw-Hill.	1967	60,725	B
17	GOLDBERG, D. E. Genetic algorithms in search, optimization, and machine learning. Reading, Mass: Addison-Wesley.	1989	59,764	B
18	ROGERS, E. M. Diffusion of Innovations. Pxiii. 367. Free Press of Glencoe, New York: Macmillan.	1962	55,738	B
19	BECKE, A.D. Density Functional Thermochemistry III The Role of Exact Exchange. *J. Chem. Phys*.	1993	54,642	J
20	LEE, C., YANG, W. & PARR, R.G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*.	1988	52,316	J
21	MURASHIGE, T. & SKOOG, F. A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiologia Plantarum*.	1962	52,011	J
22	ANDERSON, B. R. O. Imagined communities: reflections on the origin and spread of nationalism. London: Verso.	1983	51,177	B
23	FOLSTEIN, M.F., FOLSTEIN, S.E. & MCHUGH, R. Mini-mental state. *Journal of Psychiatric Research*.	1975	51,150	J
24	TOWBIN, H., STAEHELIN, T. & GORDON, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the National Academy of Sciences of the United States of America*.	1979	50,608	J
25	PAXINOS, G., & WATSON, C. The rat brain in stereotaxic coordinates. Sydney [etc.]: Academic Press.	1982	50,471	B

J: Article journal; B: Book;
* Contribution published outside the studied timeframe; fully commented on in the discussion.

The most cited document according to GS is the aforementioned article by Lowry et al, with 253,671 citations (as of May 2014), followed by Laemmly's article, with 221,680 citations.

Although the ranking is dominated by studies from the natural sciences (especially the life sciences), it also contains many works from the social sciences (especially from economics, psychology and sociology), and also from the humanities (philosophy and history). For instance, within the top 20 documents we can find "The structure of scientific revolutions (9th position; 70,662 citations) and "Diffusion of innovations" (18th; 55,738 citations).

Many of the works in this ranking are methodological in nature: they describe the steps of a certain procedure or how to handle basic tools to process and analyse data. This is exemplified by the presence of statistical manuals ("Handbook of Mathematical Functions with Formulas"), laboratory manuals ("Molecular cloning: a laboratory manual"), manuals of research methodology ("Case study research: design and methods"), and works that have become a de facto standard in professional practice ("Diagnostic and statistical manual of mental disorders").

In fact, books are the most common category among the top 1% most cited documents, constituting the 62% (395) of this subsample, followed by journal articles with 36.01% (231). Moreover, the citation average of books (2,700) is higher than that for journal articles (1,700).

Document types

The document type has been identified in 71% (45,440) of the documents sampled, whereas the typology of the other 29% (18,590) remained unknown (our automatic strategies weren't able to determine it, and manual checking would have been too costly). The distribution of document typologies is displayed in Figure 2, where we find a clear predominance of journal articles (including reviews, letters and notes as well) which represent 51% of the total 64,000 documents (72.3% of the documents with a defined document type). Book and book chapters together also make up a big part of the sample (18%; 11,240 items) while the presence of conference proceedings and other typologies (meeting abstracts, corrections, editorial material, etc.) is merely testimonial (1% each).

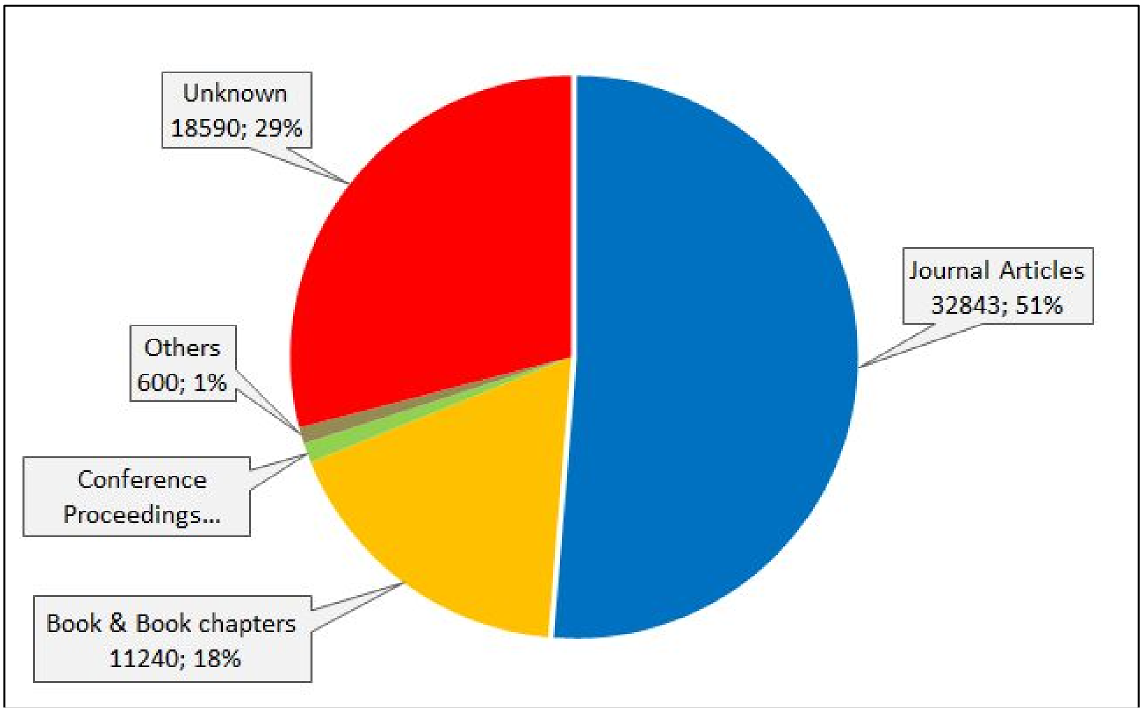


Figure 2. Document types of the highly cited documents in Google Scholar

Figure 3 represents this distribution in a longitudinal perspective, where we can observe the following three phenomena:

- Conference proceedings and "Others" categories play an insignificant role along the years, although they achieve greater presence during the last decade.
- A steady decrease over time in the number of documents with an unknown typology (from 35.4% in 1950 to 12.9% in 2013).

- A constant increase in the number of books, which become the most frequent document type in the last five years (2009-2013), monopolizing the sample. As an example, within the 1,000 results corresponding for the year 2013, we only find 27 journal articles but 842 books. The reason for this overrepresentation of the book format in the most recent years is explained in the discussion section of this article.

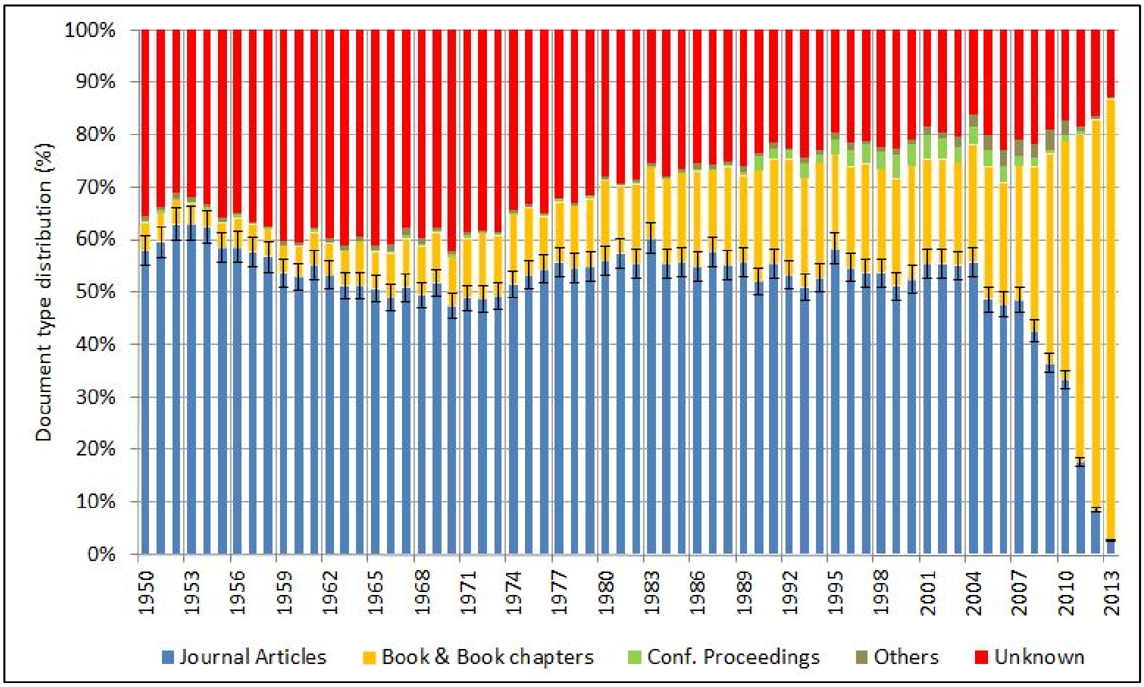


Figure 3. Document types of the highly cited documents in Google Scholar, broken down by years (1950-2013)

Language of documents

In Figure 4 we find the document distribution according to language. As we can see, English dominates over the rest of the languages as the most widely used language for scientific communication in Google Scholar, accounting for 92.5% of all the documents. The second and third places are occupied by Spanish and Portuguese respectively, but neither of them reaches even 2% of the total.

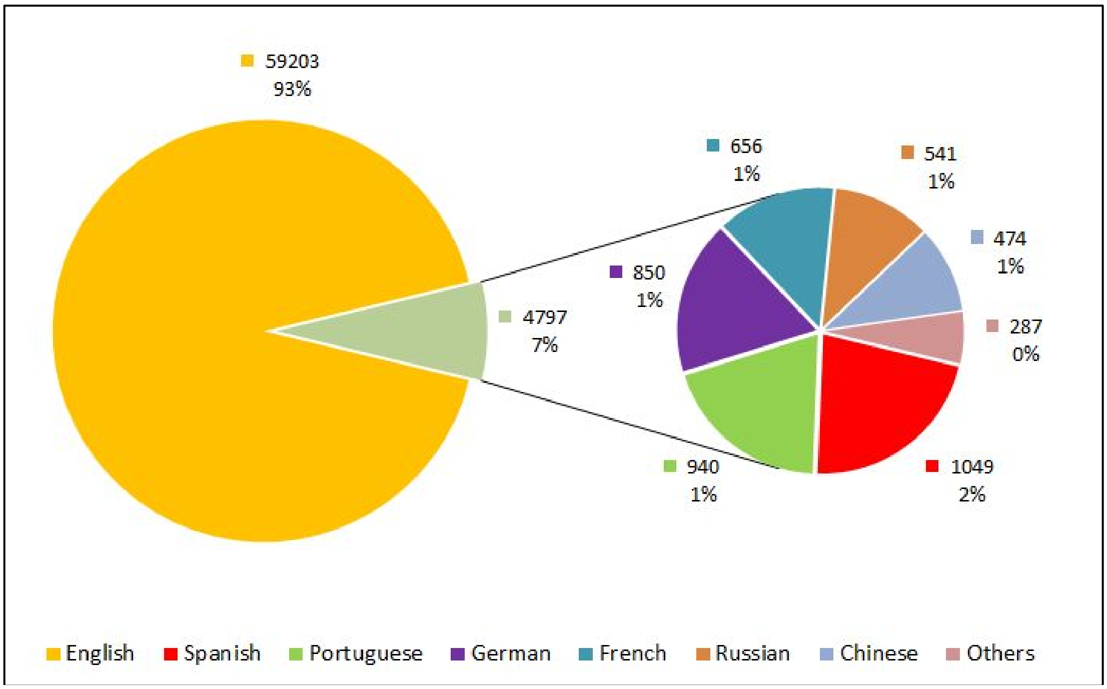


Figure 4. Distribution of languages used in the highly-cited documents in GS

In Figure 5 we can observe the longitudinal evolution of the language usage distribution, which is much more stable through the years than the ones previously found for the document types. The English language predominates during the whole period ($\widetilde{X}$= 92.5%; σ= 1.6%), with an oscillation of less than 10% between its maximum and minimum value (87% in 2013, and 95% in 1991). Data also shows a slightly decrease in English percentage in the last three years (from 92% in 2010 to 87.1% in 2013), though more data is required to determine if this change is just circumstantial or a new trend.

The "Others" category (which represents 7% of the documents) includes the following languages: Italian, Swedish, Indonesian, Finnish, Danish, Bulgarian, Polish, Norwegian, Turkish, Latin, Slovenian, Serbian, Dutch, Macedonian, Malayan, Japanese, Czech, Estonian, Slovak, Mongolian, Catalan, Croatian, Lithuanian, and Ukrainian.

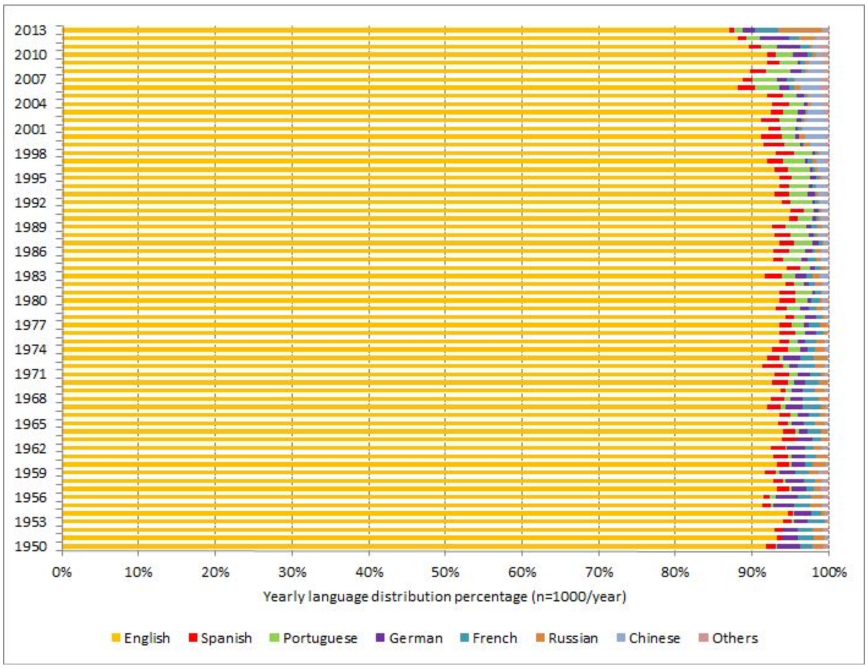


Figure 5. Distribution of languages in the highly cited documents in GS by years of publication (1950-2013)

Availability of Full text documents

A free full-text link is provided for 40% (25,849) of all the highly-cited documents retrieved (Figure 6; top). We can also observe a positive trend through the analyzed period (from 25.93% of documents with free full-text links in the period 1950-1959, to 66.84% in 2000-2009), although this trend is interrupted in the last four years (41.5% from 2010 to 2013), where the high percentage of books in these years are affecting the results (Figure 6; bottom). The journals' and publishers' embargo policies may have slight influence as well, especially for the experimental sciences.

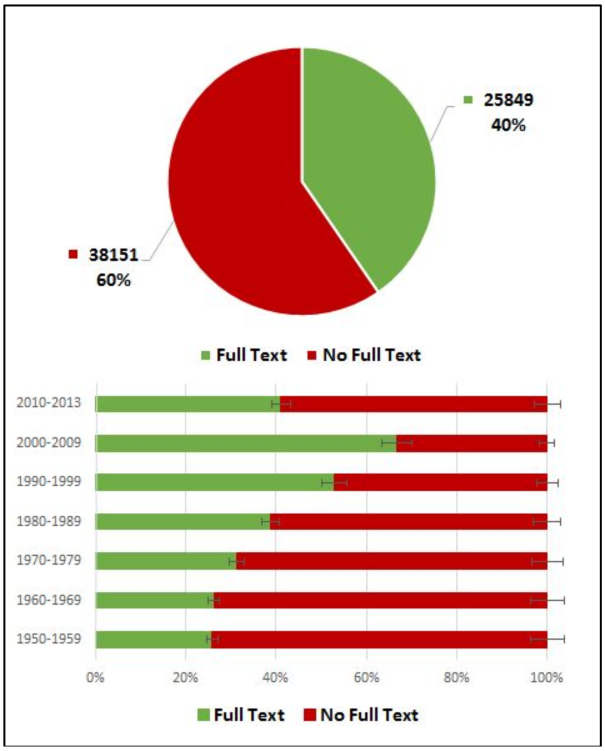


Figure 6. Percentage of freely accessible highly cited documents in Google Scholar (1950-2013). Global results (top); broken down by decades (bottom).

File types

Full-text links point to documents in a variety of formats (Figure 7). The most common one is the pdf format (86.0% of all full text documents), followed by the html format (12.1%). The remaining identified file formats (doc, ps, txt, rtf, xls, ppt) together only represent 1.9% of the freely available documents.

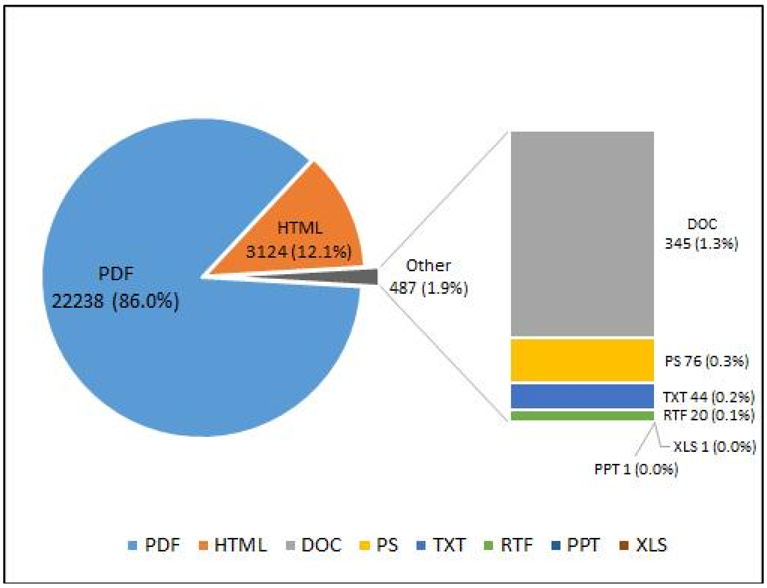


Figure 7. File Formats of the freely accessible highly cited documents in Google Scholar (1950-2013)

Figure 8 shows the same data broken down by years (1950-2013). We can see that the predominance of the pdf format is patent throughout the entire range of years. However, it is also noteworthy that the html format has started gaining more presence for documents published in the last 25 years, with a peak of almost 20% of the share in 2010.

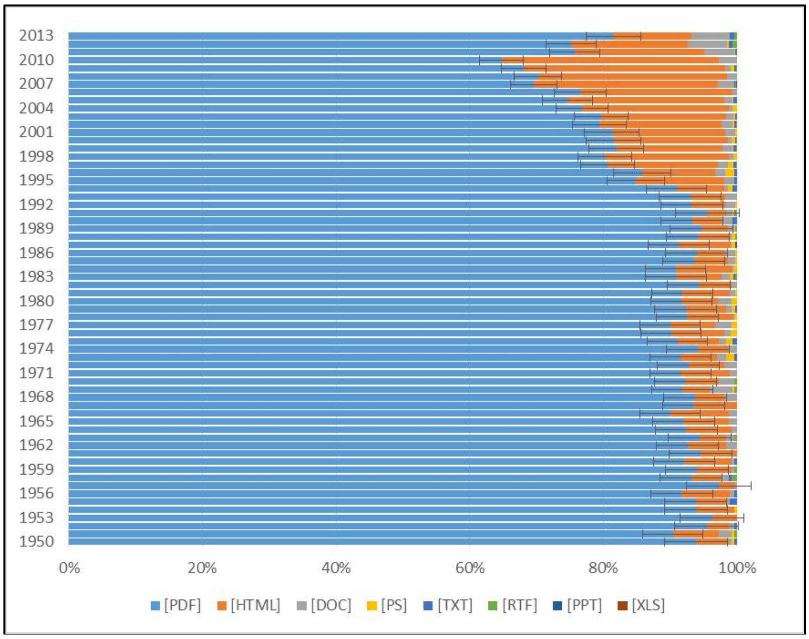


Figure 8. File Format distribution for the freely accessible highly cited documents in Google Scholar broken down by years (1950-2013)

Full-text source providers

A total of 5,715 different providers of free full-text links to highly cited documents have been found in the sample. However, a group of 35 providers (18 universities; 5 scientific societies; 4 publishers; 2 companies; 2 public administrations; 1 journal; 1 digital library; 1 repository; 1 academic social network) account for more than a third of all the links (37%).

Table 2 shows the main providers. The National Institutes of Health (NIH) hold the first position (1,405 documents), mainly due to the Pubmed central repository, hosted within the NIH website (http://www.ncbi.nlm.nih.gov/pubmed). The second position is occupied by ResearchGate (815), followed by Harvard University (495).

Table 2. Top full text source providers in Google Scholar (1950-2013)

Provider	Nº	Type of entity
nih.gov	**1,405**	Public administration
researchgate.net	**815**	Academic Social network
harvard.edu	**495**	University
pnas.org	**478**	Scientific society
oxfordjournals.org	**466**	Publisher
psu.edu	**424**	University
arxiv.org	**423**	Repository
jbc.org	**414**	Journal
sciencedirect.com	**394**	Publisher
wiley.com	**324**	Publisher
jstor.org	**322**	Digital library
rupress.org	**304**	University
royalsocietypublishing.org	**266**	Scientific society
ahajournals.org	**218**	Scientific society
dtic.mil	**208**	Public administration
stanford.edu	**203**	University
google.com	**188**	Company
mit.edu	**180**	University
tu-darmstadt.de	**177**	University
nature.com	**161**	Publisher
yale.edu	**141**	University
caltech.edu	**140**	University
physoc.org	**140**	Scientific society
cmu.edu	**122**	University
umich.edu	**120**	University
duke.edu	**118**	University
princeton.edu	**116**	University
wisc.edu	**113**	University
ucsd.edu	**112**	University
asm.org	**112**	Scientific society
berkeley.edu	**107**	University
upenn.edu	**104**	University
washington.edu	**103**	University
columbia.edu	**102**	University
yimg.com	**101**	Company
TOTAL	**9,616**	

If we analyse the top-level domains of the 25,849 links to full text available documents (Table 3), the most frequent are academic institutions (.edu; 23.74%) and organizations (.org; 21.39%). Moreover, the number of links provided by academic institutions is likely to be higher since there are many universities that use national top-level domains instead of .edu (mostly reserved for North American academic institutions). For example, Technische Universität Darmstadt (tu-darmstadt.de) provides 177 links. At a national scale, some countries use a "“ac.xx” pattern domain, such as United Kingdom (ac.uk), which provides 333 links. The most important geographic domain is Germany (.de) with only 2.62% (678) of the highly-cited documents.

Table 3. Top-level domains providing full text links in Google Scholar (1950-2013)

Domain	N	%
.edu	**6,136**	23.74
.org	**5,528**	21.39
.com	**3,466**	13.41
.gov	**1,712**	6.62
.net	**1,345**	5.20
.de	**678**	2.62
.cn	**489**	1.89
.uk	**485**	1.88
.ca	**404**	1.56
.ru	**374**	1.45
.fr	**357**	1.38
.br	**343**	1.33
.it	**275**	1.06
.ch	**214**	0.83
.mil	**210**	0.81
.nl	**186**	0.72
.es	**145**	0.56
.tw	**136**	0.53
.au	**131**	0.51
.in	**118**	0.46
Others	**3,117**	12.06
TOTAL	**25,849**	**100%**

Versions

83.17% (53,229) of the documents analyzed have more than one version (Table 4). The distribution of the number of versions is asymmetric, led by documents with 1 version (16.83; 10,771 documents) and followed by documents with 3 versions (6,903; 10.79%) and 4 versions (6,814; 10.65%). The existence of documents with a massive number of versions is also worth noting. For 281 documents, Google Scholar has found more than 100 versions, and more than 500 versions for 14 of those documents. The document with the highest number of versions in our sample has 899 versions.

Table 4. Distribution of documents according to their number of versions

Nº of versions	Nº of documents	%	Accumulated (docs)	Accumulated (%)
1	10,771	16.83	10,771	16.83
2	6,075	9.49	16,846	26.32
3	6,903	10.79	23,749	37.11
4	6,814	10.65	30,563	47.75
5	5,539	8.65	36,102	56.41
6	4,781	7.47	40,883	63.88
7	3,746	5.85	44,629	69.73
8	2,940	4.59	47,569	74.33
9	2,429	3.80	49,998	78.12
10	1,929	3.01	51,927	81.14
11-15	5,243	8.19	57,170	89.33
16-25	3,585	5.60	60,755	94.93
26-50	2,202	3.44	62,957	98.37
51-100	762	1.19	63,719	99.56
101-200	202	0.32	63,921	99.88
201-300	40	0.06	63,961	99.94
301-400	16	0.03	63,977	99.96
401-500	9	0.01	63,986	99.98
> 501	14	0.02	64,000	100

Pearson's correlation coefficient between the number of citations of a document in Google Scholar and its number of versions is low (r = 0.2; α= 0.01). However, the Spearman correlation shows a better correlation (r= 0.48; α= 0.01). This may be an effect of the highly skewed distribution of citations. For example, the average of citations for documents with at least 100 versions is high (5,878.13), although the Pearson's correlation of these highly-versioned documents with the corresponding number of citations is even lower (r= 0.13).

4. Discussion

An in-depth discussion of this radiography of highly-cited documents in Google Scholar is necessary, due to the limitations of the database. We will first consider the key parameters that may have influenced the ranking presented in Table I (essentially the dynamic of citations received, and the number of versions). Next, we'll warn about some flaws that affect the composition of the sample (related to the publication date and the language of the documents). Lastly, we will comment on some specific properties of the documents in our sample (document types, full text, file formats, and providers).

Key parameters

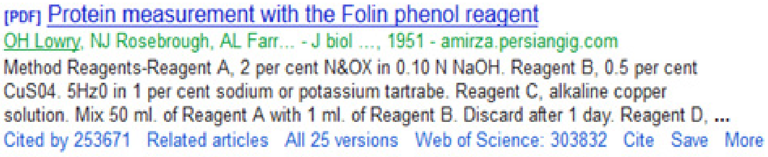
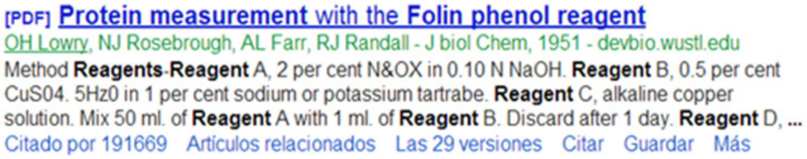
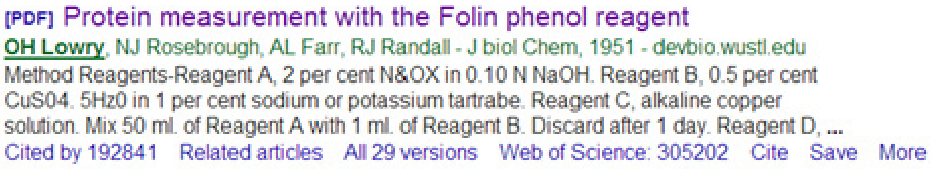
The fluctuation of citations

In this section we set aside the issues regarding the quality and the source of the citations received by the 64,000 documents analyzed, and the well-known errors related to the inaccurate attribution of citations (which is not so important when we are studying highly-cited documents). Instead, we will focus on an issue which might significantly distort the results of this kind of studies: the fluctuation of citations in Google Scholar.

Unlike in other bibliographic databases (such as Scopus or Web of Science core collection), Google Scholar reflects the number of citations considering the documents that are available on the Web at the time the search is made. Google Scholar's team warns that the database "reflects the state of the web as it is currently visible to our search robots and to the majority of users" (http://scholar.google.com/intl/en/scholar/help.html#corrections). This means that citation counts may decrease if, for some reason, a group of citing documents becomes unavailable in the Web.

In order to understand this phenomenon, we may observe the case of the most cited document in the sample (See Table I), which is Lowry's article: "Protein measurement with the Folin phenol reagent". This study suffered a severe drop in citations in the space of a few months. We observed the number of citations of this article at three different points in time: 28th May; 7th August; 21st October, 2014. As of the 28th of May, 2014 (first data sample), it was the most cited document in our sample, with 253,671 citations according to GS. However, on the 21st of October, its citation count had decreased to 192,841 (Table 5).

Table 5. Fluctuation of citations received by Lowry's article

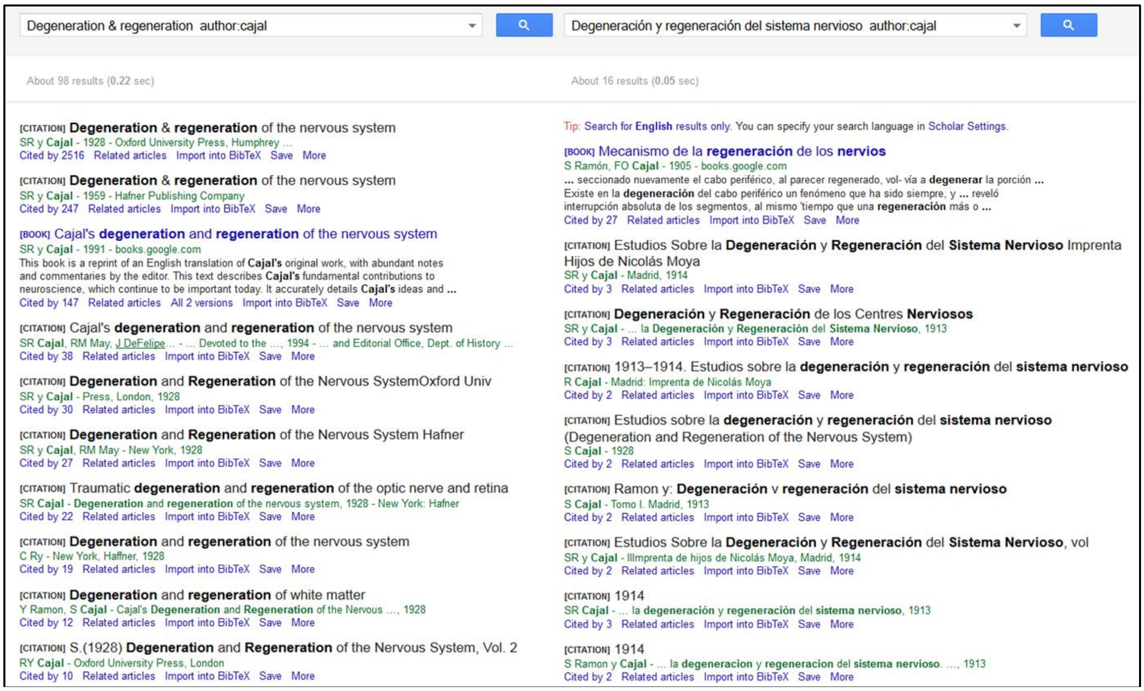
Date	WoS Citations	GS Citations	Screenshots
28th May, 2014	303,832	253,671	 [PDF] Protein measurement with the Folin phenol reagent OH Lowry, NJ Rosebrough, AL Farr... - J biol ..., 1951 - amirza.persiangig.com Method Reagents-Reagent A, 2 per cent N&OX in 0.10 N NaOH. Reagent B, 0.5 per cent CuS04. 5Hz0 in 1 per cent sodium or potassium tartrabe. Reagent C, alkaline copper solution. Mix 50 ml. of Reagent A with 1 ml. of Reagent B. Discard after 1 day. Reagent D, ... Cited by 253671 Related articles All 25 versions Web of Science: 303832 Cite Save More
7th August, 2014	304,667	191,669	 [PDF] **Protein measurement** with the **Folin phenol reagent** OH Lowry, NJ Rosebrough, AL Farr, RJ Randall - J biol Chem, 1951 - devbio.wustl.edu Method **Reagents**-**Reagent** A, 2 per cent N&OX in 0.10 N NaOH. **Reagent** B, 0.5 per cent CuS04. 5Hz0 in 1 per cent sodium or potassium tartrabe. **Reagent** C, alkaline copper solution. Mix 50 ml. of **Reagent** A with 1 ml. of **Reagent** B. Discard after 1 day. **Reagent** D, ... Citado por 191669 Artículos relacionados Las 29 versiones Citar Guardar Más
21st October, 2014	305,202	192,841	 [PDF] Protein measurement with the Folin phenol reagent OH Lowry, NJ Rosebrough, AL Farr, RJ Randall - J biol Chem, 1951 - devbio.wustl.edu Method Reagents-Reagent A, 2 per cent N&OX in 0.10 N NaOH. Reagent B, 0.5 per cent CuS04. 5Hz0 in 1 per cent sodium or potassium tartrabe. Reagent C, alkaline copper solution. Mix 50 ml. of Reagent A with 1 ml. of Reagent B. Discard after 1 day. Reagent D, ... Cited by 192841 Related articles All 29 versions Web of Science: 305202 Cite Save More

Within 5 months, Lowry's article lost approximately 60,000 citations. As a consequence, as of October, 2014, it was not the highest cited article in GS, giving way to Laemli's work, which had 223,264 citations. WoScc data seems to be much more stable, showing 303,832 citations in May and 305,202 in October. Conversely, "Diagnostic and statistical manual of mental disorders" (5th position), reported 129,473 citations in May whereas in October the count increased to 185,000 citations, that is, 55,170 more citations in just 5 months.

Presumably, this drastic change in citations took place as a consequence of a major "re-crawling" performed by Google in June 2014. In any case, we believe that this variability may affect specific positions in the ranking of Table I, but not the condition of the documents as highly-cited documents (especially in the top 1%). Of course, this phenomenon is likely to occur on highly cited items, as the number of their citations follows a skewed distribution. The impact of these errors could be large however for non-highly cited items (usual search results).

The accuracy of duplicate detection / merging versions

Google Scholar declares that they merge all versions of a same document (not only different editions or reprints published in different years but also translations to other languages), and that all their respective citations are then added (Verstak & Acharya, 2013). However, this task isn't always accomplished successfully. In Figure 9 we can see an example of two different editions (English and Spanish) for the seminal work "Degeneration and regeneration of the nervous system" by Santiago Ramón y Cajal, which haven't been merged. Even for editions in the same language, several variants can be found as well.



Degeneration & regeneration author:cajal

About 98 results (0.22 sec)

[CITATION] **Degeneration** & **regeneration** of the nervous system
SR y **Cajal** - 1928 - Oxford University Press, Humphrey ...
Cited by 2516 Related articles Import into BibTeX Save More

[CITATION] **Degeneration** & **regeneration** of the nervous system
SR y **Cajal** - 1959 - Hafner Publishing Company
Cited by 247 Related articles Import into BibTeX Save More

[BOOK] Cajal's **degeneration** and **regeneration** of the nervous system
SR y **Cajal** - 1991 - books.google.com
This book is a reprint of an English translation of **Cajal's** original work, with abundant notes and commentaries by the editor. This text describes **Cajal's** fundamental contributions to neuroscience, which continue to be important today. It accurately details **Cajal's** ideas and **...**
Cited by 147 Related articles All 2 versions Import into BibTeX Save More

[CITATION] Cajal's **degeneration** and **regeneration** of the nervous system
SR **Cajal**, RM May, J DeFelipe... - ... Devoted to the ..., 1994 - ... and Editorial Office, Dept. of History ...
Cited by 38 Related articles Import into BibTeX Save More

[CITATION] **Degeneration** and **Regeneration** of the Nervous SystemOxford Univ
SR y **Cajal** - Press, London, 1928
Cited by 30 Related articles Import into BibTeX Save More

[CITATION] **Degeneration** and **Regeneration** of the Nervous System Hafner
SR y **Cajal**, RM May - New York, 1928
Cited by 27 Related articles Import into BibTeX Save More

[CITATION] Traumatic **degeneration** and **regeneration** of the optic nerve and retina
SR **Cajal** - **Degeneration** and **regeneration** of the nervous system, 1928 - New York: Hafner
Cited by 22 Related articles Import into BibTeX Save More

[CITATION] **Degeneration** and **regeneration** of the nervous system
C Ry - New York, Haffner, 1928
Cited by 19 Related articles Import into BibTeX Save More

[CITATION] **Degeneration** and **regeneration** of white matter
Y Ramon, S **Cajal** - Cajal's **Degeneration** and **Regeneration** of the Nervous ..., 1928
Cited by 12 Related articles Import into BibTeX Save More

[CITATION] S.(1928) **Degeneration** and **Regeneration** of the Nervous System, Vol. 2
RY **Cajal** - Oxford University Press, London
Cited by 10 Related articles Import into BibTeX Save More

Degeneración y regeneración del sistema nervioso author:cajal

About 16 results (0.05 sec)

Tip: Search for **English** results only. You can specify your search language in Scholar Settings.

[BOOK] Mecanismo de la **regeneración** de los **nervios**
S Ramón, FO **Cajal** - 1905 - books.google.com
... seccionado nuevamente el cabo periférico, al parecer regenerado, vol- vía a **degenerar** la porción **...** Existe en la **degeneración** del cabo periférico un fenómeno que ha sido siempre, y **...** reveló interrupción absoluta de los segmentos, al mismo 'tiempo que una **regeneración** más o **...**
Cited by 27 Related articles Import into BibTeX Save More

[CITATION] Estudios Sobre la **Degeneración** y **Regeneración** del **Sistema Nervioso** Imprenta Hijos de Nicolás Moya
SR y **Cajal** - Madrid, 1914
Cited by 3 Related articles Import into BibTeX Save More

[CITATION] **Degeneración** y **Regeneración** de los Centres **Nerviosos**
SR y **Cajal** - ... la **Degeneración** y **Regeneración** del **Sistema Nervioso**, 1913
Cited by 3 Related articles Import into BibTeX Save More

[CITATION] 1913–1914. Estudios sobre la **degeneración** y **regeneración** del **sistema nervioso**
R **Cajal** - Madrid: Imprenta de Nicolás Moya
Cited by 2 Related articles Import into BibTeX Save More

[CITATION] Estudios sobre la **degeneración** y **regeneración** del **sistema nervioso** (Degeneration and Regeneration of the Nervous System)
S **Cajal** - 1928
Cited by 2 Related articles Import into BibTeX Save More

[CITATION] Ramon y: **Degeneración** v **regeneración** del **sistema nervioso**
S **Cajal** - Tomo I. Madrid, 1913
Cited by 2 Related articles Import into BibTeX Save More

[CITATION] Estudios Sobre la **Degeneración** y **Regeneración** del **Sistema Nervioso**, vol
SR y **Cajal** - Illmprenta de hijos de Nicolás Moya, Madrid, 1914
Cited by 2 Related articles Import into BibTeX Save More

[CITATION] 1914
SR **Cajal** - ... la **degeneración** y **regeneración** del **sistema nervioso**, 1913
Cited by 3 Related articles Import into BibTeX Save More

[CITATION] 1914
S Ramon y **Cajal** - ... la **degeneracion** y **regeneracion** del **sistema nervioso**. ..., 1913
Cited by 2 Related articles Import into BibTeX Save More

Figure 9. Example of language versions (English and Spanish) of "Degeneration and regeneration of the nervous system by Cajal in Google Scholar

This simple test suggests that book impact, measured through citations from Google Scholar, would likely be even higher if all versions were successfully merged. This would probably mean that even more books would appear in Table I.

To understand the extent of the issue of citations to a given work which are dispersed among several duplicate records, we carried out a systematic and exhaustive analysis of one book as a case study: "The Mathematical Theory of Communication", by Shannon and Weaver. This work, because of its bibliographic complexity, illustrates the challenges that the correct treatment of highly-cited documents would pose (See supplementary material).[1]

"A mathematical theory of communication" was first published by Shannon as a two-part article in 1948. This work was later expanded and reedited in book form in 1949. This new edition was co-authored by Shannon and Weaver, with a slightly different title: "(The) mathematical theory of communication". Therefore, technically there are two distinct citable items which, ideally, Google Scholar should have been able to tell apart at the moment they were indexed.

In order to learn how GS actually handled this work, we searched it with the query <"mathematical theory of communication"> and selected the result with the greater number of detected versions (830), which we will call the "main record". We downloaded the bibliographic information of all the versions GS found for the main record, which weren't actually 830, but only 763 (discrepancies between hit counts and the actual visible results are a well-known phenomenon in GS).

After this, we refined this query (adding the search command <author:Shannon>) obtaining 229 additional results. Of them, only 164 (71.6%) were actually different versions of the work. The rest were comments and reviews. These 164 records are duplicates that Google Scholar should also have merged with the main record (added to those 763 versions), but didn't.

If we consider the 165 verified records (the main record and the 164 duplicates), the main record held the larger number citations (69,738), whereas the remaining 164 duplicates together accounted for 3,714 new potential citations (not considering possible duplicates or false citations).

This analysis (search, download, and manual check) was carried out in October 2014. A complete description is provided in the supplementary material.[1]

There is a low Pearson's correlation between the number of citations and the number of versions (r= 0.2; n= 64000). This value is similar to that obtained by Jamali and Nabavi (2015), who found a weak positive correlation between the number of versions and the citation counts for full-text articles (r = 0.346; n = 4426). Pitol and De Groote (2014) found low values as well (r= 0.257; n= 982) when describing the GS versions for articles stored in institutional repositories from three US universities.

However, we found that this correlation increases when the Spearman method is used instead (r= 0.48; n= 64000), probably revealing a threshold beyond which it is unusual to find documents with a high number of versions and low citation counts. This result may also indicate that the number of missing citations (from undetected duplicates) will only be significant for highly-cited documents with a high number of versions, which in any case constitute a small portion of the records (they are mainly books). Therefore, there shouldn't be many highly-cited documents that haven't made it to our sample because of Google Scholar's duplicate detection errors.

Composition of the sample

Publication date

In Table I (highly-cited documents) we can see that the eleventh position is held by a book published outside the timeframe selected in our study (1948). This book, however, appeared in the results of the different queries we performed. Additionally, in Figure 3 we detected an uncommon increment of the presence of books in the results GS displayed for the most recent years. These issues led us to question the information about the publication date that Google Scholar provides for books.

We realized that Google Scholar lumps together all the different editions of the same book, and usually (not always) selects the latest edition as the primary version, taking the date of this version as the publication date of the book. This is the reason behind the fact that the seminal work "A mathematical theory of communication" published by Shannon in 1948 is included in the sample: Google Scholar has selected a reprint published in 2001 as the primary version.

Since Google Scholar only presents 1,000 results for any given query (and we only collected information about the primary versions of the documents), new editions of old books took the place of other publications that had really been published in those years.

The differences between the date of the first edition and the publication date used by Google Scholar for each book is shown in Figure 10 for the top 600 most cited books, where a bias in the last 10 years is evident.

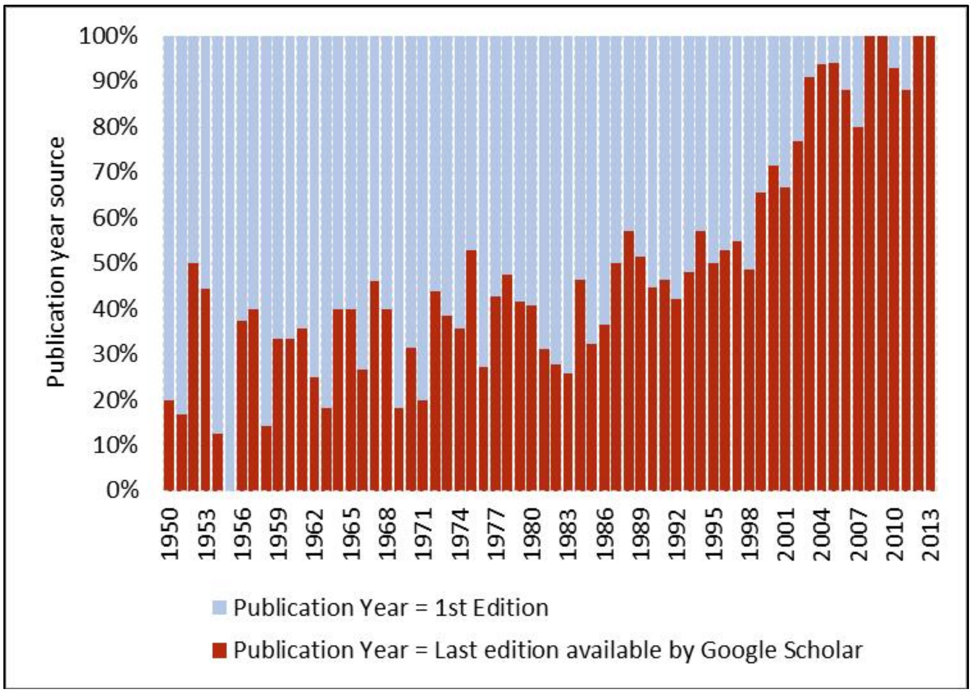


Figure 10. Number of books according to the year of publication signed by Google Scholar and to the date of the first edition (top 600)

The decision to select the publication date of the most recent edition of a book as the date of publication of the primary version makes a lot of sense from the point of view of a search engine (users will probably want to access the latest edition of a book), but it becomes a problem when the goal is to perform any kind of bibliometric analysis. This issue obviously affects our sample (it is especially noticeable in figure 2 and 3). In any case, it should be noted that this limitation doesn't affect the status of these books as highly-cited documents; only the year of publication is affected, resulting in an overrepresentation of books in the last decade, which are unfairly taking the place of other highly-cited documents that were actually published in those years.

Language of the documents

We developed a strategy to determine this information using WoScc data where possible (around 50% of the sample) as well as the title and abstract of the document in all the other cases. This approach, however, may have resulted in an overrepresentation of the English language, since it is usual for a document written in a language other than English to provide its title and abstract in English as well, for the purpose of being indexed in international databases.

Additionally, the sample may contain records that are in fact translations of other documents (which may also be present in our sample). This is the case of journals that are published both in English and in other language or books that are translated into various languages.

For this reason, the English percentage of highly-cited documents should be taken with caution and be considered only as an estimate.

Properties of the sample

The bibliographic data collected for each document (full-text availability, document type, source provider…) always comes from the version of the document Google Scholar considered as the "primary version" (the one that is displayed in the page of results of a query). This fact constitutes a limitation since one document may be freely accessible through various source providers (for example a journal and a repository) and file formats (for example html and pdf file format). For this reason, all the results obtained, especially those included in the sections 3.4, 3.5 and 3.6 must be interpreted with this limitation in mind. Additionally, it should be reminded that all the queries were performed without activating the academic Library subscriptions feature, which would have introduced a bias in the information about full-text source providers.

Document type

The great variety of document types included in Google Scholar, as well as the impossibility of filtering by this variable (Bornmann et al., 2009; Aguillo, 2012) makes document type statistics quite difficult. For this reason, three complementary methods were used in this paper to detect the typology of the 64,000 documents in the sample.

We could only determine the document types of 71% of the entire dataset. A manual inspection would have been required to ascertain the typology of the remaining 29% (18,589 documents). We believe the proportion of books and book chapters would have increased if the entire sample had been successfully categorized, since this is the typology that Google Scholar has more trouble identifying.

Free Full-text

Since the existence of a full-text link does not guarantee the disposal of the full-text (some links actually refer to publisher's abstracts), the results (40% of the documents had a free full-text link) might be somewhat overestimated. In any case, these values are consistent with those published by Archambault et al. (2013), who found that over 40% of the articles from their sample were freely accessible; higher than those by Khabsa and Giles (2014) and Björk et al. (2010), who found only a 24% and 20.4% of open access documents respectively; and much lower than Jamali and Nabavi (2015) and Pitol and De Groote (2014), who found 61.1% and 70% respectively.

The different nature of the samples makes it difficult to draw comparisons among these studies. Nonetheless, the sample used in this study (64,000 documents) is the largest ever used to date.

File format

The predominance of the pdf and the html file formats matches the results thrown by previous studies. Among others, those by Orduna-Malea et al. (2010), Aguillo et al. (2010), and Jamali and Nabavi (2015).

Source providers

The source providers for freely accessible highly-cited documents in Google Scholar are, at least as far as our sample is concerned, institutional (US universities) and subject (Pubmed central and Arxiv) repositories. Despite the fact that some commercial publishers also appear on the top positions of the ranking of source providers, their presence in absolute numbers is small. Of special note is the role of the scientific social network ResearchGate. Its presence, already noted by Jamali and Nabavi (2015), shows that a) ResearchGate contains an already large (and still growing) percentage of highly-cited documents; and b) its capacity to become the primary version of the highly-cited documents in Google Scholar.

These results differ from those obtained by Ortega (2014) who detected a high presence of publishers (constituting the source for 58.4% of all scientific documents in Google Scholar). The reason behind this difference is that Ortega used <site:> queries directly to find the number of documents hosted within the source providers' websites. The different way in which we conducted our queries makes a direct comparison impossible, but it confirms that even though most publishers now allow Google Scholar to crawl their websites, they are not becoming the main destination for users to access the full-text of highly-cited documents.

Regarding the web domains, Aguillo (2012) detected countries which intensely contribute to increase the size of Google Scholar (such as France, Japan, Brazil or China). However, these countries do not appear as the main contributors of highly-cited documents (Germany is the first country in this ranking). The comparison of a general ranking of source providers and the source providers of the highly-cited documents might serve to identify the places where these top contributions actually become freely available to final users on the Web.

5. Conclusions

In light of the results obtained, we can conclude that Google Scholar offers an original and different vision of the most influential documents in the academic/scientific environment (measured from the perspective of their citation count). These results are a faithful reflection of the all-encompassing indexing policies that enable Google Scholar to retrieve a larger and more diverse number of citations, since they come from a wider range of document types, different geographical environments, and languages.

Therefore, Google Scholar covers not only seminal research works in the entire spectrum of the scientific fields, but also the greatly influential works that scientists, teachers and professionals who are training to become practitioners use in their respective fields. This phenomenon is particularly true for works that deal with new data collecting and processing techniques.

This is reflected on the high proportion of books among the highly cited documents (62% of the top 1% most cited documents collected), as this document type is essential in the humanities and the social sciences (also as a vehicle for the communication of new results), and in the experimental sciences (as a way to consolidate and disseminate scientific knowledge).

There are still important limitations and errors when working with data extracted from Google Scholar, especially those related to the detection of duplicate documents, and the correct allocation of citations. These issues have all been discussed in-depth in this study. While these mistakes may introduce biases in the ranking of most-cited documents in Google Scholar (the specific position of a document in this list), our empirical data suggest that the influence of these errors on the characterization and description of the sample, which is the main goal of this study, would be minimal.

In conclusion, thanks to the wide and diverse list of sources from which Google Scholar feeds, this search engine covers academic documents in a broader sense, enabling the measurement of impact stemming not only from the scientific side of the academic landscape, but also from the educational side (doctoral dissertations, handbooks) and from the professional side (working papers, technical reports, patents), the last two being areas that haven't been explored as much as the first one.

Other specific findings of this study are summarized below:

- 40% of the highly cited documents in Google Scholar are freely accessible, mostly from educational institutions (mainly universities), and other non-profit organizations.
- Google Scholar has detected more than one version for 83.17% of the documents in our sample.
- The general correlation between the number of versions and the number citations they have received is low (r= 0.2) except for documents with a very high number of versions (more than 100), which also present a high number of citations.
- The average highly-cited document is a journal article (72.3% of the documents for which a document type could be ascertained) or a book (62% of the top 1% most cited documents of the sample), written in English (92.5% of all documents) and available online in PDF format (86.0% of all documents)

Acknowledgements

Alberto Martín-Martín enjoys a four-year doctoral fellowship (FPU2013/05863) granted by the Spanish Ministry of Education, Culture and Sports. Juan Manuel Ayllón enjoys a four-year doctoral fellowship (BES-2012-054980) granted by the Spanish Ministry of Economy and Competitiveness. Enrique Orduna-Malea holds a postdoctoral fellowship (PAID-10-14), from the Polytechnic University of Valencia (Spain).

Endnotes

1. Supplementary material. Available at https://dx.doi.org/10.6084/m9.figshare.1224314.v1. Accessed 25 March 2016.

References

Aguillo, Isidro F.; Ortega, J.; Fernández, M.; Utrilla, A. (2010). Indicators for a webometric ranking of open access repositories. *Scientometrics*, vol. 82 (3), 477-486. http://dx.doi.org/10.1007/s11192-010-0183-y

Aguillo, Isidro F. (2012). Is Google Scholar useful for bibliometrics? A webometric analysis. *Scientometrics*, vol. 91 (2), 343-351. http://dx.doi.org/10.1007/s11192-011-0582-8

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, vol. 12 (3), 159-170. http://dx.doi.org/10.3152/147154403781776645

Aksnes, D. W.; Sivertsen, G. (2004). The effect of highly cited papers on national citation indicators. *Scientometrics*, vol. 59 (2), 213-224. http://dx.doi.org/10.1023/b:scie.0000018529.58334.eb

Archambault, E.; Amyot, D.; Deschamps, P.; Nicol, A.; Rebout, L.; Roberge, G. (2013). Proportion of open access peer-reviewed papers at the European and world levels—2004–2011. Science-Metrix. Report. Science Matrix Inc. Disponible en: http://www.science-metrix.com/pdf/SM_EC_OA_Availability_2004-2011.pdf

Bar-Ilan, J. (2010). Citations to the “Introduction to informetrics” indexed by WOS, Scopus and Google Scholar. *Scientometrics*, vol. 82(3), 495-506. http://dx.doi.org/10.1007/s11192-010-0185-9

Beel, J.; Gipp, B.; Wilde, E. (2010). Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co. *Journal of Scholarly Publishing*, vol. 41 (2), 176-190. http://dx.doi.org/10.1353/scp.0.0082

Björk, B. C.; Welling, P.; Laakso, M.; Majlender, P.; Hedlund, T.; Gudnason, G. (2010). Open Access to the scientific journal literature: Situation 2009. *PLoS ONE*, vol. 5(6), e11273. http://dx.doi.org/10.1371/journal.pone.0011273

Bornmann, L. (2010). Towards an ideal method of measuring research performance: Some comments to the Opthof and Leydesdorff (2010) paper. *Journal of Informetrics*, vol. 4 (3), 441–443. http://dx.doi.org/10.1016/j.joi.2010.04.004

Bornmann, L.; Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: the avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics*, vol. 5 (1), 228-230. http://dx.doi.org/10.1016/j.joi.2010.10.009

Bornmann, L.; Marx, W.; Schier, H.; Rahm, E.; Thor, A.; Daniel, H. D. (2009). Convergent validity of bibliometric Google Scholar data in the field of chemistry—Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstracts. *Journal of Informetrics*, vol. 3 (1), 27-35. http://dx.doi.org/10.1016/j.joi.2008.11.001

Bornmann, L.; Moya-Anegón, F.; Leydesdorff, L. (2011). The new excellence indicator in the World Report of the SCImago Institutions Rankings 2011. *Journal of Informetrics*, vol. 6(2), 333-335. http://dx.doi.org/10.1016/j.joi.2011.11.006

Garfield, E. (1977). Introducing Citation Classics: the human side of scientific papers. *Current contents*, vol. 3 (1), 1-2.

Garfield, E. (2005). The Agony and the Ecstasy—The History and Meaning of the Journal Impact Factor. *International Congress on Peer Review and Biomedical Publication*. Chicago, 16 September. Disponible en: http://www.garfield.library.upenn.edu/papers/jifchicago2005.pdf

Glänzel, W.; Czerwon, H. J. (1992). What are highly cited publications? A method applied to German scientific papers, 1980–1989. *Research Evaluation*, vol. 2 (3), 135-141. http://dx.doi.org/10.1093/rev/2.3.135

Glänzel, W.; Schubert, A. (1992). Some facts and figures on highly cited papers in the sciences, 1981–1985. *Scientometrics*, vol. 25 (3), 373-380. http://dx.doi.org/10.1007/bf02016926

Glänzel, W.; Rinia, E. J.; Brocken, M. G. (1995). A bibliometric study of highly cited European physics papers in the 80s. *Research Evaluation*, vol. 5 (2), 113-122. http://dx.doi.org/10.1093/rev/5.2.113

Harzing, A. W. (2013). A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. *Scientometrics*, vol. 94 (3), 1057-1075. http://dx.doi.org/10.1007/s11192-012-0777-7

Harzing, A. W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics*, vol. 98 (1), 565-575. http://dx.doi.org/10.1007/s11192-013-0975-y

Harzing, A.W.; Van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, vol. 8 (1), 61-73. http://dx.doi.org/10.3354/esep00076

Jacso, P. (2005). Google Scholar: the pros and the cons. *Online information review*, vol. 29 (2), 208-214. http://dx.doi.org/10.1108/14684520510598066

Jacso, P. (2006). Deflated, inflated, and phantom citation counts. *Online Information Review*, vol. 30 (3), 297-309. http://dx.doi.org/10.1108/14684520610675816

Jacsó, P. (2008a). The pros and cons of computing the h-index using Scopus. *Online Information Review*, vol. 32 (4), 524-535. http://dx.doi.org/10.1108/14684520810897403

Jacso, P. (2008b). The pros and cons of computing the h-index using Google Scholar. *Online Information Review*, vol. 32 (3), 437-452. http://dx.doi.org/10.1108/14684520810889718

Jacso, P. (2012). Using Google Scholar for journal impact factors and the h-index in nationwide publishing assessments in academia – siren songs and air-raid sirens. *Online Information Review*, vol. 36 (3), 462-478. http://dx.doi.org/10.1108/14684521211241503

Jamali, H. R. ; Nabavi, M. (2015). Open access and sources of full-text articles in Google Scholar in different subject fields. *Scientometrics*, vol. 105 (3), 1635-1651. http://dx.doi.org/10.1007/s11192-015-1642-2

Khabsa, M.; Giles, C. L. (2014). The number of scholarly documents on the public web. *PLoS One*, vol. 9(5), e93949. http://dx.doi.org/10.1371/journal.pone.0093949

Kousha, K.; Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, vol. 74 (2), 273–294. http://dx.doi.org/10.1007/s11192-008-0217-x

Kousha, K.; Thelwall, M.; Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science*, vol. 62 (11), 2147-2164. http://dx.doi.org/10.1002/asi.21608

Kresge, N.; Simoni, R. D.; Hill, R. L. (2005). The most highly cited paper in publishing history: Protein determination by Oliver H. Lowry. *Journal of Biological Chemistry*, vol. 280 (28), e25. Disponible en: http://www.jbc.org/content/280/28/e25

Levitt, J. M.; Thelwall, M. (2009). The most highly cited Library and Information Science articles: Interdisciplinarity, first authors and citation patterns. *Scientometrics*, vol. 78 (1), 45-67. http://dx.doi.org/10.1007/s11192-007-1927-1

Maltrás Barba, B. (2003). *Los indicadores bibliométricos: fundamentos y aplicación al análisis de la ciencia*. Gijón: Trea.

Martín-Martín, A.; Ayllón, J. M.; Delgado López-Cózar, E.; Orduna-Malea, E. (2015). Nature's top 100 Re-revisited. *Journal of the Association for Information Science & Technology*, vol. 66 (12), 2714-2714. http://dx.doi.org/10.1002/asi.23570

Meho, L.; Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, vol. 58 (13), 2105–2125. http://dx.doi.org/10.1002/asi.20677

Narin, F.; Frame, J. D.; Carpenter, M. P. (1983). Highly cited Soviet papers: An exploratory investigation. *Social Studies of Science*, vol. 13 (2), 307-319. http://dx.doi.org/10.1177/030631283013002006

Oppenheim, C.; Renn, S. P. (1978). Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, vol. 29 (5), 225-231. http://dx.doi.org/10.1002/asi.4630290504

Orduna-Malea, E.; Delgado López-Cózar, E. (2014). Google Scholar Metrics evolution: an analysis according to languages. *Scientometrics*, vol. 98 (3), 2353–2367. http://dx.doi.org/10.1007/s11192-013-1164-8

Orduna-Malea, E.; Ayllón, J. M.; Martín-Martín, A.; Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, vol. 104 (3), 931-949. http://dx.doi.org/10.1007/s11192-015-1614-6

Orduna-Malea, E.; Serrano-Cobos, J.; Ontalba-Ruipérez, J. A.; Lloret-Romero, N. (2010). Presencia y visibilidad web de las universidades públicas españolas. *Revista española de documentación científica*, vol. 33 (2), 246-278. http://dx.doi.org/10.3989/redc.2010.2.740

Ortega, Jose L. (2014). *Academic Search Engines: A Quantitative Outlook*. London: Elsevier.

Persson, O. (2010). Are highly cited papers more international?. *Scientometrics*, vol. 83 (2), 397-401. http://dx.doi.org/10.1007/s11192-009-0007-0

Pitol, S. P. ; De Groote, S. L. (2014). Google Scholar versions: Do more versions of an article mean greater impact? *Library Hi Tech*, vol. 32 (4), 594–611. http://dx.doi.org/10.1108/lht-05-2014-0039

Plomp, R. (1990). The significance of the number of highly cited papers as an indicator of scientific prolificacy. *Scientometrics*, vol. 19 (3), 185-197. http://dx.doi.org/10.1007/bf02095346

Smith, D. R. (2009). Highly cited articles in environmental and occupational health, 1919–1960. *Archives of environmental & occupational health*, vol. 64 (1), 32-42. http://dx.doi.org/10.1080/19338240903286743

Tijssen, R. J.; Visser, M. S.; Van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: are highly cited research papers an appropriate frame of reference? *Scientometrics*, vol. 54 (3), 381-397.

Van Noorden, R.; Maher, B.; Nuzzo, R. (2014). The top hundred papers. *Nature*, vol. 514 (7524), 550-553. http://dx.doi.org/10.1038/514550a

Verstak, A.; Acharya, A. (2013). *Identifying multiple versions of documents*. U.S. Patent No. 8,589,784. Washington, DC: U.S. Patent and Trademark Office.

Winter, J.C.F.; Zadpoor, A.; Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics*, vol. 98 (2), 1547–1565. http://dx.doi.org/10.1007/s11192-013-1089-2

Yang, K.; Meho, L. (2006). Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology*, vol. 43 (1), 1–15. http://dx.doi.org/10.1002/meet.14504301185

Chapter 4. Nature’s Top 100 Re-Revisited

Letter to the editor. Cite as:

Martín-Martín, A., Ayllón, J. M., Delgado López-Cózar, E., & Orduna-Malea, E. (2015). Nature ’s top 100 Re-revisited. Journal of the Association for Information Science and Technology, 66(12), 2714–2714. https://doi.org/10.1002/asi.23570

Letter

Dear Sir,

In a recent letter published in the *Journal of the Association for Information Science and Technology*, Bornmann (2015) criticizes *Nature*’s top 100 ranking. Van Noorden, Maher, & Nuzzo (2014) requested this list of the most-cited research of all time from Web of Science to mark the 50th anniversary of the *Science Citation Index*. Bornmann expresses concern about the methods used to generate the list (based on raw citation counts and not-normalized bibliometric indicators).

The *Nature* article also provides an alternative list of most-cited research contributions according to Google Scholar (available in the online version of the article as complementary material). Although we acknowledge that the main focus of the article are the data extracted from Web of Science, we believe it necessary to point out some discrepancies in the Google Scholar list.

For example, it lists “Protein Measurement with the folin phenol reagent” as the second most-cited article (192,710 citations), contradicting the Web of Science ranking, which shows it to be the most cited by far (Garfield, 2005), a fact that merits a thorough discussion.

Apart from this issue (anectodal, perhaps, but worth noting), there are certain inconsistencies that do not seem to have been considered in the Google Scholar list published by *Nature*. We discovered these inconsistencies when researching highly-cited documents in Google Scholar (Martín-Martín, Orduña-Malea, Ayllón, & Delgado-López-Cózar, 2014), and they relate, in particular, to the allocation of citations and the identification and linkage of different versions of the same documents.

According to our empirical data, the aforementioned article on protein measurement had attracted, as of May 2014, a total of 253,671 citations, whereas *Nature’s* ranking (extracted from Google Scholar on October 17, 2014) records only 192,710. How can an article lose 60,961 citations in 5 months? Conversely, the *Diagnostic and Statistical Manual of Mental Disorders*, not included in the top 10 despite having 185,000 citations in Google Scholar (as of October 2014), and almost 220,000 if we merge its various versions, would seem to have attracted a remarkable 55,170 citations from May to October.

Moreover, two different editions of “Molecular Cloning” appear on the list. Adding up the two versions (and other unmerged records), the citations amount to 268,834, which would promote this work to first place in the ranking. Likewise, we found 164 additional unmerged records for “A Mathematical Theory of Communication”, where citations to the article and the subsequent book are mixed.

To what extent, therefore, can we trust this Google Scholar list?

Apart from these errors in the preparation of the list, mainly because of a lack of professional filtering (necessary if we wish to compare citations on Google Scholar with Web of Science), we wish to note two important findings: (a) Even with dirty (unfiltered), Google Scholar is capable of identifying the most-cited papers, and (b) Google Scholar is capable of providing a complementary academic landscape (including books, heavily cited in certain fields, and traditionally discriminated against).

And this is what should be borne in mind, regardless of positions or exact figures. Let us not fall into the classic trap of not seeing the wood for the trees.

References

Bornmann, L. (2015). Nature's top 100 revisited. *Journal of the Association for Information Science and Technology*, *66*(10), 2166–2166. https://doi.org/10.1002/asi.23554

Garfield, E. (2005). The Agony and the Ecstasy - The History and Meaning of the Journal Impact Factor. In *International Congress on Peer Review And Biomedical Publication*. Retrieved from http://www.garfield.library.upenn.edu/papers/jifchicago2005.pdf?wa=IPEMBI14

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2014). *Does Google Scholar contain all highly cited documents (1950-2013)?* (EC3 Working Papers No. 19). Retrieved from http://arxiv.org/abs/1410.8464

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature*, *514*(7524), 550–553. https://doi.org/10.1038/514550a

Chapter 5. Can we use Google Scholar to identify highly-cited documents?

Journal article. Cite as:

Martin-Martin, A., Orduna-Malea, E., Harzing, A.-W., & Delgado López-Cózar, E. (2017). Can we use Google Scholar to identify highly-cited documents? Journal of Informetrics, 11(1), 152–163. https://doi.org/10.1016/j.joi.2016.11.008

Abstract (English)

The main objective of this paper is to empirically test whether the identification of highly-cited documents through Google Scholar is feasible and reliable. To this end, we carried out a longitudinal analysis (1950 to 2013), running a generic query (filtered only by year of publication) to minimise the effects of academic search engine optimisation. This gave us a final sample of 64,000 documents (1,000 per year). The strong correlation between a document's citations and its position in the search results (r= -0.67) led us to conclude that Google Scholar is able to identify highly-cited papers effectively. This, combined with Google Scholar's unique coverage (no restrictions on document type and source), makes the academic search engine an invaluable tool for bibliometric research related to the identification of the most influential scientific documents. We find evidence, however, that Google Scholar ranks those documents whose language (or geographical web domain) matches with the user's interface language higher than could be expected based on citations. Nonetheless, this language effect and other factors related to the Google Scholar's operation, i.e. the proper identification of versions and the date of publication, only have an incidental impact. They do not compromise the ability of Google Scholar to identify the highly-cited papers.

Abstract (Spanish)

El objetivo principal de este artículo es comprobar empíricamente si Google Scholar es una herramienta que puede identificar documentos altamente citados de manera fácil y fiable. Para comprobar esto, llevamos a cabo un análisis de documentos publicados entre 1950 y 2013, recuperados tras realizar una serie de búsquedas en la que solo se utilizaba el campo del año de publicación, para minimizar los efectos del academic search engine optimization. De esta manera obtuvimos una muestra de 64.000 documentos (1.000 documentos por año). La alta correlación entre las citas recibidas por un documento y su posición en los resultados de búsqueda (r= -0.67) nos lleva a concluir que Google Scholar es capaz de identificar documentos altamente citados de manera efectiva. Esto, combinado con la cobertura única de Google Scholar (no tiene restricciones de tipos documentales o de fuente) convierte a este buscador académico en una herramienta de gran valor para la investigación bibliométrica en lo que respecta a la identificación de los documentos científicos más influyentes. También se ha encontrado evidencia de que Google Scholar puede posicionar los documentos que coinciden con el lenguaje de preferencia del usuario en un puesto más alto de lo que les correspondería por su número de citas. Sin embargo, el efecto de la lengua y otros factores utilizados por Google Scholar para determinar la posición de los documentos en los resultados de búsqueda no parecen comprometer la capacidad del buscador para identificar documentos altamente citados.

1. Introduction

Google Scholar is a free academic web search engine that indexes scholarly literature across a wide array of disciplines, document types and languages (Ortega, 2014). It therefore specialises in finding and identifying bibliographic scholarly material, as well as providing a number of value-added services, such as direct access to the full texts (although for legal reasons this is not possible for all documents), the number of citations received by each document, and the number of different versions of the document. Google Scholar first appeared in November 2004, coinciding almost exactly with the launch of Scopus (*Reed Elsevier*, 2004). This meant that both products entered onto the market for bibliometric databases at the same time, a market in which up till then the Web of Science (WoS) had held a monopoly. The products had diametrically opposite approaches, however. Whereas Google Scholar was conceived as an open, dynamic but uncontrolled and fully automated product (Jacsó, 2005), Scopus positioned itself as a controlled product: with human intervention, closed, more static and designed to compete directly with WoS (Jacsó, 2008c).

Web of Science and Scopus began a rivalry as databases geared to the world of academic impact assessment. In the meantime, Google Scholar operated in another, complementary, sector: searching, locating and accessing academic information, in the broad sense of the term. Before the latter had completed its first year of operation, both *Science* (Leslie, 2004) and *Nature* (Butler, 2004) had already made mention of its impact on the scientific community.

Studies of Google Scholar as a scholarly information search tool have been undertaken primarily by the library sector. Three different phases can be discerned in these studies. At first, Google Scholar was observed with curiosity and scepticism. This phase was followed by a period of systematic study when it received harsh criticism. Finally, the third phase was one of optimism about its potential to reach 100% of the information available online for an institution, person, journal or other scholarly communication channel (Howland et al, 2009).

Among these studies, we can identify, on the one hand, those which look to understand how the quality and the usefulness of the product is perceived by different types of users, such as students (Carpenter, 2012), academics (Schonfeld & Housewright, 2010; Housewright et al, 2013; Van Noorden, 2014) and information professionals (Giles, 2005; Giustini & Barsky, 2005; Ettinger, 2008). On the other hand, there are the studies comparing the performance of Google Scholar with other information search tools, such as catalogues, bibliographic databases, and discovery tools (Callicott & Vaughn, 2005; Gardner & Eng, 2005; Giustini, 2005; Levine-Clark & Kraus, 2007; Meier & Conkling, 2008; Bramer et al, 2013; Gehanno, Rollin & Darmoni, 2013; Stirbu et al, 2015; Breeding, 2015).

Beyond the analysis of Google Scholar as a search tool, its continued growth and the provision of citation counts that are not biased against the original source, language and format of the citing document, have led to a growing interest from the bibliometric and webometric community in studying this product as a tool for the evaluation of research activity (Torres-Salinas, Ruiz-Pérez & Delgado López-Cózar, 2009; Aguillo, 2011). This research has focused primarily on assessing the quality of the bibliographic and bibliometric data provided (Jacsó 2005; 2006; Bar-Ilan, 2010; Jacsó 2008a; 2008b; 2012) and its correlation with the indicators developed by Web of Science and Scopus (Bakkalbasi et al, 2006; Bar-Ilan, 2007; Bar-Ilan, Levene & Lin, 2007; Cabezas-Clavijo & Delgado López-Cózar, 2013; Harzing & Alakangas, 2016). Studies generally found significant correlations between the various data sources; however, not all studies found equally high correlations. Moreover, the fact that Google Scholar covers a far broader range of documents and that users cannot discriminate between the types of documents, compromised some of these comparative studies (Bornmann et al, 2009; Aguillo, 2012).

Other areas of research have focused on the usefulness of Google Scholar for obtaining unique citations (Yang & Meho, 2006; Meho & Yang 2007; Kousha & Thelwall, 2008; Bar-Ilan, 2010; Kousha, Thelwall & Rezaie, 2011; Harzing 2013; Orduña-Malea & Delgado López-Cózar, 2014) and, finally, on studying its coverage and its evolution over time (Aguillo, 2012; Khabsa & Giles, 2014; Harzing, 2014; Ortega, 2014; Winter, Zadpoor & Dodou, 2014; Orduna-Malea et al, 2015).

To date, empirical results have shown Google Scholar to be enormously useful when obtaining statistics on academic impact, especially for disciplines that use alternative channels of scholarly communication (in particular doctoral theses, books, book chapters, and conference proceedings), such as the Social Sciences, Humanities, and Engineering (Meho & Yang, 2007; Harzing & Van der Wal, 2008; Bar-Ilan, 2010; Kousha, Thelwall & Rezaie, 2011; Kousha & Thelwall, 2015, Martin-Martin et al, 2015). These are all disciplines in which the use of Google Scholar is deemed necessary to provide comprehensive information on academic impact. However, the literature also indicates that errors in the linking of citations and versions, and in the quality of the bibliographic data still persist, though to a lesser extent than in the early years (Winter, Zadpoor & Dodou, 2014; Orduna-Malea et al, 2015). This precludes Google Scholar's use as a standalone tool for scholarly assessment without prior filtering and processing of the data, an activity limited by the lack of options for the automated downloading of files.

Our review thus shows a significant accumulated knowledge base about Google Scholar as a search tool and a tool for evaluation of research activity, more specifically relating to its unique coverage when compared to other sources of publication and citation data. However, to the best of our knowledge there is no prior research on the capabilities of Google Scholar to identify highly-cited documents. In the context of Bibliometrics, highly-cited documents represent the most influential scientific works. Therefore, the identification of these documents allows detecting the most influential authors, topics, research methods, and sources of all times, and is thus a very important function of bibliometric research.

The identification of highly-cited documents in the Web of Science Core Collection (WoScc) or Scopus (the leading bibliographic databases for this purpose) is now quite a straightforward task. Both databases include, among the search sorting criteria (subject matter, date, author or journal), the number of times a paper is cited ("times cited" or "highest to lowest"). Therefore, it is simply a matter of selecting this option and the documents will be presented in descending order according to the number of citations received. Since there is no restriction in these databases on the number of documents that can be retrieved for a given query, identifying highly-cited papers is totally reliable. This enabled bibliometric studies of these documents to be conducted with ease. Conversely, the lack of a similar sorting feature in Google Scholar, together with the limitation of a maximum of 1000 results, i.e. document metadata, shown per query, raises the question of whether or not the identification of highly-cited papers is possible using this search engine.

Given the negative impact on the visibility of a document (and its authors) of not featuring in the top 1000 Google Scholar results for a specific query, search engine optimisation (SEO) is gaining popularity. This is an approach, already well-established in commercial environments (Ledford, 2009), whereby knowledge of the approximate sorting criteria of Google (a trade secret) has led experts to disentangle the key factors that influence the positioning of a website in the search results (Evans, 2007), one of which is the number of links that a website receives, a key indicator for webometrics (Orduña-Malea & Aguillo, 2014).

The application of these techniques to the academic environment (especially Google Scholar) has led to a new concept called Academic Search Engine Optimisation (ASEO). Beel, Gipp and Wilde (2010) define it as the creation, publication and modification of scholarly literature so as to facilitate crawling and indexing for the search engines, improving its subsequent position in the ranking. Although the number of citations seems to be one of the key indicators in this ranking process in Google Scholar (Beel & Gipp, 2009; 2010), other factors might positively or negatively influence the final rank that is achieved when a specific search is performed. We can distinguish between query factors (related to the nature of keywords), and document factors (language of articles, length of the articles, what words are used – or not used – in the title, abstract and keywords, or platforms in which documents are uploaded). In addition, the dynamism of the Web as well as the malfunction of some GS features (such as improperly linked versions) might affect the final rank as well.

These procedures (which may be artificially aimed at optimising the position of documents on the list of results to specific queries by authors) can be either a reflection of successful marketing and dissemination activities or, in contrast, the result of illicit activities designed to trick the search engines by manipulating certain data (Beel, Gipp & Wilde, 2010; Delgado López-Cózar, Robinson-García & Torres-Salinas, 2014). Existing ASEO procedures as well as the idiosyncratic way in which Google Scholar ranks results (which is kept under trade secret) mean that there is no *a priori* guarantee that users are able to retrieve all highly-cited documents. If this were the case, Google Scholar's subsequent usefulness as a tool for bibliometric evaluation would be severely limited.

Therefore, this paper has two main objectives:

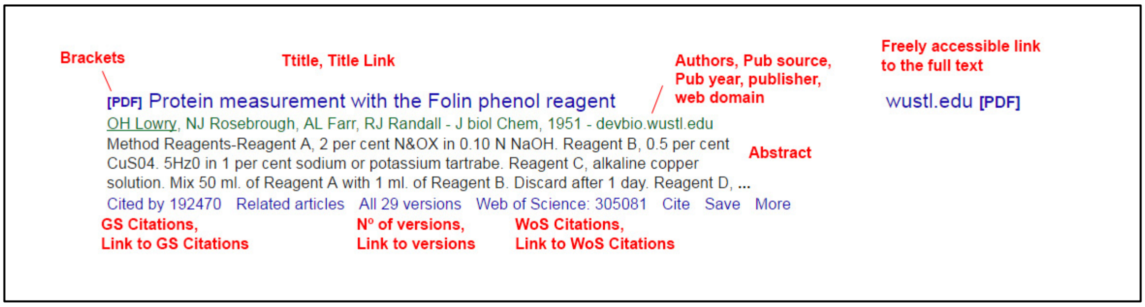
- Verify whether it is possible to reliable identify the most highly-cited papers in Google Scholar, and indirectly,
- Empirically validate whether citations are the primary result-ordering criterion in Google Scholar for generic queries or whether other factors substantially influence the rank order.

2. Methods

To accomplish the objectives formulated above, we propose first to construct generic queries (understood in this study to be a query that has not been filtered by author, journal or keywords) in order to minimize ASEO query factors, and then to calculate the correlation between the rank position achieved in the results and the number of citations the documents have received. A moderate-to-high correlation would indicate that citations are the key determinant for ordering results in Google Scholar and, as a result, we would be confident about the ability of Google Scholar to identify highly-cited documents.

We defined a generic query through conducting a null query (search box is left blank), filtering only by publication year using Google Scholar's advanced search function. In this way, we avoided the sampling bias caused by the keywords of a specific query and by other academic search engine optimisation issues. In order to work with a sufficiently large data sample, a longitudinal analysis was carried out by performing 64 generic null queries from 1950 to 2013 (one query per year). Whereas 2013 was the last complete available year when our data collection was carried out, 1950 was selected because this particular year reflected an increase in coverage in comparison to the preceding years (Orduna-Malea et al, 2015). After this, all the returned documents (a maximum of 1000 per query) were listed, obtaining a final set of 64,000 results.

This process was carried out twice (28 May and 2 June 2014). The first time, it was performed from a computer connected to a WoScc subscription via IP range to obtain WoScc data embedded in Google Scholar (http://wokinfo.com/googlescholar); the second time from a computer with a normal Internet connection. This functioned as a reliability check, as it allowed us to confirm that the two datasets contained the same records. After this reliability check, the html source code for each of the search engine results pages for each query was parsed and downloaded, and all bibliographic information for each result (taken only from the primary version of the document) was extracted (supplementary material available at https://dx.doi.org/10.6084/m9.figshare.1224314.v1). The available details for each bibliographic field are represented in Figure 1.



[PDF] Protein measurement with the Folin phenol reagent

OH Lowry, NJ Rosebrough, AL Farr, RJ Randall - J biol Chem, 1951 - devbio.wustl.edu

Method Reagents-Reagent A, 2 per cent N&OX in 0.10 N NaOH. Reagent B, 0.5 per cent CuS04. 5Hz0 in 1 per cent sodium or potassium tartrabe. Reagent C, alkaline copper solution. Mix 50 ml. of Reagent A with 1 ml. of Reagent B. Discard after 1 day. Reagent D, ...

Cited by 192470 Related articles All 29 versions Web of Science: 305081 Cite Save More

wustl.edu [PDF]

Figure 1. Google Scholar's bibliographic fields in the search engine results page.

Among the different information elements gathered, the following were processed in order to meet the objectives of this study (the remaining elements are unlikely to influence the ranking):

- **Rank**: position that each document occupies in the Google Scholar search engine results
- **GS Citations**: number of citations the document has received according to Google Scholar in the time the query was performed.
- **Number of versions**: number of versions GS has found of the documents.
- **Publication date**: year when the document was published.

Since Google Scholar doesn't provide information about the language of the documents, it was manually checked by observing WoScc data (when possible) as well as the language in which the title and abstract of the document were written.

All these data were then exported to a spreadsheet in order to be statistically analysed. Since citation data follow a skewed distribution, a Spearman correlation was calculated in order to find a relationship between citations and rank position.

3. Results

The overall correlation between the number of citations received by the 64,000 documents and the position they occupied on the results page of Google Scholar at the time of the query is r= -0.67 ($\alpha < 0.05$). Figure 2 shows the value of this correlation for each of the 64 years analysed. The aim of this year-by-year correlation analysis was both to investigate its possible evolution in time and to ascertain its value for the 1000 maximum results given for each query.

The average annual value of the correlation coefficient is very high (negative values for the correlation are due to position 1 being better than position 1000) and stable ($\bar{X}$= -0.895; σ= 0.025). In recent years (except 2013) the correlations are slightly lower than the average value; for example, the lowest values found correspond to 2006 and 2007. Even so, these correlation coefficients are still very high.

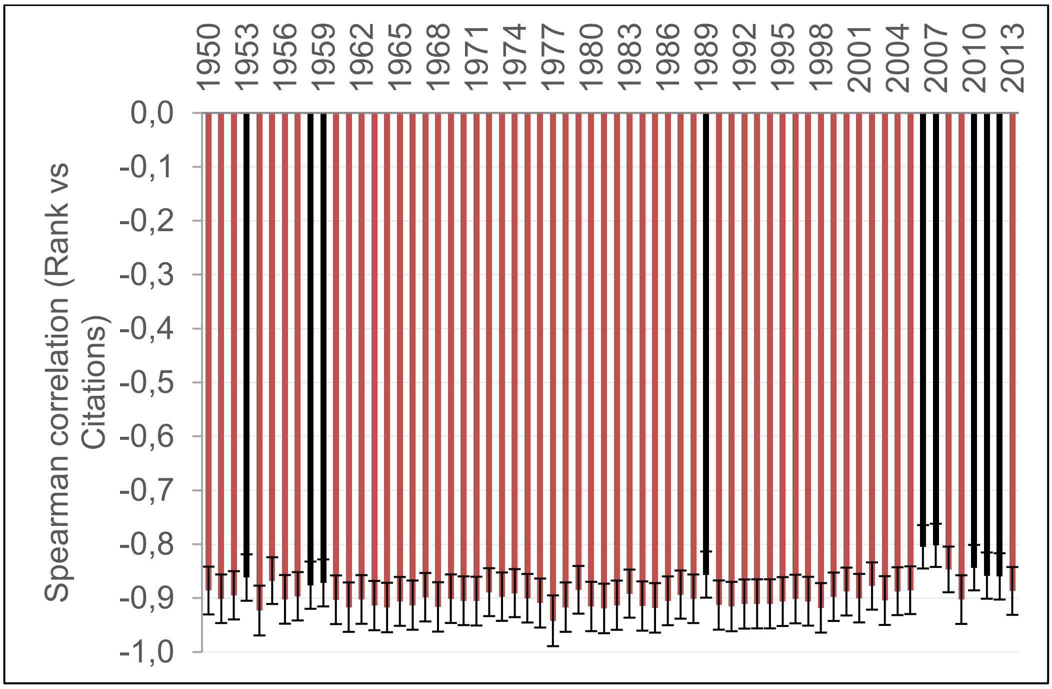


Figure 2. Spearman correlation between the number of citations received by documents in Google Scholar and the rank position they occupy in the search engine results page.

The fact that we obtained higher correlations in the annual samples than in the overall data indicates that there is a small deviation in the relationship between citations and a higher position in the list of results, which accumulates over 64 annual queries. In order to verify whether or not this deviation is concentrated in a specific area of the search engine results list, we proceeded to plot the dispersion between the position rank and citations received rank (Figure 3), marking the observation points according to the language of the document (English in green; Not English in red).

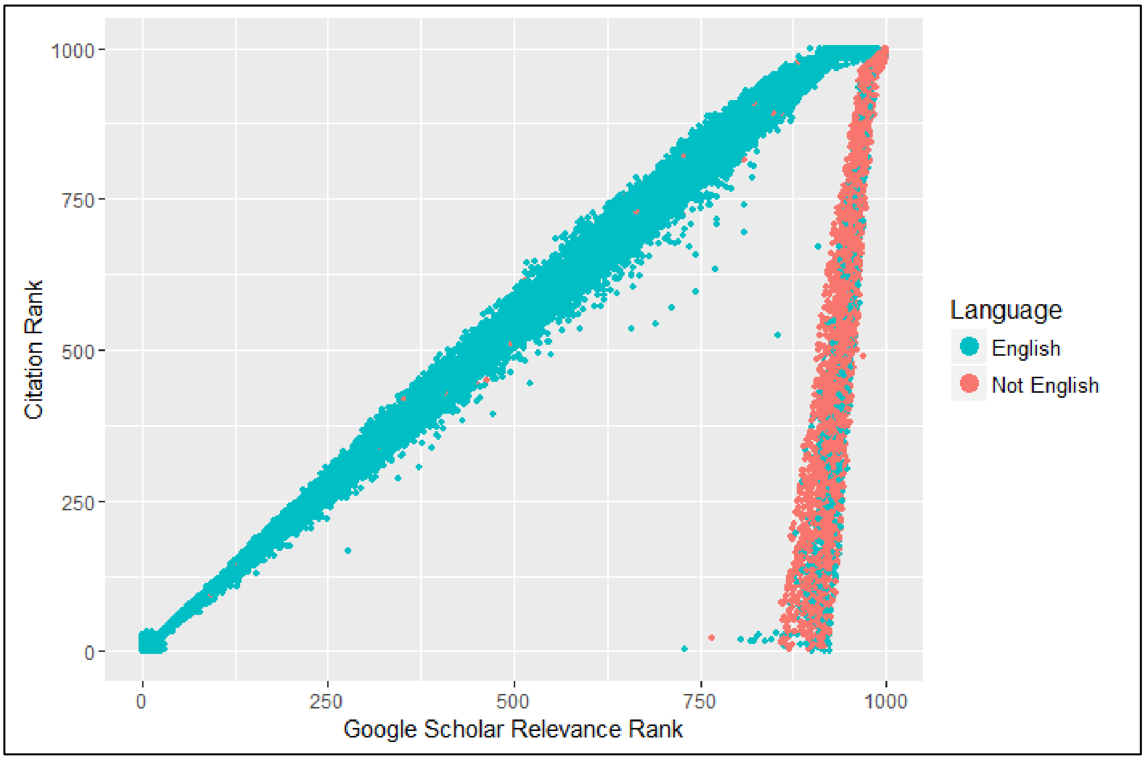


Figure 3. Relationship between the number of citations of documents in Google Scholar and the rank position they occupy in the search engine results page.

In Figure 3, the results located in the first 900 positions of each search are displayed in green, while the results in the last 100 positions are shown in red. In this way we can see clearly how, until approximately the 900th position, the Google Scholar sorting criteria are based largely on the number of citations received by each result. However, after approximately the 900th position, the data show erratic results in terms of the correlation between citations and position.

The correlation for the results placed amongst the top 900 positions is r= 0.97 ($\alpha < 0.01$). However, the correlation obtained for results in the last 100 positions is only r= 0.61 ($\alpha < 0.01$). In this case we calculated the Pearson correlation, as discrete ranking positions were being compared for both variables. Although the sample size is different in the calculation of these correlations (900 versus 100), the data indicate the existence of unexpected results for the last 100 positions of the Google Scholar results page, i.e. some highly-cited documents are found in very low positions.

The positions occupied by the documents that received the highest number of citations in each year partly corroborate this irregular behaviour in Google Scholar. Figure 4 shows how in only 11 of the 64 years analysed (17.2%), the most-cited document that year is ranked first in the results, while in 32.8% of the years, this document is among the top three. However, sometimes the most highly cited document occupies a very low position. The most extreme case was detected in 1978, where the document with the largest number of citations appeared in the 917th position.

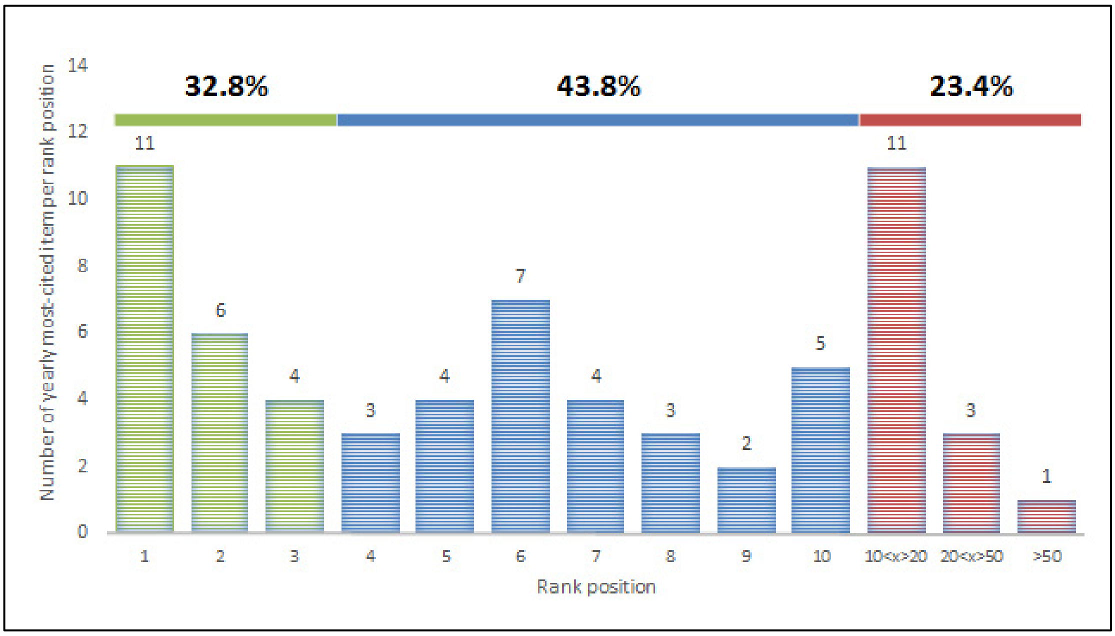


Figure 4. Frequency of position occupied by yearly most-cited document in the search engine results pages.

At the other end of the scale, the document located in the last position (1000) is the document with the fewest citations in 60.9% of the years. In the years that this is not the case, the difference between the least-cited document and the lowest-ranked document is never greater than 10 citations; therefore, the number of citations received by the document located at position 1000 is a good reflection of the citation threshold level. Figure 5 shows precisely this threshold value per year: around 50 citations during the early years of analysis (1950-1960), subsequently climbing to over 200 citations between 1985 and 2000, and from then on falling to around 50 citations again in 2010. The rise of this threshold value (especially the last years of 20th century) is attributed to the increasing scientific output worldwide. However, the decline from 2000 onwards is unexpected. Though this will need to be empirically tested, we attribute the decrease of this threshold value on the increase of citations to old documents, a phenomenon in which the Google Scholar's rank is precisely contributing (Martin-Martin et al, 2016). In 2013, an atypical value (133 citations) was obtained. The likely reason for this issue will be discussed in the next section.

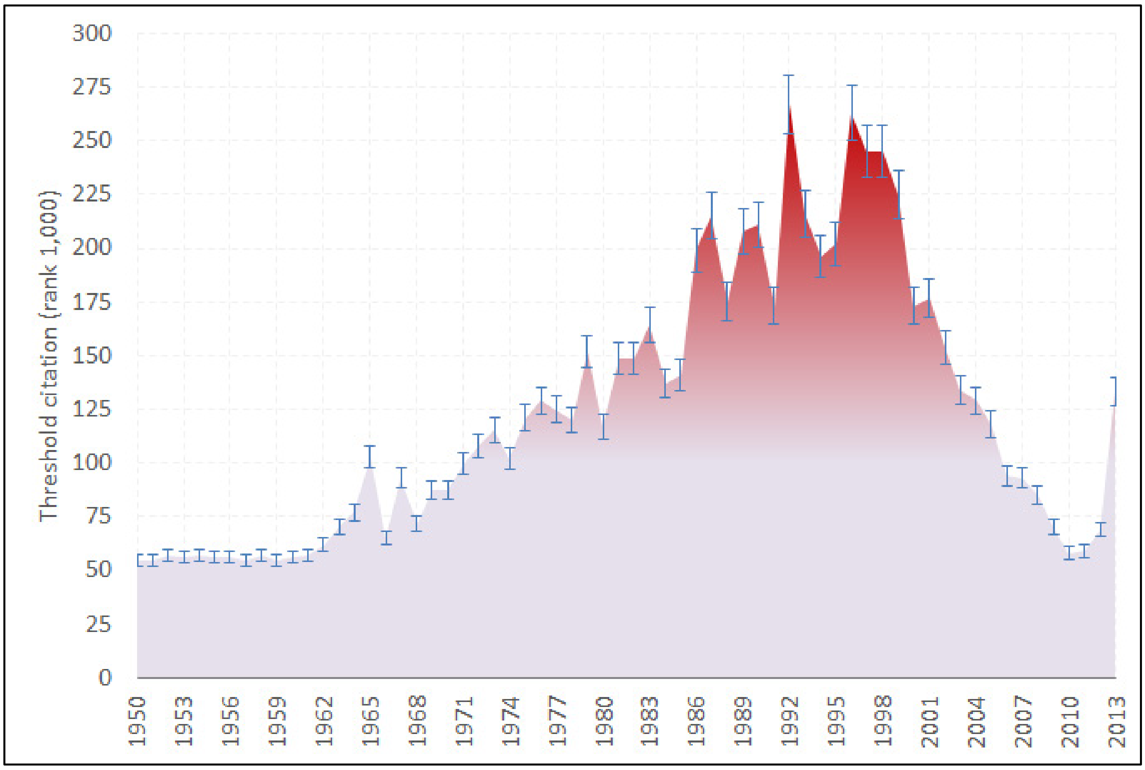


Figure 5. Number of citations received per year by the document ranked 1000.

4. Discussion

Generic queries minimise the effect of those academic search engine optimisation practices that are influenced by query terms. Unfortunately, the relationship between the rank position of the results and the citations received by them may be influenced or determined by other external variables, such as the dynamism of the Web, the malfunction of some Google Scholar features, and ASEO practices determined from specific document characteristics. These are three aspects that all require detailed discussion.

Dynamic nature of the academic search engine

The way search engines (not only academic search engines, such as Google Scholar, but also general search engines, such as Google or Bing) function can cause two identical queries, made on different computers in different geographical locations, or simply repeated after a short period of time, to generate slightly different results (Wilkinson & Thelwall, 2013). This, in turn, can cause some documents to appear or disappear, or to move to another position within the search results page. Therefore, the results of a study like our current study should be considered from a general perspective, without entering too much into individual details.

Despite this, and in order to test the potential variability of the search engine, we again compiled the sample of 64,000 documents using the same procedure described in the methodology four months later (4 October 2014), and compared both samples. Table 1 shows the number of documents from the first sample that are not retrieved in the second sample, ordered by position interval.

Table 1. Number of missing documents between the two samples of 64,000 highly-cited documents (May and October, 2014).

Rank interval	Missing documents	% Total (n= 64000)	% Partial (n= 9402)
001 – 100	402	0.6	4.3
101 – 200	340	0.5	3.6
201 – 300	319	0.5	3.4
301 – 400	373	0.6	4.0
401 – 500	450	0.7	4.8
501 – 600	588	0.9	6.3
601 – 700	778	1.2	8.3
701 – 800	1176	1.8	12.5
801 – 900	1802	2.8	19.2
901 – 1000	3174	5.0	33.8
TOTAL	**9402**	**14.7**	**100**

It is clear from Table 1 that accuracy diminishes the lower the position of the documents in the ranking. 14.7% of the 64,000 documents retrieved in the second sample (9402) are not found in the first. However of these, 65.4% (6152) are concentrated in the last 300 positions. This might have been influenced by the fact that the documents in these lower positions obtained similar or even identical values for the number of citations received. Hence even a change of one or two citations in the four-month lapse could lead a document to be included or excluded from the top-1000 results. Therefore, when considering highly-cited documents that occupy lower ranked positions results do need to be taken with a grain of salt. The same conclusion can be drawn from the low correlation coefficients obtained in Figure 3 (citation rank vs position rank) precisely for documents located in these lower ranked positions.

Google Scholar malfunction 1: Rank position and number of versions

Versions are a feature patented by Google Scholar (Verstak & Acharya, 2013) that enables all copies of the same document that are available online to be identified, and subsequently aggregated into a single result (adding up all the citations that each version may have received). We argue that the number of versions of a document could affect the relationship between citations and the position of a document mainly in two ways: multi-version effect (related to the number of versions; though it is not considered a malfunction, it is included in this section for expository clarity) and incorrect functioning effect (malfunction related to the version aggregation process).

The multi-version effect concerns the possibility of documents with a greater number of versions to appear in a higher position. Intuitively, one would expect documents with more versions to be dealing with important topics or written by outstanding researchers. This may explain why these documents are widely disseminated in several platforms. Accordingly, these documents are to be found more easily. Both their expected quality and wide discoverability may generate the multi-version documents to have more citations and, consequently, to climb in the Google Scholar's rank.

To determine the influence of the number of different versions on positioning, we calculated the dispersion between the number of versions of a document and its position on the results page (Figure 6).

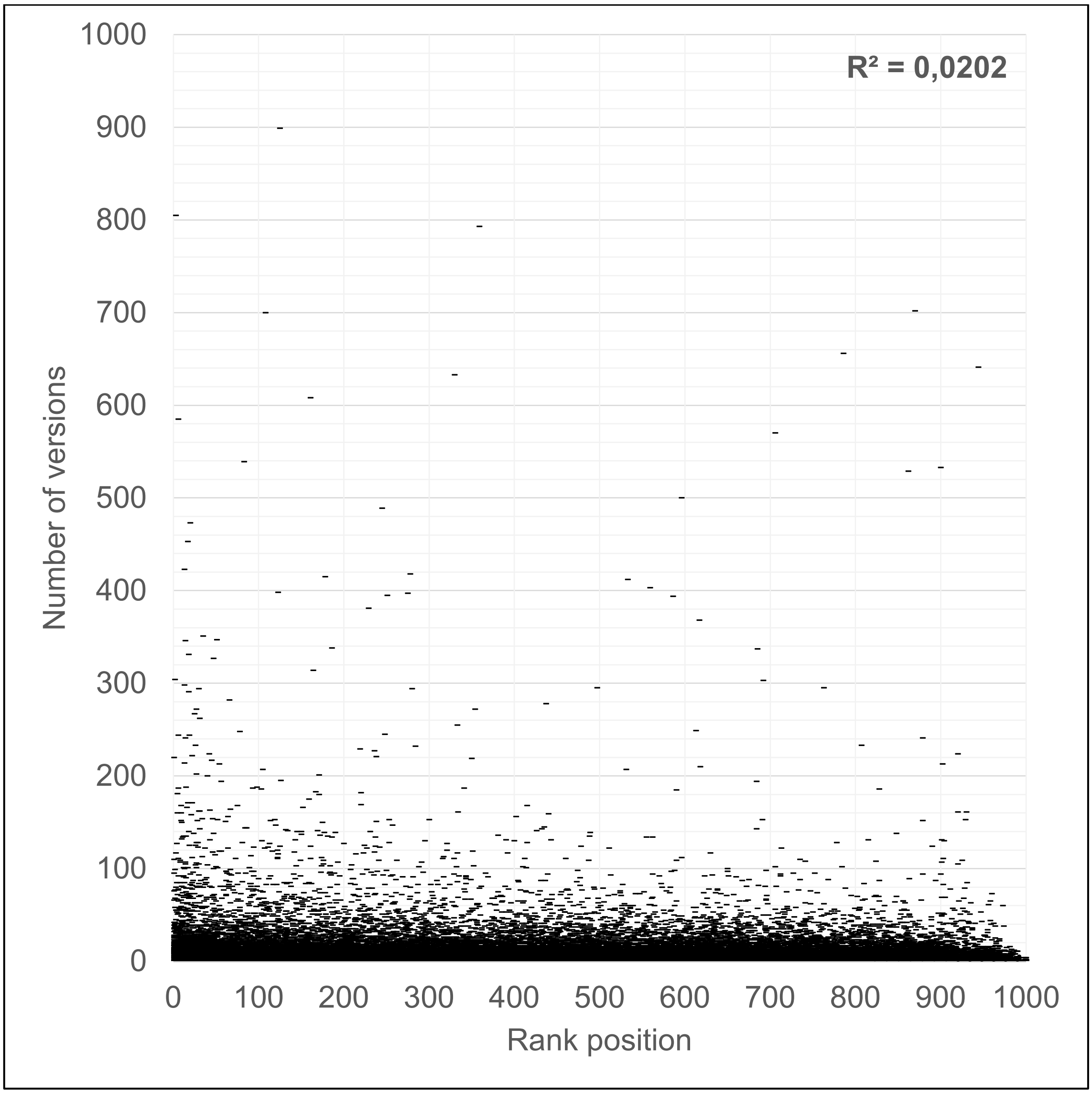


Figure 6. Scatter plot of the number of versions and rank position values for the 64,000 documents in Google Scholar.

The correlation between the position of a document and the number of versions is low, but significant (r = -0.30; α < 0.01). The average correlation per year is slightly higher (r = -0.33; σ = 0.04). Figure 6 shows that, despite the wide dispersion of data, there is a slight concentration of documents with between 100 and 300 versions amongst the first 100 rank positions. In order to analyse this observation more precisely, Table 2 gives us the average number of versions of documents in a given year depending on their location in a range of positions. High average values (with equally high standard deviations) were identified in the documents in the first 100 result positions, although this behaviour does not follow any stable pattern, and there are some notable exceptions. Hence, there does seem to be a slight positive effect of the number of versions on the rank position of a document in the top 100 result positions.

Table 2. Yearly average number of versions and standard deviation for documents according to rank position in Google Scholar (2014 – 2013).

Rank position	Versions																			
	2004		2005		2006		2007		2008		2009		2010		2011		2012		2013	
	$\tilde{X}$	σ	$\tilde{X}$	σ	$\tilde{X}$	σ	$\tilde{X}$	σ	$\tilde{X}$	σ	$\tilde{X}$	σ	$\tilde{X}$	σ	$\tilde{X}$	σ	$\tilde{X}$	σ	$\tilde{X}$	σ
1 - 100	23	35	19	20	25	30	23	28	22	50	18	25	23	63	16	20	16	25	12	11
101 - 200	22	25	23	26	21	20	26	34	19	37	17	21	13	13	18	70	12	34	15	44
201 - 300	16	19	28	52	24	32	28	59	19	24	16	20	20	57	12	12	11	24	8	12
301 - 400	18	21	22	27	17	16	15	11	21	78	15	19	17	26	12	13	10	27	8	9
401 - 500	16	19	16	14	18	18	18	16	17	19	17	32	17	20	11	9	9	17	7	7
501 - 600	13	11	16	14	14	15	16	13	17	16	13	11	16	40	10	12	8	12	9	23
601 - 700	15	16	18	14	17	15	15	11	15	18	16	21	16	16	10	12	9	19	6	6
701 - 800	14	11	14	11	16	17	15	12	13	9	16	17	12	10	9	11	8	9	12	65
801 - 900	12	10	15	15	11	17	17	54	10	9	11	12	11	9	11	25	20	87	8	19
901 -1000	10	64	5	15	3	5	3	6	4	13	7	23	7	22	5	9	4	5	6	13

In red: rank position for which the highest yearly average number of versions is obtained.

With regard to the incorrect functioning effect, we distinguish the following shortcomings:

a) Incorrect functioning leading to the omission of citations: the incorrect functioning of the version aggregation process could cause legitimate citations to a document to be omitted, causing it potentially to be excluded from the first 1000 results.
b) Incorrect functioning leading to the overestimation of citations: the incorrect functioning of the version aggregation process could cause citations to be wrongly attributed, thereby causing the document to be unjustly positioned amongst the top 1000 results.

However, given the low overall correlation detected between the number of versions and the position in the results, and without considering the exact position that each document should occupy if all existing versions were linked properly (which would require a systematic study focusing on this issue), we argue that version aggregation does not seem to affect greatly whether a document is included or excluded in the top 1000 results, i.e. the main objective of this study.

Google Scholar malfunction 2: Rank position and publication date

As was demonstrated in earlier studies, the results for Google Scholar's advanced option searches in a specific year (custom range) are not always entirely accurate (Orduna-Malea et al, 2015). This could mean that the actual year of publication of a document does not correspond with the year specified in the corresponding query. If this happens, a document may not appear among the results of a generic query (if it has no publication date) or it may appear among the results for another year (if it has an erroneous publication date). To determine the potential impact of this problem for the objectives of this study we performed two consistency tests (internal and external).

The internal consistency test verified whether the date of publication provided for each document corresponds to the date indicated in the advanced search for each of the 64 queries. The results of the test indicate that only in 2 documents (out of the 64,000 analysed) did the publication date not coincide with the date of the query. We can therefore conclude that the system works accurately at the technical level.

Another, quite different, issue is whether or not the date of publication provided by Google Scholar is correct. To this end, we conducted an external consistency test to cross-check the publication date of each document with the date provided by a controlled source independent of Google Scholar (in this case WoS). This is obviously assuming that WoS will provide correct data most of the time, though this is not always guaranteed, due to sporadic errors with online-first articles and other bibliographic data (Franceschini, Maisano & Mastrogiacomo, 2016).

We obtained a sample of the 64,000 documents (those that were linked to a WoS result), comprising 51% of the documents (32,680). The results of this process showed a match between the dates provided by both sources in 96.7% of the documents. Figure 7 displays the annual distribution of the documents in which there is no such match.

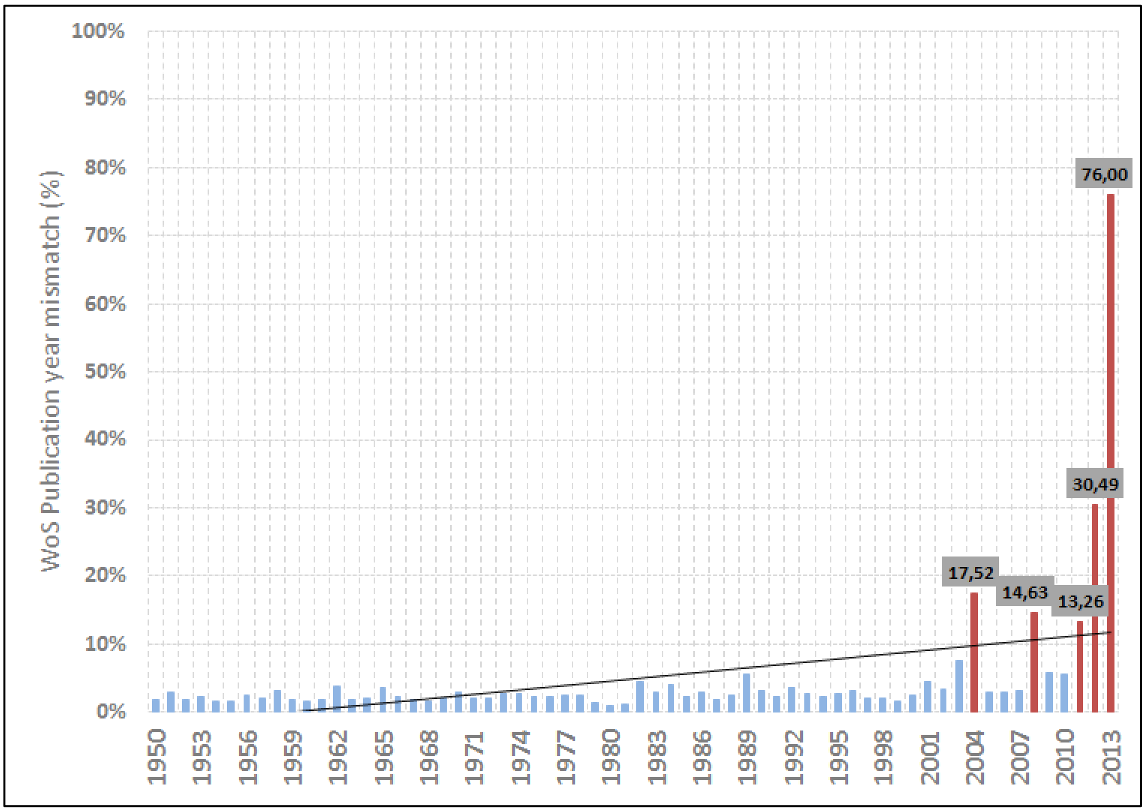


Figure 7. Percentage of publication year mismatches between documents in Google Scholar and the Web of Science between 1950 and 2013.

As shown in Figure 7, there is a concentration of errors in recent years, especially over the last two years of the period analysed (2012 and 2013), in which the error rate shoots up (30.49% and 76% respectively). This could explain the atypical value previously shown in Figure 5 for 2013.

A detailed analysis of 2013 (Figure 8) shows us how, out of the 19 errors detected this year, for 11 of them (58%) the error (difference between the year recorded by both sources) is more than 20 years, while only on two occasions is the error less than 2 years. These results therefore mean that this error cannot be attributed to the publication of preprints and/or periods during which the journal was under embargo. The Google Scholar practice of selecting the latest edition of a monograph as the main version seems to be the primary cause of these errors. Consequently, the small number of documents analysed for these two years (82 and 25 respectively) leads to a high proportion of mismatches. All different editions of a book (with their corresponding years of publication) are treated as versions by Google Scholar, after which Google Scholar selects as the primary version (the version used in this study) the document with the most recent publication date.

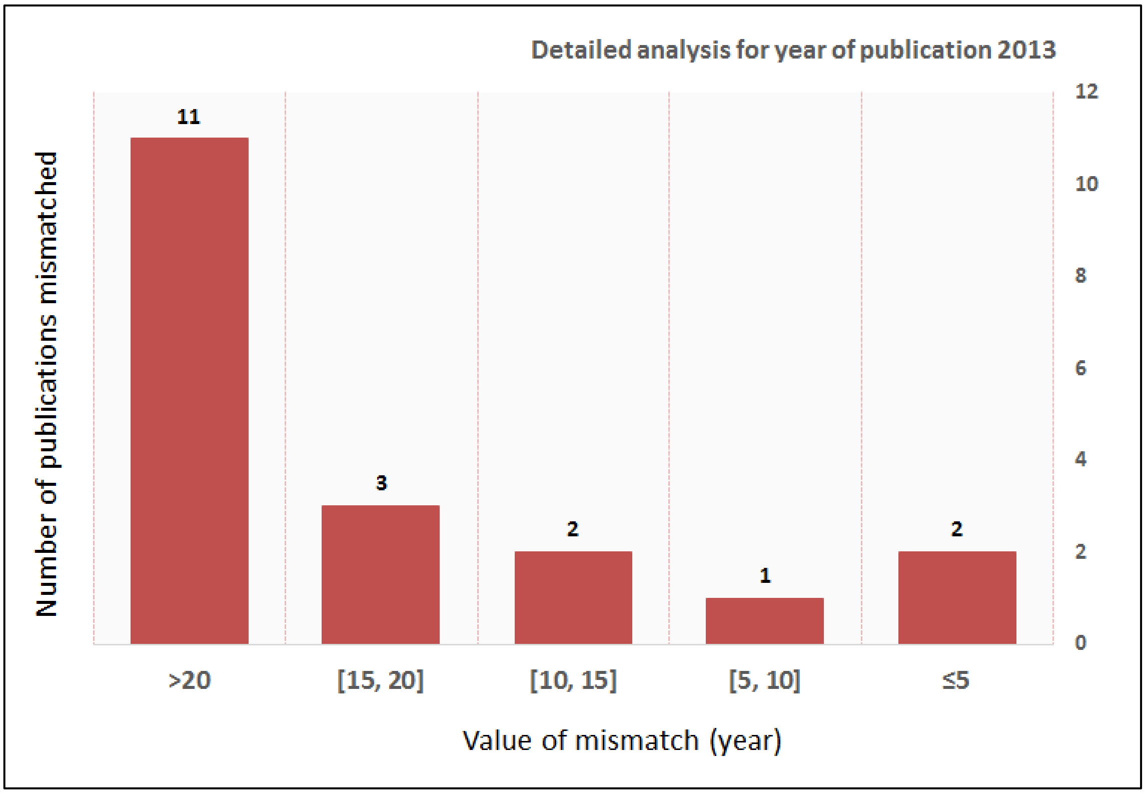


Figure 8. Publication year mismatches between documents in Google Scholar and the Web of Science (2013).

However, the percentage error for the data sample as a whole is very small. If we add to this the fact that an error in the date can cause the document to appear in the wrong year, but not exclude it from the results of a generic search, we may safely say that the publication date does not significantly affect the ability of Google Scholar to identify highly-cited documents.

ASEO document factor: Rank position and language of publication

Finally, we looked at the possible influence of the language of publication on the rank position in the results. To study this effect in detail, we have analysed the percentage of documents published in English in the first and last 100 results of each year (Figure 9).

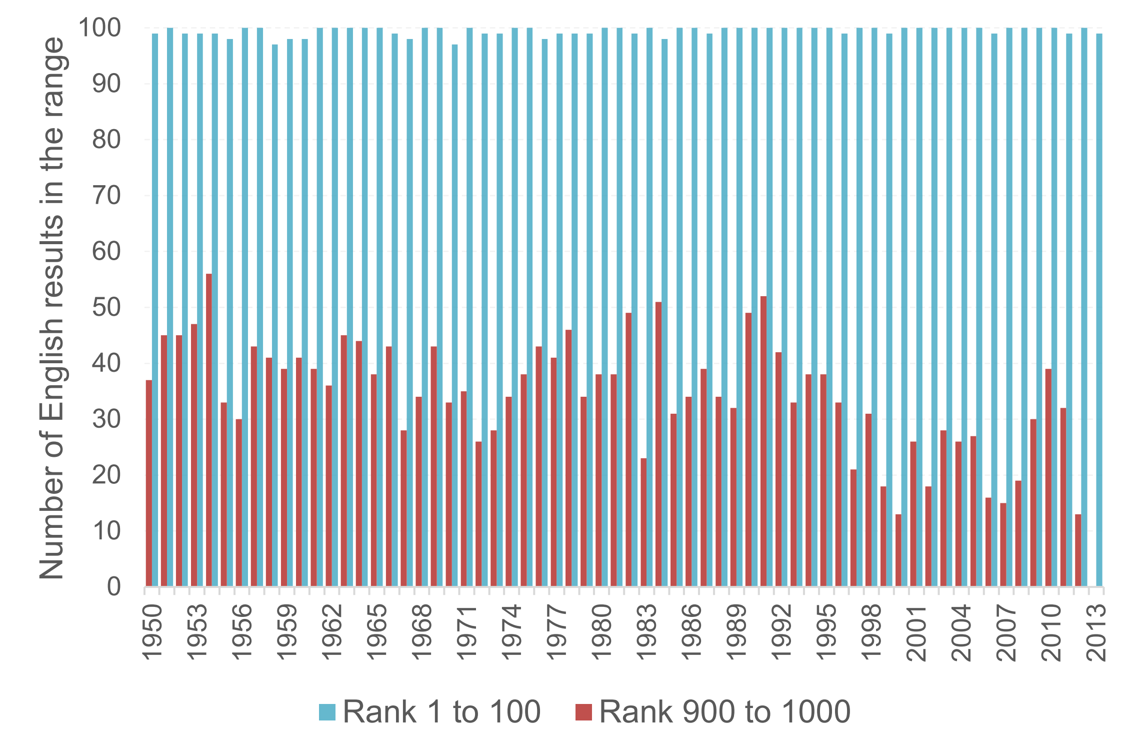


Figure 9. Yearly percentage of English documents among the first and last 100 ranked documents.

The annual average number of documents in English for results within the first 100 positions is 99.5. Therefore, the presence of documents in other languages within this range is abnormal. When analysing this same percentage for the documents in the last 100 positions, the results change significantly. The annual average drops to 34.2%.

The high presence of documents in languages other than English (often with very high levels of citations) in the last 100 positions could help explain the low correlations identified in this range between the ranking positions and citations received (Figure 3). This may have been due our choice of interface language (English was selected during the study). It should be noticed that users cannot select the actual language of documents in the Google Scholar's advanced search features but instead select the language of the website. The latter is primarily identified by detecting the geographic domain in which the document is available online (for example, .nl, .es or .jp) and does not guarantee the website is actually in that language. This may explain the fact that some documents written in English but with their primary version hosted in non-Anglophone countries' web domains do appear in lower positions in spite of receiving a large number of citations.

Therefore, if the queries had been conducted by restricting the geographic web domains to Anglophone countries, it is likely that the correlation coefficients would have been significantly higher, especially in the last quartile of the ranking results. However, this obviously would result in a biased interpretation of which publications are most highly cited, limiting the results largely to English-language publications. The effect of the choice of interface language on the results has already been partly studied in the past (Lewandowski, 2008). However, this effect should be tested empirically, and methodically, in the future to assess the impact of the interface language with greater accuracy.

5. Conclusions

A significant and high correlation between the number of citations and the ranking of the documents retrieved by Google Scholar was obtained for a generic query filtered only by year. The fact that we minimised the effects of academic search engine optimisation, together with the size of the sample analysed (64,000 documents), leads us to conclude that the number of citations is a key factor in the ranking of the results and, therefore, that Google Scholar is able to identify highly-cited papers effectively. Given the unique coverage of Google Scholar (no restrictions on document type and source), this makes it an invaluable tool for bibliometric analysis.

However, the correlation that was obtained, though high, was not excellent because of external factors (especially the language of publication and the geographic web domain where the primary version was hosted) that mainly affected the results at the bottom of the list (approximately the last 100). Restricting the language of the results to match the interface language may help to improve accuracy in the search for highly-cited papers, although this obliges us to perform as many queries as the languages we wish to analyse. Unfortunately, users can only restrict the language of the website, and this procedure is far from optimal as it mainly relies on geographic web domains.

Other factors, such as the date of publication (when erroneous) or the number of versions (multi-version effect and incorrect functioning of version aggregation effect) only have an incidental impact, and do not compromise the proven ability of Google Scholar to search for highly-cited documents.

Therefore, we conclude that Google Scholar can be used to reliably identify the most highly-cited academic documents. Given its wide and varied coverage, Google Scholar has become a useful complementary tool for Bibliometrics research concerned with the identification of the most influential scientific works.

Acknowledgements

Alberto Martín-Martín enjoys a four-year doctoral fellowship (FPU2013/05863) granted by the Ministerio de Educación, Cultura, y Deportes (Spain). Enrique Orduna-Malea holds a postdoctoral fellowship (PAID-10-14), from the Polytechnic University of Valencia (Spain).

References

Aguillo, I.F. (2012). Is Google Scholar useful for bibliometrics? A webometric analysis. *Scientometrics*, *91*(2), 343-351.

Bakkalbasi, N., Bauer, K., Glover, J. & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical digital libraries*, *3*(1), 7.

Bar-Ilan, J. (2007). Which h-index? - A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, *74*(2), 257-271.

Bar-Ilan, J. (2010). Citations to the “Introduction to informetrics” indexed by WOS, Scopus and Google Scholar. *Scientometrics*, *82*(3), 495-506.

Bar-Ilan, J., Levene, M. & Lin, A. (2007). Some measures for comparing citation databases. *Journal of Informetrics*, *1*(1): 26-34.

Beel, J. & Gipp, B. (2009). Google Scholar’s ranking algorithm: an introductory overview. In Birger Larsen & Jacqueline Leta (Eds.). *Proceedings of the 12th International Conference on Scientometrics and Informetrics* (Vol. 1, pp. 230-241). Rio de Janeiro, Brazil.

Beel, J. & Gipp, B. (2010). Academic Search Engine Spam and Google Scholar’s Resilience Against it. *JEP- the journal of electronic publishing*, *13*(3), Retrieved October 31, 2015, from http://dx.doi.org/10.3998/3336451.0013.305

Beel, J., Gipp, B. & Wilde, E. (2010). Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co. *Journal of Scholarly Publishing*, *41*(2), 176-190.

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A. & Daniel, H. D. (2009). Convergent validity of bibliometric Google Scholar data in the field of chemistry—Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstracts. *Journal of Informetrics*, *3*(1), 27-35.

Bramer, W.M., Giustini, D., Kramer, B.M.R. & Anderson, P.F. (2013). The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic reviews: a review of searches used in systematic reviews. *Systematic reviews*, *2*(1), 115.

Breeding, M. (2015). *The future of library resource discovery*. NISO Whitepapers. Baltimore: NISO.

Butler D. (2004). Science searches shift up a gear as Google starts Scholar engine. *Nature*, *432*(7016), 423.

Cabezas-Clavijo, A. & Delgado-López-Cózar, E. (2013). Google Scholar e índice h en biomedicina: la popularización de la evaluación bibliométrica. *Medicina intensiva*, *37*(5), 343-354.

Callicott, B. & Vaughn, D. (2005). Google Scholar vs. Library Scholar: Testing the Performance of Schoogle. *Internet Reference Services Quarterly*, *10*(3), 71-88.

Carpenter, J. (2012). Researchers of Tomorrow: The research behaviour of Generation Y doctoral students. *Information Services and Use*, *32*(1), 3-17.

Delgado López-Cózar, E., Robinson-García, N. & Torres-Salinas, D. (2014). The Google Scholar Experiment: how to index false papers and manipulate bibliometric indicators*. Journal of the American Society for Information Science and Technology*, *65*(3), 446-454.

Ettinger, D. (2008). The triumph of expediency: The impact of Google Scholar on library instruction. *Journal of Library Administration*, *46*(3-4), 65-72.

Evans, M. P. (2007). Analysing Google rankings through search engine optimization data. *Internet research*, *17*(1), 21-37.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics*, *10*(4), 933-953.

Gardner, S. & Eng, S. (2005). Gaga over Google? Scholar in the Social Sciences. *Library Hi Tech News*, *22*(8), 42-45.

Gehanno, J.F., Rollin, L. & Darmoni, S. (2013). Is the coverage of Google Scholar enough to be used alone for systematic reviews. *BMC Medical Informatics and Decision Making*, *13*(7).

Giles, J. (2005). Science in the web age: start your engines. *Nature*, 438(7068), 554-555.

Giustini, D. & Barsky, E. (2005). A look at Google Scholar, PubMed, and Scirus: comparisons and recommendations, *Journal of the Canadian Health Libraries Association*, *26*(3), 85-89.

Giustini, D. (2005). How Google is changing medicine. *BMJ*, *331*(7531), 1487-1488.

Harzing, A. W. (2013). A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. *Scientometrics*, *94*(3), 1057-1075.

Harzing, A. W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics*, *98*(1), 565-575.

Harzing, A. W. & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*(2), 787-804.

Harzing, A.W. & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, *8*(1), 61-73.

Housewright, R., Schonfeld, Roger C. & Wulfson, K. (2013). *UK Survey of Academics 2012.* Itaka S+R, JISC & RLUK.

Howland, J. L., Wright, T. C., Boughan, R. A., & Roberts, B. C. (2009). How scholarly is Google Scholar? A comparison to library databases. *College & Research Libraries*, *70*(3), 227-234.

Jacso, P. (2005). Google Scholar: the pros and the cons. *Online information review*, *29*(2), 208-214.

Jacso, P. (2006). Deflated, inflated, and phantom citation counts. *Online Information Review*, *30*(3), 297-309.

Jacso, P. (2008a). Testing the Calculation of a Realistic h-index in Google Scholar, Scopus, and Web of Science for F. W. Lancaster. *Library Trends*, *56*(4), 784-815.

Jacso, P. (2008b). The pros and cons of computing the h-index using Google Scholar. *Online Information Review*, *32*(3), 437-452.

Jacsó, P. (2008c). The pros and cons of computing the h-index using Scopus. *Online Information Review*, *32*(4), 524-535.

Jacso, P. (2012). Using Google Scholar for journal impact factors and the h-index in nationwide publishing assessments in academia - siren songs and air-raid sirens. *Online Information Review*, *36*(3), 462-478.

Khabsa, M. & Giles, C. L. (2014). The number of scholarly documents on the public web. *PLoS One*, 9(5), e93949.

Kousha, K. & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, *74*(2), 273–294.

Kousha, K., Thelwall, M. & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science*, *62*(11), 2147-2164.

Ledford, J. L. (2009). *Search Engine Optimization Bible* (2nd ed.). Indianapolis: John Wiley & Sons.

Leslie M. A. (2004). Google for academia. *Science*, *306*(5702), 1661-1663.

Levine-Clark, M. & Kraus, J. (2007). Finding Chemistry Information Using Google Scholar. *Science & Technology Libraries*, *27*(4), 3–17.

Lewandowski, D. (2008). Problems with the use of web search engines to find results in foreign languages. *Online information review*, *32*(5), 668-672.

Martín-Martín, A., Ayllón, J.M., Orduna-Malea, E. & Delgado López-Cózar, E. (2015). *Proceedings Scholar Metrics: H Index of proceedings on Computer Science, Electrical & Electronic Engineering, and Communications according to Google Scholar Metrics (2010-2014)*. EC3 Reports, 15.

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & López-Cózar, E. D. (2016). Back to the past: on the shoulders of an academic search engine giant. *Scientometrics*, *107*(3), 1477-1487.

Meho, L.I. & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105–2125.

Meier John J. & Conkling, Thomas W. (2008). Google Scholar's Coverage of the Engineering Literature: An Empirical Study. *Journal of Academic Librarianship*, *34*(3), 201.

Orduna-Malea, E. & Aguillo (2014). *Cibermetría: midiendo el espacio red*. Barcelona: UOC Publishing.

Orduna-Malea, E. & Delgado López-Cózar, E. (2014). Google Scholar Metrics evolution: an analysis according to languages. *Scientometrics*, *98*(3), 2353–2367.

Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A. & López-Cózar, E. D. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*,*104*(3), 931-949.

Ortega, J. L. (2014). *Academic search engines: A quantitative outlook*. Oxford: Chandos.

Reed Elsevier (2004). Scopus Comes of Age. Retrieved October 31, 2015, from https://www.elsevier.com/about/press-releases/science-and-technology/scopus-comes-of-age

Schonfeld, Roger C. & Housewright, R. (2010). *Faculty Survey 2009: Key Strategic Insights for Libraries, Publishers, and Societies*. Ithaka S+R.

Știrbu, S., Thirion, P., Schmitz, S., Haesbroeck, G. & Greco, N. (2015). The utility of Google Scholar when searching geographical literature: comparison with three commercial bibliographic databases. *The Journal of Academic Librarianship*, *41*(3), 322-329.

Torres-Salinas, D., Ruiz-Pérez, R. & Delgado-López-Cózar, E. (2009). Google Scholar como herramienta para la evaluación científica. *El profesional de la información*, *18*(5), 501-510.

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. *Nature*, 512(7513), 126-169.

Verstak, A. A. & Acharya, A. (2013). Identifying multiple versions of documents. U.S. Patent No. 8589784. Washington, DC: U.S. Patent and Trademark Office.

Wilkinson, D. & Thelwall, M. (2013). Search markets and search results: The case of Bing. *Library & Information Science Research*, *35*(4), 318-325.

Winter, J.C.F., Zadpoor, A. & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics*, *98*(2), 1547–1565.

Yang, K. & Meho, L.I. (2006). Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–15.

Chapter 6. Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison

Journal Article. Cite as:

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. Scientometrics, 116(3), 2175–2188. https://doi.org/10.1007/s11192-018-2820-9

Abstract (English)

This study explores the extent to which bibliometric indicators based on counts of highly-cited documents could be affected by the choice of data source. The initial hypothesis is that databases that rely on journal selection criteria for their document coverage may not necessarily provide an accurate representation of highly-cited documents across all subject areas, while inclusive databases, which give each document the chance to *stand on its own merits*, might be better suited to identify highly-cited documents. To test this hypothesis, an analysis of 2,515 highly-cited documents published in 2006 that Google Scholar displays in its *Classic Papers* product is carried out at the level of broad subject categories, checking whether these documents are also covered in Web of Science and Scopus, and whether the citation counts offered by the different sources are similar. The results show that a large fraction of highly-cited documents in the Social Sciences and Humanities (8.6%-28.2%) are invisible to Web of Science and Scopus. In the Natural, Life, and Health Sciences the proportion of missing highly-cited documents in Web of Science and Scopus is much lower. Furthermore, in all areas, Spearman correlation coefficients of citation counts in Google Scholar, as compared to Web of Science and Scopus citation counts, are remarkably strong (.83-.99). The main conclusion is that the data about highly-cited documents available in the inclusive database Google Scholar does indeed reveal significant coverage deficiencies in Web of Science and Scopus in several areas of research. Therefore, using these selective databases to compute bibliometric indicators based on counts of highly-cited documents might produce biased assessments in poorly covered areas.

Abstract (Spanish)

Este trabajo explora hasta qué punto a los indicadores bibliométricos basados en conteos de documentos altamente citados les podría afectar la elección de la fuente de datos. La hipótesis inicial es que las bases de datos que restringen su cobertura en función de criterios de selección de revistas podrían no proporcionar una representación precisa de los documentos altamente citados en todas las áreas de conocimiento, mientras que las bases de datos inclusivas, que dan a cada documento la oportunidad de *alzarse por sus propios méritos*, podrían ser más adecuadas para identificar documentos altamente citados. Para comprobar esta hipótesis, se realiza un análisis por categorías temáticas de los 2.515 documentos publicados en 2006 que Google Scholar muestra en su producto *Classic Papers*. En este análisis se comprueba si estos documentos también están indizados en Web of Science y Scopus, y si los conteos de citas ofrecidos por estas fuentes son similares. Los resultados muestran que una gran parte de los documentos altamente citados en Ciencias Sociales y Humanidades (8,6%-28,2%) no están indizados en Web of Science y Scopus. En las ciencias naturales, de la vida, y de la salud, la proporción de documentos altamente citados que no están en Web of Science y Scopus es mucho más baja. Además, en todas las áreas, los coeficientes de correlación de Spearman entre los conteos de citas de Google Scholar, comparados con los de Web of Science y Scopus, son

extremadamente altos (,83-,99). La principal conclusión es que, efectivamente, los datos sobre documentos altamente citados que hay disponibles en la base de datos inclusiva Google Scholar revelan deficiencias de cobertura significativas en Web of Science y Scopus en varias áreas temáticas. Por tanto, usar estas bases de datos selectivas para calcular indicadores bibliométricos basados en conteos de documentos altamente citados podría producir resultados sesgados en las áreas en las que existe cobertura deficiente.

1. Introduction

The issue of database selection for calculating bibliometric indicators

It has been proposed that bibliometric indicators based on counts of highly-cited documents are a better option for evaluating researchers than using indicators such as the h-index (Bornmann & Marx, 2014; Leydesdorff, Bornmann, Mutz, & Opthof, 2011). A recent discussion held within the journal Scientometrics brought up this issue once again (Bornmann & Leydesdorff, 2018).

It is known that database selection affects the value that a bibliometric indicator takes for a given unit of analysis (Archambault, Vignola-Gagné, Côté, Larivière, & Gingrasb, 2006; Bar-Ilan, 2008; Frandsen & Nicolaisen, 2008; Meho & Yang, 2007; Mongeon & Paul-Hus, 2016). These differences are sometimes caused by diametrically opposed approaches to document indexing: indexing based on journal selection (Web of Science, Scopus), or inclusive indexing based on automated web crawling of individual academic documents (Google Scholar, Microsoft Academic, and other academic search engines). For an exhaustive commentary and bibliography on studies that compare the coverage and bibliometric indicators available in the previously mentioned databases (especially for studies that involve Google Scholar), we refer to Halevi, Moed & Bar-Ilan (2017), and Orduna-Malea, Ayllón, Martín-Martín, & Delgado López-Cózar (2015). Lastly, Delgado López-Cózar, Orduna-Malea, & Martín-Martín (2019) presents a detailed summary of all studies published to date that discuss the differences between Google Scholar, Web of Science, and Scopus in terms of coverage and bibliometric indicators, and the correlations of citation-based indicators at various levels of aggregation[19].

Using databases in which document coverage depends on journal selection criteria (selective databases) to calculate indicators based on counts of highly-cited documents could produce biased assessments. This is because documents other than those published in journals selected by these databases could also become highly-cited. These documents could be books, reports, conference papers, articles published in non-selected journals… which could very well meet the same quality criteria as the documents covered in selective databases. Because it is not possible to predict which documents are going to become highly-cited before they are published, an inclusive database that gives each document the chance to *stand on its own merit* (Acharya, 2015), might in theory provide a better coverage of highly-cited documents than a selective database where document coverage is constricted to specific sources selected beforehand.

Compounded with the previous issue, there is the fact that Web of Science and Scopus, the most widely used selective databases for bibliometric analyses, are known to have poor coverage of areas in which research often has a local projection such as the Social Sciences and Humanities (Mongeon & Paul-Hus, 2016), as well as a bias against non-English publications (Chavarro, Ràfols, & Tang, 2018; van Leeuwen, Moed, Tijssen, Visser, & Van Raan, 2001). This goes against the principle of protecting "excellence in locally relevant research" in the Leiden Manifesto (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015).

There is evidence to show that highly-cited documents are not only being published in elite journals. Acharya et al. (2014) found that, according to data from Google Scholar, the number of highly-cited documents published in non-elite journals had significantly grown between 1995 and 2013. They posited that this change was made possible by web search and relevance rankings, which meant that nowadays "finding and reading relevant articles in non-elite journals is about as

[19] Supplementary material to book chapter containing summary tables already available at: https://doi.org/10.17605/OSF.IO/PQR53

easy as finding and reading articles in elite journals", whereas before web search, researchers were mostly limited to what they could browse in physical libraries, or to systems that only presented results in reverse chronological order. Martín-Martín, Orduna-Malea, Ayllón, and Delgado López-Cózar (2014) carried out an analysis of 64,000 highly-cited documents according to Google Scholar, published between 1950 and 2013. In this exploratory study they found that 49% of the highly-cited documents in the sample were not covered by the Web of Science. They also found that at least 18% of these 64,000 documents were books or book chapters (Martín-Martín, Orduna-Malea, Ayllón, & Delgado López-Cózar, 2016).

Google Scholar's Classic Papers

Since June 14th 2017, Google Scholar started providing a new service called *Classic papers*[20] which contains lists of highly-cited documents by discipline. Delgado López-Cózar, Martín-Martín, and Orduna-Malea (2017) explored the strengths and limitations of this new product.

The current version of Google Scholar's *Classic Papers* displays 8 broad subject categories. These broad categories contain, in total, 252 unique, more specific subject categories. Each specific subject category (from here on called subcategory) contains the top 10 most cited documents published in 2006. These documents meet three inclusion criteria: they presented original research, they were published in English, and by the time of data collection (May 2017, and therefore at least 10 years after their publication), they had at least 20 citations. Documents appear to have been categorized at the article level, judging by the fact that articles in multidisciplinary journals such as *Nature*, *Science*, or *PNAS* are categorized according to their respective topics. Appendix A provides a high-level comparison of how Google Scholar, Web of Science, and Scopus classify this sample of documents.

Despite the fact that, in line with Google Scholar's usual lack of transparency, there are many unanswered methodological questions about the product, like how the subject categorization at the document level was carried out, this dataset could shed some light on the differences in coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus. The results may provide evidence of the advantages and disadvantages of selective databases and inclusive databases for the specific purpose of finding highly-cited documents.

Research Questions

This study aims to answer the following research questions:

RQ1. How many highly-cited documents according to Google Scholar are not covered by Web of Science and Scopus? Are there significant differences at the level of subject categories?

RQ2. To the extent that coverage of highly-cited documents in these databases overlaps, are citation counts in Google Scholar similar in relative terms (rank orders) to those provided by Web of Science and Scopus?

RQ3. Which, out of Google Scholar, Web of Science, and Scopus, gives the most citations for highly-cited documents? Are there significant differences at the level of subject categories?

[20] https://scholar.googleblog.com/2017/06/classic-papers-articles-that-have-stood.html

2. Methods

In order to carry out the analysis, we first extracted all the information available in Google Scholar's *Classic Papers*. For this purpose, a custom script was developed which scraped all the relevant information, and saved it as a table in a spreadsheet file. The information extracted was:

- Broad subject categories and subcategories.
- Bibliographic information of the documents, including:
  - Title of the document, and URL pointing to the Google Scholar record for said document.
  - Authors (including URL to Google Scholar Citations profile when available), name of the publication venue, and year of publication.
  - Name and URL to Google Scholar Citations profile of showcased author (usually the first author, or the last author if the first doesn't have a public profile).
  - Number of citations the document had received when the product was developed (May 2017).

A total of 2,515 records were extracted. All subcategories display the top 10 most cited documents, except the subcategory French Studies, in which only 5 documents were found with at least 20 citations.

Once the data from *Classic Papers* had been extracted, we proceeded to check how many of those 2,515 documents were also covered by Web of Science Core Collection, and Scopus. To do this, we used the metadata embedded in the URL that pointed to the Google Scholar record of the documents. In most cases, this URL contained the DOI of the document. Those DOIs were manually searched in the respective web interfaces of the other two databases, making sure that the documents that were found were actually the ones that were searched. In the cases when a DOI wasn't available in the URL provided by Google Scholar (only 105 records out of 2,515), and also when the DOI search wasn't successful, the search was conducted using the title of the document. If the document was found, its local ID in the database (the accession number in Web of Science, and the EID in Scopus), as well as its citation count was appended to the original table extracted from *Classic Papers*. For the documents that were not found, the cause why the document was not available was identified. The reasons identified were:

- The source (journal / conference) is not covered by the database.
- Incomplete coverage of the source (only some volumes or issues were indexed). A special case of this is when the source wasn't being indexed in 2006, but it started being indexed at a later date.
- The document has not been formally published: for the few cases (4) in which reports or preprints that were not eventually published made the list of highly-cited documents.

Data collection was carried out in June 2017, shortly after *Classic Papers* was launched. At the moment of writing this piece, searches in Web of Science and Scopus were carried out again to double-check that there had been no changes. It turned out that 2 additional documents were found in the Web of Science, and 7 additional documents were found in Scopus. These documents were not added to the sample, because by the time of the second search, they had had almost one additional year to accumulate citations and therefore comparisons of citation counts between sources would have not been fair.

Lastly, in order to clean the bibliographic information extracted from Google Scholar, which often presented incomplete journal or conference titles, we extracted the bibliographic information from CrossRef and DataCite using the available DOIs and content negotiation. For the cases when no DOI was available, the information was exported from Scopus, or added manually (mostly for the 79 documents which were not available in either of the databases).

To answer RQ1, the proportions of highly-cited documents in Google Scholar that were not covered in Web of Science and/or Scopus were calculated at the level of broad subject categories. Additionally, the most frequent causes why these documents were not covered are provided.

To answer RQ2, Spearman correlation coefficients of citation counts were calculated between the pairs of databases Google Scholar/Web of Science, and Google Scholar/Scopus. Correlation coefficients are considered useful in high-level exploratory analyses to check whether different indicators reflect the same underlying causes (Sud & Thelwall, 2014). In this case, however, the goal is to find out whether the same indicator, based on different data sources, provides similar relative values. Spearman correlations were used because it is well-known that the distributions of citation counts and other impact-related metrics are highly skewed (De Solla Price, 1976).

To answer RQ3, the average log-transformed citation counts for the three databases were calculated at the level of broad subject categories, and the normal distribution formula was used to calculate 95% confidence intervals for the log-transformed data (Thelwall, 2017; Thelwall & Fairclough, 2017).

The raw data, the R code used for the analysis, and the results of this analysis are openly available (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018).

3. Results

RQ1. How many highly-cited documents according to Google Scholar are not covered by Web of Science and Scopus? What are the differences at the level of subject categories?

Out of the 2,515 documents displayed in Google Scholar's *Classic Papers*, 208 (8.2%) were not covered in Web of Science, and 87 (3.4%) were not covered in Scopus. In total, 219 highly-cited documents were not covered either by Web of Science or Scopus. Among these, 175 of them were journal articles, 40 were conference papers, one was a report, and three were preprints. Regarding these preprints, all three are in the area of Mathematics. As far as we could determine, a heavily modified version of one of the preprints was published in a journal two years after the preprint was first made public, but the other two preprints have not been published in journals.

Significant differences in coverage were found across subject categories (Table 1). The areas where there are more highly-cited documents missing from Web of Science and Scopus are *Humanities, Literature & Arts* (28.2% in Web of Science, 17.1% in Scopus), and *Social Sciences* (17.5% in Web of Science, and 8.6% in Scopus). Moreover, Web of Science seems to be missing many highly-cited documents from *Engineering and Computer Science* (11.6%), and *Business, Economics & Management* (6.0%). The coverage of these last two areas in Scopus seems to be better (2.5% and 2.7% missing documents, respectively).

Table 1. Number of highly-cited documents in Google Scholar that are not covered by Web of Science and/or Scopus, by broad subject areas

Subject category	N	Not in WoS	%	Not in Scopus	%
Humanities, Literature & Arts	245	69	28.2	42	17.1
Social Sciences	510	89 (J: 88, R: 1)	17.5	44 (J: 43, R: 1)	8.6
Engineering & Computer Science	570	66 (J: 26, C: 40)	11.6	14 (J: 10, C: 4)	2.5
Business, Economics & Management	150	9	6.0	4	2.7
Health & Medical Sciences	680	19	2.8	2	0.3
Physics & Mathematics	230	5 (J: 2, P: 3)	2.2	4 (J: 1, P: 3)	1.7
Life Sciences & Earth Sciences	380	2 (J: 1, R: 1)	0.5	2 (J: 1, R: 1)	0.5
Chemical & Material Sciences	170	0	0	0	0

Unless otherwise specified, all missing publications are journal papers
J: journal paper; C: conference paper; P: preprint; R: report

Among the causes why some highly-cited documents were not covered in Web of Science and/or Scopus (Table 2), the most frequent one is that the journal or conference where the document was published was not covered in these databases in 2006, but it started been indexed at a later date (56% of the missing documents in Web of Science, and 49% of the missing documents in Scopus). Web of Science and Scopus do not practice backwards indexing except in special cases like the Emerging Sources Citation Index Backfile for documents published between 2005 and 2014, released on October 2017 and sold separately (Clarivate Analytics, 2017). Thus, documents published in journals before they are selected are missing from the databases.

Table 2. Causes of highly-cited documents not being indexed in Web in Science and/or Scopus

The journal / conference where the document was published…	Web of Science (N = 208)	%	Scopus (N = 87)	%
… was not covered in 2006, but it was added at a later date (no backwards indexing)	117	56	43	49
… was being indexed in 2006, but coverage is incomplete (some volumes or issues are missing)	50	24	12	14
… is not covered by the database	37	18	29	33
The document is not formally published	4	2	4	5

RQ2. To the extent that coverage of highly-cited documents in these databases overlaps, are citation counts in Google Scholar similar in relative terms (rank orders) to those provided by Web of Science and Scopus?

If we focus exclusively in the documents that were covered both by Google Scholar and Web of Science, or by Google Scholar and Scopus, we find that the correlation coefficients are, in both cases, remarkably strong (Table 3).

Table 3. Spearman correlation coefficients of citation counts between Google Scholar and Web of Science, and Google Scholar and Scopus, for highly-cited documents according to Google Scholar published in 2006, by broad subject categories

	GS-WoS		GS-Scopus	
Subject category	**N**	**Spearman corr.**	**N**	**Spearman corr.**
Humanities, Literature & Arts	176	.84	203	.89
Social Sciences	421	.86	466	.91
Engineering & Computer Science	504	.83	556	.92
Business, Economics & Management	141	.89	146	.92
Health & Medical Sciences	661	.94	678	.95
Physics & Mathematics	225	.93	226	.94
Life Sciences & Earth Sciences	378	.97	378	.98
Chemical & Material Sciences	170	.99	170	.99

confidence level: 95%
p-values < 0.0001

The weakest correlations of citation counts between Google Scholar and Web of Science are found in *Engineering & Computer Science* (.83), *Humanities, Literature & Arts* (.84), *Social Sciences* (.86), and *Business, Economics & Management* (.89), but even these are strong. Between Google Scholar and Scopus, correlations are even stronger than between Google Scholar and Web of Science in all cases. The weakest one is also found in the *Humanities, Literature & Arts* (.89). In the rest of the subject categories, the correlations are always above .90, reaching their highest value in *Chemical & Material Sciences* (.99).

RQ3. Which, out of Google Scholar, Web of Science, and Scopus, gives the most citations for highly-cited documents?

Citation counts of highly-cited documents in Google Scholar are higher than citation counts in Web of Science and Scopus in all subject categories (Figure 1). Furthermore, the differences are statistically significant in all subject categories. They are larger in *Business, Economics & Management*, *Social Sciences*, and *Humanities, Literature & Arts*. The smallest difference that involves Google Scholar is found in *Chemical & Material Sciences*, where the lower bound of the 95% confidence interval for Google Scholar citation counts is closest to the higher bound of the confidence intervals for Scopus and Web of Science data.

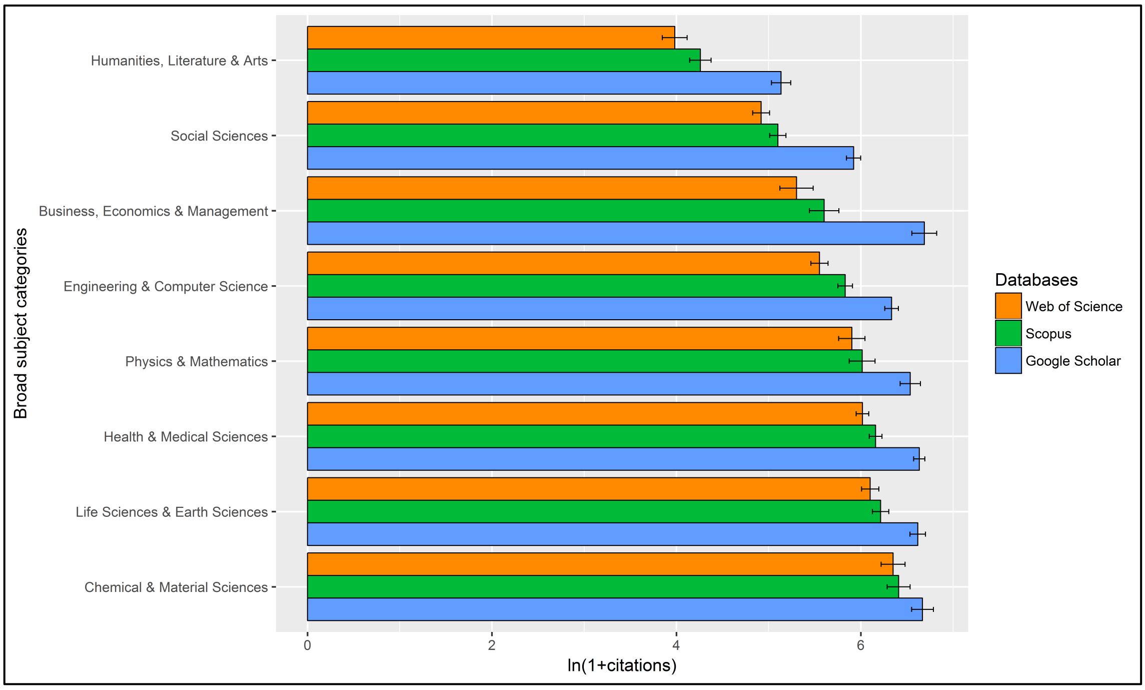


Figure 1. Average log-transformed citation counts of highly-cited documents according to Google Scholar published in 2006, based on data from Google Scholar, Web of Science, and Scopus, by broad subject categories

If we look at the differences between Web of Science and Scopus, we observe that, although the average of log-transformed citation counts is always higher in Scopus, the differences are statistically significant in only 4 out of 8 subject categories: *Engineering & Computer Science*, *Health & Medical Sciences*, *Humanities, Literature & Arts*, and *Social Sciences*. Even in these areas, the confidence intervals are very close to each other.

4. Limitations

Google Scholar's *Classic Papers* dataset suffers from a number of limitations to study highly-cited documents (Delgado López-Cózar et al., 2017). An important limitation is the arbitrary decision to only display the top 10 most cited documents in each subcategory, when it is well-known that the number of documents published in any given year greatly varies across subcategories. Moreover, the dataset only includes documents written in English which presented original research, and published in 2006. Nevertheless, these 10 documents should be well within the limits of the top 10% most cited documents suggested by Bornmann and Marx (2014) to evaluate researchers, even in the subcategories with the smallest output. Further studies could analyze whether similar effects are also found for non-English documents, and documents published in years other than 2006.

For this reason, the set of documents used in this study can be considered as an extremely conservative sample of highly-cited documents. Thus, negative results in our analysis (no missing documents in Web of Science or Scopus), especially in subcategories with a large output, should not be considered conclusive evidence that these databases cover most of the highly-cited documents that exist out there. On the other hand, positive results (missing documents in Web of Science or Scopus) in this highly exclusive set should put into question the suitability of these databases to calculate indicators based on counts of highly-cited documents, especially in some areas.

Another limitation of this study is that, although it analyzes how many highly-cited documents in Google Scholar are not covered by Web of Science and Scopus, it does not carry out the opposite analysis: how many highly-cited documents in Web of Science and Scopus are not covered by Google Scholar. This analysis deserves its own separate study, but as a first approximation, we can consider the results of a recent working paper (Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018) in which a sample of 2.6 million documents covered by Web of Science where searched in Google Scholar. The study found that 97.6% of all articles and reviews in the sample were successfully found in Google Scholar. Also, it is worth noting that this study only searched documents in Google Scholar using their DOI, and made no further efforts to find documents that were not returned by this type of search. Therefore, it is reasonable to believe that most or all the documents covered by Web of Science are also covered by Google Scholar.

5. Discussion and conclusions

The results of this study demonstrate that, even when only journal and conference articles published in English are considered, Web of Science and Scopus do not cover a significant amount of highly-cited documents in the areas of *Humanities, Literature & Arts* (28.2% in Web of Science, 17.1% in Scopus), and *Social Sciences* (17.5% in Web of Science, and 8.6% in Scopus). Additionally, a significant number of documents in *Engineering & Computer Science*, and *Business, Economics & Management* are also invisible to the Web of Science. In the case of Computer Science the cause is that Web of Science did not cover as many conference proceedings as Google Scholar and Scopus, even though this type of publication is an important part of the literature in this field. Therefore, bibliometric indicators based on counts of highly-cited documents that use data from these two databases may be missing a significant amount of relevant information.

Spearman correlation coefficients of citation counts based on Google Scholar and Web of Science, and Google Scholar and Scopus, for the 8 broad subject categories used in this study are remarkably strong: from .83 in *Business, Economics & Management* (GS-WoS), to .99 in *Chemical & Material Sciences* (both GS-WoS, and GS-Scopus). This evidence matches the results found in other studies (Delgado López-Cózar et al., forthcoming; Moed, Bar-Ilan, & Halevi, 2016), and is a step towards dispelling doubts about the possibility that documents that are highly-cited in Google Scholar but are not covered by Web of Science and/or Scopus are merely the product of unreliable citation counting mechanism in the search engine. Therefore, the notion that Google Scholar citation counts are unreliable at the macro level (Bornmann et al., 2009) does not seem to hold anymore. Although coverage of fields such as Chemistry in Google Scholar may have been poor in the past (Orduña-Malea, Martín-Martín, Ayllón, & Delgado López-Cózar, 2016; Vine, 2006), that issue seems to have been solved, as Harzing (2013) already reported, and as this study confirms.

Also, although it is well-known that Google Scholar contains errors, such as duplicate documents and citations, incomplete and incorrect bibliographic information (Delgado López-Cózar et al., forthcoming; Orduna-Malea, Martín-Martín, & Delgado López-Cózar, 2017), and that it is easy to game citation counts because document indexing is not subjected to quality control (Delgado López-Cózar, Robinson-García, & Torres-Salinas, 2014), these issues seem to have no bearing on the overall values of the citation counts of highly-cited documents. Further studies are needed to check whether these correlations hold for larger samples of documents. If that is the case, it would no longer be justified to dismiss Google Scholar’s citation counts as unreliable on account of the bibliographic errors present in this source, at least in macro-level studies.

Lastly, Google Scholar is shown to provide significantly higher citation counts than Web of Science and Scopus in all 8 areas. *Business, Economics & Management*, *Humanities, Literature & Arts*, and *Social Sciences* are the areas where the differences are larger. Previous studies also pointed in this direction (García-Pérez, 2010; Levine-Clark & Gil, 2008; Meho & Yang, 2007; Mingers & Lipitakis, 2010). This indirectly points to the existence of a much larger document base in Google Scholar for these areas of research, and provides a reasonable explanation for the weaker Spearman correlation coefficients of citation counts in these areas. Further studies could focus on identifying the sources of the citing documents. Some studies have already analysed citing documents (sources, document types, languages, unique citations) in Google Scholar and compared them to the citations found by Web of Science and Scopus (Bar-Ilan, 2010; de Winter, Zadpoor, & Dodou, 2013; Kousha & Thelwall, 2008; Meho & Yang, 2007; Rahimi & Chandrakumar, 2014). These studies reported that after journal articles, a large proportion of the citations found only by Google Scholar came from conference papers, dissertations, books, and book chapters. However, these studies focused on specific case studies, and most of them were carried out more than five years ago. Therefore, an updated, in-depth, multi-discipline analysis of the sources of citations in Google Scholar (that examines aspects such as document types, languages, peer-review status…), as compared to other citation databases like Web of Science and Scopus is now warranted, and could further elucidate the suitability of each platform as sources of data for different kinds of bibliometric analyses.

All this evidence points to the conclusion that inclusive databases like Google Scholar do indeed have a better coverage of highly-cited documents in some areas of research than Web of Science (*Humanities, Literature & Arts*, *Social Sciences*, *Engineering & Computer Science*, and *Economics & Management*) and Scopus (*Humanities, Literature & Arts*, and *Social Sciences*). Therefore, using these selective databases to compute bibliometric indicators based on counts of highly-cited documents might produce biased assessments in those poorly covered areas. In the other areas (*Health & Medical Sciences*, *Physics & Mathematics*, *Life Sciences & Earth Sciences*, *Chemical & Material Sciences*) all three databases seem to have similar coverage and citation data, and therefore the selective or inclusive nature of the database in these areas does not seem to make a difference in the calculation of indicators based on counts of highly-cited documents.

Google Scholar seems to contain useful bibliographic and citation data in the areas where coverage of Web of Science and Scopus is deficient. However, although there is evidence that it is possible to use Google Scholar to identify highly-cited documents (Martin-Martin, Orduna-Malea, Harzing, & Delgado López-Cózar, 2017), there are other practical issues that may discourage the choice of this source: lack of detailed metadata (for example, author affiliations, funding acknowledgements are not provided), or difficulty to extract data caused by the lack of an API (Else, 2018). As is often the case, the choice of data source presents a trade-off (Harzing, 2016). The suitability of each database (selective or inclusive) therefore depends on the specific requirements of each bibliometric analysis, and it is important that researchers planning to carry out these analyses are aware of these issues before making their choices, because these assessments often have direct consequences on the careers of individual researchers (hiring, promotion, or funding decisions) or institutions (university rankings).

Acknowledgements

Alberto Martín-Martín enjoys a four-year doctoral fellowship (FPU2013/05863) granted by the Ministerio de Educación, Cultura, y Deportes (Spain).

References

Acharya, A. (2015, September 21). What happens when your library is worldwide and all articles are easy to find? Retrieved from https://youtu.be/S-f9MjQjLsk?t=7m9s

Acharya, A., Verstak, A., Suzuki, H., Henderson, S., Iakhiaev, M., Lin, C. C. Y., & Shetty, N. (2014). Rise of the Rest: The Growing Impact of Non-Elite Journals. Retrieved from http://arxiv.org/abs/1410.2217

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingrasb, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, *68*(3), 329–342. https://doi.org/10.1007/s11192-006-0115-z

Bar-Ilan, J. (2008). Which h-index? — A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, *74*(2), 257–271. https://doi.org/10.1007/s11192-008-0216-y

Bar-Ilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics*, *82*(3), 495–506. https://doi.org/10.1007/s11192-010-0185-9

Bornmann, L., & Leydesdorff, L. (2018). Count highly-cited papers instead of papers with h citations: use normalized citation counts and compare "like with like"! *Scientometrics*, *115*(2), 1119–1123. https://doi.org/10.1007/s11192-018-2682-1

Bornmann, L., & Marx, W. (2014). How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics*, *98*(1), 487–509. https://doi.org/10.1007/s11192-013-1161-y

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A., & Daniel, H.-D. (2009). Convergent validity of bibliometric Google Scholar data in the field of chemistry—Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, S. *Journal of Informetrics*, *3*(1), 27–35. https://doi.org/10.1016/j.joi.2008.11.001

Chavarro, D., Ràfols, I., & Tang, P. (2018). To what extent is inclusion in the Web of Science an indicator of journal 'quality'? *Research Evaluation*, *27*(2), 106–118. https://doi.org/10.1093/reseval/rvy001

Clarivate Analytics. (2017). Emerging Sources Citation Index Backfile (2005-2014). Retrieved from https://clarivate.com/wp-content/uploads/2017/10/M255-Crv_SAR_ESCI-infographic-FA.pdf

De Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. https://doi.org/10.1002/asi.4630270505

de Winter, J. C. F., Zadpoor, A. A., & Dodou, D. (2013). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics*, *98*(2), 1547–1565. https://doi.org/10.1007/s11192-013-1089-2

Delgado López-Cózar, E., Martín-Martín, A., & Orduna-Malea, E. (2017). *Classic papers: déjà vu, a step further in the bibliometric exploitation of Google Scholar* (EC3's Working Papers No. 24). Retrieved from https://arxiv.org/abs/1706.09258

Delgado López-Cózar, E., Orduna-Malea, E., & Martín-Martín, A. (2019). Google Scholar as a data source for research assessment. In W. Glaenzel, H. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*. Springer.

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446–454. https://doi.org/10.1002/asi.23056

Else, H. (2018, April 11). How I scraped data from Google Scholar. *Nature*. https://doi.org/10.1038/d41586-018-04190-5

Frandsen, T. F., & Nicolaisen, J. (2008). Intradisciplinary differences in database coverage and the consequences for bibliometric research. *Journal of the American Society for Information Science and Technology*, *59*(10), 1570–1581. https://doi.org/10.1002/asi.20817

García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in Psychology. *Journal of the American Society for Information Science and Technology*, *61*(10), 2070–2085. https://doi.org/10.1002/asi.21372

Halevi, G., Moed, H., & Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the Literature. *Journal of Informetrics*, *11*(3), 823–834. https://doi.org/10.1016/J.JOI.2017.06.005

Harzing, A.-W. (2013). A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. *Scientometrics*, *94*(3), 1057–1075. https://doi.org/10.1007/s11192-012-0777-7

Harzing, A.-W. (2016). Sacrifice a little accuracy for a lot more comprehensive coverage. Retrieved from https://harzing.com/blog/2016/08/sacrifice-a-little-accuracy-for-a-lot-more-comprehensive-coverage

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, *74*(2), 273–294. https://doi.org/10.1007/s11192-008-0217-x

Levine-Clark, M., & Gil, E. L. (2008). A Comparative Citation Analysis of Web of Science, Scopus, and Google Scholar. *Journal of Business & Finance Librarianship*, *14*(1), 32–46. https://doi.org/10.1080/08963560802176348

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents. *Journal of the American Society for Information Science and Technology*, *62*(7), 1370–1381. https://doi.org/10.1002/asi.21534

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). *Evidence of Open Access of scientific publications in Google Scholar: a large-scale analysis*. https://doi.org/10.17605/osf.io/k54uv

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2014). *Does Google Scholar contain all highly cited documents (1950-2013)?* (EC3 Working Papers No. 19). Retrieved from http://arxiv.org/abs/1410.8464

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentacion Cientifica*, *39*(4), e149. https://doi.org/10.3989/redc.2016.4.1405

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). Data and code for: Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. https://doi.org/10.17605/OSF.IO/DNQZK

Martin-Martin, A., Orduna-Malea, E., Harzing, A.-W., & Delgado López-Cózar, E. (2017). Can we use Google Scholar to identify highly-cited documents? *Journal of Informetrics*, *11*(1), 152–163. https://doi.org/10.1016/j.joi.2016.11.008

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105–2125. https://doi.org/10.1002/asi.20677

Mingers, J., & Lipitakis, E. A. E. C. G. (2010). Counting the citations: a comparison of Web of Science and Google Scholar in the field of business and management. *Scientometrics*, *85*(2), 613–625. https://doi.org/10.1007/s11192-010-0270-0

Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics*, *10*(2), 533–551. https://doi.org/10.1016/j.joi.2016.04.017

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-

015-1765-5

Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, *104*(3), 931–949. https://doi.org/10.1007/s11192-015-1614-6

Orduña-Malea, E., Martín-Martín, A., Ayllón, J. M., & Delgado López-Cózar, E. (2016). *La revolución Google Scholar : Destapando la caja de Pandora académica*. Granada: Universidad de Granada.

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2017). Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors. *Revista Española de Documentación Científica*, *40*(4), e185. https://doi.org/10.3989/redc.2017.4.1500

Rahimi, S., & Chandrakumar, V. (2014). A comparison of citation coverage of traditional and web citation databases in medical science. *Malaysian Journal of Library and Information Science*, *19*(3), 1–11. Retrieved from http://jice.um.edu.my/index.php/MJLIS/article/view/1779

Sud, P., & Thelwall, M. (2014). Evaluating altmetrics. *Scientometrics*, *98*(2), 1131–1143. https://doi.org/10.1007/s11192-013-1117-2

Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. *Journal of Informetrics*, *11*(1), 128–151. https://doi.org/10.1016/j.joi.2016.12.002

Thelwall, M., & Fairclough, R. (2017). The accuracy of confidence intervals for field normalised indicators. *Journal of Informetrics*, *11*(2), 530–540. https://doi.org/10.1016/j.joi.2017.03.004

van Leeuwen, T. N., Moed, H. F., Tijssen, R. J. W., Visser, M. S., & Van Raan, A. F. J. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics*, *51*(1), 335–346. https://doi.org/10.1023/A:1010549719484

Vine, R. (2006). Google Scholar. *Journal of the Medical Library Association*, *94*(1), 97. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1324783/

Appendix A. Top 5 most common subject categories assigned by Web of Science and Scopus to highly-cited documents in Google Scholar, by Google Scholar broad subject categories

Google Scholar category: **Humanities, Literature & Arts**		Google Scholar category: **Social Sciences**	
Web of Science categories (176 docs.)	Scopus categories (203 docs.)	Web of Science categories (421 docs.)	Scopus categories (466 docs.)
Area Studies (24) Linguistics (21) Psychology (18) Literature (17) Social Sciences – Other Topics (16)	Arts and Humanities (138) Social Sciences (127) Psychology (17) Economics, Econometrics and Finance (11) Medicine (7)	Psychology (58) Education & Educational Research (57) Business & Economics (56) Government & Law (48) Social Sciences – Other Topics (32)	Social Sciences (285) Arts and Humanities (97) Medicine (76) Psychology (69) Economics, Econometrics and Finance (49)
Google Scholar category: **Business, Economics & Management**		Google Scholar category: **Engineering & Computer Science**	
Web of Science categories (141 docs.)	Scopus categories (146 docs.)	Web of Science categories (504 docs.)	Scopus categories (556 docs.)
Business & Economics (113) Social Sciences – Other Topics (20) Public Administration (12) Environmental Sciences & Ecology (9) Science & Technology – Other Topics (6)	Business, Management and Accounting (87) Economics, Econometrics and Finance (70) Social Sciences (36) Arts and Humanities (12) Decision Sciences (10)	Engineering (217) Computer Science (145) Materials Science (56) Chemistry (52) Science & Technology – Other Topics (44)	Engineering (223) Computer Science (158) Materials Science (72) Chemical Engineering (65) Social Sciences (61)
Google Scholar category: **Physics & Mathematics**		Google Scholar category: **Health & Medical Sciences**	
Web of Science categories (225 docs.)	Scopus categories (226 docs.)	Web of Science categories (661 docs.)	Scopus categories (678 docs.)
Physics (74) Mathematics (73) Science & Technology – Other Topics (31) Engineering (21) Mechanics (17)	Physics and Astronomy (97) Mathematics (89) General (31) Engineering (27) Computer Science (25)	General & Internal Medicine (170) Science & Technology – Other Topics (80) Surgery (53) Neurosciences & Neurology (36) Psychology (24)	Medicine (482) General (80) Biochemistry, Genetics and Molecular Biology (73) Social Sciences (32) Nursing (32)
Google Scholar category: **Life Sciences & Earth Sciences**		Google Scholar category: **Chemical & Material Sciences**	
Web of Science categories (378 docs.)	Scopus categories (378 docs.)	Web of Science categories (170 docs.)	Scopus categories (170 docs.)
Science & Technology – Other Topics (122) Environmental Sciences & Ecology (51) Biochemistry & Molecular Biology (48) Agriculture (37) Cell Biology (27)	Agricultural and Biological Sciences (122) General (118) Biochemistry, Genetics and Molecular Biology (89) Environmental Science (61) Medicine (40)	Chemistry (75) Science & Technology – Other Topics (34) Materials Science (31) Biochemistry & Molecular Biology (19) Physics (18)	Chemistry (85) Biochemistry, Genetics and Molecular Biology (53) Chemical Engineering (48) Materials Science (40) General (29)

Chapter 7. Google Scholar, Web of Science, and Scopus: a systematic comparison of citations in 252 subject categories

Journal Article. Cite as:

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. Journal of Informetrics, 12(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

Abstract (English)

Despite citation counts from Google Scholar (GS), Web of Science (WoS), and Scopus being widely consulted by researchers and sometimes used in research evaluations, there is no recent or systematic evidence about the differences between them. In response, this paper investigates 2,448,055 citations to 2,299 English-language highly-cited documents from 252 GS subject categories published in 2006, comparing GS, the WoS Core Collection, and Scopus. GS consistently found the largest percentage of citations across all areas (93%-96%), far ahead of Scopus (35%-77%) and WoS (27%-73%). GS found nearly all the WoS (95%) and Scopus (92%) citations. Most citations found only by GS were from non-journal sources (48%-65%), including theses, books, conference papers, and unpublished materials. Many were non-English (19%-38%), and they tended to be much less cited than citing sources that were also in Scopus or WoS. Despite the many unique GS citing sources, Spearman correlations between citation counts in GS and WoS or Scopus are high (0.78-0.99). They are lower in the Humanities, and lower between GS and WoS than between GS and Scopus. The results suggest that in all areas GS citation data is essentially a superset of WoS and Scopus, with substantial extra coverage.

Abstract (Spanish)

A pesar de que los conteos de citas de Google Scholar (GS), Web of Science (WoS), y Scopus son ampliamente utilizados por los investigadores, y a veces se usan en evaluación científica, no hay ninguna evidencia reciente o sistemática de las diferencias entre ellos. En respuesta, este artículo investiga 2.448.055 citas a 2.999 documentos en inglés altamente citados de 252 categorías temáticas, publicados en 2006. Se comparan sus citas en GS, WoS colección principal, y Scopus. GS encontró consistentemente los mayores porcentajes de citas en todas las áreas (93%-96%), muy por delante de Scopus (35%-77%) y WoS (27%-74%). GS encontró la gran mayoría de las citas encontradas por WoS (95%) y Scopus (92%). La mayoría de las citas que solo encontró GS venían de fuentes que no eran revistas (48%-65%), incluyendo tesis, libros, comunicaciones a congresos, y material no publicado. Muchas no estaban en inglés (19%-38%), y tendían a ser mucho menos citadas que los documentos citantes que también estaban indizados en Scopus o en WoS. A pesar de las muchas citas únicas encontradas por GS, las correlaciones Spearman entre los conteos de citas de GS y WoS, y GS y Scopus son altas (0,78-0,99). Son más bajas en las Humanidades, y más bajas entre GS y WoS que entre GS y Scopus. Los resultados sugieren que en todas las áreas los datos de citas de GS son básicamente un superconjunto de los datos disponibles en WoS y Scopus, con una sustancial cobertura extra.

1. Introduction

The launch of Google Scholar (GS) in November of 2004 brought the simplicity of Google searches to the academic environment, and revolutionized the way researchers and the public searched, found, and accessed academic information. Until that point, the coverage of academic databases depended on lists of selected sources (usually scientific journals). In contrast, and using automated methods, Google Scholar crawled the web and indexed any document with a seemingly academic structure. This inclusive approach gave GS potentially more comprehensive coverage of the scientific and scholarly literature compared to the two major existing multidisciplinary databases with selective journal-based inclusion policies, the Web of Science (WoS) and Scopus (Orduna-Malea, Ayllón, Martín-Martín, & Delgado López-Cózar, 2015).

Although citation data in Google Scholar was originally intended to be a means of identifying the most relevant documents for a given query, it could also be used for formal or informal research evaluations. The availability of free citation data in Google Scholar, together with the free software *Publish or Perish* (Harzing, 2007) to gather it made citation analysis possible without a citation database subscription (Harzing & van der Wal, 2008). Nevertheless, GS has not enabled bulk access to its data, reportedly because their agreements with publishers preclude it (Van Noorden, 2014). Thus, third-party web-scraping software is currently the only practical way to extract more data from GS than permitted by Publish or Perish.

Despite its known errors and limitations, which are consequence of its automated approach to document indexing (Delgado López-Cózar, Robinson-García, & Torres-Salinas, 2014; Jacsó, 2010), GS has been shown to be reliable and to have good coverage of disciplines and languages, especially in the Humanities and Social Sciences, where WoS and Scopus are known to be weak (Chavarro, Ràfols, & Tang, 2018; Mongeon & Paul-Hus, 2016; van Leeuwen, Moed, Tijssen, Visser, & Van Raan, 2001). Analyses of the coverage of GS, WoS, and Scopus across disciplines have compared the numbers of publications indexed or their average citation counts for samples of documents, authors, or journals, finding that GS consistently returned higher numbers of publications and citations (Harzing, 2013; Harzing & Alakangas, 2016; Mingers & Lipitakis, 2010; Prins, Costas, van Leeuwen, & Wouters, 2016). Citation counts from a range of different sources have been shown to correlate positively with GS citation counts at various levels of aggregation (Amara & Landry, 2012; De Groote & Raszewski, 2012; Delgado López-Cózar, Orduna-Malea, & Martín-Martín, 2018; Kousha & Thelwall, 2007; Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018; Meho & Yang, 2007; Minasny, Hartemink, McBratney, & Jang, 2013; Moed, Bar-Ilan, & Halevi, 2016; Pauly & Stergiou, 2005; Rahimi & Chandrakumar, 2014; Wildgaard, 2015). See the supplementary materials[21], Delgado López-Cózar et al. (2018), Orduña-Malea, Martín-Martín, Ayllón, & Delgado López-Cózar (2016), and Halevi, Moed, & Bar-Ilan (2017) for discussions of the wider strengths and weaknesses of GS.

A key issue is the ability of GS, WoS, and Scopus to find citations to documents, and the extent to which they index citations that the others cannot find. The results of prior studies are confusing, however, because they have examined different small (with one exception) sets of articles. A summary of the results found in these previous studies is presented in Table 1. For example, the number of citations that are unique to GS varies between 13% and 67%, with the differences probably being due to the study year or the document types or disciplines covered. The only multidisciplinary study (Moed et al., 2016) checked articles in 12 journals from 6 subject areas, which is still a limited set.

[21] Supplementary materials available from https://dx.doi.org/10.31235/osf.io/pqr53

Table 1. Results of studies that analysed unique and overlapping citations in GS, WoS, and Scopus

Study	**Sample**	**N citations**	**% only GS**	**% only WoS**	**% only Scopus**	**% only GS & WoS**	**% only GS & Scopus**	**% only WoS & Scopus**	**% GS & WoS & Scopus**	**% GS (all cit.)**	**% WoS (all cit.)**	**% Scopus (all cit.)**	**% WoS cit. in GS**	**% Scopus cit. in GS**
Bakkalbasi, Bauer, Glover, & Wang (2006)	50 journal articles covered in JCR Oncology	614	13	7	12	4	5	28	31	53	70	76	215/431 = **50%**	220/469 = **47%**
	50 journal articles covered in JCR Physics, Cond. Matter	296	17	20	8	9	3	22	21	50	72	54	84/212 = **40%**	72/162 = **44%**
Yang & Meho (2007)	Scientific production of two Library & Information Science (LIST) researchers	385	10	23	6	10	7	18	25	52	77	57	137/295 = **46%**	124/218 = **57%**
Meho & Yang (2007)	1,457 articles published by 25 LIS researchers	5,285	48	Only (WoS or Scopus): 21		GS-(WoS or Scopus): 31		NA	NA	79	38	44	% (WoS or Scopus) cit. in GS 1,629/2,733 = **60%**	
Kousha & Thelwall (2008)	262 WoS-covered Biology journal articles	1,554	17	28	NA	55	NA			72	83	NA	847/1288 = **66%**	NA
	276 WoS-covered Chemistry journal articles	729	8	62		30				38	92		218/668 = **33%**	
	262 WoS-covered Physics journal articles	1,734	36	24		40				76	64		690/1111 = **62%**	
	82 WoS-covered Computing journal articles	3,369	67	14		19				86	33		632/1117 = **57%**	
	Total WoS-covered journal articles (882)	7,386	43	24		32				76	57		2387/4184 = **57%**	
Jacimovic, Petrovic, & Zivkovic (2010)	158 articles published in Serbian Dental Journal	249	58	4	6	1	2	15	15	76	34	39	39/85 = **46%**	43/94 = **46%**
Bar-Ilan (2010)	Book "Introduction to Informetrics" by L. Egghe and R. Rousseau	397	27	12	2	6	5	9	39	77	66	55	177/259 = **68%**	174/218 = **80%**
Lasda Bergman (2012)	5 top journals in the field of Social Work	4,308	44	5	8	2	8	12	22	76	41	50	1042/1741 = **60%**	1285/2126 = **60%**
de Winter, Zadpoor, & Dodou (2014)	Garfield, E. (1955). Citation indexes for science. *Science*, 122(3159), 108-111.	1,309	33	41	NA	35	NA			68	76	NA	453/606 = **75%**	NA
Rahimi & Chandrakumar (2014)	2,082 WoS-covered articles in General and Internal Medicine	62,900	29	10	11	2	9	8	31	71	51	59	20532/31778 = **65%**	25180/37272 = **68%**
Moed, Bar-Ilan, & Halevi (2016)	Articles published in 12 journals from 6 subject areas	6,941	47	NA	6	NA	47	NA	NA	94	NA	53	NA	3246/3651 = **89%**

NA = not analysed in the study

Cells with more intense background color represent higher percentages of citations within the same sample of documents.

The fields previously compared for citation sources (Table 1) are Library and Information Science (5 out of 10 articles analyse case studies about LIS documents/journals/researchers), Medicine (3 papers, analysing oncology, general medicine, and dentistry), Physics (2 articles: general and condensed matter), Chemistry (2 articles: general and inorganic), Computer Science (2 articles: general, and computational linguistics), Biology (2 articles: general, and virology), Social Work, Political Science, and Chinese Studies (1 article each). From this list it is clear that most academic fields have not been analysed for Google Scholar coverage. The studies used small samples of documents and citations (9 out of 10 papers analysed less than 10,000 citations), probably because of the difficulty of extracting data from GS, caused by the lack of a public API (Else, 2018; Van Noorden, 2014). Moreover, the most recent data in these studies was collected in 2015 (three years before the current study), and the oldest data is from 2005 (13 years ago).

Given the limited nature of all prior studies of citing sources for GS and the need to update all previous research, a comprehensive analysis of citation sources in GS, WoS, and Scopus across all subject areas is needed. This information is important for those deciding whether to use GS citation counts for informal or formal research evaluations. The following research questions drive this investigation.

RQ1. How much overlap is there between GS, WoS, and Scopus in the citations that they find to academic documents and does this vary by subject?

RQ2. Do the citing documents that are only found by GS have a different type to non-unique GS citations, and does this vary by subject?

RQ3. How similar are citation counts in GS to those found in WoS and Scopus, at the level of subjects?

2. Methods

The sample used for this study is taken from GS's *Classic Papers* product (GSCP)[22]. The 2017 edition of GSCP lists 2,515 highly-cited documents written in English and published in 2006[23]. These documents were classified by GS into 252 subject categories within 8 broad subject areas. Background about GSCP can be found in Orduna-Malea, Martín-Martín, & Delgado López-Cózar (2018) and Martín-Martín, Orduna-Malea, & Delgado López-Cózar (2018). This gives a large sample of highly cited documents classified by subject. This is not a random sample of academic publications because there is no complete list of these. There is also not a complete list of documents in GS.

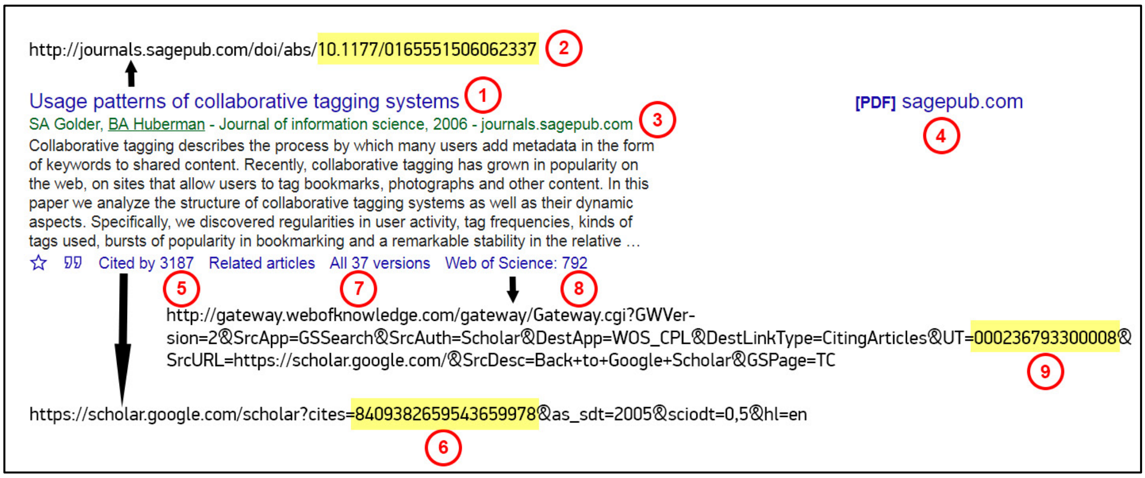
The GSCP sample is suitable because it covers all subject areas and, because the articles are classified, allows analyses by subject categories. GSCP and Google Scholar Metrics[24] (GSM) are the only products where GS provides a subject categorization. Taking a sample from one of the three sources to be compared (GS, Scopus, WoS) is not ideal because it is likely to bias the results in favour of GS. Nevertheless, the inclusion of 252 categories minimizes the chance of bias due to a subject area that is not well covered by GS. GS is also a better source than WoS or Scopus because of its more comprehensive coverage, as found by most prior studies.

2.1. Extraction of data from Google Scholar

The citations to each of the 2,515 GSCP documents were extracted from GS, WoS, and Scopus between April 22nd and May 6th, 2018. A custom script scraped all the relevant information from GS SERPs (Search Engine Results Pages) (Figure 1). Searches were submitted from Universidad de Granada IP addresses to access the additional information displayed in GS for WoS subscribers (Clarivate Analytics, 2015). CAPTCHAs were solved manually when GS

[22] https://scholar.google.com/citations?view_op=list_classic_articles&hl=en&by=2006
[23] https://osf.io/5zmk7/
[24] https://scholar.google.com/citations?view_op=top_venues&hl=en

requested them. This process found 2,415,072 citations in Google Scholar[25] to the 2,515 highly-cited documents. The number of citations is reduced to 2,301,997 for the 2,299 highly-cited documents also covered by WoS and Scopus.



http://journals.sagepub.com/doi/abs/10.1177/0165551506062337

Usage patterns of collaborative tagging systems

SA Golder, BA Huberman - Journal of information science, 2006 - journals.sagepub.com

Collaborative tagging describes the process by which many users add metadata in the form of keywords to shared content. Recently, collaborative tagging has grown in popularity on the web, on sites that allow users to tag bookmarks, photographs and other content. In this paper we analyze the structure of collaborative tagging systems as well as their dynamic aspects. Specifically, we discovered regularities in user activity, tag frequencies, kinds of tags used, bursts of popularity in bookmarking and a remarkable stability in the relative …

☆ 99 Cited by 3187 Related articles All 37 versions Web of Science: 792

[PDF] sagepub.com

http://gateway.webofknowledge.com/gateway/Gateway.cgi?GWVersion=2&SrcApp=GSSearch&SrcAuth=Scholar&DestApp=WOS_CPL&DestLinkType=CitingArticles&UT=000236793300008&SrcURL=https://scholar.google.com/&SrcDesc=Back+to+Google+Scholar&GSPage=TC

https://scholar.google.com/scholar?cites=8409382659543659978&as_sdt=2005&sciodt=0,5&hl=en

Figure 1. Metadata extracted from Google Scholar

1. Title of the document.
2. URL embedded in title of the document. The DOI of the document is sometimes embedded in this URL (depending on the host)
3. Authors, publication venue, publication year, and publisher or web domain that hosts the document.
4. URL to the freely accessible full text of the document, when available.
5. Times cited according to GS.
6. URL pointing to list of citing documents according to GS. GS's internal ID for the document is embedded in this URL.
7. Number of versions of the document found by GS.
8. Times Cited according to WoS (when the document is also covered by WoS).
9. URL pointing to list of citing documents in WoS. WoS's internal ID (UT) for the document is embedded in this URL.

The data was processed to clean and enrich the limited metadata available in GS, as follows.

- DOI were detected for as many citing documents as possible. The following techniques were used, retrieving 1,501,178 DOIs (62%).
    - Extracted from URLs for publishers like Wiley, Springer, and SAGE which embed the DOI in the article's landing page URL (Figure 1, #2).
    - Looked up with public APIs offered by the publishers (Elsevier, IEEE) or CrossRef[26] (using the *alternative-id* filter option), when the publisher landing page contained publisher document ID.
    - Extracted from a HTML Meta tag in the webpage from which Google Scholar extracted the document's metadata.
- Metadata was obtained from CrossRef and DataCite APIs when a DOI was available or otherwise from HTML Meta tags present in the website hosting the citation, when possible.

2.2. Extraction of data from Web of Science and Scopus

Each of the 2,515 highly-cited documents in GSCP was searched for in the WoS (Core Collection) web interface. The list of citations to each document was extracted (in batches of up to 500 records per download). The exported files were consolidated into a single table using a set of R functions developed for this purpose (Martín-Martín & Delgado López-Cózar, 2016). Although R has built-in functions and additional libraries to read tabulated data, none of them seemed to work

[25] https://osf.io/qg8kb/
[26] https://www.crossref.org/services/metadata-delivery/rest-api/

with data exported from WoS. A total of 1,270,225 WoS records were collected[27]. At the time of data collection FECYT[28], the Spanish organization that manages the national subscription to Clarivate Analytics' services, had not subscribed to the Emerging Sources Citation Index (ESCI) Backfile for documents published between 2005 and 2014 (Clarivate Analytics, 2017), and so the results exclude this source.

Each of the 2,515 highly-cited documents in GSCP were also searched for in the Scopus web interface. This has a limit of 2,000 records when exporting citations. When a highly-cited document had more than 2,000 citations, these could still be extracted using the alternative email service, which allows the extraction of up to 20,000 citation records in one go. A total 1,515,436 Scopus records were collected[29].

Most of the highly-cited documents (2,299 out of 2,515) were covered by all three databases, and the citations to these 2,299 documents are analysed here.

2.3. Identification of document types and languages of citing documents

Unlike WoS and Scopus, GS does not provide metadata on the document type and the language of the documents that it covers. The metadata extracted from CrossRef's API and HTML Meta tags of the hosting website gave this information for 83% of the citing documents. Adding metadata from WoS and Scopus increased this percentage to 85%. The following categories were used.

- Journal publication: article, review, letter, editorial…
- Conference paper: paper presented at conference, symposium, workshop, society meeting…
- Book or book chapter: scientific/scholarly monograph
- Thesis or dissertation: document presented by student to fulfill the requirements of a doctoral, masters', or bachelor's degree
- Other not-formally-published scientific/scholarly paper: working paper, discussion paper, other paper for which no formal publication venue could be found.
- Other: report, patent, presentation slides, syllabus, educational materials, errata…
- Unknown: document for which no document type could be identified

To identify the distribution of document types in the 15% for which metadata was not available, eight random samples of 500 citing documents with an unknown document type were selected, one for each of the broad subject categories in which GSCP are classified. The document types of these 4,000 citing documents were manually identified by accessing and perusing the full text of the documents (when possible) or the available metadata. The proportion of document types found in these random samples were applied as a correction factor to the percentage of citations with an unknown document type in each broad subject area. For example, in the Social Sciences, 33.5% of the citing documents were classified as journal articles using the available metadata, but 20% of all citing documents could not be classified with the available metadata. A random sample of documents from that unknown 20% were selected and analyzed manually, finding that 27.6% of the items in the random sample were journal articles. Therefore, the total percentage of journal articles in Social Sciences was 33.5% + (27.6% of 20% = 5.5%) = 39%.

The language of 98% of the citing documents was identified by combining data from three sources (in the order of preference shown below).

1. Metadata in CrossRef and HTML Meta tags.
2. Metadata in WoS (Scopus did not provide document language information).

[27] https://osf.io/6c7ta/

[28] https://www.fecyt.es/

[29] https://osf.io/n6k9w/

3. Google's Compact Language Detector 2[30] applied to the document title.

For RQ1, the citations extracted from GS, WoS and Scopus were matched as follows. Three pairwise matching processes were carried out: GS–Scopus; GS–WoS; and Scopus–WoS.

1. For each pair of databases *A* and *B*, and a highly-cited document from GSCP *X*, all citing documents with a DOI that cite *X* according to *A* where matched to all citing documents with a DOI that cite *X* according to *B*.
2. For each of the unmatched documents citing *X* in *A* and *B*, a further comparison was carried out. The title of each unmatched document citing *X* in *A* was compared to the titles of all the unmatched document citing *X* in *B*, using the restricted Damerau-Levenshtein distance (optimal string alignment) (Damerau, 1964; Levenshtein, 1966). The pair of citing documents which returned the highest title similarity (1 is perfect similarity) was selected as potential matches. This match was considered successful if either of the following conservative heuristics was met.
   - The title similarity was at least 0.8, and the citing document title was at least 30 characters long (to avoid matches between titles like "Introduction").
   - The title similarity was at least 0.7, and the first author of the citing document was the same in *A* and *B*.

For RQ2, the document types, languages, and citation counts of the citing documents in our sample (see Figure 2) were aggregated or averaged by GSCP broad subject areas, differentiating between unique GS citations and overlapping citations.

For RQ3, Spearman correlation coefficients were calculated for the citation counts of the citing documents in our sample (GS-WoS, and GS-Scopus), by subject category. Correlation coefficients are considered useful in high-level exploratory analyses to check whether different indicators reflect the same underlying causes (Sud & Thelwall, 2014). In this case, however, the goal is to find out whether the same indicator, based on different data sources, provides similar relative values. Spearman correlations were used because it is well-known that the distributions of citation counts and other impact-related metrics are highly skewed (De Solla Price, 1976). For the GS-WoS comparison, WoS subject categories and (for an additional check) the NOWT classification (Tijssen et al., 2010) were used. For the GS-Scopus comparison, the ASJC (All Science Journal Classification) available in the Scopus source list (Elsevier, 2018) was used.

To carry out all these processes, the R programming language (R Core Team, 2014), and several R packages and custom functions were used (Dowle et al., 2018; Larsson et al., 2018; Martín-Martín & Delgado López-Cózar, 2016; Ooms & Sites, 2018; van der Loo, van der Laan, R Core Team, Logan, & Muir, 2018; Walker & Braglia, 2018; Wickham, 2016). The resulting data files are openly available[31].

[30] https://github.com/CLD2Owners/cld2

[31] https://osf.io/gnb72/

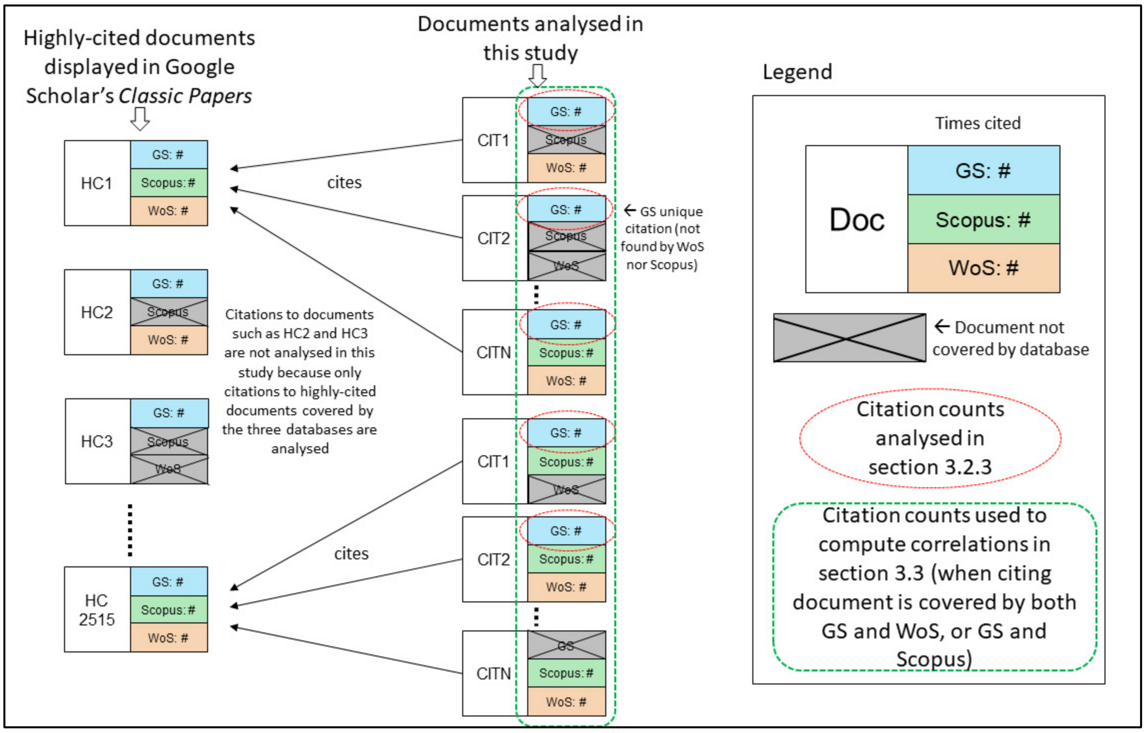

Highly-cited documents displayed in Google Scholar's *Classic Papers*
Documents analysed in this study
Legend
HC1
GS: #
Scopus: #
WoS: #
HC2
HC3
HC 2515
cites
CIT1
CIT2
CITN
← GS unique citation (not found by WoS nor Scopus)
Citations to documents such as HC2 and HC3 are not analysed in this study because only citations to highly-cited documents covered by the three databases are analysed
Times cited
Doc
← Document not covered by database
Citation counts analysed in section 3.2.3
Citation counts used to compute correlations in section 3.3 (when citing document is covered by both GS and WoS, or GS and Scopus)

Figure 2. Visual representation of the documents and citation counts analysed in this study

3. Results

3.1. RQ1: Citing source overlap

Overall, 46.9% of all citations were found by the three databases (Figure 3). GS found the most citations, including most of the citations found by WoS and Scopus. In contrast, only 6% of all citations were found by WoS and/or Scopus, and not by GS. An additional 10.2% of all citations were found by both GS and Scopus (7.7%), or GS and WoS (2.5%). Over a third (36.9%) of all citations were only found by GS.

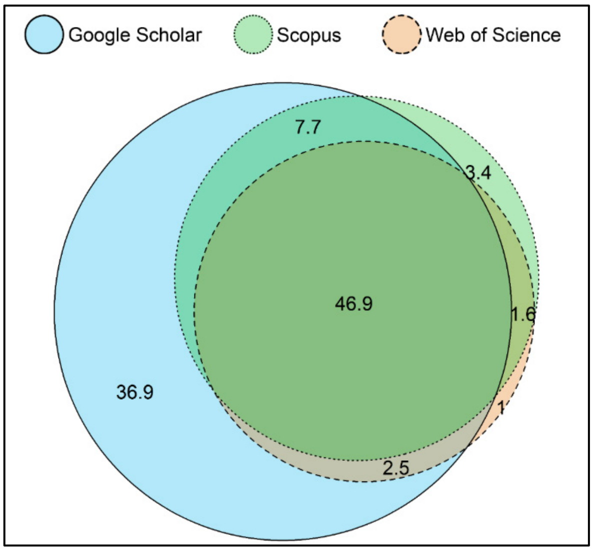


Figure 3. Percentage of unique and overlapping citations in google scholar, Scopus, and Web of Science.
n = 2,448,055 citations from all subject areas

When citations are disaggregated by the broad subject area in which the cited document was classified according to GSCP, important differences emerge (Figure 4). In *Humanities, Literature & Arts*, *Social Sciences*, and *Business, Economics & Management* the proportion of unique GS citations is well over 50% of all citations, surpassing 60% in the case of *Business, Economics & Management*. In these categories the proportion of citations found by all three databases ranges from 21.4% (*Humanities, Literature & Arts*) to 29.8% (*Social Sciences*). On the other hand, in *Engineering & Computer Science*, *Physics & Mathematics*, *Health & Medical Sciences*, *Life Sciences & Earth Sciences*, and *Chemical & Material Sciences*, the proportion of unique GS citations is much lower (20.3% - 34.3%), and the overlap is higher: percentages of citations found by all three databases range from 46.8% (*Engineering & Computer Science*) to 67.7% (*Chemical & Material Sciences*).

For the 252 specific subject categories (data and figures for each category are available in the supplementary materials[32]), there are more extreme differences (Figure 5). The highest percentages of unique citations in GS (over 70% of all citations) are found in *Educational Administration*[33], *Foreign Language Learning*[34], *Chinese Studies & History*[35], and *Finance*[36]. On the other hand, the highest percentages of overlap in the three databases (over 70% of all citations) are found in *Crystallography & Structural Chemistry*[37], *Molecular Modeling*[38], *Polymers & Plastics*[39], and *Chemical Kinetics & Catalysis*[40].

[32] https://osf.io/t3sxh/
[33] https://osf.io/xfepy/
[34] https://osf.io/wk6se/
[35] https://osf.io/q8k3u/
[36] https://osf.io/56azc/
[37] https://osf.io/ysg2j/
[38] https://osf.io/cq8j6/
[39] https://osf.io/4jwta/
[40] https://osf.io/9hmf3/

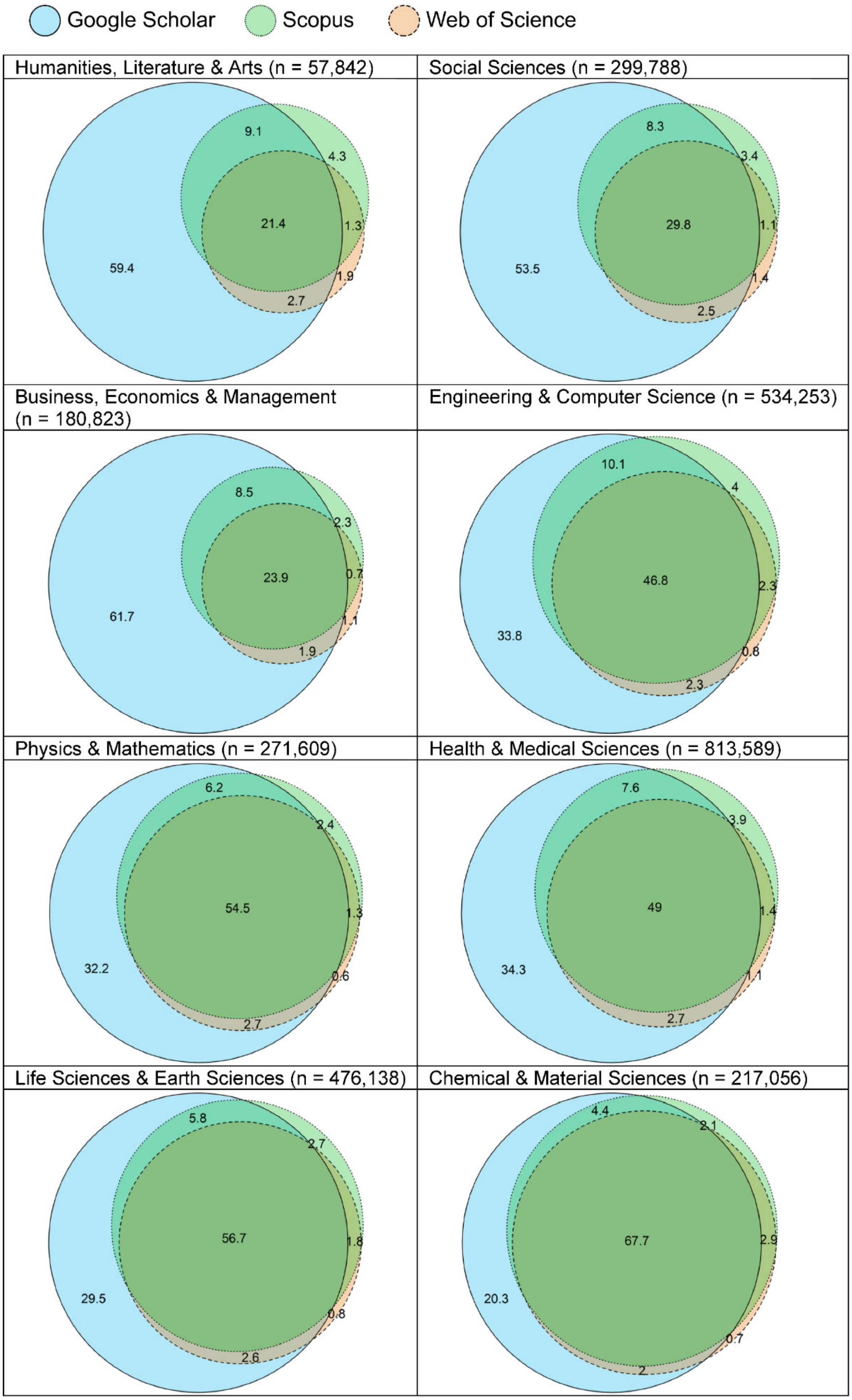


Figure 4. Percentage of unique and overlapping citations in Google Scholar, Scopus, and Web of Science, by broad subject area of cited documents

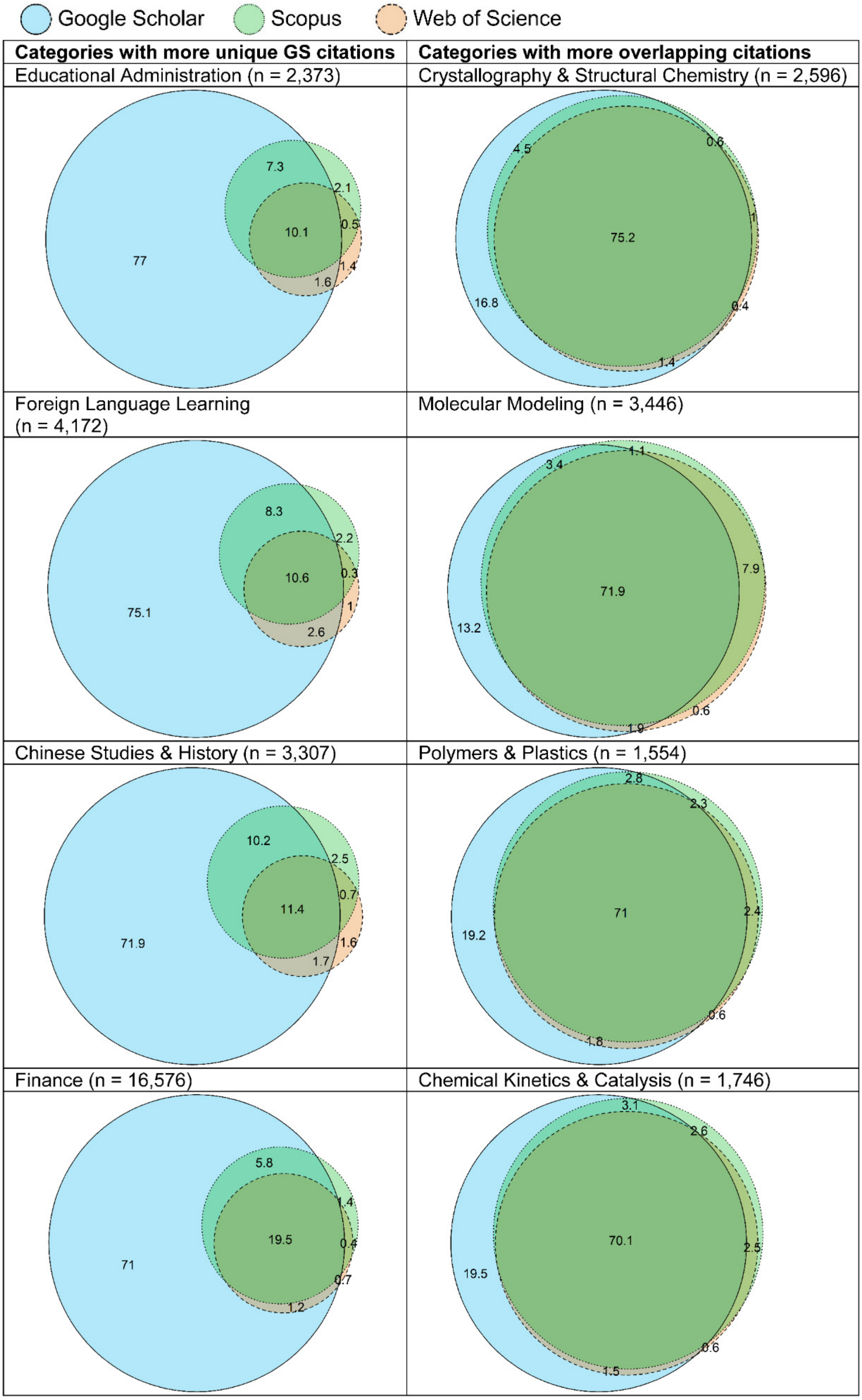


Figure 5. Categories with many unique citations or many overlapping citations

Overall, GS found 94% of all citations (93%-96% depending on the area), while WoS found 52% (ranging from 27% in *Humanities, Literature & Arts*, to 73% in *Chemical & Material Sciences*), and Scopus 60% (from 35% in *Business, Economics & Management*, to 77% in *Chemical & Material Sciences*). Additionally, GS found 95% of the citations that WoS found (88%-97% depending on the area), and 92% of the citations that Scopus found (84-94%) (Table 2). The data also shows that Scopus found 93% of the citations that Web of Science found (83-96% depending on the area).

Table 2. Percentage of citations in Google Scholar, Web of Science, and Scopus, relative to all citations, and relative to citations found by other databases

	% GS (all cit.)	**% WoS (all cit.)**	**% Scopus (all cit.)**	**% WoS cit. in GS**	**% Scopus cit. in GS**	**% WoS cit. in Scopus**
Overall	94	52	60	95	92	93
Humanities, Literature & Arts	**93**	**27**	36	**88**	**84**	**83**
Social Sciences	94	35	43	93	89	89
Business, Economics & Management	**96**	28	**35**	93	92	89
Engineering & Computer Science	**93**	52	63	94	90	94
Physics & Mathematics	96	59	64	**97**	**94**	94
Health & Medial Sciences	94	54	62	95	91	93
Life Sciences & Earth Sciences	95	62	67	96	93	95
Chemical & Material Sciences	94	**73**	**77**	95	**94**	**96**

The results for the 252 specific subject categories (available in the supplementary materials[41]) show that GS covers at least 90% of all citations in 233 out of 252 categories, the lowest value being 77% in *Visual Arts*[42], and the highest values around 98% in *Crystallography & Structural Chemistry*[43], *Evolutionary Biology*[44], *Quantum Mechanics*[45], and *Astronomy & Astrophysics*[46]. Relative to the coverage of WoS and Scopus, GS finds at least 90% of the citations that WoS and Scopus find in 221 and 164 categories, respectively, the lowest values belonging to the Humanities, such as *Film*[47], *Visual Arts*[48], and *History*[49] (56%-68%).

3.2. RQ2. Unique and non-unique citations

3.2.1. Document types

The distribution of document types of unique GS citations greatly differs from that of citations that were also found by WoS and/or Scopus. This is true across all eight broad subject categories (Figure 6). Among non-unique citations, the most common document type by far is the journal publication (from 71% in *Engineering & Computer Science*, to 94% in *Chemical & Material Sciences*). The other document types present among non-unique citations are books / book chapters and conference papers, with levels varying by subject area. Among unique GS citations, however, there is more document type diversity (including many never indexed by WoS or Scopus). Although journal publications are still the single most frequent document type, other document types comprise over 50% in all subject areas except *Health & Medical Sciences* (48%).

[41] https://osf.io/t3sxh/
[42] https://osf.io/7ea63/
[43] https://osf.io/ysg2j/
[44] https://osf.io/javkb/
[45] https://osf.io/cr3k2/
[46] https://osf.io/wmn8c/
[47] https://osf.io/7dkm3/
[48] https://osf.io/7ea63/
[49] https://osf.io/fgrp4/

The most frequent non-journal document type is the thesis or dissertation (22% in *Business, Economics & Management* – 37% in *Chemical & Material Sciences*), followed by books and book chapters (especially in *Humanities, Literature & Arts* and *Social Sciences*). This trend is different in *Engineering & Computer Science*, where conference papers are more common than books, and in *Business, Economics & Management* and *Physics & Mathematics*, where unpublished scholarly papers (such as working papers and preprints) are also more frequently used than books for scientific communication.

Considering the 252 specific subject categories[50], the percentage of known document types other than journal articles in the unique GS citations ranges from approx. 10% in Nonlinear Science, Heart & Thoracic Surgery, Natural Medicines & Medicinal Plants, and Oral & Maxillofacial Surgery, to over 55% in Special Education, and Computer Hardware & Design. However, unlike in the analysis by broad subject categories, a correction factor has not been applied (because no random samples were selected and analysed at this level), and therefore the document types of a large percentage of the citations are unknown (from approx. 20% in Special Education, and Ethnic & Cultural Studies, to over 50% in Quantum Mechanics, Geometry, and Algebra).

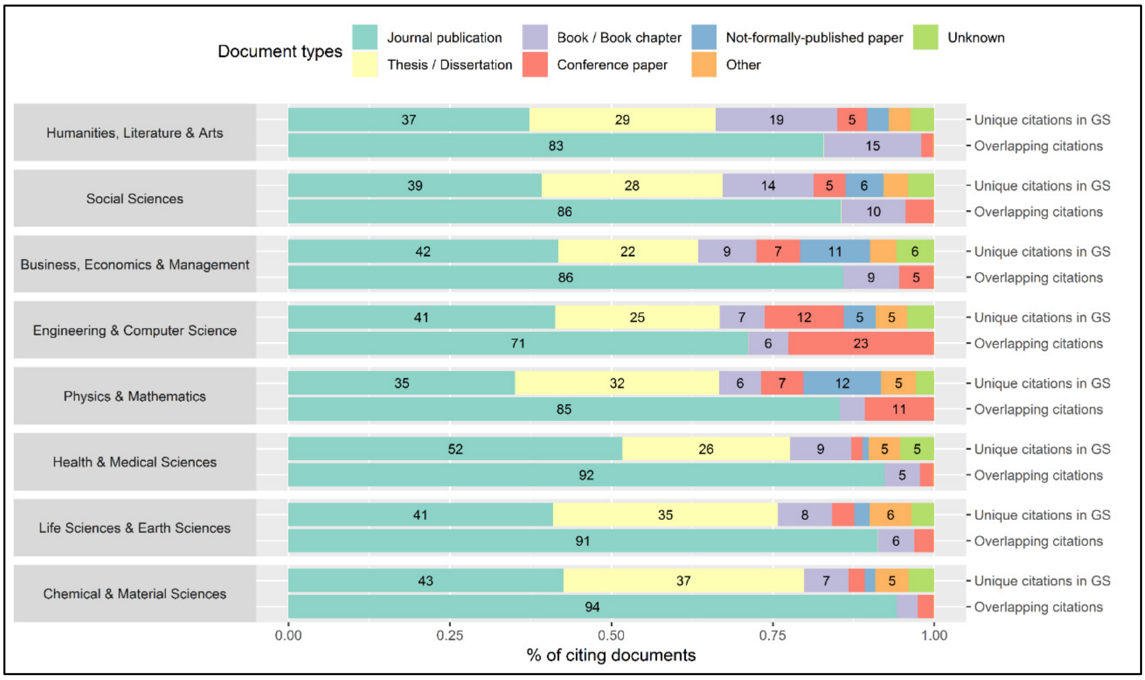
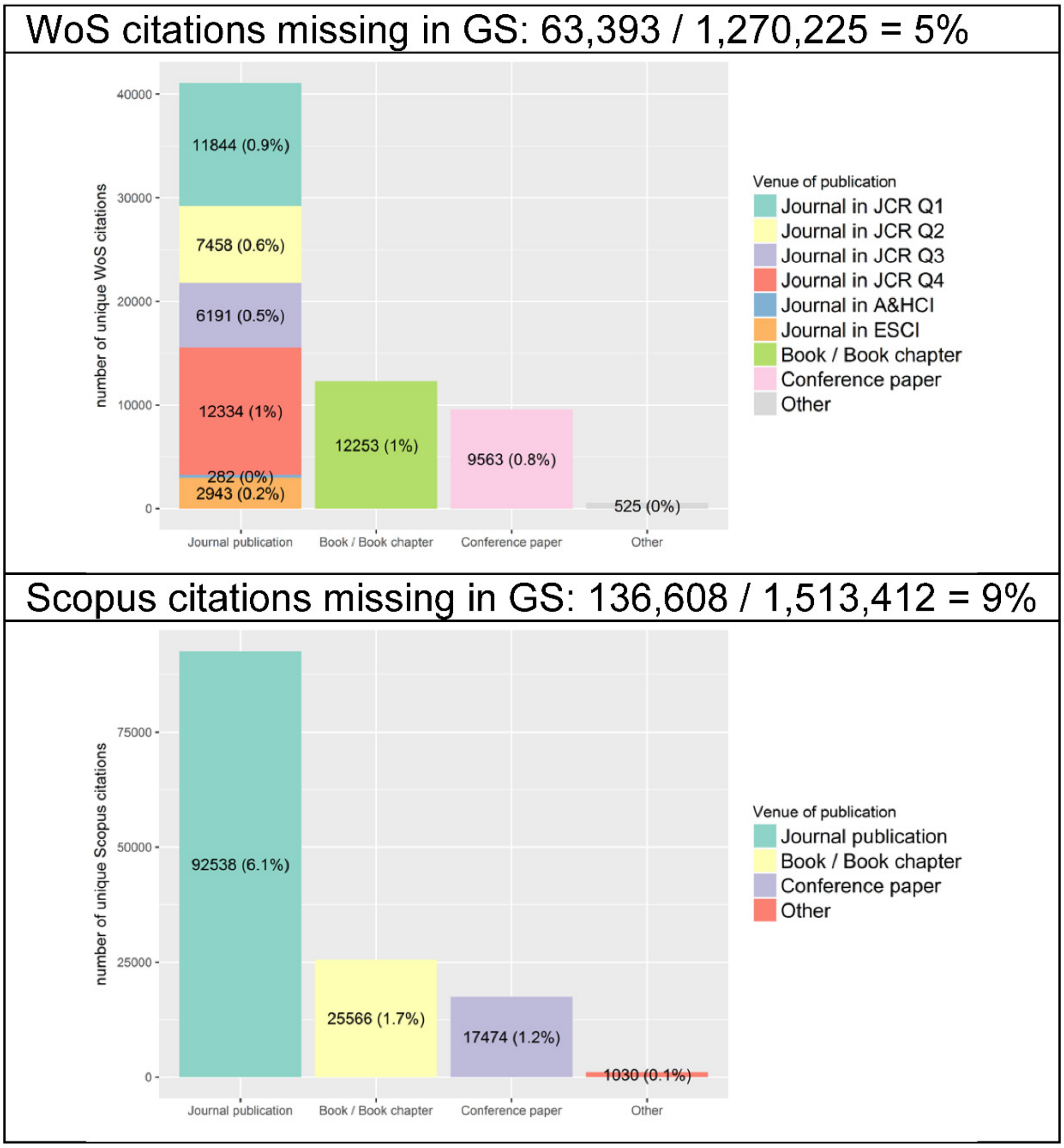


Figure 6. Distribution of document types among unique and overlapping citations in Google Scholar, by broad subject area of cited documents

Considering the citations found by WoS and/or Scopus which GS did not find (the citing document might be covered by GS without it making the connection between citing and cited document), most are from journals (Figure 7). Out of the 63,393 citations found by WoS and not by GS (5% of all citations), 41,052 (64% of the WoS citations that GS misses, or 3.2% of all citations analysed in this study) are from journals. Among citations from journal publications, there are more that were published in journals ranked in Q1 and Q4 of their respective JCR categories (0.9% and 1% of all citations), than in Q2 and Q3 (0.6% and 0.5%, respectively). The remaining missing citations come from books or book chapters (19% of WoS citations missing from GS, and 1% of all citations), and conference papers (15% of WoS citations missing from GS, and 0.8% of all citations). The proportions of Scopus citations missing from GS relative to the number of missing citations in GS (136,608) are very similar to those in WoS: 68% of journal publications, 19% books or book chapters, and 13% of conference papers. In this case, the proportion of Scopus citations missing from GS is 9%.

[50] https://osf.io/s5ndm/



WoS citations missing in GS: 63,393 / 1,270,225 = 5%

Scopus citations missing in GS: 136,608 / 1,513,412 = 9%

Figure 7. Proportion of document types among citations found by WoS and Scopus, and not by GS

3.2.2. Languages

The distribution of languages among the unique GS citations is very different from that of non-unique citations (Figure 8). Whilst for non-unique citations nearly all documents (97%-100%) were published in English, for unique GS citations the percentage ranges from 62% (*Health & Medical Sciences*) to 80% (*Humanities, Literature & Arts*). This is even though all documents in GSCP were published in English. The second most frequent language of unique GS citations was Chinese (4%-12%), and all other languages have a share of 4% or lower across all subject areas. A few (5%-10%) unique GS citations were published in languages outside the top 11 most frequently used languages overall (for all citations in our sample).

At the level of the 252 specific subject categories[51], the categories with a large proportion of non-English unique GS citations are Geochemistry & Mineralogy (59%), Surgery (56%), Radar, Positioning & Navigation (55%), and Cardiology (53%), whereas the categories with the lowest

[51] https://osf.io/xuz6w/

share of non-English citations are Astronomy & Astrophysics (10%), High Energy & Nuclear Physics (11%), Quantum Mechanics (11%), and Computer Hardware Design (11%).

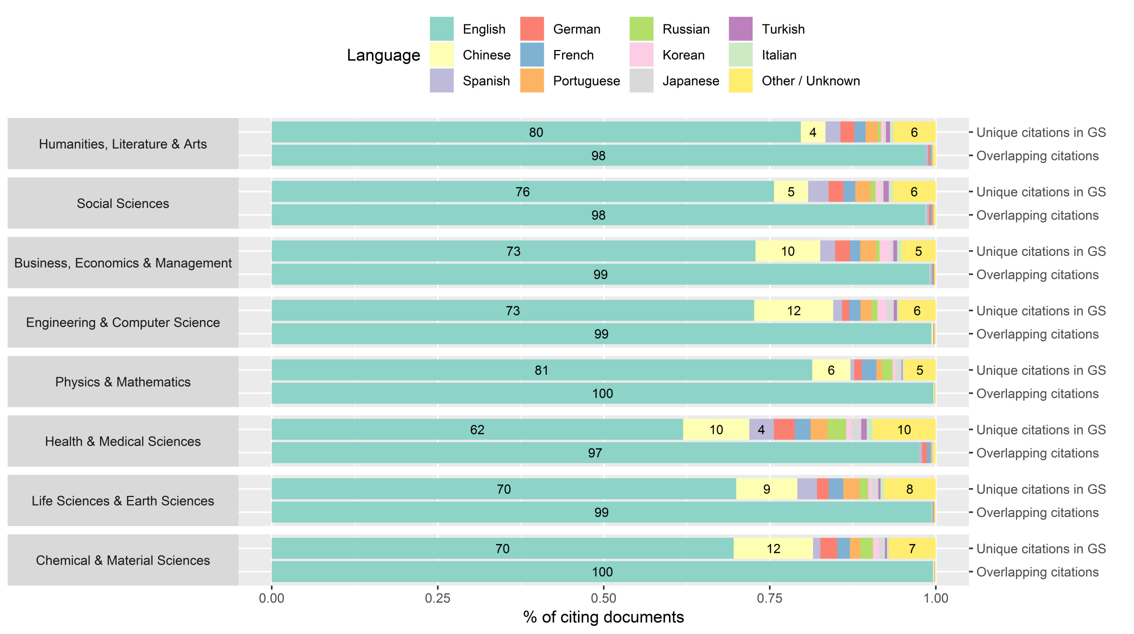


Figure 8. Distribution of languages among unique and overlapping citations in Google Scholar, by broad subject area of cited documents

3.2.3. Citation counts

This section analyses the Google Scholar citation counts of the 2,301,997 citing documents extracted from Google Scholar. The distributions of log-transformed (ln(1+x) to reduce skewing) citation counts among unique GS citations, and overlapping citations (those also found by WoS and/or Scopus) are different (Figure 9). Across all subject areas, the median log-transformed citation count is always zero and lower than the median of log-transformed citation counts of non-unique citations. The 95% confidence interval for the mean (represented as a red box in Figure 9) is also significantly lower for unique GS citations than for non-unique citations. Both unique and non-unique citations include many outliers (blue dots in Figure 9). The same pattern occurs across the 252 specific subject categories[52], although there are 29 categories in which the median of the citation counts for the unique GS citations is higher than zero (but still lower than the median for overlapping citations).

[52] https://osf.io/pm3xh/

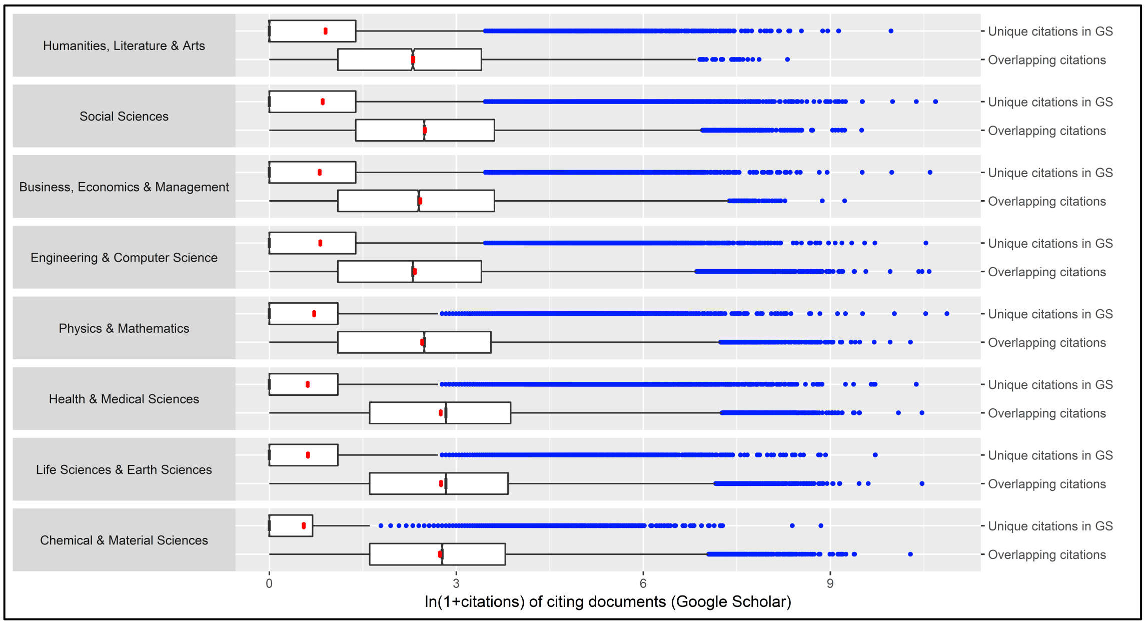


Figure 9. Distribution of citation counts among unique and overlapping citations in google scholar, by broad subject area of cited document

3.3. RQ3. Citation count comparisons

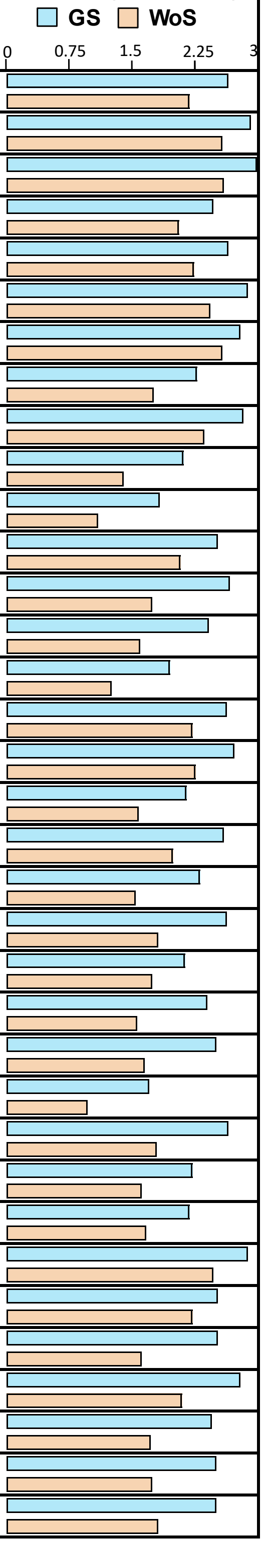
Spearman correlations between citation counts (GS-WoS, GS-Scopus) are close to 1.0 in most subject categories (Table 3 and Table 4). Correlations between GS and WoS range from .78 in *Literature*, to .98 in *Basic Life Sciences, Biomedical Sciences, Chemistry and Chemical Engineering,* and *Multidisciplinary journals*. In 30 out of the 35 areas of research in the NOWT classification (Tijssen et al., 2010), the Spearman correlation coefficient is over .90. Correlations between Google Scholar and Scopus are even stronger. The weakest correlation is .92 in *Economics, Econometrics, and Finance*, and the strongest is .99 in *Chemical Engineering, Immunology and Microbiology,* and *Multidisciplinary*. In 20 out of 27 categories in the ASJC scheme, correlation coefficients are above .95. The supplementary materials contain tables of citation count correlations computed at the level of the 252 WoS subject categories[53], and the 330 ASJC low-level categories[54], which give broadly comparable results. The weakest statistically significant correlation between GS and WoS at this level[55] is in Medieval & Renaissance Studies (.69), while the weakest correlation between GS and Scopus[56] is .74 in Classics.

On average, GS finds more citations than WoS and Scopus across all categories (see mean citation ratios in Table 3 and Table 4). This effect holds even when citation counts are log-transformed (1+ln(citations)) to reduce skewness. An inverse relationship between strength of correlation coefficients and mean citation ratios of GS over WoS/Scopus is observed. Strong correlation coefficients are associated with lower mean ratios, and vice versa.

[53] https://osf.io/x6mw7/
[54] https://osf.io/4pf9z/
[55] https://osf.io/x6mw7/
[56] https://osf.io/4pf9z/

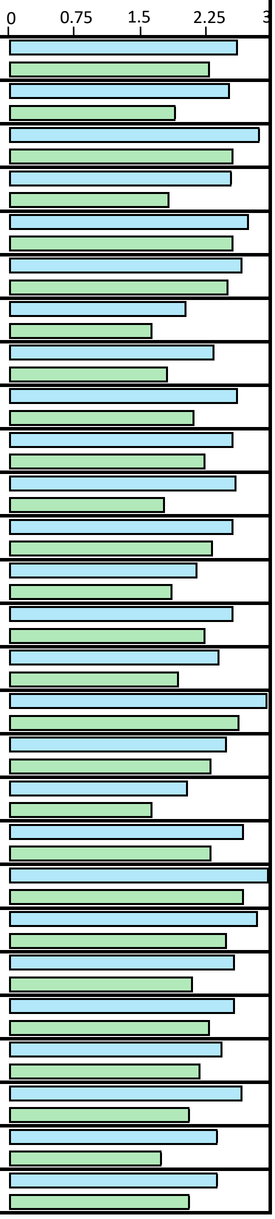
Table 3. Spearman correlation coefficients, mean ratio, and mean log-transformed citation counts of citing documents between GS and WoS, by subject category

Category (NOWT)	**N**	***r***	**Mean ratio of citation counts GS/WoS**	**Mean ln(1+citations)**
Agriculture and Food Science	24,176	.97	1.74	
Astronomy and Astrophysics	16,090	.96	1.60	
Basic Life Sciences	134,045	**.98**	1.58	
Basic Medical Sciences	23,183	.96	1.76	
Biological Sciences	62,094	.97	1.90	
Biomedical Sciences	118,817	**.98**	1.72	
Chemistry and Chemical Engineering	129,481	**.98**	**1.30**	
Civil Engineering and Construction	5,145	.95	1.87	
Clinical Medicine	223,309	.97	1.82	
Computer Sciences	61,199	.86	3.21	
Creative Arts, Culture and Music	1,145	.84	2.80	
Earth Sciences and Technology	46,536	.95	1.99	
Economics and Business	28,550	.93	3.30	
Educational Sciences	13,227	.92	2.92	
Electrical Engineering and Telecommunication	68,462	.83	3.18	
Energy Science and Technology	19,242	.95	1.86	
Environmental Sciences and Technology	64,791	.97	1.86	
General and Industrial Engineering	10,757	.92	2.37	
Health Sciences	28,371	.95	2.11	
History, Philosophy and Religion	5,062	.90	3.15	
Information and Communication Sciences	6,214	.94	2.87	
Instruments and Instrumentation	6,167	.95	1.74	
Language and Linguistics	3,149	.90	3.12	
Law and Criminology	4,348	.90	3.37	
Literature	368	**.78**	**4.02**	
Management and Planning	18,477	.94	2.83	
Mathematics	17,187	.91	2.65	
Mechanical Engineering and Aerospace	17,006	.91	2.24	
Multidisciplinary Journals	44,299	**.98**	1.63	
Physics and Materials Science	144,010	.97	1.52	
Political Science and Public Administration	8,118	.90	3.22	
Psychology	32,875	.95	2.50	
Social and Behavioral Sciences, Interdisciplinary	7,001	.93	2.77	
Sociology and Anthropology	11,504	.93	2.82	
Statistical Sciences	12,955	.92	3.16	



Confidence level of Spearman correlations: 99%; p-values < 0.01
Highest and lowest values of Spearman correlations and mean citation ratios are highlighted in bold

Table 4. Spearman correlation coefficients, mean ratio, and mean log-transformed citation counts of citing documents between GS and Scopus, by subject category

Category (ASJC)	**N**	***r***	**Mean ratio of citation counts GS/Scopus**	**Mean ln(1+citations)** 🟦 **GS** 🟩 **Scopus** 
Agricultural and Biological Sciences	109,423	.98	1.45	
Arts and Humanities	21,698	.95	2.19	
Biochemistry, Genetics and Molecular Biology	216,180	**.99**	1.43	
Business, Management and Accounting	40,539	.94	2.43	
Chemical Engineering	56,569	**.99**	1.27	
Chemistry	118,885	**.99**	**1.23**	
Computer Science	135,932	.94	1.72	
Decision Sciences	13,557	.94	2.04	
Dentistry	3,933	.97	1.78	
Earth and Planetary Sciences	52,356	.97	1.49	
Economics, Econometrics and Finance	22,273	**.93**	**2.83**	
Energy	31,166	.98	1.35	
Engineering	146,545	.96	1.49	
Environmental Science	66,212	.98	1.50	
Health Professions	12,309	.96	1.79	
Immunology and Microbiology	50,615	**.99**	1.44	
Materials Science	108,794	.98	1.27	
Mathematics	66,239	.94	1.78	
Medicine	361,217	.97	1.56	
Multidisciplinary	18,851	**.99**	1.43	
Neuroscience	46,462	.98	1.55	
Nursing	19,431	.96	1.80	
Pharmacology, Toxicology and Pharmaceutics	38,377	.98	1.42	
Physics and Astronomy	126,820	.97	1.42	
Psychology	42,037	.96	2.09	
Social Sciences	81,542	.94	2.22	
Veterinary	4,550	.98	1.47	

Confidence level of Spearman correlations: 99%; p-values < 0.01
Highest and lowest values of Spearman correlations and mean citation ratios are highlighted in bold

4. Discussion

4.1. Limitations

This study analyses a large sample of citations to highly-cited documents from all subject areas published in English. In order to generalize the results to all articles, it must be assumed that the population of documents that cite highly cited articles is not significantly different from the general population of documents that cite articles. This may not be fully true since, for example, highly cited articles are presumably more likely to be in emerging research areas and larger specialisms. Furthermore, the results may not reflect the citation coverage (in GS, WoS, and Scopus) of documents that do not usually cite scientific literature written in English, such as documents that address locally or regionally relevant topics written in vernacular languages.

Because the highly-cited documents from which our sample of citations came were all initially selected from Google Scholar, this might have provided an advantage to GS in the comparisons: GS might be better suited than WoS or Scopus to find citations for these specific documents, for unknown reasons. Nevertheless, the high citation count correlations found in section 3.3 suggest that this advantage is not substantial, as the three databases provide essentially the same citation rankings at the document level in most subject categories.

Without access to Clarivate Analytics' recently created ESCI Backfile for documents published between 2005 and 2014, an unknown number of citations in this study are listed as found only by GS and/or Scopus, when they are also captured by ESCI. Thus, the results should not be interpreted as applying to all possible WoS data.

Additionally, this article describes a methodology to match citations in GS, WoS, and Scopus at the level of cited articles. The rules chosen to classify a potential match as successful were intentionally conservative to minimize false positives (citations that are matched by the algorithm, despite being different). The matching algorithm probably created some false negatives (citations not matched by the algorithm, despite being the same), especially in categories where DOIs are less widely used and the matching had to rely more frequently on strict title similarity rules. Thus, in some cases the percentages of unique citations might be lower, and percentages of overlaps higher, than reported here.

4.2. Comparison with previous studies

The data from previous studies (Table 1) reveal a growth over time in the coverage of citations in GS. While these studies reported that GS could find 38%-94% of all citations found by any source, depending on the discipline(s) of study and the sample analysed, the current study finds values that are higher and more consistent across subject areas. The results here are more similar to those of the most recent study (Moed et al., 2016) and least similar to the earliest studies (Bakkalbasi et al., 2006; Kousha & Thelwall, 2008; Meho & Yang, 2007; Yang & Meho, 2007). For example, GS found 94.3% of all citations to GSCP in *Chemical & Material Sciences*. Although not fully comparable, this figure greatly differs from the 38% of all Chemistry citations found by GS that Kousha & Thelwall (2008) reported. This is evidence that the citation coverage of GS has become much more comprehensive over time. On the other hand, the more recent study by Moed et al., (2016) found that GS contained 94% of all citations in their sample, which is the same as the current study.

The percentages of WoS and Scopus citations that GS could find are generally higher in the current study than previously reported. While prior studies varied greatly depending on the sample (33%-75% of WoS citations, and 44%-89% of Scopus citations), in the current paper GS found 88%-97% of WoS citations, and 84%-94% of Scopus citations (depending on the area). This high relative overlap is a partial cause of the high correlations for citation counts between GS and WoS, and GS and Scopus, found by Martín-Martín et al. (2018). Lastly, this study reports lower percentages of unique citations in WoS (up to 1.9% of all citations) and Scopus (up to 4.3%) than reported in previous studies (up to 23%[57] in WoS, and 12% in Scopus).

[57] Considering studies that analysed the three databases (GS, WoS, and Scopus)

Regarding the distribution of document types and languages of GS unique citations, there were substantial percentages of theses and dissertations (from 22% in *Business, Economics & Management*, to 37% in *Chemical & Material Sciences*). These are larger than those found by Kousha & Thelwall (2008), Bar-Ilan (2010), and Lasda Bergman (2012), which found that up to 14% of GS unique citations belonged to this category. In the case of books and book chapters (from 7% in *Chemical & Material Sciences* to 19% in *Humanities, Literature & Arts*), conference proceedings (especially in *Engineering & Computer Science*: 12%), and unpublished materials such as preprints (11% in *Business, Economics & Management*, and 12% in *Physics and Mathematics*), the results are closer to those found by previous studies. The results also show a predominance of English for the citing sources, followed by Chinese (4%-12% depending on the source). These are similar to the results in Kousha & Thelwall (2008) in that Chinese is the second most used language in the sample of citations, although their study found very different percentages (approx. 35% in Biology, 25% in Chemistry, and less than 5% in Physics and Computing).

Lastly, the citation correlations between GS and WoS range from .78 in *Literature*, to .98 in *Basic Life Sciences, Biomedical Sciences, Chemistry and Chemical Engineering,* and *Multidisciplinary journals*, and the correlations between GS and Scopus range from .92 to .99. These correlations are similar to some in previous studies (Amara & Landry, 2012; Delgado López-Cózar et al., 2018; Martín-Martín et al., 2018; Minasny et al., 2013) but somewhat stronger than the ones found by others (De Groote & Raszewski, 2012; Kousha & Thelwall, 2007; Meho & Yang, 2007; Moed et al., 2016; Pauly & Stergiou, 2005; Rahimi & Chandrakumar, 2014; Wildgaard, 2015). This may be due to the disciplines of previous studies or the use of more recent data in the current paper.

5. Conclusions

This study provides evidence that GS finds significantly more citations than the WoS Core Collection and Scopus across all subject areas. Nearly all citations found by WoS (95%) and Scopus (92%) were also found by GS, which found a substantial amount of unique citations that were not found by the other databases. In the *Humanities, Literature & Arts*, *Social Sciences*, and *Business, Economics & Management*, unique GS citations surpass 50% of all citations in the area.

About half (48%-65%, depending on the area) of GS unique citations are not from journals but are theses/dissertations, books or book chapters, conference proceedings, unpublished materials (such as preprints), and other document types. These unique citations are primarily written in English, although a significant minority (19%-38% depending on the area) are in other languages. The scientific impact of these unique citations themselves is, on average, much lower than that of citations also found by WoS or Scopus, suggesting that the GS coverage advantage is mostly for low impact documents. Taken together, these results suggest caution if using GS instead of WoS or Scopus for citation evaluations. Without evidence, it cannot be assumed that the higher citation counts of GS are always superior to those of WoS and Scopus, since it is possible that the inclusion of lower quality citing documents reduces the extent to which citation counts reflect scholarly impact. For example, some of the citations from Master's theses may reflect educational impact. Therefore, depending on the type of evaluation that needs to be carried out, it might be necessary to remove certain types of citing documents from the citation counts, as suggested by Prins et al. (2016).

Spearman correlations between GS and WoS, and GS and Scopus citation counts are very strong across all subject categories but weaker in the Humanities (GS-WoS, Literature: .78) and Engineering (GS-WoS, Electrical Engineering and Telecommunication: .83). Also, correlations between GS and WoS (.78 to .98) are weaker than between GS and Scopus (.92 to .99). The weakest correlations are in the categories where there is a greater difference between the citation counts provided by GS, and the citation counts provided by WoS/Scopus. Thus, if GS is used for research evaluations then its data would be unlikely to produce large changes in the results, despite the additional citations found. It would be particularly useful when there is reason to believe that documents not covered by WoS or Scopus are important for an evaluation.

In conclusion, this study gives the first systematic evidence to confirm prior speculation (Harzing, 2013; Martín-Martín et al., 2018; Mingers & Lipitakis, 2010; Prins et al., 2016) that citation data in GS has reached a high level of comprehensiveness, because the gaps of coverage in GS found by the earliest studies that analysed GS data have now been filled. It surpasses WoS and Scopus numerically in all areas of research, and is greatly superior in the areas where WoS and Scopus have a poor coverage, including the Social Sciences and Humanities. However, at this point there is no reliable and scalable method to extract data from GS, and the metadata offered by the platform is still very limited, reducing the practical suitability of this source for large-scale citation analyses, although manual data collection is possible for small scale uses. Nevertheless, providing that a reliable method to extract citation data can be found, the lack of metadata could be solved by combining GS citation data with rich openly accessible data, such as that provided by CrossRef.

Acknowledgements

Alberto Martín-Martín is funded for a four-year doctoral fellowship (FPU2013/05863) granted by the Ministerio de Educación, Cultura, y Deportes (Spain). An international mobility grant from Universidad de Granada and CEI BioTic Granada funded a research stay at the University of Wolverhampton.

References

Amara, N., & Landry, R. (2012). Counting citations in the field of business and management: why use Google Scholar rather than the Web of Science. *Scientometrics*, *93*(3), 553–581. https://doi.org/10.1007/s11192-012-0729-2

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, *3*(1), 7. https://doi.org/10.1186/1742-5581-3-7

Bar-Ilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics*, *82*(3), 495–506. https://doi.org/10.1007/s11192-010-0185-9

Chavarro, D., Ràfols, I., & Tang, P. (2018). To what extent is inclusion in the Web of Science an indicator of journal 'quality'? *Research Evaluation*, *27*(2), 106–118. https://doi.org/10.1093/reseval/rvy001

Clarivate Analytics. (2015). Web of Science & Google Scholar collaboration. Retrieved June 5, 2018, from http://wokinfo.com/googlescholar/

Clarivate Analytics. (2017). Emerging Sources Citation Index Backfile (2005-2014). Retrieved from https://clarivate.com/wp-content/uploads/2017/10/M255-Crv_SAR_ESCI-infographic-FA.pdf

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176. https://doi.org/10.1145/363958.363994

De Groote, S. L., & Raszewski, R. (2012). Coverage of Google Scholar, Scopus, and Web of Science: a case study of the h-index in nursing. *Nursing Outlook*, *60*(6), 391–400. https://doi.org/10.1016/j.outlook.2012.04.007

De Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. https://doi.org/10.1002/asi.4630270505

de Winter, J. C. F., Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics*, *98*(2), 1547–1565. https://doi.org/10.1007/s11192-013-1089-2

Delgado López-Cózar, E., Orduna-Malea, E., & Martín-Martín, A. (2018). Google Scholar as a data source for research assessment. In W. Glaenzel, H. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*. Springer.

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446–454. https://doi.org/10.1002/asi.23056

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., … Parsonage, H. (2018). data.table: Extension of “data.frame.” Retrieved from https://cran.r-project.org/package=data.table

Else, H. (2018, April 11). How I scraped data from Google Scholar. *Nature*. https://doi.org/10.1038/d41586-018-04190-5

Elsevier. (2018). Scopus source list (April 2018). Retrieved from https://www.elsevier.com/__data/assets/excel_doc/0015/91122/ext_list_April_2018_2017_Metrics.xlsx

Halevi, G., Moed, H., & Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the Literature. *Journal of Informetrics*, *11*(3), 823–834. https://doi.org/10.1016/J.JOI.2017.06.005

Harzing, A.-W. (2013). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics*, *98*(1), 565–575. https://doi.org/10.1007/s11192-013-0975-y

Harzing, A.-W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*(2), 787–804. https://doi.org/10.1007/s11192-015-1798-9

Harzing, A. W. (2007). Publish or Perish. Retrieved from http://www.harzing.com/pop.htm

Harzing, A. W. K., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, *8*(1), 61–73. https://doi.org/10.3354/esep00076

Jacimovic, J., Petrovic, R., & Zivkovic, S. (2010). A citation analysis of Serbian Dental Journal using Web of Science, Scopus and Google Scholar. *Stomatoloski Glasnik Srbije*, *57*(4), 201–211. https://doi.org/10.2298/SGS1004201J

Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, *34*(1), 175–191. https://doi.org/10.1108/14684521011024191

Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, *58*(7), 1055–1065. https://doi.org/10.1002/asi.20584

Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, *74*(2), 273–294. https://doi.org/10.1007/s11192-008-0217-x

Larsson, J., Godfrey, A. J. R., Kelley, T., Eberly, D. H., Gustafsson, P., & Huber, E. (2018). eulerr: Area-Proportional Euler and Venn Diagrams with Circles or Ellipses. Retrieved from https://cran.r-project.org/package=eulerr

Lasda Bergman, E. M. (2012). Finding Citations to Social Work Literature: The Relative Benefits of Using Web of Science, Scopus, or Google Scholar. *The Journal of Academic Librarianship*, *38*(6), 370–379. https://doi.org/10.1016/j.acalib.2012.08.002

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).

Martín-Martín, A., & Delgado López-Cózar, E. (2016). Reading Web of Science data into R. Retrieved from https://github.com/alberto-martin/read.wos.R

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. *Scientometrics*, *116*(3), 2175–2188. https://doi.org/10.1007/s11192-018-2820-9

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105–2125. https://doi.org/10.1002/asi.20677

Minasny, B., Hartemink, A. E., McBratney, A., & Jang, H.-J. (2013). Citations and the h index of soil researchers and journals in the Web of Science, Scopus, and Google Scholar. *PeerJ*, *1*, e183. https://doi.org/10.7717/peerj.183

Mingers, J., & Lipitakis, E. A. E. C. G. (2010). Counting the citations: a comparison of Web of Science and Google Scholar in the field of business and management. *Scientometrics*, *85*(2), 613–625. https://doi.org/10.1007/s11192-010-0270-0

Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics*, *10*(2), 533–551. https://doi.org/10.1016/j.joi.2016.04.017

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Ooms, J., & Sites, D. (2018). cld2: Google's Compact Language Detector 2. Retrieved from https://cran.r-project.org/package=cld2

Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, *104*(3), 931–949. https://doi.org/10.1007/s11192-015-1614-6

Orduña-Malea, E., Martín-Martín, A., Ayllón, J. M., & Delgado López-Cózar, E. (2016). *La revolución Google Scholar : Destapando la caja de Pandora académica*. Granada: Universidad de Granada.

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2018). Classic papers: using Google Scholar to detect the highly-cited documents. In *23rd International Conference on Science and Technology Indicators*. Leiden. https://doi.org/10.31235/osf.io/zkh7p

Pauly, D., & Stergiou, K. (2005). Equivalence of results from two citation analyses: Thomson ISI's Citation Index and Google's Scholar service. *Ethics in Science and Environmental Politics*, *9*, 33–35. https://doi.org/10.3354/esep005033

Prins, A. A. M., Costas, R., van Leeuwen, T. N., & Wouters, P. F. (2016). Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation*, *25*(3), 264–270. https://doi.org/10.1093/reseval/rvv049

R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.r-project.org/

Rahimi, S., & Chandrakumar, V. (2014). A comparison of citation coverage of traditional and web citation databases in medical science. *Malaysian Journal of Library and Information Science*, *19*(3), 1–11. Retrieved from http://jice.um.edu.my/index.php/MJLIS/article/view/1779

Sud, P., & Thelwall, M. (2014). Evaluating altmetrics. *Scientometrics*, *98*(2), 1131–1143. https://doi.org/10.1007/s11192-013-1117-2

Tijssen, R., Nederhof, A., van Leeuwen, T., Hollanders, H., Kanerva, M., & van den Berg, P. (2010). *Wetenschaps- en Technologie- Indicatoren 2010*. Retrieved from http://nowt.merit.unu.edu/docs/NOWT-WTI_2010.pdf

van der Loo, M., van der Laan, J., R Core Team, Logan, N., & Muir, C. (2018). stringdist: Approximate String Matching and String Distance Functions. Retrieved from https://cran.r-project.org/package=stringdist

van Leeuwen, T. N., Moed, H. F., Tijssen, R. J. W., Visser, M. S., & Van Raan, A. F. J. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics*, *51*(1), 335–346. https://doi.org/10.1023/A:1010549719484

Van Noorden, R. (2014, November 7). Google Scholar pioneer on search engine's future. *Nature*. https://doi.org/10.1038/nature.2014.16269

Walker, A., & Braglia, L. (2018). openxlsx: Read, Write and Edit XLSX Files. Retrieved from https://cran.r-project.org/package=openxlsx

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wildgaard, L. (2015). A comparison of 17 author-level bibliometric indicators for researchers in Astronomy, Environmental Science, Philosophy and Public Health in Web of Science and Google Scholar. *Scientometrics*, *104*(3), 873–906. https://doi.org/10.1007/s11192-015-1608-4

Yang, K., & Meho, L. I. (2007). Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–15. https://doi.org/10.1002/meet.14504301185

Section 2. Reusing data from Google Scholar to create new bibliometric tools.

Chapter 8. Summary of results

Previous studies have shown how, despite its limitations, data from GS can be a useful for bibliometric analyses. However, GS's various interfaces (GS Search, GSM, GSC) give users very limited options to browse and analyse data. It was therefore interesting to explore whether GS data could be reused and reorganized to allow additional use cases which are not facilitated by the official interfaces. In this section we describe the projects in which we attempted to reuse data from GS by extracting and refactoring it for a variety of purposes, and then creating brand new interfaces where these data can be browsed. These interfaces were implemented as freely accessible web applications.

Three different types of prototype applications were developed and are presented here. The first application presents journal-level bibliometric indicators for a large collection of journals in the Arts, Humanities, and Social Sciences (AHSS). The second application presents data from a specific academic community at various levels of aggregation (author-, document-, journal-, and publisher-level), combining data not only from GS but from other sources. And lastly, in the third application, a large sample of data from GS is used to analyse Open Access levels by country, subject category, journal, and publication year.

Shining light on Arts, Humanities, and Social Sciences journals around the globe

In 2012, GS's journal ranking Google Scholar Metrics (GSM) was launched. Taking advantage of GS's extensive coverage, it enabled users to look up bibliometric indicators for a much larger number of journals than previously possible. Coverage in other journal rankings such as the Journal Citation Reports (JCR) and SCImago Journal Rank (SJR) was limited by the selective indexing approach of their "mother" databases, WoS and Scopus.

However, GSM also presented shortcomings (Martín-Martín, Ayllón, Orduña-Malea, & Delgado-López-Cózar, 2014). For example, it only includes journals published in English in its subject categories, and only presents up to 20 journals per category. This particular limitation meant that, even though much information on journals was available from GSM, most of it was not visible in the subject category and language rankings, and could only be accessed by using its search tool. This issue affected AHSS journals in particular, as in many cases, these journals are not published in English.

It is for this reason that we decided to overcome some of GSM's shortcomings by trying to extract as many of the AHSS journals it covered as possible, and presenting them in an alternative application where subject categories are not limited to journals written in English, nor to 20 journals per category (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2017a) (chapter 9 of this thesis).

First, we needed to identify as many AHSS journals as possible. For this purpose, a master list of journals was generated by combining the journals covered by a number of sources, including a general journal directory (Ulrichs' Global Serials Directory), list of journals covered by WoS and Scopus, list of journals covered by international disciplinary databases, and of course, GSM. 66,454 journals from all areas were identified. AHSS journals in this list were searched on GSM. 9,188 AHSS journals were identified. These journals were classified in one or more of 22 categories (13 in the area of Social Sciences, and 9 in the area of Arts and Humanities). In order to classify journals in categories, we based our decision on the categories where the journals were classified in other databases, and on whether the journals were covered in the appropriate disciplinary databases.

The result of this analysis is available in the web application Journal Scholar Metrics (JSM) (http://www.journal-scholar-metrics.infoec3.es), where users can browse journals either by subject category, or country of publication. Journal lists in this application provide several bibliometric indicators: H5-Index and H5-Median (provided by GSM), H Citations (sum of citation counts in documents that contribute to H5-Index), as well as H5-Index and H Citations after removing journal self-citations (citations that originate in the same journal). Lastly quartiles were also computed for lists of journals classified in a specific category.

43% (3,944) of the AHSS journals in GSM (those which are covered in JSM) are not covered by WoS/JCR or Scopus/SJR (Figure 1). And while a large number of journals in the AHSS categories of SJR seem to be missing from GSM, this is in part an artifact of not using a unified classification scheme for all sources in the comparison. In SJR's subject classification, journals such as "Science", "JAMA Psychiatry", or "Brain" are classified as Arts and Humanities journals.

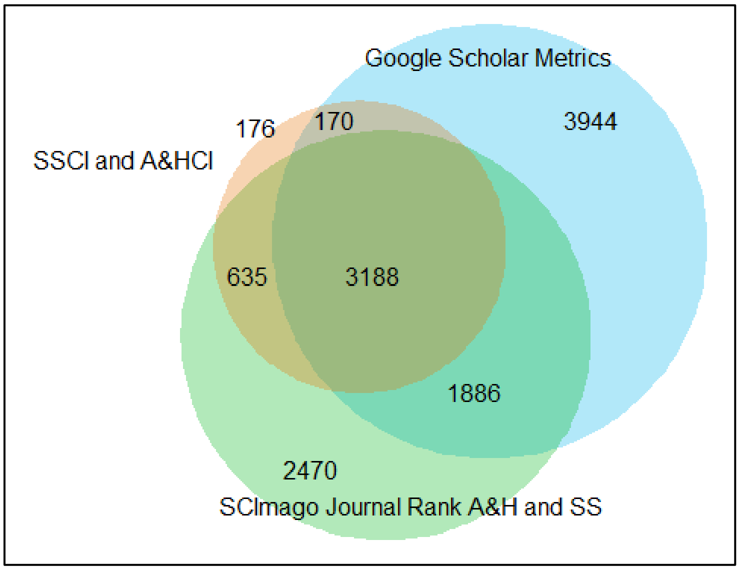


Figure 1. Journal coverage comparison of GSM, WoS, and SJR.

AHSS journals in GSM also show a greater diversity as regards countries of publication (Figure 2) and languages (Figure 3) than AHSS journals in WoS/JCR and Scopus/SJR. Across all three sources, USA and the United Kingdom are the countries where more journals are published. However, in GSM they make up for just under 50% of all journals, while in WoS, they reach almost 70% of all journals. SJR is found among these two values, with almost 60% of all journals published in USA or UK. Results by language of publication reveal a similar picture: in GSM journals published in English make up almost 60% of all journals, while in WoS, the proportion is almost 75%.

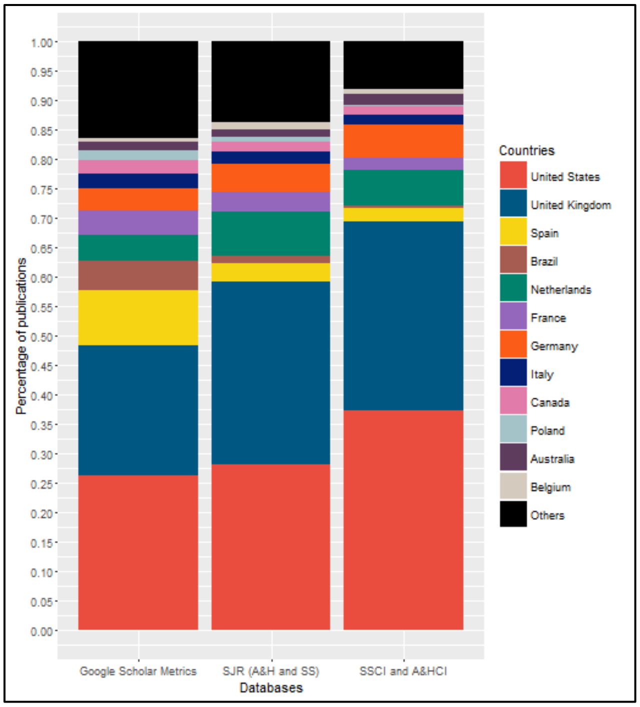


Figure 2. Distribution of journals by country of publication in GSM, SJR, and WoS

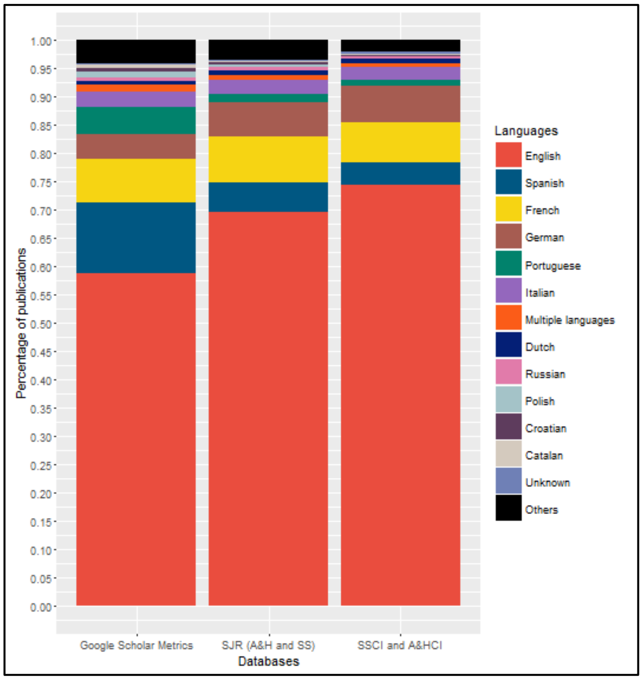


Figure 3. Distribution of journals by languages in GSM, SJR, and WoS

Producing a multifaceted representation of an academic community

The proliferation of freely available profiling tools for academic researchers, each drawing from a specific document base, each providing its own set of indicators, and more importantly, each appealing to a specific group of researchers, led us to the idea of “Scholar Mirrors”. In a House of Mirrors, each mirror presents a distorted reflection of the person that stands in front of it, the distortion depending on the imperfections of the mirror. Likewise, the indicators provided by any profiling platform depend to a large degree on the coverage of its document base (for production- and citation-based indicators) and the demographics that make up its user base (for usage and/or attention indicators). Each profile therefore provides a more or less distorted representation of a researcher’s work and the impact this work has had on its community. Faced with this scenario, we decided to try to generate a representation of an academic community by combining information from a variety of sources.

Our first attempt at this materialised as the web application “Spanish Library and Information Science in Google Scholar Citations” (http://www.biblioteconomia-documentacion-española.infoec3.es). In this first prototype, after collecting profile information from GSC about 336 LIS researchers from Spain (some of them working abroad), we also collected data from other sources such as WoS and ResearchGate. Results were displayed at the author, document, journal, publisher, and institutional levels. Users could easily observe how authors and documents changed positions in the list depending on by which indicator the list was sorted.

In our second attempt, we selected a larger community as our case study: the international community researching in the fields of Bibliometrics, Scientometrics, Informetrics, Webometrics, and Altmetrics (Martín-Martín, Orduna-Malea, Ayllón, & Delgado-López-Cózar, 2016) (Chapter 10 in this thesis). In this attempt, the method we followed to extract and reorganize the data was refined into what we called MADAP (Multifaceted Analysis of Disciplines through Academic Profiles) (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018a) (Chapter 12 of this thesis). This method includes the following steps:

- Identification of authors and their online profiles: in this first step several search strategies were combined to maximize the number of profiles identified. In total, 814 were identified using the following strategies:
    - Keyword searches in GSC: in order to identify a set of keywords that authors could have used in their profiles to describe themselves, the titles of the articles published in the core journals of the discipline were analysed.
    - Institutional affiliation: known research centers/deparments working on bibliometrics were also searched on GSC to retrieve the list of authors working in them. However, this method did not provide any author which was not found with the previous method.
    - Keyword searches in GS: searches were also carried out in GS Search in order to identify authors with a GSC profile but who might not have filled the field that contains the areas of interest. Additionally, relevant documents from authors which do not have a GSC profile were also identified with this search.
- Classification of authors: the field of Scientometrics and its various branches have the characteristic that they attract research from authors who normally work in other fields. For this reason, authors were classified as specialists (when their scientific production mainly falls within the field of Scientometrics), or occasional (authors from other fields who sometimes carry out scientometric studies). In our sample, 396 authors were classified as specialists, and 415 as occasional authors in the field of Scientometrics.
- Extraction of document-level data: the top 100 most cited documents published by each of these authors were extracted from GSC, processed, and combined with the document-level data

extracted from the previous keyword searches in GS. With this dataset, a list of the top 1,000 most cited documents in the discipline was generated. Lastly, this list of classic papers in the discipline was used to generate the list of most influential journals and book publishers.

In this attempt, other sources of author-level data apart from GS were also used (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018b) (Chapter 13 of this thesis). Specifically, we also searched the profiles of the identified authors in ResearchGate, Mendeley, ResearcherID, and Twitter. Out of all the researchers with a profile in GSC, 67% had a profile in ResearchGate, 41% were also in Mendeley, 40% were also in ResearcherID, and 30% in Twitter.

The author-level indicators provided by each platform for our sample of authors were extracted. All these indicators were compared using the Spearman correlation coefficient. The coefficients show that author-level indicators from GSC (all of them citation-based indicators) correlate well with author-level indicators from ResearchGate, such as the RG Score, number of publications, Impact Points (a now discontinued indicator that added the Impact Factor of the journals where the author has published), downloads, views, and citations. However, GSC indicators do not correlate with other RG indicators such as the number of followers of an author, or the number of people followed by the author. GSC indicators also correlate fairly well with some Mendeley indicators such as Readers, number of publications, or the average of Readers per document, and with production-based or citation-based indicators from ResearcherID. GSC indicators do not correlate well with indicators computed from Twitter data, except moderately with the sum of retweets and H retweets (the h-index formula applied to the number of retweets that tweets by an author have received). Lastly, PCA analyses confirm the differences between connectivity indicators such as the number of followers, retweets, etc., with indicators based on use or citations.

The data that was extracted and processed for these analysis was transformed into a web application called "Scholar Mirrors" (http://www.scholar-mirrors.infoec3.es) (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2017b) (Chapter 11 in this thesis), which presents the data discussed above at the level of authors, documents (top 1,000 most cited), journals, and book publishers.

Working on these prototypes enabled us to identify the issues, difficulties, limitations, and problems of each profiling platform (Martín-Martín et al., 2016) (Chapter 10 in this thesis). With this knowledge, we have begun to design and work on a much more challenging application, one that serves as a research information system for all researchers working in Spain (Chapter 14 of this thesis). Up until now, we have been able to identify all public GSC profiles of researchers working in Spain (over 43,000 in 2017 when the search took place), extract the lists of documents published by each of these researchers (over 2 million unique documents), and extract all citations to these documents (almost 25 million citations). Lastly, with the citation data and a clustering algorithm (Waltman & Van Eck, 2013) we were able to group documents in clusters of related documents, which is the first step to generate an automated classification of documents. In the future, we would like to continue with this project and create a web application that displays bibliographic information and contextualized bibliometric indicators about all researchers with a public GSC profile working in Spain.

Analysing Open Access levels using data from Google Scholar

Given how GS continuously sweeps the academic web in search of academic documents and how it merges together different versions of the same document, it is able to determine whether a document is freely accessible from one or more of the sources where it is found, even if the version of record provided by the publisher is not freely accessible. This is primarily useful to users who want to access documents, especially when the publisher version is not freely accessible. But in addition to that, this data is also useful to measure the number of documents that are freely accessible at various levels of aggregation (for example, by countries or by subject categories).

In 2016, we designed a study to conduct this type of analysis (Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018) (Chapter 16 of this thesis). The sample of study were all articles and reviews with DOI published in 2009 and 2014 covered by the three main citation indexes in WoS (Science Citation Index, Social Sciences Citation Index, and Arts & Humanities Citation Index). This sample was selected in part to facilitate analyses down the line (given the high-quality bibliographic metadata in WoS that GS lacks), and in part so that results were comparable to other studies that used similar samples of documents, but different sources of OA evidence. Over 2.3 million documents were selected in total, from all subject areas covered by WoS.

These documents were searched on GS one by one, a process that took approximately three months because of GS's limitations to extract data. 97.6% of the documents were successfully identified, and the links to freely accessible full texts displayed by GS were collected. Each link was classified as one type of OA (Gold, Hybrid, Delayed, Green) or as FA (freely available, but not belonging to any of the previous categories).

A web application was developed to aggregate results at various levels, such as by journal, by publication year, by subject category, or by country of affiliation of the authors (Chapter 15 of this thesis). The application generates a summary table using the parameters specified by the user, as well as a frequency table filtered by the documents included in the user's selection that displays the web domains where freely accessible versions of these documents can be found more often. Lastly, the application is also able to generate a graph based on the rows of the summary table that the user is more interested in.

The results of this analysis show that overall, GS had found at least one freely accessible version for 54.7% of the documents in the sample (Martín-Martín, Costas, et al., 2018) (Chapter 16 of this thesis). 7.3% were published in full OA journals (Gold OA), 1.1% were published in hybrid journals, 1.5% in journals that make articles OA after an embargo period (Delayed OA), 13.2% in journals that make their articles freely accessible, but do not attach a clear OA license to them (Bronze OA), 17.6% were available from repositories (Green OA), and 40.6% were available from other sources (freely available, but not OA), primarily the academic social network site ResearchGate (32.6%).

References

Martín-Martín, A., Ayllón, J. M., Orduña-Malea, E., & Delgado-López-Cózar, E. (2014). *Google Scholar Metrics 2014: a low cost bibliometric tool* (EC3 Working Papers No. 17). Retrieved from http://arxiv.org/abs/1407.2827

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, *12*(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2016). *The counting house, measuring those who count: Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics* *and Altmetrics in GSC (Google Scholar Citations), ResearcherID, ResearchGate, Mendeley, & Twitter* (EC3 Working Papers No. 21). Retrieved from https://arxiv.org/abs/1602.02412

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017a). Journal Scholar Metrics: building an Arts, Humanities, and Social Sciences journal ranking with Google Scholar data. In *22nd International Conference on Science, Technology & Innovation Indicators (STI)*. Paris. https://doi.org/10.17605/OSF.IO/VXNW6

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017b). Scholar Mirrors: Integrating evidence of impact from multiple sources into one platform to expedite researcher evaluation. In *22nd International Conference on Science, Technology & Innovation Indicators (STI)*. https://doi.org/10.31235/osf.io/z4bwe

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018a). A novel method for depicting academic disciplines through Google Scholar Citations: The case of Bibliometrics. *Scientometrics*, *114*(3), 1251–1273. https://doi.org/10.1007/s11192-017-2587-4

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018b). Author-level metrics in the new academic profile platforms: The online behaviour of the Bibliometrics community. *Journal of Informetrics*, *12*(2), 494–509. https://doi.org/10.1016/j.joi.2018.04.001

Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B*, *86*(11). https://doi.org/10.1140/epjb/e2013-40829-0

Capítulo 8. Resumen de resultados

Los estudios anteriores muestran cómo, a pesar de sus limitaciones, los datos de GS pueden ser útiles para realizar estudios bibliométricos. Sin embargo, las diferentes interfaces de GS (motor de búsqueda GSC, GSM) proporcionan opciones muy limitadas para explorar y analizar datos. Es por tanto interesante explorar si los datos de GS podrían ser reutilizados y reorganizados para permitir casos de uso adicionales que no son facilitados por las interfaces oficiales. En esta sección describimos proyectos en los que intentamos reutilizar datos de GS mediante la extracción de datos, su procesamiento, y su incorporación en interfaces nuevas y personalizadas que facilitaran unos casos de uso determinados. Estas interfaces fueron implementadas con aplicaciones web de acceso gratuito.

Se han diseñado tres prototipos diferentes de aplicaciones. La primera aplicación presenta indicadores bibliométricos a nivel de revista para una gran colección de revistas en las áreas de Artes, Humanidades, y Ciencias Sociales (AHSS por sus siglas en inglés). La segunda aplicación presenta datos a varios niveles de agregación (autor, documento, revista, editorial) de una comunidad científica muy específica, combinando datos no solo de GS sino de otras fuentes. Finalmente, en la tercera aplicación, una gran muestra de datos de GS se utiliza para analizar los niveles de Acceso Abierto a las publicaciones científicas por país de afiliación, categoría temática, revista, y año de publicación.

Arrojando luz sobre las revistas de Arte, Humanidades, y Ciencias Sociales de todo el mundo

En 2012 se lanzó el ranking de revistas Google Scholar Metrics (GSM). Aprovechándose de la extensa cobertura de GS, este servicio permitió a los usuarios consultar indicadores bibliométricos a nivel de revista para un número mucho mayor de revistas de lo que hasta el momento era posible. La cobertura de otros rankings de revistas como los Journal Citation Reports (JCR) y el SCImago Journal Rank (SJR) estaban limitados por la política de indización selectiva de sus respectivas bases de datos “madre”, WoS y Scopus.

Sin embargo, GSM también presentaba limitaciones (Martín-Martín, Ayllón, Orduña-Malea, & Delgado-López-Cózar, 2014). Por ejemplo, solo clasifica en categorías temáticas las revistas que se publican en inglés, y solo muestra 20 revistas en cada categoría. Esto significa que, aunque existe información sobre un gran número de revistas en GSM, la mayoría de esta información no es accesible a través de los rankings por categorías y por idiomas, sino que solo es accesible al utilizar la herramienta de búsqueda del servicio. Este aspecto afecta a las revistas de AHSS en particular, ya que en muchos casos, estas revistas no se publican en inglés.

Por esta razón decidimos intentar superar algunas de las limitaciones de GSM creando una aplicación alternative donde la categorización no estuviera limitada a 20 revistas por categoría ni a revistas publicadas en inglés. (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2017a) (capítulo 9 de esta tesis).

En primer lugar necesitábamos identificar tantas revistas de AHSS como fuera posible. Para conseguir esto, se generó una master list de revistas científicas que se alimentó de varias fuentes, como un directorio general de revistas (Ulrichs’ Global Serials Directory), la lista de revistas cubiertas por WoS y Scopus, las listas de revistas cubiertas por una serie de bases de datos especializadas de ámbito internacional, y por supuesto, GSM. Se identificaron 66.454 revistas de todas las áreas científicas. Las revistas de AHSS en esta lista fueron buscadas en GSM, donde se identificaron un total de 9,188 revistas. Cada una de estas revistas fue clasificada en una o más de categorías de entre un listado de 22 (13 en el área de Ciencias Sociales, y 9 en el área de Arte y Humanidades). Para llevar a cabo la clasificación, basamos nuestra decisión en las categorías donde las revistas habían sido clasificada en otras bases de datos, y atendiendo a en qué bases de datos especializadas estaban indizadas las revistas.

El resultado de este análisis se puede visualizar en la aplicación web Journal Scholar Metrics (JSM) (http://www.journal-scholar-metrics.infoec3.es), donde los usuarios pueden explorar las revistas por categoría temática o por país de publicación. Para cada revista se proporcionan varios indicadores bibliométricos: índice h5 y mediana h5 (proporcionados por GSm), número de citas h (suma de citas de los documentos que contribuyen al índice h5), así como el índice h5 y el número de citas h después de haber eliminado las autocitas de revista (referencias citadas a la misma revista donde se ha publicado el documento citante). Por último, también se calcularon los cuartiles de revistas en cada categoría.

El 43% (3.944) de las revistas AHSS de GSM (aquellas que aparecen en JSM) no están cubiertas por WoS/JCR o Scopus/SJR (Figura 1). Además, aunque un gran número de revistas en las categorías AHSS de SJR parecen no estar indizadas en GSM, esto es en realidad un artefacto del método de análisis, en el que no se está utilizando una clasificación unificada de revistas para las tres fuentes. En la clasificación de categorías de SJR, revistas como *Science*, *JAMA Psychiatry*, or *Brain*, están clasificadas como revistas de Arte y Humanidades, mientras en la clasificación de revistas JSM estas revistas no fueron clasificadas como AHSS, aunque obviamente también están indizadas en GSM.

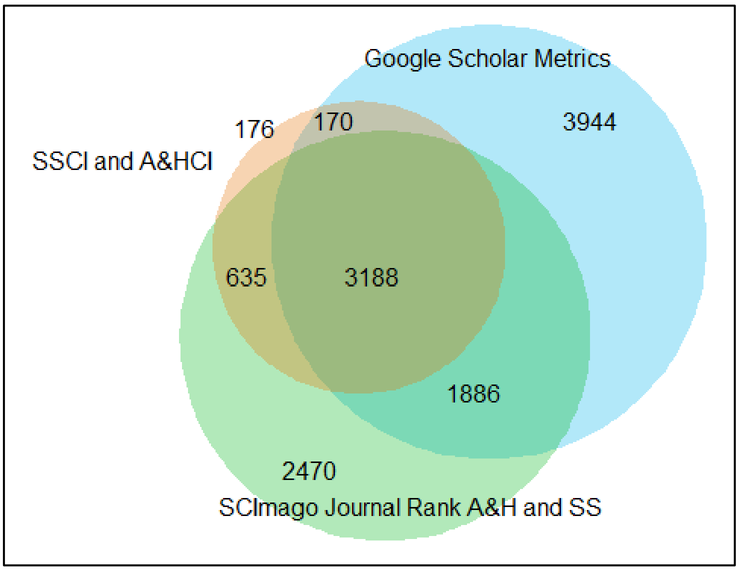


Figura 1. Comparación de la cobertura de revistas en GSM, WoS, y SJR.

Las revistas AHSS en GSM presentan una mayor diversidad en lo que respecta a países de publicación (Figura 2) e idiomas de publicación (Figura 3) que las revistas AHSS de WoS/JCR y Scopus/SJR. En las tres fuentes, Estados Unidos y el Reino Unido son los países donde se publican más revistas. Sin embargo, en GSM el porcentaje de revistas publicadas en estos países está por debajo del 50%, mientras que en WoS alcanza casi el 70%. SJR se encuentra entre estos dos valores, con casi el 60% de sus revistas publicadas ya sea en EE.UU. o el Reino Unido. Los resultados por idioma de publicación revelan una situación parecida: en GSM hay casi un 60% de revistas publicadas en inglés, mientras que en WoS la proporción es casi el 75%.

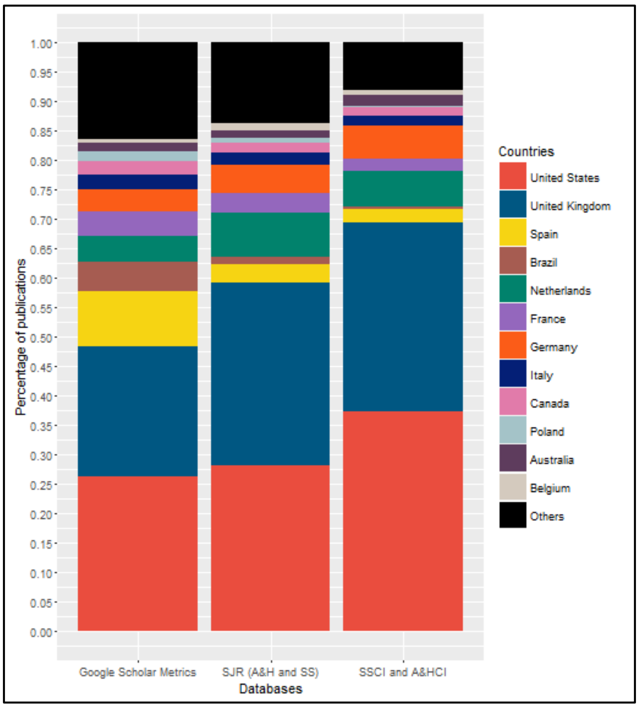


Figura 2. Distribución de revistas por país de publicación en GSM, SJR, y WoS

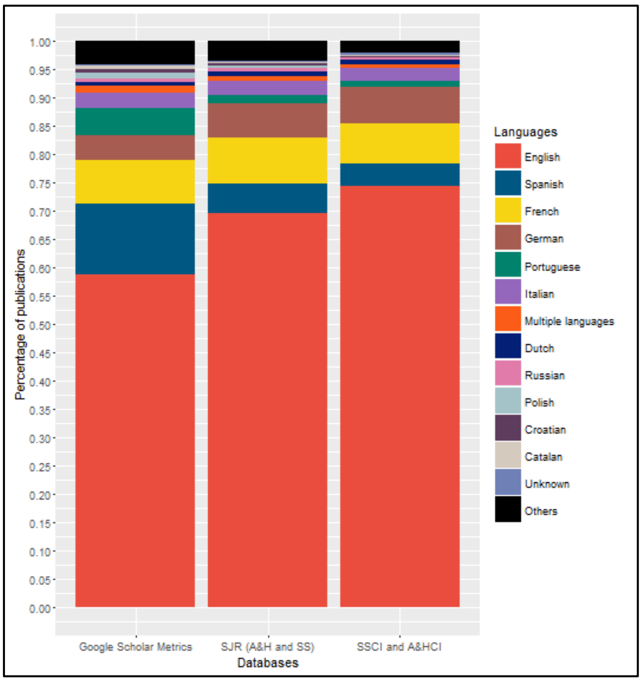


Figura 3. Distribución de revistas por idioma de publicación en GSM, SJR, y WoS

Generación de una representación multifacetada de una comunidad académica

En los últimos años ha habido una proliferación de herramientas gratuitas para general perfiles de investigadores. Cada de estas herramientas tiene una base documental específica, proporciona un conjunto determinado de indicadores, y es capaz de atraer a un grupo de investigadores diferente. Este escenario nos condujo a la idea de "Scholar Mirrors" (espejos académicos). En una casa de espejos (también llamada laberinto de espejos), cada espejo muestra un reflejo de la persona que está enfrente del mismo, pero este reflejo siempre está distorsionado de una manera u otra, dependiendo de las imperfecciones del propio espejo. De igual manera, los indicadores proporcionados por cualquier herramienta de generación de perfiles dependen en gran medida de la cobertura de su base documental (para indicadores de producción y de citas), y del sector demográfico que predomine en su base de usuarios (para indicadores de uso y/o de atención). Cada perfil por tanto proporciona una representación más o menos distorsionada del trabajo de un autor y del impacto que este ha tenido en su comunidad. Enfrentados con esta situación, decidimos intentar generar una representación de una comunidad académica mediante la combinación de información de varias fuentes.

Nuestro primer intento se materializó en la aplicación web "La Biblioteconomía y Documentación española en Google Scholar Citations" (http://www.biblioteconomia-documentacion-española.infoec3.es). En este primer prototipo, tras recoger información en GSC de 336 investigadores españoles en el área de Biblioteconomía y Documentación (algunos de ellos trabajando fuera de España), también recogimos datos de otras fuentes como WoS y ResearchGate. Los resultados se mostraban a nivel de autor, documento, revista, editorial, e institución. Los usuarios podían observar fácilmente cómo los autores y documentos cambiaban posiciones en la lista dependiendo de por qué indicadores la lista fuera ordenada.

En nuestro segundo prototipo seleccionamos una comunidad más grande como objeto de estudio: la comunidad internacional que trabaja en el campo de la bibliometría, cienciometría, informetría, webmetría, y altmetría (Martín-Martín, Orduna-Malea, Ayllón, & Delgado-López-Cózar, 2016) (capítulo 10 de esta tesis). En este proyecto, el método que seguimos para extraer y reorganizar los datos fue refinado en lo que llamamos MADAP (Análisis Multifacetado de Disciplines a través de Perfiles Académicos, por sus siglas en inglés) (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018a) (capítulo 12 de esta tesis). Este método incluye los siguientes pasos:

- Identificación de los autores y de sus perfiles online: en este paso se combinaron varias estrategias para maximizar el número de perfiles identificados. En total, se identificaron 814 perfiles utilizando las siguientes estrategias:
    - Búsqueda por palabras clave en GSC: para identificar las palabras clave que los autores podrían utilizar en sus perfiles para describirse a si mismos, se analizaron los títulos de los artículos publicados en las revistas core de la disciplina.
    - Afiliación institucional: se buscaron los perfiles con afiliación a centros de investigación y departamentos conocidos por trabajar en el campo. Sin embargo, este método no proporcionó ningún autor que no hubiera sido ya encontrado con la estrategia anterior.
    - Búsqueda por palabras clave en GS: se llevaron a cabo búsquedas en GS para identificar autores con un perfil GSC pero que pudieran no haber rellenado sus áreas de interés en su perfil. Además, los documentos relevantes de autores que no tienen un GSC profile fueron también identificados en estas búsquedas.
- Clasificación de los autores: el campo de la cienciometría y sus ramas derivadas tiene la característica de que atrae investigación de autores que no trabajan normalmente en este tema. Por esta razón, los autores fueron clasificados como especialistas (cuando su producción científica se dedica principalmente a la cienciometría) u ocasionales (autores de otros campos que en ocasiones publican estudios bibliométricos). En nuestra muestra, 396 autores fueron clasificados como especialistas, y 415 como ocasionales.

- Extracción de los datos a nivel de documento: el top 100 de los documentos más citados en los perfiles de GSC de cada uno de los autores fueron extraídos, procesados, y combinados con los documentos extraídos en las búsquedas por palabras clave llevadas a cabo en GS. Con este dataset, se generó un listado del top 1.000 de los documentos más citados en la disciplina. Finalmente, esta lista de artículos clásicos en la disciplina fue usada para generar el listado de las revistas y lo editores de libros más influyentes de la disciplina.

En este proyecto también se utilizaron datos de otras fuentes aparte de GS (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018b) (capítulo 13 de esta tesis). Específicamente, también se buscaron los perfiles de los autores ya identificados en ResearchGate, Mendeley, ResearcherID, y Twitter. De todos los investigadores con un perfil en GSC, el 67% también tenían perfil en ResearchGate, el 41% también en Mendeley, el 40% en ResearcherID, y el 30% en Twitter.

Se extrajeron los indicadores bibliométricos a nivel de autor proporcionados por cada plataforma, y se compararon mediante el coeficiente de correlación Spearman. Los coeficientes muestran que los indicadores a nivel de autor de GSC (todos ellos basados en citas) correlacionan bien con los indicadores a nivel de autor de ResearchGate, como el RG Score, número de publicaciones, Impact Points (un indicador ya descontinuado que sumaba los factores de impacto de las revistas donde el autor había publicado), descargas, visualizaciones, y total de citas. Sin embargo, los indicadores de GSC no correlacionan bien con otros indicadores de RG como el número de seguidores de un investigador en la plataforma, o el número de personas seguidas por el investigador. Los indicadores de GSC también correlacionan relativamente bien con algunos indicadores de Mendeley como el de Readers, número de publicaciones, o la media de Readers por documento, y con los indicadores de producción y de citación de ResearcherID. Los indicadores de GSC no correlacionan bien con los indicadores calculados a partir de datos de Twitter, excepto moderadamente con el total de retweets de un autor y el H retweets (fórmula del índice h aplicada al número de retweets de los tweets publicados por un investigador). Por último, el análisis de componentes principales confirma las diferencias entre los indicadores de conectividad como el número de followers, retweets, etc., con los indicadores basados en uso y los indicadores basados en citas.

Los datos extraídos y procesados para este análisis fueron transformados en una aplicación web llamada "Scholar Mirrors" (http://www.scholar-mirrors.infoec3.es) (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2017b) (capítulo 11 de esta tesis). que presenta los datos mencionados arriba a nivel de autores, documentos (top 1.000 más citados), revistas, y editoriales de libros.

Trabajar en estos prototipos nos permitió identificar las dificultades, limitaciones y problemas de cada plataforma de perfiles (Martín-Martín et al., 2016) (capítulo 10 de esta tesis). Con este conocimiento, hemos empezado a diseñar y a trabajar en una aplicación más compleja que pueda funcionar como sistema de información científica para todos los investigadores que trabajan en España (capítulo 14 de esta tesis). Hasta ahora, hemos sido capaces de identificar todos los perfiles públicos en GSC de investigadores que trabajan en España (más de 43.000 en 2017 cuando realizamos la búsqueda), extraer la lista de documentos publicados por cada uno de los investigadores (más de dos millones de documentos únicos), y extraer todas las citas a estos documentos (casi 25 millones de citas). Por último, con el set de citas y un algoritmo de clustering (Waltman & Van Eck, 2013) hemos sido capaces de agrupar los documentos en clusters de documentos relacionados, que es el primer paso para generar una clasificación automática de documentos. En el futuro nos gustaría continuar con este proyecto y crear una aplicación web que muestre información bibliográfica e indicadores bibliométricos contextualizados por categorías temáticas sobre todos los investigadores que trabajen en España con un perfil público en GSC.

Análisis de niveles de Acceso Abierto usando datos de Google Scholar

GS está continuamente rastreando la web académica en búsqueda de documentos académicos, y es capaz de agrupar versiones de un mismo documento bajo un mismo registro principal. Estas circunstancias permiten que GS sea capaz de determinar si existe alguna versión del texto completo de un documento que sea accesible gratuitamente, incluso si la versión publicada por la editorial no es gratuita. Esto es muy útil para los usuarios que quieren acceder a los documentos, especialmente si no tienen suscripciones al contenido de pago de las editoriales. Pero, además, estos datos también son útiles para medir el número de documentos que están disponibles gratuitamente a varios niveles de agregación (por ejemplo, por países de afiliación, o por categorías temáticas).

En 2016 diseñamos un estudio para realizar este tipo de análisis (Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018) (capítulo 16 de esta tesis). La muestra de estudio fueron todos los artículos y revisiones con DOI publicadas en 2009 o 2014 y cubiertas por los tres principales índices de citas de WoS (Science Citation Index, Social Sciences Citation Index, and Arts & Humanities Citation Index). Esta muestra fue seleccionada en parte para facilitar su análisis una vez los datos fueran recogidos (dada la alta calidad de los metadatos bibliográficos que proporciona WoS, y de la que GS carece), y en parte también para que los resultados fueran comparables con otros estudios que usaban muestras de documentos similares, pero diferentes métodos para medir los niveles de Acceso Abierto. Más de 2,3 milliones de documentos fueron seleccionados en total, de todas las áreas temáticas cubiertas por WoS.

Estos documentos se buscaron en GS uno a uno, proceso que llevó aproximadamente tres meses debido a las limitaciones de GS para extraer datos. El 97,6% de los documentos fueron identificados satisfactoriamente, y los links a versiones gratuitas del texto completo que GS había identificado fueron recogidos. Cada link se clasificó en uno de los tipos de Acceso Abierto (dorado, híbrido, con retraso, o verde) o bien como FA (disponible gratuitamente, pero sin pertenecer a ninguna de las categorías anteriores).

Se desarrolló una aplicación web para agregar los resultados a varios niveles: por revista, por año de publicación, por categoría temática, y por país de afiliación de los autores (capítulo 15 de esta tesis). La aplicación genera una tabla resumen usando los parámetros especificados por el usuario, y también genera una tabla de frecuencias filtrada por los documentos incluidos en la selección del usuario, que muestra los dominios web donde se han encontrado más versiones gratuitas de los documentos. Finalmente, la aplicación también es capaz de generar una gráfica de columnas apiladas que represente parte de la información de la tabla resumen (aquellas filas en las que el usuario esté más interesado).

Los resultados del análisis de estos datos muestran que, en general, GS encontró al menos una versión gratuita para el 54,7% de los documentos en la muestra (Martín-Martín, Costas, et al., 2018) (capítulo 16 de esta tesis). El 7,3% estaban disponibles gratuitamente al ser publicados en revistas de Acceso Abierto (ruta dorada), el 1,1% estaban publicados en revistas híbridas, el 1,5% en revistas que convierten los artículos a Acceso Abierto tras un periodo de embargo (Acceso Abierto con retraso), el 13,2% estaban publicados en revistas que todos o algunos de sus artículos accesibles de manera gratuita, pero no establecen una licencia compatible con el Acceso Abierto (recientemente denominado Bronce), el 17,6% estaban disponibles desde repositorios (ruta verde), y el 40,6% estaban disponibles gratuitamente desde otras fuentes sin ajustarse a la definición de Acceso Abierto. Principalmente en esta categoría encontramos los documentos disponibles desde ResearchGate (32,6% del total de la muestra).

References

Martín-Martín, A., Ayllón, J. M., Orduña-Malea, E., & Delgado-López-Cózar, E. (2014). *Google Scholar Metrics 2014: a low cost bibliometric tool* (EC3 Working Papers No. 17). Retrieved from http://arxiv.org/abs/1407.2827

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, *12*(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2016). *The counting* *house, measuring those who count: Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in GSC (Google Scholar Citations), ResearcherID, ResearchGate, Mendeley, & Twitter* (EC3 Working Papers No. 21). Retrieved from https://arxiv.org/abs/1602.02412

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017a). Journal Scholar Metrics: building an Arts, Humanities, and Social Sciences journal ranking with Google Scholar data. In *22nd International Conference on Science, Technology & Innovation Indicators (STI)*. Paris. https://doi.org/10.17605/OSF.IO/VXNW6

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017b). Scholar Mirrors: Integrating evidence of impact from multiple sources into one platform to expedite researcher evaluation. In *22nd International Conference on Science, Technology & Innovation Indicators (STI)*. https://doi.org/10.31235/osf.io/z4bwe

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018a). A novel method for depicting academic disciplines through Google Scholar Citations: The case of Bibliometrics. *Scientometrics*, *114*(3), 1251–1273. https://doi.org/10.1007/s11192-017-2587-4

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018b). Author-level metrics in the new academic profile platforms: The online behaviour of the Bibliometrics community. *Journal of Informetrics*, *12*(2), 494–509. https://doi.org/10.1016/j.joi.2018.04.001

Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B*, *86*(11). https://doi.org/10.1140/epjb/e2013-40829-0

Chapter 9. Journal Scholar Metrics: building an Arts, Humanities, and Social Sciences journal ranking with Google Scholar data

Conference communication. Cite as:

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017). Journal Scholar Metrics: building an Arts, Humanities, and Social Sciences journal ranking with Google Scholar data. In 22nd International Conference on Science, Technology & Innovation Indicators (STI). Paris. https://doi.org/10.17605/OSF.IO/VXNW6

Abstract (English)

This paper describes the creation of "Journal Scholar Metrics" (JSM), a prototype web application that ranks journals in the areas of Arts, Humanities, and Social Sciences (AH&SS) on the basis of the citations their articles have received according to Google Scholar Metrics (GSM). To identify as many AH&SS journals as possible, a master list of 66,454 journals covered by various databases was developed. All AH&SS journals in that list were searched on GSM. Additionally, a series of keyword searches were carried out to identify journals covered by GSM which weren't present in the master list. A total of 9,188 AH&SS journals with names written in Latin characters were found in the 2015 edition of GSM (which displays data about articles published between 2010 and 2014). Besides the journal-level indicators provided by GSM (H5-index and H5-median), several additional indicators were computed (H5-citations, H5-index and H5-citations without journal self-citations, and journal self-citation rate). Journals are displayed by subject categories and by country of publication. Quartiles were computed for each category, and journals in a category were further classified either as core (high affinity to the category) or related (partial affinity). A detail page for each journal is also available, displaying journal indicators, as well as a list of other databases were the journal is indexed.

Abstract (Spanish)

Esta comunicación describe la creación de "Journal Scholar Metrics" (JSM), un prototipo de aplicación web que ofrece rankings de revistas en las áreas de Arte, Humanidades, y Ciencias Sociales basado en las citas que han recibido según la herramienta Google Scholar Metrics (GSM). Para identificar el mayor número de revistas posible, primero se desarrolló una master list de 66,454 revistas a partir de la cobertura de varias bases de datos. Todas las revistas de Arte, Humanidades y Ciencias Sociales en este master list fueron buscadas en GSM. Además, se llevaron a cabo una serie de búsquedas por palabras clave, para identificar revistas en GSM que pudieran no estar cubiertas en la master list. Se encontraron un total de 9.188 revistas con nombres escritos en caracteres latinos en la edición de 2015 de GSM (que proporciona datos sobre artículos publicados entre 2010 y 2014). Además de los indicadores a nivel de artículo proporcionados por GSM (H5-index y H5-median), se calcularon una serie de indicadores adicionales (H5-citations, H5-index y H5-citations sin autocitas de revista, y el ratio de autocitación de revista). Las revistas se presentan por categorías temáticas y por país de publicación. También se calcularon cuartiles para cada categoría. Las revistas de cada categoría fueron clasificadas como "core" (alta afinidad con la categoría) y relacionadas (afinidad parcial). Cada revista tiene una página de detalle, donde se muestran sus indicadores, además de una lista de base de datos donde está indexada.

1. Introduction

In April 2012 Google Scholar launched Metrics (commonly known as Google Scholar Metrics, or GSM for short), a complementary tool based on Google Scholar data and designed to become an "easy way for authors to quickly gauge the visibility and influence of recent articles in scholarly publications" (http://googlescholar.blogspot.com.es/2012/04/google-scholar-metrics-for-publications.html). Despite its original purpose, GSM was immediately perceived by the scientific community as a new bibliometric tool, since it computed the H-index for a wide range of scientific journals and other bibliographic sources (conferences and repository collections). In order to be included in GSM, journals or conferences must meet certain requirements: their articles must be indexed in Google Scholar; they must have published at least 100 articles over a period of five years; lastly, those articles must have received at least one citation. These criteria represent an effort (albeit arguable a crude one) to filter periodical publications from other types of documents indexed in Google Scholar.

Journal rankings in GSM are presented by languages. In the first two editions of the product there were ten different languages available (English, Chinese, Portuguese, German, Spanish, French, Korean, Japanese, Dutch, and Italian). In the 2015 edition (which displayed indicators calculated from documents published between 2010 and 2014) there were nine, because the Korean ranking was discontinued. In the 2016 edition (2011-2015) there were twelve languages, bacause five additional languages were added (Russian, Korean, Polish, Ukrainian & Indonesian) and two were removed: Italian, and Dutch.

For publications written in English, however, GSM also groups journals in 8 broad subject categories (Business, Economics & Management; Chemical & Material Sciences; Engineering & Computer Science, Health & Medical Sciences; Humanities, Literature & Arts; Life Sciences & Earth Sciences; Physics & Mathematics; Social Sciences), and 261 subcategories. For each journal, the list of documents with a citation count that is equal or higher than the h5-index of the journal can also be consulted. For each one of these documents, in turn, it is also possible to consult the list of citing documents.

Unfortunately, Google Scholar Metrics presents a rather restrictive visualization system. Only the top 100 sources according to their h5-index are displayed when selecting any of the language or broad subject category rankings. As for the subcategory rankings and the queries that can be made through the available search box, only the top 20 sources according to their h5-index are displayed. This effectively means that there is no straightforward method to learn how many journals are indexed in GSM, and it also means that most of the journals in GSM haven't been assigned to any subject category or subcategory (at least publicly). What's more, GSM doesn't allow grouping and ordering journals according to their country of publication.

In order to overcome these limitations, we took advantage of the various search features available in GSM's search box, and set out to collect all the Art, Humanities, and Social Science journals indexed in this product that we could find. The main goal of this project is, therefore, to gauge the extent of the journal coverage in Google Scholar Metrics, focusing our efforts in the areas of Arts, Humanities, and Social Sciences, which are also the areas that have been historically neglected by other journal rankings (JCR, SJR). The result of this work is available through a freely accessible web application which we called Journal Scholar Metrics (http://www.journal-scholar-metrics.infoec3.es).

Journal Scholar Metrics focus exclusively on journals belonging to the areas of Arts, Humanities, and Social Sciences, since these are the areas that have traditionally presented more difficulties in terms of bibliometric assessment, and the ones for which there is a greater lack of international, geographically and linguistically unbiased tools. In these disciplines, where research is often oriented towards local interests and where cultural peculiarities are determinant, researchers usually use national channels - and their native language- to communicate their results. This is why Google Scholar, thanks to its robots that automatically index all seemingly scientific publications without any kind of geographic or linguistic restriction, is currently the most appropriate source of data to find evidence of scientific impact in these areas.

2. Methods

Journal Scholar Metrics focus exclusively on Arts, Humanities, and Social Sciences journals indexed in GSM. It covers journals from all around the world and in all languages, providing that the name of the journal is displayed in Latin characters in GSM. Thus, names of journals written only in Arabic, Cyrillic, Chinese, Korean, or Japanese characters were excluded.

2.1 Making a master list of journals

In order to identify all journals that could potentially have been indexed in GSM, the following journal databases were consulted:

a) Ulrichsweb: Global Serials Directory. It is considered the largest directory of periodic publications in the world. We retrieved the list of all existing scientific journals (academic/scholarly) indexed in the 162 categories («subjects») concerning Arts, Humanities, and Social Sciences. Data retrieved on September, 2013.

b) Web of Science Master Lists: journals indexed in the 84 subject categories included in the Arts & Humanities Citation Index and the Social Science Citation Index. Data retrieved on May, 2015.

c) SCImago Journal Rank: journals indexed in the 64 subject categories included in the areas related to the Humanities and Social Sciences. Data retrieved on December, 2015.

d) The journal coverage lists of various international disciplinary-based databases: Anthropological Index Online, ARTBibliographies Modern, Avery Index, Communication & Mass Media, Econlit, Education Resources Information Center (ERIC), Geobase, Historical Abstracts, Index Islamicus, L'Année philologique, Library and Information Science Abstracts (LISA), MLA (Linguistics and Literature), Philosopher's Index, PsychINFO, Répertoire International de Littérature Musicale (RILM), Social Work Abstracts, Sociological Abstracts, SPORTDiscus, and Worldwide Political Science Abstracts. Data retrieved between October 2015 and February 2016.

e) Google Scholar Metrics: All journals indexed in the categories "Humanities, Literature & Arts", "Social Sciences", and "Business, Economics & Management" were downloaded.

Using the information retrieved from these databases, a master list of journals was developed where titles were unified and duplicates were merged, resulting in an exhaustive database that combined the information contained in all the databases described above. The bibliographic information contained in Ulrichsweb was given precedence in case of any discrepancies in the data available from other sources.

2.2 Finding journals in GSM

First, we searched all journals in our master list on GSM by their names (GSM doesn't support searching by ISSN, publisher, or any other field). Secondly, in order to find journals in GSM which weren't already in our master list, we performed topic searches using meaningful keywords from each discipline. The selection of these keywords was based on the analysis of the frequency of words which appeared in the names of the journals in the master list. These keywords were also translated to several languages: English, French, Spanish, German, Italian, Portuguese, Polish, and Czech.

2.3 Data processing

Once we identified all the available journals in GSM and matched them to the journals in our master list (relying on journal title comparisons), we proceeded to download the following data:

- Journal metrics (H5-index and H5-median).
- Bibliographic information of the articles published in each of those journals, including the number of citations these articles have received. Only the articles that contribute to the H5-index of the journals are displayed in GSM (nº of citations ≥ H5-index), and therefore only those were downloaded.
- Bibliographic information of the documents that cite the aforementioned articles.

After downloading this information, journal self-citations were identified and computed.

2.4 Journal classification

All journals found in GSM were assigned to at least one of 13 custom subject categories in Social Sciences (Anthropology, Communication, Economics, Business, & Management, Education, Geography & Urbanism, Law, Library & Information Science, Political Science, Administration, & International Relations, Psychology, Social Work, Sociology, Sports Science, Multidisciplinary) and 9 custom subject categories in Arts & Humanities (Archaeology & Prehistory, Arts, Classical Studies, History, Literature, Language & Linguistics, Philosophy, Religion, Multidisciplinary).

A journal assigned to a subject category is considered either a core journal, or a related journal. Core journals are those publications which are considered essential in a category. Related journals are those which are linked to the category because some (but not all) of the articles they publish are relevant to the category. In order to make these distinctions we relied on the classifications made by other databases (Ulrichsweb, Web of Science, Scopus, disciplinary-based databases, etc.). As a general rule, journals were considered as core in a given category when they met at least one of the following criteria:

a) The name of the journal contains meaningful keywords from the discipline. Word frequency lists extracted from lists of journal names were used. Polysemic words (such as “information”) were excluded.

b) The journal is considered as a core by the specialized database in the field, and it is also classified in Ulrichsweb in the corresponding category. Only some specialized databases make the core/related distinction (or similar ones). Therefore, none the journals covered by the databases that don’t make this distinction met this specific criterion.

c) The journal is classified in the corresponding categories in at least three of the following databases: Ulrichsweb, WoS, SJR, GSM (only journals written in English), and the specialized database for the field (as core or related).

Journals that didn’t meet any of the previous criteria but were classified in the corresponding categories in two of the aforementioned databases were considered as related.

Lastly, a manual revision was also carried out to check and fix incongruities. Changes were made on a case-by-case basis after examining the scope of the journals (articles published, relationships with other journals…).

2.5 Bibliometric indicators

Six bibliometric indicators are used:

1. *H5-index*: h-index for articles published in the last 5 complete years. It is the largest number h such that h articles published in 2010-2014 have at least h citations each.
2. *H5-median*: median number of citations for the articles that make up a journal’s h5-index.

3. *H5-citations*: sum of the number of citations received by all the articles that make up the journal's h5-index.

4. *H5-index without journal self-citations*: computed in the same way as the H5-index, but excluding citations that come from articles published in the same journal.

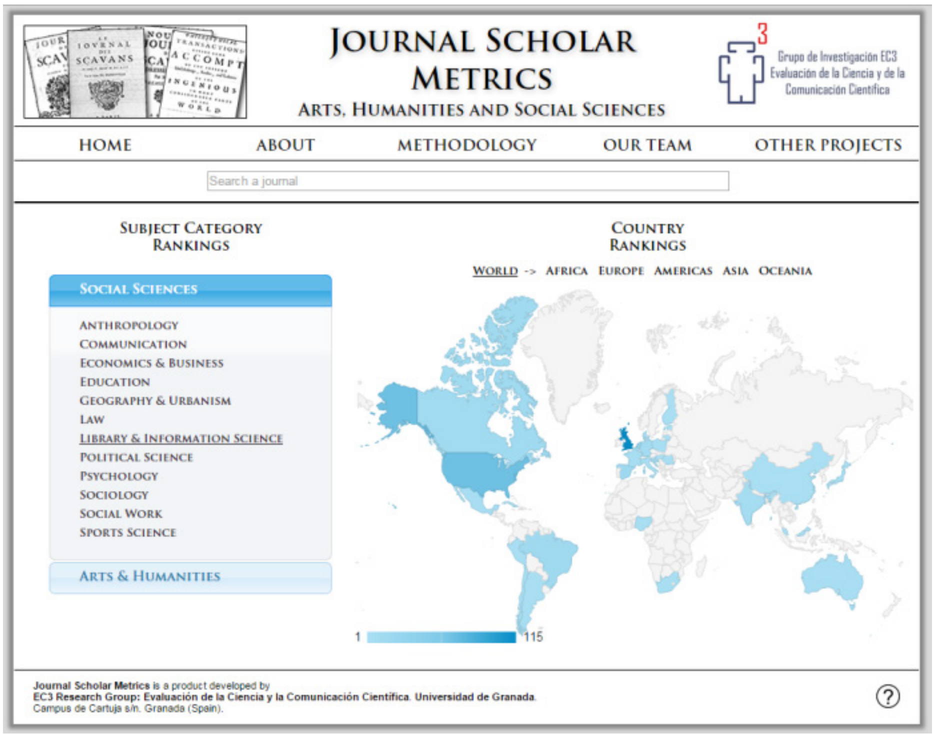
5. *H5-citations without journal self-citations*: computed in the same way as the H5-citations, but excluding citations that come from articles published in the same journal.

6. *Journal self-citation rate*: Percentage of citations that come from articles published in the same journal.

3. Results

We found 9,188 Arts, Humanities, and Social Sciences journals with titles written in Latin characters in the 2015 edition of Google Scholar Metrics (which displays data from articles published between 2010 and 2014). We are confident we found most of them, but it is quite possible that we may have missed some of them, especially if their h-index is low (5 or lower) or if they use obscure names for the title of the journal (proper nouns, words coming from Latin or ancient Greek, acronyms, etc.).

Journals are presented both by subject category and country rankings. These options can be selected directly from the Home Page of the web application (Figure 1).

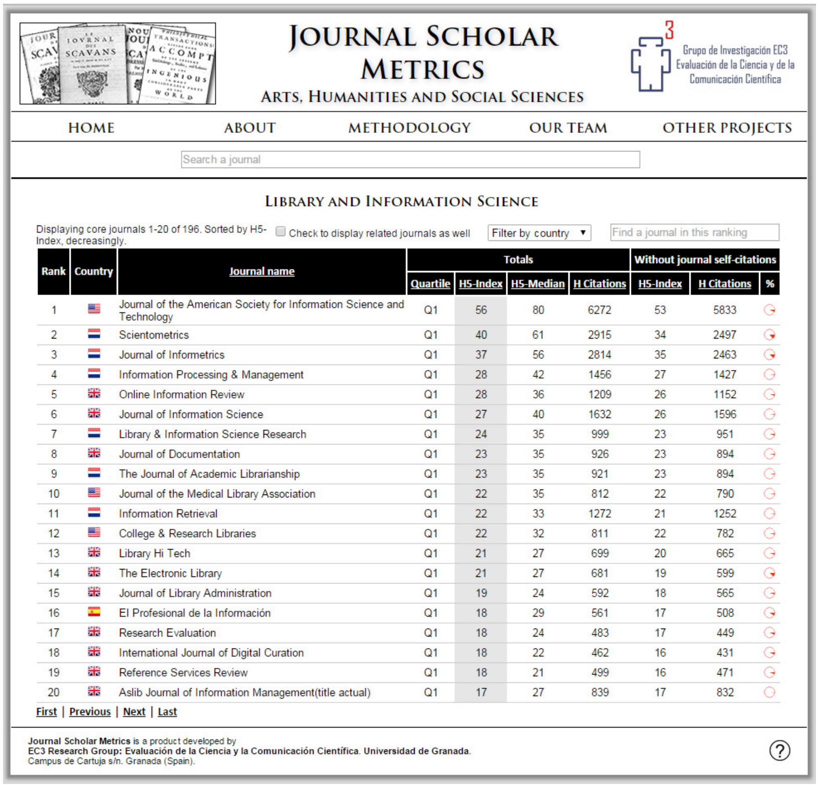


Journal Scholar Metrics is a product developed by
EC3 Research Group: Evaluación de la Ciencia y la Comunicación Científica. Universidad de Granada.
Campus de Cartuja s/n. Granada (Spain).

Figure 1: Home Page Journal Scholar Metrics

3.1 Browsing by subject category

Subject rankings can be sorted by any of the six bibliometric indicators previously described, as well as by the country of publication, just by clicking on their respective column headers (Figure 2).


Journal Scholar Metrics
Arts, Humanities and Social Sciences
Grupo de Investigación EC3 Evaluación de la Ciencia y de la Comunicación Científica
HOME
ABOUT
METHODOLOGY
OUR TEAM
OTHER PROJECTS
Search a journal
Filter by country ▾
Find a journal in this ranking
First | Previous | Next | Last

Library and Information Science

Displaying core journals 1-20 of 196. Sorted by H5-Index, decreasingly.

□ Check to display related journals as well

Rank	Country	Journal name	Totals				Without journal self-citations		
			Quartile	H5-Index	H5-Median	H Citations	H5-Index	H Citations	%
1	🇺🇸	Journal of the American Society for Information Science and Technology	Q1	56	80	6272	53	5833	○
2	🇳🇱	Scientometrics	Q1	40	61	2915	34	2497	○
3	🇳🇱	Journal of Informetrics	Q1	37	56	2814	35	2463	○
4	🇳🇱	Information Processing & Management	Q1	28	42	1456	27	1427	○
5	🇬🇧	Online Information Review	Q1	28	36	1209	26	1152	○
6	🇬🇧	Journal of Information Science	Q1	27	40	1632	26	1596	○
7	🇳🇱	Library & Information Science Research	Q1	24	35	999	23	951	○
8	🇬🇧	Journal of Documentation	Q1	23	35	926	23	894	○
9	🇳🇱	The Journal of Academic Librarianship	Q1	23	35	921	23	894	○
10	🇺🇸	Journal of the Medical Library Association	Q1	22	35	812	22	790	○
11	🇳🇱	Information Retrieval	Q1	22	33	1272	21	1252	○
12	🇺🇸	College & Research Libraries	Q1	22	32	811	22	782	○
13	🇬🇧	Library Hi Tech	Q1	21	27	699	20	665	○
14	🇬🇧	The Electronic Library	Q1	21	27	681	19	599	○
15	🇬🇧	Journal of Library Administration	Q1	19	24	592	18	565	○
16	🇪🇸	El Profesional de la Información	Q1	18	29	561	17	508	○
17	🇬🇧	Research Evaluation	Q1	18	24	483	17	449	○
18	🇬🇧	International Journal of Digital Curation	Q1	18	22	462	16	431	○
19	🇬🇧	Reference Services Review	Q1	18	21	499	16	471	○
20	🇬🇧	Aslib Journal of Information Management(title actual)	Q1	17	27	839	17	832	○

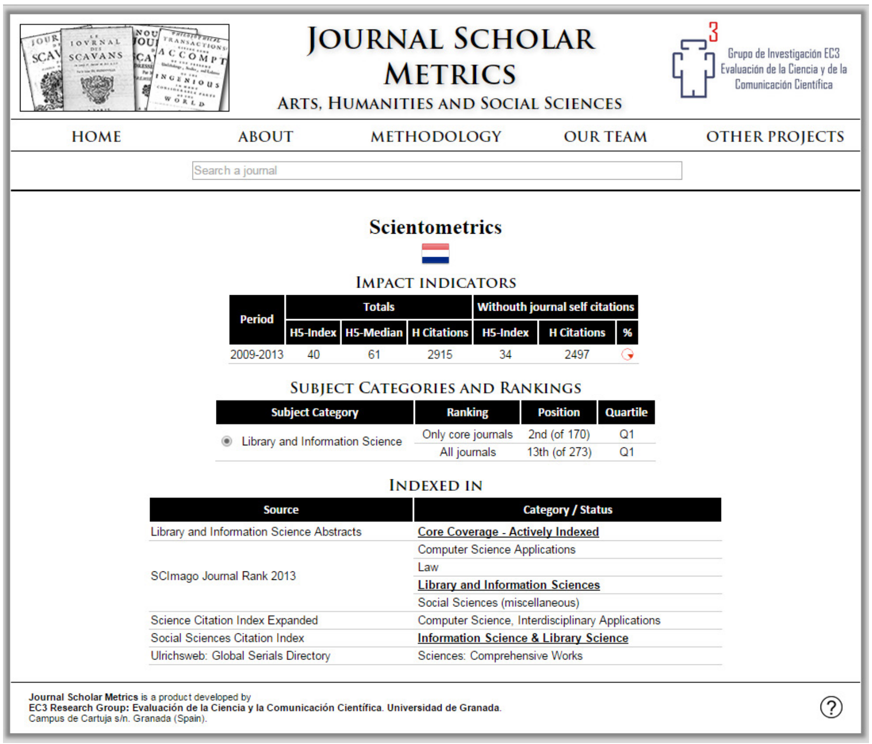
Journal Scholar Metrics is a product developed by
EC3 Research Group: Evaluación de la Ciencia y la Comunicación Científica. Universidad de Granada.
Campus de Cartuja s/n. Granada (Spain).

Figure 2: Page Subject categories

Above the results table users will find three elements:

- A check box to show core or related journals. When the box is checked, core journals are highlighted using bold letters, and related journals are slightly translucent. When the box is not checked, only core journals are shown.
- A drop-down list to filter by country of publication.
- A search box from which it is possible to search any journal in the current category.

Clicking on the title of a journal in any ranking takes users to the profile page of that journal (Figure 3). That page will show basic information for that journal, like: Name of the journal and country of publication, metrics, subject categories to which the journal has been assigned and position it occupies in the category, and a list of databases where this journal is also indexed, as well as the categories these databases have assigned to the journal, according to their respective classification schemes.



Journal Scholar Metrics

Arts, Humanities and Social Sciences

Grupo de Investigación EC3
Evaluación de la Ciencia y de la Comunicación Científica

HOME ABOUT METHODOLOGY OUR TEAM OTHER PROJECTS

Search a journal

Scientometrics

Impact indicators

Period	Totals			Withouth journal self citations		
	H5-Index	H5-Median	H Citations	H5-Index	H Citations	%
2009-2013	40	61	2915	34	2497	

Subject Categories and Rankings

Subject Category	Ranking	Position	Quartile
◉ Library and Information Science	Only core journals	2nd (of 170)	Q1
	All journals	13th (of 273)	Q1

Indexed in

Source	Category / Status
Library and Information Science Abstracts	**Core Coverage - Actively Indexed**
SCImago Journal Rank 2013	Computer Science Applications
	Law
	Library and Information Sciences
	Social Sciences (miscellaneous)
Science Citation Index Expanded	Computer Science, Interdisciplinary Applications
Social Sciences Citation Index	**Information Science & Library Science**
Ulrichsweb: Global Serials Directory	Sciences: Comprehensive Works

Journal Scholar Metrics is a product developed by
EC3 Research Group: Evaluación de la Ciencia y la Comunicación Científica. Universidad de Granada.
Campus de Cartuja s/n. Granada (Spain).

Figure 3: Journal Profile Page

3.2 Browsing by country

The Home Page also presents a world map from which users can access all journals in this product published in a given country, regardless of their topic. Since there might be difficulty in clicking some countries in the world map because of their size, it is also possible to display continent maps, as well as smaller territories. The shade of blue indicates the quantity of journals in JSM published in that country, and a box with the name of the country and the exact number of journals collected so far in JSM will pop up when you hover over it with the cursor. When a country is selected, all journals published in that country will be shown.

4. Conclusions

In previous studies (Martín-Martín et al. 2014; Orduna-Malea et al. 2016), we have deeply described the underlying philosophy embedded in all Google's academic products. These tools have been created in the image and likeness of Google's general search engine: fast, simple, easy to use, and last but not least, accessible to everyone free of charge. GSM follows all these precepts, and it is a hybrid between a bibliometric tool (indicators based on citation counts), and a minimalist information product with few features, closed (it cannot be customized by the user), and simple (navigating it only takes a few clicks) (Jacsó 2012; Delgado López-Cózar & Cabezas 2012; 2013).

Journal Scholar Metrics was designed to prove that some of GSM's shortcomings can be overcome. This product uses both advanced query search procedures and journal data processing to solve GSM's limitations. Moreover, this product is focused in the Arts, Humanities and Social Sciences, precisely the areas that have been more neglected in terms of coverage by other traditional bibliographic databases, thus surfacing the wealth of data available in Google Scholar that would have otherwise stayed unknown to most users.

As a proof of this, the number of journals found in the subject categories ("Humanities, Literature & Arts", "Social Sciences", and "Business, Economics & Management") offered by GSM (2010-2014 edition) only amounts to approximately 1,800. By perusing the language rankings we can find around 300 more.

However, Journal Scholar Metrics displays 9,188 journals in these categories. Moreover, the journals available in GSM are only a fraction of the journals available in Google Scholar, because GSM's inclusion criteria leave out a great number of Humanities and Social Sciences journals that don't publish at least 100 in five consecutive years.

In addition to giving more visibility to these journals, Journal Scholar Metrics has offered to the community the following methodological findings:

A. A procedure to compute the impact of journals excluding journal self-citations using Google Scholar data, presenting two additional indicators in this regard: sum of citations to articles that contribute to the h5-index (including and excluding journal self-citations), as well as the percentage of citations that are journal self-citations. The h5-index excluding journal self-citations is also displayed.

B. A new journal classification system: journals are classified as core or related on a given subject category depending on the level to which they are linked to said subject category. We based this scheme on the classification systems of other multidisciplinary and specialized bibliographic databases and journal directories.

References

Delgado López-Cózar, E. & Cabezas-Clavijo, A. (2013). Ranking journals: could Google Scholar Metrics be an alternative to Journal Citation Reports and Scimago Journal Rank? *Learned Publishing*, 26, 101-113.

Delgado López-Cózar, E. & Cabezas-Clavijo, Á. (2012). Google Scholar Metrics: an unreliable tool for assessing scientific journals. *El Profesional de la Información*, 21, 419–427.

Jacsó, P. (2012). Google Scholar Metrics for Publications: the software and content features of a new open access bibliometric service. *Online Information Review*, 36, 604–619.

Martín-Martín, A., Ayllón, J.M., Orduna-Malea, E. & Delgado López-Cózar, E. (2014). Google Scholar Metrics 2014: a low cost bibliometric tool. *EC3 Working Papers*,17. Retrieved April 28, 2017 from: http://arxiv.org/pdf/1407.2827

Orduna-Malea, E.; Martín-Martín, A.; Ayllón, J.M. & Delgado López-Cózar, E. (2016). *La revolución Google Scholar: destapando la caja de Pandora académica*. Granada: UNE.

Chapter 10. The counting house, measuring those who count: presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in Google Scholar Citations, ResearcherID, ResearchGate, Mendeley, & Twitter

Working paper. Cite as:

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2016). The counting house, measuring those who count: Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in GSC (Google Scholar Citations), ResearcherID, ResearchGate, Mendeley, & Twitter (EC3 Working Papers No. 21). Retrieved from https://arxiv.org/abs/1602.02412

Abstract (English)

Following in the footsteps of the model of scientific communication, which has recently gone through a metamorphosis (from the Gutenberg galaxy to the Web galaxy), a change in the model and methods of scientific evaluation is also taking place. A set of new scientific tools are now providing a variety of indicators which measure all actions and interactions among scientists in the digital space, making new aspects of scientific communication emerge. In this work we present a method for “capturing” the structure of an entire scientific community (the Bibliometrics, Scientometrics, Informetrics, Webometrics, and Altmetrics community) and the main agents that are part of it (scientists, documents, and sources) through the lens of Google Scholar Citations (GSC).

Additionally, we compare these author “portraits” to the ones offered by other profile or social platforms currently used by academics (ResearcherID, ResearchGate, Mendeley, and Twitter), in order to test their degree of use, completeness, reliability, and the validity of the information they provide. A sample of 814 authors (researchers in Bibliometrics with a public profile created in GSC) was subsequently searched in the other platforms, collecting the main indicators computed by each of them. The data collection was carried out on September, 2015. The Spearman correlation (α= 0.05) was applied to these indicators (a total of 31), and a Principal Component Analysis was carried out in order to reveal the relationships among metrics and platforms as well as the possible existence of metric clusters.

We found that it is feasible to depict an accurate representation of the current state of the Bibliometrics community using data from GSC (the most influential authors, documents, journals, and publishers). Regarding the number of authors found in each platform, GSC takes the first place (814 authors), followed at a distance by ResearchGate (543), which is currently growing at a vertiginous speed. The number of Mendeley profiles is high, although 17.1% of them are basically empty. ResearcherID is also affected by this issue (34.45% of the profiles are empty), as is Twitter (47% of the Twitter accounts have published less than 100 tweets). Only 11% of our sample (93 authors) have created a profile in all the platforms analyzed in this study. From the PCA, we found two kinds of impact on the Web: first, all metrics related to academic impact. This first group can further be divided into usage metrics (views and downloads) and citation metrics. Second, all metrics related to connectivity and popularity (followers). ResearchGate indicators, as well as Mendeley readers, present a high correlation to all the indicators from GSC, but only a moderate correlation to the indicators in ResearcherID. Twitter indicators achieve only low correlations to the rest of the indicators, the highest of these being to GSC (0.42-0.46), and to Mendeley (0.41-0.46).

Lastly, we present a taxonomy of all the errors that may affect the reliability of the data contained in each of these platforms, with a special emphasis in GSC, since it has been our main source of data. These errors alert us to the danger of blindly using any of these platforms for the assessment of individuals, without verifying the veracity and exhaustiveness of the data.

In addition to this working paper, we also have made available a website where all the data obtained for each author and the results of the analysis of the most cited documents can be found: Scholar Mirrors (http://www.scholar-mirrors.infoec3.es/).

Abstract (Spanish)

Al igual que el modelo de comunicación científica, que recientemente ha sufrido una metamorfosis (de la Galaxia Gutenberg a la Galaxia Web), el modelo y los métodos de evaluación científica también está pasando por cambios. Un conjunto de nuevas herramientas proporcionan una variedad de indicadores que miden todas las acciones e interacciones entre científicos en el espacio digital, haciendo emerger nuevos aspectos de la comunicación científica. En este trabajo presentamos un método para “capturar” la estructura de una comunidad científica (la comunidad que trabaja en el campo de Bibliometría, Cienciometría, Informetría, Webometría, y Altmetría), de los principales agentes y elementos que forman parte de la misma (investigadores, documentos, y fuentes), a través de la lente de Google Scholar Citations (GSC).

Además, comparamos los “retratos” de autores mostrados en otras plataformas sociales o de perfiles que actualmente usan los investigadores (ResearcherID, ResearchGate, Mendeley, and Twitter), con el objetivo de analizar su grado de uso, exhaustividad, fiabilidad, y validez de la información que proporcionan. Para esto, una muestra de 814 autores (investigadores en el campo de la Bibliometría que tienen un perfil público en GSC) fue buscada en las otras plataformas, y se extrajeron los principales indicadores calculados por cada una de ellas. Este proceso de extracción de datos fue llevado a cabo en septiembre de 2015. Se aplicó la correlación Spearman (α= 0.05) a estos indicadores (un total de 31), y un análisis de componentes principales fue llevado a cabo para revelar las relaciones entre los indicadores de cada plataforma, así como para detectar posibles agrupaciones de indicadores.

El experimento sugiere que es factible generar una representación del estado actual de la comunidad bibliométrica usando datos de GSC (los autores, documentos, revistas, y editoriales con más influencia). Sobre el número de investigadores encontrados en cada plataforma, GSC está en primer lugar (814), seguida a cierta distancia por ResearchGate (543), que actualmente está creciendo a un ritmo vertiginoso. El número de perfiles en Mendeley es alto, pero muchos de ellos están completamente vacíos. En ResearcherID pasa algo parecido (el 34,45% de los perfiles está vacío). En Twitter, el 47% de las cuentas habían publicado menos de 100 tweets. Solo el 11% de los investigadores en nuestra muestra (93 autores) habían creado un perfil en todas las plataformas analizadas en este estudio. En el análisis PCA, encontramos dos tipos de indicadores en la Web: primero, todos los indicadores de impacto. Este grupo se puede subdividir en métricas de uso (visualizaciones y descargas), y métricas de citas. El segundo grupo son los indicadores de conectividad y popularidad (seguidores). Los indicadores de ResearchGate, así como el indicador Mendeley readers, presentan una alta correlación con los indicadores de GSC, pero solo una correlación moderada con los indicadores de ResearcherID. Los indicadores de Twitter alcanzan solo bajas correlaciones con el resto de los indicadores, siendo las más altas con GSC (0,42-0,46) y con Mendeley (0,41-0,46).

Finalmente, se presenta una taxonomía de todos los errores que afectan a la fiabilidad de los datos ofrecidos por estas plataformas, haciendo especial énfasis en GSC, ya que ha sido nuestra fuente principal de datos. Estos errores nos alertan del peligro de usar ciegamente cualquiera de estas plataformas para evaluar a individuos, sin antes verificar la veracidad y exhaustividad de sus datos.

Además de este working paper, también hemos desarrollado una aplicación web donde se puede navegar y visualizar todos los datos a nivel de autor, así como los resultados del análisis de los documentos más altamente citados: Scholar Mirrors (http://www.scholar-mirrors.infoec3.es/).

1. Introduction

1.1 Disciplines and scientific communities: territories and the tribes of Science

Science, in order to be properly investigated, grasped, and taught, has usually been organized in various areas of knowledge. Over time, each of these areas has been further divided into fields, subfields, disciplines, and specialties, as a result of the ever faster growth of knowledge and the parallel increase in the number of people who form the scientific communities within each of these areas. This process of scientific budding follows the life cycle of a living being (birth, growth, reproduction, and death), and is subject to endless metamorphosis, each discipline displaying its own idiosyncrasies.

Each of these units in which scientific knowledge is structured has its own epistemological properties (its object, its principles, and its methods) that endow them with a characteristic identity as well as boundaries that demarcate their cognitive territory. The inner and outer boundaries are not always clearly defined. There is overlapping between disciplines, gaps, and loops, sometimes quite vague and difficult to trace.

The different areas of knowledge are populated by communities of scientists and professionals, each group using their own tools, methodologies and techniques. These are social groups that share - with more or less consensus - professional practices, forms of work organization, living conditions, social expectations, principles, values, and beliefs.

Whitley (1984) dissected with a precision close to that of a surgeon's scalpel the process by which the academic communities - and their disciplines and specialties - become socially and cognitively institutionalized: how they create organizations that allow them to associate in order to defend their interests, how they erect spaces for the exchange of ideas and social development (conferences, seminars, forums, etc.), how they institute professional (newsletter, discussion list) or scientific (journals) means of communication, how they obtain academic standing by teaching the subject at the university (courses in graduate and postgraduate programs, including Master and PhD degrees), how they create groups, departments, laboratories, and companies dedicated to advance research, how they define research agendas where not only research problems but also ways to address and solve them are addressed, or how to create a common language to establish ideas and principles. Not to mention that the process of social and cognitive institutionalization of disciplines is directly influenced by the geographic location and the different levels of economic and cultural development of the countries where they are based.

As masterfully formulated by Becher and Trowler (2001), there is a close relationship between the disciplines (territories of knowledge) and people who advance them (scientific tribes); between the epistemic properties of the forms of scientific knowledge and the social aspects of academic communities. This is why any analysis of a discipline cannot ignore these two areas: the cognitive (disciplines) and social (community); you cannot understand one without the other.

Therefore, the ultimate aim of this Working Paper is to portray a discipline (Bibliometrics) and those who practice it, because a discipline is what is performed by those who cultivate it. Consequently, identifying the members of the Bibliometric tribe is one of the goals of this work.

1.2. A discipline with many names

There are numerous works which address the history of our field of knowledge (Broadus 1987a; Hertzel 1987; Shapiro 1992; Godin, 2006; De Bellis 2009). Its denomination, object of study, and scope have been addressed as well (Lawani, 1981; Bonitz, 1982; Peritz, 1984; Broadus, 1987b; Brookes, 1988; 1990; Sengupta, 1992; Glänzel & Schoepflin 1994; Braun 1994, Gorbea, 1995; Hood & Wilson, 2001; Cronin, 2001; Thelwall, 2008; Larriviere, 2012). There are also several literature reviews about this subject (Narin & Moll, 1977; White & McCain, 1989; Van Raan, 1997; Wilson, 1999; Borgman & Furner, 2002).

Bibliometrics can be synthetically defined as the discipline responsible for measuring communication and, in enlarged form, as the specialty responsible for quantitatively study the production, distribution, dissemination and consumption of information conveyed in any type of document (book, journal, conference, patent, or website) and any intellectual field, but with special attention to scientific information. It is a discipline with peculiar features:

- It is a very young discipline: although rooted in the early twentieth century in the library environment with the idea of measuring the production of knowledge (bibliographic statistics) and to properly manage library collections, it is not after World War II that Bibliometrics really starts to set its foundations. Its epistemic fundamentals are still boiling (they are not fully settled yet).
- It is a discipline best defined by its methods than by the thematic areas covered (the so-called “metrics”: quantitative data analysis applying various statistical techniques).
- It has a strong interdisciplinary character which arises from the incorporation of methods and techniques developed in other fields, and by its application to the study of any subject area. This makes Bibliometrics an open discipline willing to be fertilized by ideas from the most diverse origin and accept scientists from the most diverse disciplinary environments. This is the reason why Bibliometrics resembles a crossroads, a place where different scientific traditions meet.

The young age of the discipline and its interdisciplinary and instrumental character is the reason why this discipline is known by many different names. However, this fact does not mean the subject of study or the borders of the discipline are not clearly defined. Rather, it is a sign of the coexistence of different traditions that have shaped the development of the discipline.

Bibliometrics is the original and most widespread name. It stems from the bibliographic tradition represented by Paul Otlet with his proposal for a "bibliometrie", a Science for measuring all the dimensions of books and other documents, and from the library tradition concerned since ancient times about measuring the growth of knowledge and usage of its holdings.

Scientometrics is oriented towards the quantitative analysis of scientific and technical literature. It comes from the tradition of the science of science (space of confluence of Sociology, History, and Philosophy of science), to which science policy is also linked. It was crucial for this scientometric orientation the creation of the Citation indexes (databases dedicated to the collection of scientific production).

Informetrics is focused on the discovery of mathematical models that explain the properties of information. It is connected with the modern information science. It is a designation so close to Scientometrics that sometimes it is difficult to find differences among them.

Webometrics and Altmetrics are the most recent denominations. They started to gain momentum as the use of the new information and communication technologies began to spread. They are being developed in the tradition of the modern Library and Information Science, a discipline increasingly dedicated to computer science and to computing itself. These new names are strongly influenced by the medium in which information is conveyed rather than by the content itself. They come also to highlight the traditional technological aspect that the different metric specialties have enjoyed since their inception.

An analysis of the terms used in the titles of documents in our field published between 1969 and 2015 and indexed in Google Scholar (Figure 1) shows a clear predominance of the term “Bibliometrics”, followed by “Scientometrics”. However, in the last three years the term “Altmetrics” is being increasingly used, as a result of the novelty of the new social media communication technologies.

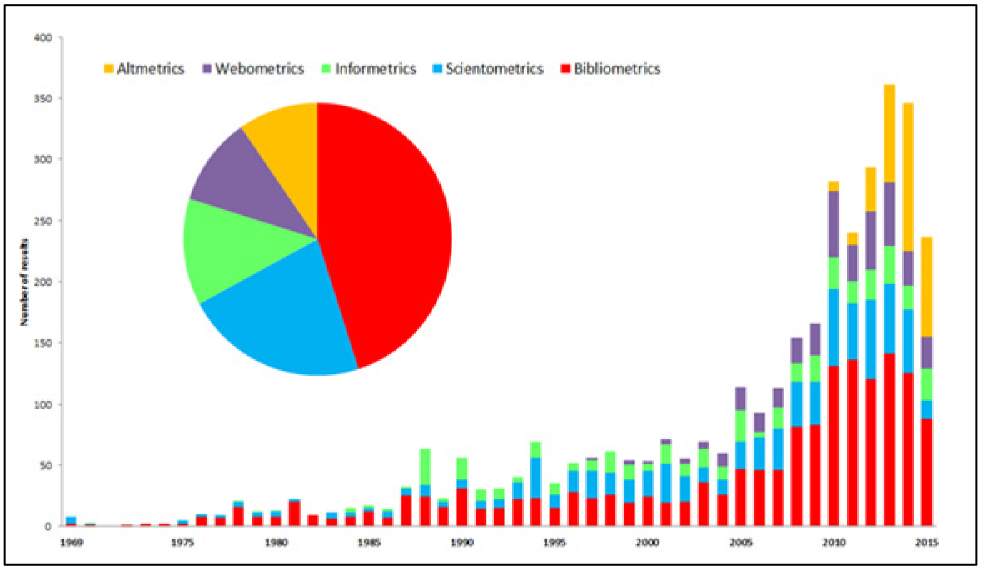


Figure 1. Number of results returned by Google Scholar for the terms Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics contained only within the document titles by year (1969-2015)

A similar result is obtained when the keywords used by the 814 scientists specialized in Bibliometrics or working sporadically in this field with a public profile on Google Scholar Citations are analyzed (Figure 2). The prevalence of Scientometrics and Bibliometrics is clear, although the weight of the latter would be higher had the terms been properly standardized.

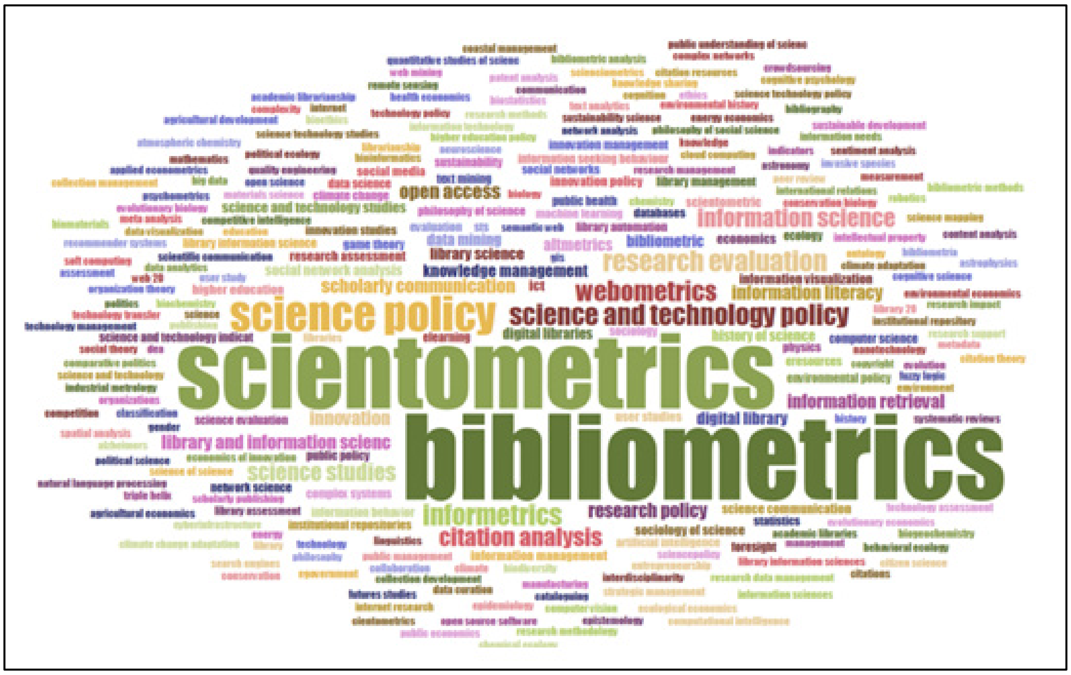


Figure 2. Word cloud of the keywords used by the researchers with a public Google Scholar Citations profile analyzed in this product (size indicates frequency of use in the sample)

Furthermore, it is of great interest to know which other terms are used by bibliometricians. The list of terms associated with the Library and Information Science are very numerous, which shows how this discipline was the area where Bibliometrics stemmed from. Similarly, the relationship with science and technology studies (and specifically with science policy) is obvious. Lastly, there are also many terms related to research evaluation and citation analysis.

1.3. New mirrors and meters of Science: new media and new metrics

There is no better way to learn about a discipline than analysing its scientific literature. The best mirror of a scientific discipline is precisely the intellectual production that its academic community generates. This is the assumption in which Bibliometrics is based when it is used to examine the traits that define other disciplines and specialties.

Knowing the scope of a discipline will not only help characterize and determine its perspective and scientific nature, but it will also indirectly delineate its internal structure, its coherence, its contours, and its location in the overall picture of Sciences. This will enable an understanding of what the research is and has been about in a particular discipline, and how it may evolve in the future.

Today the number of venues in which research results produced in any discipline are published has been remarkably increased. The "Gutenberg paradigm", which limited research products to the printed world (and more specifically to the journal, the main communication channel), has been challenged since the end of the twentieth century by a plethora of new channels of communication that are created, indexed, searched, located, read, and mentioned in the shared hyperspace (Castells, 2002). All this, of course, made possible by the development and worldwide use of the Internet, and the social web in particular. These are the new mirrors where the disciplines and communities are reflected. Revealing and evaluating the role of these new channels in Bibliometrics is another goal of this paper.

Following in the footsteps of the model of scientific communication, which has recently gone through a metamorphosis (from the Gutenberg galaxy to the web galaxy), a change in the model and methods of scientific evaluation is also taking place. The new media, due to its electronic nature, are supplied with multiple indicators measuring all actions and interactions among scientists in the digital space. In this work we open the door to new platform providers of metric indicators (whose nature is still unknown because of its youth) and snoop inside to see what they tell us about the various facets of scientific communication, complementing in this way some recent works in the topic (Jamali, Nicholas & Herman, 2015; Mikki et al, 2015), where not only the potential of these new mirrors but also their limitations and perceptions are considered.

We intend to bring attention to some of these new metrics and look into their meaning. In this way we position ourselves in the debate about "Altmetrics", but using a different perspective: the perspective of individuals and not just the documents they produce. We observe what these new metrics measure by taking as the object of study precisely those researchers who measure others (bibliometricians). In short, Bibliometrics, and those who measure, are measured.

Following our research line oriented on discovering the inner depths of Google Scholar while testing its suitability as a tool for research evaluation, this time we have turned our efforts to investigate new uses for Google Scholar Citations (sometimes also known as Google Scholar Profiles). We present in this new Working Paper a method to learn about the impact of an entire scientific specialty: a very specific scientific and professional community (the Bibliometrics, Scientometrics, Informetrics, Webometrics, and Altmetrics community), and the main agents that are part of it (scientists, professionals, the documents they produce, and the journals and publishers that publish these documents).

From the scientific output of the members of the metrics and quantitative information science studies community who have made public their profile on *Google Scholar Citations (GSC)*, we can develop a picture of this discipline.

Once we've seen the picture of the discipline that can be observed through the data available in GSC, we also want to compare it to its counterparts in other academic web services, like *ResearcherID*, a researcher identification system launched by *Thomson Reuters*, mainly built upon data from *Web of Science* (which has been and still is the go-to source for many researchers in the field of research evaluation), and other profiling services which arose in the wake of the Web 2.0 movement: *ResearchGate*, an academic social network, and *Mendeley*, a social reference manager which also offers profiling features. These are the most widely known tools worldwide for academic profiling. [58, 59]

These tools offer researchers the chance to create an academic profile, as well as the chance to upload their publications, which are therefore available for other researchers to access, download, and comment upon. Researchers can also feed these databases with other kinds of data (tagging and following profiles, asking and answering specific questions) which might be useful for the rest of users in the platform.

In addition, we also include the links to the authors' homepages (the first tool researchers used to showcase their scientific activities on the Web), and *Twitter*, the popular microblogging site, in order to learn how much presence bibliometricians have in this platform and the kind of communication activities in which they take part there.

In short, our aim is to present a multifaceted and integral perspective of the discipline, as well as to provide the opportunity for an easy and intuitive comparison of these products and the reflections of scientific activity each of them portrays.

This project can also be considered as an attempt to deconstruct traditional journal, author, and institutional (mainly university) rankings, which are usually built upon data from traditional citation databases (Web of Science, Scopus) and are based exclusively on journal impact indicators. In this product, we are using a bottom-up approach by analyzing the documents that are either published by a group of authors associated with the discipline, those which are published in the main journals of the discipline, or those which use the most common and significant keywords in the discipline.

This is done in keeping with the widespread notion that the impact of the various scientific units (documents, individuals, organizations, subject domains) should be evaluated directly, using appropriate indicators for each unit, and not by using proxies like, for example, the average impact of the journals where a researcher's or an institution's documents are published to evaluate that researcher or institution.

In short, the objectives of this study are essentially the following:

1. Applying Google Scholar Citations to radiograph Bibliometrics as a discipline, identifying the core authors, documents, journals, and most influential publishers in the field.
2. Comparing the user metric portraits generated by *Google Scholar Citations* to those offered by new platforms for the management of personal bibliographic profiles (*ResearcherID*, *ResearchGate*, and *Mendeley*) and content dissemination and communication (*Twitter*).
3. Testing the completeness, reliability and validity of the information provided by *Google Scholar Citations* (to generate disciplinary rankings), and by the remaining social platforms (to generate complementary academic mirrors of the scientific community).

[58] http://www.nature.com/news/online-collaboration-scientists-and-the-social-network-1.15711

[59] https://101innovations.wordpress.com/2015/06/23/first-1000-responses-most-popular-tools-per-research-activity

2. Methods

2.1. Search and indentification of relevant authors

The first step was to identify all authors who have published in the areas of Bibliometrics, Scientometrics, Informetrics, Webometrics or Altmetrics, and for whom a *Google Scholar Citations* (GSC) public profile could be found at the time the data was collected (24/07/2015).

In order to locate the set of authors relevant to our study (i.e., those who have published in Bibliometrics and have a public profile in GSC), the following search strategies were used:

a) Keywords
A search was conducted in core selected journals: *Scientometrics*, *Journal of Informetrics*, *Research Evaluation*, *Cybermetrics*, and the *ISSI conferences* (*International Conference on Scientometrics and Informetrics*) with the goal of extracting the most frequently used and representative words in the discipline. The selected keywords were:

- Altmetrics
- Bibliometrics
- Citation Analysis
- Citation Count
- H-Index
- Impact Factor
- Informetrics
- Patent Citation
- Quantitative Studies of Science and Technology
- Research Assessment
- Research Evaluation
- Research Policy
- Science and Technology Policy
- Science Evaluation
- Science Policy
- Science Studies
- Scientometrics
- Webometrics

All public GSC profiles containing any of these keywords as one of the research interests were selected (GSC allows authors to display up to five research interests).

The lack of normalization in the use of keywords sometimes forced us to search alternatives of these keywords. These variants included misspelled words, the same keywords in other languages, etc. As an example, these are all the variants we found of the keyword “bibliometrics”: bibliometric; bibliometría; bibliometria; bibliometric analysis; bibliometric methods; bibliometics; bibliometircs; bibliometric analysis in mining sciences; bibliometric mapping; bibliometric studies; bibliometric visualization; bibliometric.; bibliometrics methodology; bibliometrics of social sciences and…; bibliometrics.; bibliometrics...; bibliométrie; bibliometry.

b) Institutional affiliation

All the profiles associated with research centers working on Bibliometrics were also selected. As an example, the profiles with these verified e-mail domains were selected: cwts.leidenuniv.nl, cwts.nl, science-metrix.com, etc.

c) Additional searches

Since there may be some authors working in the discipline who have created a public GSC profile, but who haven't added significant keywords or appropiately filled the institution field in their profile, we also conducted a topic search on *Google Scholar* (using the same keywords as before) as well as a journal search (all the documents indexed in *Google Scholar* published in the following journals: *Scientometrics*, *Journal of Informetrics*, *Research Evaluation*, *Cybermetrics*, and *ISSI proceedings*), with the aim of finding authors we might have missed with the previous two strategies. These searches returned roughly 15,000 documents. Additionally, these searches allowed us to find documents written by authors with no public GSC profile, but which are nonetheless extremely relevant to the discipline.

All these searches were conducted on the 24th of July, 2015.

2.2. Filtering and classification of author profiles

Since *Google Scholar Citations* gives authors complete control over how to set their profile (personal information, institutional affiliation, research interests, as well as their scientific production), a systematic manual revision was carried out in order to:

- Detect false positives: authors whose scientific production doesn't have anything to do with this discipline, even though they labeled themselves with one or more of the keywords associated with it.
- Classify authors in two categories:
    - **Core**: authors whose scientific production substantially falls within the field of Bibliometrics.
    - **Related**: authors who have sporadically published bibliometric studies, or whose field of expertise is closely related to Scientometrics (social, political, and economic studies about science), and therefore they can't be strictly considered bibliometricians.

In order to set the limit between the two categories (core and related authors), we decided to consider as "core authors" those who meet a certain criterion: at least half of the documents which contribute to their h-index should fall within the limits of the field of Bibliometrics.

We considered the titles of the documents, as well as the publishing channel where they appeared, focusing our attention in the journals. Our Bradford-like core of journals about Bibliometrics consisted of six journals (*Scientometrics*, *Journal of Informetrics*, *JASIST*, *Research Evaluation*, *Research Policy*, and *Cybermetrics*), followed by other LIS journals which also publish numerous bibliometric studies (*Journal of Information Science*, *Information Processing & Management*, *Journal of Documentation*, *College Research Libraries*, *Library Trends*, *Online Information Review*, *Revista Española de Documentación Científica*, *Aslib Proceedings*, and *El Profesional de la Información*) and lastly, journals devoted to social and political studies about science (*Social Studies of Science*, *Science and Public Policy*, *Minerva*, *Journal of Health Services Research Policy*, *Technological Forecasting and Social Change*, *Science Technology Human Values*, *Environmental Science Policy*, and *Current Science*).

In the end, we selected a total of 814 GSC profiles. 398 of them have been classified as core authors, and the remaining 416 as related authors.

2.3. Expansion to a multi-faceted approach: units of scientific analysis

Once we defined the set of authors, we automatically extracted the top 100 most cited documents for each author from their GSC profile. To this set of documents, we added the documents we found on our previous topic and journal searches (the third strategy we used to find authors who work on Bibliometrics).

After deleting duplicates, a set of roughly 41,000 documents remained. In the cases where various versions of the same document were found with different number of citations, the one with the highest citation count was selected. This list was sorted according to the number of citations.

For each of the top 1,000 most cited documents in this list, both the basic bibliographic information (especially the sources: journals and book publishers) and the number of citations according to WoS (*Web of Science*) were collected. For those documents that were not indexed in WoS Core Collection (mostly books), the number of citations in WoS was calculated by searching the document in WoS's Cited Reference Search. By doing this we're trying to highlight the (until now mostly neglected) potential of this tool, which truly offers a wealth of citation data that could be used for the evaluation of non-WoS documents.

Lastly, in the cases when a book is a collective work, the number of citations is the sum of the citations to each of the chapters, in addition to the citations directed to the book as a whole.

2.4. Expansion to a multi-faceted approach: social media mirrors

The original 814 authors selected in the previous step (with a public profile created in Google Scholar Citations) were subsequently searched by name in *ResearcherID*, *ResearchGate*, *Mendeley*, and *Twitter*. In the cases where a profile was found in any of these platforms, the main indicators provided by the platform were collected. The data collection from these new academic mirrors was carried out between the 4th and 10th of September, 2015.

Since the maturity of each platform is an important issue to adequately consider its degree of use, the official release date of each platform can be found below:

- *Google Scholar Citations*: a restricted beta release was made on the 20th of July, 2011. It was opened to the general public on the 16th of November, 2011.
- *ResearcherID*: author identification system developed by Thomson Reuters. Released in January 2008.
- *ResearchGate*: academic social network created in May 2008.
- *Mendeley*: social reference manager created in August 2008.
- *Twitter*: online social networking service that enables users to send and read short 140-character messages. Released on the 15th of July, 2006.

The URLs to personal homepages were searched and collected as well. In this case, this information was retrieved from the field "homepage" included in the *Google Scholar Citations* profiles of the authors considered. Since there is not any restriction about the kind of URL an author may use in this field, some authors choose to save the URL of their profile in other platforms (such as ResearchGate), or the URL of the research group, institution, or company they work for, among other cases. In this case, this information was filtered and only personal or institutional websites managed directly by the authors are analyzed.

2.5. Author-level metrics: list and scope

All the metrics collected from each of the social media platforms analyzed, as well as their definition and scope can be found in Table 1.

Table 1. List and explanation of author-level indicators

Google Scholar Citations	
INDICATOR	DEFINITION
Citations	Number of citations to all publications. Computed for citations from all years, and citations since 2010
h-index	The largest number h such that h publications have at least h citations. Computed for citations from all years, and citations since 2010
i10 index	Number of publications with at least 10 citations. Computed for citations from all years, and citations since 2010
INDICATOR	DEFINITION
Total Articles in Publication List	The number of items in the publication list
Articles with Citation Data	Only articles added from *Web of Science Core Collection* can be used to generate citation metrics. The publication list may contain articles from other sources. This value indicates how many articles from the publication list were used to generate the metrics
Sum of the Times Cited	The total number of citations to any of the items in the publication list from *Web of Science Core Collection*. The number of citing articles may be smaller than the sum of the times cited because an article may cite more than one item in the set of search results
Average Citations per Item	The average number of citing articles for all items in the publication list from *Web of Science Core Collection*. It is the sum of the times cited divided by the number of articles used to generate the metrics
h-index	h is the number of articles greater than h that have at least h citations. For example, an h-index of 20 means that there are 20 items that have 20 citations or more
INDICATOR	DEFINITION
RG Score	It's a metric that measures scientific reputation based on how an author's research is received by his/her peers. The exact method to calculate this metric has not been made public, but it takes into account how many times the contributions (papers, data, etc.) an author uploads to *ResearchGate* are visited and downloaded, and also by whom (reputation)
Publications	Total number of publications an author has added to his/her profile in *ResearchGate* (full-text or no)
Views	Total number of times an author's contributions to *ResearchGate* have been visualized. This indicator has recently been combined with the "Downloads" indicator to form the new "Reads" indicator, but the data collection for this product was made before this change came into effect
Downloads	Total number of times an author's contributions to *ResearchGate* have been downloaded. This indicator has recently been combined with the "Views" indicator to form the new "Reads" indicator, but the data collection for this product was made before this change came into effect
Citations	Total number of citations to the documents uploaded to the profile. *ResearchGate* generates its own citation database, and they warn this number might not be exhaustive

Impact Points	Sum of the JCR impact factors of the journals where the author has published articles
Profile views	Number of times the author’s profile has been visited
Following	Number of *ResearchGate* users the author follows (the author will receive notifications when those users upload new material to *ResearchGate*)
Followers	Number of *ResearchGate* users who follow the author (those *ResearchGate* will receive notifications when the author uploads new materials to *ResearchGate*)

INDICATOR	DEFINITION
Readers	This number represents the total number of times a *Mendeley* user has added a document by this author to his/her personal library
Publications	Number of publications the author has uploaded to *Mendeley* and classified as “My Publications”
Followers	Number of *Mendeley* users who follow the author
Following	Number of *Mendeley* users the author follows

INDICATOR	DEFINITION
Tweets	Total number of tweets an author has published according to his profile
Followers	Number of *Twitter* users who follow the tweets published by the author
Following	Number of *Twitter* users the author follows
Days registered	Number of days since the author created an account on *Twitter*
Sum Retweets	Number of Retweets obtained for the author.
H Retweets	An author has a h-Retweet of “n” when “n” of its tweets has achieved at least “n” Retweets.

2.6. Limitations

Projects of a bibliographic nature like this one can’t ever reach perfection, and it is entirely possible that we may have missed relevant authors. The criteria for selecting the authors were two: first, the existence of a public GSC profile about the author by 24/07/2015 (when the data collection was made), and second, that the author works on the fields of Bibliometrics, Scientometrics, Informetrics, Webometrics, or Altmetrics.

We’re completely aware that these lists don’t include all the researchers in the area, since some haven’t created a profile, or they haven’t made it public. We should note that we made an exception with Eugene Garfield, one of the fathers of Bibliometrics. Despite the fact that he doesn’t have a public GSC profile, we manually searched his production on *Google Scholar* and computed the same indicators GSC displays. We believe this Working Paper would be incomplete without him.

We strongly encourage researchers without a GSC profile, and especially those who have made important contributions to the development of this field, to bring together the scattered bibliographic information *Google Scholar* has already compiled about their works. Sharing this information would not only greatly benefit their online visibility; it would also be very useful to the rest of the scientific community.

2.7. Statistical analysis

Spearman correlation (α= 0.05) was applied to all 31 metrics considered in each of the platforms (excluding personal webpages), and finally a Principal Component Analysis (Spearman similarity with varimax rotation of axes and uniform weighting) was applied in order to reveal the relationships among metrics and platforms as well as the possible existence of metric clusters.

3. Results

3.1. The actors of Bibliometrics as a discipline, according to Google Scholar Citations: authors, documents, journals and publishers

a) Authors

By analyzing the list of most influential authors of the discipline (Table 2) we noticed that the most prominent positions (top ten) include the founders of the discipline (Price and Garfield) and the most influential bibliometricians, almost all of them holders of the Price medal (all except Chen), a prize that recognizes scientists who have contributed with their work to the development of Bibliometrics.

Table 2. Top 25 influential core authors in Bibliometrics according to Google Scholar Citations

AUTHOR	GS CITATIONS	H-INDEX
Loet Leydesdorff	**26,484**	73
Eugene Garfield	**22,622**	55
Mike Thelwall	**13,840**	61
Derek J. de Solla Price	**13,263**	33
Francis Narin	**11,297**	45
Wolfgang Glänzel	**10,796**	54
Ronald Rousseau	**9,570**	42
Chaomei Chen	**9,512**	43
Anthony (Ton) F.J. van Raan	**9,200**	53
Ben R Martin	**8,975**	39
András Schubert	**8,655**	45
Peter Ingwersen	**8,356**	35
Henk F. Moed	**8,256**	46
Blaise Cronin	**7,347**	43
Henry Small	**7,307**	32
Tibor Braun	**7,231**	41
Vasily V. Nalimov	**6,343**	31
Lutz Bornmann	**6,108**	40
Belver C. Griffith	**5,695**	26
Howard D. White	**5,569**	30
Johan Bollen	**5,394**	33
Katy Borner	**5,326**	31
Félix de Moya Anegón	**5,074**	35
Koenraad Debackere	**4,933**	32
Jose Maria López Piñero	**4,823**	31

Bibliometrics received a decisive boost from the personality and the work of both Price and Garfield, who can be considered the fathers of this discipline. On the one hand, Price, armed with the theoretical foundations laid by John Desmond Bernal and Robert K. Merton, set out to systematically apply quantitative techniques to the History and social studies of Science, developing the theoretical foundations of Scientometrics, born from the combination of the Sociology of science, History, Philosophy of science, and Information science. This approach is characterized by the analysis of the life and activity of Science and scientists from a quantitative perspective. The numbers were used to characterize the production of knowledge and scientists' lives: what they create and produce, with whom they relate to, the sources they used, and the impact and influence they provide/receive to/from other scientists, etc.

On the other hand, Garfield made possible that Bibliometrics became a reality (Mccain 2010; Bensman, 2007): the creation of the "citation index" made possible the quantification of scientific activity through its main output: the publications and citations they generate. Since then, citation analysis and all its variants have become the most widespread analysis technique of this new specialty (this is evidenced by the significant presence of highly cited documents that deal with this topic). Garfield defined the phenotype of the discipline: technology (the basis for the storage and circulation of information) is at the heart of all its tools. That is, Bibliometrics will evolve at the same rate the technologies of information and communication do.

The map of Bibliometrics can also be discerned by analyzing the rest of the authors in the list: the Hungarian school (both Eastern Europe and Russia, like Nalimov), the Dutch school (with its various branches in Leiden and Amsterdam), the Belgian school (with Egghe and Rousseau), the North American School (Small, Griffith, and White), the Spanish school (with López Piñero, Spanish translator of Price's work, and the one who introduced Bibliometrics in Spain), and the new authors that represent the technological transformation of the discipline (mainly Thelwall).

b) Documents

An analysis of the list of the 25 most cited documents according to Google Scholar (Table 3) reveals several issues:

- The importance of the documents that first introduce new techniques and citation-based indicators, like the ones by Hirsch (3rd), Garfield (9th and 10th), Small (12th), Egghe (23rd), and Griffith and White (37th). Among them we find the most widely known indicator in Bibliometrics (the impact factor) and the one that has come to replace it while extending its capabilities (h-index).
- The excellence both in the work in which Hirsch proposes the h-index and in the articles about the impact factor highlights the strong orientation of Bibliometrics towards evaluation in general and the assessment of the performance of individuals, journals, and institutions... This reveals a clear link between Bibliometrics and Science policy, and explains the use of bibliometric indicators and other bibliometric tools by policymakers.
- As we would expect, among the most cited documents we find texts that have served as textbooks for the discipline (written by Moed, Van Raan, Eghhe, Rousseau, etc.).
- The anomalous institutionalization process of the discipline. The main "bibliometric laws" which still hold true today where established at the dawn of the discipline, even before it was fully instituted (Lotka, Zipf, Bradford), and were developed by authors working outside the discipline. The same happened with the proposal of the h-index by Hirsch, elaborated by this physician in his "leisure time". Bibliometrics is often revolutionized from outside Bibliometrics.
- The great relevance of some topics such as the "Triple Helix" by Leydersdorff, or the social networks by Barabási, which make a big impact outside the borders of our discipline (Management and Economy in the first case, and sociometrics and computer science in the second).

Table 3. Top 25 most influential documents in Bibliometrics according to Google Scholar Citations

TITLE	AUTHORS	SOURCE	YEAR	GS CITATIONS
Little science, big science	de Solla Price	Columbia University Press	1963	**5,410**
An index to quantify an individual's scientific research output	Hirsch	PNAS	2005	**4,860**
The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university-industry-government relations	Etzkowitz & Leydesdorff	Research Policy	2000	**4,414**
Universities and the global knowledge economy: a triple helix of university-industry-government relations	Etzkowitz & Leydesdorff	Pinter Press	1997	**2,585**
Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems	Moed; Glänzel & Schmoch (ed.)	Springer	2005	**2,261**
Citation analysis as a tool in journal evaluation. Journals can be ranked by frequency and impact of citations for science policy studies	Garfield	Science	1972	**2,166**
Citation indexing: Its theory and application in science, technology, and humanities	Garfield	Wiley	1979	**2,130**
The frequency distribution of scientific productivity	Lotka	Journal of Washington Academy Sciences	1926	**2,090**
Co-citation in the scientific literature: A new measure of the relationship between two documents	Small	JASIS	1973	**1,988**
Links and impacts: The influence of public research on industrial R&D	Cohen; Nelson & Walsh	Management Science	2002	**1,881**
Evolution of the social network of scientific collaborations	Barabasi; Jeong; Neda; Ravasz; Schubert & Vicsek	Physica A	2002	**1,851**
Citation indexes for science. A new dimension in documentation through association of ideas	Garfield	Science	1955	**1,783**
What is research collaboration?	Katz & Martin	Research Policy	1997	**1,591**
Handbook of quantitative studies of science and technology	Van Raan (ed.)	North-Holland	1988	**1,510**
The history and meaning of the journal impact factor	Garfield	JAMA	2006	**1,487**
The increasing linkage between US technology and public science	Narin; Hamilton & Olivastro	Research Policy	1997	**1,211**
A general theory of bibliometric and other cumulative advantage processes	de Solla Price	JASIST	1976	**1,148**
Statistical bibliography or bibliometrics?	Pritchard	Journal of Documentation	1969	**1,134**
Theory and practise of the g-index	Egghe	Scientometrics	2006	**1,113**
The Web of knowledge: a Festschrift in honor of Eugene Garfield	Garfield; Cronin & Atkins (ed).	Information Today	2000	**1,102**
Visualizing a discipline: An author co-citation analysis of information science, 1972-1995	White & McCain	JASIS	1998	**1,100**
CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature	Chen	JASIST	2006	**1,083**
Citation analysis in research evaluation	Moed	Springer	2005	**1,060**
Citation frequency and the value of patented inventions	Harhoff; Narin; Scherer & Vopel	Review of Economics and Statistics	1999	**1,023**
Maps of random walks on complex networks reveal community structure	Rosvall & Bergstrom	PNAS	2008	**992**

If we pay attention to the distribution of documents according to their typology (Figure 3), the journal article stands out overwhelmingly (89% of all 1,069 documents processed), showing that formal papers published in peer reviewed journals stand as the main scientific vehicles in this social science discipline.

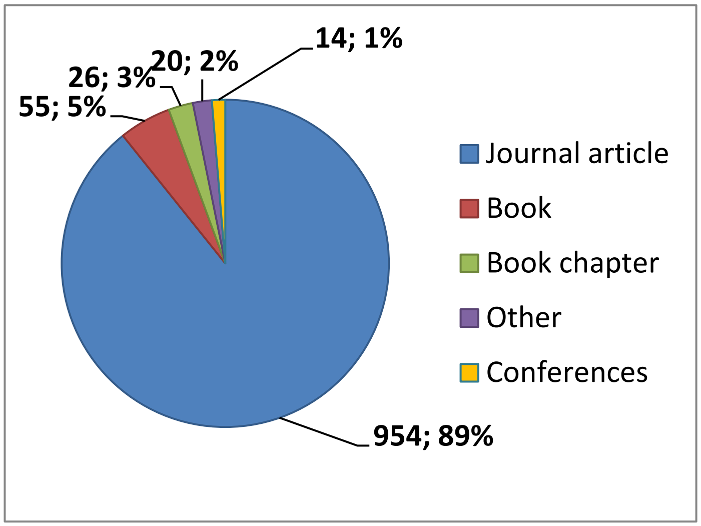


Figure 3. Distribution of highly cited documents in Bibliometrics according to Google Scholar citations (n= 1,069)

The presence of books (5%) is smaller, but this figure may be misleading. An analysis of the top highly cited documents according to Google Scholar citations shows that within the top 25 documents, 8 of them are books (14 within the top 50). Obviously, the number of published books is lower than the number of articles. The presence of documents from the remaining categories is lower: Book chapters (3%), and other material (including dissertation theses, reports, etc.; 2%). Lastly, the results obtained for conference proceedings (1%) reveal a low impact of this scientific communication channel.

c) Journals

The third unit analyzed is the journals in which highly cited documents have been published (i.e., considering only the top 1,000 most cited documents). In Table 4 we provide the top 25 journals according to the number of highly cited documents published. Additionally, we show the total number of citations received by these articles, the percentage of citations per article (C/A), the percentage of highly cited documents in the sample (HCA) and the distribution of citations.

Scientometrics is the journal with more articles published within the 1,000 most cited documents (284 articles). It is thus the most influential journal in the discipline. Its birth in 1978 was a milestone in the process of institutionalization of the discipline.

The second place is occupied by *JASIST* (137 articles). This fact shows the important role of this journal in Bibliometrics, although its scope is broader. This journal has maintained since its inception a strong link between Information Science and Bibliometrics, though some authors have noticed a slight specialization towards Bibliometrics over time (Nicolaisen & Frandsen, 2015).

Journal of informetrics, focused exclusively on Bibliometrics, Scientometrics, Webometrics, and Altmetrics, appears in the fourth position (36 articles). The young age of this journal (it was created in 2007) explains why there isn't a greater number of articles published in this journal among the most cited documents in the discipline.

The connection between Library and Information Science and Bibliometrics is noticeable through the presence of other important LIS journals in the list, such as *Journal of Documentation*, *Journal of Information Science*, *Library Trends*, or *Aslib Proceedings*. This connection has been a matter of public record for a long time now (White & McCain 1998; Larivière, Sugimoto, Cronin 2012, Larivière 2012).

Its connections with the field of web technologies from an information science perspective is strongly marked as well (*Cybermetrics*, *Online Information Review*). Additionally, we can see that journals oriented towards the Social Studies of Science (such as *Research Policy*, *Social Studies of Science*, and *Science and Public Policy*) also have strong ties to Bibliometrics.

Lastly, the role of multidisciplinary journals (such as *Nature*, *Science*, *PNAS* or *PLoS One*) should not be forgotten. If we analyze the number of citations instead of the number of articles published, we find the same first three journals occupying the first positions (*Scientometrics*, *JASIST*, and *Research Policy*), but the data also shows a great impact of articles published outside the core journals of the discipline. *Science* gets 9,219 citations from only 8 articles whereas *PNAS* achieves 7,642 citations from 9 articles, and *PLoS One* gets 2,376 citations from 13 articles (the figures for *Nature* are lower, with "only" 1,871 citations from 10 articles).

Table 4. Top 25 most influential journals in Bibliometrics according to Google Scholar Citations

JOURNAL	ARTICLES	CITATIONS	C/A	HCA(%)	CITATIONS(%)
Scientometrics	284	44,384	156	29.8	22.5
JASIST	137	27,021	197	14.4	13.7
Research Policy	57	18,866	330	6.0	9.6
Journal of Informetrics	36	5,052	140	3.8	2.6
Journal of Documentation	25	5,538	221	2.6	2.8
Information Processing & Management	24	4,404	183	2.5	2.2
Journal of Information Science	20	3,815	190	2.1	1.9
Research Evaluation	18	2,126	118	1.9	1.1
ARIST	14	3,621	258	1.5	1.8
Social Studies of Science	13	3,204	246	1.4	1.6
Science and Public Policy	13	2,875	221	1.4	1.5
Plos One	13	2,376	182	1.4	1.2
Nature	10	1,871	187	1.0	1.0
Current Contents	10	1,696	169	1.0	0.9
PNAS	9	7,642	849	0.9	3.9
Science	8	9,219	1,152	0.8	4.7
Library Trends	7	1,230	175	0.7	0.6
Medicina Clinica	6	958	159	0.6	0.5
Online Information Review	6	806	134	0.6	0.4
Science Technology & Human Values	5	946	189	0.5	0.5
Aslib Proceedings	5	765	153	0.5	0.4
Cybermetrics	5	627	125	0.5	0.3
American Psychologist	4	1,026	256	0,4	0,5
World Patent Information	4	726	181	0.4	0.4
Ethics in Science and Environmental Politics	4	687	171	0.4	0.3

d) Book publishers

The last unit of analysis is the book publishers. Table 5 shows the top 20 publishers according to the percentage of highly cited documents (top 1,000). Additionally, the number of documents, citations (total and percentage of citations respect to the total) and citations per document are offered.

The first position is occupied by Springer, with 10 documents positioned within the set of highly cited books, and receiving 5,766 citations (14.3% of all citations to book publishers). Information Today (10.9%) and Wiley (9.1%) stand on the second and third position respectively.

Table 5. Top 25 most influential publishers in Bibliometrics according to Google Scholar Citations

PUBLISHER	HC	HC(%)	CITATIONS	CITATIONS(%)	C/A
Springer	10	18,2	5,766	14,3	576.60
Information Today	6	10,9	1,635	4,0	272.50
Wiley	5	9,1	3,121	7,7	624.20
Lexington	4	7,3	1,627	4,0	406.75
Sage	4	7,3	1,324	3,3	331.00
UFMG	4	7,3	845	2,1	211.25
University of Chicago Press	3	5,5	6,874	17,0	2,291.33
Russell Sage Foundation	3	5,5	3,836	9,5	1,278.67
North-Holland	3	5,5	2,130	5,3	710.00
Blackwell	2	3,6	1,132	2,8	566.00
Elsevier	2	3,6	1,071	2,7	535.50
Taylor Graham	2	3,6	688	1,7	344.00
Scarecrow Press	2	3,6	416	1,0	208.00
ISSI	2	3,6	276	0,7	138.00
Ablex	2	3,6	193	0,5	96.50
FECYT	2	3,6	193	0,5	96.50
Columbia University Press	1	1,8	5,410	13,4	5,410.00
Pinter Press	1	1,8	2,585	6,4	2,585.00
Yale University Press	1	1,8	936	2,3	936.00
MIT Press	1	1,8	710	1,8	710.00

We can observe that all publishers achieve high numbers of citations per document. In this case, we should highlight the performance of university presses (such as University of Chicago, Columbia, Yale, or MIT), with a very low presence in terms of productivity but an impressive impact in the number of citations. The ability to attract well-established authors in order to edit specialized books makes a great difference in book publisher rankings.

3.2. Online presence of the bibliometric community

Scientists traditionally communicated with their communities both through informal means (letters, meetings, seminars, conferences ...) and formal means (books, journal articles, patents, patents, etc.), and in both of them the scope of these communications was limited by the printed technology in which the contents were transmitted. Today, since the birth of the Web, which brought the chance to create personal pages, and with the emergence of academic social networks, researchers can display their work through a rich variety of channels and electronic formats.

Studies of the level of web presence and impact of scientists' through their personal websites have already been carried out. Barjak, Li & Thelwall (2007) analyzed data from 456 scientists from five scientific disciplines in six European countries, whereas Mas-Bleda, Aguillo, (2013) and Más-Bleda et al (2014) put their focus on 1,498 highly cited researchers working at European institutions, distributed in 22 different countries, using data extracted from the ISIHighlyCited.com database.

In the field of Bibliometrics, the pioneer work by Haustein et al (2014) should also be highlighted. In this study, 1,136 documents authored by the 57 presenters of the 2010 STI conference in Leiden (57 researchers, who together had authored 1,136 papers) were collected using WoS and Scopus. After this, the scholarly and professional social media presence of these authors in several platforms was measured (Google Scholar Citations, LinkedIn, Twitter, Academia.edu, ResearchGate and ORCID).

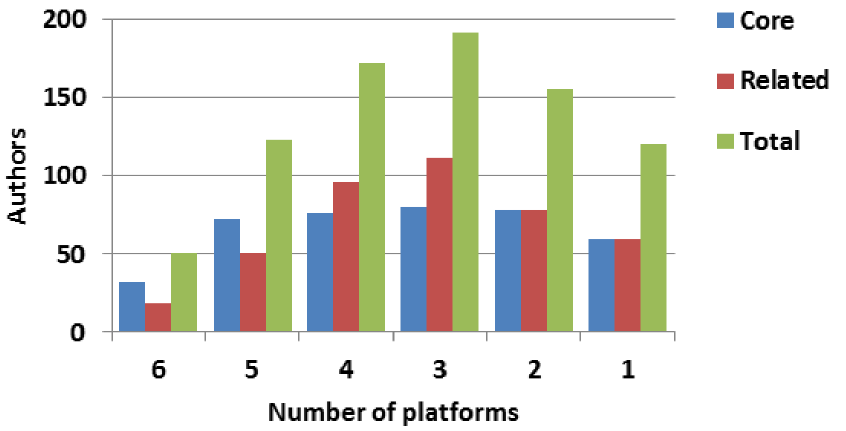
In this work we intend to expand this sample by considering the social presence of the whole bibliometric community as well as other researchers who are related to the discipline in some way. A total of 814 researchers (398 bibliometricians and 416 researchers who have sporadically published bibliometric studies) have been analyzed.

In Table 6 we find the distribution of authors according to the number of platforms in which they have created a personal profile, regardless of their impact or the degree to which these profiles are updated. We highlight the following points:

- The degree of social presence is high. All 814 authors have at least a personal profile created in one platform; 14.7% of the authors are visible in only one platform.
- Authors with two (19.1%), three (23.5%), or four (21.1%) profiles are the more numerous groups.
- No significant differences between core and related authors are found.
- There is a small group of authors (6.2%) with high media visibility (presence in all social media analyzed), being among them some of the most influential bibliometricians (such as Loet Leydesdorff, Mike Thelwall, Chaomei Chen, Lutz Bornmann, Félix de Moya Anegón, Katy Borner, Judit Bar-Ilan, Nees Jan van Eck, or Isidro F. Aguillo, among others).

Table 6. Social presence of the bibliometric community

NUMBER OF PLATFORMS	AUTHORS		
	CORE	RELATED	TOTAL
6	32	19	51
5	72	51	123
4	76	96	172
3	80	112	192
2	78	78	156
1	60	60	120
TOTAL	398	416	814



The use of each specific social platform is shown in Table 7. The main results derived from these data are the following:

- *ResearchGate* is (after *Google Scholar Citations*) the second most used platform by these authors (66.7%), followed at some distance by *Mendeley* (41.28%) and homepages (41.15%).
- The number of *Mendeley* profiles is high, although this data by itself is misleading, since 17.1% of the profiles (68 out of 397) are basically empty. *ResearcherID* is also affected by this issue (34.45% of the profiles are empty); as is *Twitter* (47% of the 240 authors with a *Twitter* profile have published less than 100 *tweets*).
- *ResearcherID* presents a wider acceptance among core authors (45.7%) than related authors (35.1%)
- *Twitter* is the least used platform, since only 33.17% of core authors (and 25.96% of related authors) have created a *Twitter* profile.
- Personal homepages are widely used by authors, although this denomination covers a wide range of different website typologies (personal websites outside institutions, institutional websites not managed by authors). The use of social platforms as personal sites is common (22 authors considered their profiles in other academic social sites such as ResearchGate, Academia.edu, Mendeley, and ImpactStory as their personal websites).
- Core and related authors present similar behavior as regards their presence on these social platforms, although there is a slightly higher rate of core authors on *Twitter*, *ResearcherID,* and *Mendeley* than there is of related authors.

Table 7. Degree of use of social platforms by type of author

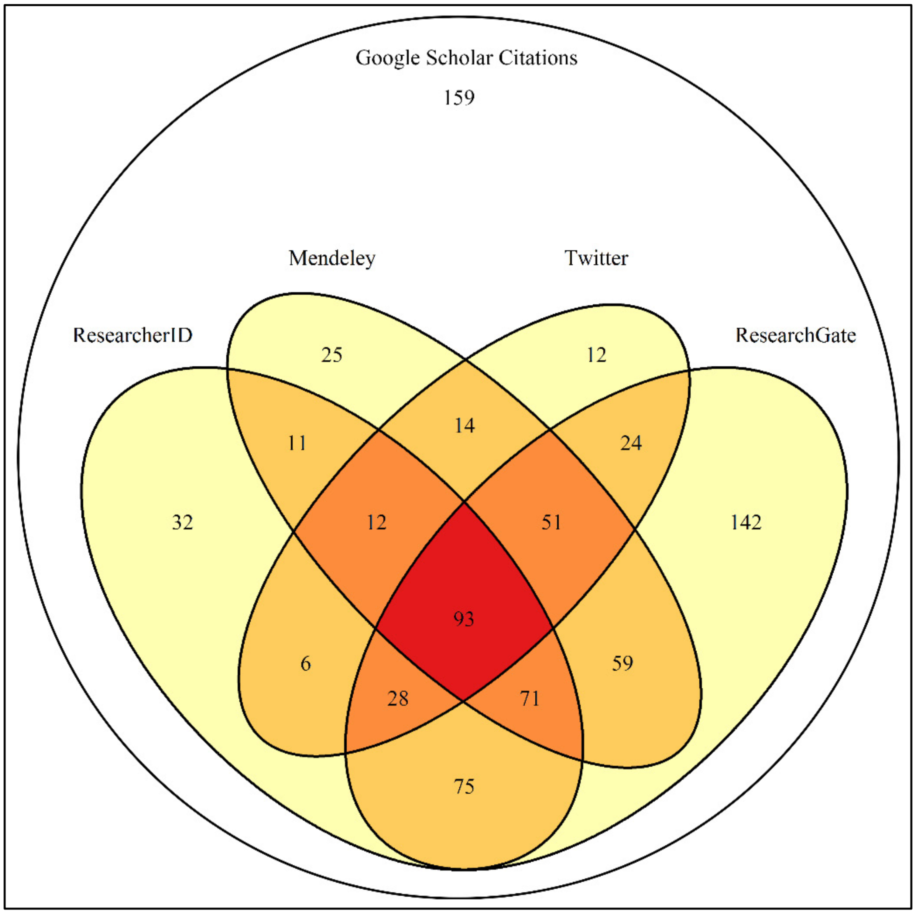
WEB PLATFORMS	AUTHORS					
	CORE	%	RELATED	%	TOTAL	%
* Google Scholar Citations	398	100	416	100	814	100
ResearcherGate	260	65.33	283	68.03	543	66.71
Mendeley	171	42.96	165	39.66	336	41.28
** Homepage	158	39.69	177	42.54	335	41.15
ResearcherID	182	45.73	146	35.10	328	40.29
Twitter	132	33.17	108	25.96	240	29.48

* All authors in the sample have a profile in GSC. ** *ResearchGate* and *Academia.edu* URLs were discarded.

Figure 4 shows the combination of profiles used by the authors (core and related) of the bibliometric community. It should be reminded that all authors in our sample have Google Scholar Citation profiles (this was the main selection criteria).

Personal webpages have been omitted from this analysis since they represent another dimension of web presence, different from those offered by social platforms and academic profiles.

Figure 4. Combination of profiles used by the bibliometricians in our sample



As we can see in Figure 4, there is great number of researchers who only have a profile in *Google Scholar Citations* (159). There are also many authors who only have a profile in GSC and *ResearchGate* (142). The number of researchers who have an account in all the platforms analyzed in this study (GSC, *ResearcherID*, *Mendeley*, *Twitter*, and *ResearchGate*) is 93 (11.4% of our sample).

The remaining combinations seem to be more unusual. For example, there are only 12 authors who use only GSC and *Twitter*, and 14 authors who use only GSC, *Mendeley*, and *Twitter*. In a similar manner, there are only 11 authors who use only GSC, *ResearcherID*, and *Mendeley*.

These results are similar to the ones offered by Van Noorden (2014) about the presence of scientists in social networks and the provisional results by Bosman and Kramer (2015).

Despite the fact that this sample suffers from a bias in favor of *Google Scholar Citations* because of how the data were collected, there is no doubt that GSC is the platform authors currently prefer to display their publications, followed at a distance by *ResearchGate*, but a distance that is increasingly shorter. 66.7% of the authors with a GSC profile have also a *ResearchGate* profile. This is significant enough, although these results must be tempered by the degree of use and update frequency of each platform, aspects which will be discussed later in greater detail.

These results should be especially contextualized within the bibliometric community, which undoubtedly has a certain bias towards using these platforms, because these platforms are sometimes objects of study themselves. Differences in the degree of presence on social platforms in different fields of knowledge should be expected, as González-Díaz, Iglesias-García, and Codina (2015) have recently proved in their analysis of the discipline of Communication.

3.3. Comparing social platform metrics: from citations to followers

After analyzing the academic output and impact for the bibliometric community using *Google Scholar Citations*, and describing the preferences of the members of this scientific community for social interaction in the Web, in this section we are going to analyze the correlation between these metrics. Firstly, all metrics associated with *Google Scholar Citation* profiles, and secondly, all metrics associated and offered by each of the social platforms analyzed (*Mendeley*, *ResearcherID*, *ResearchGate*, and *Twitter*). Personal webpages have been excluded from this analysis.

By way of illustration, in Table 8 we show the median of the main metrics evaluated so that we can compare the performance or size of similar indicators in each web platform. In this sense, we highlight the following issues:

- Regarding the “Total citations received”, the higher median value corresponds to *Google Scholar* (156), followed by *ResearchGate* (85) and *ResearcherID* (63).
- As to the h-index, Google Scholar obtains a score of 6; whereas in *ResearcherID* this value is lower (4).
- Regarding academic output, *ResearchGate* achieves the first position (27), followed at a distance by *ResearcherID* (15) and *Mendeley* (9). The number of records stored in each *Google Scholar Citation* profile is not available in this work.
- Regarding the social interaction features (“following / followed by”), users in both *ResearchGate* and *Mendeley* show a slightly passive behavior: users tend to be followed by many people, but they do not follow many other users. Interestingly, the opposite behavior is found in *Twitter*, where scholars tend to follow many users, but it seems harder to be followed by others. Since *ResearchGate* and *Mendeley* deal exclusively with academic audiences, a logical explanation may be that respected scholars who create an account are widely followed, but they do not tend to follow other users. Nonetheless, in the open space defined by *Twitter*, the situation is just the opposite: gaining followers implies an active participation in the platform.

Table 8. Median of principal metrics

SOURCE	METRIC	MEDIAN
Google Scholar (n=811)	Citations_total	156
	Citations_last5	117
	H-index_total	6
	H-index_last5	5
	i10_total	4
	i10_last5	3
ResearcherID (n=275)	Total_articles	15
	Articles_cited	11
	Times_cited	63
	Average_citations	5.75
	H-index	4
ResearchGate (n=515)	RG Score	13.82
	Publications	27
	Impact_points	12.97
	Followers	38
	Following	23
	Downloads	802
	Views	1845
	Citations	85
	Profile_views	696
Mendeley (n= 185)	Publications	9
	Readers	93
	Followers	3
	Following	2
Twitter (n=226)	Tweets	153.5
	Followers	99
	Following	130

In Table 9 we show all correlations achieved among each of the 31 metrics considered in this study (α= 0.05), whereas in Figure 5 we show the results of a Principal Component Analysis (PCA).

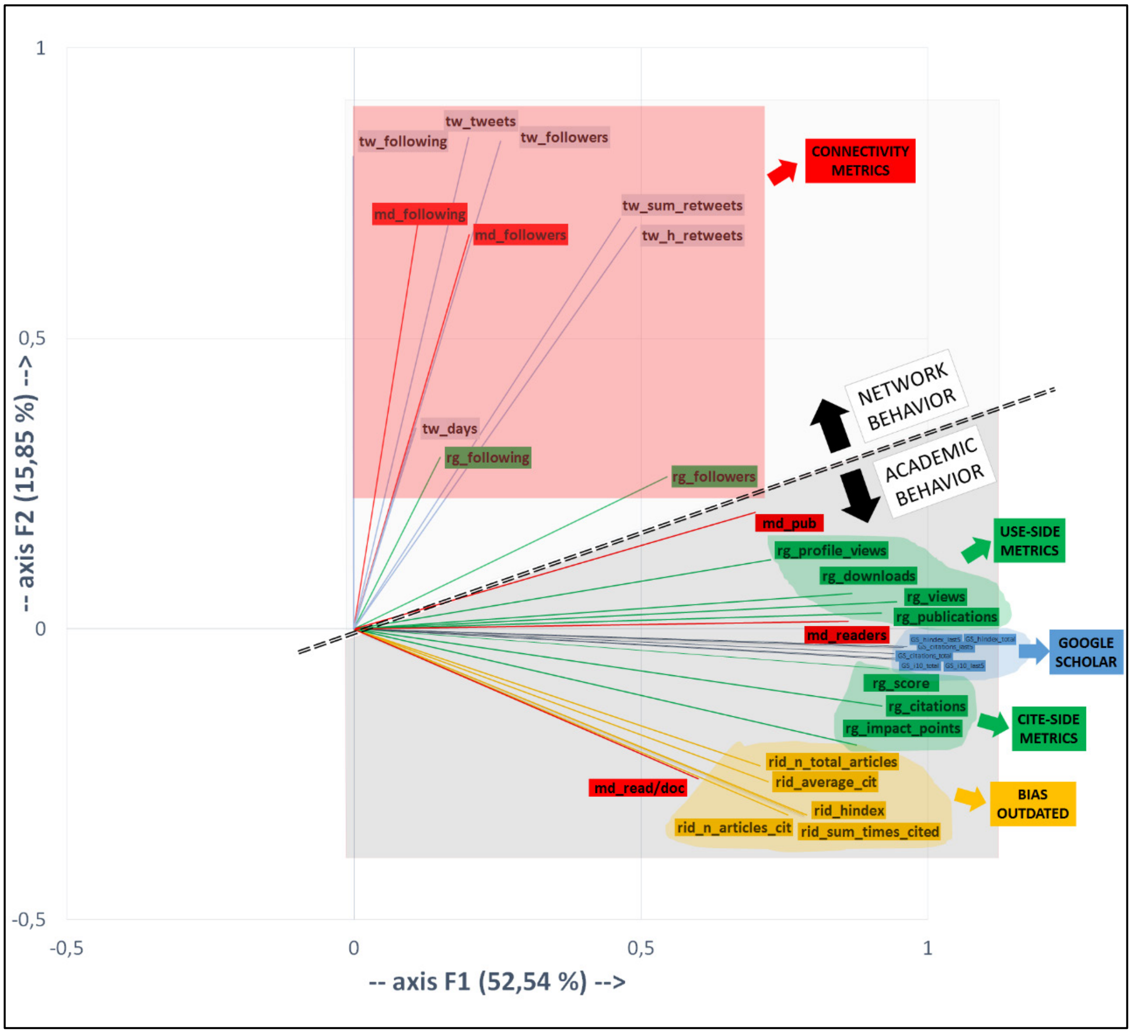


Figure 5. Principal Component Analysis for 31 metrics associated with bibliometricians' social platform profiles

The main results are:

- We find two clear dimensions: at the top we can see all metrics related to connectivity and popularity (followers), and at the bottom, all metrics related to academic performance. This second group can further be divided into usage metrics (views and downloads) and citation metrics. *ResearchGate* provides examples for these two faces of academic performance, since *Google Scholar Citations* profiles do not offer data about downloads or reads.
- All metrics provided by *Google Scholar* (both citations and h-index) correlate strongly among themselves.
- We find a clear separation between the usage (views and downloads) and citation metrics (Citations, Impact Points) provided by *ResearchGate*. The RG Score for example displays a high correlation to metrics from *Google Scholar Citations*: i.e. total citations (r= 0.89) and the h-index (r= 0.92).
- The number of readers in *Mendeley* is connected to the usage metrics offered by *ResearchGate*, and strongly correlates to *Google's* total citations (r= 0.77), Google's h-index (r= 0.82), and the RG Score (r= 0.75). The number of documents in *Mendeley* is far from the *Mendeley* readers in this PCA, probably because *Mendeley* profiles aren't updated as regularly as GSC profiles. Of course, this also affects the combined metric "readers per document".
- Indicators from *ResearcherID* strongly correlate among themselves, but are slightly separated from other citation metrics (those from *Google Scholar* and *ResearchGate*). This issue can probably be

explained by the low regularity with which *ResearcherID* profiles are updated. In view of the results, this isolation may be used as a mechanism to check the "currentness" (or lack thereof) of a profile in *ResearcherID*.

- All metrics associated with the number of followers (all *Twitter* metrics and their counterparts in *ResearchGate* and *Mendeley*) correlate among themselves, and are separated from the citation metrics. Curiously enough, the number of followers offered by *ResearchGate* is, within the group of connectivity metrics, the one which is closest to the usage metrics, serving in fact as a bridge between the two groups. This may mean that networking metrics from academic social networks correlate better with usage metrics than networking metrics from *Twitter*. *Mendeley*'s networking metrics, however, are placed closer to *Twitter*'s metrics.
- The impact of *Tweets* (measured by Retweets) is closer to the academic side. In any case, their correlation with impact measures is statistically significant (α=0.05). The correlation of Sum Retweets and H-Retweets with *Google Scholar* total citations is 0.44 and 0.45 respectively.
- The number of days that a *Twitter* account has been active does not seem to correlate with any other *Twitter* metric. Unlike in online marketing, time is not a critic factor to achieve followers. Academic prestige and activity (number of *Tweets* tweeted) may be the most important parameters to achieve a great number of Twitter followers.

Table 9. Correlation analysis (Spearman) for 31 metrics associated with bibliometricians' social platform profiles

PLATFORM		GOOGLE SCHOLAR						RESEARCHERID					RESEARCHGATE									MENDELEY					TWITTER					
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
GOOGLE SCHOLAR	1	1.0	.99	.97	.96	.97	.97	.57	.61	.67	.62	.66	.89	.86	.86	.05	.43	.78	.87	.95	.60	.57	.77	.57	.11	.02	.17	.21	-.06	.08	.45	.46
	2	.99	1.0	.97	.97	.96	.97	.56	.62	.67	.63	.67	.91	.88	.88	.06	.46	.81	.90	.94	.63	.58	.79	.61	.13	.05	.18	.21	-.04	.07	.44	.46
	3	.97	.97	1.0	.99	.97	.98	.60	.64	.67	.60	.68	.92	.91	.86	.10	.50	.84	.91	.92	.66	.62	.82	.61	.12	.02	.18	.22	-.05	.06	.44	.46
	4	.96	.97	.99	1.0	.97	.98	.57	.63	.66	.59	.67	.93	.90	.87	.07	.50	.84	.91	.92	.66	.59	.81	.63	.12	.03	.17	.22	-.05	.07	.43	.45
	5	.97	.96	.97	.97	1.0	.99	.58	.63	.65	.57	.66	.88	.88	.86	.06	.46	.80	.88	.93	.61	.59	.80	.60	.10	.01	.17	.21	-.08	.09	.42	.43
	6	.97	.97	.98	.98	.99	1.0	.56	.62	.65	.58	.66	.90	.87	.87	.05	.47	.81	.88	.94	.62	.59	.80	.62	.10	.01	.17	.21	-.08	.08	.42	.44
RESEARCHERID	7	.57	.56	.60	.57	.58	.56	1.0	.91	.88	.78	.89	.59	.59	.61	.03	.22	.48	.59	.57	.44	.54	.58	.37	.06	.01	-.04	.02	-.11	.08	.21	.25
	8	.61	.62	.64	.63	.63	.62	.91	1.0	.96	.85	.97	.67	.65	.71	-.02	.23	.55	.65	.62	.45	.53	.59	.40	.01	-.04	-.08	-.02	-.18	.11	.14	.20
	9	.67	.67	.67	.66	.65	.65	.88	.96	1.0	.95	.99	.69	.63	.73	-.03	.20	.56	.67	.69	.48	.52	.62	.45	.01	-.05	-.06	.00	-.15	.08	.17	.22
	10	.62	.63	.60	.59	.57	.58	.78	.85	.95	1.0	.93	.60	.54	.65	-.04	.14	.50	.61	.65	.46	.50	.58	.41	.03	-.01	-.03	.02	-.10	.09	.21	.25
	11	.66	.67	.68	.67	.66	.66	.89	.97	.99	.93	1.0	.70	.65	.73	-.02	.22	.57	.68	.68	.50	.52	.62	.46	.01	-.05	-.07	.00	-.15	.09	.16	.21
RESEARCHGATE	12	.89	.91	.92	.93	.88	.90	.59	.67	.69	.60	.70	1.0	.87	.89	.15	.51	.83	.91	.90	.69	.52	.75	.62	.11	.02	.12	.20	-.02	.01	.37	.39
	13	.86	.88	.91	.90	.88	.87	.59	.65	.63	.54	.65	.87	1.0	.78	.26	.63	.89	.94	.83	.70	.67	.77	.43	.19	.12	.18	.20	-.04	.10	.38	.40
	14	.86	.88	.86	.87	.86	.87	.61	.71	.73	.65	.73	.89	.78	1.0	-.04	.32	.68	.79	.89	.48	.45	.69	.59	.02	-.07	.01	.09	-.15	.05	.34	.37
	15	.05	.06	.10	.07	.06	.05	.03	-.02	-.03	-.04	-.02	.15	.26	-.04	1.0	.70	.34	.26	.06	.42	.30	.09	-.24	.17	.14	.16	.13	.25	-.12	.09	.11
	16	.43	.46	.50	.50	.46	.47	.22	.23	.20	.14	.22	.51	.63	.32	.70	1.0	.69	.63	.42	.71	.56	.49	.16	.29	.20	.21	.23	.08	-.03	.24	.29
	17	.78	.81	.84	.84	.80	.81	.48	.55	.56	.50	.57	.83	.89	.68	.34	.69	1.0	.95	.75	.82	.64	.74	.44	.25	.15	.16	.20	-.01	.02	.32	.34
	18	.87	.90	.91	.91	.88	.88	.59	.65	.67	.61	.68	.91	.94	.79	.26	.63	.95	1.0	.86	.80	.65	.78	.49	.24	.16	.18	.23	.00	.10	.40	.42
	19	.95	.94	.92	.92	.93	.94	.57	.62	.69	.65	.68	.90	.83	.89	.06	.42	.75	.86	1.0	.58	.53	.78	.61	.07	-.02	.07	.13	-.12	.06	.35	.36
	20	.60	.63	.66	.66	.61	.62	.44	.45	.48	.46	.50	.69	.70	.48	.42	.71	.82	.80	.58	1.0	.54	.61	.38	.22	.13	.18	.23	.06	.09	.28	.32
MENDELEY	21	.57	.58	.62	.59	.59	.59	.54	.53	.52	.50	.52	.52	.67	.45	.30	.56	.64	.65	.53	.54	1.0	.83	.27	.43	.36	.24	.21	.12	.06	.35	.39
	22	.77	.79	.82	.81	.80	.80	.58	.59	.62	.58	.62	.75	.77	.69	.09	.49	.74	.78	.78	.61	.83	1.0	.72	.26	.17	.17	.19	.00	.00	.35	.38
	23	.57	.61	.61	.63	.60	.62	.37	.40	.45	.41	.46	.62	.43	.59	-.24	.16	.44	.49	.61	.38	.27	.72	1.0	-.10	-.17	-.05	.04	-.15	-.06	.14	.14
	24	.11	.13	.12	.12	.10	.10	.06	.01	.01	.03	.01	.11	.19	.02	.17	.29	.25	.24	.07	.22	.43	.26	-.10	1.0	.96	.46	.43	.42	.24	.42	.43
	25	.02	.05	.02	.03	.01	.01	.01	-.04	-.05	-.01	-.05	.02	.12	-.07	.14	.20	.15	.16	-.02	.13	.36	.17	-.17	.96	1.0	.46	.41	.45	.27	.41	.41
TWITTER	26	.17	.18	.18	.17	.17	.17	-.04	-.08	-.06	-.03	-.07	.12	.18	.01	.16	.21	.16	.18	.07	.18	.24	.17	-.05	.46	.46	1.0	.87	.77	.29	.71	.69
	27	.21	.21	.22	.22	.21	.21	.02	-.02	.00	.02	.00	.20	.20	.09	.13	.23	.20	.23	.13	.23	.21	.19	.04	.43	.41	.87	1.0	.81	.40	.78	.77
	28	-.06	-.04	-.05	-.05	-.08	-.08	-.11	-.18	-.15	-.10	-.15	-.02	-.04	-.15	.25	.08	-.01	.00	-.12	.06	.12	.00	-.15	.42	.45	.77	.81	1.0	.18	.55	.53
	29	.08	.07	.06	.07	.09	.08	.08	.11	.08	.09	.09	.01	.10	.05	-.12	-.03	.02	.10	.06	.09	.06	.00	-.06	.24	.27	.29	.40	.18	1.0	.30	.32
	30	.45	.44	.44	.43	.42	.42	.21	.14	.17	.21	.16	.37	.38	.34	.09	.24	.32	.40	.35	.28	.35	.35	.14	.42	.41	.71	.78	.55	.30	1.0	.98
	31	.46	.46	.46	.45	.43	.44	.25	.20	.22	.25	.21	.39	.40	.37	.11	.29	.34	.42	.36	.32	.39	.38	.14	.43	.41	.69	.77	.53	.32	.98	1.0

COD	METRIC	COD	METRIC	COD	METRIC
1	GS_citations_total	**12**	RG_score	**23**	MEND_readers / document
2	GS_citations_last5	**13**	RG_publications	**24**	MEND_followers
3	GS_hindex_total	**14**	RG_impact_points	**25**	MEND_following
4	GS_hindex_last5	**15**	RG_following	**26**	TW_tweets
5	GS_i10_total	**16**	RG_followers	**27**	TW_followers
6	GS_i10_last5	**17**	RG_downloads	**28**	TW_following
7	RID_n_total_articles	**18**	RG_views	**29**	TW_dias
8	RID _n_articles_cit	**19**	RG_citations	**30**	TW_sum_retweets
9	RID _sum_times_cited	**20**	RG_profile_views	**31**	TW_h_retweets
10	RID _average_cit	**21**	MEND_pub		
11	RID _hindex	**22**	MEND_readers		

3.4. Data reliability

After describing the multifaceted presence (authors, documents, and sources) of the bibliometric community in *Google Scholar Citations*, describing the presence of the authors of this community in other social platforms, and analyzing the possible correlation between all metrics offered by these platforms, it is absolutely essential to face the discussion about the reliability of these metrics and platforms. In Science, if the data source and the instrument (that stores that data and computes the measures) are not reliable, the results achieved are meaningless and scientifically irrelevant; such groundless results should not be considered as proper scientific results until their validity is proven.

In Bibliometrics, there is a large tradition of studies addressing the errors related to the correct assignment of citations to documents in bibliometric databases, as well as the deficiencies in the design or application of bibliometric indicators (Sher, Garfield & Elias, 1966; Poyer, 1979; Garfield, 1983; Moed & Vriens, 1989; Garfield, 1990; Garcia-Perez, 2010; Franceschini, Maisano & Mastrogiacomo, 2015).

Since these platforms are quite new, there are still few in-depth empirical studies using representative samples which may allow us to make informed assertions about the reliability of these platforms. So far, there are only a few isolated analyses pointing out errors, inaccuracies and inconsistencies. Regrettably, there are not many of these interesting works, and they don't often go beyond reporting a few anecdotal issues. In this respect, we must highlight the great impact of Peter Jacsó's works, who analyzed the strengths and specially the weaknesses of Google Scholar (Jacsó 2005; 2006a; 2006b; 2008; 2010).

In order to contextualize all the data offered previously in this work, we present a final section providing insights about the different kinds of errors found in each of the platforms, with a special emphasis in *Google Scholar*, since it has been our main source of data.

3.4.1. The uncontrolled giant: Google Scholar & Google Scholar Citations

The errors that can compromise the metric portrait of an author offered by *Google Scholar* can be grouped into two main sections. First, the errors *Google Scholar* sometimes makes when it indexes a document or when it assigns citations to it. Second, the specific errors that are sometimes made during the creation of a *Google Scholar Citations* profile.

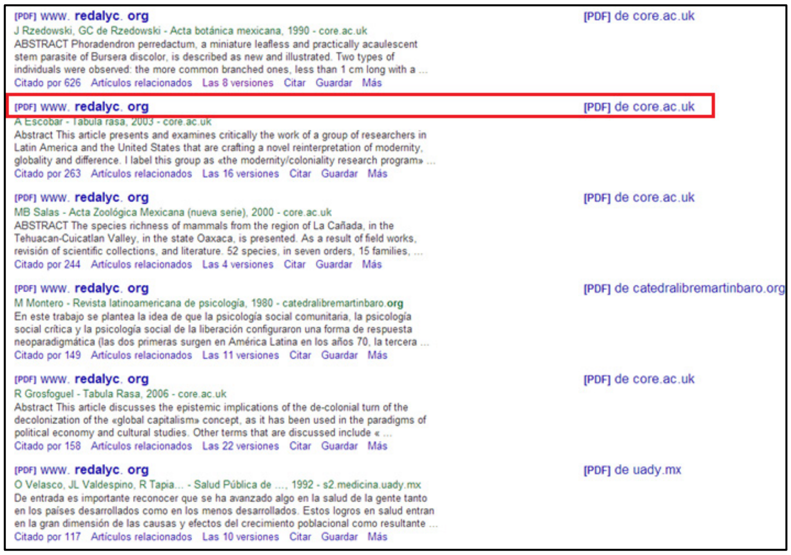
The former are a logical consequence of the tricky and complex task that is automatically searching the current academic papers available in the net. This task also involves merging in only one record all possible versions of the same work, and linking to it all documents in which it is cited (keeping in mind that these documents and references can be presented in the most varied formats). The latter are the ultimate responsibility of the author, who must periodically revise his/her profile in order to eliminate misattributed documents which might been included in the automatic weekly updates, clean the records by merging different versions of the same document when Google Scholar's algorithms are not able to detect their similarity, as well as improve and complete the bibliographic references of these documents (filling in blank fields in a document when Google Scholar hasn't been able to find that information).

Next, we classify, describe, and illustrate some of the most common mistakes in Google Scholar:

a) Incorrect identification of the title of the document

Google Scholar always tries to extract bibliographic information from the HTML Meta tags in a webpage. When there are no Meta tags available, it parses the webpage itself (the HTML code of the page, or even PDFs themselves). Even though its spiders are able to successfully parse pages with a quite broad range of different structures, and despite the fact that they have published a very clear set of inclusion guidelines, some parsing errors occasionally arise for documents extracted from websites with unusual layouts. It is not rare in these cases that an incorrect text string is selected as the title of the document. In Figure 6 we illustrate an example in which an incorrect string ("www.redalyc.org") has been selected as the title of the document in several records, probably because it is the string that is featured with a higher font size in the first page of the PDF document from which Google Scholar has parsed the bibliographic information. Note that the authors and the source publications are correctly assigned.

Figure 6. Document titles improperly identified in Google Scholar: URLs



[PDF] www. redalyc. org [PDF] de core.ac.uk
J Rzedowski, GC de Rzedowski - Acta botánica mexicana, 1990 - core.ac.uk
ABSTRACT Phoradendron perredactum, a miniature leafless and practically acaulescent stem parasite of Bursera discolor, is described as new and illustrated. Two types of individuals were observed: the more common branched ones, less than 1 cm long with a ...
Citado por 626 Artículos relacionados Las 8 versiones Citar Guardar Más

[PDF] www. redalyc. org [PDF] de core.ac.uk
A Escobar - Tabula rasa, 2003 - core.ac.uk
Abstract This article presents and examines critically the work of a group of researchers in Latin America and the United States that are crafting a novel reinterpretation of modernity, globality and difference. I label this group as «the modernity/coloniality research program» ...
Citado por 263 Artículos relacionados Las 16 versiones Citar Guardar Más

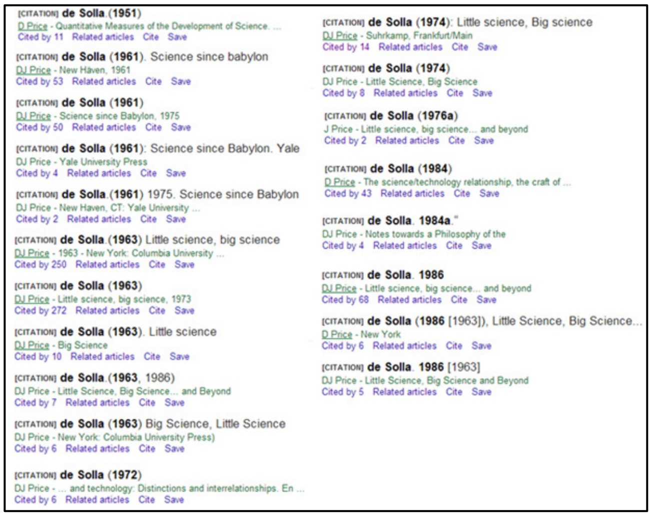
[PDF] www. redalyc. org [PDF] de core.ac.uk
MB Salas - Acta Zoológica Mexicana (nueva serie), 2000 - core.ac.uk
ABSTRACT The species richness of mammals from the region of La Cañada, in the Tehuacan-Cuicatlan Valley, in the state Oaxaca, is presented. As a result of field works, revisión of scientific collections, and literature. 52 species, in seven orders, 15 families, ...
Citado por 244 Artículos relacionados Las 4 versiones Citar Guardar Más

[PDF] www. redalyc. org [PDF] de catedralibremartinbaro.org
M Montero - Revista latinoamericana de psicología, 1980 - catedralibremartinbaro.org
En este trabajo se plantea la idea de que la psicología social comunitaria, la psicología social crítica y la psicología social de la liberación configuraron una forma de respuesta neoparadigmática (las dos primeras surgen en América Latina en los años 70, la tercera ...
Citado por 149 Artículos relacionados Las 11 versiones Citar Guardar Más

[PDF] www. redalyc. org [PDF] de core.ac.uk
R Grosfoguel - Tabula Rasa, 2006 - core.ac.uk
Abstract This article discusses the epistemic implications of the de-colonial turn of the decolonization of the «global capitalism» concept, as it has been used in the paradigms of political economy and cultural studies. Other terms that are discussed include « ...
Citado por 158 Artículos relacionados Las 22 versiones Citar Guardar Más

[PDF] www. redalyc. org [PDF] de uady.mx
O Velasco, JL Valdespino, R Tapia... - Salud Pública de ..., 1992 - s2.medicina.uady.mx
De entrada es importante reconocer que se ha avanzado algo en la salud de la gente tanto en los países desarrollados como en los menos desarrollados. Estos logros en salud entran en la gran dimensión de las causas y efectos del crecimiento poblacional como resultante ...
Citado por 117 Artículos relacionados Las 10 versiones Citar Guardar Más

In many other occasions, other text strings, such as the author's name and/or the year of publication, are incorrectly selected as the title of the document. In Figure 7 we can observe how "de Solla" has been selected as the title in many records.



[CITATION] **de Solla.(1951)**
D Price - Quantitative Measures of the Development of Science. ...
Cited by 11 Related articles Cite Save

[CITATION] **de Solla (1961)**. Science since babylon
DJ Price - New Häven, 1961
Cited by 53 Related articles Cite Save

[CITATION] **de Solla (1961)**
DJ Price - Science since Babylon, 1975
Cited by 50 Related articles Cite Save

[CITATION] **de Solla (1961)**: Science since Babylon. Yale
DJ Price - Yale University Press
Cited by 4 Related articles Cite Save

[CITATION] **de Solla.(1961)** 1975. Science since Babylon
DJ Price - New Haven, CT: Yale University ...
Cited by 2 Related articles Cite Save

[CITATION] **de Solla.(1963)** Little science, big science
DJ Price - 1963 - New York: Columbia University ...
Cited by 250 Related articles Cite Save

[CITATION] **de Solla (1963)**
DJ Price - Little science, big science, 1973
Cited by 272 Related articles Cite Save

[CITATION] **de Solla (1963)**. Little science
DJ Price - Big Science
Cited by 10 Related articles Cite Save

[CITATION] **de Solla.(1963**, 1986)
DJ Price - Little Science, Big Science... and Beyond
Cited by 7 Related articles Cite Save

[CITATION] **de Solla (1963)** Big Science, Little Science
DJ Price - New York: Columbia University Press)
Cited by 6 Related articles Cite Save

[CITATION] **de Solla (1972)**
DJ Price - ... and technology: Distinctions and interrelationships. En ...
Cited by 6 Related articles Cite Save

[CITATION] **de Solla (1974)**: Little science, Big science
DJ Price - Suhrkamp, Frankfurt/Main
Cited by 14 Related articles Cite Save

[CITATION] **de Solla (1974)**
DJ Price - Little Science, Big Science
Cited by 8 Related articles Cite Save

[CITATION] **de Solla (1976a)**
J Price - Little science, big science... and beyond
Cited by 2 Related articles Cite Save

[CITATION] **de Solla (1984)**
D Price - The science/technology relationship, the craft of ...
Cited by 43 Related articles Cite Save

[CITATION] **de Solla. 1984a**."
DJ Price - Notes towards a Philosophy of the
Cited by 4 Related articles Cite Save

[CITATION] **de Solla. 1986**
DJ Price - Little science, big science... and beyond
Cited by 68 Related articles Cite Save

[CITATION] **de Solla (1986** [1963]), Little Science, Big Science...
D Price - New York
Cited by 6 Related articles Cite Save

[CITATION] **de Solla. 1986** [1963]
DJ Price - Little Science, Big Science and Beyond
Cited by 5 Related articles Cite Save

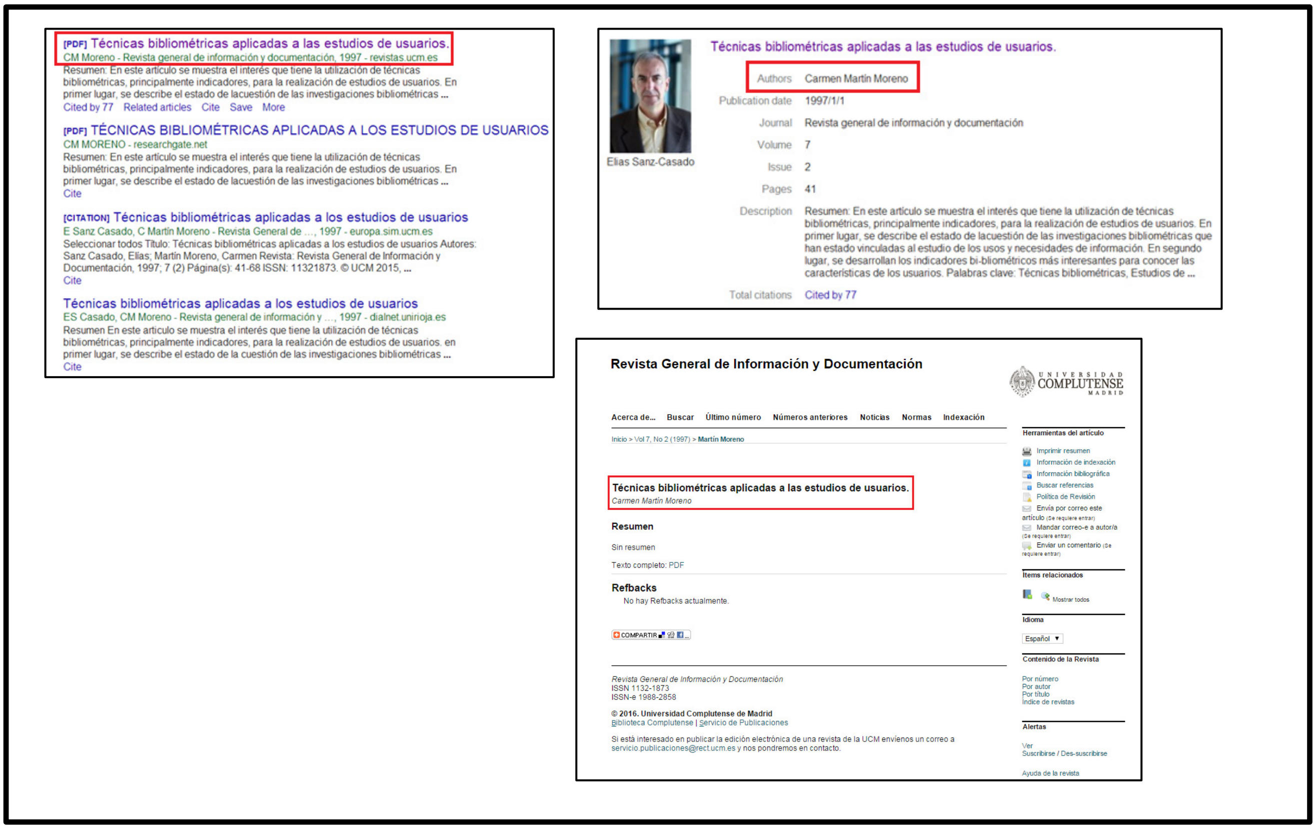
Figure 7. Author names incorrectly selected as document titles in Google Scholar

Source: https://scholar.google.com/scholar?start=0&q=allintitle:+%22de+solla%22+-Moravcsik+-gulls+-comments+-1922+-foreword+-Toward+-tribute+-space+-pensamento+-address+-appreciation&hl=en&as_sdt=0,5

b) Ghost authors

The topic of ghost authors, citations, and documents was approached by Jacsó in numerous works, mostly before *Google Scholar Citations* was launched. Although profiles have served to filter and correct many mistakes, some of them still persist, especially if authors do not clean their personal profiles. In Figure 8 we can see one such example. In this case, the record only displays one person as the author of the article (Carmen Martín Moreno), when in fact the article was written by two authors (Elías Sanz-Casado and Carmen Martín Moreno).

In this case, Google Scholar extracted the bibliographic information from the HTML Meta tags in the website of the journal where the article was published, but, as we can see in Figure 8 (bottom image), these metadata were already incorrect (the title should read "Técnicas bibliométricas aplicadas a **los** estudios de usuarios"), and incomplete (Elías Sanz-Casado is missing from the record). Nonetheless, thanks to Google Scholar Citations, Elías was able to add the document to his profile, even if his name is still missing from the authors field (Figure 8, top left).



[PDF] Técnicas bibliométricas aplicadas a las estudios de usuarios.
CM Moreno - Revista general de información y documentación, 1997 - revistas.ucm.es
Resumen: En este artículo se muestra el interés que tiene la utilización de técnicas bibliométricas, principalmente indicadores, para la realización de estudios de usuarios. En primer lugar, se describe el estado de lacuestión de las investigaciones bibliométricas ...
Cited by 77 Related articles Cite Save More

[PDF] TÉCNICAS BIBLIOMÉTRICAS APLICADAS A LOS ESTUDIOS DE USUARIOS
CM MORENO - researchgate.net
Resumen: En este artículo se muestra el interés que tiene la utilización de técnicas bibliométricas, principalmente indicadores, para la realización de estudios de usuarios. En primer lugar, se describe el estado de lacuestión de las investigaciones bibliométricas ...
Cite

[CITATION] Técnicas bibliométricas aplicadas a los estudios de usuarios
E Sanz Casado, C Martín Moreno - Revista General de ..., 1997 - europa.sim.ucm.es
Seleccionar todos Título: Técnicas bibliométricas aplicadas a los estudios de usuarios Autores: Sanz Casado, Elías; Martín Moreno, Carmen Revista: Revista General de Información y Documentación, 1997; 7 (2) Página(s): 41-68 ISSN: 11321873. © UCM 2015, ...
Cite

Técnicas bibliométricas aplicadas a los estudios de usuarios
ES Casado, CM Moreno - Revista general de información y ..., 1997 - dialnet.unirioja.es
Resumen En este articulo se muestra el interés que tiene la utilización de técnicas bibliométricas, principalmente indicadores, para la realización de estudios de usuarios. en primer lugar, se describe el estado de la cuestión de las investigaciones bibliométricas ...
Cite

Elias Sanz-Casado

Técnicas bibliométricas aplicadas a las estudios de usuarios.

Authors	Carmen Martín Moreno
Publication date	1997/1/1
Journal	Revista general de información y documentación
Volume	7
Issue	2
Pages	41
Description	Resumen: En este artículo se muestra el interés que tiene la utilización de técnicas bibliométricas, principalmente indicadores, para la realización de estudios de usuarios. En primer lugar, se describe el estado de lacuestión de las investigaciones bibliométricas que han estado vinculadas al estudio de los usos y necesidades de información. En segundo lugar, se desarrollan los indicadores bi-bliométricos más interesantes para conocer las características de los usuarios. Palabras clave: Técnicas bibliométricas, Estudios de ...
Total citations	Cited by 77

Revista General de Información y Documentación

UNIVERSIDAD COMPLUTENSE MADRID
Acerca de... Buscar Último número Números anteriores Noticias Normas Indexación

Inicio > Vol 7, No 2 (1997) > Martín Moreno

Técnicas bibliométricas aplicadas a las estudios de usuarios.
Carmen Martín Moreno

Resumen

Sin resumen

Texto completo: PDF

Refbacks

No hay Refbacks actualmente.

Revista General de Información y Documentación
ISSN 1132-1873
ISSN-e 1988-2858

© 2016. Universidad Complutense de Madrid
Biblioteca Complutense | Servicio de Publicaciones

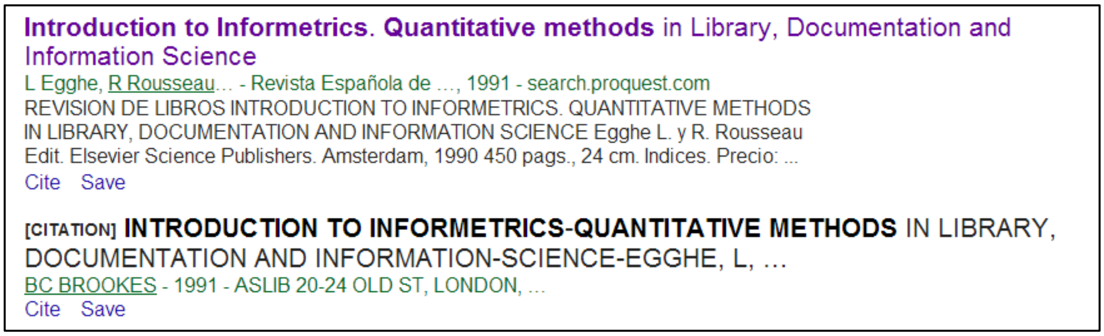
Si está interesado en publicar la edición electrónica de una revista de la UCM envíenos un correo a servicio.publicaciones@rect.ucm.es y nos pondremos en contacto.

Herramientas del artículo
Imprimir resumen
Información de indexación
Información bibliográfica
Buscar referencias
Política de Revisión
Envía por correo este artículo (Se requiere entrar)
Mandar correo-e a autor/a (Se requiere entrar)
Enviar un comentario (Se requiere entrar)
Ítems relacionados
Mostrar todos
Idioma
Español ▼
Contenido de la Revista
Por número
Por autor
Por título
Índice de revistas
Alertas
Ver
Suscribirse / Des-suscribirse
Ayuda de la revista

Figure 8. Missing authors in primary versions of documents in Google Scholar

c) Book reviews indexed as books

Among the most common mistakes in document identification is mistaking the review of a book for the book itself. In Figure 9 we show two different records which correspond with book reviews of the work “Introduction to informetrics. Quantitative methods in Library, Documentation and Information Science” by Egghe and Rousseau. At a first glance the first record (Figure 9; top) looks like a normal record, since the title and authors of the book have been correctly identified. However, the record actually points to a review of the book published in *Revista Española de Documentación Científica*. The second record (Figure 9; bottom), is also a review of the book which was published in *Aslib Proceedings*. In this case, the author of the review is the one who appears in the GS record (Brookes).



Introduction to Informetrics. Quantitative methods in Library, Documentation and Information Science

L Egghe, R Rousseau… - Revista Española de …, 1991 - search.proquest.com

REVISION DE LIBROS INTRODUCTION TO INFORMETRICS. QUANTITATIVE METHODS IN LIBRARY, DOCUMENTATION AND INFORMATION SCIENCE Egghe L. y R. Rousseau Edit. Elsevier Science Publishers. Amsterdam, 1990 450 pags., 24 cm. Indices. Precio: ...

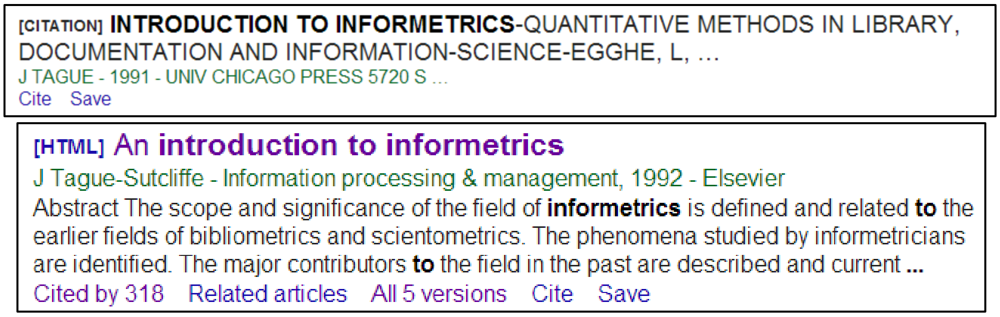
[CITATION] **INTRODUCTION TO INFORMETRICS-QUANTITATIVE METHODS** IN LIBRARY, DOCUMENTATION AND INFORMATION-SCIENCE-EGGHE, L, …

BC BROOKES - 1991 - ASLIB 20-24 OLD ST, LONDON, …

Figure 9. Authorship and attribution of book reviews

c) Incorrect attribution of documents to authors

Somewhat related to the previous error is the attribution of a document to the wrong authors. In Figure 10 we observe a special case: the book “Introduction to informetrics. Quantitative methods in Library, Documentation and Information Science” by Egghe and Rousseau, is wrongly attributed to Tague-Sutcliffe, probably because this author has a short publication in the journal Information Processing & Management (Figure 10; bottom) with a similar title (“An introduction to informetrics”).



[CITATION] **INTRODUCTION TO INFORMETRICS**-QUANTITATIVE METHODS IN LIBRARY, DOCUMENTATION AND INFORMATION-SCIENCE-EGGHE, L, …

J TAGUE - 1991 - UNIV CHICAGO PRESS 5720 S …

[HTML] An **introduction to informetrics**

J Tague-Sutcliffe - Information processing & management, 1992 - Elsevier

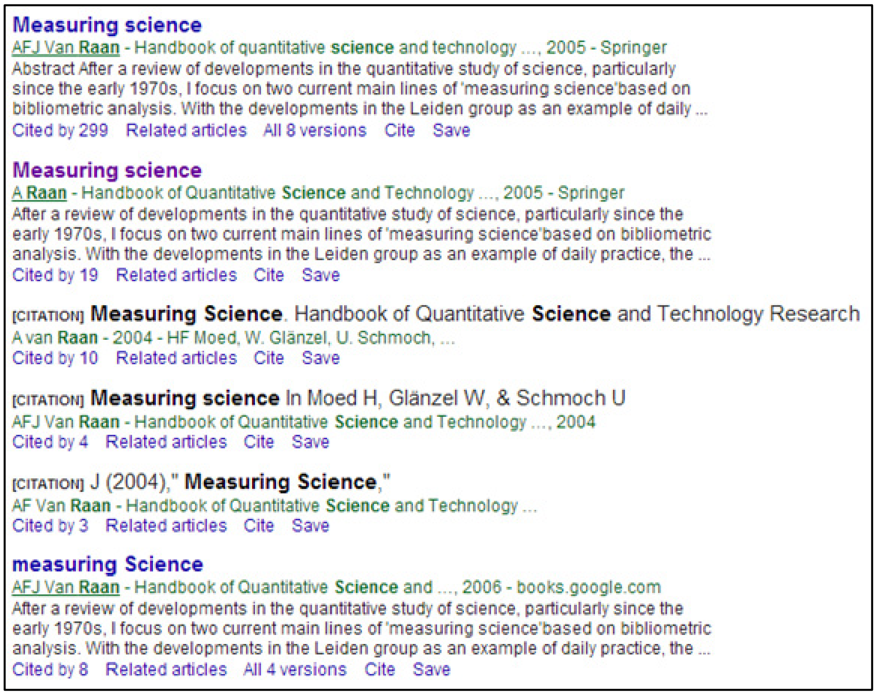
Abstract The scope and significance of the field of **informetrics** is defined and related **to** the earlier fields of bibliometrics and scientometrics. The phenomena studied by informetricians are identified. The major contributors **to** the field in the past are described and current **...**

Figure 10. Authorship improperly assigned in Google Scholar

d) Failing to merge all versions of a same document into one record

Although the algorithms for grouping versions work well in most cases, Google Scholar sometimes fails to realize that two or more records it has indexed actually represent the same document. This happens when there are enough formal differences between the metadata of the two versions (differences in the way the name of the authors have been stored, in the title, the year of publication…), that Google Scholar judges they’re not similar enough to be the same document. This issue mostly affects document types other than journal articles (books, book chapters, reports), but duplicate articles also exist. Articles translated into one or more languages are an extreme example: in those cases, the title of the original version is completely different to that of the translated version, so it is understandable that Google Scholar doesn’t realize they are the same document. From a bibliometric perspective, however, their citation counts shouldn’t be split.

This issue obviously affects the citation count of some documents. In Figure 11 we can observe how this phenomenon affects a book chapter: “Measuring science”, by Van Raan.



Measuring science
AFJ Van **Raan** - Handbook of quantitative **science** and technology ..., 2005 - Springer
Abstract After a review of developments in the quantitative study of science, particularly since the early 1970s, I focus on two current main lines of 'measuring science'based on bibliometric analysis. With the developments in the Leiden group as an example of daily ...
Cited by 299 Related articles All 8 versions Cite Save

Measuring science
A **Raan** - Handbook of Quantitative **Science** and Technology ..., 2005 - Springer
After a review of developments in the quantitative study of science, particularly since the early 1970s, I focus on two current main lines of 'measuring science'based on bibliometric analysis. With the developments in the Leiden group as an example of daily practice, the ...
Cited by 19 Related articles Cite Save

[CITATION] **Measuring Science**. Handbook of Quantitative **Science** and Technology Research
A van **Raan** - 2004 - HF Moed, W. Glänzel, U. Schmoch, ...
Cited by 10 Related articles Cite Save

[CITATION] **Measuring science** In Moed H, Glänzel W, & Schmoch U
AFJ Van **Raan** - Handbook of Quantitative **Science** and Technology ..., 2004
Cited by 4 Related articles Cite Save

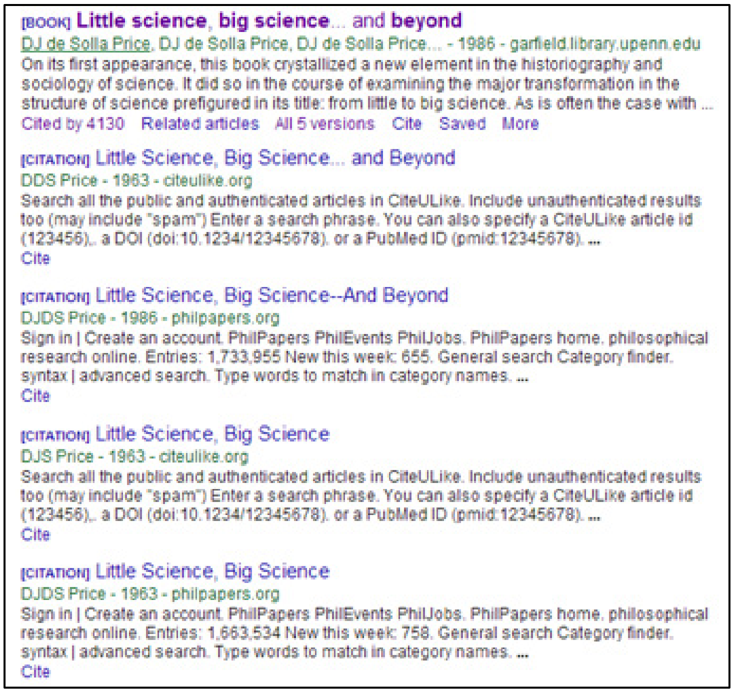
[CITATION] J (2004)," **Measuring Science**,"
AF Van **Raan** - Handbook of Quantitative **Science** and Technology ...
Cited by 3 Related articles Cite Save

measuring Science
AFJ Van **Raan** - Handbook of Quantitative **Science** and ..., 2006 - books.google.com
After a review of developments in the quantitative study of science, particularly since the early 1970s, I focus on two current main lines of 'measuring science'based on bibliometric analysis. With the developments in the Leiden group as an example of daily practice, the ...
Cited by 8 Related articles All 4 versions Cite Save

Figure 11. Versions of book chapters improperly tied in Google Scholar

d) Grouping different editions of the same book in a single record

Conversely to the previous error, Google Scholar sometimes groups together records that should stay separate, for example in the cases when there are different editions of the same book (a new book edition provides new content, contrary to a reprinting of a book, which is identical to the previous printing). Figure 12 illustrates the case of “Little Science, big Science”, written by Price. This book was first published in 1963 by Columbia University Press, and reedited in 1986 under the title “Little science, big science… and beyond”, an edition that contained the original text of the book, as well as seven of his most famous articles.



[BOOK] Little science, big science... and beyond
DJ de Solla Price, DJ de Solla Price, DJ de Solla Price... - 1986 - garfield.library.upenn.edu
On its first appearance, this book crystallized a new element in the historiography and sociology of science. It did so in the course of examining the major transformation in the structure of science prefigured in its title: from little to big science. As is often the case with ...
Cited by 4130 Related articles All 5 versions Cite Saved More

[CITATION] Little Science, Big Science... and Beyond
DDS Price - 1963 - citeulike.org
Search all the public and authenticated articles in CiteULike. Include unauthenticated results too (may include "spam") Enter a search phrase. You can also specify a CiteULike article id (123456),. a DOI (doi:10.1234/12345678). or a PubMed ID (pmid:12345678). ...
Cite

[CITATION] Little Science, Big Science--And Beyond
DJDS Price - 1986 - philpapers.org
Sign in | Create an account. PhilPapers PhilEvents PhilJobs. PhilPapers home. philosophical research online. Entries: 1,733,955 New this week: 655. General search Category finder. syntax | advanced search. Type words to match in category names. ...
Cite

[CITATION] Little Science, Big Science
DJS Price - 1963 - citeulike.org
Search all the public and authenticated articles in CiteULike. Include unauthenticated results too (may include "spam") Enter a search phrase. You can also specify a CiteULike article id (123456),. a DOI (doi:10.1234/12345678). or a PubMed ID (pmid:12345678). ...
Cite

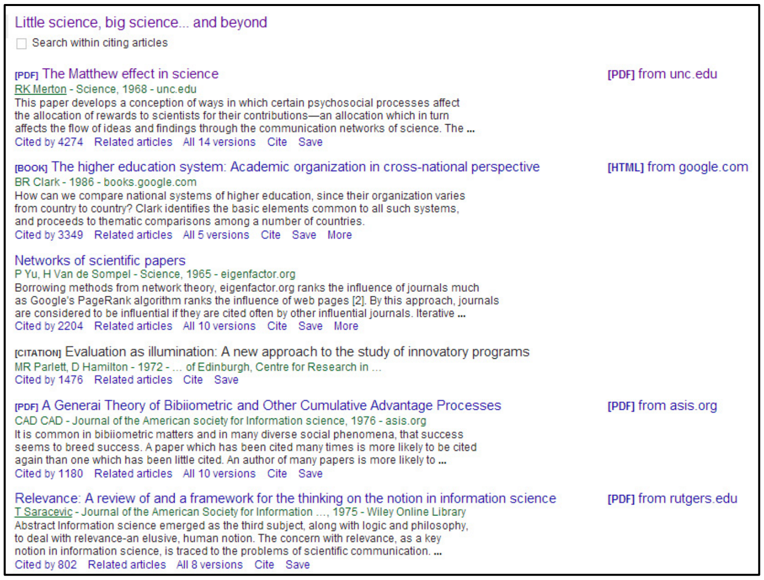
[CITATION] Little Science, Big Science
DJDS Price - 1963 - philpapers.org
Sign in | Create an account. PhilPapers PhilEvents PhilJobs. PhilPapers home. philosophical research online. Entries: 1,663,534 New this week: 758. General search Category finder. syntax | advanced search. Type words to match in category names. ...
Cite

Figure 12. Different book editions tied in Google Scholar

The primary version (which has received 4,130 citations) is the edition from 1986, but among its versions are several records pointing to the version from 1963. Different editions of the same book should be treated as separate documents when computing citations because their content may be very different.

Of course, aumatically detecting and managing these details is a very complex task, and only a very tiny fraction of the documents indexed in Google Scholar (the most influential manuals and seminal works) would benefit from this thorough treatment. We must not forget that *Google Scholar* is, first of all, a search tool devoted to helping researchers find academic information. A great percentage of users probably don't care about the different editions of a book, and those who do probably just want the most recent one. That may be the reason why Google Scholar usually displays the most recent edition of a book as the primary version. The use of separate entries for different editions is something just a few people, like librarians, would be interested in.

In any case, this may have an important effect in citation counts because citations to different editions (providing different content) are added together. In Figure 13 we can see how the 1986 edition of the book is receiving citations that were actually made to the original work published in 1963.



Little science, big science... and beyond

□ Search within citing articles

[PDF] The Matthew effect in science — [PDF] from unc.edu
RK Merton - Science, 1968 - unc.edu
This paper develops a conception of ways in which certain psychosocial processes affect
the allocation of rewards to scientists for their contributions—an allocation which in turn
affects the flow of ideas and findings through the communication networks of science. The ...
Cited by 4274 Related articles All 14 versions Cite Save

[BOOK] The higher education system: Academic organization in cross-national perspective — [HTML] from google.com
BR Clark - 1986 - books.google.com
How can we compare national systems of higher education, since their organization varies
from country to country? Clark identifies the basic elements common to all such systems,
and proceeds to thematic comparisons among a number of countries.
Cited by 3349 Related articles All 5 versions Cite Save More

Networks of scientific papers
P Yu, H Van de Sompel - Science, 1965 - eigenfactor.org
Borrowing methods from network theory, eigenfactor.org ranks the influence of journals much
as Google's PageRank algorithm ranks the influence of web pages [2]. By this approach, journals
are considered to be influential if they are cited often by other influential journals. Iterative ...
Cited by 2204 Related articles All 10 versions Cite Save More

[CITATION] Evaluation as illumination: A new approach to the study of innovatory programs
MR Parlett, D Hamilton - 1972 - ... of Edinburgh, Centre for Research in ...
Cited by 1476 Related articles Cite Save

[PDF] A Generai Theory of Bibiiometric and Other Cumulative Advantage Processes — [PDF] from asis.org
CAD CAD - Journal of the American society for Information science, 1976 - asis.org
It is common in bibiiometric matters and in many diverse social phenomena, that success
seems to breed success. A paper which has been cited many times is more likely to be cited
again than one which has been little cited. An author of many papers is more likely to ...
Cited by 1180 Related articles All 10 versions Cite Save

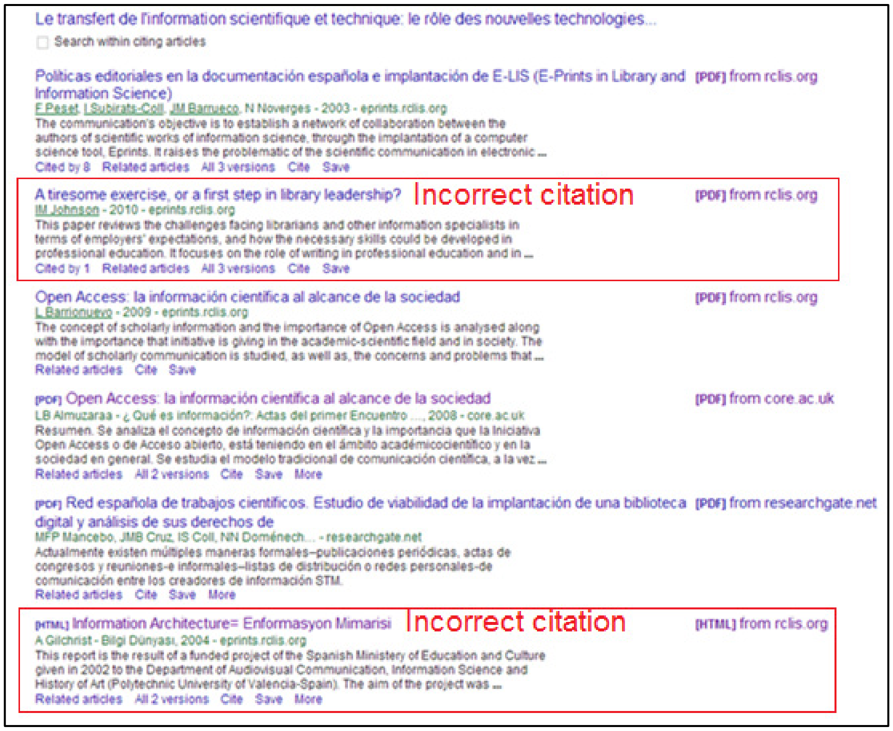
Relevance: A review of and a framework for the thinking on the notion in information science — [PDF] from rutgers.edu
T Saracevic - Journal of the American Society for Information ..., 1975 - Wiley Online Library
Abstract Information science emerged as the third subject, along with logic and philosophy,
to deal with relevance-an elusive, human notion. The concern with relevance, as a key
notion in information science, is traced to the problems of scientific communication. ...
Cited by 802 Related articles All 8 versions Cite Save

Figure 13. Citations to different book editions tied in Google Scholar

e) Improper attribution of citations to a document

Document citation counts in Google Scholar are also affected by the attribution of “ghost” citations to documents, that is, citations that aren’t actually there when we examine the citing document. Figure 14 shows an example of this issue: the work “Le transfert de l'information scientifique et technique: le rôle des nouvelles technologies de l'information face à la crise du modèle actuel de communication écrite” has allegedly received eight citations, but if we manually examine the second document in the list (marked in red), we can’t find any mention of the cited work. This phenomenon has been frequently observed in documents stored in the E-LIS repository.[60]

[60] http://eprints.rclis.org



Le transfert de l'information scientifique et technique: le rôle des nouvelles technologies...

□ Search within citing articles

Políticas editoriales en la documentación española e implantación de E-LIS (E-Prints in Library and Information Science) [PDF] from rclis.org
F Peset, I Subirats-Coll, JM Barrueco, N Noverges - 2003 - eprints.rclis.org
The communication's objective is to establish a network of collaboration between the authors of scientific works of information science, through the implantation of a computer science tool, Eprints. It raises the problematic of the scientific communication in electronic ...
Cited by 8 Related articles All 3 versions Cite Save

A tiresome exercise, or a first step in library leadership? Incorrect citation [PDF] from rclis.org
IM Johnson - 2010 - eprints.rclis.org
This paper reviews the challenges facing librarians and other information specialists in terms of employers' expectations, and how the necessary skills could be developed in professional education. It focuses on the role of writing in professional education and in ...
Cited by 1 Related articles All 3 versions Cite Save

Open Access: la información científica al alcance de la sociedad [PDF] from rclis.org
L Barrionuevo - 2009 - eprints.rclis.org
The concept of scholarly information and the importance of Open Access is analysed along with the importance that initiative is giving in the academic-scientific field and in society. The model of scholarly communication is studied, as well as, the concerns and problems that ...
Related articles Cite Save

[PDF] Open Access: la información científica al alcance de la sociedad [PDF] from core.ac.uk
LB Almuzaraa - ¿ Qué es información?: Actas del primer Encuentro ..., 2008 - core.ac.uk
Resumen. Se analiza el concepto de información científica y la importancia que la Iniciativa Open Access o de Acceso abierto, está teniendo en el ámbito académicocientífico y en la sociedad en general. Se estudia el modelo tradicional de comunicación científica, a la vez ...
Related articles All 2 versions Cite Save More

[PDF] Red española de trabajos científicos. Estudio de viabilidad de la implantación de una biblioteca digital y análisis de sus derechos de [PDF] from researchgate.net
MFP Mancebo, JMB Cruz, IS Coll, NN Doménech... - researchgate.net
Actualmente existen múltiples maneras formales–publicaciones periódicas, actas de congresos y reuniones-e informales–listas de distribución o redes personales-de comunicación entre los creadores de información STM.
Related articles Cite Save More

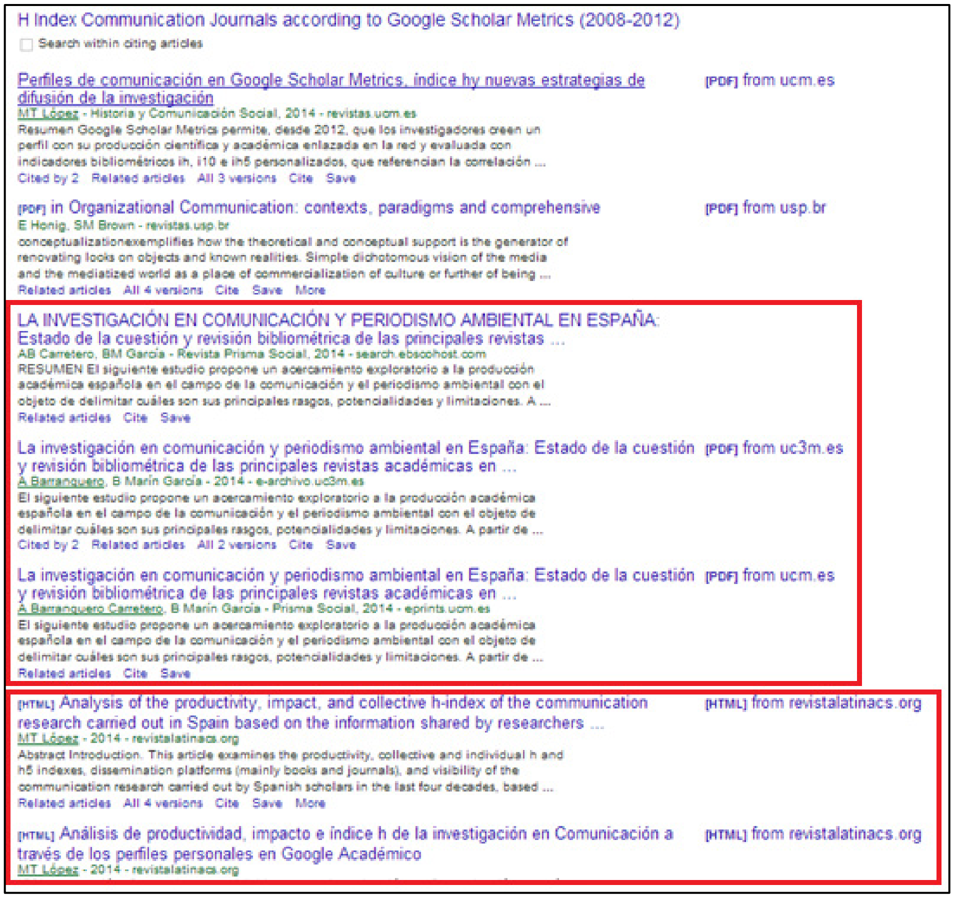
[HTML] Information Architecture= Enformasyon Mimarisi Incorrect citation [HTML] from rclis.org
A Gilchrist - Bilgi Dünyası, 2004 - eprints.rclis.org
This report is the result of a funded project of the Spanish Ministery of Education and Culture given in 2002 to the Department of Audiovisual Communication, Information Science and History of Art (Polytechnic University of Valencia-Spain). The aim of the project was ...
Related articles All 2 versions Cite Save More

Figure 14. Appearance of false citations

f) Duplicate citations

This phenomenon is a consequence of an issue previously discussed. When Google Scholar fails to realise that two records are actually versions of the same document, these versions are stored as if they were different documents. Therefore, each of them provides its own set of citations to the citation pool. Since the two sets of citations are probably identical, each cited document will receive two citations from what is actually only one document, thus falsely inflating their citation counts.

In Figure 15 we observe a double example of this phenomenon. In the first case (first red rectangle), there are three versions of the same document. Note the differences in the way the authors' names are stored, since this is probably the reason why the records weren't merged into one. In the second case (second red rectangle), the two records refer to the same document (the first one is the English version of the article, and the second one is the Spanish version).



H Index Communication Journals according to Google Scholar Metrics (2008-2012)

□ Search within citing articles

Perfiles de comunicación en Google Scholar Metrics, índice hy nuevas estrategias de difusión de la investigación [PDF] from ucm.es
MT López - Historia y Comunicación Social, 2014 - revistas.ucm.es
Resumen Google Scholar Metrics permite, desde 2012, que los investigadores creen un perfil con su producción científica y académica enlazada en la red y evaluada con indicadores bibliométricos ih, i10 e ih5 personalizados, que referencian la correlación ...
Cited by 2 Related articles All 3 versions Cite Save

[PDF] in Organizational Communication: contexts, paradigms and comprehensive [PDF] from usp.br
E Honig, SM Brown - revistas.usp.br
conceptualizationexemplifies how the theoretical and conceptual support is the generator of renovating looks on objects and known realities. Simple dichotomous vision of the media and the mediatized world as a place of commercialization of culture or further of being ...
Related articles All 4 versions Cite Save More

LA INVESTIGACIÓN EN COMUNICACIÓN Y PERIODISMO AMBIENTAL EN ESPAÑA: Estado de la cuestión y revisión bibliométrica de las principales revistas ...
AB Carretero, BM García - Revista Prisma Social, 2014 - search.ebscohost.com
RESUMEN El siguiente estudio propone un acercamiento exploratorio a la producción académica española en el campo de la comunicación y el periodismo ambiental con el objeto de delimitar cuáles son sus principales rasgos, potencialidades y limitaciones. A ...
Related articles Cite Save

La investigación en comunicación y periodismo ambiental en España: Estado de la cuestión y revisión bibliométrica de las principales revistas académicas en ... [PDF] from uc3m.es
A Barranquero, B Marín García - 2014 - e-archivo.uc3m.es
El siguiente estudio propone un acercamiento exploratorio a la producción académica española en el campo de la comunicación y el periodismo ambiental con el objeto de delimitar cuáles son sus principales rasgos, potencialidades y limitaciones. A partir de ...
Cited by 2 Related articles All 2 versions Cite Save

La investigación en comunicación y periodismo ambiental en España: Estado de la cuestión y revisión bibliométrica de las principales revistas académicas en ... [PDF] from ucm.es
A Barranquero Carretero, B Marín García - Prisma Social, 2014 - eprints.ucm.es
El siguiente estudio propone un acercamiento exploratorio a la producción académica española en el campo de la comunicación y el periodismo ambiental con el objeto de delimitar cuáles son sus principales rasgos, potencialidades y limitaciones. A partir de ...
Related articles Cite Save

[HTML] Analysis of the productivity, impact, and collective h-index of the communication research carried out in Spain based on the information shared by researchers ... [HTML] from revistalatinacs.org
MT López - 2014 - revistalatinacs.org
Abstract Introduction. This article examines the productivity, collective and individual h and h5 indexes, dissemination platforms (mainly books and journals), and visibility of the communication research carried out by Spanish scholars in the last four decades, based ...
Related articles All 4 versions Cite Save More

[HTML] Análisis de productividad, impacto e índice h de la investigación en Comunicación a través de los perfiles personales en Google Académico [HTML] from revistalatinacs.org
MT López - 2014 - revistalatinacs.org

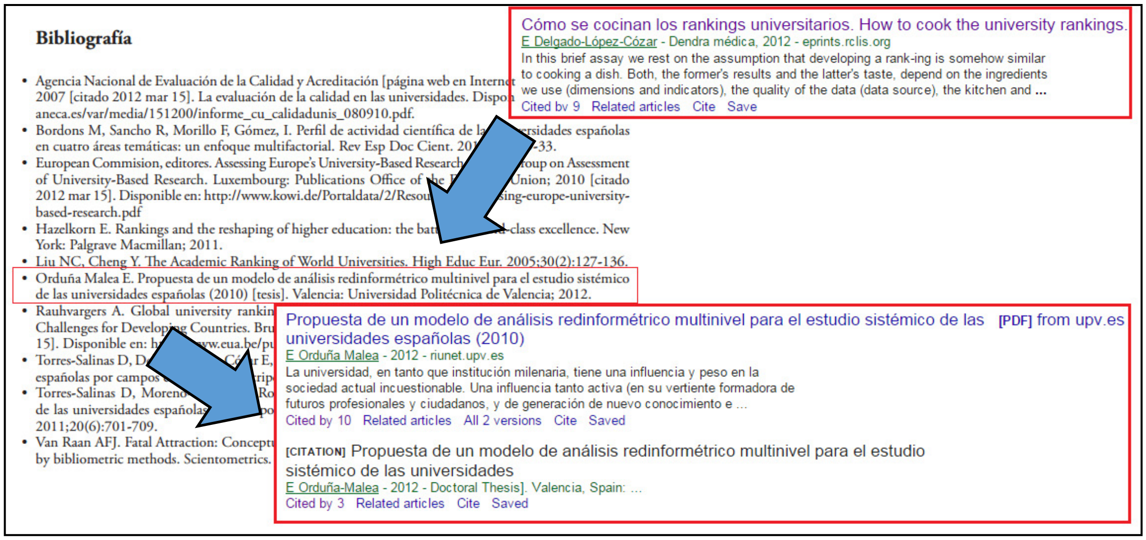
Figure 15. Duplicate citations in Google Scholar

g) Missing citations

There are cases when Google Scholar's parser fails to match a cited reference inside document, with the record of the document it is citing. When Google Scholar parses the reference section within an article, it tries to find a match for these references in its records, but if for some reason the reference hasn't been correctly recorded (authors of the citing article may have made a mistake when citing it or used an uncommon reference format Google Scholar doesn't understand) the system will be unable to make the connection between the two documents.

However, we also find examples in which no apparent mistake has been made in the citing document, but still the citation isn't attributed to the cited document.

In order to illustrate this issue, in Figure 16 we show how a document ("How to cook the university rankings") is citing in its reference section other document (a doctoral thesis). However, this citation doesn't appear as one of the 13 citations that the thesis has received according to *Google Scholar*. The reason is unknown. At the time the citing document was first indexed, the connection wasn't made for some reason, and this error hasn't been solved since. Typos in the PDF can also generate this kind of error.



Bibliografía

- Agencia Nacional de Evaluación de la Calidad y Acreditación [página web en Internet 2007 [citado 2012 mar 15]. La evaluación de la calidad en las universidades. Dispon aneca.es/var/media/151200/informe_cu_calidadunis_080910.pdf.
- Bordons M, Sancho R, Morillo F, Gómez, I. Perfil de actividad científica de la ersidades españolas en cuatro áreas temáticas: un enfoque multifactorial. Rev Esp Doc Cient. 201 -33.
- European Commision, editores. Assessing Europe's University-Based Research roup on Assessment of University-Based Research. Luxembourg: Publications Office of the Union; 2010 [citado 2012 mar 15]. Disponible en: http://www.kowi.de/Portaldata/2/Resou sing-europe-university-based-research.pdf
- Hazelkorn E. Rankings and the reshaping of higher education: the batt -class excellence. New York: Palgrave Macmillan; 2011.
- Liu NC, Cheng Y. The Academic Ranking of World Universities. High Educ Eur. 2005;30(2):127-136.
- Orduña Malea E. Propuesta de un modelo de análisis redinformétrico multinivel para el estudio sistémico de las universidades españolas (2010) [tesis]. Valencia: Universidad Politécnica de Valencia; 2012.
- Rauhvargers A. Global university rankin Challenges for Developing Countries. Bru 15]. Disponible en: h ww.eua.be/pu
- Torres-Salinas D, D C r E, españolas por campos rip
- Torres-Salinas D, Moreno Ro de las universidades españolas po 2011;20(6):701-709.
- Van Raan AFJ. Fatal Attraction: Conceptu by bibliometric methods. Scientometrics.

Cómo se cocinan los rankings universitarios. How to cook the university rankings.
E Delgado-López-Cózar - Dendra médica, 2012 - eprints.rclis.org
In this brief assay we rest on the assumption that developing a rank-ing is somehow similar to cooking a dish. Both, the former's results and the latter's taste, depend on the ingredients we use (dimensions and indicators), the quality of the data (data source), the kitchen and ...
Cited bv 9 Related articles Cite Save

Propuesta de un modelo de análisis redinformétrico multinivel para el estudio sistémico de las universidades españolas (2010) [PDF] from upv.es
E Orduña Malea - 2012 - riunet.upv.es
La universidad, en tanto que institución milenaria, tiene una influencia y peso en la sociedad actual incuestionable. Una influencia tanto activa (en su vertiente formadora de futuros profesionales y ciudadanos, y de generación de nuevo conocimiento e ...
Cited by 10 Related articles All 2 versions Cite Saved

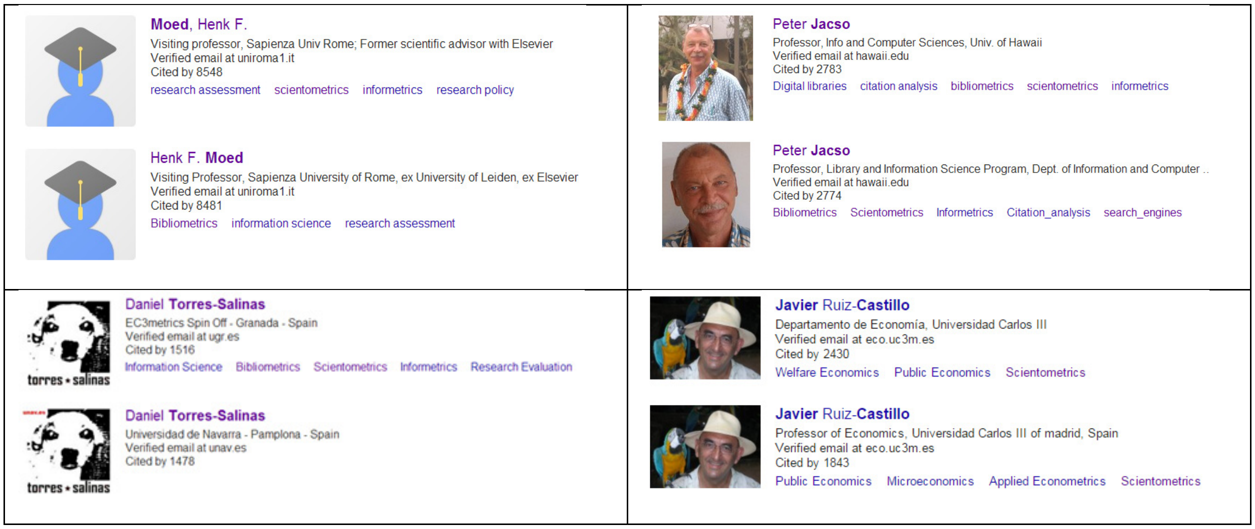
[CITATION] Propuesta de un modelo de análisis redinformétrico multinivel para el estudio sistémico de las universidades
E Orduña-Malea - 2012 - Doctoral Thesis]. Valencia, Spain: ...
Cited by 3 Related articles Cite Saved

Figure 16. Citations unrevealed in Google Scholar

All the errors previously described are related directly with the *Google Scholar* database (and are concerned with how the automatic parser works). Next we show some of the mistakes identified in the elaboration of bibliographic profiles through *Google Scholar Citations*:

a) Duplicate profiles

Since the only restriction to create a public academic profile in *Google Scholar Citations* is to provide a valid email, an author (or anyone really) may create as many profiles as he/she wants. This opens the door to the existence of duplicate profiles, that is, different profiles about the same person. In Figure 17 we present some examples of duplicate profiles of authors related to the field of Bibliometrics. The differences in citation counts between profiles are sometimes quite high (for example, one of the profiles belonging to Ruiz-Castillo achieves 1,843 citations whereas in the second profile the figure goes up to 2,430).



Moed, Henk F.
Visiting professor, Sapienza Univ Rome; Former scientific advisor with Elsevier
Verified email at uniroma1.it
Cited by 8548
research assessment scientometrics informetrics research policy

Henk F. Moed
Visiting Professor, Sapienza University of Rome, ex University of Leiden, ex Elsevier
Verified email at uniroma1.it
Cited by 8481
Bibliometrics information science research assessment

Peter Jacso
Professor, Info and Computer Sciences, Univ. of Hawaii
Verified email at hawaii.edu
Cited by 2783
Digital libraries citation analysis bibliometrics scientometrics informetrics

Peter Jacso
Professor, Library and Information Science Program, Dept. of Information and Computer ..
Verified email at hawaii.edu
Cited by 2774
Bibliometrics Scientometrics Informetrics Citation_analysis search_engines

Daniel Torres-Salinas
EC3metrics Spin Off - Granada - Spain
Verified email at ugr.es
Cited by 1516
Information Science Bibliometrics Scientometrics Informetrics Research Evaluation

Daniel Torres-Salinas
Universidad de Navarra - Pamplona - Spain
Verified email at unav.es
Cited by 1478

Javier Ruiz-Castillo
Departamento de Economía, Universidad Carlos III
Verified email at eco.uc3m.es
Cited by 2430
Welfare Economics Public Economics Scientometrics

Javier Ruiz-Castillo
Professor of Economics, Universidad Carlos III of madrid, Spain
Verified email at eco.uc3m.es
Cited by 1843
Public Economics Microeconomics Applied Econometrics Scientometrics

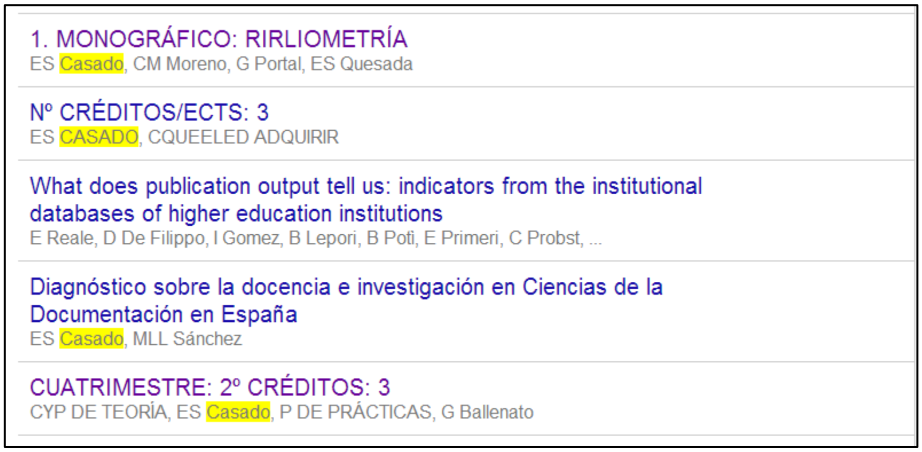
Figure 17. Duplicate profiles in Google Scholar Citations

A real problem can arise when one of the profiles has been created by someone other than the author the profile is about. The author may send a request to Google Scholar to delete the profile, but this kind of requests might take a while to be processed, generating a feeling of helplessness in the author.

b) Variety of document types (including non-academic documents)

One of the main criticisms to the profiles in *Google Scholar Citations* (when considering whether they're suited for evaluation purposes) is the inclusion of a wide variety of document types: from peer-reviewed articles to posters. An author can add any kind of work to his profile, and sometimes they aren't even academic works: teaching materials, software, online resources, etc. (Figure 18).

While this is a true shortcoming from the research evaluation perspective, these profiles are designed to showcase any material that the author considers appropriate, especially if these materials could potentially generate some kind of impact through citations. The possibility to select the document typology (as *ResearchGate* does) may help solve this problem. However, the selection of document type is only an internal mechanism not reflected in the public profile.



1. MONOGRÁFICO: RIRLIOMETRÍA
ES Casado, CM Moreno, G Portal, ES Quesada

Nº CRÉDITOS/ECTS: 3
ES CASADO, CQUEELED ADQUIRIR

What does publication output tell us: indicators from the institutional databases of higher education institutions
E Reale, D De Filippo, I Gomez, B Lepori, B Poti, E Primeri, C Probst, ...

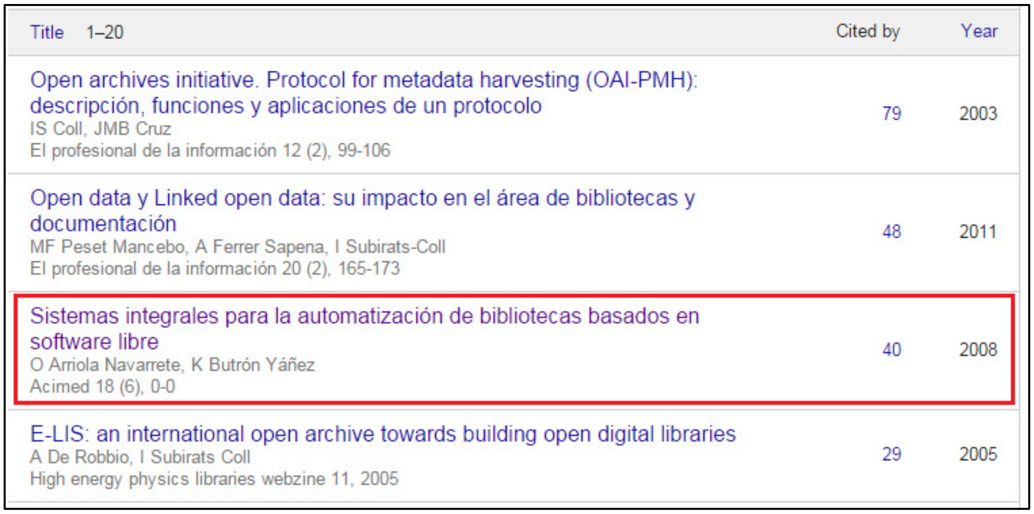
Diagnóstico sobre la docencia e investigación en Ciencias de la Documentación en España
ES Casado, MLL Sánchez

CUATRIMESTRE: 2º CRÉDITOS: 3
CYP DE TEORÍA, ES Casado, P DE PRÁCTICAS, G Ballenato

Figure 18. Teaching materials in Google Scholar Citations

c) Inclusion of misattributed documents in the profile

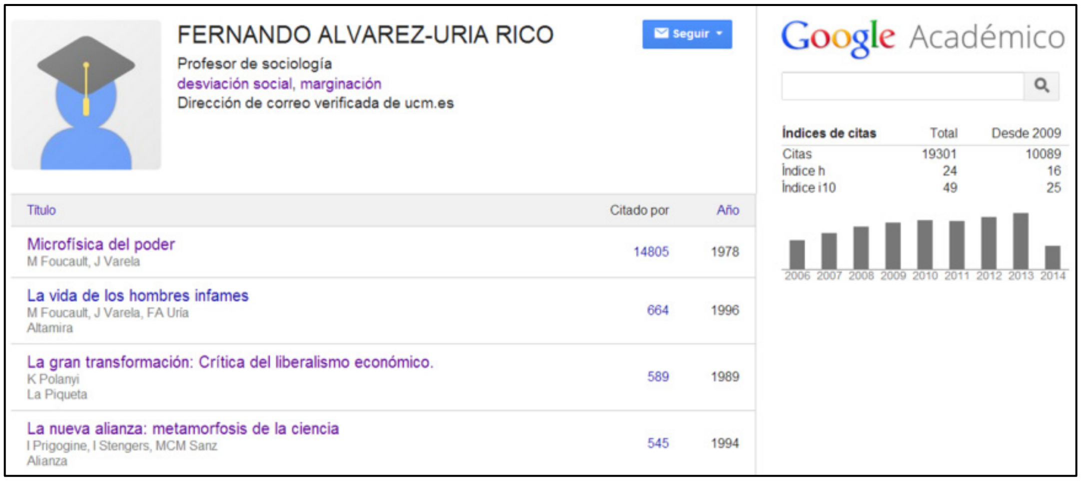
The Google Scholar team doesn't oversee the validity of all the information available in Google Scholar Citations. Therefore, it is the sole responsibility of the author that the information visible in his/her profile is accurate. Profiles can be set to be updated automatically (when the system finds an article that it's reasonably sure it's yours, it is automatically added to your profile), or by asking the author for confirmation first when the system thinks an addition or a change should be made. If the user selects the automatic updates, there is a risk that the system will add documents to the profile that the author hasn't actually written, thus falsely increasing the author's bibliometric indicators. The author will probably be completely oblivious to this issue if he or she doesn't check the profile regularly. If that is the case, it shouldn't be considered an active attempt to fake one's bibliometric indicators, but it is still a matter that should be fixed as soon as it comes to the author's knowledge. In Figure 19 we can see an example: the third document (marked in red), which has received 40 citations, hasn't been written by the owner of the profile (Imma Subirats-Coll).



Title 1–20	Cited by	Year
Open archives initiative. Protocol for metadata harvesting (OAI-PMH): descripción, funciones y aplicaciones de un protocolo IS Coll, JMB Cruz El profesional de la información 12 (2), 99-106	79	2003
Open data y Linked open data: su impacto en el área de bibliotecas y documentación MF Peset Mancebo, A Ferrer Sapena, I Subirats-Coll El profesional de la información 20 (2), 165-173	48	2011
Sistemas integrales para la automatización de bibliotecas basados en software libre O Arriola Navarrete, K Butrón Yáñez Acimed 18 (6), 0-0	40	2008
E-LIS: an international open archive towards building open digital libraries A De Robbio, I Subirats Coll High energy physics libraries webzine 11, 2005	29	2005

Figure 19. Misattributed documents in Google Scholar Citations

We can find examples where the owner of the profile has participated as a translator or editor of a work (Figure 20). The assignation of the citation counts of a work to the people who have fulfilled this kind of roles is controversial. At the very least, they should make sure that their role is clearly stated and visible in the profile.



FERNANDO ALVAREZ-URIA RICO
Profesor de sociología
desviación social, marginación
Dirección de correo verificada de ucm.es

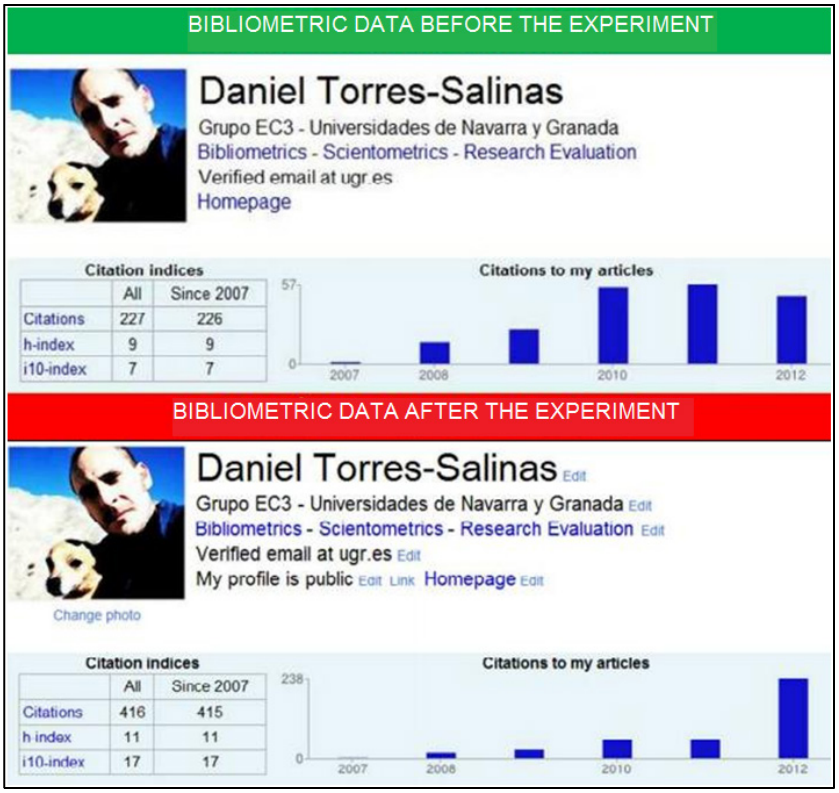
Índices de citas	Total	Desde 2009
Citas	19301	10089
Índice h	24	16
Índice i10	49	25

Título	Citado por	Año
Microfísica del poder M Foucault, J Varela	14805	1978
La vida de los hombres infames M Foucault, J Varela, FA Uría Altamira	664	1996
La gran transformación: Crítica del liberalismo económico. K Polanyi La Piqueta	589	1989
La nueva alianza: metamorfosis de la ciencia I Prigogine, I Stengers, MCM Sanz Alianza	545	1994

Figure 20. Edition and translation roles in Google Scholar Citations

d) Deliberate manipulation of documents and citations in Google Scholar

Another issue is that of the conscious manipulation of profiles by their owners. The fact that anyone, without advanced technical skills, can manipulate his/her own bibliometric indicators, or other people's (Delgado López-Cózar, Robinson-García & Torres-Salinas, 2014) may affect the credibility of GSC academic profiles if no action to control this issue is taken by the *Google Scholar* team. In Figure 21 we observe how uploading a set of fake documents to a repository (with nonsensical text, and a list of references which include the set of documents whose impact you want to boost) will, in just a few days, cause the desired adulteration of citation scores in the profiles of the authors of the referenced documents.



BIBLIOMETRIC DATA BEFORE THE EXPERIMENT

Daniel Torres-Salinas

Grupo EC3 - Universidades de Navarra y Granada
Bibliometrics - Scientometrics - Research Evaluation
Verified email at ugr.es
Homepage

Citation indices

	All	Since 2007
Citations	227	226
h-index	9	9
i10-index	7	7

BIBLIOMETRIC DATA AFTER THE EXPERIMENT

Daniel Torres-Salinas Edit

Grupo EC3 - Universidades de Navarra y Granada Edit
Bibliometrics - Scientometrics - Research Evaluation Edit
Verified email at ugr.es Edit
My profile is public Edit Link Homepage Edit

Change photo

Citation indices

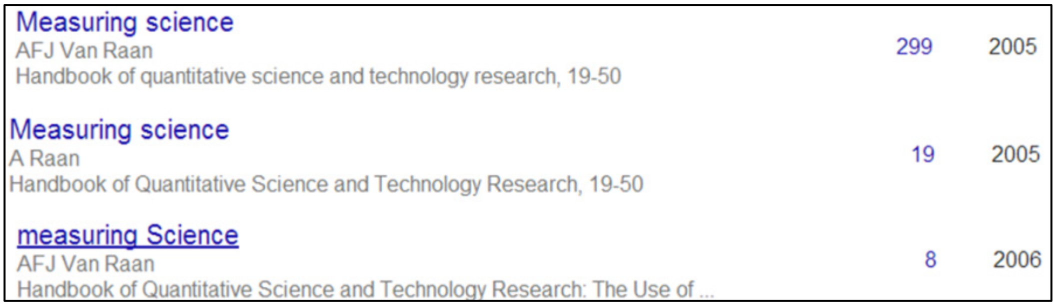
	All	Since 2007
Citations	416	415
h-index	11	11
i10-index	17	17

Figure 21. Effect of data manipulation in Google Scholar Citations

Source: Delgado López-Cózar, Robinson-García & Torres-Salinas, 2014

e) Duplicate documents in profiles

This is also a side effect of the cases when Google Scholar fails to group together different versions of the same document. The consequence for the profiles is that the different versions will also be added as different records in the profile, which might affect (positively or negatively) indicators like the h-index and the i-index, which are computed automatically. Fortunately, profile users can manually merge records in their profile, which will solve this issue (Figure 22). This merge only affects the author's profile. It doesn't alter Google Scholar search query results in any way, that is, there will still be two (or more) records for that document in Google Scholar's index, at least until the error gets fixed in a future update.

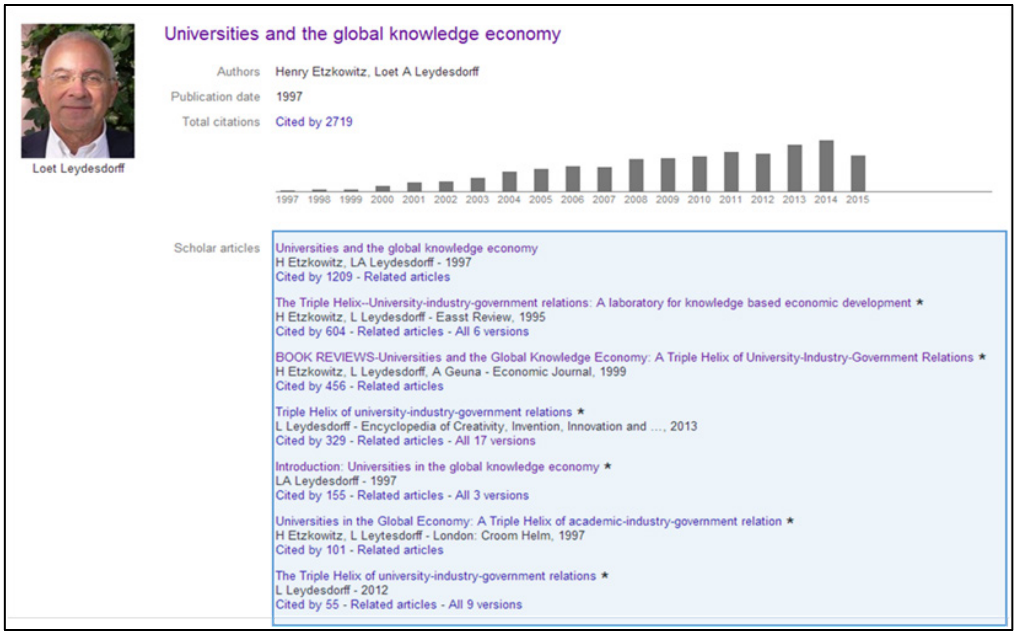


Measuring science AFJ Van Raan Handbook of quantitative science and technology research, 19-50	299	2005
Measuring science A Raan Handbook of Quantitative Science and Technology Research, 19-50	19	2005
measuring Science AFJ Van Raan Handbook of Quantitative Science and Technology Research: The Use of ...	8	2006

Figure 22. Versions not tied in Google Scholar Citations

f) Incorrectly merged documents

The downside to the fact that an author can freely merge documents in his/her profile is, obviously, that incorrect merges (of different documents) can also be made. As we discussed before, Google Scholar doesn't run any validity or accuracy checks on the information displayed in these profiles. Of course, this can also have a distorting effect on the automatically generated author-level indicators.



Universities and the global knowledge economy

Authors Henry Etzkowitz, Loet A Leydesdorff
Publication date 1997
Total citations Cited by 2719

Scholar articles

Universities and the global knowledge economy
H Etzkowitz, LA Leydesdorff - 1997
Cited by 1209 - Related articles

The Triple Helix--University-industry-government relations: A laboratory for knowledge based economic development ★
H Etzkowitz, L Leydesdorff - Easst Review, 1995
Cited by 604 - Related articles - All 6 versions

BOOK REVIEWS-Universities and the Global Knowledge Economy: A Triple Helix of University-Industry-Government Relations ★
H Etzkowitz, L Leydesdorff, A Geuna - Economic Journal, 1999
Cited by 456 - Related articles

Triple Helix of university-industry-government relations ★
L Leydesdorff - Encyclopedia of Creativity, Invention, Innovation and ..., 2013
Cited by 329 - Related articles - All 17 versions

Introduction: Universities in the global knowledge economy ★
LA Leydesdorff - 1997
Cited by 155 - Related articles - All 3 versions

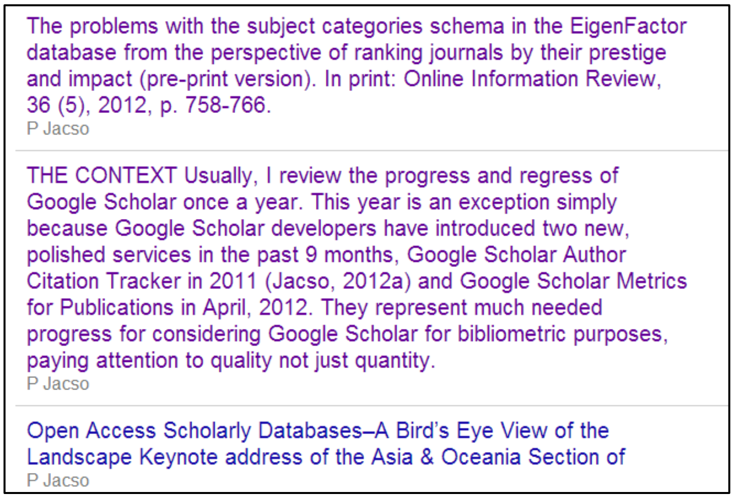
Universities in the Global Economy: A Triple Helix of academic-industry-government relation ★
H Etzkowitz, L Leytesdorff - London: Croom Helm, 1997
Cited by 101 - Related articles

The Triple Helix of university-industry-government relations ★
L Leydesdorff - 2012
Cited by 55 - Related articles - All 9 versions

Figure 23. Incorrectly merged records in Google Scholar Citations

g) Unclean document titles

This error is also inherited from Google Scholar's metadata parsing errors. Google Scholar Citations allows authors to modify almost all aspects of a record in their profile, including the title of the documents. Unfortunately, not all authors pay attention to such details, and so these errors persist (Figure 24).



The problems with the subject categories schema in the EigenFactor database from the perspective of ranking journals by their prestige and impact (pre-print version). In print: Online Information Review, 36 (5), 2012, p. 758-766.
P Jacso

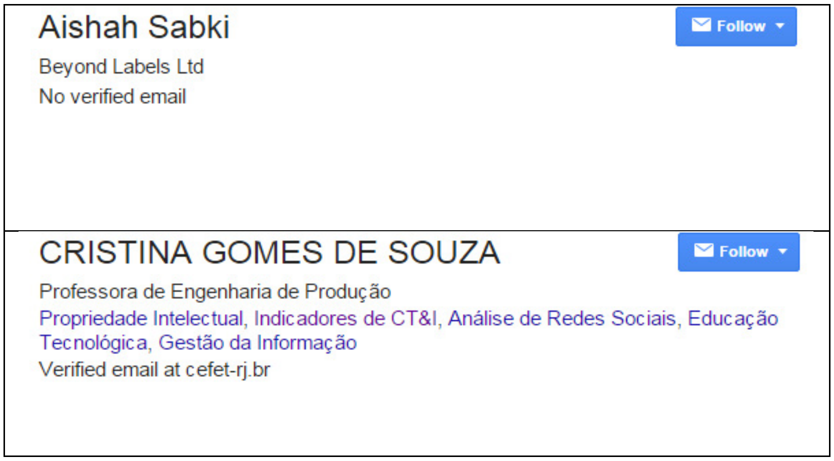
THE CONTEXT Usually, I review the progress and regress of Google Scholar once a year. This year is an exception simply because Google Scholar developers have introduced two new, polished services in the past 9 months, Google Scholar Author Citation Tracker in 2011 (Jacso, 2012a) and Google Scholar Metrics for Publications in April, 2012. They represent much needed progress for considering Google Scholar for bibliometric purposes, paying attention to quality not just quantity.
P Jacso

Open Access Scholarly Databases–A Bird's Eye View of the Landscape Keynote address of the Asia & Oceania Section of
P Jacso

Figure 24. Parse errors in identifying document titles in Google Scholar Citations

h) Missing or uncommon areas of interest

One last limitation that may affect the results of this Working Paper is related to the areas of interest declared by the authors in their profiles (a maximum of five areas can be provided). Researchers in bibliometrics with a public profile in Google Scholar Citations, but haven't declared any area of interest (Figure 24, top), those who use uncommon keywords, or keywords in a language other than English (Figure 25, bottom) may have been overlooked.


Aishah Sabki
Beyond Labels Ltd
No verified email

CRISTINA GOMES DE SOUZA
Professora de Engenharia de Produção
Propriedade Intelectual, Indicadores de CT&I, Análise de Redes Sociais, Educação Tecnológica, Gestão da Informação
Verified email at cefet-rj.br

Figure 24. Missing (top) and uncommon (bottom) areas of interest in Google Scholar Citations

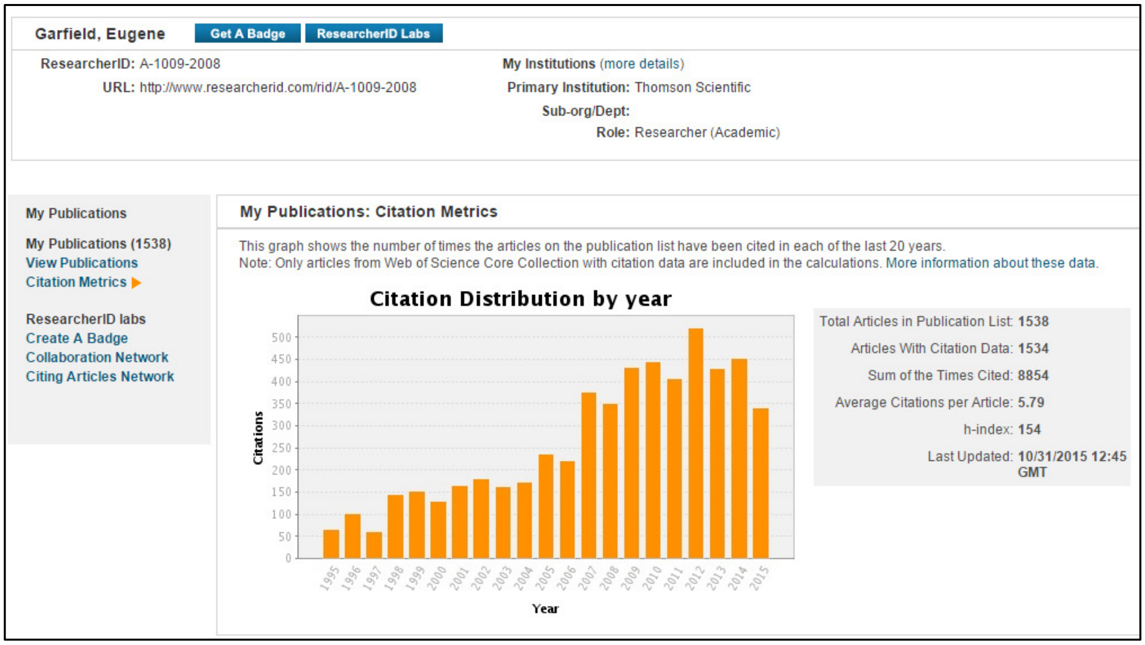
3.4.2. ResearcherID

One of the main shortcomings that characterize *ResearcherID* is the need to manually update the profiles. An author needs to synchronize his/her account with a search in *Web of Science Core Collection* in order to update the list of publications, unlike in Google Scholar and ResearchGate, where the process is largely carried out by the system, and authors only need to confirm new additions or modifications when the system prompts them to do so.

The fact that active manual intervention is needed on the author's part to keep the profile up to date results in a very inconsistent set of data. Authors concerned with online visibility will regularly update their profile, but in the majority of cases, authors will rarely visit their profile again after setting it up the first time. This may explain the results previously shown in Figure 5.

Moreover, we have found additional shortcomings in the system, caused by incorrectly attributed citations in Web of Science, which affect *ResearcherID* profiles.

Let's illustrate this issue with an example in which Dr. Eugene Garfield will be our test subject. In figure 25 we can see the citation metrics for Eugene Garfield's academic profile according to *ResearcherID*, which displays the number of articles published, the sum of times cited, the h-index, and other bibliometric indicators based on data from Web of Science Core Collection. Since Dr. Garfield hasn't created a Google Scholar Citations profile for himself, we generated a private profile in GSC (only accessible by us) in order to compare the indicators provided by the two profile platforms. A screenshot of this profile can be seen in Figure 26.



Garfield, Eugene

ResearcherID: A-1009-2008
URL: http://www.researcherid.com/rid/A-1009-2008

My Institutions (more details)
Primary Institution: Thomson Scientific
Sub-org/Dept:
Role: Researcher (Academic)

My Publications
My Publications (1538)
View Publications
Citation Metrics

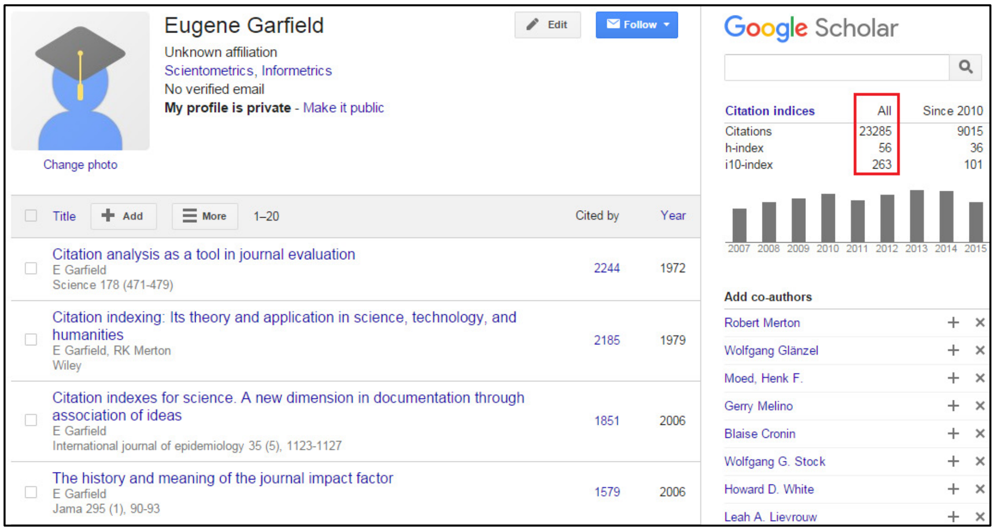
ResearcherID labs
Create A Badge
Collaboration Network
Citing Articles Network

My Publications: Citation Metrics

This graph shows the number of times the articles on the publication list have been cited in each of the last 20 years.
Note: Only articles from Web of Science Core Collection with citation data are included in the calculations. More information about these data.

Total Articles in Publication List:	1538
Articles With Citation Data:	1534
Sum of the Times Cited:	8854
Average Citations per Article:	5.79
h-index:	154
Last Updated:	10/31/2015 12:45 GMT

Figure 25. Eugene Garfield's academic profile in ResearcherID



Eugene Garfield
Unknown affiliation
Scientometrics, Informetrics
No verified email
My profile is private - Make it public

1–20

Title	Cited by	Year
Citation analysis as a tool in journal evaluation E Garfield Science 178 (471-479)	2244	1972
Citation indexing: Its theory and application in science, technology, and humanities E Garfield, RK Merton Wiley	2185	1979
Citation indexes for science. A new dimension in documentation through association of ideas E Garfield International journal of epidemiology 35 (5), 1123-1127	1851	2006
The history and meaning of the journal impact factor E Garfield Jama 295 (1), 90-93	1579	2006

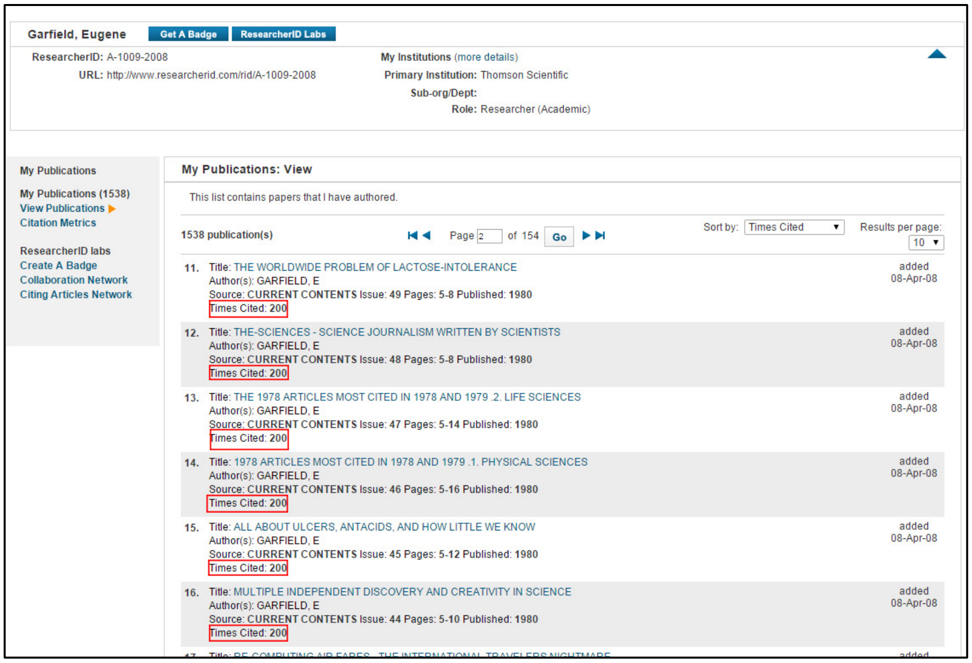
Citation indices	All	Since 2010
Citations	23285	9015
h-index	56	36
i10-index	263	101

Add co-authors

- Robert Merton
- Wolfgang Glänzel
- Moed, Henk F.
- Gerry Melino
- Blaise Cronin
- Wolfgang G. Stock
- Howard D. White
- Leah A. Lievrouw

Figure 26. Eugene Garfield's academic profile in Google Scholar Citations

As we can see, there is a huge difference between Dr. Eugene Garfield's h-index according to *ResearcherID* (154) and his h-index according to *Google Scholar* (55). This is caused by a technical error in the data provided by Web of Science. Dr. Garfield's *ResearcherID* profile contains a great number of works published in *Current Contents*, many of them with exactly 200 citations (Figure 27), an odd phenomenon. There is another large group of documents with exactly 155 citations, and other groups of documents which also share the same number of citations.



Garfield, Eugene

ResearcherID: A-1009-2008
URL: http://www.researcherid.com/rid/A-1009-2008

My Institutions (more details)
Primary Institution: Thomson Scientific
Sub-org/Dept:
Role: Researcher (Academic)

My Publications: View

This list contains papers that I have authored.

1538 publication(s) Page 2 of 154 Go

Sort by: Times Cited Results per page: 10

11. Title: THE WORLDWIDE PROBLEM OF LACTOSE-INTOLERANCE
Author(s): GARFIELD, E
Source: CURRENT CONTENTS Issue: 49 Pages: 5-8 Published: 1980
Times Cited: 200
added 08-Apr-08

12. Title: THE-SCIENCES - SCIENCE JOURNALISM WRITTEN BY SCIENTISTS
Author(s): GARFIELD, E
Source: CURRENT CONTENTS Issue: 48 Pages: 5-8 Published: 1980
Times Cited: 200
added 08-Apr-08

13. Title: THE 1978 ARTICLES MOST CITED IN 1978 AND 1979 .2. LIFE SCIENCES
Author(s): GARFIELD, E
Source: CURRENT CONTENTS Issue: 47 Pages: 5-14 Published: 1980
Times Cited: 200
added 08-Apr-08

14. Title: 1978 ARTICLES MOST CITED IN 1978 AND 1979 .1. PHYSICAL SCIENCES
Author(s): GARFIELD, E
Source: CURRENT CONTENTS Issue: 46 Pages: 5-16 Published: 1980
Times Cited: 200
added 08-Apr-08

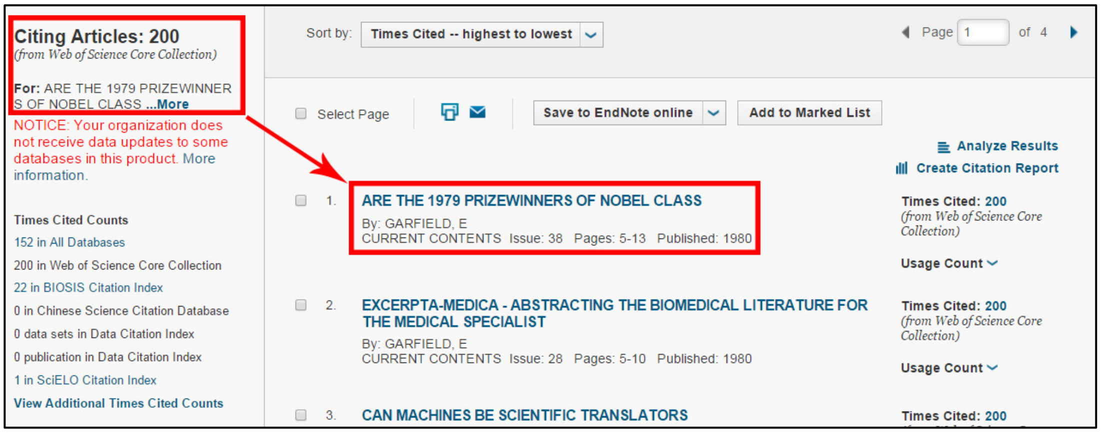
15. Title: ALL ABOUT ULCERS, ANTACIDS, AND HOW LITTLE WE KNOW
Author(s): GARFIELD, E
Source: CURRENT CONTENTS Issue: 45 Pages: 5-12 Published: 1980
Times Cited: 200
added 08-Apr-08

16. Title: MULTIPLE INDEPENDENT DISCOVERY AND CREATIVITY IN SCIENCE
Author(s): GARFIELD, E
Source: CURRENT CONTENTS Issue: 44 Pages: 5-10 Published: 1980
Times Cited: 200
added 08-Apr-08

17. Title: RE-COMPUTING AIR FARES - THE INTERNATIONAL TRAVELERS NIGHTMARE
added

Figure 27. Eugene Garfield's publication view in ResearcherID

The examination of any of these documents on the *Web of Science* database reveals that all these citations have been incorrectly attributed. In fact, there are some cases where, according to *Web of Science*, a document cites itself (Figure 28). The cause for this error is yet unknown to us and further research is needed to ascertain how often this kind of error occurs throughout the *Web of Science* database.



Citing Articles: 200
(from Web of Science Core Collection)

For: ARE THE 1979 PRIZEWINNERS OF NOBEL CLASS ...More

NOTICE: Your organization does not receive data updates to some databases in this product. More information.

Times Cited Counts
152 in All Databases
200 in Web of Science Core Collection
22 in BIOSIS Citation Index
0 in Chinese Science Citation Database
0 data sets in Data Citation Index
0 publication in Data Citation Index
1 in SciELO Citation Index
View Additional Times Cited Counts

Sort by: Times Cited -- highest to lowest Page 1 of 4

□ 1. ARE THE 1979 PRIZEWINNERS OF NOBEL CLASS
By: GARFIELD, E
CURRENT CONTENTS Issue: 38 Pages: 5-13 Published: 1980
Times Cited: 200 (from Web of Science Core Collection)
Usage Count

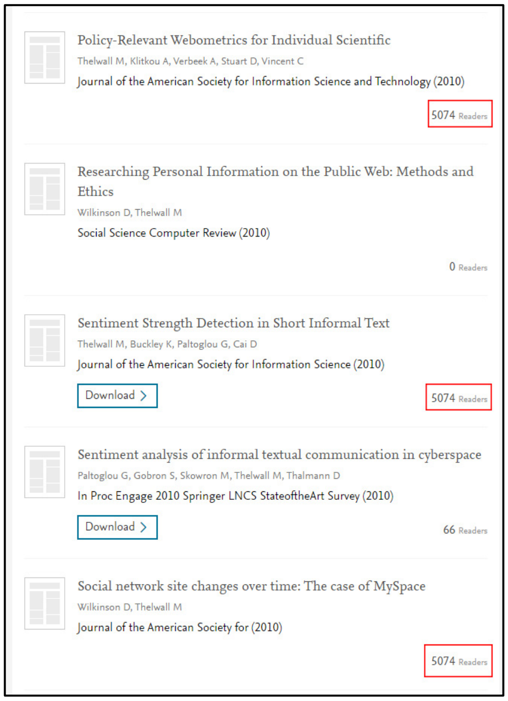
□ 2. EXCERPTA-MEDICA - ABSTRACTING THE BIOMEDICAL LITERATURE FOR THE MEDICAL SPECIALIST
By: GARFIELD, E
CURRENT CONTENTS Issue: 28 Pages: 5-10 Published: 1980
Times Cited: 200 (from Web of Science Core Collection)
Usage Count

□ 3. CAN MACHINES BE SCIENTIFIC TRANSLATORS
Times Cited: 200

Figure 28. Eugene Garfield's citing articles in Web of Science

3.4.3. Mendeley

An unusual phenomenon was detected while perusing some bibliometricians' profiles in *Mendeley*: many papers published in the *Journal of the American Society for Information Science and Technology* had abnormally high reader counts (number of *Mendeley* users who have saved a certain paper to their collection of references). On November 6th, 2015, a group of *JASIST* articles all exhibited exactly 5,074 readers. Figure 29, a snapshot taken from Mike Thelwall's Mendeley profile, illustrates this phenomenon.



Policy-Relevant Webometrics for Individual Scientific
Thelwall M, Klitkou A, Verbeek A, Stuart D, Vincent C
Journal of the American Society for Information Science and Technology (2010)
5074 Readers

Researching Personal Information on the Public Web: Methods and Ethics
Wilkinson D, Thelwall M
Social Science Computer Review (2010)
0 Readers

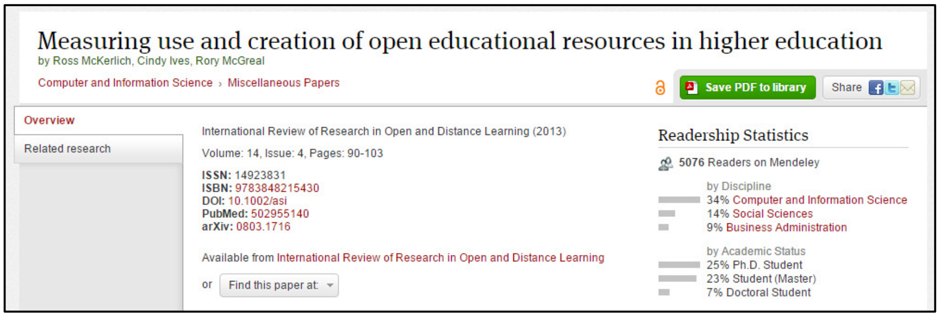
Sentiment Strength Detection in Short Informal Text
Thelwall M, Buckley K, Paltoglou G, Cai D
Journal of the American Society for Information Science (2010)
Download
5074 Readers

Sentiment analysis of informal textual communication in cyberspace
Paltoglou G, Gobron S, Skowron M, Thelwall M, Thalmann D
In Proc Engage 2010 Springer LNCS StateoftheArt Survey (2010)
Download
66 Readers

Social network site changes over time: The case of MySpace
Wilkinson D, Thelwall M
Journal of the American Society for (2010)
5074 Readers

Figure 29. Mike Thelwall's publications with incorrect reader counts in Mendeley

The immediate cause of this issue seems to be that all of these articles had been incorrectly linked to the same paper (Figure 30), which had precisely 5,074 readers. This paper - which doesn't have anything to do with the *JASIST* articles shown previously - could be accessed by clicking on any of the titles of the *JASIST* papers from their authors' profiles. The technical reason why this could've happened is yet unknown.



Measuring use and creation of open educational resources in higher education
by Ross McKerlich, Cindy Ives, Rory McGreal
Computer and Information Science › Miscellaneous Papers

International Review of Research in Open and Distance Learning (2013)
Volume: 14, Issue: 4, Pages: 90-103
ISSN: 14923831
ISBN: 9783848215430
DOI: 10.1002/asi
PubMed: 502955140
arXiv: 0803.1716

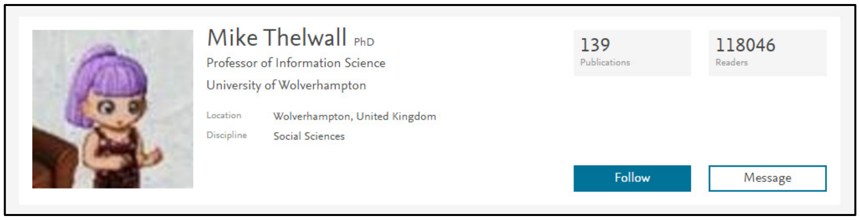
Available from International Review of Research in Open and Distance Learning

or Find this paper at:

Readership Statistics
5076 Readers on Mendeley
by Discipline
34% Computer and Information Science
14% Social Sciences
9% Business Administration
by Academic Status
25% Ph.D. Student
23% Student (Master)
7% Doctoral Student

Figure 30. Publication causing readership metrics misleading in Mendeley

The fact is that this phenomenon has affected several researchers in our study, greatly distorting their aggregate reader counts. The most noticeable case is that of Dr. Mike Thelwall, who has 23 articles affected by this issue in his personal profile, rising his aggregate reader count to 118,046 readers on November 6th (Figure 31), much higher than the count we collected on September (7,423). The error hasn't been fixed yet, and this count keeps growing every day (144,319 by January the 14th, 2016).



Mike Thelwall PhD

Professor of Information Science

University of Wolverhampton

Location Wolverhampton, United Kingdom

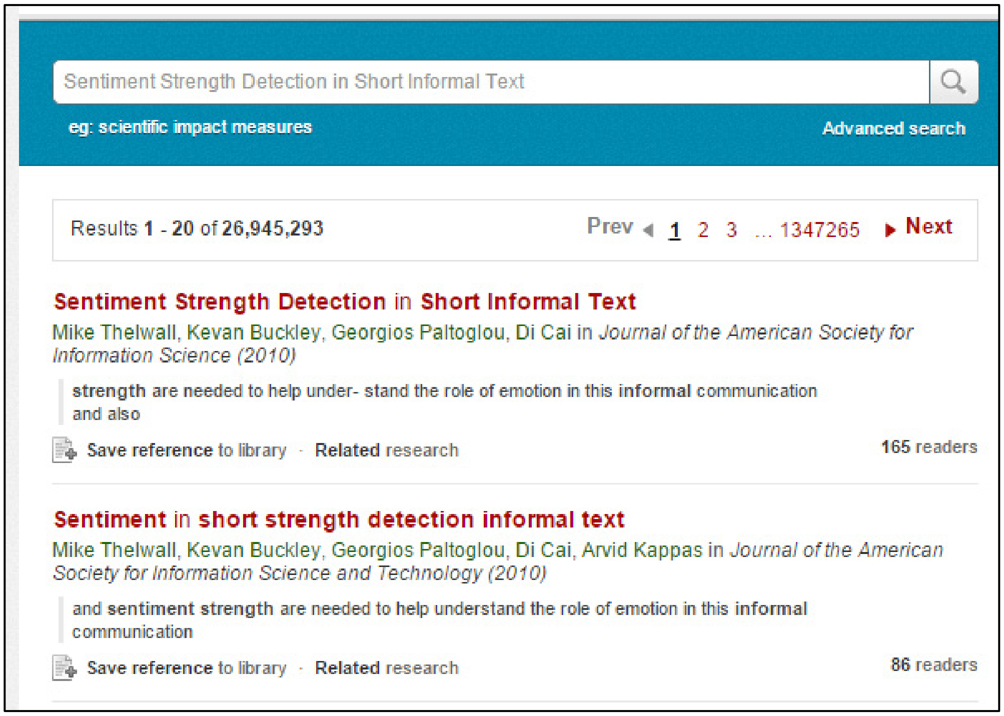
Discipline Social Sciences

139 Publications

118046 Readers

Figure 31. Mike Thelwall's personal profile metrics in Mendeley (6th November 2015)

Lastly, it is important to note that if you search any of these documents directly on *Mendeley*'s search feature, the results show the correct (or at least more plausible) reader count for the articles (Figure 32).



Sentiment Strength Detection in Short Informal Text

Results **1 - 20** of **26,945,293**

Sentiment Strength Detection in **Short Informal Text**

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai in *Journal of the American Society for Information Science (2010)*

strength are needed to help under- stand the role of emotion in this **informal** communication and also

Save reference to library · Related research — 165 readers

Sentiment in **short strength detection informal text**

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, Arvid Kappas in *Journal of the American Society for Information Science and Technology (2010)*

and **sentiment strength** are needed to help understand the role of emotion in this **informal** communication

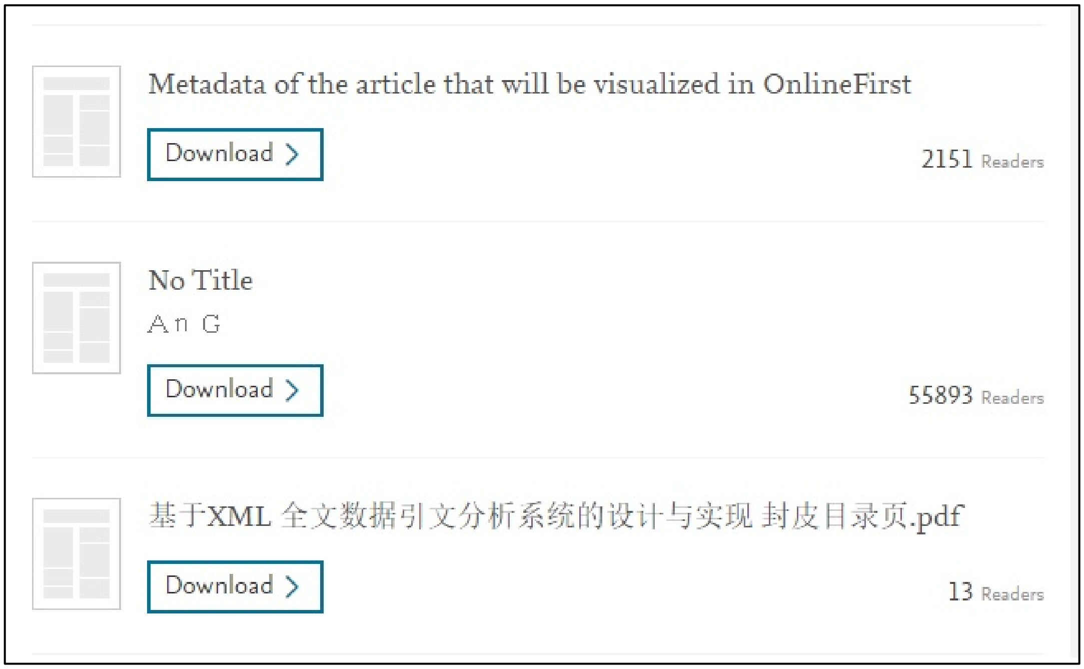
Save reference to library · Related research — 86 readers

Figure 32. Direct search of documents in Mendeley

Apart from these anomalous readership metrics in *Mendeley* (that should be understood as an anecdotal mistake that Mendeley will fix soon), we have found other malfunctions caused by errors in the metadata of the references added to the platform, which also affect readership metrics.

In Figure 32 we can see how one author (Arvid Kappas) is missing from one of the two versions of the article "Sentiment in short strength detection informal text". Probably for this reason, *Mendeley* didn't consider them to be the same document, and thus, at some point it created a second record for the document instead of merging it with the version it already had. This, in turn, meant that the reading counts would be split between the two versions of the document (a similar scattering effect to the one found *Google Scholar Citations* with versions and citations, as we previously described).

Not only incorrect metadata can lead to erroneous reader counts, missing metadata can also be dangerous. In Figure 33, taken from Zhigang Hu's *Mendeley* profile on November the 6th, 2015, there are examples of both incorrect metadata (the first article) and missing metadata (the second article) leading to inaccurate reader counts.



Metadata of the article that will be visualized in OnlineFirst

2151 Readers

No Title

A n G

55893 Readers

基于XML 全文数据引文分析系统的设计与实现 封皮目录页.pdf

13 Readers

Figure 33. Documents with incorrect or missing metadata affecting Mendeley reader counts

In the first case, the title of one of this researcher's articles wasn't correctly parsed from the PDF of the article, and an incorrect string was selected as the title instead. This is a relatively common issue, so all the articles which have been incorrectly parsed in a similar way and share the same incorrect title "Metadata of the article that will be visualized in *OnlineFirst*" have been lumped together by *Mendeley*, which explains the high reader count for that article. The same explanation could probably be applied to the second document. All documents with a missing title or with the incorrect title "No Title" must have been merged by *Mendeley* to obtain such a high reader count (55,893).

3.4.4. ResearchGate

ResearchGate (RG), the academic profiling and sharing platform created by Dr. Ijad Madisch[61] and Dr. Sören Hofmayer[62] in 2008, is currently gaining momentum as one most used services of this kind among researchers. In May 2015 they announced they had reached 7 million users,[63] and just five months later, in October, they claimed to have reached 8 million.[64]

The reasons behind the success of this platform are undoubtedly related to the constant stream of new (and usually very convenient) features the platform has been introducing during the past months, but probably also to the constant flow of ego-boosting e-mails that users receive informing them about the great impact their work is having on the scientific community.

Like the rest of platforms fulfilling similar needs, RG computes a set of indicators which are designed to measure the popularity, impact, and degree of use of the documents a researcher uploads to the system (Thelwall and Kousha, 2015). In section 3.3 we observed how these metrics (especially the RG Score) achieved a high correlation with impact metrics provided by *Google Scholar Citations* (especially total citations and h-index). Moreover, this platform was, at the moment we collected the data, the only one who provided both citation and usage metrics for articles (until the Web of Science began to offer usage metrics

[61] https://www.researchgate.net/profile/Ijad_Madisch
[62] https://www.researchgate.net/profile/Soeren_Hofmayer
[63] https://www.researchgate.net/blog/post/celebrating-seven-million-members-and-seven-years-of-researchgate
[64] https://www.researchgate.net/blog/post/8-out-of-8-million

in November 2015). All these impressive results are partly a consequence of this momentum in terms of user growth.

However, we must point out some important shortcomings related to the lack of transparency in the way all these metrics are computed, a lack of transparency that makes them currently unsuitable for scientific evaluation. It looks like *ResearchGate* is acting like a modern "alchemist", in the sense that it produces its own "concoctions", but without revealing their ingredients and method of preparation to anyone, an issue that, of course, has not gone unnoticed by the scientific community.[65]

First, we may consider the RG Score, which is the indicator they display more prominently in the researchers' profiles, situated right next to the name of the researcher. According to *ResearchGate*[66], this author-level indicator measures "scientific reputation based on how all of your research is received by your peers". The main concern with this indicator - in terms of usefulness for scientific evaluation - is that the way it's calculated hasn't been made public. Therefore, even though this indicator may be a good way to attract researchers who enjoy going on ego trips once in a while, the fact that only *ResearchGate* knows how to calculate it renders it ill-suited for research assessment before the discussion about its intrinsic merits and defects can even begin.

Another matter is that, at the end of September 2015, that is, a few weeks after we collected our data about bibliometric researchers (results offered in sections 3.1 and 3.2), *ResearchGate* combined two of the indicators they used to display on its users' profiles (document views and downloads) into one (Reads).[67]

According to them, "a read is counted each time someone reads the summary or full-text, or downloads one of your publications from *ResearchGate*". However, the "document views" and "download counts" collected in September don't match the "read counts" available after that change (Table 10). We can easily see how "Reads" are clearly lower than the combination of downloads and views. The separation of summary views and document views may have something to do with this issue, and it's a matter that should be further analyzed.

[65] http://blogs.lse.ac.uk/impactofsocialsciences/2015/12/09/the-researchgate-score-a-good-example-of-a-bad-metric

[66] https://www.researchgate.net/publicprofile.RGScoreFAQ.html

[67] https://www.researchgate.net/blog/post/introducing-reads

Table 10. Top 10 authors with the highest Reads counts on ResearchGate (9th of November, 2015), compared to their Downloads and Views counts on the 10th of September, 2015.

AUTHOR NAME	**SEPTEMBER 10th (2015)**		**NOVEMBER 9th (2015)**	**MISMATCH (%)**
	DOWNLOADS	**VIEWS**	**READS**	
Loet Leydesdorff	32,165	42,926	21,013	27.98
Mike Thelwall	24,989	34,376	17,748	29.90
Chaomei Chen	31,579	26,734	13,452	23.07
Nader Ale Ebrahim	31,853	23,144	10,282	18.70
Lutz Bornmann	13,556	22,987	9,863	26.99
Maite Barrios	14,234	7,600	9,439	43.23
Wolfgang Glänzel	10,572	20,145	9,439	30.73
Félix Moya Anegón	18,691	23,583	8,625	20.40
Cassidy Sugimoto	13,079	8,081	8,458	39.97
Ronald Rousseau	8,066	19,118	6,934	25.51

The same thing can be said about the "profile views" indicator: the counts obtained back in September are always higher than the ones available two months later on November the 9th (Table 11). To the best of our knowledge, there has not been an announcement regarding any changes in the profile views indicator.

Table 11. Top 10 authors with the highest profile view counts on ResearchGate (9th of November, 2015), compared to the same indicator on the 10th of September, 2015.

AUTHOR NAME	**SEPTEMBER 10th (2015)**	**NOVEMBER 9th (2015)**	**MISMATCH (%)**
	PROFILE VIEWS	**PROFILE VIEWS**	
Nader Ale Ebrahim	19,821	13,281	67.00
Chaomei Chen	7,760	3,937	50.73
Loet Leydesdorff	4,227	1,758	41.59
Bakthavachalam Elango	2,883	1,756	60.91
Zaida Chinchilla	5,840	1,569	26.87
Mike Thelwall	4,297	1,568	36.49
Lutz Bornmann	3,129	1,439	45.99
Wolfgang Glänzel	3,012	1,301	43.19
Kevin Boyack	3,256	1,135	34.86
Peter Ingwersen	2,335	1,025	43.90

In any case, a high Pearson correlation between the sum of *Downloads* and *Views*, and the new *Reads* indicator (r= 0.93, n = 499; α = 0.95; p-value < 2.2e-16) is observed; and also between the Profile View counts collected in September and the ones collected in November (r= 0.93; n = 535; α = 0.95; p-value < 2.2e-16).

3.4.5. General strengths and shortcomings of academic profiles

Lastly, Table 12 summarizes the main strengths and weaknesses of each of the platforms analyzed in this study.

Table 12. Advantages and disadvantages of academic profiles provided by social platforms

GOOGLE SCHOLAR CITATIONS	
ADVANTAGES	**DISADVANTAGES**
• Widest coverage (all languages, sources and disciplines) • User-friendly • High growth rate • Automatic updates • Alerts (new citations to your work, or publications from other authors)	• Scarce quality control • Open to manipulation • Inherits mistakes from Google Scholar
RESEARCHERID	
ADVANTAGES	**DISADVANTAGES**
• Offers advanced bibliometric indicators	• No automatic updates • Not very user-friendly • Inherits mistakes from WoS • Not used by many authors • Only WoS CC publications count towards citation metrics
RESEARCHGATE	
ADVANTAGES	**DISADVANTAGES**
• Increasingly used by the scientific community: very high growth rate • Offers usage data (views and downloads) • User-friendly • Correlates with citation data • Social functions to contact other authors	• No automatic updates (one co-author must upload the document) • Lack of transparency in its indicators • Still not used by many authors • Sends too many e-mails (by default)
MENDELEY PROFILES	
ADVANTAGES	**DISADVANTAGES**
• Increasingly used by community • Offers usage data (reads) • Correlates with citation data • Allows discipline analysis • Social functions (follow other authors)	• No automatic updates • Quality of metadata depends on user input

It is clear that none of the platforms considered and analyzed in this working paper is without its problems and limitations. At the same time, all of them offer new insights for measuring scientific impact.

Google Scholar offers the widest coverage, situated on approximately 160 million hits on May 2014 (Orduna-Malea et al, 2014). Its indexing criteria (all academic documents openly stored in the academic web space) makes this database the only place where every academic document is indexed regardless of its typology (not only journal articles but also books, book chapters, reports, thesis dissertations, conference proceedings, etc.), its language, or its discipline. Thanks to this wide variety of sources, *Google Scholar* is able to measure not only scientific but also educational and professional impact in the broadest sense of the term. At the same time, as regards strict scientific impact, there is a high correlation (r = 0.8) between the number of citations of these documents in GS and their citations in WoS (Martin-Martin et al, 2014).

Google Scholar Citations includes citation scores for authors, areas of interest, and institutional information. Additionally, in this platform, the owner of the profile can improve the bibliographic information provided by Google Scholar, and merge duplicates Google Scholar hasn't been able to detect. This impressive collection of data, together with the development of functionalities (such as detecting and merging duplicates), makes *Google Scholar* the best tool for the bibliometric analysis of some disciplines, especially those within the areas of the Humanities, Social Sciences, and Engineering.

Unfortunately, Google Scholar is not without its problems. The possibility to edit records in the profiles does not solve its parsing problems, for which there doesn't seem to be a clear explanation sometimes. We must point out however that the system is improving year by year. Moreover, in an academic big data environment, these errors (which we deem affect less than 10% of the records in the database) are of no great consequence, and do not affect the core system performance significantly.

On the other hand, the philosophy of the product (oriented to the user, lacking any bibliographic control) makes the tool rather open to confusing data, mistakes (described in section 3.4.1), and to manipulation, a really serious problem in the academia at the moment. Scientific misconduct should not be disregarded as mere spam.

Moreover, *Google Scholar* is user-friendly but not bibliometrician-friendly. Google Scholar's agreements with big publishers to collect data from their servers and present them in the search engine come at a price: among other things, the impossibility of offering an API which would no doubt be highly welcomed by the scientific community. An API would allow us to keep working on our understanding the production, dissemination, and consumption of scientific information worldwide.

ResearchGate is the second most-used platform among the tools analyzed in this work. The high number of users that this platform is currently attracting reinforces the validity of the metrics it provides (essentially because of the great amount of documents that have been already uploaded to the system). This is reflected in the extraordinary correlation that RG Score achieves with the h-index and total citations from *Google Scholar*. Moreover, there is no better platform to calculate number of downloads per document.

We believe this is a logic result, because the RG Score is basically made up of the number of publications an author has published, the citations to these publications, and the JCR Impact Factor of the journals where these articles are published. Usage indicators may also have some weight, but not much yet.

Nonetheless, the lack of transparency in the calculation of the different metrics (especially the RG Score) prevents it from being useful, since they cannot be replicated.

This the reason why the following questions still arise: what was *ResearchGate* really measuring before the changes in the View and Download indicators took place? What is it really measuring now? Why isn't *ResearchGate* more open about the way it computes the indicators they display?

Moreover, the introduction of subjective values (such as the participation in question & answers in the platform) may introduce some bias (high participation in the social platform does not have anything to do with academic impact, though it serves to incentive the use of the platform). In any case, the weight of this parameter doesn't seem to be significant.

Likewise, changes in the company policies, such as the elimination of some services (the complete list of documents ranked according to number of reads is no longer available), makes this platform unpredictable and unreliable at the moment. Other specific limitations are related to the quantity of documents indexed in the platform; references not properly identified, or incorrectly attributed citations.

Regarding *Mendeley*, we should acknowledge the validity of the Readers indicator, which strongly correlates to both the Downloads indicator provided by *ResearchGate* (different sides of usage) and to citation-based metrics from *Google Scholar*. However, we found some limitations in this platform while studying the Bibliometric community (which may be extrapolated to other academic communities).

First, calling the number of users that have saved a bibliographic record in their personal collection "readers" is absolutely incorrect (Delgado López-Cózar and Martín-Martín, 2015). The term should be changed to one that more accurately represents the nature of the indicator, because the current one can lead to misunderstandings and misinterpretations.[68]

Second, the fact that there are no automatic profile updates makes the system completely dependent on user activity. A total of 149 out of the 336 profiles analyzed (44.3%) didn't include a single document (Figure 34), and only 23% of the researchers have an effective presence in the platform. This fact strongly limits the use of *Mendeley* for the purpose of evaluating authors.

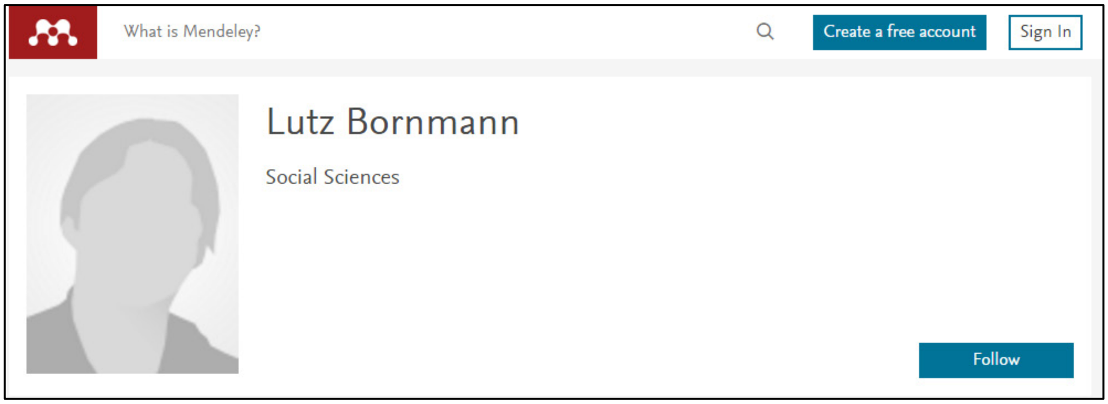
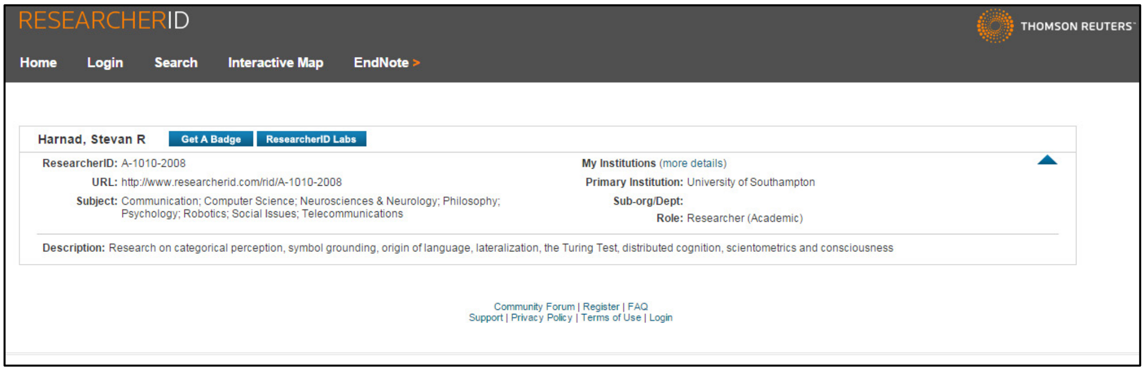


Figure 34. Example of empty academic profile in Mendeley

The last academic profiling service we analyzed was *ResearcherID*. There is no automatic profile updates in this platform, and a great percentage of user profiles (34.4%) have no public publications displayed (Figure 35), that is, the profile only contains basic information about the subject interests of the author and its affiliation. Only 26% of the authors in our sample had a ResearcherID profile with at least one document, and most of these profiles were out of date.

[68] The new "Reads" metric provided by *ResearchGate* suffers from the same problem, as it is combining online accesses to the document and downloads, which are not the same even though they claim they are.



Harnad, Stevan R

ResearcherID: A-1010-2008
URL: http://www.researcherid.com/rid/A-1010-2008
Subject: Communication; Computer Science; Neurosciences & Neurology; Philosophy; Psychology; Robotics; Social Issues; Telecommunications

My Institutions (more details)
Primary Institution: University of Southampton
Sub-org/Dept:
Role: Researcher (Academic)

Description: Research on categorical perception, symbol grounding, origin of language, lateralization, the Turing Test, distributed cognition, scientometrics and consciousness

Community Forum | Register | FAQ
Support | Privacy Policy | Terms of Use | Login

Figure 35. Example of an empty academic profile in ResearcherID

Apart from this lack of real use, we found several errors which had been inherited from the citation scores available in the Web of Science. That is, WoS is not error-free in the attribution of citation scores. For all these reasons, we do not consider *ResearcherID* a valuable tool for bibliometric purposes.

4. Conclusions

Although this work is focused on the analysis of a specific academic community (Bibliometrics), the results obtained allowed us to obtain a number of important findings, summarized below.

Firstly, *Google Scholar* (with its associated platform for academic profiles *Google Scholar Citations*), provides a very precise and accurate picture of the bibliometric community. The data collected, not only at the author-level but also at the document-level and source-level (journal and books), clearly responds to our mental image of the field. That is, *Google Scholar* helped identify the most influential authors (core and related) and sources (journals and publishers) in the discipline.

The level of use of other social platforms is quite far from the one found for *Google Scholar Citations*, not only in the number of user profiles created, but also in the regularity with which they are updated. *ResearchGate*'s growth rate is impressive and currently stands as the second most used profile platform by the bibliometric community. Its usage indicators (*Downloads* and *Views*) and its social network features (communication and information sharing among users) provide a perspective that *Google Scholar Citations* lacks.

The social tools analyzed here have a number of significant limitations, which clearly get in the way of generating academic mirrors complementary to those based merely on citations. In the case of *ResearchGate* these limitations are caused by the opacity of the indicators and unexpected changes in the policies of the company, whereas in *Mendeley* and *ResearcherID* the problems arise from the existence outdated profiles. This issue has a negative effect on the accuracy of the information provided by these platforms, as seen in Figure 5.

Twitter, on the other hand, presents a completely different picture. Its author-level indicators do not correlate with citation-based indicators (from *Google Scholar*) nor with usage indicators (provided by *ResearchGate* and *Mendeley*), but they do correlate with other network indicators (which measure an author's participation in the community as well as his/her ability to connect with other users). This lack of correlation should however not be understood negatively. Instead, we should interpret it as a sign that it these indicators measure a different dimension of the author's impact on the Web.

Two different kinds of indicators were found in these platforms: first, all metrics related to academic performance. This first group can further be divided into usage metrics (views and downloads) and citation metrics. Second, all metrics related to connectivity and popularity (followers). *ResearchGate* provides examples for these two sides of academic performance, since *Google Scholar Citations* profiles do not offer data about downloads or reads.

In the process of conducting this analysis, we identified a series of errors that allowed us to outline the main limitations of each product. May this serve as a sign that this study hasn't been made with an intention to exalt a particular database over the others. On the contrary, the intention was to thoroughly, comprehensively, conscientiously, and neutrally test the possibilities of *Google Scholar* as a tool for scientific evaluation.

In this sense, the empirical results indicate that *Google Scholar* should be the preferred source for relational and comparative analyses in which the emphasis is put on author clusters. Individual data should be taken with some caution as it may be subject to some errors. Despite these errors (as well as the lack of more advanced filtering features), *Google Scholar* has been able to measure the academic community dedicated to measuring; and has done it successfully: detecting "those who count" (bibliometricians).

Lastly, the results should be understood within the context of the bibliometric community. They may be different in other academic communities where the greater or lesser use of technologies can clearly influence the data. Furthermore, there is a certain positive bias in the use of these platforms because within the bibliometric community, these platforms are part of the object of study of the discipline, as is the case of this work.

References

Barjak, F., Li, X., & Thelwall, M. (2007). "Which factors explain the web impact of scientists' personal homepages?" *Journal of the American Society for Information Science and Technology, 58*(2), 200–211.

Becher, T., & Trowler, P. (2001). *Academic tribes and territories: Intellectual enquiry and the culture of disciplines*. McGraw-Hill Education (UK).

Bensman, S. J. (2007). "Garfield and the impact factor". *Annual Review of Information Science and Technology*, *41*(1), 93-155.

Bonitz, M. (1982). "Scientometrie, Bibliometrie, Informetrie". *Zentralblatt für Bibliothekswesen*, *96* (2):19–24.

Borgman, C. L. & Furner, J. (2002). "Scholarly Communication and Bibliometrics". *Annual Review of Information Science and Technology*, *36*, 3-72.

Braun, T. (1994). "Little scientometrics, big scientometrics… and beyond?" *Scientometrics*, *30*, 373–537.

Broadus, R.N. (1987a). "Early approaches to bibliometrics". *Journal of the American Society for Information Science*, *38*, 127–129.

Broadus, R. N. (1987b). "Toward a definition of 'bibliometrics'". *Scientometrics*, 12, 373–379.

Brookes, B. C. (1988). "Comments on the scope of bibliometrics". In: L. Egghe, R. Rousseau (Eds). *Informetrics 87/88. Select Proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval*. Amsterdam, Elsevier Science, 29–41.

Brookes, B. C. (1990). "Biblio-, Sciento-, Infor-metrics??? What are we talking about?". In: L. Egghe, R. Rousseau (Eds). *Informetrics 89/90. Selection of Papers Submitted for the Second International Conference on Bibliometrics, Scientometrics and Informetrics*. Amsterdam, Netherlands, Elsevier, 31–43.

Castells, M. (2002). *La galaxia internet*. Barcelona: Plaza & Janés.

Cronin, B. (2001). “Bibliometrics and beyond: some thoughts on web-based citation analysis”. *Journal of Information science*, *27*(1), 1-7.

De Bellis, N. (2009). *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Maryland: Scarecrow Press.

Delgado López-Cózar, E. & Martín-Martín, A. (2015). “Thomson Reuters coquetea con las altmetrics: usage counts para los artículos indizados en la Web of Science”. *EC3 Working Papers*, 20.

Delgado López-Cózar, E., Robinson-García, N. & Torres-Salinas, D. (2014). “The Google Scholar Experiment: how to index false papers and manipulate bibliometric indicators”. *Journal of the Association for Information Science and Technology*, *65*(3), 446-454.

Franceschini, F., Maisano, D. & Mastrogiacomo, L. (2015). “Research quality evaluation: comparing citation counts considering bibliometric database errors”. *Quality & Quantity*, *49*(1), 155-165.

García-Pérez, M. A. (2010). “Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in Psychology”. *Journal of the American Society for Information Science and Technology*, *61*(10), 2070-2085.

Garfield, E. (1983). “Idiosyncrasies and errors, or the terrible things journals do to us”. *Current Contents*, *2*, 5-11

Garfield, E. (1990). “Journal editors awaken to the impact of citation errors-how we control them at ISI”. *Current Contents*, *41*, 5-13

Glänzel, W. & Schoepflin, U. (1994). “Little scientometrics, big scientometrics … and beyond?” *Scientometrics*, *30*, 375–384.

Godin, B. (2006). “On the origins of bibliometrics”. *Scientometrics*, *68*(1), 109-133.

González-Díaz, C.; Iglesias-García, M.; Codina, L. (2015). “Presencia de las universidades españolas en las redes sociales digitales científicas: caso de los estudios de comunicación”. *El profesional de la información*, *24*(5), 640-647.

Gorbea Portal, S. (1994). “Principios teóricos y metodológicos de los estudios métricos de la información. *Investigación Bibliotecológica*, *8*, 23-32.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H. & Terliesner, J. (2014). “Coverage and adoption of altmetrics sources in the bibliometric community”. *Scientometrics*, *101*(2), 1145-1163.

Hertzel, D.H. (1987). “History of the development of ideas in bibliometrics”. In: A. Kent, (Ed.). *Encyclopedia of library and information sciences*, Vol. 42 (Supplement 7), Marcel Dekker, New York, 144–219

Hood, W. & Wilson, C. (2001). “The literature of bibliometrics, scientometrics, and informetrics”. *Scientometrics*, *52*(2), 291-314.

Jacsó, P. (2005). “Google Scholar: the pros and the cons”. *Online information review*, *29*(2), 208-214.

Jacso, P. (2006a). “Deflated, inflated and phantom citation counts”. *Online information review*, *30*(3), 297-309.

Jacsó, P. (2006b). “Dubious hit counts and cuckoo's eggs”. *Online Information Review*, *30*(2), 188-193.

Jacsó, P. (2008). “Google scholar revisited”. *Online information review*, *32*(1), 102-114.

Jacsó, P. (2010). “Metadata mega mess in Google Scholar”. *Online Information Review*, *34*(1), 175-191.

Jamali, H. R., Nicholas, D. & Herman, E. (2015). “Scholarly reputation in the digital age and the role of emerging platforms and mechanisms”. *Research Evaluation*, rvv032.

Kramer, Bianca; Bosman, Jeroen (2015): 101 Innovations in Scholarly Communication - the Changing Research Workflow. Available at: https://101innovations.wordpress.com/

Larivière, V. (2012). “The decade of metrics? Examining the evolution of metrics within and outside LIS”. *Bulletin of the American Society for Information Science and Technology*, *38*(6), 12-17.

Larivière, V., Sugimoto, C. & Cronin, B. (2012). “A bibliometric chronicling of library and information science’s first hundred years”. *Journal of the American Society for Information Science and Technology*, *63*(5), 997-1016.

Lawani, S. M. (1981), “Bibliometrics: its theoretical foundations, methods and applications”. *Libri*, *31*, 294–3

Martín-Martín, A., Orduña-Malea, E., Ayllón, J.M. & Delgado López-Cózar, E. (2014). “Does Google Scholar contain all highly cited documents (1950-2013)?”. *EC3 Working Papers*, 19.

Más-Bleda, A. & Aguillo, I. F. (2013). “Can a personal website be useful as an information source to assess individual scientists? The case of European highly cited researchers”. *Scientometrics*, *96*(1), 51-67.

Mas-Bleda, A., Thelwall, M., Kousha, K. & Aguillo, I. F. (2014). “Do highly cited researchers successfully use the social web?”. *Scientometrics*, *101*(1), 337-356.

McCain, K. W. (2010). “The view from Garfield’s shoulders: Tri-citation mapping of Eugene Garfield’s citation image over three successive decades”. *Annals of Library and Information Studies*, *57*, 261-270.

Mikki, S., Zygmuntowska, M., Gjesdal, Ø. L. & Al Ruwehy, H. A. (2015). “Digital Presence of Norwegian Scholars on Academic Network Sites—Where and Who Are They?”. *PloS one*, *10*(11), e0142709.

Moed, H. F. & Vriens, M. (1989). “Possible inaccuracies occurring in citation analysis”. *Journal of Information Science*, *15*(2), 95-107.

Narin, F. & Moll, J.K. (1977). "Bibliometrics"*. Annual Review of Information Science and Technology*, *12*, 35-58.

Nicolaisen, J. & Frandsen, T. F. (2015). "Bibliometric evolution: Is the journal of the association for information science and technology transforming into a specialty Journal?". *Journal of the Association for Information Science and Technology*, *66*(5), 1082-1085.

Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A. & López-Cózar, E. D. (2015). “Methods for estimating the size of Google Scholar”. *Scientometrics*,*104*(3), 931-949.

Peritz, B.C. (1984). “On the careers of terminologies; the case of bibliometrics”, Libri, *34*: 233–242

Poyer, R. K. (1979). “Inaccurate references in significant journals of science”. *Bulletin of the Medical Library Association*, *67*(4), 396.

Sengupta, I.N. (1992). “Bibliometrics, informetrics, scientometrics and librametrics: an overview”, *Libri*, *42*, 75–98.

Shapiro, Fred R. (1992). “Origins of Bibliometrics, Citation Indexing, and Citation Analysis: The Neglected Legal Literature”. *Journal of the American Society for Information Science*, *43*(5), 337–39.

Sher, I. H., Garfield, E., & Elias, A. W. (1966). “Control and Elimination of Errors in ISI Services”. *Journal of Chemical Documentation*, *6*(3), 132-135.

Thelwall, M. (2008). “Bibliometrics to webometrics”. *Journal of Information Science*, *34*(4), 605-621.

Thelwall, M., & Kousha, K. (2015). “ResearchGate: Disseminating, communicating, and measuring Scholarship?”. *Journal of the Association for Information Science and Technology*, *66*(5), 876-889.

Van Raan, A. (1997). “Scientometrics: State-of-the-art”. *Scientometrics*, *38*(1), 205-218.

Van-Noorden, R. (2014). “Online collaboration: Scientists and the social network”. *Nature news, 512*(7513), 126-129.

White, H. D. & McCain, K. W. (1998). “Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972–1995”. *Journal of the American Society for Information Science*, *49*(4), 327-355.

White, H.D. & McCain, K.W. (1989). “Bibliometrics”. *Annual review of information science and technology*, *24*, 119-186.

Whitley, R. (1984). *The intellectual and social organization of the sciences*. UK: Oxford University Press.

Wilson, C.S. (1999). “Informetrics”. *Annual Review of Information Science and Technology*, *34*, 107-247.

Chapter 11. Scholar Mirrors: Integrating evidence of impact from multiple sources into one platform to expedite researcher evaluation

Submitted as conference communication. Accepted as poster. Cite as:

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017). Scholar Mirrors: Integrating evidence of impact from multiple sources into one platform to expedite researcher evaluation. In 22nd International Conference on Science, Technology & Innovation Indicators (STI). https://doi.org/10.17605/OSF.IO/z4bwe

Abstract (English)

This paper describes the creation of "Scholar Mirrors", a prototype web application that aims to provide a quick but accurate representation of the situation of a scientific discipline by integrating data from multiple online platforms. We chose the discipline of Bibliometrics / Scientometrics as a case study. After carrying out a series of keywords searches in Google Scholar Citations (GSC) and Google Scholar (GS), 813 relevant researchers were identified. Researchers were further classified as core (those who work mainly on Scientometrics) or related (those who work in other disciplines, with occasional incursions into Scientometrics). Additional information about these researchers was collected from other platforms (ResearcherID, ResearchGate, Mendeley, and Twitter). Up to 28 author-level indicators were collected about each researcher, as well as data about up to 100 of the most cited documents displayed in their GSC profile. The document-level data from all GSC profiles, as well as the data extracted from the keyword searchers in GS, was aggregated to create a list of the top 1000 most cited documents in the discipline. This document collection was further processed to generate a list of the most influential journals and publishers in the discipline. The results are accessible from the "Scholar Mirrors" website, which presents the results in four sections: authors, documents, journals, and book publishers. Lastly, the paper presents the main features of the web application, and the main limitations and future challenges of the product.

Abstract (Spanish)

Esta comunicación describe la creación de "Scholar Mirrors", un prototipo de aplicación Web cuyo objetivo es proporcionar una rápida, pero precisa representación de la situación de una disciplina científica mediante la integración de datos de múltiples plataformas online. Elegimos la disciplina Bibliometría/Cienciometría como caso de estudio. Después de realizar una serie de búsquedas por palabras clave en Google Scholar Citations (GSC) y Google Scholar (GS), 813 investigadores relevantes fueron identificados. Los investigadores fueron clasificados como core (aquellos que trabajan principalmente en cienciometría), o relacionados (aquellos que trabajan en otras disciplinas, pero realizan incursiones ocasionales en la cienciometría). También se recogió información adicional sobre estos investigadores de otras plataformas (ResearcherID, ResearchGate, Mendeley, y Twitter). Para cada investigador se extrajeron 28 indicadores a nivel de autor, así como datos sobre los 100 documentos más citados en su perfil GSC. Los datos a nivel de artículo, así como los datos extraídos de las búsquedas por palabras clave en GS fueron agregados para generar una lista de los 1000 documentos más citados de la disciplina. Estos datos fueron procesados para generar una lista de las revistas y editoriales más influyentes en la disciplina. Los resultados están disponibles en la web "Scholar Mirrors", que tiene cuatro secciones: autores, documentos, revistas, y editoriales de libros. Finalmente, este trabajo presenta las principales características de la aplicación web, sus principales limitaciones, y retos futuros para mejorar el producto.

1. Introduction

In the last few years there has been a proliferation of platforms that enable researchers to disseminate their publications on the Web and track the degree to which they are used by other people. These for the most part previously unavailable indicators might become a good complement to the ones currently used in evaluative bibliometrics. These platforms, though similar in some ways, are very diverse as regards the sources of their data, their purpose for which they were designed, their features, their user base… all of which affects the impact indicators they present. Thus, each platform shows its own – more or less distorted – reflection of the performance of a researcher, not unlike what happens to a person who enters a house of mirrors attraction at an amusement park (hence the name "Scholar Mirrors").

The objective of this work is to present a prototype web application (Scholar Mirrors http://www.scholar-mirrors.infoec3.es) which collects mainly author-level indicators from several of these platforms for a specific community of researchers (in this case, the Bibliometrics/Scientometrics community). This study will address the issues of researcher selection, data collection and processing, and the design of the database and web interface used to visualize the data.

2. Methods

The first task was to identify the set of researchers we wanted to study. To do this, we selected Google Scholar Citations (GSC) as our main source of data, and followed two different approaches so as to be as exhaustive as possible:

a) The keyword approach: A search was conducted in the most important journals of the field: Scientometrics, Journal of Informetrics, Research Evaluation, Cybermetrics, and the ISSI conferences (International Conference on Scientometrics and Informetrics) with the goal of extracting the most frequently used and representative words in the discipline. Table 1 shows the selected keywords. All public GSC profiles containing these keywords were selected. In addition, the lack of normalization in the use of keywords sometimes forced us to search variants of these keywords. These variants included misspelled words, the same keywords in other languages, etc. As an example, these are all the variants we found of the keyword "bibliometrics": bibliometric, bibliometría, bibliometria, bibliometric analysis, bibliometric methods, bibliometics, bibliometircs, bibliometric analysis in mining sciences, bibliometric mapping, bibliometric studies, bibliometric visualization, bibliometric., bibliometrics methodology, bibliometrics of social sciences and…, bibliometrics., bibliometrics..., bibliométrie, bibliometry.

b) The topic search approach: since there may be some authors working in this discipline who have created a public GSC profile, but who haven't added significant keywords or filled the institution field in their profile, we also conducted a topic search on Google Scholar (using the same keywords as before), and a journal search (all the documents indexed in Google Scholar published in the journals Scientometrics, Journal of Informetrics, Research Evaluation, Cybermetrics, as well as the ISSI conference proceedings), with the aim of finding authors we might have missed with the previous approach. This was possible because Google Scholar results display a link to the GSC profile of the author of the articles whenever a profile is available.

Table 1. Keywords selected to find authors

Altmetrics	Research Assessment
Bibliometrics	Research Evaluation
Citation Analysis	Research Policy
Citation Count	Science and Technology Policy
H Index	Science Evaluation
Impact Factor	Science Policy
Informetrics	Science Studies
Patent Citation	Scientometrics
Quantitative Studies of Science and Technology	Webometrics

The searches were conducted on the 24th of July, 2015. Researchers that didn't have a public GSC profile on that date are not included in this study.

Since Google Scholar Citations gives the author complete control over how to set their profile (personal information, institutional affiliation, research interests, as well as their scientific production), a systematic manual revision was carried out in order to:

a) Detect false positives: authors whose scientific production doesn't have anything to do with this discipline, even though they labelled themselves with one or more of the keywords associated with it.

b) Classify authors in two categories:

    i. Core authors: those authors whose scientific production substantially falls within the field of Bibliometrics.

    ii. Related authors: those authors who have sporadically published bibliometric studies, or whose field of expertise is closely related to Scientometrics (social, political, and economic studies about science), and therefore they can't be strictly considered bibliometricians.

In order to set a limit between the two categories, we decided to consider as core authors those who met the following criterion: at least half of the documents which contributed to their h index had to be related to Bibliometrics. We considered the titles of the documents, as well as the publishing channel where they appeared, focusing our attention in the journals. Our Bradford-like core of journals about Bibliometrics consisted of six journals (Scientometrics, Journal of Informetrics, JASIST, Research Evaluation, Research Policy, Cybermetrics), followed by other LIS journals which also publish numerous bibliometric studies (Journal of Information Science, Information Processing & Management, Journal of Documentation, College Research Libraries, Library Trends, Online Information Review, Revista Española de Documentación Científica, Aslib Proceedings, El Profesional de la Información) and lastly, journals devoted to social and political studies about science (Social Studies of Science, Science and Public Policy, Minerva, Journal of Health Services Research Policy, Technological Forecasting and Social Change, Science Technology Human Values, Environmental Science Policy, Current Science).

After this process, 813 relevant GSC profiles were identified. 397 of them were considered core authors, and the rest (416) as related authors. The data collection process was carried out using a custom web scraper written in Python. From each profile, this scraper extracted the researcher's personal information, all the author-level indicators available, and the bibliographic information of up to the 100 most cited documents in the profile, including the number of times cited. The data was initially saved as a two-table spreadsheet, one containing the personal information and author-level indicators for each author, and one containing the article references.

These 813 authors were searched by name in ResearcherID, ResearchGate, Mendeley, and Twitter, and in the cases where a profile was found, the indicators provided by these platforms were downloaded. We selected these sources because they are the most popular and widely used (Van Noorden, 2015; Bosman & Gramer, 2016). The data collection for these platforms was carried out between the 4th and 10th of September, 2015. Custom web scrapers were developed to extract the relevant author-level indicators from each platform.

A total of 28 author-level indicators were extracted from these sources:

a) Google Scholar Citations: sum of citations, h-index, and i10-index (for all years, and only for citations since 2010)

b) ResearchGate: RG Score, number of publications, sum of times cited, views, downloads, impact points, profile views, following (number of users the researcher follows), followers

c) Mendeley: number of publications, sum of readers, following, followers

d) ResearcherID (powered by Web of Science data): total number of articles in publication list, number of articles with citation data, sum of times cited, average citations per item, h-index

e) Twitter: number of tweets, days since registered, following, followers

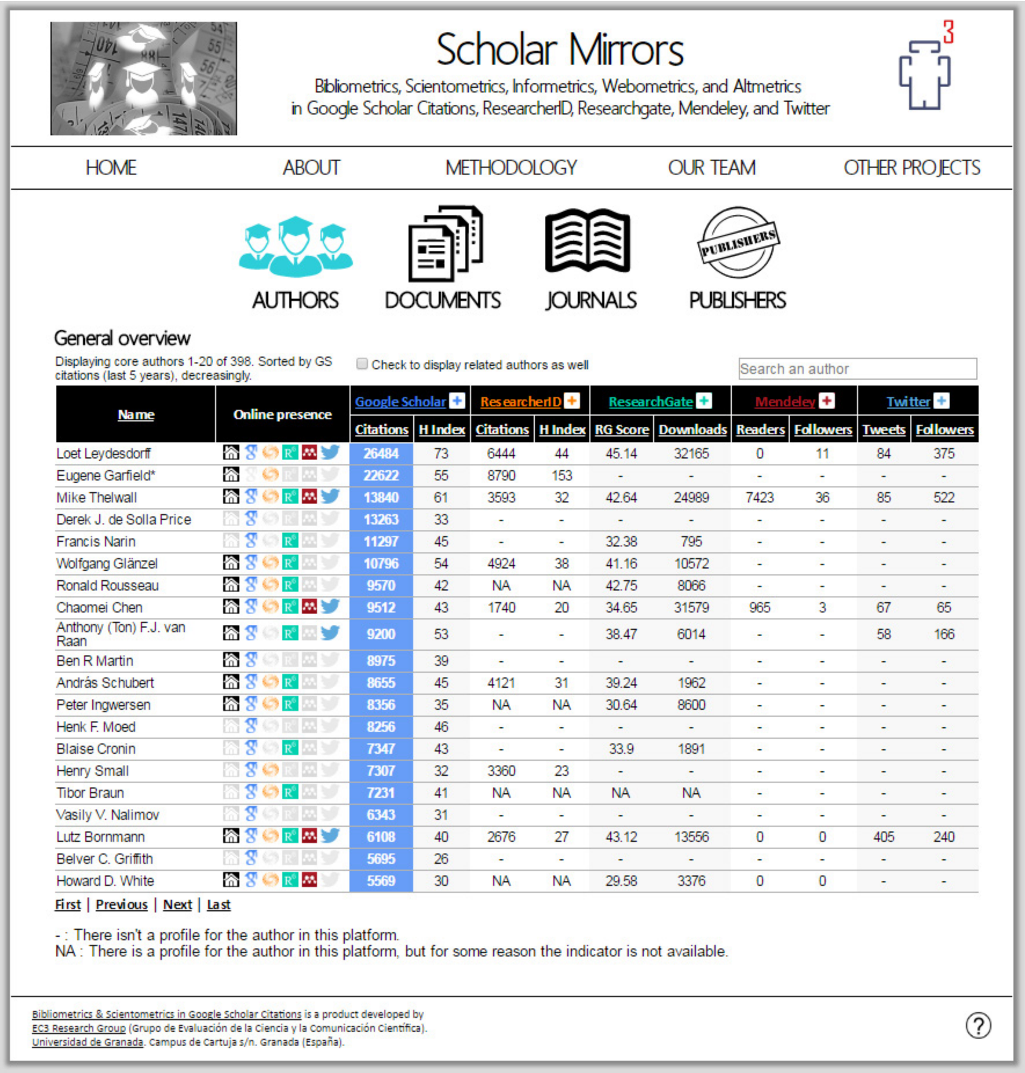
Additionally, from the article data extracted from the GSC profiles, as well as from the articles found in the "topic search approach" to select authors, a list of the top 1000 most cited documents of the field according to Google Scholar was generated. The citation counts according to Web of Science (WoS) were also collected, thanks to the Google Scholar / Web of Science integration available to subscribing institutions. In the case of books and other materials which are not covered by WoS, manual searches were carried out using the cited reference search tool available in WoS.

Using this set of top of highly cited documents, rankings of the most relevant journals and book publishers in the discipline were generated (according to the percentage of articles/books published by each journal/book publisher in the sample).

Once all the data had been processed, the resulting tables were saved to a SQLite database using Python. This database is the core component of a custom web application written mostly in PHP (and a little bit of JavaScript for some components), available at http://www.scholar-mirrors.infoec3.es.

3. Description of the web application

The application is structured in four sections: authors, documents, journals, and book publishers.



Scholar Mirrors

Bibliometrics, Scientometrics, Informetrics, Webometrics, and Altmetrics
in Google Scholar Citations, ResearcherID, Researchgate, Mendeley, and Twitter

General overview

Displaying core authors 1-20 of 398. Sorted by GS citations (last 5 years), decreasingly.

□ Check to display related authors as well

Name	Online presence	Google Scholar +		ResearcherID +		ResearchGate +		Mendeley +		Twitter +	
		Citations	H Index	Citations	H Index	RG Score	Downloads	Readers	Followers	Tweets	Followers
Loet Leydesdorff		26484	73	6444	44	45.14	32165	0	11	84	375
Eugene Garfield*		22622	55	8790	153	-	-	-	-	-	-
Mike Thelwall		13840	61	3593	32	42.64	24989	7423	36	85	522
Derek J. de Solla Price		13263	33	-	-	-	-	-	-	-	-
Francis Narin		11297	45	-	-	32.38	795	-	-	-	-
Wolfgang Glänzel		10796	54	4924	38	41.16	10572	-	-	-	-
Ronald Rousseau		9570	42	NA	NA	42.75	8066	-	-	-	-
Chaomei Chen		9512	43	1740	20	34.65	31579	965	3	67	65
Anthony (Ton) F.J. van Raan		9200	53	-	-	38.47	6014	-	-	58	166
Ben R Martin		8975	39	-	-	-	-	-	-	-	-
András Schubert		8655	45	4121	31	39.24	1962	-	-	-	-
Peter Ingwersen		8356	35	NA	NA	30.64	8600	-	-	-	-
Henk F. Moed		8256	46	-	-	-	-	-	-	-	-
Blaise Cronin		7347	43	-	-	33.9	1891	-	-	-	-
Henry Small		7307	32	3360	23	-	-	-	-	-	-
Tibor Braun		7231	41	NA	NA	NA	NA	-	-	-	-
Vasily V. Nalimov		6343	31	-	-	-	-	-	-	-	-
Lutz Bornmann		6108	40	2676	27	43.12	13556	0	0	405	240
Belver C. Griffith		5695	26	-	-	-	-	-	-	-	-
Howard D. White		5569	30	NA	NA	29.58	3376	0	0	-	-

- : There isn't a profile for the author in this platform.
NA : There is a profile for the author in this platform, but for some reason the indicator is not available.

Bibliometrics & Scientometrics in Google Scholar Citations is a product developed by
EC3 Research Group (Grupo de Evaluación de la Ciencia y la Comunicación Científica).
Universidad de Granada. Campus de Cartuja s/n. Granada (España).

Figure 1. Screen capture of the authors section. General overview.

Authors

This is the main section of the application. By default, users are presented with a list of the top 20 most cited core authors in the field, according to Google Scholar (Figure 1). This is the General Overview page, and it displays the most relevant indicators available from each platform (Google Scholar, ResearcherID, ResearchGate, Mendeley, and Twitter), as well as a column called "Online presence" which presents the links to the profiles in those platforms, whenever available.

The navigation options in this page are the following:

a) Sorting tables: It is possible to sort the author tables by any of the displayed indicators, just by clicking on the name of the indicator. It is also possible to sort the table by names (alphabetically). By default, indicators will be sorted in descending order. When the table is sorted by a given indicator, clicking again in the name of the same indicator will sort the table in ascending order. Text fields will be sorted in ascending order by default.

b) Navigating to following or previous pages: to facilitate visualization, only 20 authors are visible in each page. However, it is possible to navigate the entire set of authors by making use of the "First / Previous / Next / Last" links at the bottom-left side of the table.

c) Search Box: a search box is available to facilitate the task of finding a specific author. If a name or surname is entered in the box, a list of up to five names will appear just below it. Selecting one of them will automatically take the user to the page where that author is found (taking into account the current sorting criteria), and it'll be easily distinguishable from the rest because the background will be highlighted in yellow.

d) Navigating to platform-specific author tables: for each of the platforms there is a separate table that displays all the author indicators available in the platform. These tables can be accessed from the General Overview table by clicking in the appropriate header for each platform.

e) Core/Related authors: By default, only core authors are displayed. In order to display all authors, it is necessary to check the box with the label "Check to display related authors as well". When that box is checked, it'll be possible to tell core and related authors apart because the rows for core authors will be displayed with a grey background. Unchecking the box will hide related authors once again.

On the top-left part of the table a string of text will always inform of the current configuration parameters.

Documents

The documents section displays the top 1000 most cited documents in the field according to Google Scholar, along with the citation counts according to Web of Science. The bibliographic information in this table is structured in four columns: title of the document, authors, publication information (name of the journal, volume, issue, and pages in the case of journal articles, and publisher in the case of books or book chapters), and year of publication. This table can only be sorted by year of publication, times cited according to GS, and times cited according to WoS.

Journals

This section presents a ranking of journals according to the percentage of articles in the set of the top 1000 most cited documents in the field that are published in each journal. The percentage of citations of each journal (out of the sum of citations in those 1000 documents) is also displayed.

Book Publishers

This section presents a ranking of book publishers according to the percentage of books or book chapters in the set of the top 1000 most cited documents in the field that are published by each publisher. The percentage of citations of each book publisher (out of the sum of citations in those 1000 documents) is also displayed.

4. Limitations and future challenges

As we advanced in the introduction, this product is only just an early prototype, with which we wanted to test the feasibility of developing a product that integrates impact indicators from diverse sources. We are aware of its many shortcomings, and believe that there is still a long way to go, if ever, before a product such as this one should be considered for use in evaluative processes.

The limitations we have detected are:

- Only researchers with a public profile in Google Scholar Citations at the time of data collection are considered.

- The website is not easily updatable: the data collection scripts are not integrated into the web application. This means that a lot of human intervention is needed to update the data (running the web scrapers to extract the updated data from each platform, processing it, and adding the new data to the database). Ideally, these processes should be carried out automatically by the web application on a regular basis. When a more straightforward system is in place, it will also be possible to study the evolution of the indicators for a particular researcher over time.

- Incomplete or incorrect data in the profiles: most profile platforms leave on the hands of the users the responsibility of keeping their profile up to date and free of errors. However, many researchers don't consider this an important task, and so there are many outdated profiles, or profiles with incorrect information, which obviously affects the impact indicators. This issue is difficult to address, but at the very least, a few mechanisms to detect profiles that are likely to contain errors and warn about this should be implemented.

- Subject classification has been done manually. Ideally, an automatic classification, not at the author-level, but at the level of documents themselves, would allow a much more precise representation of the importance of an author in a specific area, field, or subfield. This is especially true in Bibliometrics, where researchers come from many different areas. A document-level classification would allow the calculation of author-level indicators using only the documents that are relevant to the field that is the object of study. The use of the core/related classification for authors in this product is just a rudimentary way of addressing this issue.

- The documents section (as well as the journals and book publishers sections) are just rough drafts of what could be done. In this product, the only sources of article-level indicators are Google Scholar and Web of Science. This could be extended to cover indicators from many other platforms (usage indicators, altmetrics…). Additionally, if more detailed information could be obtained, such as the references of the citing articles themselves, instead of only the citation counts, other issues like detecting unusual levels of self-citations could be addressed.

- Another aspect that this product doesn't address is collaboration. Only with the information available in Google Scholar it is possible to generate collaboration networks among researchers and in some cases even among institutions and countries, although richer metadata would probably be necessary to address the last two.

References

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. *Nature*, *512*(7513), 126–129. https://doi.org/10.1038/512126a

Bosman J, Kramer B. (2016). *Global survey on research tool usage*. [Data set]. https://doi.org/10.5281/zenodo.49583

Chapter 12. A novel method for depicting academic disciplines through Google Scholar Citations: The case of Bibliometrics

Journal article. Cite as:

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). A novel method for depicting academic disciplines through Google Scholar Citations: The case of Bibliometrics. *Scientometrics*, 114(3), 1251–1273. https://doi.org/10.1007/s11192-017-2587-4

Abstract (English)

This article describes a procedure to generate a snapshot of the structure of a specific scientific community and their outputs based on the information available in Google Scholar Citations (GSC). We call this method MADAP (Multifaceted Analysis of Disciplines through Academic Profiles). The international community of researchers working in Bibliometrics, Scientometrics, Informetrics, Webometrics, and Altmetrics was selected as a case study. The records of the top 1,000 most cited documents by these authors according to GSC were manually processed to fill any missing information and deduplicate fields like the journal titles and book publishers. The results suggest that it is feasible to use GSC and the MADAP method to produce an accurate depiction of the community of researchers working in Bibliometrics (both specialists and occasional researchers) and their publication habits (main publication venues such as journals and book publishers). Additionally, the wide document coverage of Google Scholar (specially books and book chapters) enables more comprehensive analyses of the documents published in a specific discipline than were previously possible with other citation indexes, finally shedding light on what until now had been a blind spot in most citation analyses.

Abstract (Spanish)

Este artículo describe un procedimiento para generar una foto fija de la estructura de una comunidad científica específica y de sus publicaciones, basada en información disponible en Google Scholar Citations (GSC). Llamamos a este método MADAP (Multifaceted Analysis of Disciplines through Academic Profiles). Seleccionamos a la comunidad internacional de investigadores que trabajan en Bibliometría, Cienciometría, Informetría, Webometría, y Altmetría como estudio de caso. Se procesaron manualmente los registros del top 1.000 de documentos más altamente citados publicados por estos autores según GSC, con el objetivo de añadir cualquier información faltante, y de normalizar campos como el nombre de la revista, y la editorial de los libros. Los resultados sugieren que es factible utilizar GSC y el método MADAP para generar una representación precisa de la comunidad de investigadores del área de Bibliometría (tanto de los especialistas como de los investigadores ocasionales). Además, la amplia cobertura de Google Scholar (especialmente en libros y capítulos de libro) permite realizar análisis más exhaustivos de los documentos publicados en una disciplina específica comparado con las posibilidades que ofrecen otros índices de citas, de manera que al fin se puede arrojar luz sobre lo que hasta ahora había sido un punto ciego en la mayoría de los análisis de citas.

1. Introduction

Science, in order to be properly investigated, grasped, and taught, has usually been organized in various areas of knowledge. Over time, each of these areas has been further divided into fields, subfields, disciplines, and specialties, as a result of the ever faster growth of knowledge and the parallel increase in the number of people who form the scientific communities within each of these areas. This process of scientific budding resembles the life cycle of a living being (birth, growth, reproduction, and death), and is subject to an endless metamorphosis.

Each of these units in which scientific knowledge is structured has its own idiosyncrasies and epistemological properties (its object, its principles, and its methods) that endow them with a characteristic identity as well as boundaries that demarcate their cognitive territory. However, the inner and outer boundaries are not always clearly defined due to overlaps between disciplines, gaps, and loops, sometimes quite vague and difficult to trace.

The different areas of knowledge are populated by communities of scientists and professionals, each group using their own tools, methodologies and techniques. These are social groups that share – with more or less consensus – professional practices, forms of work organization, living conditions, social expectations, principles, values, and beliefs.

Whitley (1984) dissected the process by which academic communities – and their disciplines and specialties – become socially and cognitively institutionalized: how they create organizations that allow them to associate in order to defend their interests; how they erect spaces for the exchange of ideas and social development (conferences, seminars, forums, etc.); how they institute professional (newsletters, discussion lists) or scientific means (journals) of communication; how they obtain academic standing by teaching the subject at the university (courses in graduate and postgraduate programs, including Master and PhD degrees); how they create groups, departments, laboratories, and companies dedicated to advance research; how they define research agendas where not only research problems but also ways to solve them are addressed; or how to create a common language to establish ideas and principles. Not to mention that the process of social and cognitive institutionalization of disciplines is directly influenced by the geographic location and the different levels of economic and cultural development of the countries where researchers are based.

As formulated by Becher and Trowler (2001), there is a close relationship between the disciplines (territories of knowledge) and people who advance them (scientific tribes); between the epistemic properties of the forms of scientific knowledge and the social aspects of academic communities. This is why any analysis of a discipline cannot ignore the cognitive (disciplines themselves) and social (community) areas. A discipline is what is performed by those who cultivate it.

Being aware of the scope of a discipline will not only help characterize and determine its perspective and scientific nature, but it will also indirectly delineate its internal structure, its coherence, its contours, and its location in the overall picture of the Sciences. This will enable an understanding of what the research is and has been about in a particular discipline, and how it may evolve in the future.

Although there is no unanimity yet about what the most appropriate methods to describe disciplines are, this work intends first to depict one scientific discipline and those who practice it (through a multifaceted approach based on the intellectual production generated by its academic community), and second, to carry out this procedure using both semi-supervised (Google Scholar Citations) and unsupervised environments (Google Scholar).

Therefore, the main goal of this work is to investigate the suitability of Google Scholar (GS) and Google Scholar Citations (GSC) to provide a comprehensive and multifaceted picture of the structure of an entire scientific specialty through the main agents that are part of it (scientists, professionals, the documents they produce, and the venues where these documents are published).

While classic citation indexes (Scopus and Web of Science) have been traditionally used to analyse scientific disciplines, their particular coverage and principles (controlled sets of journals that represent the elite, based on a Bradford-like core) have probably constrained the pictures that could be obtained. These databases provide a better coverage in areas like Science, Medicine and Technology, but they lack many relevant sources in areas like the Social Sciences and Humanities. Academic search engines like GS practice a radically different approach when it comes to selecting sources to cover and index, and therefore it might be useful to explore the wider view of academic outputs that they provide (Martín-Martín et al. 2016). To the best of our knowledge, there have not yet been any attempts to comprehensively analyse an entire discipline using GS and GSC.

Both GS and GSC present a series of well-known shortcomings and restrictions that hinder the use of these platforms for bibliometric analyses (Jacsó 2005; 2008; 2012; Meho and Yang 2007; Aguillo 2012; Prins 2016). Therefore, the development of a method that enables the use of these platforms for bibliometric purposes would facilitate studies that are not limited by the document coverage biases of other citation indexes.

In this line, this study intends to answer the following questions:

RQ1: Can GSC and GS be used to generate a representation of the community of authors that work in any given academic discipline, and their outputs?

RQ2: Is it possible to apply a multi-faceted approach to analyse a discipline with the data available in GS and GSC?

A positive answer to these questions would mean that it is possible to carry out bibliometric analyses of disciplines using Google Scholar Citations, a source of data that is free to access and semi-automatically updated. The data from this source could at the very least complement the data available in other subscription-based citation indexes.

In order to answer this research question, this work takes as a case study a very specific scientific and professional community (Bibliometrics) along with its close-related areas (Scientometrics, Informetrics, Webometrics, and Altmetrics). The reason behind the selection of this discipline is that the authors are familiar with this field. This expertise is considered necessary in order to assess the results of the analyses and be able to detect the potential shortcomings of the method.

2. Research background

The object of study. Bibliometrics: A discipline with many names

There are numerous works which address the history of Bibliometrics (Broadus 1987a; Hertzel 1987; Shapiro 1992; Godin 2006; De Bellis 2009). Its denomination, object of study and scope have been addressed as well (Lawani 1981; Bonitz 1982; Peritz 1984; Broadus 1987b; Brookes 1988; 1990; Sengupta 1992; Glänzel and Schoepflin 1994; Braun 1994; Gorbea 1994; Hood and Wilson 2001; Cronin 2001; Thelwall 2008; Lariviere 2012). There are also several literature reviews about this subject (Narin and Moll 1977; White and McCain 1989; Van Raan 1997; Wilson 1999; Borgman and Furner 2002).

Bibliometrics can be synthetically defined as the discipline responsible for measuring communication and, more specifically, as the specialty responsible for quantitatively studying the production, distribution, dissemination and consumption of information conveyed in any type of document (book, journal, conference, patent, or website) and across all spheres of activity, but with special attention to scientific information. This discipline has various peculiar features:

a) It is a very young discipline, and its epistemic foundations are still not fully defined.
b) It is a discipline best defined by its methods than by the thematic areas that it covers.
c) It has a strong interdisciplinary nature, which arises from the incorporation of methods and techniques developed in other fields, and by its application to the study of any subject area.

It is probably because of these reasons that this discipline is known by many different names. However, this fact does not mean that the subject of study or the borders of the discipline are not clearly defined. Rather, it is a sign of the coexistence of different traditions that have shaped the development of the discipline.

Bibliometrics is the original and most widely-used term to refer to it. It stems from the bibliographic tradition represented by Paul Otlet with his proposal for a "bibliometrie", a Science for measuring all the dimensions of books and other documents (Otlet 1934), and from the library tradition concerned since ancient times with measuring the growth of knowledge and the usage of its holdings (Ranganathan 1969).

Scientometrics is oriented towards the quantitative analysis of scientific and technical literature. It comes from the tradition of the science of science (space of confluence of Sociology, History, and Philosophy of science), to which science policy is also linked. It was crucial for this scientometric orientation the creation of the citation indexes (Garfield 1970).

Informetrics is focused on the discovery of mathematical models that explain the properties of information (Egghe and Rousseau 1990; Tague-Sutcliffe 1992; Bar-Ilan 2008). It is connected with the modern information science. Webometrics (Almind and Ingwersen 1997; Thelwall, Vaughanand Björneborn 2005; Thelwall 2009) and Altmetrics (Priem and Hemminger 2010) are the most recent denominations. They started to gain momentum as the use of the new information and communication technologies began to spread. They are being developed in the tradition of the modern Library and Information Science, a discipline increasingly dedicated to computer science and to computing itself. These new names are strongly influenced by the medium in which information is conveyed rather than by the content itself.

The terms used as well as their conceptual domains and boundaries have been already described in the literature (Björneborn and Ingwersen 2004; Milojević and Leydesdorff 2013; Stuart 2014). However, there is no consensus on the precise relation among them. By way of illustration, an analysis of the five selected terms (Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics) used in the titles of documents published between 1969 and 2016 and indexed in GS (Figure 1) shows a clear predominance of the term “Bibliometrics”, followed by “Scientometrics”.

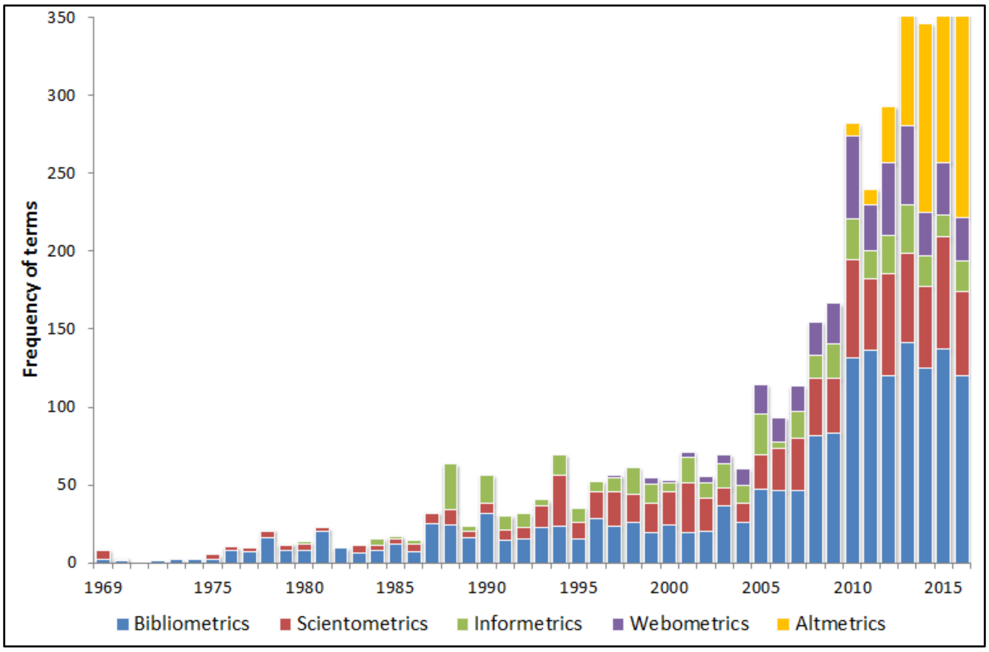


Figure 1. Frequency of the terms “Bibliometrics”, “Scientometrics”, “Informetrics”, “Webometrics” and “Altmetrics” in the title of documents indexed in Google Scholar (1969-2016)

The term “Altmetrics” is being increasingly used (Figure 2) in the last three years as a result of the novelty of the new social media communication technologies. Another reason why Altmetrics is currently a hot topic in the field is the relatively unknown role that the metrics that this term encompasses can play in the quantification and evaluation of academic impact, both at the article (Lin and Fenner 2013) and author levels (Bar-Ilan et al. 2012; Orduna-Malea et al. 2016).

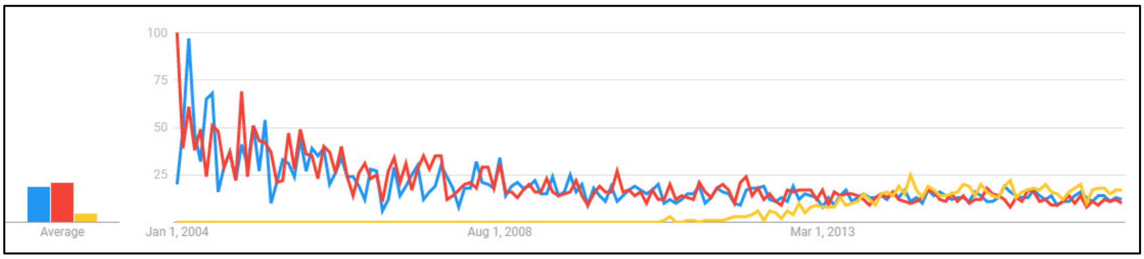


Figure 2. Interest measured in search queries frequency of the terms Bibliometrics, Scientometrics and Altmetrics

Source: Google Trends
Blue: Bibliometrics; Red: Scientometrics; Yellow: Altmetrics

The unit of analysis. Google Scholar Citations: an unmoderated academic profile

GSC was launched in 2011 (Jacsó 2012) and currently stands out as one of the preferred academic profiles by scholars. Kramer and Bosman (2015) released a comprehensive report about the use of academic communication tools, finding that GSC was used by 62% of the surveyed users (about 20,000), in second place just after ResearchGate (66%). The fact that GSC is linked to Google Scholar, currently the most comprehensive academic bibliographic database (Orduna-Malea et al. 2015), as well as the preferred source to start academic information discovery processes (Orduna-Malea et al. 2016), makes this service an essential professional tool for academics.

Several studies have recently used data extracted from GS for bibliometric purposes (Bornmann, Thor, Marx, and Schier 2016; Martín-Martín, Orduna-Malea, Ayllón and López-Cózar 2014; Mingers and Meyer 2017; Mingers, O'Hanley and Okunola in press). However, since the information contained in GSC is better structured than in GS, this platform has recently started to be used as a new source for bibliometric studies. Ortega and Aguillo (2012) used GSC to map the labels included in each profile to build a Science map as well as to construct country and institutional collaboration networks using co-authors lists of these profiles (Ortega and Aguillo 2013). The issue of its coverage has been addressed as well, finding not only an unbalanced subject coverage (with an important bias in favour of Computing Sciences and Engineering) but also a bias in favor of young researchers and specific institutions and countries (Ortega 2015a). Despite this, Ortega and Aguillo (2014) acknowledge that GSC has an interesting potential for research evaluation, such as a wider coverage of academic outputs and therefore a broader coverage of research impact**.**

Some other studies have applied GSC data to specific research environments. For example, Ortega (2015b) focused on the researchers affiliated to the Spanish National Research Council (CSIC), and Mikki et al. (2015) focused on the researchers at the University of Bergen. Nevertheless, these studies only analyse specific institutions.

Haustein et al. (2014) studied the social media presence of attendees at the 2010 STI conference celebrated in Leiden (57 researchers, who together had authored 1,136 papers). However, to the best of our knowledge there has not yet been any exhaustive study focused on one academic discipline (in this case Bibliometrics), which addresses not only author-level metrics but also documents and sources. Therefore, the main objective of this paper is to identify and describe a scientific discipline through the data available in GSC on the authors who work in said discipline.

3. Methods

We developed and tested a method to capture, classify and measure data from the different scientific agents of one discipline. We called this method MADAP (Multifaceted Analysis of Disciplines through Academic Profiles).

3.1. Author profiles search and identification

The first step was to identify all authors who have published in the areas of Bibliometrics, Scientometrics, Informetrics, Webometrics or Altmetrics, and for whom a GSC public profile could be found at the time of data collection (July 24th 2015). In order to identify the set of authors relevant to our study, an iterative snowball process was conceived, which consisted on the following search strategies.

a) Keywords

A search was conducted in four core selected journals (Scientometrics, Journal of Informetrics, Research Evaluation, and Cybermetrics) as well as the ISSI conferences (International Conference on Scientometrics and Informetrics) with the goal of extracting the most frequently used and representative words in the discipline. This process was driven by the need of capturing keywords describing the discipline. Among these terms we expected to find the most common keywords that authors use to describe their scientific interests in their GSC profiles. For this reason, we considered that these four purely bibliometric sources were sufficient for this purpose. The inclusion of other important sources which publish bibliometric studies, but also publish studies in other topics (for example, JASIST) might have introduced too much noise (keywords related to information retrieval, for example) and we think it unlikely that they would have provided any relevant terms that could not be extracted from the other journals.

To do this, the bibliographic records from all indexed articles published by these four sources were automatically retrieved using the Web of Science (n= 7143). This database was used due to its data export features, which facilitated the extraction of the documents' keyword field, a field that is not available in the metadata presented by GS. Next, all significant terms from the documents' titles and keywords (when available) were extracted. A pool of 619 terms (458 from titles and 161 from keywords) with a minimum frequency of occurrence of five in our set of documents was obtained. This vocabulary was manually processed to merge variants of the same term (for example, bibliometric and bibliometrics), delete duplicates, and exclude irrelevant terms (e.g., credit, editorial board, Nobel price, item, program, content, etc.), which were highly mentioned but useless for our purpose of representing a discipline.

After obtaining the list of terms, we checked for the existence of GS profiles in which the authors had selected one or more of these terms as their areas of interest (GSC allows authors to display up to five areas of interests). For example, the term "citation index" appeared in the title of 89 articles. However, no one had selected this term in their GS profile. Terms that no author had selected as a research interest were therefore ignored from this point on.

Lastly, the data available in all public GSC profiles that contained one or more of the selected terms as areas of interest were collected. The lack of normalization in the use of keywords sometimes forced us to search alternative keywords. These variants included misspelled words, the same keywords in other languages, etc.

b) Institutional affiliation

All the profiles associated with research centres working on Bibliometrics were also selected regardless the research interest keywords used by authors. As an example, profiles with verified e-mail domains such as <cwts.leidenuniv.nl>, <cwts.nl>, or <science-metrix.com> were selected.

c) Additional searches

Since there may have been some authors working in the discipline and who have created a public GSC profile, but who haven't added significant keywords or appropriately filled the affiliation field in their profile, we also conducted a topic search on GS (using the same previously selected terms) as well as a journal search (all the documents indexed in Google Scholar published by the core journals previously mentioned), with the aim of finding authors we might have missed with the previous two strategies.

The last two search strategies provided profiles with new keywords, some of them quite important to the discipline though they did not appear in the sample of 7143 document titles (e.g., Science and Technology Policy; 0 mentions in Titles, 72 authors including this term). These keywords were included in the final master list of disciplinary keywords. All terms that are not exclusively related to the discipline (Information Science: 61 profiles; Open Access: 41 profiles; Information literacy, 36 profiles) were excluded. The final master list of keywords consisted of 18 keywords. Table 1 displays the frequency of occurrence of these terms in the sample of documents (in Title and Keywords) and the number of authors that use that keyword in their GSC profile to describe their research interests.

Table 1. List of Keywords describing Bibliometrics discipline

Term	WoS source		GSC Profile source
	Title	Article Keyword	Author Keywords
Bibliometrics	**640**	313	444
Scientometrics	**372**	127	382
H-Index	**152**	144	1
Impact Factor	**135**	149	1
Citation Analysis	**124**	199	58
Informetrics	**108**	21	75
Research Evaluation	**62**	104	74
Webometrics	**38**	49	68
Patent Citation	**30**	17	1
Research Assessment	**26**	28	13
Citation Count	**25**	0	0
Research Policy	**17**	16	37
Science Policy	**16**	21	148
Altmetrics	**11**	27	29
Science Studies	**9**	0	57
Quantitative Studies of Science and Technology	**6**	0	1
Science Evaluation	**3**	0	7
Science and Technology Policy	**0**	21*	72

* Occurrences for “Science and Technology”

3.2. Filtering and classification of author profiles

GSC gives authors complete control over how to set their profile (personal information, institutional affiliation, research interests, as well as their scientific production). For this reason, a systematic manual revision was carried out in order to:

- Detect false positives: authors whose scientific production doesn’t have anything to do with this discipline, even though they labelled themselves with one or more of the keywords associated with it.
- Classify authors in two categories:

  a) *Specialists*: authors whose scientific production substantially falls within the field of Bibliometrics.
  b) *Occasional*: authors who have sporadically published bibliometric studies, or whose field of expertise is closely related to Scientometrics (social, political, and economic studies about science), and therefore they can’t be strictly considered bibliometricians.

In order to set the boundaries between the two categories (specialist and occasional authors), we decided to consider as “specialist authors” those who meet the following criterion: at least half of the documents which contribute to their h-index should fall within the limits of the field of Bibliometrics.

In order to establish the limits of the field we considered the titles of the documents as well as the venue where they were published, focusing our attention in the journals. Our Bradford-like core of journals about Bibliometrics consisted of six journals (Scientometrics, Journal of Informetrics, JASIST, Research Evaluation, Research Policy, and Cybermetrics), followed by other LIS journals which also publish numerous bibliometric studies (Journal of Information Science, Information Processing & Management, Journal of Documentation, College Research Libraries, Library Trends, Online Information Review, Revista Española de Documentación Científica, Aslib Proceedings, and El Profesional de la Información). Lastly, journals devoted to social and political studies about science (Social Studies of Science, Science and Public Policy, Minerva, Journal of Health Services Research Policy, Technological Forecasting and Social Change, Science Technology Human Values, Environmental Science Policy, and Current Science) were also searched.

811 GSC profiles were identified, out of which 48.83% (396) were classified as specialists, and the remaining 51.17% (415) as occasional authors in Bibliometrics.

3.3. A multi-faceted approach: units of scientific analysis

Once the set of 811 authors had been identified, we extracted the number of citations received by each of them directly from their GSC profiles (see Table 2). Additionally, we automatically extracted – by means of an ad hoc web scraper – the top 100 most cited documents for each specialist author from their GSC profile. To this set of documents (39,600), we manually added the documents we found through the additional keyword and journal queries that had been previously performed in Google Scholar (15,000 documents authored by researchers with or without a public profile in GSC).

After deleting duplicates, a set of roughly 41,000 documents remained. In the cases where various versions of the same document were found with different number of citations, the one with the highest citation count was selected. This list was sorted according to the number of citations. For each of the top 1,000 most cited documents in this list, the basic bibliographic information (especially the sources: journals and book publishers) were collected (see Tables 3, 4, and 5).

For the sake of clarity we should point out that in those cases when a book is a collective work, the number of citations is the sum of the citations to each of the chapters, in addition to the citations directed to the book as a whole.

A graphical visualization of the MADAP procedure can be found in Figure 3

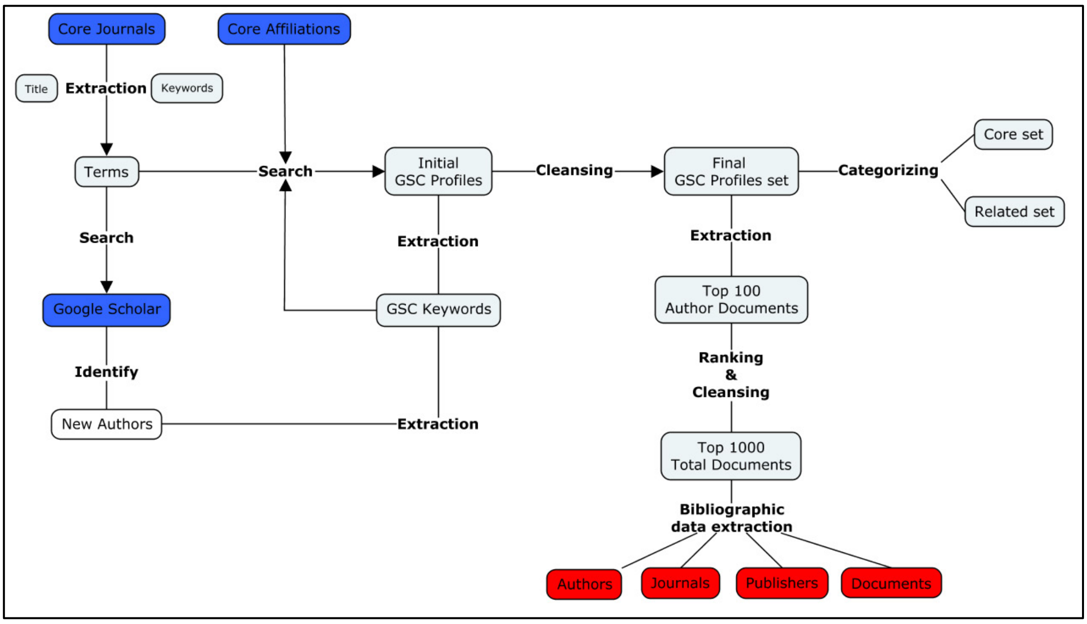


Figure 3. Description of MADAP method

4. Results

4.1. The actors of Bibliometrics according to Google Scholar Citations, through the MADAP method

a) Authors

The list of most influential authors of the discipline is available in the Table 2.

Table 2. Top 25 influential specialist/occasional authors in Bibliometrics according to Google Scholar Citations

SPECIALIST AUTHORS	CITATIONS	H INDEX	OCCASIONAL AUTHORS	CITATIONS	H INDEX
Loet Leydesdorff	**26,484**	73	Robert K. Merton	**109,507**	104
Eugene Garfield	**22,622**	55	Francisco Herrera	**38,407**	101
Mike Thelwall	**13,840**	61	Keith Pavitt	**35,521**	65
Derek J. de Solla Price	**13,263**	33	Peter Willett	**25,758**	74
Francis Narin	**11,297**	45	Richard S J Tol	**21,851**	77
Wolfgang Glänzel	**10,796**	54	Stevan Harnad	**17,330**	62
Ronald Rousseau	**9,570**	42	Collins Harry	**16,355**	49
Chaomei Chen	**9,512**	43	Enrique Herrera-Viedma	**16,154**	62
Anthony F.J. van Raan	**9,200**	53	George Kingsley Zipf	**14,745**	15
Ben R Martin	**8,975**	39	Alfred J. Lotka	**14,706**	30
András Schubert	**8,655**	45	Barry Bozeman	**13,764**	56
Peter Ingwersen	**8,356**	35	John Mingers	**11,997**	49
Henk F. Moed	**8,256**	46	Daniele Archibugi	**11,996**	48
Blaise Cronin	**7,347**	43	William C. Clark	**11,915**	41
Henry Small	**7,307**	32	Bart Verspagen	**11,490**	56
Tibor Braun	**7,231**	41	Stan Metcalfe	**10,829**	50
Vasily V. Nalimov	**6,343**	31	Reinhilde Veugelers	**10,581**	41
Lutz Bornmann	**6,108**	40	David I. Stern	**9,695**	39
Belver C. Griffith	**5,695**	26	Yannis Manolopoulos	**9,557**	45
Howard D. White	**5,569**	30	Andy Stirling	**8,989**	45
Johan Bollen	**5,394**	33	Christine L. Borgman	**8,893**	41
Katy Borner	**5,326**	31	Anne-Wil Harzing	**8,839**	44
Félix de Moya Anegón	**5,074**	35	Kal Jarvelin	**8,669**	32
Koenraad Debackere	**4,933**	32	Johan Schot	**8,639**	32
Jose Maria López Piñero	**4,823**	31	John P. Walsh	**8,500**	29

b) Documents

The equivalent list of most influential documents according to GSC in the field of Bibliometrics is available in Table 3.

Table 3. Top 25 most influential documents in Bibliometrics according to Google Scholar Citations

TITLE	AUTHORS	SOURCE	YEAR	CITATIONS
Little science, big science	Price	Columbia University Press	1963	**5,410**
An index to quantify an individual's scientific research output	Hirsch	PNAS	2005	**4,860**
The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university-industry-government relations	Etzkowitz & Leydesdorff	Research Policy	2000	**4,414**
Universities and the global knowledge economy: a triple helix of university-industry-government relations	Etzkowitz & Leydesdorff	Pinter Press	1997	**2,585**
Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems	Moed, Glänzel & Schmoch (ed.)	Springer	2005	**2,261**
Citation analysis as a tool in journal evaluation. Journals can be ranked by frequency and impact of citations for science policy studies	Garfield	Science	1972	**2,166**
Citation indexing: Its theory and application in science, technology, and humanities	Garfield	Wiley	1979	**2,130**
The frequency distribution of scientific productivity	Lotka	J. of Washington Academy Sciences	1926	**2,090**
Co-citation in the scientific literature: A new measure of the relationship between two documents	Small	JASIS	1973	**1,988**
Links and impacts: The influence of public research on industrial R&D	Cohen, Nelson & Walsh	Management Science	2002	**1,881**
Evolution of the social network of scientific collaborations	Barabasi et al	Physica A	2002	**1,851**
Citation indexes for science. A new dimension in documentation through association of ideas	Garfield	Science	1955	**1,783**
What is research collaboration?	Katz & Martin	Research Policy	1997	**1,591**
Handbook of quantitative studies of science and technology	Van Raan (ed.)	North-Holland	1988	**1,510**
The history and meaning of the journal impact factor	Garfield	JAMA	2006	**1,487**
The increasing linkage between US technology and public science	Narin, Hamilton & Olivastro	Research Policy	1997	**1,211**
A general theory of bibliometric and other cumulative advantage processes	Price	JASIST	1976	**1,148**
Statistical bibliography or bibliometrics?	Pritchard	J. of Documentation	1969	**1,134**
Theory and practise of the g-index	Egghe	Scientometrics	2006	**1,113**
The Web of knowledge: a Festschrift in honor of Eugene Garfield	Garfield, Cronin & Atkins (ed).	Information Today	2000	**1,102**
Visualizing a discipline: An author co-citation analysis of information science, 1972-1995	White & McCain	JASIS	1998	**1,100**
CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature	Chen	JASIST	2006	**1,083**
Citation analysis in research evaluation	Moed	Springer	2005	**1,060**
Citation frequency and the value of patented inventions	Harhoff et al	R. of Economics and Statistics	1999	**1,023**

Maps of random walks on complex networks reveal community structure	Rosvall & Bergstrom	PNAS	2008	**992**

c) Journals

The third unit analysed was the journals in which highly cited documents had been published (i.e., considering only the top 1,000 most cited documents). Table 4 contains the top 25 journals according to the number of highly cited documents published. Additionally, we show the total number of citations received by these articles, the percentage of citations per article (C/A), the percentage of highly cited documents in the sample (HCD) and the distribution of citations.

Table 4. Top 25 most influential journals in Bibliometrics according to Google Scholar Citations

JOURNAL	DOCUMENTS	CITATIONS	C/A	HCD (%)	CITATIONS (%)
Scientometrics	**284**	44,384	156	29.8	22.5
JASIST	**137**	27,021	197	14.4	13.7
Research Policy	**57**	18,866	330	6.0	9.6
Journal of Informetrics	**36**	5,052	140	3.8	2.6
Journal of Documentation	**25**	5,538	221	2.6	2.8
Information Processing & Management	**24**	4,404	183	2.5	2.2
Journal of Information Science	**20**	3,815	190	2.1	1.9
Research Evaluation	**18**	2,126	118	1.9	1.1
ARIST	**14**	3,621	258	1.5	1.8
Social Studies of Science	**13**	3,204	246	1.4	1.6
Science and Public Policy	**13**	2,875	221	1.4	1.5
Plos One	**13**	2,376	182	1.4	1.2
Nature	**10**	1,871	187	1.0	1.0
Current Contents	**10**	1,696	169	1.0	0.9
PNAS	**9**	7,642	849	0.9	3.9
Science	**8**	9,219	1,152	0.8	4.7
Library Trends	**7**	1,230	175	0.7	0.6
Medicina Clinica	**6**	958	159	0.6	0.5
Online Information Review	**6**	806	134	0.6	0.4
Science Technology & Human Values	**5**	946	189	0.5	0.5
Aslib Proceedings	**5**	765	153	0.5	0.4
Cybermetrics	**5**	627	125	0.5	0.3
American Psychologist	**4**	1,026	256	0,4	0,5
World Patent Information	**4**	726	181	0.4	0.4
Ethics in Science and Environmental Politics	**4**	687	171	0.4	0.3

C/A: Citations per article; HCD (%): Percentage of highly cited articles (top 1,000 most cited documents in the sample; Citations (%): Distribution of citations in the sample

d) Book publishers

The last unit of analysis is the book publishers. The top 20 publishers according to the percentage of highly cited books or book chapters (top 1,000) are presented in Table 5. Additionally, the number of documents, citations (total and percentage of citations respect to the total) and citations per document are displayed.

Table 5. Top 20 most influential book publishers in Bibliometrics according to Google Scholar Citations

PUBLISHER	HCD	HCD (%)	CITATIONS	CITATIONS (%)	C/D
Springer	10	**18,2**	5,766	14,3	576.60
Information Today	6	**10,9**	1,635	4,0	272.50
Wiley	5	**9,1**	3,121	7,7	624.20
Lexington	4	**7,3**	1,627	4,0	406.75
Sage	4	**7,3**	1,324	3,3	331.00
UFMG	4	**7,3**	845	2,1	211.25
University of Chicago Press	3	**5,5**	6,874	17,0	2,291.33
Russell Sage Foundation	3	**5,5**	3,836	9,5	1,278.67
North-Holland	3	**5,5**	2,130	5,3	710.00
Blackwell	2	**3,6**	1,132	2,8	566.00
Elsevier	2	**3,6**	1,071	2,7	535.50
Taylor Graham	2	**3,6**	688	1,7	344.00
Scarecrow Press	2	**3,6**	416	1,0	208.00
ISSI	2	**3,6**	276	0,7	138.00
Ablex	2	**3,6**	193	0,5	96.50
FECYT	2	**3,6**	193	0,5	96.50
Columbia University Press	1	**1,8**	5,410	13,4	5,410.00
Pinter Press	1	**1,8**	2,585	6,4	2,585.00
Yale University Press	1	**1,8**	936	2,3	936.00
MIT Press	1	**1,8**	710	1,8	710.00

HCD: Highly cited documents; C/D: Citations per document

4.2. The map of the discipline

To visualise the relations between the main actors of Bibliometrics and related fields, a network connecting the main authors and journals/publishers has been generated (Figure 4). Since the set of 1,000 highly cited documents is too big to be easily visualised, only the Top 200 documents have been considered. For each of these documents all authors and sources have been extracted and linked. In this case, all the co-authors of each of the 200 documents have been analysed, discarding authors not related with the discipline, and including authors that are related but do not have a public GSC profile (this approach allows the consideration of this important set of authors, although data from the GS database was needed in addition to the data available in GSC).

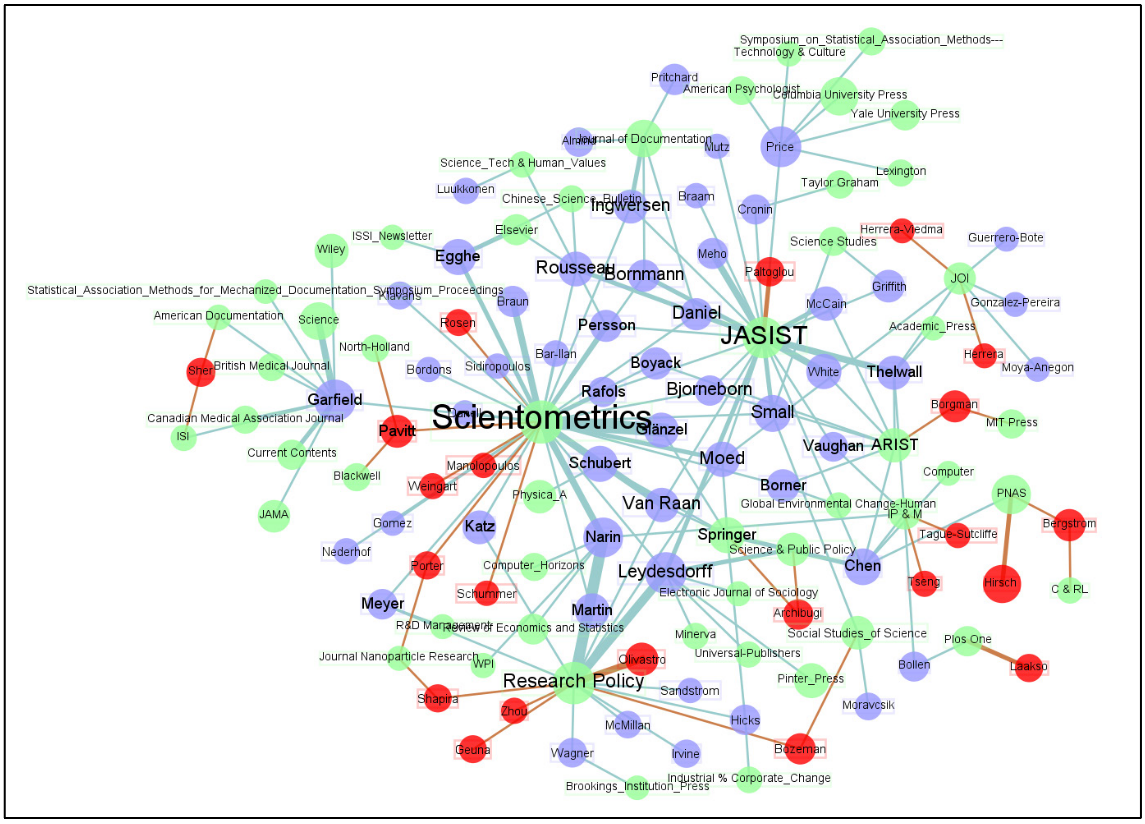


Figure 4. Network of the Bibliometrics discipline through the MADAP method in Google Scholar (author-journal)

Blue nodes: core authors; Red nodes: related authors; Green nodes: sources
N= 174 nodes (80 sources, 63 core authors, 31 related authors)
Map energysed by Noverlap algorithm with Gephi

The journals with a higher eigenvector centrality are Scientometrics, JASIST and Research Policy. Henk Moed, Loet Leydesdorff, and Anthony Van Raan are the most central specialist authors. Occasional authors (Pavitt, Porter, and Manolopoulos are those with a higher eigenvector centrality score) play a less central role although their influence is notable, especially in relation to some journals (e.g, Research Policy).

Although Figure 4 can reflect author-journal relationships, this map is less informative when it comes to describing sub-disciplines and research fronts. For this reason, an alternative map (Figure 5) has been generated showing author-keyword relationships. In this case, we consider the Top 100 highly cited specialist authors according to GSC public profiles (blue nodes), and all normalized research field keywords included in each of the author profiles (red nodes).

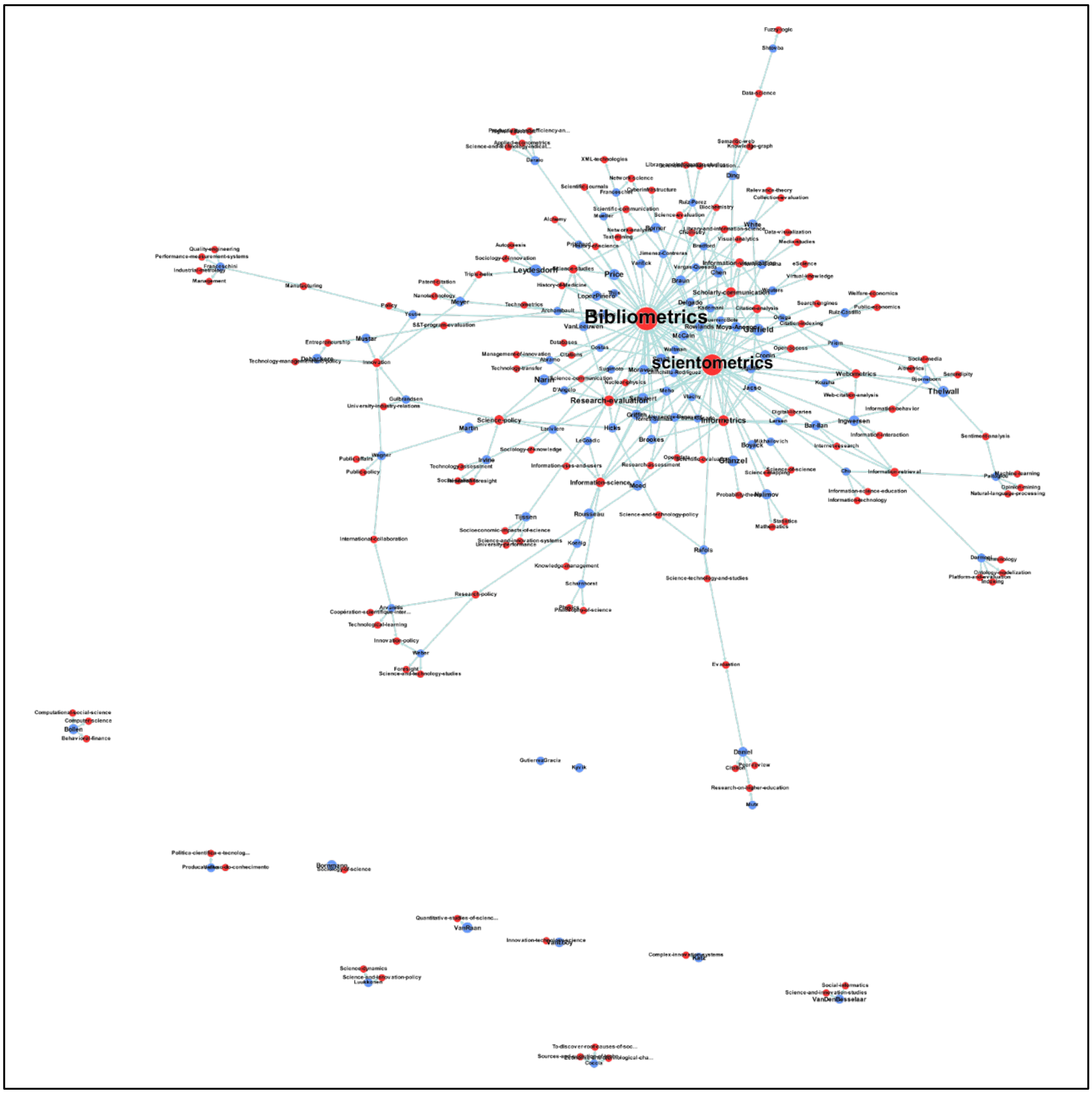


Figure 5. Network of the Bibliometrics discipline through the MADAP method in Google Scholar (author-keyword)

Blue nodes: core authors; Red nodes: topic keywords
N= 239 nodes (100 authors, 139 keywords).
Author node size: times cited; Keyword node size: number of authors sharing the keyword
Map energysed by Force Atlas algorithm with Gephi

This new map groups authors according to the keywords (main research interests) that they selected in their profile, showing sub-disciplinary relationships of the authors (Bibliometrics, Scientometrics, Webometrics, Research evaluation, Science policy, etc.), and at the same time identifying leaders in each front. Additionally, we can observe that some prominent authors with unusual field keywords (e.g., Van Raan or Bornman) are separated from the core, which shows the importance of using appropriate keywords for positioning authors among their peers and creating more accurate disciplinary maps.

5. Discussion

5.2. About the method (MADAP)

Projects of a bibliographic nature like this one can't ever reach perfection, and it is entirely possible that we may have missed relevant authors. The criteria for selecting the authors were two: first, the existence of a public GSC profile of the author on 24 July 2015 (when the data collection was made), and second, that the author works on the fields of Bibliometrics, Scientometrics, Informetrics, Webometrics, or Altmetrics. Hence, in order to avoid possible confusion, we stress that the ranking of authors (Table 2) was constructed exclusively from the set of 811 authors with a GSC public profile at the time of data collection.

We're well aware that these lists don't include all the researchers in the area. On the one hand some scholars have not created a profile, or they haven't made it public (this is the case of Leo Egghe, an essential figure in the discipline). We should note however that users can create and curate GSC profiles (private preferably) for any researcher, not only for themselves, which may help solving this coverage limitation. Using Harzing's Publish or Perish (PoP) (https://harzing.com/resources/publish-or-perish) in combination with CleanPoP (http://cleanpop.ifris.net) can be an alternative in the cases when a public profile is not available. In addition to this limitation, other scholars may have created a public profile but have included obscure or inadequate keywords to describe their research interests, thus making it impossible to find them using the more common keywords that we used in our approach. We tried to ameliorate this limitation by running the additional topic searches in Google Scholar.

Working with the top cited documents of the discipline – instead of only the authors with a public GSC profile – as the unit of analysis enabled us to capture all relevant authors (whether or not they had a public profile). Documents, journals and publishers rankings (Tables 3, 4, and 5) were constructed following this approach. However, this method requires using Google Scholar in addition to GSC, which adds complexity to the process, is time consuming, and requires a prior in-depth knowledge of the discipline under study. For example, in the case of the network presented in Figure 4 we only analysed the top 200 most cited documents because of these limitations.

Another important point of discussion is the one concerned with the accuracy of data provided by GSC. GSC feeds from GS, which is known to contain errors related both to citation and bibliographic data (recently summarized by Orduna-Malea et al. 2016). These errors are inherited by GSC. However, in GSC authors have the power to edit the bibliographic records and fix these errors. Although it is not likely that many researchers in general bother to do this, the composition of our sample (bibliometricians) makes us think that the data in this particular case might be of a slightly better quality than average. Of course, errors may persist in some profiles. Nevertheless, the manual cleaning process applied in this study prevents bibliographic errors from significantly affecting the general findings.

Another source of errors comes from profile manipulation. Metrics in GSC have been proved to be easily gamed by authors who want to boost their citation counts by abusing self-citations, or by uploading fake academic documents to the Web (Delgado López-Cózar, Robinson-García and Torres-Salinas 2014). Additionally, since GSC profiles can be set to be automatically populated by the system, they may sometimes contain documents that have not been actually authored by the researcher in question (and the researcher may even not be aware of this).

Regarding false citations (caused either by GSC malfunctions or manipulation), their effect in the results obtained in this study is considered to be low, especially on the top positions (the core intellectual map of the discipline). We would like to emphasize that the specific rank positions and metrics in the lists provided (authors, documents, journals, and publishers) should not be considered especially significant. It is the general shape of the discipline that is important. The purpose of this study was to reveal the main agents in the discipline according to the data available in GSC, not to generate micro-level research evaluations.

The main limitation of this method is that it is highly time-consuming. The process of searching, extracting, and cleaning bibliographic data from GS and GSC cannot be completely automated, and much manual labour is required. Carrying out discipline studies with other citation indexes such as Scopus or Web of Science is easier, because they provide more and better metadata. The difference, of course, is that while GS and GSC can be accessed for free, access to Scopus and Web of Science is subject to paying hefty subscription fees. Therefore, each platform presents a tradeoff: with Google Scholar it is possible to freely extract unrefined data. These data requires intensive human intervention to clean in order for it to be useful, which is costly in person-hours. On the other hand, with Scopus and Web of Science it is possible to carry out similar and even more detailed analyses in less time, providing that the necessary (and extremely high) subscription fees have been covered, which is costly in money. The decision of which source is more cost-effective will depend on the type of analyses that need to be carried out, but generally speaking, for small to medium-size projects, the cost of cleaning data extracted from Google Scholar should be several orders of magnitude lower than the subscription costs of the other citation indexes.

Limited time and the availability of just a small workforce are the main reasons why most of this analysis has focused on the most cited documents in the discipline (top 1,000 most cited documents). Thus, this specific analysis mainly presents information on the documents and researchers with the highest impact in the discipline. With more resources (people, time) the analysis could be expanded to cover a larger portion of the data, which would provide insight on the rest of the researchers and their publications. Nevertheless, the method described seems to be a very cost-effective way to accurately represent the structure of the discipline, specially suitable in the cases when accessing other subscription-based citation indexes is not an option.

The extensive coverage in Google Scholar (geographic, linguistic, document types…) is a clear advantage when it comes to developing discipline studies. Particularly, the inclusion of books (see Table 3 and 5) provides a wider vision of the discipline than the one offered by Scopus and Web of Science, where book coverage is merely testimonial (Martin-Martin et al. 2016). In our case, however, 10.5% of the top 200 most highly cited bibliometrics documents according to GS are books (mostly manuals describing techniques and procedures). These documents are not covered by WoS or Scopus.

This method could be used to analyse other disciplines and fields, although as noted before, an in-depth knowledge of the discipline under study may be necessary to identify and contextualize the results obtained. Obviously, the accuracy of the results depends on the level of uptake of the platform by researchers who work in the discipline. It has been reported that coverage of GSC at the discipline level can vary significantly (Ortega, 2015a).

Lastly, the data for this analysis was collected on 2015, and the results would undoubtedly be different if they were collected again now. However, this issue does not compromise the findings of the current study, which were to test the suitability of GSC and GS as sources of data to generate a comprehensive picture of the structure of a discipline, using the procedures previously described (MADAP method).

5.2. About the bibliometric actors (the discipline studied)

The accuracy of the method should be discussed not only from a technical/conceptual point of view but also from an empirical perspective. Therefore, we believe it is best to discuss the results obtained from applying the MADAP method to the Bibliometrics field from different points of view (authors, documents, journals, and book publishers).

Authors

The top cited authors in Bibliometrics according to GSC (Table 2) accurately represent the map of the discipline, including the founders of the discipline (Price and Garfield) as well as the most influential bibliometricians, almost all of them recipients of the Price medal, a prize that recognizes scientists who have exceptionally contributed with their work to the development of Bibliometrics.

On the one hand, Price, armed with the theoretical foundations laid by John Desmond Bernal and Robert K. Merton, set out to systematically apply quantitative techniques to the History and social studies of Science, developing the theoretical foundations of Scientometrics, born from the combination of the Sociology of science, History, Philosophy of science, and Information science. This approach is characterized by the analysis of the life and activity of Science and scientists from a quantitative perspective. The numbers were used to characterize the production of knowledge and scientists' lives: what they create and produce, to whom they relate to, the sources they used, and the impact and influence they provide/receive to/from other scientists, etc.

On the other hand, Garfield made possible that Bibliometrics became a reality (Bensman 2007; McCain 2010; Small 2017; Wouters 2017): the creation of the "citation index" made possible the quantification of scientific activity through its main output: the publications and citations they generate. Since then, citation analysis and all its variants have become the most widespread analysis technique of this new specialty. This is evidenced by the significant presence of highly cited documents that deal with this topic. Garfield defined the phenotype of the discipline: technology (the basis for the storage and circulation of information) is at the heart of all its tools.

As for the occasional authors of the discipline, these have been included solely as a matter of illustration. Obviously, many of the citations they have received belong to non-bibliometric publications. Nevertheless, the table reflects those important scholars who, despite belonging to other disciplines, provided important contributions to the field. This should be kept in mind when interpreting Table 2.

Lastly, the Bibliometrics map is useful to analyse the rest of the authors in the list: the Hungarian school (both Eastern Europe and Russia, like Nalimov), the Dutch school (with its various branches in Leiden and Amsterdam), the Belgian school (with Egghe and Rousseau), the North American School (Small, Griffith, and White), the Spanish school (with López Piñero, who introduced Price's work in Spain), and the new authors that represent the technological transformation of the discipline (mainly Thelwall).

Documents

The top documents in Bibliometrics according to Google Scholar Citations (Table 3) embody the main findings of the field. Among the top documents we can highlight those that first introduced new techniques and citation-based indicators, like the ones by Hirsch (3rd), Garfield (9th and 10th), Small (12th), and Egghe (23rd). Among them we find the most widely known indicator in Bibliometrics (the Impact Factor) and the one that has come to replace it while extending its capabilities (h-index).

The strong orientation of Bibliometrics towards evaluation in general and the assessment of the performance of individuals, journals, and institutions in particular, reveals a clear link between Bibliometrics and Science policy, and explains the use of the aforementioned indicators and other bibliometric tools by policymakers.

Additionally, this list is also a proof of the anomalous institutionalization process of the discipline. The main "bibliometric laws" which still hold true today where established at the dawn of the discipline, even before it was fully instituted (Lotka, Zipf, Bradford), and were developed by authors working outside the discipline. The same happened with the proposal of the h-index by Hirsch, elaborated by this physicist in his "leisure time". Bibliometrics is often revolutionized from outside Bibliometrics.

We can also distinguish the great relevance of some topics such as the "Triple Helix" by Leydersdorff, or the social networks by Barabási, which have had a strong impact outside the borders of our discipline.

Lastly, as we would expect, we can find among the most cited documents those texts that have served as textbooks for the discipline (written by Moed, Van Raan, Eghhe, Rousseau, etc.).

Journals

The top journals in Bibliometrics according to GSC (Table 4) illustrate in this case the main communication channels of the discipline.

Scientometrics is the journal with more articles published within the 1,000 most cited documents (284 articles). It is thus the most influential journal in the discipline. Its birth in 1978 was a milestone in the process of institutionalization of the discipline. The second place is occupied by JASIST (137 articles). This fact shows the important role of this journal in Bibliometrics, although its scope is broader. This journal has maintained since its inception a strong link between Information Science and Bibliometrics, though some authors have noticed a slight specialization towards Bibliometrics over time (Nicolaisen and Frandsen 2015). Journal of informetrics, focused exclusively on Bibliometrics, Scientometrics, Webometrics, and Altmetrics, appears in the fourth position (36 articles). The young age of this journal (it was created in 2007) explains why there isn’t a greater number of articles published in this journal among the most cited documents in the discipline.

The connection between Library and Information Science (LIS) and Bibliometrics is noticeable through the presence of other important LIS journals in the list, such as Journal of Documentation, Journal of Information Science, Library Trends, or Aslib Proceedings. This connection has been a matter of public record for a long time now (White and McCain 1998; Larivière, Sugimoto and Cronin 2012; Larivière 2012). Its connections with the field of web technologies from an information science perspective is strongly marked as well (Cybermetrics, Online Information Review). Additionally, we can see that journals oriented towards the Social Studies of Science (such as Research Policy, Social Studies of Science, and Science and Public Policy) also have strong ties to Bibliometrics.

If we analyse the number of citations instead of the number of articles published, we find the same first three journals occupying the first positions (Scientometrics, JASIST, and Research Policy), but the data also shows a great impact of articles published outside the core journals of the discipline, revealing the role of multidisciplinary journals. Science gets 9,219 citations from only 8 articles whereas PNAS gets 7,642 citations from 9 articles, and PLoS One gets 2,376 citations from 13 articles (the figures for Nature are lower, with 1,871 citations from 10 articles).

As regards the contributions published outside both the core and multidisciplinary journals (primarily bibliometric studies of specific fields published in the journals of the field), the MADAP method is able to capture both the documents and journals only if at least one of the co-authors of these manuscripts have been previously identified by the search and identification process (See section 3.1), and have created a GSC public profile. In this sense, the method does not exclude these contributions by default.

Book publishers

In this case, output is low (the first position is occupied by Springer, with only 10 documents positioned within the set of highly cited documents), although we observe that all publishers achieve high numbers of citations per document (Springer receives 5,766 citations to 10 documents). Also remarkable is the performance of university presses in the dissemination of bibliometric research results (such as the University of Chicago, Columbia, Yale or MIT), with a very low presence in terms of productivity but an impressive impact in the number of citations. The ability to attract well-established authors in order to publish specialized books makes a great difference in book publisher rankings.

6. Conclusions

By virtue of the results obtained, the research question (RQ1) can be answered positively. GSC (in combination with Google Scholar) is able to provide a precise and accurate picture of the Bibliometrics community. Moreover, the data collected, not only at the author-level but also at the document-level and source-level, clearly responds to our mental image of the field. That is, it is possible to identify the most influential authors (both specialists and occasional researchers), documents (articles and books) and sources (journals and publishers) in the discipline using data from GSC. Therefore, the MADAP method has been proved not only feasible but also accurate and valid (RQ2).

The application of the procedures followed in this work (the MADAP method) to study other fields and disciplines through GSC challenges new research on this front.

Acknowledgements

Funding was provided by Ministerio de Educación, Cultura y Deporte (FPU2013/05863), Universitat Politècnica de València (PAID-10-14).

References

Aguillo, Isidro F. (2012). Is Google Scholar useful for bibliometrics? A webometric analysis. *Scientometrics*, *91*(2), 343-351.

Almind, T. C. & Ingwersen, P. (1997). Informetric analyses on the world wide web: methodological approaches to ‘webometrics’. *Journal of documentation*, *53*(4), 404-426.

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century—A review. *Journal of informetrics*, *2*(1), 1-52.

Bar-Ilan, J., Haustein, S., Peters, I., Priem, J., Shema, H. & Terliesner, J. (2012). Beyond citations: Scholars' visibility on the social Web. In É. Archambault, Y. Gingras, & V. Larivière (Eds.). *Proceedings of the International Conference on Science and Technology Indicators (STI 2012)* (pp. 99-109).

Becher, T. & Trowler, P. (2001). *Academic tribes and territories: Intellectual enquiry and the culture of disciplines*. UK: McGraw-Hill Education.

Bensman, S. J. (2007). Garfield and the impact factor. *Annual Review of Information Science and Technology*, *41*(1), 93-155.

Björneborn, L. & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American society for information science and technology*, *55*(14), 1216-1227.

Bonitz, M. (1982). Scientometrie, Bibliometrie, Informetrie. *Zentralblatt für Bibliothekswesen*, *96*(2), 19-24.

Bornmann, L., Thor, A., Marx, W. & Schier, H. (2016). The application of bibliometrics to research evaluation in the humanities and social sciences: An exploratory study using normalized Google Scholar data for the publications of a research institute. *Journal of the Association for Information Science and Technology*, *67*(11), 2778-2789.

Borgman, C. L. & Furner, J. (2002). Scholarly Communication and Bibliometrics. *Annual Review of Information Science and Technology*, *36*, 3-72.

Braun, T. (1994) (ed.) *Little scientometrics, big scientometrics… and beyond?*, *Scientometrics*, *30*(2-3), 373-537.

Broadus, R.N. (1987a). Early approaches to bibliometrics. *Journal of the American Society for Information Science*, *38*(2), 127-129.

Broadus, R. N. (1987b). Toward a definition of ‘bibliometrics’. *Scientometrics*, *12*(5-6), 373-379.

Brookes, B. C. (1988). Comments on the scope of bibliometrics. In: L. Egghe, R. Rousseau (Eds). *Informetrics 87/88. Select Proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval* (pp. 29-41). Amsterdam, Elsevier Science.

Brookes, B. C. (1990). Biblio-, Sciento-, Infor-metrics??? What are we talking about?”. In: L. Egghe, R. Rousseau (Eds). *Informetrics 89/90. Selection of Papers Submitted for the Second International Conference on Bibliometrics, Scientometrics and Informetrics* (pp. 31-43). Amsterdam, Netherlands: Elsevier.

Cronin, B. (2001). Bibliometrics and beyond: some thoughts on web-based citation analysis. *Journal of Information science*, *27*(1), 1-7.

De Bellis, N. (2009). *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Maryland: Scarecrow Press.

Delgado López-Cózar, E., Robinson-García, N. & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446-454.

Egghe, L. & Rousseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Netherlands: Elsevier.

Garfield, E. (1970). Citation indexing for studying science. *Nature*, *227*(5259), 669-671.

Glänzel, W. & Schoepflin, U. (1994). Little scientometrics, big scientometrics … and beyond?, *Scientometrics*, *30*(2-3), 375-384.

Godin, B. (2006). On the origins of bibliometrics. *Scientometrics*, *68*(1), 109-133.

Gorbea Portal, S. (1994). Principios teóricos y metodológicos de los estudios métricos de la información. *Investigación Bibliotecológica*, *8*, 23-32.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H. & Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics*, *101*(2), 1145-1163.

Hertzel, D.H. (1987). History of the development of ideas in bibliometrics. In: A. Kent, (Ed.). *Encyclopedia of library and information sciences* (pp. 144-219). Marcel Dekker, New York, 144–219.

Hood, W. & Wilson, C. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, *52*(2), 291-314.

Jacsó, P. (2005). Google Scholar: the pros and the cons. *Online information review*, *29*(2), 208-214.

Jacsó, P. (2008). Google scholar revisited. *Online information review*, *32*(1), 102-114.

Jacsó, P. (2012). Google Scholar Author Citation Tracker: is it too little, too late?. *Online Information Review*, *36*(1), 126-141.

Kramer, B. & Bosman, J. (2015). *101 Innovations in Scholarly Communication - the Changing Research Workflow*. Available at: https://101innovations.wordpress.com

Larivière, V. (2012). The decade of metrics? Examining the evolution of metrics within and outside LIS. *Bulletin of the American Society for Information Science and Technology*, *38*(6), 12-17.

Larivière, V., Sugimoto, C. & Cronin, B. (2012). A bibliometric chronicling of library and information science's first hundred years. *Journal of the American Society for Information Science and Technology*, *63*(5), 997-1016.

Lawani, S. M. (1981). Bibliometrics: its theoretical foundations, methods and applications. *Libri*, *31*(1), 294-315.

Lin, J. & Fenner, M. (2013). Altmetrics in evolution: Defining and redefining the ontology of article-level metrics. *Information standards quarterly*, *25*(2), 20-26.

Martín-Martín, A., Orduna-Malea, E., Ayllón, Juan M. & Delgado López-Cózar, E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentación Científica*, *39*(4), e149.

Martín-Martín, A., Orduna-Malea, E., Harzing, A. W., & Delgado López-Cózar, E. (2017). Can we useGoogle Scholar to identify highly-cited documents? Journal of Informetrics, 11(1), 152–163.

McCain, K. W. (2010). The view from Garfield's shoulders: Tri-citation mapping of Eugene Garfield's citation image over three successive decades. *Annals of Library and Information Studies*, *57*, 261-270.

Meho, L. I. & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the Association for Information Science and Technology*, *58*(13), 2105-2125.

Mikki, S., Zygmuntowska, M., Gjesdal, Ø. L. & Al Ruwehy, H. A. (2015). Digital Presence of Norwegian Scholars on Academic Network Sites—Where and Who Are They?'. *PloS ONE*, *10*(11), e0142709.

Milojević, S. & Leydesdorff, L. (2013). Information metrics (iMetrics): a research specialty with a socio-cognitive identity?. *Scientometrics*, *95*(1), 141-157.

Mingers, J. & Meyer, M. (2017). Normalizing Google Scholar data for use in research evaluation. *Scientometrics*, *112*(2), 1111-1121.

Mingers, J., O'Hanley, J. & Okunola, M. (in press). Using Google Scholar institutional level data to evaluate the quality of university research. *Scientometrics*. https://doi.org/10.1007/s11192-017-2532-6

Narin, F. & Moll, J.K. (1977). "Bibliometrics". *Annual Review of Information Science and Technology*, *12*, 35-58.

Nicolaisen, J. & Frandsen, T. F. (2015). Bibliometric evolution: Is the journal of the association for information science and technology transforming into a specialty Journal?. *Journal of the Association for Information Science and Technology*, *66*(5), 1082-1085.

Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A. & Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, *104*(3), 931-949.

Orduna-Malea, E., Martín-Martín, A., Ayllón, Juan M. & Delgado López-Cózar, E. (2016). *La revolución Google Scholar: destapando la caja de Pandora académica*. Granada: UNE.

Orduna-Malea, E., Martín-Martín, A. & Delgado-López-Cózar, E. (2016). The next bibliometrics: ALMetrics (Author Level Metrics) and the multiple faces of author impact. *El profesional de la información*, *25*(3), 485-496.

Ortega, Jose L. (2015a). How is an academic social site populated? A demographic study of Google Scholar Citations population. *Scientometrics*, *104*(1), 1-18.

Ortega, Jose L. (2015b). Relationship between altmetric and bibliometric indicators across academic social sites: The case of CSIC's members. *Journal of Informetrics*, *9*(1), 39-49.

Ortega, Jose L. & Aguillo, Isidro F. (2012). Science is all in the eye of the beholder: Keyword maps in Google Scholar Citations. *Journal of the American Society for Information Science and Technology*, *63*(12), 2370-2377.

Ortega, Jose L. and Aguillo, Isidro F. (2013). Institutional and country collaboration in an online service of scientific profiles: Google Scholar Citations. *Journal of Informetrics*, *7*(2), 394-403.

Ortega, Jose L. & Aguillo, Isidro F. (2014). Microsoft academic search and google scholar citations: Comparative analysis of author profiles. *Journal of the Association for Information Science and Technology*, *65*(6), 1149-1156.

Otlet, P. (1934). *Traité de documentation: le livre sur le livre, théorie et pratique*. Brussels: Editiones Mundaneum,

Peritz, B. (1984). On the careers of terminologies; the case of bibliometrics. *Libri*, *34*(1), 233-242.

Priem, J. & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, *15*(7).

Prins, A.A.M., Costas, R., Van Leeuwen, T. N. & Wouters, Paul F. (2016). Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation*, *25*(3), 264-270.

Ranganathan, S. R. (1969). Librametry and its scope'. In *Documentation Research and Training Centre: Annual Seminar (7) (1969): Subject analysis for document finding systems: Quantification and librametric studies: Management of translation service* (pp. 285-301). Bangalore: Documentation Research and Training Centre.

Sengupta, I.N. (1992). Bibliometrics, informetrics, scientometrics and librametrics: an overview. *Libri*, *42*(2), 75-98.

Shapiro, Fred R. (1992). Origins of Bibliometrics, Citation Indexing, and Citation Analysis: The Neglected Legal Literature. *Journal of the American Society for Information Science*, *43*(5), 337-339.

Small, H. (2017). A tribute to Eugene Garfield: Information innovator and idealist. *Journal of Informetrics*, *1*(3), 599-612.

Stuart, D. (2014). *Web metrics for library and information professionals*. London: Facet publishing.

Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information processing & management*, *28*(1), 1-3.

Thelwall, M. (2008). Bibliometrics to webometrics. *Journal of Information Science*, *34*(4), 605-621.

Thelwall, M. (2009). *Introduction to webometrics: Quantitative web research for the social sciences*. San Diego: Morgan & Claypool.

Van Raan, A. (1997). Scientometrics: State-of-the-art. *Scientometrics*, *38*(1), 205-218.

Thelwall, M., Vaughan, L. & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology*, *39*(1), 81-135.

White, H. D. & McCain, K. W. (1998). Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972–1995. *Journal of the American Society for Information Science*, *49*(4), 327-355.

White, H.D. & McCain, K.W. (1989). Bibliometrics. *Annual review of information science and technology*, *24*, 119-186.

Whitley, R. (1984). *The intellectual and social organization of the sciences*. Oxford: Oxford University Press.

Wilson, C.S. (1999). Informetrics. *Annual Review of Information Science and Technology*, *34*. 107-247.

Wouters, P. (2017). Eugene Garfield (1925-2017). *Nature*, *543*(7642), 492.

Chapter 13. Author-level metrics in the new academic profile platforms: The online behaviour of the Bibliometrics community

Journal article. Cite as:

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). Author-level metrics in the new academic profile platforms: The online behaviour of the Bibliometrics community. Journal of Informetrics, 12(2), 494–509. https://doi.org/10.1016/j.joi.2018.04.001

Abstract (English)

The new web-based academic communication platforms do not only enable researchers to better advertise their academic outputs, making them more visible than ever before, but they also provide a wide supply of metrics to help authors better understand the impact their work is making. This study has three objectives: a) to analyse the uptake of some of the most popular platforms (Google Scholar Citations, ResearcherID, ResearchGate, Mendeley and Twitter) by a specific scientific community (bibliometrics, scientometrics, informetrics, webometrics, and altmetrics); b) to compare the metrics available from each platform; and c) to determine the meaning of all these new metrics. To do this, the data available in these platforms about a sample of 811 authors (researchers in bibliometrics for whom a public profile Google Scholar Citations was found) were extracted. A total of 31 metrics were analysed. The results show that a high number of the analysed researchers only had a profile in Google Scholar Citations (159), or only in Google Scholar Citations and ResearchGate (142). Lastly, we find two kinds of metrics of online impact. First, metrics related to connectivity (followers), and second, all metrics associated to academic impact. This second group can further be divided into usage metrics (reads, views), and citation metrics. The results suggest that Google Scholar Citations is the source that provides more comprehensive citation-related data, whereas Twitter stands out in connectivity-related metrics.

Abstract (Spanish)

Las nuevas plataformas de comunicación en la web no solo facilitan que los investigadores puedan realizar una mejor promoción de sus publicaciones, haciéndolas más visibles que nunca, sino que también proporcionan un amplio abanico de indicadores que permiten a los autores entender mejor cómo su trabajo está teniendo impacto. Este trabajo tiene tres objetivos: a) analizar el nivel de adopción de algunas de las plataformas académicas más populares (Google Scholar Citations, ResearcherID, ResearchGate, Mendeley, and Twitter) en una comunidad científica específica (bibliometría, cienciometría, informetría, webometría, y altmetría), b) comparar los indicadores disponibles en cada plataforma, y c) determinar el significado de todos estos indicadores. Se analizan un total de 31 indicadores. Los resultados muestran que un alto número de los investigadores analizados solo tenían un perfil en Google Scholar Citations (159), o solamente en Google Scholar Citations y ResearchGate (142). Finalmente, encontramos dos tipos de indicadores académicos. Primero, las métricas relacionadas con conectividad (seguidores), y segundo, las métricas asociadas al impacto. Este segundo grupo puede subdividirse en métricas de uso (lecturas, visualizaciones), y métricas de citas. Los resultados sugieren que Google Scholar Citations es la fuente que proporciona los datos de citas más exhaustivos, mientras que Twitter sobresale por sus métricas relacionadas con conectividad.

1. Introduction

Last decade has witnessed the emergence of a plethora of new communication channels and social collaboration platforms where academic outputs are susceptible of being indexed, searched, located, read, and mentioned (Priem & Hemminger, 2010; Piwowar, 2013). The degree to which these online platforms are used (as well as the metrics they offer) provide new insights about the current dynamics of research activity, not without introducing some methodological concerns (Bornmann, 2014; 2016; Sugimoto et al., 2017).

1.1. Online academic communication channels: from personal websites to academic profiles

Academic personal websites are probably the first venues where scholars started to disseminate their personal information, current activities and projects, and their lists of academic contributions. Scientists' personal websites have been extensively studied from a webometric approach (Barjak, Li & Thelwall, 2007; Mas-Bleda & Aguillo, 2013; Más-Bleda et al, 2014). The literature on this issue describes significant differences in web visibility according to disciplines, countries, gender, and age. Additionally, several studies find an overall low presence (number of researchers with personal website) and lack of essential information in these websites (Chen et al, 2009; Mas-Bleda & Aguillo, 2013).

However, despite their initial important role as the first venues where researchers could make their work available to others, personal websites did not allow much interaction between researchers, nor were they suitable tools to record and provide metrics about the authors' academic contributions. Online academic profile platforms are now filling this gap. These online environments usually supply a variety of metrics that capture the diversity of actions and interactions that can occur amongst scientists in the digital space (Haustein, 2016), actions that already contribute to reshape the scholarly reputation of authors (Jamali, Nicholas & Herman, 2015).

Naturally, the features available in these platforms vary slightly from one to another (Jordan, 2014a), but they usually provide researchers with the choice to create an academic profile, and upload their research outputs (not only published materials, but also posters, presentations, software, and other kinds of unpublished materials). These outputs can then be accessed by other researchers, who can download them or comment them. Researchers can interact in other ways with these platforms (tagging and following profiles, asking and answering questions). The most popular academic platforms that provide social features and author profiles are ResearchGate (Kadriu, 2013; Thelwall & Kousha, 2015; Nicholas, Clark & Herman, 2016), Academia.edu (Thelwall & Kousha, 2014), and Mendeley (Li, Thelwall & Giustini, 2011; Mohammadi & Thelwall, 2014). Additionally, academic databases have also developed platforms that enable authors to create profiles to list their publications, based on the coverage of each platform. Among these, we can find ResearcherID (Web of Science), Scopus ID (Scopus), and Google Scholar Citations (GSC) (Google Scholar).

The increasing use of all these social platforms by scholars was reported in the results of a survey carried out by the Nature Publishing Group (Van Noorden, 2014). The data collected in this survey was also made openly available (Nature Research, 2014). Jordan (2014b) re-used this dataset and found significant differences in the perceived usefulness of social network platforms depending on the respondents. Of the 480 researchers in the Humanities, Arts, and Social Sciences who responded, more than 70% declared that they were aware of Google Scholar (either the search engine or the profile service) and visited it regularly. This figure decreased to 61% for the respondents from Science and Engineering (n ~ 3,000).

Mas-Bleda et al. (2014) analysed the presence of 1,517 highly-cited European researchers in several platforms (GSC, Microsoft Academic Search, Mendeley, Academia.edu, Linkedin, and SlideShare). She found that the use of online academic profile services by these top-cited researchers was still low (only 9% of the researchers in the sample had a public profile in GSC). Additionally, this study also reported high inter-disciplinary differences (24% of the social scientists had a profile created in GSC; 8% in Mendeley). This study did not consider ResearchGate due to the low number of researchers in that platform at the time. However, ResearchGate later proved to be the most used academic profile platform according to a survey about the use of academic communication tools carried out by Kramer and Bosman (2016). According to this survey, which was responded by over 20,000 people related to academia (mostly researchers, but also librarians, publishers, and people from the industry and government), the most used academic profile platform overall was ResearchGate (66%), followed by GSC (62%).

1.2. Online academic metrics: from citations to author-level metrics

Since the classic study by Bollen et al. (2009), where the data came primarily from usage logs provided by publishers, many papers have been published on the nature of online article-level metrics. Some of these works intended to shed light on the correlation between traditional citation-based metrics and the flourishing array of altmetrics across disciplines and platforms (Priem, Piwowar & Hemminger, 2012; Thelwall et al, 2013; Costas, Zahedi & Wouters, 2015; De Winter, 2015). Nevertheless, prudence has been advised when interpreting the meaning of these correlations (Thelwall, 2016).

Similar studies, but focused on author-level metrics, instead of article-level metrics, are not as frequent (Bar-Ilan et al, 2012; Wildgaard, Schneider & Larsen, 2014; Orduna-Malea, Martín-Martín & Delgado López-Cózar, 2016a). The few studies on this front have found that the author-level metrics in a given platform tend to correlate with one another. This may be related to the claim that authors primarily create these profiles to advertise themselves and not to collaborate, in line with the Diogenes Club analogy proposed by Ortega (2016a). On the other hand, reports of high correlations between metrics across different platforms have been scarce to date.

Some earlier studies have addressed the similarity of the metrics reported by GSC to those displayed by other platforms. Ortega (2015b) analysed researchers working at the Spanish National Research Council (CSIC), and Mikki et al. (2015) chose to study researchers at the University of Bergen. Despite using very different samples, the results of these two studies agree on the low correlations found between citation-based indicators and altmetrics (based on social interactions) at the level of authors. Nevertheless, these studies were limited to authors from two particular institutions.

1.3. Towards a disciplinary study of author-level-metrics

At the discipline level, Bar-Ilan et al. (2012) sampled 57 presenters at the 2010 Leiden STI Conference (an event that is mainly devoted to discuss issues in the field of bibliometrics and related areas), collecting publication and citation counts from a variety of web platforms. The authors found that 70% of presenters had a LinkedIn account, 23% of them had public GSC profiles, and 16% of them were on Twitter, as of 2012. Later, Haustein et al. (2014) studied the same sample of authors, finding that 58% (33) of the authors had a profile in ResearchGate, and 53% (30) had a public profile in GSC.

However, an exhaustive multi-platform study that analyses an entire academic discipline has not been carried out yet. It is reasonable to assume that, given that each platform has its own specific userbase and document coverage (which inevitably affects the metrics provided by the platform), none of these platforms, by themselves, are able to provide a complete and accurate portrayal of an author's impact. For this reason, we believe a multi-platform approach is necessary if the goal is to collect all the available evidence of an author's impact.

By working with a homogeneous sample (researchers that work on the same discipline), we believe we might find existing patterns regarding the use of public academic profiles, and shed light on the question of the meaning of the metrics that are displayed in these profiles.

1.4. Research questions

The main goals of this study are to analyse the uptake of some of the most popular platforms (Google Scholar Citations, ResearcherID, ResearchGate, Mendeley and Twitter) by a specific scientific community (bibliometrics, scientometrics, informetrics, webometrics, and altmetrics), to compare the metrics available from each platform, and to determine the meaning of all these new metrics. The reason behind the selection of this discipline is that this discipline is the one the authors know best, and expertise in the field is in this case necessary to find patterns that arise from the metrics that will be analysed.

The following research questions (RQ) are proposed:

(RQ1): Is there a large enough volume of data at the author-level (author-level metrics) to allow the development of bibliometric analysis? Are there significant differences in the quantity of data available across platforms?

(RQ2): Which are the main dimensions of online impact captured by author-level metrics? Is there any significant correlation between the different author-level metrics offered by the social platforms analysed when considering authors that belong to one specific academic discipline?

2. Methods

The construction of the data sample followed the MADAP (Multifaceted Analysis of Disciplines through Academic Profiles) method (Martín-Martín, Orduna-Malea & Delgado López-Cózar, 2018), which consisted basically of four steps: identification of authors, location of academic profiles for those authors, extraction of author-level metrics from each platform, and, lastly, statistical data analysis.

2.1. Identification of authors

We first established two required criteria to include authors in the sample:

a) Authors who have published in the areas of bibliometrics, scientometrics, informetrics, webometrics or altmetrics, and
b) Authors who have created a GSC public profile.

Regarding the first criterion, authors from all these subfields were included, because all of them are related, to a greater or lesser extent, with the quantitative studies of science. The second criterion was established mainly because a previous study (Haustein et al., 2014) had established that GSC was the most popular platform among researchers in this discipline. Bosman and Kramer (2016) confirmed that GSC and ResearchGate were the most widely used academic profile platforms.

In order to identify the corpus of authors, three complementary procedures were carried out, described below.

All queries and data extraction were performed on July 24th 2015.

Method 1: topical keywords

We obtained a list of frequently-used descriptive terms in the discipline. To do this, the bibliographic records from all indexed articles published in an initial seed of five core journals exclusively devoted to the discipline were automatically retrieved using the Web of Science and Scopus. This set of sources was composed by the Journal of Informetrics, Research Evaluation and Scientometrics (the three journals with a declared scope mainly devoted to the field with higher impact factor in the Journal Citation Reports, 2016 edition). In addition to this, Cybermetrics (in order to cover the Webometrics subfield) and ISSI Conference Proceedings (in order to cover the major worldwide conference on the topic) were also included as initial sources seeds.

Next, all significant terms from the documents' titles and keywords (when available) were extracted and analysed to find out their frequency of use. This vocabulary was later cleaned, merging variants (for example, scientometric and scientometrics), deleting duplicates, and excluding generic terms that did not unequivocally describe the discipline or its main topics (e.g., credit, item, program, content, etc.). All the profiles in GSC that included at least one of these descriptive keywords (see Table 1) was included in the sample. Keyword variants (such as misspelled words, the same terms in other languages, etc.) were also considered. Terms not included as GSC keywords in any profile were excluded.

Method 2: Additional searches in Google Scholar

In order to capture authors relevant to the discipline but who have not included any of the relevant terms selected in method 1 or do not belong to any of the institutions mentioned in method 2, we also conducted a series of searches in Google Scholar (the search engine) with the intention to find these stragglers. Two types of searches were carried out: a) keyword searches using the terms obtained in method 1 (18 queries in total); journal name searches, to find obtain the documents published in the journals mentioned in method 1 (5 queries in total). The keyword searches were limited so that only documents with those terms in the titles were retrieved (to increase precision).

All the results to these queries (a maximum of 1,000 results per query, as per Google Scholar's limitations) were extracted. Then, although the information provided by GS is not always complete, we extracted all the instances of profile URLs that are displayed in the list of authors of each article (see Figure 1). All the profiles that had already been identified in previous steps were removed. Given how Google Scholar ranks results, which is basically according to the number of citations a document has received (Martín-Martin et al, 2017), we are reasonably confident that extracting the first 1,000 results of each query is enough to identify most of the relevant authors we might have been missing in the previous steps.

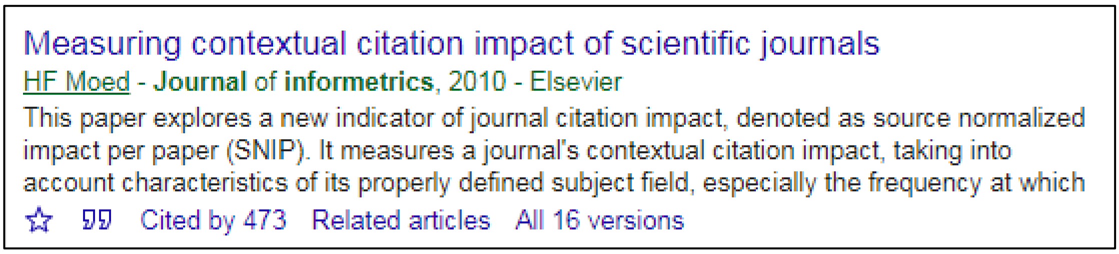

Measuring contextual citation impact of scientific journals
HF Moed - Journal of informetrics, 2010 - Elsevier
This paper explores a new indicator of journal citation impact, denoted as source normalized impact per paper (SNIP). It measures a journal's contextual citation impact, taking into account characteristics of its properly defined subject field, especially the frequency at which

Figure 1. Example of a Google Scholar bibliographic record.

Table 1. List of terms used as author keywords in GSC profiles representing the field of bibliometrics.

Topic Keywords	
Altmetrics	Research Assessment
Bibliometrics	Research Evaluation
Citation Analysis	Research Policy
Citation Count	Science and Technology Policy
H-Index	Science Evaluation
Impact Factor	Science Policy
Informetrics	Science Studies
Patent Citation	Scientometrics
Quantitative Studies of Science and Technology	Webometrics

Method 3: Institutional affiliation

As a supplementary search strategy, we considered searching authors affiliated to centres or departments that produce research in the area of bibliometrics, regardless of the research keywords used by authors in their public profiles. To do this, we obtained the list of researchers from the websites of these centers and searched them manually. The research centres we considered were: CWTS (Centre for Science and Technology Studies) (cwts.leidenuniv.nl; cwts.nl), Cybermetrics Lab (webometrics.info; ipp.csic.es), DZHW (Deutsches Zentrum für Wissenschafts- und Hochschulforschung) (dzhw.eu), ECOOM (Expertisecentrum Onderzoek en Ontwikkelingsmonitoring) (ecoom.be), EC3 Research Group: Evaluación de la Ciencia y de la Comunicación Científica (ec3.ugr.es), Science-Metrix (science-metrix.com), Scimago (scimagojr.com; scimagolab.com), SciTech Strategies (scitechstrategies.com), Statistical Cybermetrics Research Group (cybermetrics.wlv.ac.uk). However, this method did not add any additional authors to our sample, and therefore, the list of centers were not exhaustively expanded and this method was not ultimately used. However, it may be useful as a complement search in other fields.

Since GSC gives authors complete control over how their profile is set (personal information, institutional affiliation, research interests, as well as their scientific production), a systematic manual revision was carried out in order to:

- Detect false positives: authors whose scientific production doesn't have anything to do with this discipline, even though they labelled themselves with one or more of the keywords associated with it.

- Classify authors in two categories:

  c) *Specialists*: authors whose scientific production substantially falls within the field of bibliometrics.
  d) *Occasional*: authors who have sporadically published bibliometric studies, or whose field of expertise is closely related to Scientometrics (social, political, and economic studies about science), and therefore they can't be strictly considered bibliometricians.

We decided to consider as specialists those who meet the following criterion: at least half of the documents which contribute to their h-index in their GSC profile should fall within the limits of the field of bibliometrics.

To help delimiting the limits of the field, we considered not only the titles of the documents but also the venue where they had been published. To do this, we first defined a Bradford-like core of journals about bibliometrics (Scientometrics, Journal of Informetrics, JASIST, Research Evaluation, Research Policy, and Cybermetrics), followed by other LIS journals which also publish numerous bibliometric studies (Journal of Information Science, Information Processing & Management, Journal of Documentation, College Research & Libraries, Library Trends, Online Information Review, Revista Española de Documentación Científica, Aslib Proceedings, and El Profesional de la Información). Lastly, journals devoted to social and political studies about science (Social Studies of Science, Science and Public Policy, Minerva, Journal of Health Services Research Policy, Technological Forecasting and Social Change, Science Technology Human Values, Environmental Science Policy, and Current Science) were also searched. The selection of these journals was mainly based on our expert judgement, which also matches to a large degree the empirical results obtained by Hood and Wilson (2001), who analysed the bibliometrics literature and reported a list of the most productive journals in the field of bibliometrics.

In the end, we gathered a total of 811 GSC profiles, out of which the 48.8% (396) were classified as specialists, and the remaining 51.2% (415) as occasional authors.

2.2. Searching in other platforms

In addition to GSC, the academic profile services we considered in this analysis were ResearcherID[69], ResearchGate[70], and Mendeley[71]. Other academic profile services were not considered due to several reasons. Preliminary explorations showed that there was a very low coverage of authors in the discipline in platforms like Academia.edu and Loop. AMiner was discarded because it was found to be outdated, and Microsoft Academic was still in beta when this study was carried out.

We decided to include Twitter[72] because, although this platform is not designed to set up academic profiles, participation in this platform affects the level of dissemination of research papers (Ortega, 2016b), thus capturing an important dimension of an author's online visibility.

2.3. Obtaining the metrics

For each of the 811 authors in our sample, we manually checked whether they had also created profiles in ResearcherID, ResearchGate, Mendeley, and Twitter, by searching their names in each of the profile services' search features, and by searching their names in Google in combination with the name of the profile platform. In the case of Twitter, additional searches through well identified authors' followers were used in order to find authors whose profile names did not correspond with their personal names.

When a profile was found, all available author-level metrics were extracted. Custom automated parsers were developed for this purpose. The data collection in these platforms was carried out between the 4th and the 10th of September, 2015.

A total of 31 author-level metrics were extracted from GSC and the rest of profile platforms. These were the metrics that were available in each platform at the time of data collection. Some platforms might now offer new metrics, or they might have stopped displaying some of the metrics we discuss in this analysis. Their scope and definition can be found in Table 2. Additionally, we categorize each metric according to its nature: total (size-dependent; 22 metrics), average (size-independent; 3 metrics), and hybrid (composite indicator; 6 metrics).

Table 2. List of Author-Level metrics.

PLATFORM	METRIC	DEFINITION	CATEGORY
GSC	Citations	Number of citations to all publications. Computed for citations from all years, and citations received since 2010	Total
	h-index	The largest number h such that h publications have at least h citations. Computed for citations from all years, and citations received since 2010	Hybrid
	i10 index	Number of publications with at least 10 citations. Computed for citations from all years, and citations received since 2010	Total
RESEARCHER ID	Total Articles	Number of items in the publication list	Total
	Articles with Citation Data	Only articles added from *Web of Science Core Collection* can be used to generate citation metrics, even though the publication list may contain articles from other sources. This value indicates how many articles from the publication list were used to generate the metrics	Total

[69] https://www.researcherid.com
[70] https://www.researchgate.net
[71] https://www.mendeley.com
[72] https://twitter.com

PLATFORM	METRIC	DEFINITION	CATEGORY
	Sum of Times Cited	Total number of citations to any of the items in the publication list from *Web of Science Core Collection*. The number of citing articles may be smaller than the sum of the times cited because an article may cite more than one item in the set of search results	Total
	Average Citations per Item	Average number of citing articles for all items in the publication list from *Web of Science Core Collection*. It is the sum of the times cited divided by the number of articles used to generate the metrics	Average
	h-index	An author has a h-index of “h” when “h” of its articles has achieved at least “h” citations.	Hybrid
RESEARCH GATE	RG Score	It is a composite indicator that according to RG measures scientific reputation based on how an author’s research is received by his/her peers. The exact method to calculate this metric has not been made public, but it takes into account how many times the contributions (papers, data, etc.) an author uploads to *ResearchGate* are visited and downloaded, and also by whom (reputation)	Hybrid
	Publications	Total number of publications an author has added to his/her profile in *ResearchGate* (full-text or no)	Total
	Views	Total number of times an author’s contributions to *ResearchGate* have been visualized. This was later combined with the “Downloads” metric to form the new “Reads” indicator, but the data collection for this product was made before this change came into effect	Total
	Downloads	Total number of times an author’s contributions to *ResearchGate* have been downloaded. This metric was later combined with the “Views” indicator to form the new “Reads” indicator, but the data collection for this product was made before this change came into effect	Total
	Citations	Total number of citations to the documents uploaded to the profile.	Total
	Impact Points	Sum of the JCR impact factors of the journals where the author has published articles. This metric is no longer available in public RG profiles.	Hybrid
	Profile views	Number of times the author’s profile has been visited. This indicator is no longer public. Currently, users can only see their own profile views count, but not other users’.	Total
	Following	Number of *ResearchGate* users the author follows (friends)	Total
	Followers	Number of *ResearchGate* users who follow the author	Total
MENDELEY	Readers	This number represents the total number of times a *Mendeley* user has added a document by this author to his/her personal library	Total
	Publications	Number of publications the author has uploaded to *Mendeley* and classified as “My Publications”	Total

PLATFORM	METRIC	DEFINITION	CATEGORY
	Readers per document	Number of Readers divided by the number of publications per each author	Average
	Followers	Number of *Mendeley* users who follow the author in *Mendeley*	Total
	Following	Number of *Mendeley* users the author follows in *Mendeley*	Total
TWITTER	Tweets	Total number of tweets an author has published according to his/her profile	Total
	Followers	Number of *Twitter* users who follow the tweets published by the author	Total
	Following	Number of *Twitter* users the author follows	Total
	Days registered	Number of days since the author created his/her account on *Twitter*	Total
	Sum Retweets	Number of Retweets received for the author. These data was extracted using the software Webometric Analyst (Statistical Cybermetrics Research Group, 2011), and is limited to the data that the Twitter API allowed us to extract, meaning that it was not possible to extract all the tweets from all authors, especially those that are more active on Twitter. Therefore, the sum of retweets for these authors are incomplete as well.	Total
	H-Retweets	An author has a h-Retweet of "n" when "n" of its tweets has achieved at least "n" Retweets.	Hybrid

* This metric is currently available only

2.4. Statistical data analysis

Because the goal of this analysis is to find any potential relationships between metrics, this analysis has no preconceptions regarding the nature of each of these metrics. For this reason, the Spearman correlation (α < 0.05) was computed for all pairs of the 31 metrics under study, and a Principal Component Analysis (PCA) was applied in order to display by means of two-dimensional axis the relatedness between the variables analysed with the aim to synthetize components whose relatedness may illustrate different web impact dimensions (RQ2). Both for the correlations and the PCA, all the observations with null values were removed.

Correlations are considered useful in high-level exploratory analyses to check whether different indicators reflect the same underlying causes (Sud & Thelwall, 2014). Spearman correlations were used because it is well-known that citation counts and other impact-related metrics are highly skewed (de Solla Price, 1965). The PCA, on the other hand, has been proved as a valid technique to reduce the dimensionality of the dataset through the identification of principal components (Jollife, 1986). In this case, due to the nature of the web data distribution, Spearman similarity (with varimax rotation of axes and uniform weighting to simplify the data interpretation) was applied.

3. Results

3.1. Online presence of the bibliometrics community

The distribution of authors according to the number of platforms in which they have created a personal profile shows a high degree of social presence (Table 3). Authors with two (26.3%) and three (23.3%) profiles are the more numerous groups whereas there is a small group (11.3%) of authors (most of them specialists) with presence in all five platforms analysed.

Table 3. Social presence (number of authors) of the bibliometrics community.

NUMBER OF	AUTHORS			
PLATFORMS	SPECIALIST	OCCASIONAL	TOTAL	%
5	58	34	92	11.3
4	82	80	162	20.0
3	83	106	189	23.3
2	99	114	213	26.3
1	74	81	155	19.1
TOTAL	**396**	**415**	**811**	

The use of each specific social platform reveals that ResearchGate is, after GSC, the second most used platform by the authors in our sample (67%), followed at some distance by Mendeley (41.2%). However, the number of Mendeley profiles is misleading, since 17.1% of them are basically empty. ResearcherID profiles suffer from the same issue (34.5% of the profiles are empty). Twitter is the least used platform, since only 33.2% of specialist authors (and 26% of occasional authors) have created a Twitter profile. Additionally, most of the authors in our sample that have presence in Twitter, ResearcherID or Mendeley are specialists, while most of the authors in ResearchGate are only occasional authors in bibliometrics (Table 4).

Table 4. Degree of use of social platforms according to the type of author (specialist and occasional).

WEB	AUTHORS					
PLATFORMS	SPECIALIST	%	OCCASIONAL	%	TOTAL	%
*** GSC**	396	100	415	100	**811**	100
ResearcherGate	260	65.7	283	68.2	**543**	67.0
Mendeley	169	42.7	165	39.8	**334**	41.2
ResearcherID	182	46.0	146	35.2	**328**	40.4
Twitter	132	33.3	108	26.0	**240**	30.0

* All authors in the sample have a profile in GSC.

The combination of profiles used by the authors in our sample (specialists and occasional) is shown in Figure 2.

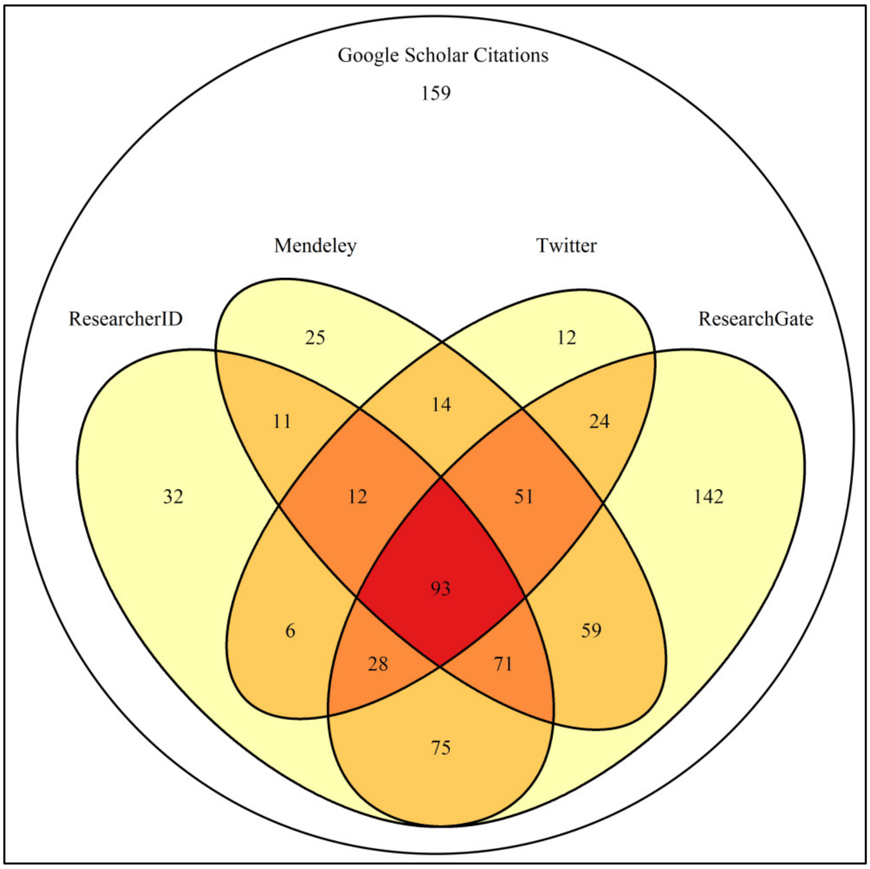


Figure 2. Combination of academic profiles used by the bibliometrics community.

We can observe a great number of researchers who only have a profile in GSC (159) whereas the preferred combination corresponds to GSC and ResearchGate (142). Of the four platforms that we studied (other than Google Scholar Citations), ResearchGate is the one with the highest uptake among the authors in our sample (543 authors had a profile in this platform 66% of the sample). The remaining combinations seem to be more unusual. For example, there are only 12 authors who use only GSC and Twitter or only 11 authors who use only GSC, ResearcherID, and Mendeley.

As it was previously stated, the are many available venues where authors can showcase their work and themselves. As is natural, each author has his/her own preferences, and as a consequence, each profile service offers a different array of products (authors). This issue can be observed just by considering the top 5 authors according to each of the metrics (Table 5). While GSC, ResearcherID, and ResearchGate (all more academic-oriented) seem to portray a similar picture of the discipline, Mendeley, and particularly Twitter, provide a quite different snapshot.

Table 5. Top 5 Author performers according to each of the metrics in each of the academic profiles.

GSC

Citations (5 years)	H-Index (5 years)	I10 index (5 years)	Citations (all)	H-Index (all)	i10 index (all)
L Leydesdorff	L Leydesdorff	L Leydesdorff	L Leydesdorff	L Leydesdorff	L Leydesdorff
M Thelwall	M Thelwall	M Thelwall	E Garfield	M Thelwall	E Garfield
E Garfield	W Glänzel	W Glänzel	M Thelwall	E Garfield	M Thelwall
W Glänzel	L Bornmann	R Rousseau	DJS Price	W Glänzel	R Rousseau
R Rousseau	E Garfield	L Bornmann	F Narin	AF.J. van Raan	W Glänzel

RESEARCHER ID

Articles with cit. data	Citations / Item	Total Articles	Citations	h-index
AK Sahu	H Small	AK Sahu	AK Sahu	E Garfield
E Garfield	L Meho	E Garfield	E Garfield	AK Sahu
L Leydesdorff	I Rafols	HD White	L Leydesdorff	L Leydesdorff
W Glänzel	M Meyer	L Leydesdorff	W Glänzel	W Glänzel
P Jacso	CS Wagner	F Moya	A Schubert	M Thelwall

RESEARCHGATE

RG Score	Impact Points	Publications	Citations	Downloads	Views	Profile Views	Followers	Following
L Leydesdorff	L Leydesdorff	L Leydesdorff	L Leydesdorff	L Leydesdorff	L Leydesdorff	NA Ebrahim	NA Ebrahim	NA Ebrahim
L Bornmann	L Bornmann	R Rousseau	W Glänzel	NA Ebrahim	M Thelwall	C Chen	L Leydesdorff	G. Rathinasabapathy
R Rousseau	R Rousseau	M Thelwall	M Thelwall	C Chen	C Chen	Z Chinchilla	M Thelwall	A Keramatfar
M Thelwall	A Schubert	S Darmoni	F Narin	M Thelwall	S Darmoni	M Thelwall	Z Chinchilla	IF Aguillo
W Glänzel	M Thelwall	C Chen	A Schubert	F Moya	F Moya	L Leydesdorff	IF Aguillo	OB Onyancha

MENDELEY

Readers	Publications	Followers	Following
M Thelwall	RSJ Tol	H Aziz	H Aziz
RSJ Tol	P Mayr	J Pacheco	J Pacheco
J Vanclay	J Vanclay	C Neylon	E Romero
M Pautasso	M Thelwall	I Michán	C Neylon
AW Harzing	IF Aguillo	L Adriaanse	I Michán

TWITTER

Tweets	Days registered	Followers	Following
S Fausto	D Hendrix	J Priem	IF Aguillo
A Ramos	J Delasalle	IF. Aguillo	A Ramos
D Giustini	Á Cabezas	D Giustini	J Pacheco
IF Aguillo	S Konkiel	S Konkiel	NA Ebrahim
S Konkiel	K Holmberg	Á Cabezas	Y Milanes

3.2. Data available to generate Author-Level Metrics

Table 6 provides an indication of the volume of data available in each platform, by comparing the median values of similar indicators across different platforms.

Table 6. Median of principal online metrics broken down by category.

TYPE	SOURCE	MEDIAN
Citations	GSC	156
	ResearchGate	85
	Researcher ID	63
Publications	ResearchGate	27
	Researcher ID	15
	Mendeley	9
H Index	GSC	6
	Researcher ID	4
Followers	Twitter	99
	ResearchGate	38
	Mendeley	3
Following	Twitter	130
	ResearchGate	23
	Mendeley	2

Population:
GSC (n = 811); ResearchGate (n = 515); Researcher ID (n = 275); Twitter (n = 226); Mendeley (n = 185).

As we can see, the median h-index in GSC for authors in our sample ($\tilde{x}$=6) is higher than the median h-index according to ResearcherID ($\tilde{x}$=4). This is most likely a consequence of the higher document coverage in Google Scholar (GS) as compared to the Web of Science. This is also visible if we look at the total number of citations received. The median value according to GS is $\tilde{x}$=156, almost twice the median value according to ResearchGate ($\tilde{x}$=85), which is still higher than the median value of citations according to ResearcherID ($\tilde{x}$=63). Regarding the raw total number of publications, this information was not readily available in GSC profiles, so we could only extract it from ResearchGate (median $\tilde{x}$=27), ResearcherID ($\tilde{x}$=15), and Mendeley ($\tilde{x}$=9). In terms of social interaction metrics, Twitter is the platform that contains more information about followers and followees. It is also worth noting that ResearchGate accumulates more information about social interactions in its platform than Mendeley.

It is important to know that in some cases, even when a profile has been created in one of these platforms, some of the indicators that the platform usually provides are not available. This might be caused by a number of reasons. Table 7 shows the number of profiles in which each metric was available, equal to zero, or not available. The platforms that presented a larger number of profiles in which metrics were unavailable were ResearcherID and Mendeley. ResearchGate also presented a large number of profiles in which the RG Score, the total number of citations, and the Impact Points indicator were not available. In GSC, 25% of the sample had an i-index of 0. This is because these authors did not have any document with at least 10 citations.

Table 7. Number of authors for whom metrics are either or not available in each of the social platforms.

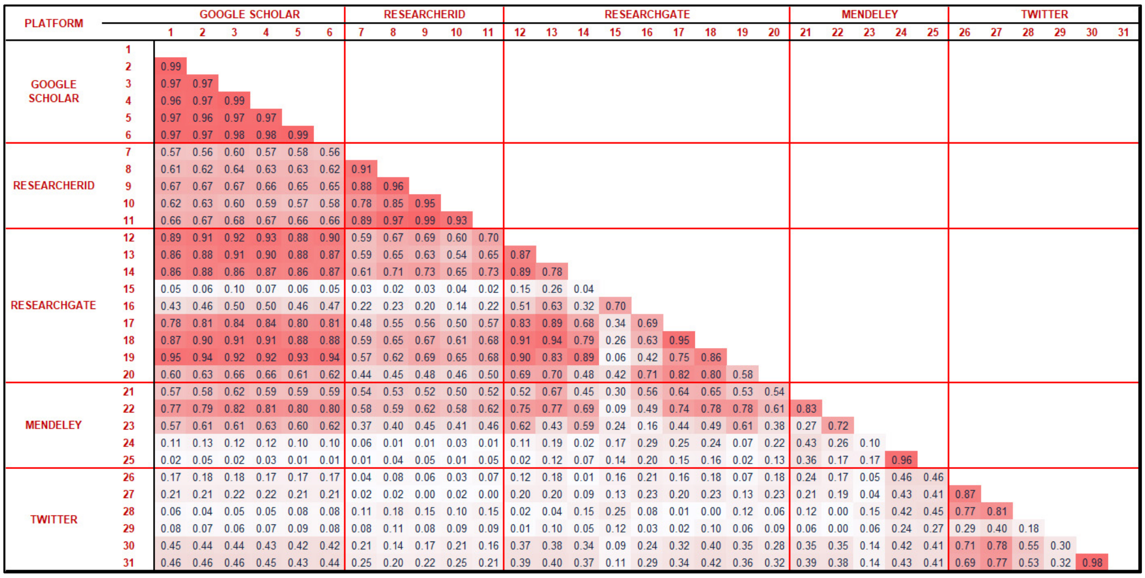
PLATFORM	METRIC	Available	Equal to zero	Not available	% (zero)	% (av.)
GSC	Citations	811	42	0	5.2	0
	h-index	811	42	0	5.2	0
	i10 index	811	203	0	25.0	0
	Citations (Last 5 years)	811	46	0	5.7	0
	h-index (Last 5 years)	811	46	0	5.7	0
	i10 index (Last 5 years)	811	216	0	26.6	0
RESEARCHER ID	Total Articles	328	113	483	34.5	59.5
	Articles with Citation Data	328	131	483	39.9	59.5
	Sum of the Times Cited	328	140	483	42.7	59.5
	Average Citations per Item	328	140	483	42.7	59.5
	h-index	328	140	483	42.7	59.5
RESEARCH GATE	RG Score	543	61	268	11.2	33..0
	Publications	543	28	268	5.2	33..0
	Views	543	22	268	4.1	33..0
	Downloads	543	40	268	7.4	33..0
	Citations	543	56	268	10.3	33..0
	Impact Points	543	118	268	21.7	33..0
	Profile views	543	3	268	0.6	33..0
	Following	543	29	268	5.3	33..0
	Followers	543	16	268	2.9	33..0
MENDELEY	Readers	334	156	477	32.7	58.8
	Publications	334	149	477	31.2	58.8
	Readers per document	185	7	626	3.8	77.2
	Followers	334	122	477	25.6	58.8
	Following	334	156	477	32.7	58.8
TWITTER	Tweets	240	14	571	5.8	70.4
	Followers	240	1	571	0.4	70.4
	Following	240	3	571	1.3	70.4
	Days	240	0	571	0	70.4
	Sum Retweets	240	82	571	34.2	70.4
	H-Retweets	240	82	571	34.2	70.4

Available: metric available in the platform (including zeroes)
Equal to zero: number of authors with the corresponding metric equal to “0”
Not av.: metric not available in the profile.
% (zero): percentage of profiles in which the metric is ZERO, respect to the total number of authors with a profile in the corresponding platform
% (tot): percentage of profiles in which the metric is NOT available, respect to the total number of authors in our sample (811).

3.3. From citations to followers: a comparison of academic profile metrics

All the metrics displayed by GSC correlate strongly with one another, which makes sense because all of them are based on citations (Figure 3). In ResearchGate, however, we find a clear separation between usage (views and downloads) and citation-based metrics (total received citations, impact points, RG Score), and social interaction indicators. This separation can be observed in several platforms. Additionally, moderate to very high correlations are found between citation-based indicators in different platforms, i.e. correlations between the RG Score and all indicators in GSC are very high, whereas metrics in ResearcherID correlate only moderately with

GSC and RG metrics, maybe because metrics in ResearcherID profiles are only updated when the user updates his/her profile, and therefore the data we collected was probably outdated.



PLATFORM		GOOGLE SCHOLAR						RESEARCHERID					RESEARCHGATE									MENDELEY					TWITTER					
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
GOOGLE SCHOLAR	1																															
	2	0.99																														
	3	0.97	0.97																													
	4	0.96	0.97	0.99																												
	5	0.97	0.96	0.97	0.97																											
	6	0.97	0.97	0.98	0.98	0.99																										
RESEARCHERID	7	0.57	0.56	0.60	0.57	0.58	0.56																									
	8	0.61	0.62	0.64	0.63	0.63	0.62	0.91																								
	9	0.67	0.67	0.67	0.66	0.65	0.65	0.88	0.96																							
	10	0.62	0.63	0.60	0.59	0.57	0.58	0.78	0.85	0.95																						
	11	0.66	0.67	0.68	0.67	0.66	0.66	0.89	0.97	0.99	0.93																					
RESEARCHGATE	12	0.89	0.91	0.92	0.93	0.88	0.90	0.59	0.67	0.69	0.60	0.70																				
	13	0.86	0.88	0.91	0.90	0.88	0.87	0.59	0.65	0.63	0.54	0.65	0.87																			
	14	0.86	0.88	0.86	0.87	0.86	0.87	0.61	0.71	0.73	0.65	0.73	0.89	0.78																		
	15	0.05	0.06	0.10	0.07	0.06	0.05	0.03	0.02	0.03	0.04	0.02	0.15	0.26	0.04																	
	16	0.43	0.46	0.50	0.50	0.46	0.47	0.22	0.23	0.20	0.14	0.22	0.51	0.63	0.32	0.70																
	17	0.78	0.81	0.84	0.84	0.80	0.81	0.48	0.55	0.56	0.50	0.57	0.83	0.89	0.68	0.34	0.69															
	18	0.87	0.90	0.91	0.91	0.88	0.88	0.59	0.65	0.67	0.61	0.68	0.91	0.94	0.79	0.26	0.63	0.95														
	19	0.95	0.94	0.92	0.92	0.93	0.94	0.57	0.62	0.69	0.65	0.68	0.90	0.83	0.89	0.06	0.42	0.75	0.86													
	20	0.60	0.63	0.66	0.66	0.61	0.62	0.44	0.45	0.48	0.46	0.50	0.69	0.70	0.48	0.42	0.71	0.82	0.80	0.58												
MENDELEY	21	0.57	0.58	0.62	0.59	0.59	0.59	0.54	0.53	0.52	0.50	0.52	0.52	0.67	0.45	0.30	0.56	0.64	0.65	0.53	0.54											
	22	0.77	0.79	0.82	0.81	0.80	0.80	0.58	0.59	0.62	0.58	0.62	0.75	0.77	0.69	0.09	0.49	0.74	0.78	0.78	0.61	0.83										
	23	0.57	0.61	0.61	0.63	0.60	0.62	0.37	0.40	0.45	0.41	0.46	0.62	0.43	0.59	0.24	0.16	0.44	0.49	0.61	0.38	0.27	0.72									
	24	0.11	0.13	0.12	0.12	0.10	0.10	0.06	0.01	0.01	0.03	0.01	0.11	0.19	0.02	0.17	0.29	0.25	0.24	0.07	0.22	0.43	0.26	0.10								
	25	0.02	0.05	0.02	0.03	0.01	0.01	0.01	0.04	0.05	0.01	0.05	0.02	0.12	0.07	0.14	0.20	0.15	0.16	0.02	0.13	0.36	0.17	0.17	0.96							
TWITTER	26	0.17	0.18	0.18	0.17	0.17	0.17	0.04	0.08	0.06	0.03	0.07	0.12	0.18	0.01	0.16	0.21	0.16	0.18	0.07	0.18	0.24	0.17	0.05	0.46	0.46						
	27	0.21	0.21	0.22	0.22	0.21	0.21	0.02	0.02	0.00	0.02	0.00	0.20	0.20	0.09	0.13	0.23	0.20	0.23	0.13	0.23	0.21	0.19	0.04	0.43	0.41	0.87					
	28	0.06	0.04	0.05	0.05	0.08	0.08	0.11	0.18	0.15	0.10	0.15	0.02	0.04	0.15	0.25	0.08	0.01	0.00	0.12	0.06	0.12	0.00	0.15	0.42	0.45	0.77	0.81				
	29	0.08	0.07	0.06	0.07	0.09	0.08	0.08	0.11	0.08	0.09	0.09	0.01	0.10	0.05	0.12	0.03	0.02	0.10	0.06	0.09	0.06	0.00	0.06	0.24	0.27	0.29	0.40	0.18			
	30	0.45	0.44	0.44	0.43	0.42	0.42	0.21	0.14	0.17	0.21	0.16	0.37	0.38	0.34	0.09	0.24	0.32	0.40	0.35	0.28	0.35	0.35	0.14	0.42	0.41	0.71	0.78	0.55	0.30		
	31	0.46	0.46	0.46	0.45	0.43	0.44	0.25	0.20	0.22	0.25	0.21	0.39	0.40	0.37	0.11	0.29	0.34	0.42	0.36	0.32	0.39	0.38	0.14	0.43	0.41	0.69	0.77	0.53	0.32	0.98	

Figure 3. Correlation matrix (Spearman) for 31 social platform profile metrics associated with the bibliometrics community.

	GSC		ResearcherID		ResearchGate				Mendeley		Twitter
1	Citations	7	Articles	12	RG Score	17	Downloads	21	Publications	26	Tweets
2	Citations (5Y)	8	Articles cited	13	Publications	18	Views	22	Readers	27	Followers
3	h-index	9	Sum of times cited	14	Impact Points	19	Citations	23	Readers/doc	28	Following
4	h5-index	10	Average citations	15	Following	20	Profile Views	24	Followers	29	Days
5	i10-index	11	h-index	16	Followers			25	Following	30	Sum reTweets
6	i10-index (5Y)									31	H reTweets

The number of readers in Mendeley exhibits a very particular behaviour. While it correlates with the usage metrics offered by ResearchGate, it also achieves moderately high correlations with Google Scholar's total citations (r= 0.77) and h-index (r= 0.82), and with the RG Score (r= 0.75).

Regarding Twitter, Figure 3 shows that the number of tweets published, and the number of followers or followees do not correlate well with metrics from other platforms, but only among themselves. Only the sum of retweets and the H retweets have a moderate correlation with citation-based metrics from other platforms (r= 0.44 for total citations in GSC and sum of retweets, and r= 0.45 for total citations in GSC and h-retweets). However, the number of days that a Twitter account has been active (a variable included to check whether it may influence other Twitter metrics) does not seem to correlate with any other metric, not even with the other metrics extracted from Twitter. This suggests that time (whether an author is veteran or rookie in the platform) is not a critic factor to achieve a high number of followers.

Perhaps surprisingly, follower counts across different platforms do not seem to correlate well. Correlations between follower counts in ResearchGate and Twitter (0.23), and ResearchGate and Mendeley (0.29) can only be considered to be low. Only between Twitter and Mendeley did we find a correlation that could be considered moderate (0.43). What it is clear is that connectivity metrics, such as follower counts, do not correlate well at all with citation-based metrics. This separation can be visualized through the Principal Component Analysis (PCA) available in Figure 4.

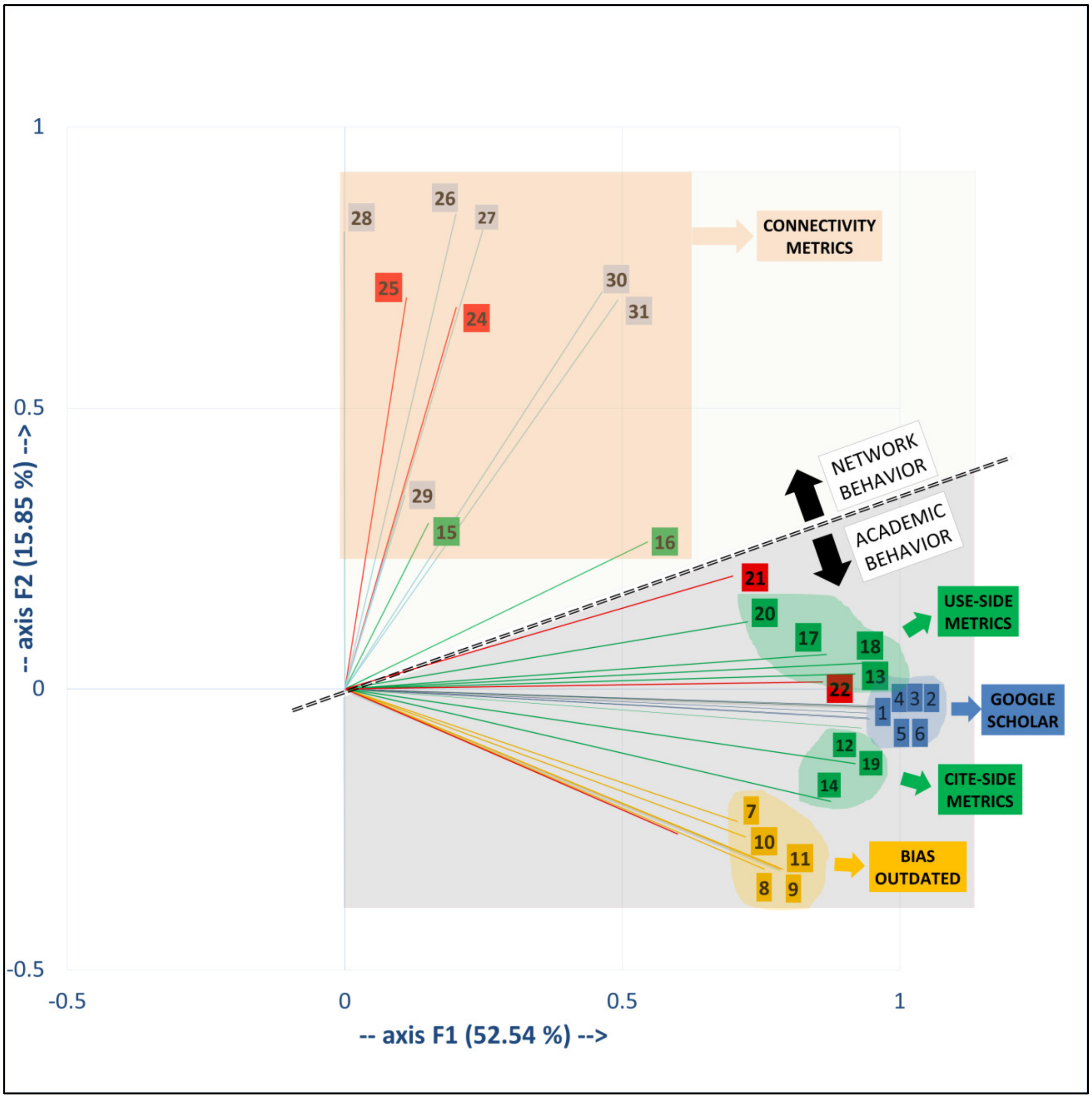


Figure 4. Principal Component Analysis (PCA) for 31 author-level metrics in the bibliometrics community.

	GSC		ResearcherID		ResearchGate				Mendeley		Twitter
1	Citations	7	Articles	12	RG Score	17	Downloads	21	Publications	26	Tweets
2	Citations (5Y)	8	Articles cited	13	Publications	18	Views	22	Readers	27	Followers
3	h-index	9	Sum of times cited	14	Impact Points	19	Citations	23	Readers/doc	28	Following
4	h5-index	10	Average citations	15	Following	20	Profile Views	24	Followers	29	Days
5	i10-index	11	h-index	16	Followers			25	Following	30	Sum reTweets
6	i10-index (5Y)									31	H reTweets

The existence of differences in correlations according to the nature of metrics (total, average, and hybrid scores) may introduce a bias in the previous PCA (Figure 4). In this work, out of the 31 metrics included in the PCA, only 3 are average metrics and 6 hybrid metrics (the remaining 22 are total metrics), minimizing the effect. By way of illustration, a PCA composed only by the 6 hybrid metrics can be observed in Figure 5, reinforcing in this case the previous results. As we can see, while h-index values from Google Scholar are related with RG score and Impact Points in ResearchGate, the h-index from Mendeley is slightly separated. The h-index of ReTweets clearly shows a different impact dimension instead.

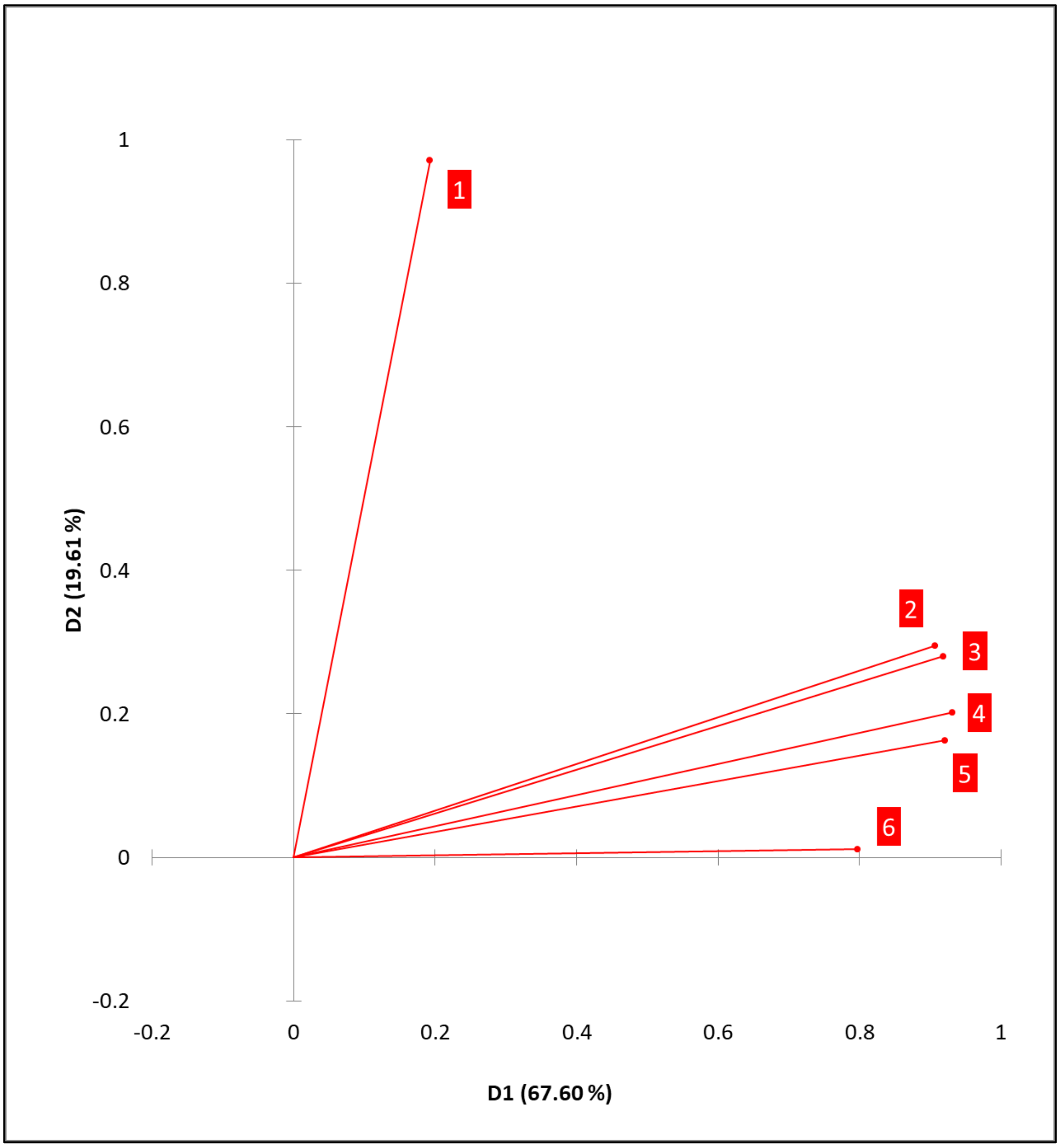


Figure 5. Principal Component Analysis (PCA) for 6 hybrid author-level metrics in the bibliometrics community.

1: Twitter_H ReTweets; 2: GS_H-index; 3: GS_H5 index; 4: RG_Score; 5: RG_Impact Points; 6: ResearcherID_H-Index

Lastly, the PCA related to the 22 total metrics is offered in Figure 6. Although some logical differences with Figure 4 can be observed, the main findings remain unaltered: following/follower metrics form one separated dimension, and citation-based metrics form another one. Mendeley (Readers) is located halfway between the dimentions mentioned above. In this case, however, the number of citations from ResearcherID seems to be located near citation-like metrics, confirming on the one hand that citation metrics make up a dimensions of their own, and on the other hand, the dependence on platforms' coverage and updating mechanisms.

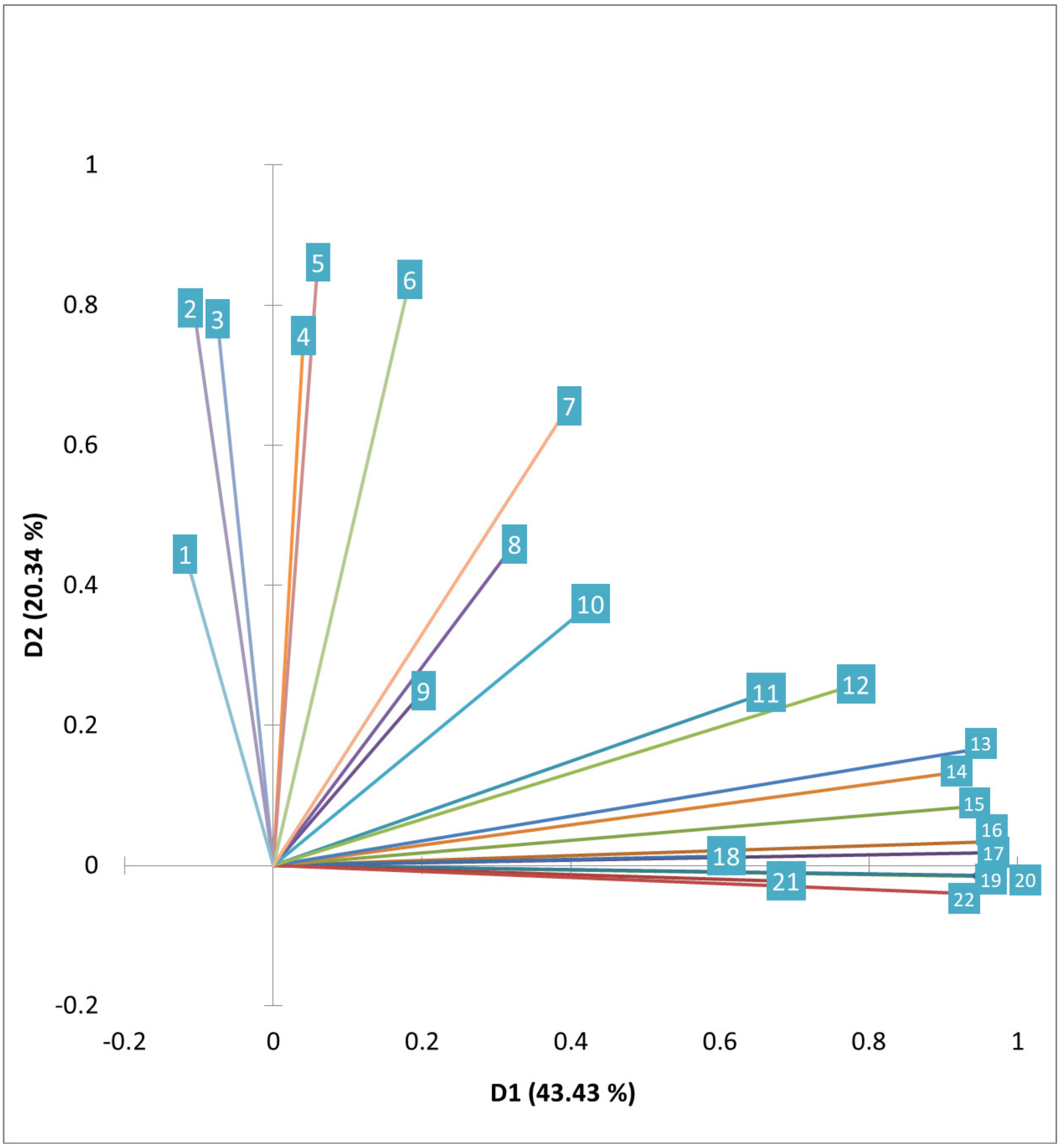


Figure 6. Principal Component Analysis (PCA) for 22 total author-level metrics in the bibliometrics community

1: Twitter_Days; 2: Twitter_Following; 3: Mendeley_Following; 4: Mendeley_Followers; 5: Twitter_Tweets; 6: Twitter_Followers; 7: Twitter_Sum ReTweets; 8: Mendeley_Publications; 9: RG_Following; 10: Mendeley_Readers; 11: RG_Followers; 12: RG_Profile Views; 13: RG_Views; 14: RG_Downloads; 15: RG_Publications; 16: GS_i10-index(5Y); 17: GS_Citations (5Y); 18: ResearcherID_Total articles; 19: GS_i10-index; 20: ResearcherID_Sum of times cited 21: GS_Citations; 22: RG_Citations;

4. Discussion

One of the main limitations of the sample of authors used in this study is the fact that it does not consider all the researchers in the field, because selection was dependent on having created and made public a profile in Google Scholar Citations. Moreover, although there are evidences that suggest that Google Scholar Citations was the most popular profile service in the field under study (Haustein et al., 2014), there could be some researchers in the area who have not created a profile in GSC, but have profiles in other platforms. Our sample would also miss these cases. For this reason, we acknowledge that the sample has a strong bias towards GSC. Our methodology to find author profiles relied on manual searching, and therefore we might have missed some profiles, especially if the authors did not use the same names they use in their research. This could be especially problematic in Twitter. Future studies could make use of more elaborate methodologies for author profile detection, like the one described in Costas, van Honk, and Franssen (2017). Moreover, the number of people who use academic profiles, regardless of the platform they choose, is still only a fraction of the total number of people working in any given discipline. In spite of this, this sample (811 researchers) is the largest to date in studies that aim to analyse the discipline of bibliometrics.

Overall, the results obtained in this work regarding the level of online presence of scientists in social networks are similar to the results found by Van Noorden (2014), and Bosman and Kramer (2016). However, the difference in the uptake of GSC and ResearchGate is higher in this work (only 67% of the authors with a GSC profile also have a ResearchGate profile). The most likely reason for this is that this work only studies bibliometrics researchers, which might have a stronger preference for GSC in detriment of ResearchGate. Previously, Haustein et al. (2014) had found more similar figures in the uptake of these two platforms among bibliometrics researchers. Also, the data collected by Kramer and Bosman (2016) show that ResearchGate and Google Scholar Citations are the most used profile platforms by the respondents to their survey, with a small advantage of ResearchGate over GSC.

Time is also an important factor to consider, because as Haustein et al. (2014) already found, there was a significant increase in the number of researchers who had a profile in the relatively short period of time between the collection of their two samples of data. In the past two years, ResearchGate has emerged as one of the most popular options for researcher to create an academic profile. It is therefore possible that the percentage of researchers that currently have a profile in these platforms has changed since the data for this analysis was collected. There are already studies that show the increasing preference of ResearchGate over Google Scholar Citations, like the already mentioned survey by Kramer and Bosman (2016), and also Mikki et al. (2015), who found ResearchGate to be the platform with more profiles of researchers from the University of Bergen (76% had a profile in this platform).

Regarding the use of multiple profile by the same researchers, the results in this study differ from those found in Mikki et al. (2015) and Ortega (2015a). While Mikki et al. found that 77% of the researchers in their sample only maintained a profile in one platform, (72% in the case of Ortega's study), our analysis found that only 19% of the researchers analysed had only one profile (in GSC). According to our results, 26% had two profiles, 23% three, 20% four, and 11% had a profile in the five platforms that we analysed. The differences might be explained by the particular circumstances of our sample of researchers, who may be more naturally inclined to explore these tools because they are closely related to their field of study. These differences notwithstanding, the results in this study match the finding in Mikki et al. (2015) and Ortega (2015a) that the most common pairwise combination of profiles for authors to have is the combination of GSC and ResearchGate.

RQ 1: Data volumes

It is important to note that some of the metrics analysed in this study are no longer available from the platforms that previously displayed them. This is the case of the Views and Downloads metrics in ResearchGate, which were combined into the Reads[73] metric shortly after we collected our data. The Impact Points metric was also hidden, in this case without giving a public explanation. Conversely, some platforms have added new metrics to their portfolio. ResearchGate started computing the h-index (with and without self-citations). Mendeley also updated its Stats page, which for a long time was also visible to the owner of the profile, giving the choice to make it public to everyone. This stats page displays Media mentions (powered by Newsflo), the h-index and total number of citations received (powered by Scopus), the total number of times an author's papers have been viewed, according to ScienceDirect, and of course, the total number of Mendeley readers.

[73] https://www.researchgate.net/blog/post/introducing-reads

There are several limitations that complicate making comparisons between the author-level and article-level indicators calculated by each platform (e.g. number of people with a profile, number of citing documents found by the platform, and the actual level of interaction among scientists in the platform). Nevertheless, there are significant differences in similar indicators depending on the source of the data: the median number of citations received by authors according to GSC is twice that of RG; the median number of total publications published by authors according to RG is twice that of Mendeley (see Table 6). Lastly, the median number of people that follow an author (followers) and the number of people the author follows (followees) in Twitter are higher by far than the same indicators in ResearchGate and Mendeley. Despite the methodological limitations, these results suggest that some platforms are able to provide more information than others depending on what aspect of an author's academic activities one is interested in.

RQ 2: Author-level metrics

The usefulness of the metrics that were obtained (which were subsequently used to compute correlations and other comparative measurements) depends on the accuracy of the data provided by each platform. GSC is known to suffer from bibliographic inconsistencies (Jacsó, 2012). It has also been reported that it introduces biases against some disciplines and institutions (Ortega, 2015b), and its citation metrics can be easily gamed (Delgado López-Cózar, Robinson-Garcia & Torres-Salinas, 2014). Regarding ResearchGate, some studies have discussed that its flagship indicator, the RG Score, is not an accurate measure of an author's authority in a scientific community, but a measure of how much the author engages in the platform itself, which also opens the metric to manipulation (Kraker and Lex, 2015; Jordan, 2015; Orduna-Malea, Martin-Martin and Delgado López-Cózar, 2016b; Orduna-Malea et al, 2017). Nevertheless, for the purposes of this specific analysis, the consequences of these issues are considered to be low, and should not have affected the correlations.

The correlations found between the different metrics reinforce previous findings obtained at the article level. Priem, Piwowar, and Hemminger (2012) also found that citation-based indicators cluster closely together. Schlögl et al. (2014) argued that downloads, Mendeley readership, and citations reflect different aspects of impact, and Glänzel and Gorraiz (2015) claimed that there are also differences between usage metrics and alternative metrics. The PCA analysis carried out in this study (Figure 4) indeed reflects these differences: Views, Downloads, and Mendeley readers are grouped together, but apart from the social interaction metrics. Therefore, these results agree with the findings by Naude (2016), who suggested that download counts especially, but also Mendeley readership, can be a useful complement to the citation data available in GSC.

Our results also agree with previous findings (Mikki et al., 2015) when we look at author-level metrics. The indicators provided by the same platform tend to correlate, even when they reflect different aspects of impact (e.g. a strong correlation was found between downloads and citations in ResearchGate: 0.75). Additionally, citation-based metrics tend to correlate across platforms: e.g. between GSC and ResearchGate, and to a lower but still considerable degree, between the previous two and ResearcherID. The reason for the lower correlations when ResearcherID is involved has to do with the lack of information in the profiles in this platform: 34.5% of the profiles are empty, and citation indicators in this platform can only be updated manually by the owner of the profile.

The data we extracted from Twitter yielded similar results to those found by De Winter (2015), who concludes that the scientific citation process acts relatively independently of the social dynamics on Twitter. On the other hand, Ortega (2016) concludes that the number of followers indirectly influences the citation impact because participation on Twitter affects the dissemination of research papers, and therefore it may indirectly favour the likelihood of academic outputs being cited. Our findings seem to contradict Ortega's claims (no correlation between number of followers and number of citations received) but we do not think the results can disprove the claim, because correlation does not imply causation. The number and exact composition of followers is a factor that may decisively influence the degree of dissemination of academic outputs (whether the follower base is actually the target audience of the publications). This issue was not addressed in this study.

Lastly, the obtained correlations and PCA results should be taken cautiously since metrics have not been normalized. For example, the age of authors has not been controlled. In this sense, more experienced authors may exhibit a distinctive behaviour compared to emerging authors. This fact does not jeopardize the main findings of this work (identifying different web impact dimensions through raw author-level metrics provided by social networks at discipline level). However, the identification of author clusters (sharing common attributes such as similar academic age range, online activity, gender, language, etc.) constitutes a future line of research.

5. Conclusions

The results indicate that an important number of the researchers in our sample only had a profile in GSC (159), although many of them (543, 67%) also had a profile in ResearchGate, which made it the second most used platform at the time of data collection. The usage indicators (currently, Reads) and the networking capabilities provided by ResearchGate are features that GSC lack.

The analysis finds two main dimensions of online impact (RQ1). There is a cluster of metrics related to academic performance, which can be further subdivided into two subclusters: usage metrics (views, downloads), and citation metrics (citation counts, h-index). The other cluster contains metrics related to social connectivity and popularity (followers).

The authors in our sample seem to prefer Twitter over Mendeley and ResearchGate when it comes to engaging in social interactions. In Mendeley, the researchers in our sample attract a significant amount of followers, but do not tend to use this platform to keep informed about new publications: they mainly use Twitter for this purpose. ResearchGate seemed to be emerging as a source for researchers to keep themselves informed about the latest published research in their fields.

The data suggest that GSC is still the source that is able to provide the highest volumes of citation-related data (RQ2). Regarding social connectivity, Twitter seemed to be the platform of choice, even if it is not a purely academic platform. Nevertheless, ResearchGate also showed a significant amount of social interaction among researchers (followers/followees), much higher than in Mendeley. For this type of comparisons, it is interesting to put side-by-side similar metrics provided by different platforms. Doing this can provide insight into the different levels of uptake of the different platforms by a specific scientific community.

One general conclusion that can be extracted from this study is that despite the general preference towards some platforms (GSC and ResearchGate) there is not any platform in which all researchers in a discipline are present, or one that collects and provides the best data across all dimensions (citations, usage, social connectivity). Nevertheless, we found that publication and citation metrics correlate more or less consistently across platforms.

Lastly, these results should always be interpreted within the context of the bibliometrics community. The study of other communities might very well yield different results. For example, penetration rates of these platforms in other communities might be different, because for the bibliometrics community, academic profile platforms are an increasingly interesting object of study.

Acknowledgements

Alberto Martín-Martín enjoys a four-year doctoral fellowship (FPU2013/05863) granted by the Ministerio de Educación, Cultura, y Deportes (Spain). We also thank Juan Manuel Ayllón for his valuable help cleaning data in the early stages of this study.

References

Bar-Ilan, J., Haustein, S., Peters, I., Priem, J., Shema, H. & Terliesner, J. (2012). Beyond citations: Scholars' visibility on the social Web. In Éric Archambault, Yves Gingras and Vincent Larivière (Eds.). *Proceedings of the 17th International Conference on Science and Technology Indicators* (pp. 98-109). Montréal, Canada.

Barjak, F., Li, X. & Thelwall, M. (2007). Which factors explain the web impact of scientists' personal homepages? *Journal of the American Society for Information Science and Technology*, *58*(2), 200-211.

Bollen, J., Van de Sompel, H., Hagberg, A. & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS one*, *4*(6), e6022. Retrieved March 11, 2017, from http://dx.doi.org/10.1371/journal.pone.0006022

Bornmann, L. (2014). Do altmetrics point to the broader impact ofresearch? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, *8*(4), 895–903.

Bornmann, L. (2016). *Scientific revolution in scientometrics: The broad-ening of impact from citation to societal*. In C.R. Sugimoto (Ed.). Theories of informetrics and scholarly communication (pp. 347–359). Berlin: de Gruyter Mouton.

Bosman, J., & Kramer, B. (2016). Innovations in scholarly communication - data of the global 2015-2016 survey. https://doi.org/10.5281/ZENODO.49583

Chen, C., Tang, Q., Huang, X., Wu, Z., Hua, H., Yu, Y. & Chen, S. (2009). An assessment of the completeness of scholarly information on the internet. *College & Research Libraries*,*70*(4), 386-401.

Costas, R., van Honk, J., & Franssen, T. (2017). Scholars on Twitter: who and how many are they? *Proceedings of the 16th International Conference on Scientometrics and Informetrics*. Wuhan, China.

Costas, R., Zahedi, Z. & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, *66*(10), 2003-2019.

de Solla Price (1965). Networks of scientific papers. *Science*, *149*(3683), pp. 510-515

de Winter, J. C. F. (2015). The relationship between tweets, citations, and article views for PLOS ONE articles. *Scientometrics*, *102*(2), 1773-1779.

Delgado López-Cózar, E., Robinson-García, N. & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446-454.

Glänzel, W. & Gorraiz, J. (2015). Usage Metrics Versus Altmetrics: Confusing Terminology? *Scientometrics*, *102*(3), 2161-2164.

Haustein, S. (2016). Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics*, *108*(1), 413-423.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H. & Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics*, *101*(2), 1145-1163.

Hood, W. W., & Wilson, C. S. (2001). The Literature of Bibliometrics, Scientometrics, and Informetrics. *Scientometrics*, *52*(2), 291–314.

Jacsó, P. (2012). Google Scholar Author Citation Tracker: is it too little, too late?. *Online Information Review*, *36*(1), 126-141.

Jamali, H. R., Nicholas, D. & Herman, E. (2015). Scholarly reputation in the digital age and the role of emerging platforms and mechanisms. *Research Evaluation*, 25(1), 37-49.

Jolliffe, I. T. (1986). Principal component analysis. New York: Springer.

Jordan, K. (2014a). Academics and their online networks: Exploring the role of academic social networking sites. *First Monday*, *19*(11). Retrieved March 11, 2017, from http://dx.doi.org/10.5210/fm.v19i11.4937

Jordan, K. (2014b). Academics' awareness, perceptions and uses of social networking sites: Analysis of a social networking sites survey dataset. Retrieved March 11, 2017, from http://dx.doi.org/10.2139/ssrn.2507318

Jordan, K. (2015). Exploring the ResearchGate score as an academic metric: reflections and implications for practice. *Quantifying and Analysing Scholarly Communication on the Web (ASCW'15)*, 30 June 2015, Oxford. Retrieved March 11, 2017, from http://ascw.know-center.tugraz.at/wp-content/uploads/2015/06/ASCW15_jordan_response_kraker-lex.pdf

Kadriu, A. (2013). Discovering value in academic social networks: A case study in ResearchGate. *Proceedings of the ITI 2013 - 35th Int. Conf. on Information Technology Interfaces Information Technology Interfaces* (pp. 57-62). Zagreb, Croatia.

Kraker, P. & Lex, E. (2015). A critical look at the ResearchGate score as a measure of scientific reputation. *Proceedings of the Quantifying and Analysing Scholarly Communication on the Web workshop (ASCW'15), Web Science conference 2015*. Retrieved March 11, 2017 from http://ascw.know-center.tugraz.at/wp-content/uploads/2016/02/ASCW15_kraker-lex-a-critical-look-at-the-researchgate-score_v1-1.pdfLi, X., Thelwall, M. & Giustini, D. (2011). Validating online reference managers for scholarly impact measurement. *Scientometrics*, *91*(2), 461-471.

Martín-Martín A., Orduna-Malea E., Harzing, A-W. & Delgado López-Cózar, E. (2017). Can we use Google Scholar to identify highly-cited documents?. *Journal of Informetrics*, *11*(1), 152-163.

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). A novel method for depicting academic disciplines through Google Scholar Citations: The case of Bibliometrics. *Scientometrics*, *114*(3), 1251-1273.

Más-Bleda, A. & Aguillo, Isidro F. (2013). Can a personal website be useful as an information source to assess individual scientists? The case of European highly cited researchers. *Scientometrics*, *96*(1), 51-67.

Mas-Bleda, A., Thelwall, M., Kousha, K. & Aguillo, I. F. (2014). Do highly cited researchers successfully use the social web?. *Scientometrics*, *101*(1), 337-356.

Mikki, S., Zygmuntowska, M., Gjesdal, Ø. L. & Al Ruwehy, H. A. (2015). Digital Presence of Norwegian Scholars on Academic Network Sites—Where and Who Are They?. *PloS one*, *10*(11), e0142709. Retrieved March 11, 2017, from http://dx.doi.org/10.1371/journal.pone.0142709

Mohammadi, E. & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology*, *65*(8), 1627-1638.

Nature Research (2014). NPG 2014 Social Networks survey. *Figshare*. https://doi.org/10.6084/m9.figshare.1132584.v4

Naude, F. (2016). Comparing Downloads, Mendeley Readership and Google Scholar Citations as Indicators of Article Performance. *The Electronic Journal of Information Systems in Developing Countries*, *78*(4), 1-25.

Nicholas, D., Clark, D. & Herman, E. (2016). ResearchGate: Reputation uncovered. *Learned Publishing*, *29*(3), 173-182.

Orduna-Malea, E., Martín-Martín, A. & Delgado-López-Cózar, E. (2016a). The next bibliometrics: ALMetrics (Author Level Metrics) and the multiple faces of author impact. *El profesional de la información*, *25*(3), 485-496.

Orduna-Malea, E., Martín-Martín, A. & Delgado López-Cózar, E. (2016b). ResearchGate como fuente de evaluación científica: desvelando sus aplicaciones bibliométricas. *El profesional de la información*, *25*(2), 303-310.

Orduna-Malea, E., Martín-Martín, A., Thelwall, M., & Delgado López-Cózar, E. (2017). Do ResearchGate Scores create ghost academic reputations?. *Scientometrics*, *112*(1), 443-460.

Ortega, Jose L. (2015a). Relationship between altmetric and bibliometric indicators across academic social sites: The case of CSIC's members. *Journal of Informetrics*, *9*(1), 39-49.

Ortega, Jose L. (2015b). How is an academic social site populated? A demographic study of Google Scholar Citations population. *Scientometrics*, *104*(1), 1-18.

Ortega, Jose L. (2016a). *Social Network Sites for Scientists: A Quantitative Survey*. Cambridge, MA: Chandos Publishing.

Ortega, Jose L. (2016b). To be or not to be on Twitter, and its relationship with the tweeting and citation of research papers. *Scientometrics*, *109*(2), 1353-1364.

Piwowar, H. (2013). Altmetrics: Value all research products. *Nature*, *493*(7431), 159-159.

Priem, J. & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, *15*(7). Retrieved March 11, 2017, from http://dx.doi.org/10.5210/fm.v15i7.2874

Priem, J., Piwowar, H.& Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. Retrieved March 11, 2017, from https://arxiv.org/html/1203.4745

Schlögl, C., Gorraiz, J., Gumpenberger, C., Jack, K. & Kraker, P. (2014). Comparison of Downloads, Citations and Readership Data for Two Information Systems Journals. *Scientometrics*, *101*(2), 1113-1128.

Statistical Cybermetrics Research Group (2011). Webometric Analyst 2.0 [online]. Retrieved 7 November from http://lexiurl.wlv.ac.uk

Sud, P., & Thelwall, M. (2014). Evaluating altmetrics. Scientometrics, 98(2), 1131–1143. https://doi.org/10.1007/s11192-013-1117-2

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: a review of the literature. *Journal of the Association for Information Science and Technology*, *68*(9), 2037-2062.

Thelwall, M. (2016). Interpreting correlations between citation counts and other indicators. *Scientometrics*, *108*(1), 337-347.

Thelwall, M. & Kousha, K. (2014). Academia. edu: social network or academic network?. *Journal of the Association for Information Science and Technology*, *65*(4), 721-731.

Thelwall, M. & Kousha, K. (2015). ResearchGate: Disseminating, communicating, and measuring Scholarship?. *Journal of the Association for Information Science and Technology*, *66*(5), 876-889.

Thelwall, M., Haustein, S., Larivière, V. & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLoS One*, *8*(5), e64841. Retrieved March 11, 2017, from http://dx.doi.org/10.1371/journal.pone.0064841

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. *Nature news*, *512*(7513), 126-129.

Wildgaard, L., Schneider, J. W. & Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics*, *101*(1), 125-158.

Chapter 14. Work in progress: web application that displays exhaustive and detailed bibliographic and bibliometric data about researchers working in Spain who have a Google Scholar Citations profile (and their publications)

1. Introduction

Google Scholar Citations (GSC) profiles are currently one of the most convenient ways for researchers to keep an exhaustive and up-to-date publication profile (with citation data attached) on the Web, with relatively low effort required for setting it up and maintaining it. This is because GSC can draw from Google Scholar's (GS), which have been proved to have a much more extensive coverage than commercial citation databases such as Web of Science and Scopus (Martín-Martín, Orduna-Malea, Thelwall, & Delgado López-Cózar, 2018), and because it is automatically, or semi-automatically updated (depending on the user's preference), unlike other profile services such ORCID, ResearcherID, Mendeley's user profiles, and Academia.edu, where users have to manually enter their publications, or import them from other data sources.

According to the results of a survey answered by 20,663 academics from a variety of countries and positions where the respondents were asked about their preferences on researcher profile platforms (Bosman & Kramer, 2016), Google Scholar Citations profiles were the second most used profile platform (10,692 respondents, or 51% of all respondents), closely following ResearchGate (11,378, 55%), and at a large distance from the third platform (ORCID: 5,867, 28%). In October 2018, over 56,000 GSC profiles of researchers working in Spain were identified (Aguillo, 2018).

GSC profiles display the following information about an author:

1. Personal information: name, photo, current affiliation, domain of verified institutional email, and list of topics the researcher is interested in.
2. List of publications where the researcher is author or co-author. For each publication, the basic bibliographic information, as well as its citation count, is displayed on the profile. Furthermore, users can display more detailed information about any publication by clicking on its title, or access the list of citing documents by clicking on the citation count.
3. Using the citation counts of the publications listed in the profile, a series of author-level indicators are computed and displayed in the profile. Each of the indicators is computed twice: once considering all citations received by the researcher, and once considering only the citations received in the last five years.
    a. citations: sum of citation counts of publications listed in the profile
    b. h-index: the largest number h such that h publications have at least h citations (Hirsch, 2005)
    c. i10-index: the number of publications with at least 10 citations
4. A list of co-authors: other authors with a GSC profile with whom the researcher has collaborated in one or more publications.

Despite the usefulness of this service, the bibliometric data it displays is limited, and its interface does not allow much in the way of comparisons among researchers working in the same fields, or among documents concerning similar topics.

Because of these limitations, we think it would be useful to create an application that processes the data in GSC profiles and generates enhanced profiles with more complete bibliographic and bibliometric data. These enhanced profiles would serve various purposes:

- detecting unclean profiles: profiles that contain erroneous information, such as listing documents that were not actually authored by the researcher
- displaying information that is not displayed in regular GSC profiles, such as
    - subject classification at the level of documents (not a journal- or author- level classifications)
    - self-citation indicators (at author- and document-level),
    - publication profile (most frequently used journals, most frequent document types, most frequent co-authors, most frequent topics of publication, number of publications in each year…)
    - citation profile (most common sources of citations to work published by the researcher, by journal, by document type, by authors)
    - reference profile (most common journals, authors, document types cited by the researcher)
    - information on the open access status of the author's publications
    - other author- and article-level bibliometric and altmetric indicators (in some cases normalized by subject, year of publication, academic age of the researcher, etc.)
- enabling users of the application to carry out comparisons among researchers and/or documents in the same area
- Allowing users to export (via a public API) and reuse these enhanced author- and document-level data to carry out other bibliometric analyses, or to populate information in other services

We see a clear benefit in making available enhanced and more transparent versions of researcher profiles that anyone can access and browse on the Web, or reuse for other purposes. Providing that a sufficient level of accuracy and completeness of the data can be attained, a potential reuse case could be research evaluation exercises. In Spain, most official research evaluation exercises are carried out at the level of individual researchers. Researchers are responsible for filling out their own applications, which must include both bibliographic data about their research outputs, and evidence of the impact these outputs have had in their respective scientific communities (usually in the form of bibliometric indicators). At the very least, the data in these enhanced profiles could be used to streamline the process of applying to these evaluation exercises by automatically populating researchers' applications, thus saving hours of work of entering and completing data. In addition to this, the article-level indicators available in the application might be deemed useful to inform evaluators' decisions in these exercises.

2. Methods

During a research stay on CWTS Leiden on August-October 2017, and under the supervision of Ludo Waltman and Nees Jan van Eck, we started identifying all GSC profiles of researchers working in Spain. We followed several strategies to identify as many profiles as possible:

- First, we used a list of normalized Spanish institutions in GSC, which we had obtained from a previous project. We used a script to iterate through the lists of profiles that GSC displayed for each of these institutions, and extracted these lists.
- Since there are many profile that are not included in a normalized institution (Orduña-Malea, Ayllón, Martín-Martín, & Delgado López-Cózar, 2017), we also used the list of Spanish e-mail domains of academic institutions. We also added the top level domains .es, .cat, .eus, and .gal.
- Lastly, we extracted the list of researchers working in Spain released by Isidro Aguillo (edition of April 2017), to check whether we had missed some of them, and added those we were missing to our list.

After processing the lists of researchers obtained through these methods, the total number of profiles identified was 43,873.

Next, a Python script was used to scrape the data displayed in each of these profiles. We extracted personal data, complete publication lists, author-level indicators, and co-author lists for each of the 43,873 profiles. The list of documents in these profiles were processed using Python and R scripts, obtaining a total of 2,031,711 unique documents (after removing duplicates). The process of identifying the profiles of researchers working in Spain and extracting their profile data took approximately a week.

Of these ~2M documents, roughly half had received at least one citation according to GS. We ran a Python script to extract the list of citations to these documents. This was the most time-consuming data-collection task, taking approximately five weeks to extract the citations of all the documents in our sample. In the end, 24,894,896 citations were extracted. However, this task was absolutely necessary to implement the document-level classification process.

GSC profiles doesn't provide a subject classification for the documents in its profiles beyond the general research interests that authors can assign to themselves. To generate a document-level classification for the documents in our sample, the Smart Local Moving algorithm developed by Ludo Waltman and Nees Jan van Eck (2013)[74] was applied to the ~25M citations to those documents extracted from GS (the citation data had to be converted to a suitable format that the algorithm could understand first). The algorithm clustered the documents in the GSC profiles in 3,000 clusters, which were later visualized using the software VOSviewer (van Eck & Waltman, 2010). Each node in Figure 1 is a cluster of documents. The size of the node represents the number of documents in the cluster (larger node, more documents in the cluster) and the color of the node denotes belonging to a higher-level cluster (a cluster of clusters). Thus, the red nodes all belong to biomedicine-related topics, bright green nodes all belong to computer science-related topics, the pink nodes at the top belong to particle physics and astrophysics, the dark yellow nodes in the low right belong to Humanities and Social Sciences (History, Etnography, Law…), and the soft green next to the dark yellow belongs to Education. The number that labels each cluster is currently only the ID of the cluster. For example, the cluster with the ID 24 close to the Social Sciences and Humanities belongs to Bibliometrics and Information Sciences in general.

[74] In 2018, an improved clustering algorithm called the Leiden algorithm was published (Traag, Waltman, & van Eck, 2018). Once this project is resumed, we plan to use this new algorithm instead of the Smart Local Moving algorithm.

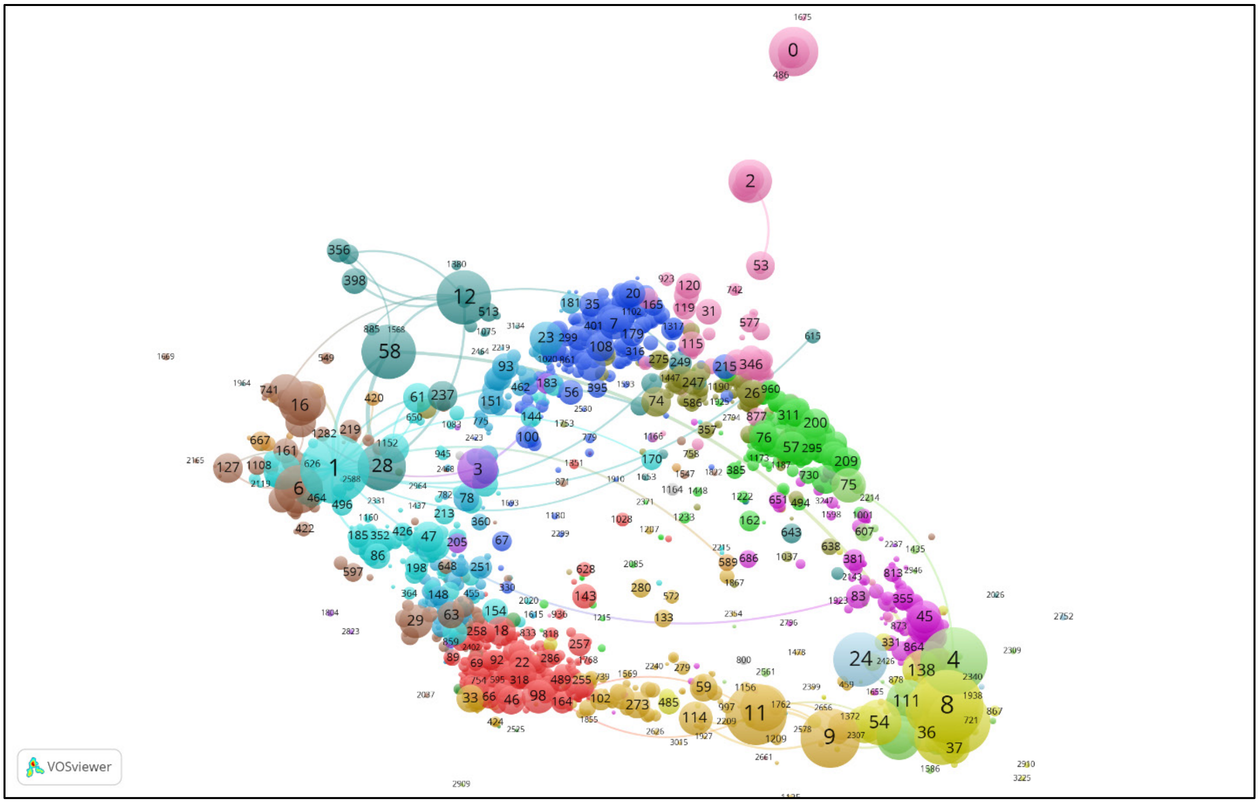


Figure 2. Clusters of documents displayed in the Google Scholar Citations profiles of researchers working in Spain

The next step in the process would be to extract meaningful text labels from the corpus of documents that describe the topic of the documents contained in each cluster. To this end, using Natural Language Processing (NLP) techniques has been proven to be a useful approach. In this regard, our sample has the particularity that it contains documents written in English and documents written in Spanish, which is something we'll have to keep in mind, because NLP techniques usually take into consideration the frequency of occurrence of terms in the corpus.

Additional steps would be necessary to develop a multi-level classification scheme, that is, documents would be aggregated at various levels of aggregation. The lowest-level classification would be the one with 3,000 clusters, but we would ideally like to generate two other levels of classification, one with a few hundred categories (with a similar granularity to the classification scheme in Web of Science, but applied to documents instead of journals), and a high-level classification with a few dozens of categories.

Once all data is adequately processed, a web application that enables users to access, browse it, and export it will be implemented.

Due to the magnitude and complexity of this project, it was not possible to complete it before the presentation of this thesis. However, we plan to continue working on it after the thesis is defended, and we plan to apply for a research grant that allows us to take it to completion. An interesting option that could be explored would be the possibility of combining data from Google Scholar with the increasingly comprehensive reference data available in CrossRef.

References

Aguillo, I. (2018). Ranking of Spanish researchers and researchers working in Spanish Institutions (Spain) according to their Google Scholar Citations public profiles (Eleventh Edition). Retrieved January 16, 2019, from http://www.webometrics.info/en/GoogleScholar/Spain

Bosman, J., & Kramer, B. (2016). Innovations in scholarly communication - data of the global 2015-2016 survey. https://doi.org/10.5281/ZENODO.49583

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2017). The lost academic home: institutional affiliation links in Google Scholar Citations. *Online Information Review*, *41*(6), 762–781. https://doi.org/10.1108/OIR-10-2016-0302

Traag, V., Waltman, L., & van Eck, N. J. (2018). From Louvain to Leiden: guaranteeing well-connected communities. Retrieved from http://arxiv.org/abs/1810.08473

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B*, *86*(11). https://doi.org/10.1140/epjb/e2013-40829-0

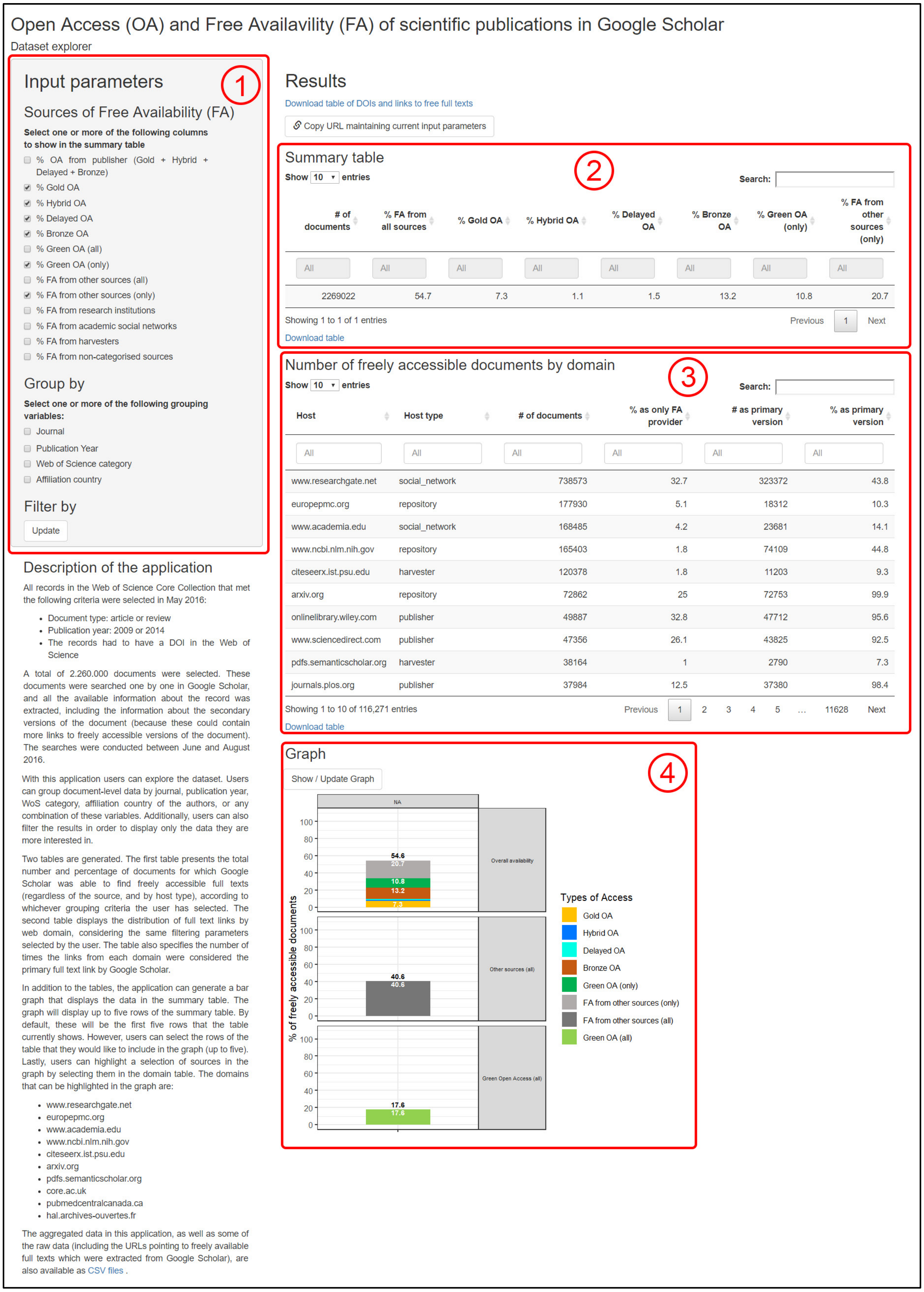
Chapter 15. Description of a web application that presents data on Open Access of scientific publications at various levels of aggregation, based on data from Google Scholar

In order to carry out the study on Open Access (OA) and free availability (FA) levels using evidence extracted from Google Scholar (Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018), a large amount of data was compiled. The article only presented results at large levels of aggregation (by broad categories, and by countries of affiliation), but it is possible to carry out other types of analyses with the data that was collected. For example:

- someone might want to know exactly how many articles in a specific journal are OA or FA, and from which sources.
- someone might be interested in analysing the countries with a larger proportion of OA or FA for a specific discipline.
- someone might want to compare OA and FA levels for a specific set of journals
- someone might want to know the difference in OA and FA levels of papers published in specific journals, subject categories, and by specific countries, across various publication years.

This type of information could be useful to many stakeholders involved with scientific communication, such as journal editors, librarians, policy makers, and researchers themselves.

To facilitate the exploration of this dataset, a web application was developed using the *shiny* framework (Chang, Cheng, Allaire, Xie, & McPherson, 2018). The application has not been made public because of lack of funding necessary to cover the costs of a server that meets the requirements that the application needs to run. A screenshot of the application interface is available in Figure 1. The application has four sections:



Open Access (OA) and Free Availavility (FA) of scientific publications in Google Scholar

Dataset explorer

Input parameters

1

Sources of Free Availability (FA)

Select one or more of the following columns to show in the summary table

- ☐ % OA from publisher (Gold + Hybrid + Delayed + Bronze)
- ☑ % Gold OA
- ☑ % Hybrid OA
- ☑ % Delayed OA
- ☑ % Bronze OA
- ☐ % Green OA (all)
- ☑ % Green OA (only)
- ☐ % FA from other sources (all)
- ☑ % FA from other sources (only)
- ☐ % FA from research institutions
- ☐ % FA from academic social networks
- ☐ % FA from harvesters
- ☐ % FA from non-categorised sources

Group by

Select one or more of the following grouping variables:

- ☐ Journal
- ☐ Publication Year
- ☐ Web of Science category
- ☐ Affiliation country

Filter by

Update

Description of the application

All records in the Web of Science Core Collection that met the following criteria were selected in May 2016:

- Document type: article or review
- Publication year: 2009 or 2014
- The records had to have a DOI in the Web of Science

A total of 2.260.000 documents were selected. These documents were searched one by one in Google Scholar, and all the available information about the record was extracted, including the information about the secondary versions of the document (because these could contain more links to freely accessible versions of the document). The searches were conducted between June and August 2016.

With this application users can explore the dataset. Users can group document-level data by journal, publication year, WoS category, affiliation country of the authors, or any combination of these variables. Additionally, users can also filter the results in order to display only the data they are more interested in.

Two tables are generated. The first table presents the total number and percentage of documents for which Google Scholar was able to find freely accessible full texts (regardless of the source, and by host type), according to whichever grouping criteria the user has selected. The second table displays the distribution of full text links by web domain, considering the same filtering parameters selected by the user. The table also specifies the number of times the links from each domain were considered the primary full text link by Google Scholar.

In addition to the tables, the application can generate a bar graph that displays the data in the summary table. The graph will display up to five rows of the summary table. By default, these will be the first five rows that the table currently shows. However, users can select the rows of the table that they would like to include in the graph (up to five). Lastly, users can highlight a selection of sources in the graph by selecting them in the domain table. The domains that can be highlighted in the graph are:

- www.researchgate.net
- europepmc.org
- www.academia.edu
- www.ncbi.nlm.nih.gov
- citeseerx.ist.psu.edu
- arxiv.org
- pdfs.semanticscholar.org
- core.ac.uk
- pubmedcentralcanada.ca
- hal.archives-ouvertes.fr

The aggregated data in this application, as well as some of the raw data (including the URLs pointing to freely available full texts which were extracted from Google Scholar), are also available as CSV files .

Results

Download table of DOIs and links to free full texts

Copy URL maintaining current input parameters

Summary table

2

Show 10 ▾ **entries** **Search:**

# of documents	% FA from all sources	% Gold OA	% Hybrid OA	% Delayed OA	% Bronze OA	% Green OA (only)	% FA from other sources (only)
All	All	All	All	All	All	All	All
2269022	54.7	7.3	1.1	1.5	13.2	10.8	20.7

Showing 1 to 1 of 1 entries Previous 1 Next

Download table

Number of freely accessible documents by domain

3

Show 10 ▾ **entries** **Search:**

Host	Host type	# of documents	% as only FA provider	# as primary version	% as primary version
All	All	All	All	All	All
www.researchgate.net	social_network	738573	32.7	323372	43.8
europepmc.org	repository	177930	5.1	18312	10.3
www.academia.edu	social_network	168485	4.2	23681	14.1
www.ncbi.nlm.nih.gov	repository	165403	1.8	74109	44.8
citeseerx.ist.psu.edu	harvester	120378	1.8	11203	9.3
arxiv.org	repository	72862	25	72753	99.9
onlinelibrary.wiley.com	publisher	49887	32.8	47712	95.6
www.sciencedirect.com	publisher	47356	26.1	43825	92.5
pdfs.semanticscholar.org	harvester	38164	1	2790	7.3
journals.plos.org	publisher	37984	12.5	37380	98.4

Showing 1 to 10 of 116,271 entries Previous 1 2 3 4 5 … 11628 Next

Download table

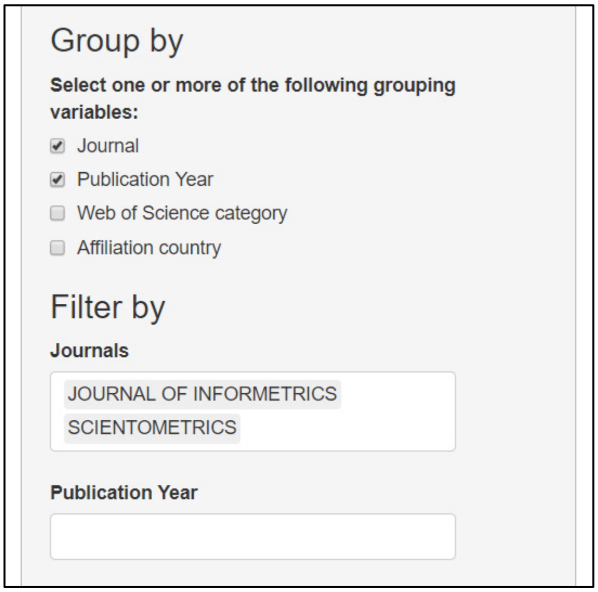
Graph

4

Show / Update Graph

Figure 3. Interface of the web application

1. Input parameters: this is where users can set up their query. This section has three subsections:
   1.1. Sources of Free Availability: here users can select exactly which types of OA and FA will be displayed in the summary table. By default, Gold, Hybrid, Delayed, Bronze, Green (only), and FA from other sources (all) are selected.
   1.2. Group by: here users can choose the level at which the application will display results. The available options are by journal, publication year, Web of Science (WoS) category, and Affiliation country. Two or more grouping levels can be selected at once (Figure 2).
   1.3. Filter by: once a grouping variable has been selected, it is also possible to filter by one or more specific cases within that variable. For example, if the journal grouping variable is selected, it is possible to filter the results to show only those relative to specific journals (Figure 2).



Group by

Select one or more of the following grouping variables:

☑ Journal
☑ Publication Year
□ Web of Science category
□ Affiliation country

Filter by

Journals

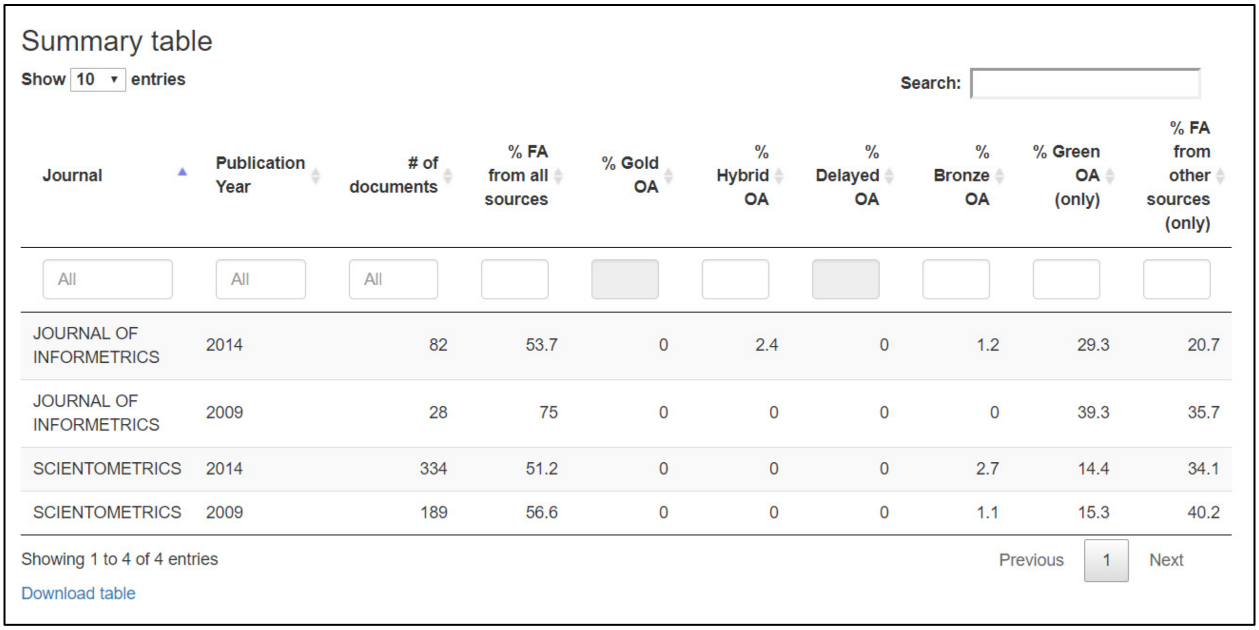
JOURNAL OF INFORMETRICS
SCIENTOMETRICS

Publication Year

Figure 2. Example of how grouping variables can be selected and filters can be applied

2. Summary table: the summary table takes the input parameters, and computes the percentages of the different types of OA and FA selected, grouped by the desired variables, and applying the specified filters. Following with the example in Figure 2, the resulting summary table is available in Figure 3. The summary table displays the following columns:
   2.1. Grouping variables: first, a column is displayed for each grouping variable selected. The rows of the summary table therefore correspond to each of the possible combinations of the selected grouping variables.
   2.2. # of documents: the absolute number of articles that are found for each element of the grouping variables.
   2.3. Percentages of OA and FA: the remaining columns are those that were selected in the subsection "Sources of Free Availability" in the Input parameters.

Users can display up to 100 rows at a time (with the option to navigate through pages if necessary), sort the data in the summary table by any of its columns, and they can further filter the results using the search function and the filtering option that can be found below the heading of each column. For numeric values, users can filter by range (less than, more than) as well as by specific values. Lastly, users can export the table in CSV format.



Summary table

Show 10 entries

Search:

Journal	Publication Year	# of documents	% FA from all sources	% Gold OA	% Hybrid OA	% Delayed OA	% Bronze OA	% Green OA (only)	% FA from other sources (only)
All	All	All							
JOURNAL OF INFORMETRICS	2014	82	53.7	0	2.4	0	1.2	29.3	20.7
JOURNAL OF INFORMETRICS	2009	28	75	0	0	0	0	39.3	35.7
SCIENTOMETRICS	2014	334	51.2	0	0	0	2.7	14.4	34.1
SCIENTOMETRICS	2009	189	56.6	0	0	0	1.1	15.3	40.2

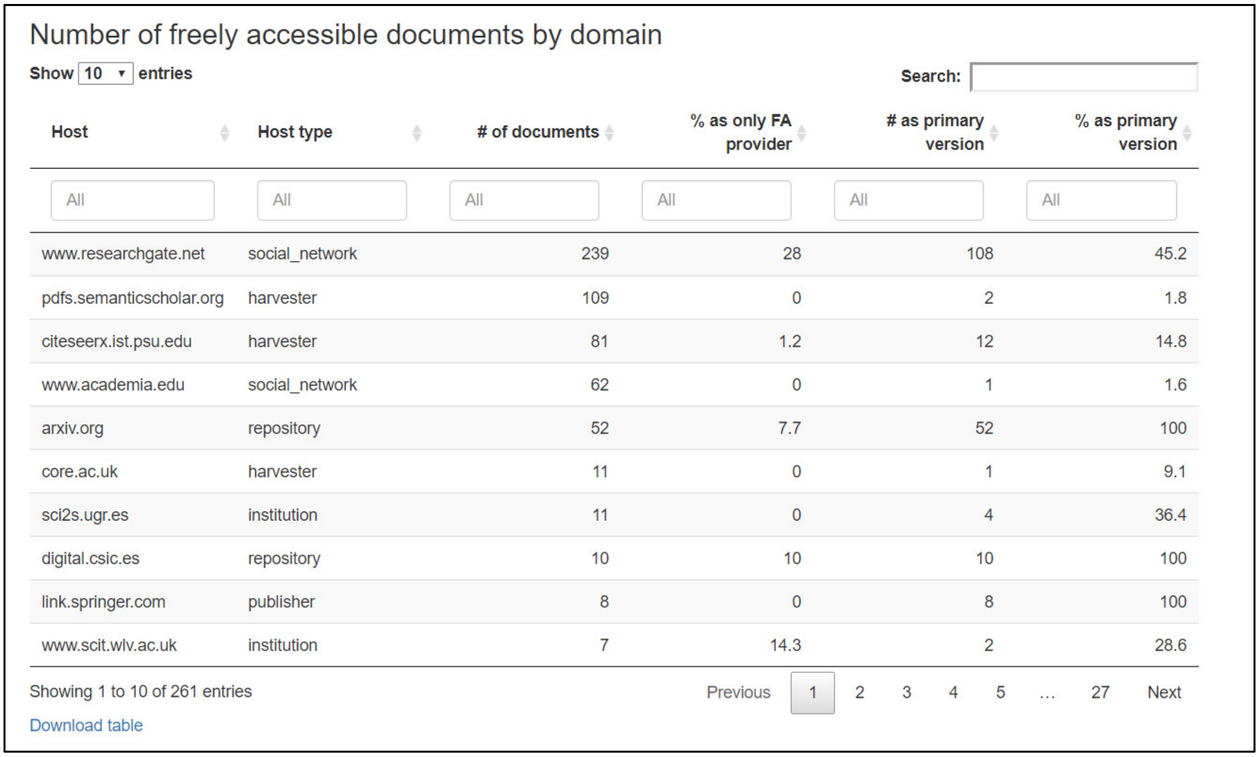
Showing 1 to 4 of 4 entries

Download table

Figure 3. Example of summary table

3. Number of freely accessible documents by domain: here the application takes all the articles that meet the requirements of the input parameters, and computes a frequency table of the sources where GS found OA or FA versions of these articles. The information displayed in this table is:
   3.1. Host: website domain that hosts the freely accessible version of the article
   3.2. Host type: whether the host is a publisher website, a repository, an institutional website, a harvester, or a host of unknown type.
   3.3. # of documents (that meet the input parameters) that have been found in the host.
   3.4. % as only FA provider: proportion (relative to # of documents) of cases where the version provided by this host was the only freely accessible version available in GS.
   3.5. # as primary version: absolute number of times when the freely accessible version in a specific host was the version that GS displays in the primary record (not within the list of secondary versions).
   3.6. % as primary version: proportion (relative to # of documents) where the freely accessible version from a specific host was the primary version.

Figure 4 shows the frequency table that is computed with the input parameters established in Figure 2. As in the summary table, in this table users can set a higher number of entries displayed at a time, navigate through the subsequent pages, sort by any column, filter, and download the table as a CSV file.



Number of freely accessible documents by domain

Show 10 entries

Search:

Host	Host type	# of documents	% as only FA provider	# as primary version	% as primary version
All	All	All	All	All	All
www.researchgate.net	social_network	239	28	108	45.2
pdfs.semanticscholar.org	harvester	109	0	2	1.8
citeseerx.ist.psu.edu	harvester	81	1.2	12	14.8
www.academia.edu	social_network	62	0	1	1.6
arxiv.org	repository	52	7.7	52	100
core.ac.uk	harvester	11	0	1	9.1
sci2s.ugr.es	institution	11	0	4	36.4
digital.csic.es	repository	10	10	10	100
link.springer.com	publisher	8	0	8	100
www.scit.wlv.ac.uk	institution	7	14.3	2	28.6

Showing 1 to 10 of 261 entries

Figure 4. Example of host frequency table

4. Graph: some of the data in the summary table is also displayed as a stacked column bar. However, because the summary table can have many rows, the graph is limited to five rows. By default, these are the first five rows in the summary table, but users can change the default behaviour by selecting whichever rows they want in the summary table (up to five). For each row of the summary table, three stacked column graphs are generated. The first one shows overall availability. This includes all types of publisher OA (Gold, Hybrid, Delayed, Bronze), as well as Green (only), and other sources (only). The second one displays all non-publisher and non-repository sources, and the third one displays all Green sources. Additionally, in the columns that display all other sources and all green it is possible to single out one specific host, which can be selected from the host frequency table. This option is only available for some hosts (those with volume of articles that is large enough to be appreciated in a graph):
    - www.researchgate.net
    - europepmc.org
    - www.academia.edu
    - www.ncbi.nlm.nih.gov
    - citeseerx.ist.psu.edu
    - arxiv.org
    - pdfs.semanticscholar.org
    - core.ac.uk
    - pubmedcentralcanada.ca
    - hal.archives-ouvertes.fr

The graph that is generated after selecting the input parameters in Figure 2 can be observed in Figure 5. Before generating this graph, the host www.researchgate.net was selected from the host frequency table, and for this reason we can see this host singled out in the “Other sources (all)” columns.

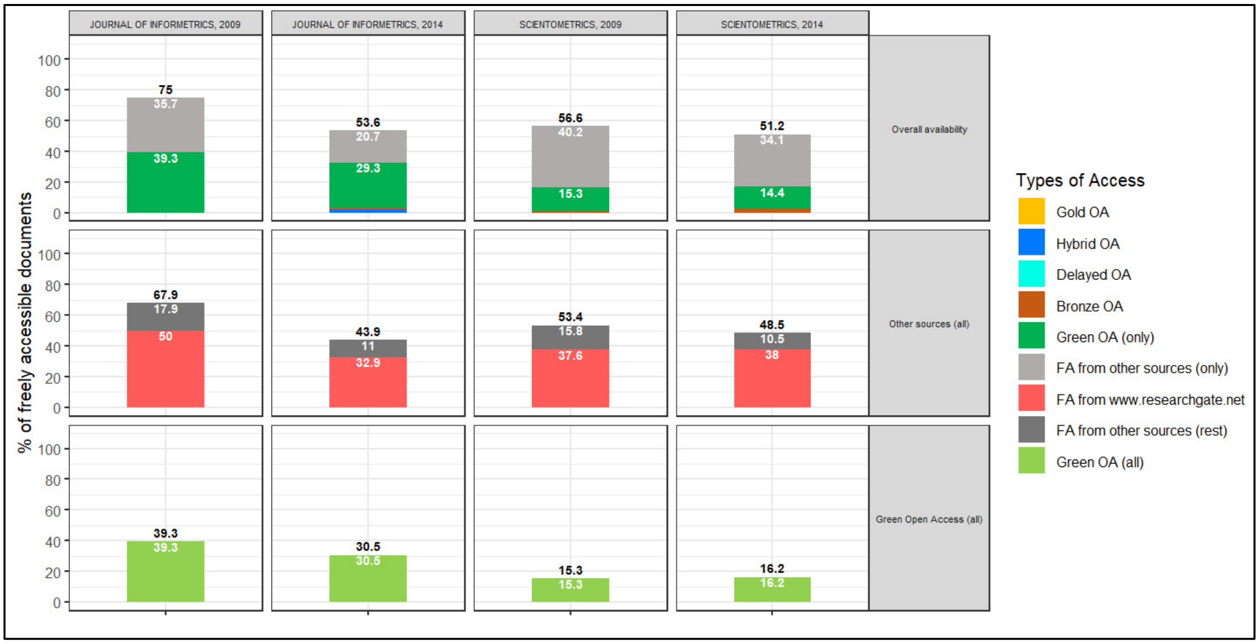


Figure 5. Example of a stacked column graph generated by the application

Lastly the application also lets users export raw article-level data in CSV format. However, because of the limitations in the use license of WoS data, the raw data is limited to the DOI of the article, the full text URLs found by GS, and basic bibliographic metadata such as the publication year. Author affiliations and WoS categories are not included in the exported article-level data.

Chapter 16. Evidence of Open Access of scientific publications in Google Scholar: a large-scale analysis

Journal article. Cite as:

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. Journal of Informetrics, 12(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012

Abstract (English)

This article uses Google Scholar (GS) as a source of data to analyse Open Access (OA) levels across all countries and fields of research. All articles and reviews with a DOI and published in 2009 or 2014 and covered by the three main citation indexes in the Web of Science (2,269,022 documents) were selected for study. The links to freely available versions of these documents displayed in GS were collected. To differentiate between more reliable (sustainable and legal) forms of access and less reliable ones, the data extracted from GS was combined with information available in DOAJ, CrossRef, OpenDOAR, and ROAR. This allowed us to distinguish the percentage of documents in our sample that are made OA by the publisher (23.1%, including Gold, Hybrid, Delayed, and Bronze OA) from those available as Green OA (17.6%), and those available from other sources (40.6%, mainly due to ResearchGate). The data shows an overall free availability of 54.6%, with important differences at the country and subject category levels. The data extracted from GS yielded very similar results to those found by other studies that analysed similar samples of documents, but employed different methods to find evidence of OA, thus suggesting a relative consistency among methods.

Abstract (Spanish)

En este artículo se usa Google Scholar (GS) como una fuente de datos para analizar niveles de Open Access (OA) a nivel de países y campos de investigación. Se analizan todos los artículos y revisiones bibliográficas con DOI publicadas en 2009 o 2014 e indizadas por los tres principales índices de citas de la Web of Science (2.269.022 documentos). Se extrajeron de GS los enlaces a versiones gratuitamente accesibles de estos documentos. Para diferenciar entre formas de acceso más fiables (sostenibles y legales) de otras menos fiables, los datos extraídos de GS se combinaron con la información disponible en DOAJ, CrossRef, OpenDOAR, y ROAR. Esto nos permitió distinguir los porcentajes de documentos en nuestra muestra que estaban en OA desde la editorial (23,1%, incluyendo OA por la ruta dorada, revistas híbridas, con embargo, y Bronce) de los que estaban disponibles por la ruta verde (17,6%), y de los que estaban disponibles por otros medios (40,6%, principalmente debido a ResearchGate). Los datos muestran una disponibilidad general del 54,6%, con importantes diferencias a nivel de países productores, y categorías temáticas. Los datos extraídos de GS muestran resultados similares a los encontrados en otros estudios que analizaban muestras similares de documentos, pero empleaban otros métodos para encontrar evidencias de OA, lo que sugiere una consistencia relativa entre métodos.

1. Introduction

1.1. Beginnings of the Open Access movement

The widespread adoption of web technologies removed most of the physical impediments for accessing scientific information (Harnad, 2001). Since then, the issue of Open Access (henceforth referred to as OA) to the scholarly literature has been hotly debated by all sorts of actors in the academic community, including researchers, publishers, funding institutions, librarians, and policy makers. Many of these discussions revolved around the ways in which the system of scholarly communication should change, taking advantage of this new virtual environment to become more effective and efficient and thus hopefully solve problems like the affordability and accessibility to scientific information that afflict many research institutions.

One of the first crystallizations of these intentions to change the scholarly communication system was the Budapest Open Access Initiative (Chan et al., 2002) (BOAI). This was the first time the term "Open Access" was used, although the practices described in that document had already been taking place in some scientific communities long before that date. The BOAI defined OA to the literature as:

"*free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited*".

Additionally, the BOAI also described the two main ways to realise the goal of OA: by self-archiving documents in public archives (which later came to be known as Green OA), or by publishing in OA journals (later dubbed as Gold OA). Poynder (2018) provides a historic overview of OA since the BOAI declaration.

Since the original BOAI declaration was first published, the discussion has continued and the panorama of scholarly publishing and OA has greatly changed. All actors have had to adapt in some way to the new reality. In addition, the Web gave rise to new types of academic platforms, which further complicated the issue of access to scientific information by expanding the access points to scientific content (e.g. Google Scholar, ResearchGate, etc.). Some of these new platforms were quickly adopted by the scientific community and have already become an important part of the system. These platforms will be discussed later on.

1.2. Reactions of academic institutions, funders and publishers to OA

In the beginnings of the OA movement, a great emphasis was put on the importance of authors self-archiving their own publications on public repositories (Harnad, 2001). Many research institutions, which saw in self-archiving a potential solution to the *journal affordability problem* (the problem of selecting which journals to subscribe to, when economic resources are limited), put systems in place to allow researchers to self-archive and make public their research. These institutional repositories are under the direct control of the institution, and are usually managed by the libraries. Additionally, other subject-specific repositories were launched. Apart from arXiv[75], the physics repository created in 1991 in Cornell University, many other repositories are now available to researchers. ROAR[76] (Registry of Open Access Repositories) and OpenDOAR[77] (Directory of Open Access Repositories) provide an exhaustive list of these institutional and subject based OA repositories. More recently, there has been an explosion in the growth of the so-called preprint servers, largely enabled by the infrastructure developed by the Open Science Framework[78], a project launched by the Center for Open Science, which is a non-profit organization founded in 2013 to "increase the openness, integrity, and reproducibility of scientific research" (Mellor, 2016, para. 6). These servers are designed to share manuscripts that still have not gone through a process of peer review, although they usually welcome accepted manuscripts as well.

One of the notions that has served to justify the need of OA is that money from public institutions to fund research was not realizing its true potential, because most publicly-funded research ended up behind publishers' paywalls, and other researchers who could make use of that research had no access to it. For these reasons, many funding institutions, governments, and policy makers started to issue OA mandates to force researchers who use their funding to make their results OA. Among these we can find the National Institutes of Health (NIH) in the USA, the Research Councils in the United Kingdom, or the European Research Council. In 2016, the European Union announced its resolve to make all scientific publications based on publicly-funded research freely accessible by 2020 (Enserink, 2016).

[75] https://arxiv.org/

[76] http://roar.eprints.org/

[77] http://www.opendoar.org/

[78] https://osf.io/

ROARMAP[79] (Registry of Open Access Repository Mandates and Policies) provides an exhaustive database of OA mandates issued by all kinds of organizations worldwide.

Largely because of these mandates, most publishers adapted their business models, which previously relied almost exclusively on journal subscriptions paid by academic institutions, to business models compatible with the OA requirements mandated by funders:

- Gold OA journals publish all their articles as OA. Their revenue usually comes from charging Article Processing Charges (APC) to authors instead of charging subscription fees to academic libraries. There is much controversy concerning the price of these APCs, which range from a few hundreds of dollars, to over $5,000 per article. There are also Gold OA journals that do not charge APCs to authors, and instead absorb publishing costs in other ways (like via member subscriptions fees in the cases of academic societies that also publish journals). These are sometimes called Diamond OA or Platinum OA journals (Fuchs & Sandoval, 2013; Haschak, 2007).
- Hybrid OA journals maintain the subscription model, but give authors the choice to make their article OA, also by paying an APC (Prosser, 2003; Walker, 1998). This model has also been controversial, because in addition to charging APCs to authors to make the articles OA, they still charge libraries ever-increasing subscription costs for access to the entire collection of articles published by the journals. This phenomenon has been dubbed “double-dipping”, because publishers seem to be charging twice for the same content. Some publishers, like Elsevier, claim that Hybrid OA articles are excluded when calculating subscription costs[80], while other publishers compensate institutions “for the extra money they are putting into the system through payment of APCs” (Kingsley, 2017, para. 3) by means of the so-called “offset agreements”, which can take many forms. Lawson (2018) reports on the offset agreements made with publishers by the organization JISC Collections, which works on behalf of UK academic libraries.
- Delayed OA journals are subscription journals that convert their articles to OA once a specific amount of time has passed after publication. Laakso and Björk (2013) analyzed a sample of 111,312 articles published in 492 journals and found that 77.8% of them were available from the publisher website twelve months after publication. The percentage reached 85.4% 24 months after publication.
- Gratis Access Journals (Suber, 2008a, 2008b): journals that make their articles free-to-read, but don’t extend other rights to users (such as reuse or distribution) apart from the right to read. The publisher retains the copyright of these articles. This type of access is also referred to as “public access”, especially by the publishing industry (Crotty, 2017). Sometimes publishers intend to maintain access to these documents free indefinitely, but sometimes access is only free for a specific period of time (promotional access). Therefore, this type should not be conflated with Gold, Hybrid, or Delayed OA.

The costs of subscriptions and APCs are continually increasing (Tickell et al., 2017). This fact has led a number of institutions and governments to re-negotiate the so-called Big Deals (flat rates to access large numbers of journals published by a single publisher) so that they also include flat rates or considerable discounts for the APCs of the articles their researchers publish (Elsevier, 2015). In other cases, governments have refused to pay the increasing costs that large commercial publishers demanded. This was the case with Germany and the publisher Elsevier. A coalition of German institutions (grouped under the name project DEAL[81]) decided not to renew their license to Elsevier content at the end of 2016. Elsevier subsequently stopped allowing them to access its content, but decided to restore access shortly after, “in good faith” while negotiations lasted. By June 2018 an agreement had still not been reached. After Germany, other countries have followed suit: in March 2018,

[79] http://roarmap.eprints.org/
[80] https://www.elsevier.com/about/our-business/policies/pricing#dipping
[81] https://www.projekt-deal.de/about-deal/

the Couperin consortia in France decided to not to renew their agreement with Springer-Nature, and in May 2018 the Bibsam Consortium in Sweden decided not to renew their agreement with Elsevier (Else, 2018).

Most journal publishers also offer alternative sharing policies for the articles that they do not publish as OA. The freedom these policies give to researchers to self-archive their content greatly varies by publisher and by specific journal. These policies often include embargo periods that prohibit authors to share their research on public repositories for a period of time after publication (from less than a year, to over two years). Despite initiatives like Sherpa/Romeo[82] or the publisher-backed How Can I Share It[83], which try to aggregate and standardise publisher's sharing policies, it is difficult to keep track of them because they change over time, usually to become more restrictive regarding how, where, and when self-archiving is permitted (Gadd & Troll Covey, 2016; Kingsley, 2013). These policies are often arbitrary and complicated, for example allowing to share an article immediately upon publication from the author's personal website, but imposing an embargo to share the same article from an institutional repository (Bolick, 2017; Tickell et al., 2017).

1.3. New players in the system

Other types of platforms, different from repositories and publishers but also with a large impact in the free availability of scholarly literature, have been launched since the BOAI declaration. In 2007 the academic search engine CiteSeerX[84] (based on an even earlier version called CiteSeer) was launched by Pennsylvania State University. In 2008, the academic social networks ResearchGate[85] and Academia.edu[86] were launched. In 2015, the search engine Semantic Scholar[87], developed by the Allen Institute for Artificial Intelligence, was launched, focusing mostly on the areas of Computer Science, and recently also Biomedicine. All these platforms share the characteristic that they host copies of the full texts of scholarly documents (automatically harvested from other sources or uploaded by users themselves) and make them available to their users, thus becoming another source from which readers can access scientific information.

Academic social networks (ASN) in particular have attracted a lot of attention because of how quickly users have taken to sharing their work on them (Björk, 2016). Borrego (2017) found that researchers from 13 Spanish universities used ResearchGate much more frequently to upload and share their research than the repositories available at their institutions. Martín-Martín, Orduna-Malea, Ayllón, and Delgado López-Cózar (2014; 2016), and Jamali and Navabi (2015) studied the free accessibility to a sample of documents covered by Google Scholar. Both studies found that ResearchGate was the source that provided the highest number of freely available full texts. However, full text documents in ASNs are uploaded by researchers themselves and, unlike OA repositories, these platforms do not carry out any kind of checks to guarantee copyright compliance. This resulted in a large portion of documents being accessible from ASN in violation of their copyright. Jamali (2017) found that 51.3% of the non-OA documents in a sample of 500 random documents were available from ResearchGate in violation of their copyright.

Moreover, despite some similarities, academic social networks engage in practices that clearly set them apart from OA repositories. The ongoing dispute between publishers and ResearchGate (Coalition for Responsible Sharing, 2017a, 2017b, 2017c) is unequivocal proof of the instability of these platforms as sources of full texts. A related issue is that in ResearchGate users are allowed to delete full texts of documents they have uploaded, even in the cases when the platform generates a DOI for the document (through their collaboration with DataCite[88]). This entirely differs from the policies of repositories such as arXiv or socArxiv, where the academic record is always maintained (authors cannot delete files but retain the right to issue a retraction notice if they feel a document they deposited should no longer be

[82] http://www.sherpa.ac.uk/romeo
[83] http://www.howcanishareit.com/
[84] http://citeseerx.ist.psu.edu
[85] https://www.researchgate.net
[86] http://www.academia.edu
[87] https://www.semanticscholar.org
[88] https://www.datacite.org/

used). Full texts uploaded by a user to ResearchGate are also deleted if the user deletes his/her account in the platform. Academia.edu also engages in practices that make it different from repositories. This academic social network requires users to log in to their platform to access full texts. However, perhaps because this contravenes Google Scholar's indexing policies[89], they left open a back door so that users coming from a Google Scholar search[90] would be allowed to access full texts without the need to log in. Presumably, they did this to avoid being dropped as a source by Google Scholar, a large source of web traffic given its huge user-base. These cases raise the need to distinguish between merely uploading a document to the Internet (to ResearchGate, Academia.edu or to any privately managed personal website) and depositing or archiving a document in a repository, which usually provides more guarantees as to the long-term preservation of the documents that they host.

There is another player who is currently having a major influence in the accessibility to scholarly literature: Sci-Hub. This website was launched in 2011 by a graduate student called Alexandra Elbakyan, and it illegally provides access to over 60 million research articles. Elbakyan developed a system that automatically accesses publisher websites using credentials *donated* by users who work at institutions with access to paywalled journal articles. There are reports, however, that claim that some of these credentials might have been stolen rather than donated (Bohannon, 2016). The system then copies the full texts of articles to the Library Genesis database (LibGen), which is the platform that hosts the articles that in turn are provided to the users. The kind of copyright-infringing access that Sci-Hub provides is sometimes called Robin Hood OA, Rogue OA, and Black OA (Archambault et al., 2014; Björk, 2017; Green, 2017). Despite the efforts made by large commercial publishers like Elsevier to shut down Sci-Hub's operations, the website remained functional at the time of this writing, providing access to the vast majority of recently-published paywalled articles (Himmelstein et al., 2018) and virtually providing access to all scientific publications worldwide.

1.4. Current landscape of free availability of scientific information

To summarise the scenario described above, Figure 1 provides a representation of the main paths by which a journal article may become freely available on the Web.

[89] https://scholar.google.com/intl/en/scholar/inclusion.html#content

[90] Technically speaking, users who accessed Academia.edu with the Referer HTML request header "https://scholar.google.com"

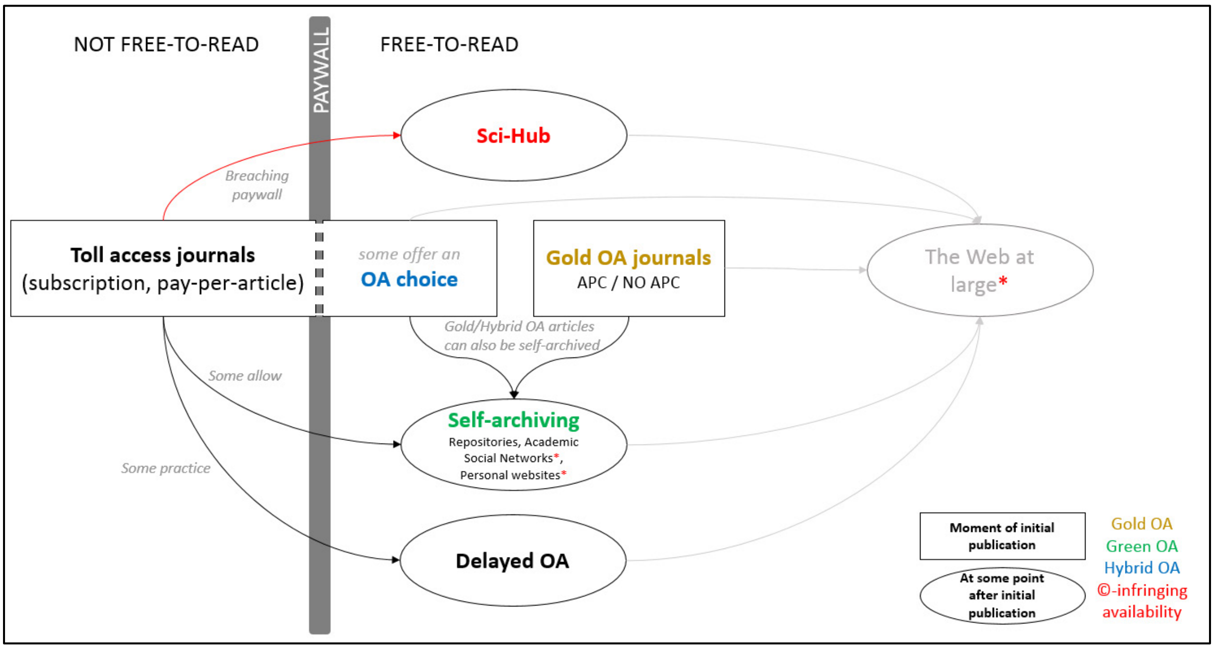

NOT FREE-TO-READ
PAYWALL
FREE-TO-READ
Sci-Hub
Breaching paywall
Toll access journals
(subscription, pay-per-article)
some offer an
OA choice
Gold OA journals
APC / NO APC
The Web at large*
Gold/Hybrid OA articles can also be self-archived
Some allow
Self-archiving
Repositories, Academic Social Networks*, Personal websites*
Some practice
Delayed OA
Moment of initial publication
At some point after initial publication
Gold OA
Green OA
Hybrid OA
©-infringing availability

Figure 1. Model of free availability of academic journal articles: Where are freely available journal articles hosted?

The figure divides articles in two different spaces: the space in which articles are *not free-to-read* (to the left of the paywall) and the space in which articles are *free-to-read* (to the right of the paywall).

Articles published in Gold OA journals (regardless of whether they charge APCs or not), and articles published in OA in Hybrid journals are immediately made OA, hence their placement in the *free-to-read* section of Figure 1.

Articles that are initially *not free-to-read* (published in toll access journals) may become *free-to-read* in several ways (represented by lines going from the *toll access journals* box to the *free-to-read* space in Figure 1):

- By breaching the paywall, generating copyright-infringing availability (represented in the figure with a red continuous line and red asterisks). This is the case of Sci-hub, which cannot really be considered as a sustainable form of OA (van Leeuwen, Meijer, Yegros-Yegros, & Costas, 2017).
- Via self-archiving, when the journal allows it (represented by a line from the *toll access journals* box to the *free-to-read* space). Self-archiving mostly takes place in repositories, academic social networks, and personal websites. Repositories (both institutional or subject-specific) usually check for copyright compliance when articles are submitted. In personal websites and academic social networks, however, no such checks are made. Therefore, these venues might also contain articles in violation of their copyright (Jamali, 2017). This is represented with red asterisks in Figure 1. There is also a line from Gold and Hybrid OA journals to the self-archiving section, because OA articles can always be self-archived.
- Delayed OA, which is practiced by some journals (also represented by a line from the *toll access journals* box to the *free-to-read* section).

Once articles are *free-to-read* in any of the ways described above, they may be distributed (legally or not) to any other part of the Web at large. For example, some platforms, like the academic search engines CiteSeerX and Semantic Scholar harvest the full texts of articles available in other sources, and provide a copy from their own servers.

Lastly, apart from being freely available, documents must also be discoverable in order to be used. There are several services that address the *discoverability problem*, like the academic search engines

BASE[91] and Google Scholar, or the browser extension Unpaywall[92]. Google Scholar and Unpaywall are described in more detail in the following section. Coverage of freely available documents varies by platform. Google Scholar, the focus of this article, serves as a gateway for all types of sources described in Figure 1, with the exception of Sci-Hub.

1.5. Quantification of OA levels

In a scenario like the one described above, it is not surprising that the question of how much of the scholarly literature is openly accessible (or at least freely available) has attracted much attention, because many agents of the scholarly community are interested in its answer. Funders are interested in the degree to which their OA mandates are being obeyed. Libraries need to decide how to best use their acquisitions budget (whether to renew, renegotiate, or cancel license agreements with publishers). Publishers routinely monitor how the documents they publish are shared on the Web in order to protect their business. Countries, for their part, want to know how much of the scientific literature published by its researchers is openly accessible. Researchers may also be interested in the proportion of their publications that is openly accessible, especially if this is an issue that is taken into account in the performance evaluations to which they are subjected in their country.

Numerous studies have analyzed the levels of OA for different samples of documents, presenting results at various levels of aggregation (publication year, subject areas, countries of authors' affiliations, OA types...). Methods to ascertain levels of OA include using data collected by custom crawlers (1science database, Unpaywall data) and carrying out searches in diverse search engines (BASE, Google, Google Scholar...). Table 1 contains information on the sample of documents analyzed, source of OA evidence used, and OA levels found by studies that used a source of OA evidence other than Google Scholar.

[91] https://www.base-search.net/

[92] http://unpaywall.org/

Table 1.Studies that analyse OA levels using sources of OA evidence other than Google Scholar

Study	Sample of documents					OA evidence				OA levels
	Source	Field	Pub. Year	Doc types	Size	Source	Date of data collection	Levels of aggregation	Methodological Observations	
Björk et al., 2010	Scopus (random)	All fields	2008	Articles	1,837	Searches on Google	2009/10	Subject areas, OA types		20.4% freely accessible (8.5% from publisher)
Gargouri, Larivière, Gingras, Carr, & Harnad, 2012	Web of Science (random)	11 fields	1998-2006	Articles	110,212	Custom crawler (no details given)	2009	Publication year, subject areas, OA types		20% freely accessible (average of entire period)
	=	14 fields	2005-2010	=	107,052	=	2011	=		24% freely accessible (average of entire period). 21.4% as Green OA, 2.4% as Gold OA
Archambault et al., 2014	Scopus (random)	All fields	1996-2013	Articles	~ 245,000	Custom crawler: Scielo, PubMed Central, ResearchGate, CiteSeerX, publisher websites, arXiv, repositories in ROAR and OpenDOAR	2013/04, 2014/04	OA types	Calibration factor (1.146) applied to account for limited recall of custom crawler	Over 50% of articles published 2007-2012 were freely available in 2014
	Scopus (random)	22 fields	2008-2013	Articles	~ 1 million	=	2014/04	Subject areas, countries (ERA)	=	Top OA field (2011-2013): General Science & Technology (90%) Top OA countries (2008-2013): Netherlands, Croatia, Estonia, and Portugal (>70%)
van Leeuwen et al., 2017	Web of Science (all records)	All fields	2009-2014	All types	Not declared	DOAJ, ROAD, CrossRef, PubMed Central, OpenAIRE	2017	Publication year, OA evidence source, countries		Almost 30% of articles were OA. Top countries: Netherlands (37%), Sweden, Ireland, and UK (34%)
Smith et al., 2017	PubMed (selected subject heading)	Global Health	2010-2014	Articles	3,366	PubMed, manual searches on Google	2016	OA types		29.2% OA from publisher, 27.2% Green OA, 1.3% OA from other sources. Total OA: ~ 58%
Science-Metrix Inc., 2018	Web of Science (all records)	All fields	2006-2015	Not declared	Not declared	1science database: scholarly material indexed in over 180,000 websites	2016/07-09	Publication year, countries, Subject areas, OA types	Calibration factor (1.2) applied. PubMed Central considered Gold OA; ResearchGate considered Green OA	Pub. Year 2006: 50%, pub. year 2011: 60% Top countries 2014: Brazil (74%), Netherlands (68%) Top fields: Health Sciences (59%)
Piwowar et al., 2018	CrossRef (random)	All fields	All years	Articles	100,000	Unpaywall data	2017/05	Publication year, publisher, OA types	ResearchGate not included in Unpaywall	27.9% are OA; 44.7% for pub. year 2015
	Web of Science (random)	All fields	2009-2015	Articles and reviews	100,000	Unpaywall data	2017/05	Subject areas, OA types	=	36.1% are OA
	Unpaywall use logs	All fields	All years	All types	100,000	Unpaywall data	2017/06/05-11	OA types	=	47% of documents accessed by users via Unpaywall are OA
Bosman & Kramer, 2018	Web of Science (all records)	All fields	2010-2017	Articles and reviews	12.3 million	Unpaywall data integrated in Web of Science	2017/12/20 – 2018/01/05	Publication year, Subject areas, languages, countries, institutions, funders	ResearchGate not included. Preprints not included	Almost 30% OA for pub. year 2016

1.5.1 Google Scholar as a source of OA evidence

Google Scholar has become one of the most widely used tools for researchers to search scientific information (Bosman & Kramer, 2016; Mussell & Croft, 2013; Nicholas et al., 2017; Van Noorden, 2014a). By automatically parsing the entire academic web instead of indexing only some specific sources, Google Scholar's coverage is much more extensive than the coverage of any other multidisciplinary commercial databases like Web of Science and Scopus. Although there are not official figures on the size of its document base, it was estimated in approximately 170 million records in 2014 (Orduna-Malea, Ayllón, Martín-Martín, & Delgado López-Cózar, 2015). Recently, Google Scholar's chief engineer, Anurag Acharya, has declared that the size of its document base is "larger than the estimates that are out there" (Rogers, 2017).

An important feature of Google Scholar is that it usually provides links to freely available versions of the documents displayed in its results page, also when the document is not openly accessible from the publisher website. Unfortunately, despite the wealth of information available in Google Scholar, the platform does not provide a way to easily extract and analyse its data (something like an open API), reportedly because the agreements that Google Scholar had to reach with publishers to access their content preclude this (Van Noorden, 2014b). Perhaps because of this limitation, all OA-related studies based on Google Scholar data either used very small samples of documents, mostly focusing on specific case studies, or the samples of documents they analyzed were not random because the selection of documents relied on searches in the platform, and Google Scholar is known to rank documents primarily, although not only, on descending order of citations (Martin-Martin, Orduna-Malea, Harzing, & Delgado López-Cózar, 2017). Moreover, most of these studies only analyzed the links to freely accessible full texts that are displayed beside the primary version of the document in Google Scholar, but not the links available in the secondary versions (see Figure 2). Table 2 contains information on the sample of documents analyzed, source of OA evidence used, and OA levels found by studies that used Google Scholar as a source of OA evidence.

These studies all pointed to the value of Google Scholar as a source of free availability of scientific literature, but were limited in scope and thematically. Thus, it is still missing in the literature a relatively large-scale study of the free availability of scientific publications that can be identified through Google Scholar. This paper aims at filling this gap.

Table 2. Studies that analyse OA levels using Google Scholar as a source of OA evidence

Study	Sample of documents					OA evidence			OA levels
	Source	Field	Pub. Year	Doc types	Size	Source	Date of data collection	Levels of aggregation	
Christianson, 2007	Journals in CSA's Ecology Abs. and JCR: Ecology (random)	Ecology	1945-2005	Articles	840	Google Scholar	2005/03	Only total figure	9% of the articles were freely accessible from Google Scholar
Norris, Oppenheim, & Rowland, 2008	Web of Science (selected journals)	Ecology, Appl. Math., Sociology, Economics	2003	Articles	4,633	OAIster, OpenDOAR, Google, Google Scholar	Not declared	Subject area	Economics: 65%; Appl. Math.: 59%; Ecology: 53%; Sociology: 21%. Overall OA: 49%
Pitol & De Groote, 2014	Web of Science (organization search)	Psychology, Chemistry, Electrical Engineering, Earth Sciences	2006-2011	Articles	982	Google Scholar	Not declared	OA version provider, OA type	70% of documents were freely accessible in some form
Khabsa & Giles, 2014	Microsoft Academic Search (random sample)	All fields	All years	Not specified	1,500 (100x15)	Google Scholar	2013/01	Subject areas	Top OA categories: Computer Science (50%), Multidisciplinary (43%), Economics & Business (42%). Overall OA: 24%
Jamali & Nabavi, 2015	Google Scholar (topic search)	All fields	2004-2014	All except citations and patents	8,310	Google Scholar	2014/04	Subject areas, OA types	Top OA category: Life Sciences (66.9%). Lowest OA category: Health Sciences (59.7%). Overall OA: 57.3%
Laakso & Lindman, 2016	Scopus (selected journals)	Information Systems	2010-2014	Articles	1,515	Google, Google Scholar	2015/02	Journal, OA types	60% of the articles were freely accessible from Google Scholar
Martín-Martín et al., 2016	Google Scholar (pub. year search)	All fields	1950-2013	All types	64,000	Google Scholar	2014/05	Publication year	40% of documents were freely accessible for the whole period. Over 66% considering only pub. years 2000-2009
Teplitzky, 2017	Pangaea (topic search)	Earth Sciences	2010, 2015	All types	744+482 = 1,226	Google Scholar	2016/05	OA types	75% of documents in pub. year 2010, and 72% in pub. year 2015
Abad-García, González-Teruel, & González-Llinares, 2018	Web of Science (funding search)	Health	2012-2014	Articles	762	OpenAIRE, BASE, Recolecta, Google Scholar	Not declared	Only total figures	46.3% of the documents were freely available from some source. Recall of Google was 93.5%
Mikki, Ruwehy, Gjesdal, & Zygmuntowska, 2018	Web of Science (topic search)	Climate and ancient societies	All years	All types	639	Google Scholar	Not declared	Publication years	74% of the documents were freely accessible
Laakso & Polonioli, 2018	Publication lists of ethics researchers	Ethics	2010-2015	Articles	1,682	Google Scholar	2017	Publication years, OA types	56% of the documents were freely accessible

1.6. Research questions

This paper mainly intends to ascertain the suitability of the data available in Google Scholar to gauge the levels of adoption of OA in scientific journal articles, across all subject categories and countries, thus overcoming the limitations related to sample selection and sample size of the previous OA-related studies that used this source of data. Specifically, this article aims to answer the following questions:

RQ1. How much of the recently published scientific literature is freely available according to the data available in Google Scholar, by year of publication, subject categories, and country of affiliation of the authors?

RQ2. How much is openly accessible in a sustainable and legal way, and what proportion is freely available but does not meet these criteria?

RQ3. What is the distribution of freely available documents by web domains?

2. Methods

The three main citation indexes of the Web of Science Core Collection (Science Citation Index Expanded [SCIE], Social Sciences Citation Index [SSCI], and Arts & Humanities Citations Index [A&HCI]) were used to select the sample of documents analysed in this study. All documents with a DOI indexed in either the SCIE, SSCI, or the A&HCI, and published in 2009 or 2014 were selected on the 19th of May, 2016. The rationale behind choosing these two years was that we wanted to analyse a large sample of documents from various publication years, but we also wanted to keep the sample manageable because of the difficulty of extracting data from Google Scholar. At the time of data collection, 2014 was the most recent year in which most articles scheduled to become OA after an embargo (Delayed OA) had already become OA. The data from articles published in 2009 would give us information on the trend.

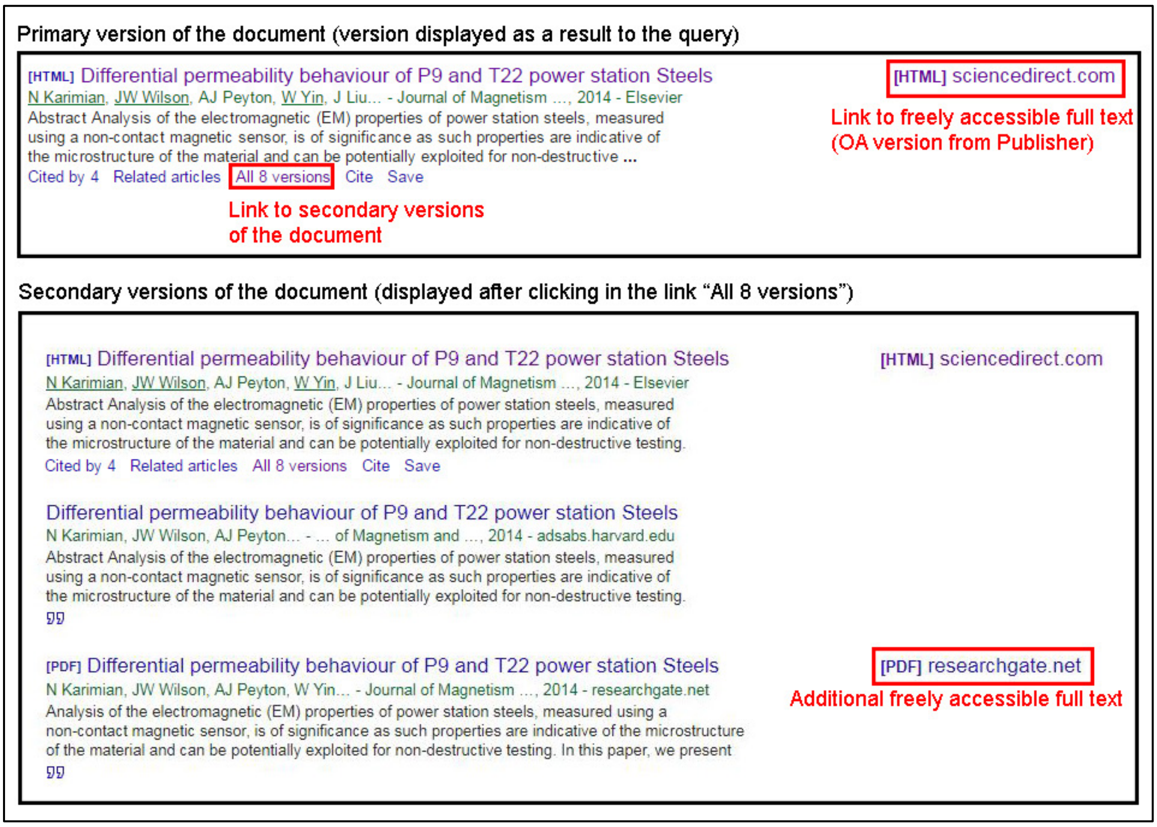
The records of these documents were extracted from the local version of the Web of Science database available at the Centre for Science and Technology Studies (CWTS) in Leiden University. A total of 2,610,305 records were extracted, 1,080,199 from 2009, and 1,530,106 from 2014. We decided to use this source (as opposed to the CrossRef registry) because it would later enable us to carry out detailed analyses of the data, with breakdowns by subject categories, country affiliations, publication years, and journals.

It is worth noting that the number of Web of Science documents in these two years (2009 and 2014) at the time of writing this article had increased from 2,610,305 to 2,893,175. This could have been caused by backwards indexing of new documents, or by the addition of DOIs to records that previously did not contain one in the Web of Science database.

Each of these documents was searched on Google Scholar, using a non-documented method to search documents by their DOI. Example of query for the document with DOI “10.1010/j.jmmm.2013.09.059”:

https://scholar.google.com/scholar_lookup?doi=10.1010/j.jmmm.2013.09.059

Given that Google Scholar does not provide an API to query its database, a custom Python script was developed to carry out a query for each of the DOIs in our sample and scrape the data from the results page. Queries were distributed across a pool of different IP addresses to minimise the amount of CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) that Google Scholar requests users to solve from time to time. However, this approach did not entirely suppress the appearance of CAPTCHAs, which were solved manually when the system requested them. Additionally, when it was detected that Google Scholar provided a link to a freely accessible full text of a document, the link to the secondary versions of the same document was also followed through, in order to extract all the additional links to freely accessible full texts of the document that Google Scholar might have found (Figure 2). Searches were carried out off-campus to avoid retrieving links to full texts that are only accessible through library subscriptions. The process of extracting the data from Google Scholar was very time-consuming, taking over three months (from the end of May to the end of August of 2016) to collect data for the 2,610,305 selected documents.



Primary version of the document (version displayed as a result to the query)

[HTML] Differential permeability behaviour of P9 and T22 power station Steels
N Karimian, JW Wilson, AJ Peyton, W Yin, J Liu... - Journal of Magnetism ..., 2014 - Elsevier
Abstract Analysis of the electromagnetic (EM) properties of power station steels, measured using a non-contact magnetic sensor, is of significance as such properties are indicative of the microstructure of the material and can be potentially exploited for non-destructive ...
Cited by 4 Related articles All 8 versions Cite Save

[HTML] sciencedirect.com

Link to freely accessible full text (OA version from Publisher)

Link to secondary versions of the document

Secondary versions of the document (displayed after clicking in the link "All 8 versions")

[HTML] Differential permeability behaviour of P9 and T22 power station Steels
N Karimian, JW Wilson, AJ Peyton, W Yin, J Liu... - Journal of Magnetism ..., 2014 - Elsevier
Abstract Analysis of the electromagnetic (EM) properties of power station steels, measured using a non-contact magnetic sensor, is of significance as such properties are indicative of the microstructure of the material and can be potentially exploited for non-destructive testing.
Cited by 4 Related articles All 8 versions Cite Save

[HTML] sciencedirect.com

Differential permeability behaviour of P9 and T22 power station Steels
N Karimian, JW Wilson, AJ Peyton... - ... of Magnetism and ..., 2014 - adsabs.harvard.edu
Abstract Analysis of the electromagnetic (EM) properties of power station steels, measured using a non-contact magnetic sensor, is of significance as such properties are indicative of the microstructure of the material and can be potentially exploited for non-destructive testing.

[PDF] Differential permeability behaviour of P9 and T22 power station Steels
N Karimian, JW Wilson, AJ Peyton, W Yin... - Journal of Magnetism ..., 2014 - researchgate.net
Analysis of the electromagnetic (EM) properties of power station steels, measured using a non-contact magnetic sensor, is of significance as such properties are indicative of the microstructure of the material and can be potentially exploited for non-destructive testing. In this paper, we present

[PDF] researchgate.net

Additional freely accessible full text

Figure 2. Example of primary and secondary versions of an article in Google Scholar

Using the search strategy described above, Google Scholar retrieved results for 99.3% of the documents searched. The system did not retrieve any results for 0.7% of the DOIs searched. However, this does not necessarily mean that these documents were not covered by Google Scholar. These documents might have been covered by Google Scholar without a DOI, and therefore they might have been found using other search strategies, for example, searching by the title of the document. However, we did not try other search strategies, as we considered the results could not be overly affected by these missing documents.

A test was also carried out to assess the accuracy of the results retrieved from Google Scholar. That is, whether or not we had actually retrieved data about the documents we were looking for. In order to do this, we compared the bibliographic information available from Web of Science, with the data extracted from Google Scholar. The match was considered successful if at least one of the following criteria were met:

- Similarity of document titles in the two sources of data (based on the Levenshtein distance of the two strings of text) was equal or greater than 0.8 (similarity is 1 when the titles are exactly the same, and 0 when they are completely different).
- Similarity of document titles was between 0.6 and 0.8 AND the documents shared the same first author AND the same publication year.
- Same first author and same publication year, and title of document in Google Scholar was not in English. In some cases when the journal publishes in a language other than English, the title provided by Google Scholar is the original title, whereas in Web of Science, the title of the document is always displayed in English (even when the document itself is not written in English). In these cases the title similarity was very low, and using it resulted in a significant number of false negatives.

Based on these criteria, we classified as good matches 96% of the documents in our sample (2.51 million documents). The proportion of good matches was slightly higher if we only considered documents of the type "article" or "review" (97.6%). Therefore, we decided to analyze only the articles and reviews in our sample that we had considered as good matches, a total of 2,269,022 documents.

Google Scholar does not provide any information on the type of source that is providing free access to the full text of a document. For this reason, we combined information from a variety of sources in order to provide more detailed information about the type of free access that Google Scholar had been able to detect. We classified each *full text link* in one of the following categories:

- Publisher: when the full text is hosted on a publisher website, or on journal aggregators such as JSTOR or SciELO. Data from the oaDOI dataset from 18 August 2017, DOAJ (Directory of Open Access Journals), and the Ulrich's Directory of Journals was used to create a list of websites where journal publishers make their articles available.
- Repository: when the full text is hosted in a repository, as defined by the Registry of Open Access Repositories (ROAR), and the Directory of Open Access Repositories (openDOAR).
- Research Institutions: when the full text is hosted in the web domain of a research institution (universities, research centers, institutes), excluding the website of the institutional repository. That is to say, this category mostly contains personal websites of individual researchers, research groups, departments, etc. inside an academic domain. In order to determine which domains belonged to academic institutions, a list of academic domains was also extracted from openDOAR.
- Academic Social Networks: in this category we only classified the full texts available from ResearchGate and Academia.edu.
- Harvesters: websites that copy full texts from other sources and make them available from their own servers. In this category we classified full texts hosted in the search engines CiteSeerX and Semantic Scholar, and the British CORE service.
- Non-categorized: any website that could not be classified in the previous categories.

After combining the information from the sources described above, there were still thousands of web domains that had not been classified. Therefore, we decided to manually check the hosts with a higher number of occurrences in our sample that still had not been categorised. Specifically, we checked the domains in which Google Scholar had found 100 or more full texts of documents in our sample, and the hosts that Google Scholar more frequently selected as the primary full text version (because these hosts would likely be publishers, as declared in Google Scholar's publisher guidelines[93]). Thus, approximately 1,000 hosts were classified after visiting the website and checking it manually. The rest of the web domains that had not been classified were considered as "non-categorized". The specific categorisation of hosts used in this study is available in the complementary material to this article[94].

In this article we make a distinction between Freely Available (FA) documents, and OA documents. We consider that all documents for which Google Scholar provided a link to a FA version of the document, regardless of the legality under which they were shared and their sustainability over time, are FA. When FA documents meet certain additional criteria (described below) they were also considered OA.

Unfortunately, there is no clear consensus regarding the minimum rights that any user should have in order to be able to consider a document OA. Some definitions, like the one declared by the BOAI or the Open Definition[95] are clear in that mere right to access the document free of charge is not enough to consider a document OA. They consider it necessary that the license extends other rights to all users, like redistribution, modification, or application for any lawful purpose. The reality, however, is that in many cases documents are made FA under licenses that fail to meet one or several of these criteria. For example, there are Creative Commons licenses that include Non-Commercial and/or Non-derivatives clauses, thus limiting the ways in which a document can be reused. The Elsevier user

[93] https://scholar.google.com/intl/en/scholar/publishers.html#policies
[94] https://osf.io/fsujy/
[95] http://opendefinition.org/

license[96] (the license under which Elsevier makes FA after an embargo period articles published in journals included in its Open Archive[97]) prohibits redistribution of the documents and reuse for commercial purposes. Moreover, there is a large portion of articles that publishers make available free of charge, without extending any other rights to users other than access. This is usually called "public access" in the publishing industry (Crotty, 2017). These issues have led some researchers to think in terms of degrees of openness, instead of considering OA a binary quality (Chen & Olijhoek, 2016).

Apart from the conceptual issues, there are also practical limitations for classifying documents as OA. In many cases, especially when we are talking about Green OA, there is no license attached to the document, or it is attached in a way that cannot be easily detected by automated systems. Fortunately, publishers are increasingly taking to sending license information to CrossRef (which makes these data openly accessible) or they display it as metadata in their own websites.

For the reasons described above, in this article we use a more inclusive definition of OA than the one declared by the BOAI or the Open Definition, and we instead set our focus on sustainability and legality. Specifically, this article considers the following types of OA:

- Gold OA: when the journal that published the article was listed in DOAJ.
- Hybrid OA: when the journal was not listed in DOAJ but an OA license was recorded in the metadata available in CrossRef, and the Open license came into effect at the same time the article was published (OA immediately upon publication). We considered as OA licenses all Creative Commons licenses, the Elsevier OA user license, and other OA licenses registered in CrossRef by publishers like the ASPB[98], ACS[99], and IEEE[100]. Our operational definition of "OA immediately upon publication" was that the value recorded in the *delay-in-days* field of the License element available in the CrossRef metadata (defined as the "[n]umber of days between the publication date of the work and the start date of this license"[101]), should be less than 30 (one month). We decided to set this limit instead of *delay-in-days* = 0 because we noticed that for some articles published as OA, the Open license came into effect a few days after publication, and we considered that these articles should also be classified as "OA immediately upon publication".
- Delayed OA: when the journal was not listed in DOAJ but an Open Access license was recorded in the metadata available in CrossRef, and the Open license came into effect more than 30 days after the publication of the article.
- Bronze OA: when the full text is FA from the publisher, but the journal is not listed in DOAJ and no OA license could be found. This category includes gratis / public access from the publisher (free to read but the publisher retains copyright), but might also contain masked Hybrid or Delayed OA (when the publishers fail to disclose an OA license in machine-readable form), and possibly even some masked Gold OA (if a full OA journal is not listed in DOAJ and the publisher does not discloses an OA license).
- Green OA: the documents that are FA from institutional or subject-based repositories, as listed in ROAR and OpenDOAR.

All the documents that were available from sources other than the publisher website and repositories (such as websites of research institutions excluding the repository, academic social networks, harvesters, and the rest) were only considered as FA, and not OA. We took this conservative measure because we wanted to make a distinction between more legally sound and sustainable sources (publishers and repositories) which are more likely to be copyright-compliant and usually implement long-term preservation plans for the documents they host, and less stable sources (personal websites,

[96] https://www.elsevier.com/open-access/userlicense/1.0
[97] https://www.elsevier.com/about/open-science/open-access/open-archive
[98] https://aspb.org
[99] https://pubs.acs.org
[100] https://www.ieee.org
[101] https://github.com/CrossRef/rest-api-doc/blob/master/api_format.md

academic social networks…) where any document (regardless of its copyright status) can be uploaded and deleted at any time.

Lastly, Google Scholar does not provide data on the publication stage of the freely accessible versions that it finds: that is, whether the free version is a preprint (before peer-review), an author's accepted manuscript (after peer-review, but before typesetting), or the journal's version of record (final published article). Although this is an interesting aspect of OA publishing, identifying the type of version would have required accessing the full text of each individual article, and so it falls outside the scope of this study.

Data was processed and analyzed using the R programming language. The percentages of OA documents were computed by publication year, subject category, country of affiliation (considering all co-authors), and journal. The data used in this study is openly available (Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018). This will facilitate the creation of custom analyses that focus on the research done in specific countries, specific fields, specific journals, etc.

3. Results

3.1. General overview

Google Scholar provided links to FA full texts for 54.7% of our sample of documents (Figure 3). If we break down the results by year of publication, documents published in 2014 show a slightly higher percentage of FA documents (55.8%) than documents published in 2009 (53%), even though the number of documents published in 2014 (1,331,795) was larger than the number of documents published in 2009 (937,227), and the fact that at the time of data collection documents from 2014 had had considerably less time to be made freely available on the Web than documents from 2009.

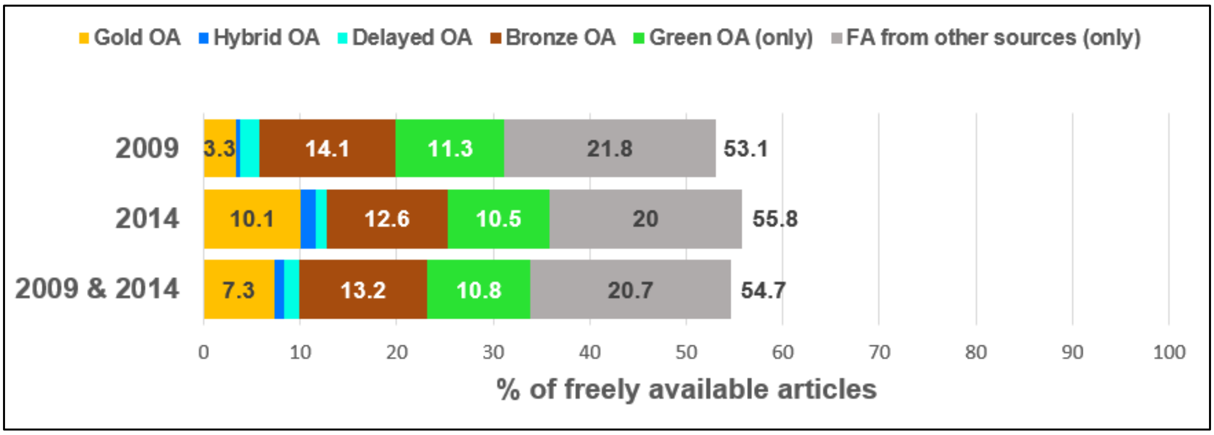


Figure 3. Overall OA and FA levels found in Google Scholar, by year of publication and both years combined

If we consider the two years under study (Figure 3), we can see that 23.1% of the documents are FA from publisher websites (Gold + Hybrid + Delayed + Bronze). It is worth noting that most of the documents available from publishers are Bronze OA, which are usually made accessible under very restrictive reuse terms. However, it seems like Gold and Hybrid are gaining importance, judging by the increment from 3.3% to 10.1% from 2009 to 2014 of Gold OA, and from 0.5% to 1.5% for Hybrid OA. Bronze OA decreased from 14.1% to 12.6%, and Delayed OA decreased as well (from 2% in 2009 to 1.1% in 2014).

Figure 3 displays OA provided by the publisher (Gold, Hybrid, Delayed, Bronze), Green OA, and FA from other sources. However, in the cases where a document is available from several types of sources, publisher-provided OA is given preference over Green OA and FA from other sources. In a similar manner, Green OA versions are given preference over FA from other sources. Therefore, Figure 3 does not display the total percentages of Green OA and FA from other sources. These are displayed in Figure 4.

The proportion of documents available as Green OA (repositories) was higher in the publication year 2014 (18.9%) than in 2009 (15.7%), as displayed in Figure 4. However, the number of documents that were available from repositories and not from the publisher (displayed in Figure 3) was slightly higher in the publication year 2009 (11.3%) than in 2014 (10.5%).

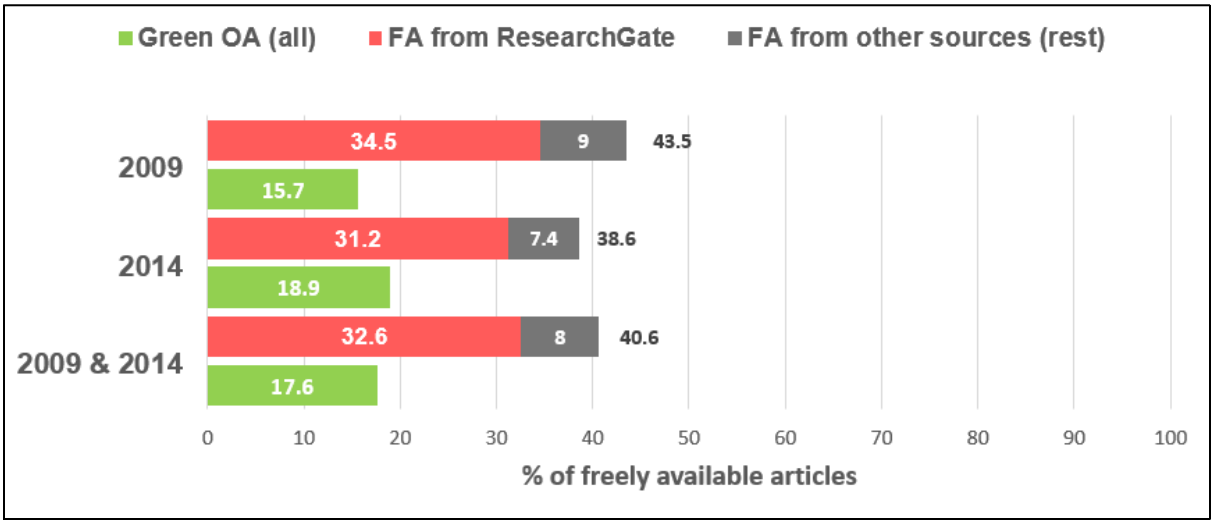


Figure 4. Total percentage of Green OA and FA found in Google Scholar, by year of publication and both years combined

Apart from publisher websites and repositories, there is a large fraction of documents that are available from other sources (mainly the academic social network ResearchGate, but also personal websites, and harvesters). Google Scholar found that 43.5% of the documents in the sample published in 2009 were available from other sources (Figure 4). This percentage was lower in the publication year 2014 (38.6%). Nevertheless, in both years this percentage is larger than the sum of what all publishers and repositories together provided. Moreover, a considerable portion of these documents are FA only from these other sources (that is, these documents are not openly accessible from the publisher or from repositories). This figure remains relatively stable in the two publication years (21.8% in 2009, and 20% in 2014), as can be observed in Figure 3.

The predominance of sources other than publishers and repositories can also be observed if we take a look at the number of freely available documents by website (Table 3). By far, the source that provided more freely available full texts was the academic social network ResearchGate, which by itself provided access to 32.6% of the documents in our sample (738,573). If we compare this figure to the percentage of documents provided as OA by publishers available in Figure 3 (23.1%, approx. 525,000 documents), we see that ResearchGate provided access to more documents in our sample than all publishers together. Moreover, 32.7% of the documents available from ResearchGate (over 240,000) were not freely available from any other source.

Table 3 also shows how often Google Scholar displays links from each host as the primary full text links. This is interesting because the primary link is likely to be the link that most Google Scholar users click to access the full text of an article. Again, ResearchGate is first in the rank, followed by Pubmed Central (www.ncbi.nlm.nih.gov) and arXiv. However it is worth noting that some hosts that provided many FA documents (europepmc.org, academia.edu, citeseerx.ist.psu.edu) are rarely selected by Google Scholar as the primary full text links (only in 10.3%, 14.1%, and 9.3% of the cases, respectively), meaning that the documents these platforms provide are also available from other platforms which are placed higher in Google Scholar's host precedence rules. Regarding these precedence rules, the data in Table 3 shows that Google Scholar does indeed tend to select the publisher version as the primary version whenever it is an option (as stated in its indexing policies). Most publisher websites are selected as the primary full text version in over 90% of the cases. The exceptions seem to be Springer and BioMed Central, which are only selected as the primary version in about 45% of the cases. Lastly, it appears that Google Scholar chooses the arXiv repository even over most publishers, as this repository is selected as the primary source of full text in 99.9% of the cases. This means that when an article is openly accessible from arXiv, Google Scholar always chooses the arXiv version as the primary full text version, presumably even when the article is also openly accessible from the publisher.

Table 3. Top 20 websites according to the number of FA full texts they host.

Host	Type	# of FA documents	% as only FA provider	# of FA as primary version	% as primary version
www.researchgate.net	Social network	738,573	32.7	323,372	43.8
europepmc.org	Repository	177,930	5.1	18,312	10.3
www.academia.edu	Social network	168,485	4.2	23,681	14.1
www.ncbi.nlm.nih.gov	Repository	165,403	1.8	74,109	44.8
citeseerx.ist.psu.edu	Harvester	120,378	1.8	11,203	9.3
arxiv.org	Repository	72,862	25.0	72,753	99.9
onlinelibrary.wiley.com	Publisher	49,887	32.8	47,712	95.6
www.sciencedirect.com	Publisher	47,356	26.1	43,825	92.5
pdfs.semanticscholar.org	Harvester	38,164	1.0	2,790	7.3
journals.plos.org	Publisher	37,984	12.5	37,380	98.4
link.springer.com	Publisher	35,295	6.2	15,335	43.4
www.biomedcentral.com	Publisher	27,400	2.1	12,328	45.0
www.nature.com	Publisher	23,726	26.1	21,699	91.5
downloads.hindawi.com	Publisher	18,566	38.8	18,565	100.0
core.ac.uk	Harvester	15,344	1.4	769	5.0
pubmedcentralcanada.ca	Repository	14,286	1.0	461	3.2
hal.archives-ouvertes.fr	Repository	11,293	10.7	5,530	49.0
www.mdpi.com	Publisher	11,084	12.9	11,083	100.0
www.infona.pl	Repository	10,060	41.4	6,132	61.0
www.tandfonline.com	Publisher	8,973	61.2	8,730	97.3

3.2. Analysis by disciplines

We mapped the original WoS subject categories to more general classification schemes: one containing 7 broad subject areas, and the other containing 35 scientific disciplines. The schemes were introduced by Tijssen et al. (2010), and the specific correspondence with WoS categories is available in the complementary materials.

There is a high inter-area variability, ranging from 60% overall availability in the Medical and Life Sciences, to 32.3% overall availability in Law, Arts, and Humanities (Figure 5). Multidisciplinary journals achieve a 93.6% overall availability, which is natural if we consider that this category includes Gold OA multidisciplinary mega-journals such as PLOS ONE.

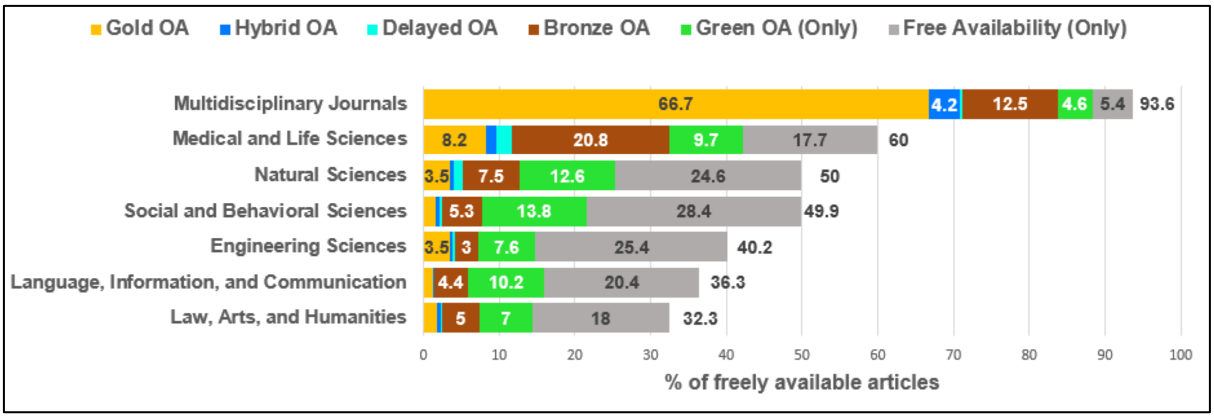


Figure 5. OA and Free Availability levels found in Google Scholar, by broad subject areas.

If we descend to the level of disciplines (Figure 6) we can see that Bronze OA is usually the predominant type in which publishers provide OA. In 28 out of the 35 disciplines shown in Figure 6, the percentage of Bronze OA is higher than the sum of Gold, Hybrid, and Delayed OA. Bronze OA is especially important in Basic Life Sciences, Biomedical Sciences, and Clinical Medicine.

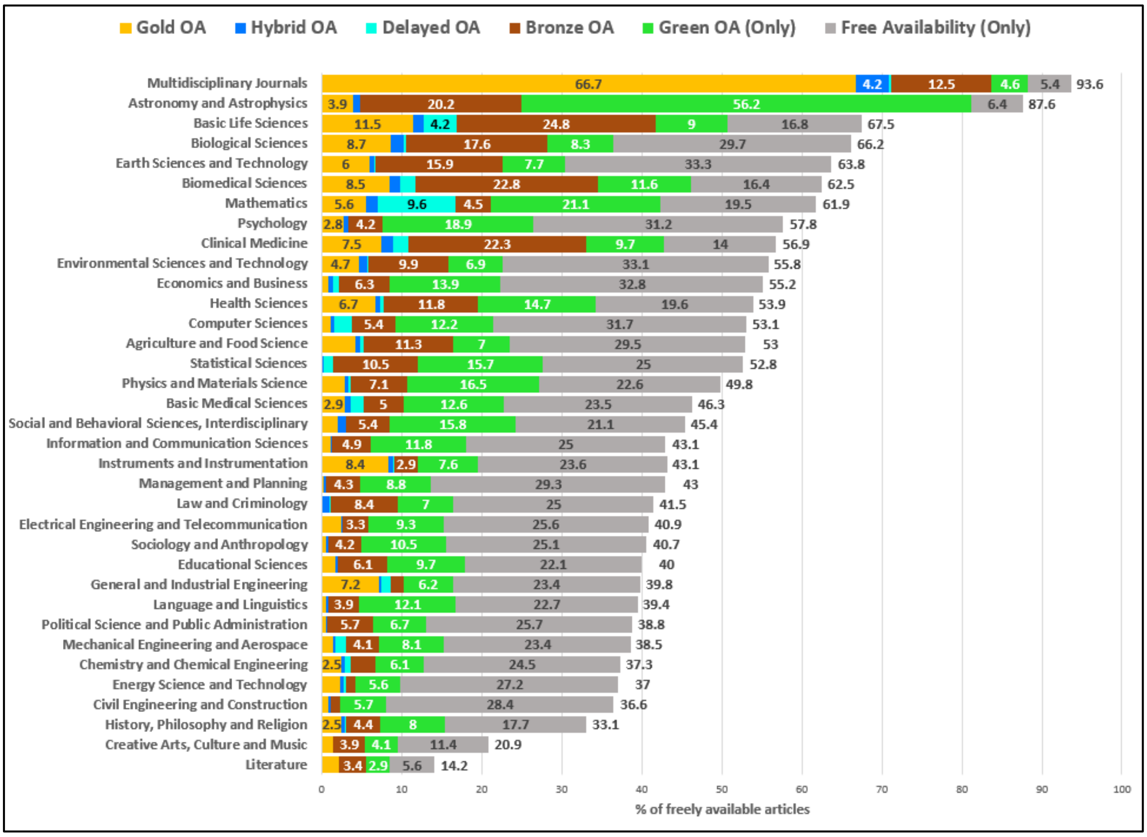


Figure 6. OA and Free Availability levels found in Google Scholar, by scientific discipline

Figure 6 also shows the percentage of articles in Green OA that are not openly accessible from the publisher: Green OA (only)[102]. In 19 out of the 35 disciplines, the number of documents that are accessible only through Green OA was higher than the sum of Gold, Hybrid, Delayed, and Bronze OA. The disciplines with a larger share of documents in the Green OA (only) category are Astronomy and Astrophysics (56.2%), and Mathematics (21.1%).

If we consider FA only (the cases when documents were only available from sources other than publishers and repositories), Figure 6 shows that this is the most frequent type of availability in most disciplines. In 23 out of the 35 disciplines, FA (only) achieves higher percentages than Gold, Hybrid, Delayed, Bronze, and Green combined. In four of these disciplines (Management and Planning, Political Science and Public Administration, Energy Science and Technology, and Civil Engineering and Construction), more than two thirds of the documents that were FA in some form, were only available from sources other than the publisher or repositories.

Lastly, it is worth noting that there is a large degree of intra-discipline variability as well. Figure A2 in the complementary materials[103] displays the correspondence between the 35 disciplines in Figure 6, and the subject categories used by the Web of Science. This figure shows that in many cases there are important differences among the categories of a discipline, regarding not only the overall free availability of documents, but also the types of availability. If we take Clinical Medicine (56.9% overall free availability), for example, the subject categories with the highest overall availability are *Tropical*

[102] Total percentages of Green OA by subject categories (including the cases when the article is also openly accessible from the publisher) are available from the complementary materials and in the web application.
[103] https://osf.io/fsujy/

Medicine (85.9%), and *Andrology* (84.7%). Both categories also present high levels of OA provided by the publisher (over 70%). *Dermatology,* however, presents a completely different behavior: only 37% of the documents are freely available in some way, and the most common type of availability is FA from other sources (14.5%).

3.3. Analysis by countries of affiliation

Table 4 displays OA and FA levels of countries with an output equal or higher than 1% of the total, considering only documents published in 2014 (the most recent year in our sample). The affiliation of all co-authors of the articles were considered (each article was considered once for each different country of affiliation). It distinguishes between OA provided by the publisher, OA from repositories (when OA from publisher is not available), and FA from any other sources (when OA from publisher or from repositories is not available). A green background in one of the cells of the table indicates that the value in that cell is higher than the World value (visible in the first row below the headers). A red background indicates a value lower than the World value. Higher color intensity indicates a higher distance relative to the World value. The last column (% OA + FA) highlights the top three countries with a higher overall availability (in green) and the top three countries with a lower overall availability (in red).

Table 4. OA and Free Availability (FA) levels for documents published in 2014 by researchers in countries with high output (>1% of the total)

Country	Documents	% OA from publisher	% OA from repositories*	% OA Total	% FA other sources†	% OA + FA†
World	1,331,795	25.3	10.5	35.8	20.0	55.7
USA	360,889	29.1	18.2	47.3	18.9	66.2
Peoples R China	231,162	22.9	4.3	27.2	18.7	**46.0**
Germany	96,265	28.6	13.4	42.0	19.2	61.3
England	89,996	35.0	15.9	50.9	17.3	68.3
Japan	71,587	26.6	9.9	36.5	13.4	49.9
France	66,648	26.5	17.4	43.9	23.5	67.4
Canada	60,342	28.1	10.5	38.6	23.1	61.7
Italy	58,397	26.2	11.9	38.1	25.6	63.7
Australia	53,822	26.2	10.5	36.7	24.9	61.7
Spain	51,586	25.3	13.9	39.2	24.7	63.9
South Korea	51,036	26.2	5.4	31.6	17.9	49.5
India	50,468	15.7	7.4	23.1	25.6	48.7
Netherlands	36,228	33.7	14.2	47.9	22.9	**70.8**
Brazil	34,517	37.0	8.8	45.8	25.8	**71.6**
Russia	28,108	10.6	9.7	20.3	23.9	**44.3**
Switzerland	26,580	33.8	14.9	48.7	21.8	**70.5**
Taiwan	25,492	27.3	8.4	35.7	17.5	53.2
Sweden	24,286	35.3	14.9	50.2	19.2	69.4
Iran	23,387	14.5	4.1	18.6	26.4	**45.0**
Turkey	21,516	22.8	5.8	28.6	23.9	52.5
Poland	20,496	33.4	9.6	43.0	20.7	63.8
Belgium	19,809	29.5	15.7	45.2	24.2	69.4
Denmark	15,853	34.9	12.4	47.3	20.2	67.5
Scotland	13,813	38.3	18.3	56.6	16.4	73.0
Austria	13,514	34.9	12.2	47.1	19.3	66.4

***** Accessible from repository but not from publisher

† Only available from other sources

All countries in Table 4 present higher percentages of OA from publishers than of OA only from repositories. 18 out of the 25 high output countries displayed in Table 4 present OA levels (sum of OA from publisher and OA from repositories) that are higher than the World level (35.8%). 13 of these countries are in Europe. The other five are the USA, Japan, Canada, Australia, and Brazil. The countries with the highest percentages of OA come very close to or slightly surpass 50% of the total amount of documents published by researchers in that country (United Kingdom, Sweden, Switzerland, Netherlands, USA, Denmark, Austria). All these countries present percentages of OA from publishers and from repositories that are higher than the average world percentages. Japan, Brazil, and Poland also have higher than average OA levels, with the particularity that most of their OA is available from publishers, and their percentage of OA from repositories is lower than the World level. The opposite, however, does not occur: there are no countries in Table 4 with a lower than average percentage of OA from publishers that manage to achieve a higher than average total percentage of OA thanks to OA from repositories.

7 out of the 25 high output countries displayed in Table 4 present OA levels that are lower than the World level (35.8%). Chief among them is China, with only 27.2% of its documents accessible either from the publisher or from repositories, even though it is the second country in terms of output (231,162 articles and reviews published in 2014). The other six countries are also located in Asia (South Korea, India, Russia, Taiwan, Iran, and Turkey).

At the world level, 20% of the documents are only freely available through sources other than publishers and repositories. At the country level there is some variation: from the 13.4% percent of documents written by Japanese researchers that are only available from these other sources, to the cases of Italy, India, Brazil, and Iran, where the percentage is slightly over 25%.

If we consider overall availability (the sum of OA and FA only), the countries with a higher percentage of availability are Brazil (71.6%), the Netherlands (70.8%), and Switzerland (70.5%). Scotland deserves a special mention, because if considered separately from the rest of the United Kingdom (which is the way the Web of Science presents authors' affiliations), it achieves 73% overall free availability. The United Kingdom as a whole presents a slightly lower percentage (68.7%). In the lowest positions of the rank we can find China (46%), Iran (45%), and Russia (44.3%).

Table A1, available in the complementary materials[104], extends Table 4, displaying the same information for 40 additional countries, those with an output larger than 0.1% and lower than 1% of the World total. The countries with a higher overall availability in this output tier are Kenya (1,504 documents, 80.6% overall availability), Chile (5,812 documents, 76% overall availability), and Norway (11,601 documents, 67.9% overall availability), and the countries with a lower overall availability are Tunisia (3,008 documents, 50.3% overall availability), Ukraine (4,397 documents, 49.1% overall availability), and Algeria (2,139 documents, 43.1% overall availability).

4. Discussion

4.1. Limitations and further lines of study

The analysis carried out in this study suffers from a number of limitations. These are related either to the sample selection, to the data available in Google Scholar, to the categorisation of OA / FA of the documents in the sample, or to the replicability of the study.

The first limitation of this article related to sample selection is that it only analyses scientific journal articles and reviews published in journals indexed in Clarivate Analytics' SCIE, SSCI, and A&HCI. These three citation indexes are known to have limited coverage of journals in the Social Sciences, Arts, and Humanities (SSAH), and to suffer from a bias towards English-language journals (Mongeon & Paul-Hus, 2016; Van Leeuwen, Moed, Tijssen, Visser, & Van Raan, 2001). Therefore, results might have been different if more articles published in journals in the SSAH that are not covered by these indexes, and/or more articles from journals that publish in languages other than English had been included in the sample. Furthermore, this study focuses on the OA levels of articles and reviews, and not on the OA levels of other document types such as books, conference papers, or scientific reports. Further studies could focus on the free availability of these other document typologies, which Google Scholar also covers.

An additional limitation is that this article only considers articles and reviews for which a DOI was available in Clarivate Analytics' citation indexes at the time of data collection. Documents without a DOI, or documents for which a DOI had been minted but was not recorded in these databases at the time of data collection, have not been considered in this study.

Regarding the data extracted from in Google Scholar, this study has the following limitations:

1. This study only analyses OA evidence in Google Scholar of documents published in 2009 and 2014 at a specific moment in time: summer of 2016. Therefore, no extrapolation should be made regarding OA levels of other publication years. Furthermore, OA levels of documents published in 2009 and 2014 might have changed by the time of this writing, caused by OA

[104] https://osf.io/fsujy/

backfilling: documents that have become OA after we collected the data, either because the publisher practices Delayed OA, or because authors have self-archived their articles. It also may be the case that some documents that were available when we collected the data are no longer available. The dispute between the Coalition for Responsible Sharing and ResearchGate, in which ResearchGate was forced to remove from public view a significant number of articles that infringed copyright, may have affected the current levels of free availability of the documents in our sample. Additionally, some documents hosted in other unstable sources, such as personal websites, may have also been removed.

2. In some cases, Google Scholar failed to recognize that an article was freely available from a source that the search engine indexes. In practice, this takes the form of a record in which no FA link is provided to the right of the main bibliographic information (see Figure 2), but if users would follow the link available in the title of the document, they would find that the article is in fact freely available. Our study only considers the links that Google Scholar provides to the right of the bibliographic information, and therefore, our results undercount free availability in these cases. We are aware that some journals (for example, some Gold OA journals published by *Frontiers*, and also *eLife*) were affected by this problem. We have also noticed that Google Scholar has fixed these errors for the most part, and at the time of this writing, FA links are correctly displayed to the right of the bibliographic information of the articles published in the aforementioned journals.
3. In some cases, Google Scholar is not able to successfully merge all the different versions of an article that can be found on the Web (Martín-Martín et al., 2014; Orduna-Malea, Martín-Martín, & Delgado López-Cózar, 2017), and as a result, two or more entries might exist in Google Scholar for documents that are actually the same. This might happen for a number of reasons, but is more frequent in journals that publish several versions of the same document (i.e. versions in several languages), and also for journals that, even though they publish only in one language, create versions of the article metadata in several languages. In these cases, Google Scholar's algorithms to detect duplicate documents usually fail. For our study this means that in some cases, the record we retrieved from Google Scholar might be one that does not provide a link to a freely available version, even though other entries of the same document in Google Scholar might contain such links. Therefore, our study undercounts free availability in these cases as well. One journal in our sample that is affected by this problem is *Revista Espanola de Documentacion Cientifica*, a Gold OA journal for which our data shows FA links in only 56% of the documents it published in 2009 and 2014.

Regarding the categorization of documents as OA / FA, and its specific subtypes (Gold OA, Hybrid OA, Delayed OA, Bronze OA, and Green OA, as well as FA only from sources other than the publisher and repositories), there are several limitations that should be taken into account.

1. We considered as Gold OA only the articles published in journals included in the Directory of Open Access Journals (DOAJ). There are, however, journals that adhere to the Gold OA model that are not included in this directory, like, for example, some journals owned by the Korean Association of Medical Journal Editors (*Korean Journal of Radiology*, and *Korean Journal of Physiology & Pharmacology*, for example). Because our study relies on DOAJ, it suffers from this limitation, and articles published in these journals are miscategorized either as Hybrid OA (when an Open License could be found) or as Bronze OA (if the journal does not deposit license information in CrossRef) Therefore, our study might be overestimating Hybrid and Bronze OA in detriment of Gold OA. Nevertheless, the error introduced by this issue in our calculation of Hybrid OA is estimated to be fairly small, as the total sum of articles in journals where more than 70% of the articles have been categorized as Hybrid OA (those that could be affected by this problem) is only 9,211 (0.4% of the sample).
2. The license information provided by CrossRef is incomplete. We found that for approximately 85,000 out of 163,000 articles classified as Gold OA (because the journal where they are

published is listed in DOAJ), no open license was reported via CrossRef, suggesting that a large number of journals still do not deposit license information in CrossRef. If the proportion of Hybrid or Delayed OA journals that do not deposit license information in CrossRef is any similar, our results would be affected in that some Gold, Hybrid, and/or Delayed OA articles would have been erroneously classified as Bronze OA. Therefore, further analyses are needed to ascertain the specific composition of the Bronze OA category. It may turn out that Bronze OA is only a mix of Gratis Access provided by the publishers, and Gold, Hybrid or Delayed OA in journals that do not declare licenses in a easily identifiable way. In that case, the term "Bronze OA" will stop being necessary once these practical limitations are overcome.

3. Regarding Green OA, in this article we make the assumption that documents available from repositories are sustainable and legal. This might not be true in some cases, and therefore a more in-depth study of the sustainability and legality of subject and institutional repositories all over the world would be helpful to advance our knowledge of OA.
4. This study does not differentiate between the various versions of the articles that may have been made available on the Web: preprints that still have not gone through peer-review, authors' accepted manuscripts, and the publisher's version of record. Further studies are needed to detect the extent to which preprints are prevalent in specific subject areas, and whether this could affect the quality and validity of research that cites preprints, rather than accepted manuscripts or the publisher's version of record, which have been vetted by peer-review panels.

Lastly, perhaps one of the most important limitations of this study is that it is not easily replicable because of the limitations on data extraction imposed by Google Scholar. Extracting a large amount of data from this source is still only possible if one is willing to commit an inordinate amount of time to the task (three months, in our case). However, the goal of this study was not to describe a replicable method to analyze OA levels using Google Scholar, but to find out whether the data available in Google Scholar could in fact be useful for this purpose. If it turns out that the data *is* useful, a request could be made to Google Scholar to reconsider making their data (at least to the parts related to the free availability of documents) more open for reuse. Repositories have traditionally been in favor of interoperability (as proven by the OAI-PMH initiative), and publishers are slowly but steadily making article metadata more open through platforms like CrossRef and also thanks to initiatives like I4OC[105] (Initiative for Open Citations), so it is not clear who, if anyone, would be against opening these data nowadays. Of course, this would implicate a change of direction for a platform that has traditionally been quite reluctant to provide its data in bulk. It is possible that the Google Scholar team prefers to spend its efforts in the same problem they have been trying to solve up to now: connecting users with the academic documents they need to help them solve important problems. Nevertheless, as worthy as that goal is, it is also beyond doubt that these data would be of great interest to all actors in the scientific community, and might also be able to save duplicated efforts to other OA-related initiatives.

Despite the limitations described above, this study analyses the largest sample of data extracted from Google Scholar to date, and by combining these data with the data available in other sources such as DOAJ, CrossRef, OpenDOAR, and ROAR, it offers insights into all the variants of OA (Gold, Hybrid, Delayed, Bronze, Green). It also provides information on the free availability of documents from other sources (FA), thus providing a holistic, large-scale, and detailed depiction of the status of OA of scientific publications across all scientific fields and countries.

4.2. Comparison of results with similar studies

The report recently published by Science-Metrix (2018), and the studies published by Piwowar et al. (2018), Bosman & Kramer (2018), and van Leeuwen et al. (2017) are perhaps the ones that offer more opportunities for comparison with this study. This is because they all extracted samples of documents from the Web of Science. Moreover, they all analysed documents from 2009 and 2014 (among other publication years). In the case of Science-Metrix's report, they declare to have carried out their data

[105] https://i4oc.org/

collection in the third quarter of 2016, roughly the same months in which we carried out our own data collection. The two studies based on Unpaywall as well as van Leeuwen's study used data extracted more recently (2017), and thus differences between our study and theirs may be attributable at least in part to the backfilling that has occurred between the time of our data collection and theirs.

Science-Metrix reports 55% overall free availability both in 2009 and 2014. These results are very similar to ours (53.1% and 55.8% in 2009 and 2014, respectively). Their percentages on OA provided by the publisher are also very similar to ours: 20.2% in 2009, and 23.3% in 2014 in the Science-Metrix report, while our study shows 19.9% in 2009 and 25.3% in 2014 (Figure 3). The figures on Green OA differ in the two studies. The Science-Metrix report finds 33.3% and 31.5% of Green OA in 2009 and 2014, whereas our study only finds 15.7% and 18.9% in these years (Figure 4). The reason of this difference is probably that the Science-Metrix report considered documents available from ResearchGate as Green OA, and our study does not. However, our study shows that 34.5% and 31.2% of the documents in 2009 and 2014, respectively, are available from ResearchGate (which we label as FA only), which matches the results found by Science-Metrix.

As regards the results at the country level, the country tables available in the Science-Metrix report offer strikingly similar results to the ones displayed in Table 4, although the percentages in our study are roughly 3 points higher for each country than in the Science-Metrix report (except in the case of Brazil, which has a higher percentage in the Science-Metrix report). The case of Brazil reveals other possible differences between Science-Metrix's approach, and ours, because they declare that SciELO is almost tied to ResearchGate in the number of freely accessible documents they offer, whereas our data shows that ResearchGate offers over 24,000 documents published by Brazilian researchers, and SciELO only 6,000.

As for the results at the level of subject areas, their results also agree with our study in that the areas with a higher percentage of free availability are the Health and Natural Sciences (over 50%), followed by Applied and the Social Sciences (between 40% and 50%), and lastly, the Arts & Humanities, with lower percentages (less than 40%).

Lastly, it is worth noting that the Science-Metrix study applies a calibration factor of 1.2 to the counts of freely available documents found by the 1science database, because the recall of this source is considered to be low. Therefore, although the results in this study closely match the results in the Science-Metrix study (at least at the levels of countries and broad subject categories), Google Scholar seems to have a better recall than the 1science database, because no calibration factor was applied in this case.

As stated in the literature review, the study by Piwowar et al. (2018) used three different samples, and one of them was a sample of documents covered by the Web of Science. Although their paper only reports the percentage of overall availability for documents in this sample (36.1%), the supplemental data they released alongside the paper (Piwowar et al., 2017) provides the necessary data to calculate OA percentages by year and type of OA in their WoS sample. Their data shows that 33.1% of the documents published in 2009, and 37.4% of the documents published in 2014 in their sample of WoS documents were freely accessible in some way according to Unpaywall. These results are very similar to ours (31,2% in 2009 and 35.8% in 2014), if we disregard the percentage of documents that we considered FA only (available only from sources other than publishers and repositories), which their study does not analyze. The slightly higher percentage in their study might be caused by slightly better coverage of OA sources in the Unpaywall system than in Google Scholar, but might also be explained by sampling issues (they use a sample, rather than the entire collection of documents), by small methodological differences regarding OA labelling (our study does not consider as Green OA documents hosted in personal or department websites inside academic domains, while theirs does), or by the fact that their study analyses data extracted in 2017, and therefore OA levels might have increased because of backfilling since the data in our study was collected (summer of 2016). In any case, the specific percentages of the different types of OA in the two studies are remarkably similar, as can be observed in Table 5.

Table 5. Comparison of OA levels found by Google Scholar in this study, and by Piwowar et al. (2018) using Unpaywall data

	2009		2014	
	Google Scholar	**Unpaywall**	**Google Scholar**	**Unpaywall**
% Gold	3.3	3.1	10.1	9.4
% Hybrid	0.5	3.4	1.5	5.2
% Delayed	2	-	1.1	-
% Bronze	14.1	14.7	12.6	11.6
% Green (only)	11.3	11.9	10.5	11.2
% Total OA	31.2	33.1	35.8	37.4

Bosman & Kramer (2018) analysed the data from Unpaywall that the Web of Science has integrated into its system. They found an overall 28% of OA for documents published in 2014 (Kramer & Bosman, 2018). However, the Web of Science only provides OA information when the version that is FA is the author accepted manuscript (AAM), or the publisher's version of record (VOR). Therefore, this suggests that almost 10% of the documents covered by Unpaywall (and probably also Google Scholar, given the similarities found above) are preprints, that is, manuscripts that still have not gone through peer-review.

The results of this study show significantly higher percentages of OA (up to 15 points higher) than those found by van Leeuwen et al. (2017). That study reports overall OA levels of roughly 21% in 2009 and 27% in 2014. These differences may be explained by the more restricted approach of van Leeuwen's method, focused on OA sources related with the idea of legality and sustainability such as OpenAIRE, DOAJ, PubMed Central, etc., and with a strong focus on Gold and Green OA; while Google Scholar, Unpaywall and Science Metrix identify also Hybrid, Delayed and particularly Bronze OA. Considering together the Gold and Green OA shares in 2014 in this study (10.1%+10.5%) we come up with a closer value to the 27% observed in van Leeuwen's study, thus suggesting the relative consistency among methods, but also highlighting the role that Hybrid, Delayed and Bronze OA (together with FA only) play in the overall consideration of what is OA.

Lastly, the results from this study somewhat differ from those found by Jamali & Navabi (2015), who carried out a series of subject queries in Google Scholar to analyze OA levels in 277 minor subject categories extracted from Scopus. They found approximately 60% of free availability for documents published between 2004 and 2014 in all areas of research (Life, Physical, Social, and Health Sciences). This differs from our study, where we found significantly less free availability in the Social Sciences and Applied Sciences, than in the Natural and Health Sciences. The difference might be explained at least in part by the fact they only analyzed the first ten hits of each query, and Google Scholar is known to rank documents in a search based primarily on the number of citations that the documents have received (Martin-Martin et al., 2017). Highly cited documents might have different patterns of behavior regarding OA availability than a randomly selected sample of documents. Moreover, their study was not limited to documents covered by the Web of Science, which might also have influenced the results.

5. Conclusions

5.1. Answers to research questions

RQ1. How much of the recently published scientific literature is freely available according to the data available in Google Scholar, by year of publication, subject categories, and country of affiliation of the authors?

Google Scholar provided links to freely available versions of documents indexed in the Web of Science and published in 2009 or 2014 in approximately 54.6% of the cases. The percentage is slightly lower for documents published in 2009 (53%) than for documents published in 2014 (55.8%). However, there are important differences at the subject level and at the country level.

Categories related to the Natural and Health sciences achieve the highest percentages of free availability (Basic Life Sciences: 67.5%; Biomedical Sciences: 62.5%). Categories related to the Social Sciences, excepting Psychology (57.8%) and Economics & Business (55.2%) reach lower percentages (Sociology and Anthropology: 40.7%; Social and Behavioral Sciences, Interdisciplinary: 45.4%; Educational Sciences: 40%). Categories in the Arts and Humanities achieve the lowest percentages (Language and Linguistics: 39.4%; Creative Arts, Culture, and Music: 20.9%; Literature: 14.2%).

At the country level the percentages range from approximately 70% overall availability (Brazil, the Netherlands, Switzerland) to approximately 45% (China, Iran, and Russia), if we consider the top 25 countries with a higher output.

These results are remarkably similar to the ones found in other recent large-scale studies that analyse similar datasets but use different mechanisms to find evidence of OA (Piwowar et al., 2018; Science-Metrix Inc., 2018; van Leeuwen et al., 2017).

RQ2. How much is openly accessible in a sustainable and legal way, and what proportion is freely available but does not meet these criteria?

We consider that sustainability and legality in OA is important from a policy perspective. For this reason in this study we made a distinction between what we considered reasonably sustainable and legal sources (publishers and repositories), and sources that did not meet these criteria (academic social networks, personal websites, harvesters, and other websites).

Considering the two publication years under study (2009 and 2014), only 33.9% of the documents are openly accessible from sustainable and legal sources. This percentage is formed by the sum of all forms of OA provided by the publisher (Gold, Hybrid, Delayed, and Bronze: 23.1%), and OA provided by repositories that is not also available from the publisher (Green only: 10.8%). Bronze OA is the most common form of OA provided by the publishers. 13.2% of all documents in our sample were available as Bronze OA, while the combination of Gold, Hybrid, and Delayed only made up for 10.1% of the total number of documents. In the Bronze variety of OA, no Open License is available, and publishers usually extend very few rights to the user apart from free access. Therefore, Bronze OA articles cannot be redistributed or reused by anyone without explicit permission from the publisher, thus introducing a legal restriction in the OA consideration of Bronze OA publications.

As for Green OA, 17.6% of the documents in our sample were available from repositories according to Google Scholar.

Using Google Scholar as source of data made it possible to detect that 40.6% of the documents in our sample are freely available from sources that are not considered to meet the criteria of sustainability and legality. This means that more documents are freely available in unsustainable sources and/or in violation of their copyright, than through sustainable and legal ways. In addition to that, 20.7% (of all the documents in our sample) are only freely available from these other sources.

RQ3. What is the distribution of freely available documents by web domains?

As other studies had previously hinted (Jamali & Nabavi, 2015; Martín-Martín et al., 2014), the main source of freely available documents according to Google Scholar is, by far, the academic social network ResearchGate, which provided free access to 32.6% of all the documents in our sample (almost the same amount as all publishers and repositories put together). ResearchGate has a strong presence in Google Scholar, demonstrated by the fact that Google Scholar selects the ResearchGate version of an article as the primary version (see Figure 2) in 43.8% of the cases.

After ResearchGate, among the first places of the rank of websites that provided more freely available documents we can find the repositories PubMed Central and arXiv, the academic social network Academia.edu, harvesters like CiteSeerX and Semantic Scholar. After those, we find the largest commercial publishers (Wiley, Elsevier, PloS, Springer-Nature, BioMed Central, Hindawi, MDPI, and Taylor & Francis). In the majority of the cases when there is a freely accessible version of a document from the publisher, Google Scholar selects that version as the primary version.

5.2. Final remarks

From the answers to the research questions posed by this study, some general remarks can be drawn about the current status of OA to scientific publications:

The data available in Google Scholar, combined with the data available in other open resources such as CrossRef, DOAJ, OpenDOAR, and ROAR, can provide a faithful representation of OA levels of scientific publications. The results obtained with Google Scholar are similar to other existing approaches of OA identification (e.g. Unpaywall, Science Metrix or van Leeuwen's) thus suggesting some degree of agreement among the different approaches depending on how OA and FA are defined. However, as long as the data available in Google Scholar is not made available to the scientific community, Google Scholar cannot be considered a viable option to analyze OA levels on a regular basis. That said, the fact that Google Scholar, currently the most widely used academic search engine, is able to direct users to freely available versions of documents even when they are not freely accessible from the publisher or from repositories, is something that should not be ignored if one is to truly understand how scientific information is being accessed throughout the world nowadays. Unpaywall can be seen as a strong alternative to find only legal sources of OA, although future research should focus on how the concurrence of several methods could help to depict the most exhaustive landscape of multiple and diverse forms of OA (and FA).

Regarding the prevalence of the different variants of OA, this study confirms that most of the documents that publishers make freely accessible (e.g. Bronze OA) do not specify a clear OA-compatible license. Although this category might contain some masked Gold, Hybrid, or Delayed OA because of practical limitations (see section 4.1 above), it is likely that most of the documents categorised as Bronze OA were intended to be released by the publishers as Gratis Access. If this is the case, it would mean that continued free access over time to a large fraction of documents is entirely dependent on the publishers. This is a precarious situation, because even if publishers' original intention is to maintain Gratis Access status in perpetuity, as sole copyright holders nothing could stop them if they decided to revoke that status in the future. Moreover, in the best-case scenario (where Gratis Access status is maintained over time), the rights extended to users in these cases are very limited (for example, no redistribution and/or limited or no reuse rights), far from what the BOAI initially envisioned. This situation calls for a discussion among all stakeholders regarding the minimum requirements for OA status in scientific articles. But, even if an agreement is not reached, policy makers and funders should still strive to be clear in their OA mandates about the specific accessibility criteria that the outputs of research done with their funds should meet.

As for the OA levels at the country level, this study shows that even though many high-output European countries have OA levels that are above the world average, most of them are still far from complete OA adoption. This means that the goal of reaching 100% OA of scientific publications by 2020 proposed by the European Union in 2016 (Enserink, 2016) is probably unrealistic for most EU countries.

As other studies previously found (Borrego, 2016), the results of this study suggest that even with the current limitations that publishers impose on self-archiving (Gadd & Troll Covey, 2016; Tickell et al., 2017), there is much room for the growth of Green OA, because most publishers do not set limitations for archiving preprints, and some allow the archiving of author's accepted manuscripts at least in some types of websites with no embargo. However, the reality is that many authors still do not do this. What's more, this study confirms that when authors self-archive their documents, they vastly prefer ResearchGate over repositories. ResearchGate has succeeded in convincing researchers from all fields and all over the world to upload massive amounts of documents to its platform, something that institutional repositories have not managed to do. This matches the findings by Borrego (2017) for a sample of Spanish universities. There are several reasons that may have motivated this: the added-value services that ResearchGate provides (e.g. automatically updating the profiles of researchers, the easiness to upload publications, detailed impact and usage indicators that allow the 'quantification of the self' (Hammarfelt, de Rijcke, & Rushforth, 2016; Orduna-Malea, Martín-Martín, & Delgado López-Cózar, 2016), etc.), the prominence with which documents hosted in ResearchGate are displayed in Google Scholar (which might have served as a way to introduce users to the platform), the lack of awareness by researchers of the existence of repositories at their institutions, ignorance on how to use them, usability problems, the increasing barriers to self-archiving imposed by publishers (by which, unlike ResearchGate, repositories usually abide), as well as the lack of academic incentives for scholars to self-archive their work, in opposition to the "immediate feedback and gratification" provided by these academic networks (Hammarfelt et al., 2016). Whatever the reasons, this presents a problem for the advancement of a sustainable and legal system of OA and Open Science in general, because researchers are dedicating their efforts to feeding a proprietary platform that does not make its data available to the scientific community and which may disappear the moment it is not considered profitable.

Lastly, this study confirms that article metadata that contains license information is still not readily available for many articles, making it difficult to categorize the various variants of OA accurately. The appearance of the term "Bronze OA" (Piwowar et al., 2018), which is likely a mix of different variants of publisher-provided OA (Gratis, Gold, Hybrid, Delayed) that cannot be correctly identified because of the lack of license metadata, is a testament of this. CrossRef, currently the largest open source of license information at the article level, strongly recommends publishers to deposit license information, but they are not required to fill this field when they deposit metadata about an article. The system would benefit from the implementation of a standard metadata scheme that defines the specific rights that the license of an article extends to users. This should include the cases of Gratis Access provided by the publishers. This would be a way for publishers to declare their commitment to provide sustainable free access to these articles. In the cases of non-OA documents, self-archiving policies should also be recorded at the article level in machine-readable form, specifying how (under which license), when (specific date for the end of embargo period), where (in what kind of websites), and in what form (preprint, author's accepted manuscript, or version of record) an article can be self-archived. Among other things, this would allow funders and policy makers to check whether a published article meets the terms of a specific OA mandate, and it would allow institutional repositories to monitor the status of the documents published by its researchers more efficiently, and to automate the public release of these documents from the repository under the conditions specified by the license of each article.

In fact, the system suggested above would provide the same functionality as the automated system that the International Association of Scientific, Technical and Medical Publishers (STM) offered to implement in a letter they sent to ResearchGate (STM, 2017). This letter was an attempt to make the social network agree to check for copyright compliance when users upload documents to the platform. However, according to an undated STM announcement, ResearchGate rejected this offer (STM, n.d.), forcing publishers to continue issuing takedown notices when they detect that documents are made freely available in violation of their copyright (Coalition for Responsible Sharing, 2017b). As far as we know, the system has not been mentioned in public again after this exchange, despite its potential usefulness for repository managers all over the world, who, unlike ResearchGate, are usually willing to comply with copyright during the process of deposit.

References

Abad-García, M.-F., González-Teruel, A., & González-Llinares, J. (2018). Effectiveness of OpenAIRE, BASE, Recolecta, and Google Scholar at finding spanish articles in repositories. *Journal of the Association for Information Science and Technology*, *69*(4), 619–622. https://doi.org/10.1002/asi.23975

Archambault, É., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, L., & Roberge, G. (2014). *Proportion of Open Access papers published in peer-reviewed journals at the European and world levels: 1996-2013*. Retrieved from http://science-metrix.com/en/publications/reports/proportion-of-open-access-papers-published-in-peer-reviewed-journals-at-the

Björk, B.-C. (2016). The open access movement at a crossroad: Are the big publishers and academic social media taking over? *Learned Publishing*, *29*(2), 131–134. https://doi.org/10.1002/leap.1021

Björk, B.-C. (2017). Gold, green, and black open access. *Learned Publishing*, *30*(2), 173–175. https://doi.org/10.1002/leap.1096

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Guðnason, G. (2010). Open Access to the Scientific Journal Literature: Situation 2009. *PLoS ONE*, *5*(6), e11273. https://doi.org/10.1371/journal.pone.0011273

Bohannon, J. (2016, April 28). Who's downloading pirated papers? Everyone. *Science*. https://doi.org/10.1126/science.aaf5664

Bolick, J. (2017). Exploiting Elsevier's Creative Commons License Requirement to Subvert Embargo. *Poster Session Presented at the Kraemer Copyright Conference*. Retrieved from http://hdl.handle.net/1808/24107

Borrego, Á. (2016). Measuring compliance with a Spanish Government open access mandate. *Journal of the Association for Information Science and Technology*, *67*(4), 757–764. https://doi.org/10.1002/asi.23422

Borrego, Á. (2017). Institutional repositories versus ResearchGate: The depositing habits of Spanish researchers. *Learned Publishing*, *30*(3), 185–192. https://doi.org/10.1002/leap.1099

Bosman, J., & Kramer, B. (2016). Innovations in scholarly communication - data of the global 2015-2016 survey. https://doi.org/10.5281/ZENODO.49583

Bosman, J., & Kramer, B. (2018). *Open access levels: a quantitative exploration using Web of Science and oaDOI data*. https://doi.org/10.7287/peerj.preprints.3520v1

Chan, L., Cuplinskas, D., Eisen, M., Friend, F., Genova, Y., Guédon, J.-C., … Velterop, J. Budapest Open Access Initiative (2002). Retrieved from http://www.budapestopenaccessinitiative.org/read

Chen, X., & Olijhoek, T. (2016). Measuring the Degrees of Openness of Scholarly Journals with the Open Access Spectrum (OAS) Evaluation Tool. *Serials Review*, *42*(2), 108–115. https://doi.org/10.1080/00987913.2016.1182672

Christianson, M. (2007). Ecology Articles in Google Scholar: Levels of Access to Articles in Core Journals. *Issues in Science and Technology Librarianship*. https://doi.org/10.5062/F4MS3QPD

Coalition for Responsible Sharing. (2017a). Coalition for Responsible Sharing issues take down notices to ResearchGate to address remaining violations. Retrieved from http://www.responsiblesharing.org/2017-10-18-coalition-for-responsible-sharing-issues-take-down-notices-to-researchgate-to-address-remaining-violations/

Coalition for Responsible Sharing. (2017b). Publishers and societies take action against ResearchGate's copyright infringements. Retrieved from http://www.responsiblesharing.org/coalition-statement/

Coalition for Responsible Sharing. (2017c). ResearchGate Removed Significant Number of Copyrighted Articles. Retrieved from http://www.responsiblesharing.org/2017-10-10-ResearchGate-removed-articles/

Crotty, D. (2017). Study Suggests Publisher Public Access Outpacing Open Access; Gold OA Decreases Citation Performance [Blog Post]. Retrieved from https://scholarlykitchen.sspnet.org/2017/10/04/study-suggests-publisher-public-access-outpacing-open-access-gold-oa-decreases-citation-performance/

Else, H. (2018, May 17). Europe's open-access drive escalates as university stand-offs spread. *Nature*, pp. 479–480. https://doi.org/10.1038/d41586-018-05191-0

Elsevier. (2015). Dutch Universities and Elsevier Reach Agreement in Principle on Open Access and Subscription [Press release]. Retrieved from https://www.elsevier.com/about/press-releases/corporate/dutch-universities-and-elsevier-reach-agreement-in-principle-on-open-access-and-subscription

Enserink, M. (2016, May 27). In dramatic statement, European leaders call for 'immediate' open access to all scientific papers by 2020. *Science*. https://doi.org/10.1126/science.aag0577

Fuchs, C., & Sandoval, M. (2013). The Diamond Model of Open Access Publishing: Unions and the Publishing World Need to Take Non-Commercial, Non-Profit Open Access Serious. *TripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, *11*(2), 428–443. Retrieved from https://www.triple-c.at/index.php/tripleC/article/view/502

Gadd, E., & Troll Covey, D. (2016). What does 'green' open access mean? Tracking twelve years of changes to journal publisher self-archiving policies. *Journal of Librarianship and Information Science*. https://doi.org/10.1177/0961000616657406

Gargouri, Y., Larivière, V., Gingras, Y., Carr, L., & Harnad, S. (2012). *Green and Gold Open Access Percentages and Growth, by Discipline*. Retrieved from http://arxiv.org/abs/1206.3664

Green, T. (2017). We've failed: Pirate black open access is trumping green and gold and we must change our approach. *Learned Publishing*, *30*(4), 325–329. https://doi.org/10.1002/leap.1116

Hammarfelt, B., de Rijcke, S., & Rushforth, A. D. (2016). Quantified academic selves: the gamification of research through social networking services. *Information Research*, *21*(2). Retrieved from http://www.informationr.net/ir/21-2/SM1.html

Harnad, S. (2001). The self-archiving initiative. *Nature*, *410*(6832), 1024–1025. https://doi.org/10.1038/35074210

Haschak, P. G. (2007). The "platinum route" to open access: a case study of E-JASL: The Electronic Journal of Academic and Special Librarianship. *Information Research*, *12*(4). Retrieved from http://www.informationr.net/ir/12-4/paper321.html

Himmelstein, D. S., Rodriguez Romero, A., Levernier, J. G., Munro, T. A., McLaughlin, S. R., Greshake Tzovaras, B., & Greene, C. S. (2018). Sci-Hub provides access to nearly all scholarly literature. *ELife*, *7*. https://doi.org/10.7554/eLife.32822

Jamali, H. R. (2017). Copyright compliance and infringement in ResearchGate full-text journal articles. *Scientometrics*, *112*(1), 241–254. https://doi.org/10.1007/s11192-017-2291-4

Jamali, H. R., & Nabavi, M. (2015). Open access and sources of full-text articles in Google Scholar in different subject fields. *Scientometrics*, *105*(3), 1635–1651. https://doi.org/10.1007/s11192-015-1642-2

Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. *PloS One*, *9*(5), e93949. https://doi.org/10.1371/journal.pone.0093949

Kingsley, D. (2013). Walking in quicksand – keeping up with copyright agreements [Blog Post]. Retrieved from https://aoasg.org.au/2013/05/23/walking-in-quicksand-keeping-up-with-copyright-agreements/

Kingsley, D. (2017). Whose money is it anyway? Managing offset agreements [Blog Post]. Retrieved from https://unlockingresearch-blog.lib.cam.ac.uk/?p=1458

Kramer, B., & Bosman, J. (2018, January 9). Data from: Open access levels: a quantitative exploration using Web of Science and oaDOI data. https://doi.org/10.5281/ZENODO.1143707

Laakso, M., & Björk, B.-C. (2013). Delayed open access: An overlooked high-impact category of openly available scientific literature. *Journal of the American Society for Information Science and Technology*, *64*(7), 1323–1329. https://doi.org/10.1002/asi.22856

Laakso, M., & Lindman, J. (2016). Journal copyright restrictions and actual open access availability: a study of articles published in eight top information systems journals (2010–2014). *Scientometrics*, *109*(2), 1167–1189. https://doi.org/10.1007/s11192-016-2078-z

Laakso, M., & Polonioli, A. (2018). Open access in ethics research: an analysis of open access availability and author self-archiving behaviour in light of journal copyright restrictions. *Scientometrics*, 1–27. https://doi.org/10.1007/s11192-018-2751-5

Lawson, S. (2018). *Report on offset agreements: evaluating current Jisc Collections deals. Year 2 – evaluating 2016 deals*. Retrieved from http://hdl.handle.net/10760/31711

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Dataset: sources of free full text found by Google Scholar for documents in Web of Science published in 2009 and 2014 (raw and aggregated). https://doi.org/10.17605/OSF.IO/FSUJY

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2014). *Does Google Scholar contain all highly cited documents (1950-2013)?* (EC3 Working Papers No. 19). Retrieved from http://arxiv.org/abs/1410.8464

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentacion Cientifica*, *39*(4), e149. https://doi.org/10.3989/redc.2016.4.1405

Martin-Martin, A., Orduna-Malea, E., Harzing, A.-W., & Delgado López-Cózar, E. (2017). Can we use Google Scholar to identify highly-cited documents? *Journal of Informetrics*, *11*(1), 152–163. https://doi.org/10.1016/j.joi.2016.11.008

Mellor, D. (2016). Rewarding Transparent and Reproducible Scholarship. In *8th Conference on Open*

Access Scholarly Publishing. Retrieved from https://osf.io/6wsc2/wiki/home/

Mikki, S., Ruwehy, H. A. Al, Gjesdal, Ø. L., & Zygmuntowska, M. (2018). Filter bubbles in interdisciplinary research: a case study on climate and society. *Library Hi Tech*, LHT-03-2017-0052. https://doi.org/10.1108/LHT-03-2017-0052

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Mussell, J., & Croft, R. (2013). Discovery Layers and the Distance Student: Online Search Habits of Students. *Journal of Library & Information Services in Distance Learning*, *7*(1–2), 18–39. https://doi.org/10.1080/1533290X.2012.705561

Nicholas, D., Boukacem-Zeghmouri, C., Rodríguez-Bravo, B., Xu, J., Watkinson, A., Abrizah, A., … Świgoń, M. (2017). Where and how early career researchers find scholarly information. *Learned Publishing*, *30*(1), 19–29. https://doi.org/10.1002/leap.1087

Norris, M., Oppenheim, C., & Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the American Society for Information Science and Technology*, *59*(12), 1963–1972. https://doi.org/10.1002/asi.20898

Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, *104*(3), 931–949. https://doi.org/10.1007/s11192-015-1614-6

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2016). Metrics in academic profiles: a new addictive game for researchers? *Revista Espanola de Salud Publica*, *90*, e1–e5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27653216

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2017). Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors. *Revista Española de Documentación Científica*, *40*(4), e185. https://doi.org/10.3989/redc.2017.4.1500

Pitol, S. P., & De Groote, S. L. (2014). Google Scholar versions: do more versions of an article mean greater impact? *Library Hi Tech*, *32*(4), 594–611. https://doi.org/10.1108/LHT-05-2014-0039

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., … Haustein, S. (2017, August 1). Data from: The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles. https://doi.org/10.5281/ZENODO.837902

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., … Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375. https://doi.org/10.7717/peerj.4375

Poynder, R. (2018). Preface. In U. Herb & J. Schöpfel (Eds.), *Open Divide? Critical Studies on Open Access*. Sacramento, CA: Litwin Books, LLC. Retrieved from https://poynder.blogspot.com.es/2018/01/preface-open-divide.html

Prosser, D. C. (2003). From here to there: a proposed mechanism for transforming journals from closed to open access. *Learned Publishing*, *16*(3), 163–166. https://doi.org/10.1087/095315103322110923

Rogers, A. (2017, March 12). It's gonna get a lot easier to break science journal paywalls. *Wired*. Retrieved from https://www.wired.com/story/its-gonna-get-a-lot-easier-to-break-science-journal-paywalls/

Science-Metrix Inc. (2018). *Open access availability of scientific publications*. Retrieved from http://www.science-metrix.com/en/oa-report

Smith, E., Haustein, S., Mongeon, P., Shu, F., Ridde, V., & Larivière, V. (2017). Knowledge sharing in global health research – the impact, uptake and cost of open access to scholarly literature. *Health Research Policy and Systems*, *15*(1), 73. https://doi.org/10.1186/s12961-017-0235-3

STM. (n.d.). STM publishers and ResearchGate. Retrieved from http://www.stm-assoc.org/stm-publishers-researchgate/stm-and-researchgate/

STM. (2017). STM proposal – RG platform to become consistent with usage and access rights for article sharing. Retrieved from https://www.elsevier.com/__data/assets/pdf_file/0010/509068/STM_letter_ResearchGate.20170916.pdf

Suber, P. (2008a, April 29). Strong and weak OA. Retrieved from http://legacy.earlham.edu/~peters/fos/2008/04/strong-and-weak-oa.html

Suber, P. (2008b, August 2). Gratis and libre open access. *SPARC Open Access Newsletter*. Retrieved from https://sparcopen.org/our-work/gratis-and-libre-open-access/

Teplitzky, S. (2017). Open Data, [Open] Access: Linking Data Sharing and Article Sharing in the Earth Sciences. *Journal of Librarianship and Scholarly Communication*, *5*(General Issue), eP2150. https://doi.org/10.7710/2162-3309.2150

Tickell, A., Jubb, M., Plume, A., Oeben, S., Brammer, L., Johnson, R., … Pinfield, S. (2017). *Monitoring the Transition To Open*. Retrieved from http://www.universitiesuk.ac.uk/policy-and-analysis/reports/Pages/monitoring-transition-open-access-2017.aspx

Tijssen, R., Nederhof, A., van Leeuwen, T., Hollanders, H., Kanerva, M., & van den Berg, P. (2010). *Wetenschaps- en Technologie- Indicatoren 2010*. Retrieved from http://nowt.merit.unu.edu/docs/NOWT-WTI_2010.pdf

van Leeuwen, T., Meijer, I., Yegros-Yegros, A., & Costas, R. (2017). Developing indicators on Open Access by combining evidence from diverse data sources. In *STI 2017. Open indicators: innovation, participation and actor-based STI Indicators*. Retrieved from http://arxiv.org/abs/1802.02827

van Leeuwen, T. N., Moed, H. F., Tijssen, R. J. W., Visser, M. S., & Van Raan, A. F. J. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics*, *51*(1), 335–346. https://doi.org/10.1023/A:1010549719484

Van Noorden, R. (2014a). Online collaboration: Scientists and the social network. *Nature*, *512*(7513), 126–129. https://doi.org/10.1038/512126a

Van Noorden, R. (2014b, November 7). Google Scholar pioneer on search engine’s future. *Nature*. https://doi.org/10.1038/nature.2014.16269

Walker, T. J. (1998). Free Internet Access to Traditional Journals. *American Scientist*, *86*(5), 463–471. https://doi.org/10.2307/27857100

Final discussion and conclusions

The results of this thesis consistently find that GS data, and especially its citation data, can be useful for bibliometric analyses. Nevertheless, throughout all the analyses that have been performed, it has also become clear that there are important limitations that have to be considered when deciding whether to use data from GS for these purposes. Many of these limitations arise from the desire to use this tool for a purpose that falls outside the original scope intended by its creators.

Strengths of Google Scholar as a source of data for bibliometric analyses

The studies in this thesis show that GS has an extensive coverage of academic documents that includes, but is not limited to, most of the documents covered in the multidisciplinary citation databases WoS and Scopus. For example, GS has a considerably more extensive coverage of documents in the areas of Arts, Humanities, and Social Sciences than the other citation databases, where these areas have been traditionally neglected. GS covers types of documents that have been traditionally excluded from bibliometric analyses such as theses and dissertations, books, conference communications, and not-peer-reviewed materials such as reports, working papers, and preprints. It also has a more diverse distribution of languages among its sources. GS is also better positioned to work in the current scenario where documents are increasingly becoming living entities that change over time (for example, preprints that go through various modifications and end up being published in a journal, or not), rather than static objects that don't change once they are first published. This is evidenced in all the samples that have been analysed:

- Of the 64,000 highly-cited documents published between 1950-2013 which were extracted from GS, only 51% were covered by WoS. At least 18% of those highly-cited documents were books (Martín-Martín, Orduña-Malea, Ayllón, & Delgado-López-Cózar, 2014; Martín-Martín, Orduna-Malea, Ayllón, & Delgado López-Cózar, 2016).

- Out of the 9,188 journals found in GSM in the areas of Arts, Humanities, and Social Sciences (AHSS), almost four thousand were not covered in WoS or Scopus. The distribution of countries of publication and languages of the journals in GSM is also more diverse (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2017).

- WoS and Scopus have limited coverage of AHSS even when samples are circumscribed to very highly-cited documents (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018).

- Of the 2.45 million citations found by either GS, WoS, and/or Scopus, to highly-cited documents displayed in GSCP, 94% (2.3 million) were found by GS, while WoS found 52%, and Scopus 60%. Citations from non-journal sources and in languages other than English were much more common among citations only found by GS than among citations that were also found by WoS or Scopus (Martín-Martín, Orduna-Malea, Thelwall, & Delgado López-Cózar, 2018).

- Of the 2.32 million articles and reviews with a DOI published in 2009 or 2014 and covered by WoS, 2.27 million (97.6%) were successfully found in GS (Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018).

GS's citation graph is probably one of the most comprehensive in existence at the moment (if not the most comprehensive). This was observed in all the case studies included in this thesis (samples of documents from specific fields, samples of highly-cited documents) and was confirmed in the most recent one, where we carried out a systematic analysis of citations across all subject areas in GS, WoS, and Scopus using one of the (as far as we know) largest samples of citation data from GS used in a citation analysis. GS citation data was found to be a superset of WoS and Scopus citation data. Two conclusions can be extracted from this: first, that GS covers the vast majority of the source documents that WoS and Scopus cover, in addition to a substantial number of sources missing from the other two databases; and second, that the performance of GS's citation matching algorithm (its ability to detect citation relationships among documents) is at least as adequate as that of WoS or Scopus.

Even more interestingly, the results showed that despite the many unique citing sources covered by GS (document types other than articles and conference proceedings, documents in languages other than English), and more importantly, despite the various types of errors in citation data that can be observed in GS (missing, duplicate, or incorrectly assigned citations), Spearman correlations between citation counts in GS and WoS or Scopus were high across all samples analysed in this thesis. Depending on the characteristics of the samples, correlations could be extremely high (up to 0.99), or somewhat lower (down to 0.63). The highest correlations were found in the subject areas where the databases had a higher coverage overlap (STEM fields). The lowest correlations were found in samples that had one or more of the following characteristics: samples of documents in the areas of the Humanities of Social Sciences, samples where the majority of documents were published in languages other than English, and samples of documents that only contained highly-cited documents.

This seems to indicate that, at least at a macro level, the errors in GS citation data do not seem to have a large influence in the results of bibliometric analyses. Of course, at a micro level even a single important mistake in the citation data of a document could generate unfair comparisons. Therefore, because there is no single infallible database, data from various sources should always be considered when carrying out micro-level analyses. This ultimately led us to the idea of "Scholar Mirrors", and to the development of web applications that combine bibliometric indicators from various sources.

The correlations also suggest that GS citation data seems to be as useful as WoS and Scopus citation data for bibliometric analyses in the STEM fields, and significantly more useful than the other two in the areas of Humanities and Social Sciences, where WoS and Scopus have a lower coverage (they are missing a part of the picture), and studies where there is an interest in analysing academic publications other than journal articles.

Lastly, while in this thesis GS has been benchmarked against the most widely used databases that contain citation data (WoS and Scopus), several new important players in this arena have emerged recently: Microsoft Academic (in February 2016), Dimensions (in January 2018), and COCI (the OpenCitations Index of Crossref open DOI-to-DOI citations, first released on June 2018). Although we were very interested in analysing how GS compared to these new sources, such a study could not be fit within our schedule, since they were released once the thesis was well under way. Nevertheless, some studies have started to suggest that in some disciplines, Microsoft Academic has a similar publication and citation coverage to GS (Harzing & Alakangas, 2017), and citation data in Dimensions has been found to be very similar to Scopus (Thelwall, 2018).

Limitations of Google Scholar as a source of data for bibliometric analyses

In this thesis we define limitations as the characteristics of GS that do not align well with the purpose of carrying out bibliometric analyses. The most important limitations identified during this thesis are listed below:

- Lack of transparency regarding its size and coverage: GS does not declare which sources it covers (publishers, journals, repositories, aggregators, websites of academic institutions…). All the information we know on its coverage comes from empirical studies carried out by researchers not affiliated with GS. This limitation also means that carrying out a true random selection of documents (or journals, or authors) is not possible in GS.
- Lack of support for advanced searches and filtering options: this makes it difficult to control exactly which documents are displayed in the results, and therefore, it introduces limitations in how researcher can select documents for a bibliometric analysis. In GS, users have two general choices:
    - they can extract the results that GS displays for a given query: this method usually provides a good number of results relative to the effort invested, but the downside is that users never fully know how the black-box relevance algorithm used by GS might affect results. Therefore, random selections are not possible with this method. The use of keywords in combination with the advanced search operators (OR, -, “”, intitle, source, site), and filtering by year of publication can help to specify which documents the user wants to retrieve, but does not fully solve the issue.
    - they can carry out the document selection beforehand in a source other than GS, and then search the selected documents in GS: with this method, a truly random selection can be accomplished, provided that the full list of documents in the other source is available. On the other hand, searching small groups of documents, or even searching one document per query, drastically reduces the speed of data extraction. Moreover, selecting documents in other sources, especially if their coverage is lower than that of GS, can introduce coverage limitations in the study that would not truly reflect GS's true coverage.
- Dynamic coverage: GS's document base is always in flux. New documents are added several times a week, but unlike WoS or Scopus, documents and whole sources can also be dropped from GS. This occurs when a website that hosts documents is taken down, or when GS detects that there is something wrong with its metadata. This happened in 2017: GS detected that the large aggregator of bibliographic metadata Dialnet, specialized in Spanish and Iberoamerican academic publications, had some faulty metadata in a batch of old records. Because of this, the whole source was silently dropped from GS. This had the consequence that a large number of AHSS journals that had no other web exposure became effectively invisible in GS. This became apparent when a large number of Spanish journals were found to be missing from the 2012-2016 edition of GSM (from 1,101 journals in the previous edition, to 599) (Delgado López-Cózar & Martín-Martín, 2018). This kind of issue obviously has an effect in citation data, because citation counts can decrease when citing sources are dropped from GS. This has become a source of frustration to authors who pay attention to their bibliometric indicators in GSC. Combined with the lack of transparency, this means that it is not possible to know at the beginning of an analysis whether important and relevant sources have been dropped for GS for technical reasons, or whether this could affect the results of the analysis. This problem is difficult to detect, because in order to know that a source has been dropped from GS, it is necessary to keep track of the sources it covered before.

- Very limited document metadata: unlike WoS or Scopus, the bibliographic information displayed in GS for any given document is very limited: title, (some) authors, (part of the) name of the journal, and (not always) the publication year. Necessary information for advanced bibliometric analyses, such as the institutional affiliation of the authors, the document type, the language of the document, funding institutions, type of OA available… is not offered by GS.
- Lack of options to export data in bulk (public API): GS has never provided, nor, apparently, intends to provide in the future, a system that allows user to export data in bulk. Therefore, the only way to export data from GS is to scrape it from search results pages (SERP). In GS, a SERP can display a maximum of 20 results per page, and for any query, only a maximum of 1,000 results can be displayed in total. What's more, GS, as a part of Google, has strict security measures in place to detect bots that try to extract content. Specifically, users are asked to solve CAPTCHAs when the system detects a higher than normal volume of queries in a short amount of time. The threshold for triggering the CAPTCHA is at this point so low that it does not only affect automated bots, it also affects regular users.
- More open to manipulation: unlike selective citation indexes like WoS or Scopus, users can easily fabricate fake documents that, once indexed by GS (which is easily achievable, just by uploading those fake documents to an academic site or repository), will boost the author's own bibliometric indicators, or those of any other researcher (Delgado López-Cózar, Robinson-García, & Torres-Salinas, 2014). While this ease of manipulation might be considered a downside to GS, it is worth noting that the citation data in GS can be easily audited (anyone can check the origin of the citing documents). GS's chief engineer considers that the potential career-ending consequences for people who engage in these practices (which he considers SPAM) are the best deterrent against them (Van Noorden, 2014).

In this thesis, some of the limitations described above have been overcome, at least in part. For example, the lack of rich metadata in GS can be overcome by combining data from GS with data from other freely available sources, such as CrossRef, the metadata available as HTML meta tags in the webpages where GS finds the documents (publisher websites, repositories), and public APIs. However, other limitations cannot be easily solved, like the lack of bulk export capabilities. Up to now, it has been possible to extract small quantities of data (data about an author, or a journal, or a keyword query) from GS in a short time, but extracting large quantities of data in a centralized manner is a very time-intensive task because it requires making a large number of queries, which trigger the CAPTCHA system. Moreover, GS could decide to strengthen its security measures at any time, and make it even more difficult that it is now to extract data. Therefore, at this point GS is not a source from which large quantities of data can be extracted sustainably over time in order to power large bibliometric applications.

The limitations listed above, together with the errors that GS makes in some cases (Martín-Martín, Ayllón, Delgado López-Cózar, & Orduna-Malea, 2015; Martín-Martín et al., 2014; Martín-Martín, Orduna-Malea, Ayllón, & Delgado-López-Cózar, 2016; Orduna-Malea, Martín-Martín, & Delgado López-Cózar, 2017), make it difficult for a small team of people with no funding (which has been our case) to carry out medium- and large-scale analyses with GS data. Moreover, in the case of web applications, these analyses would need to be repeated on a regular basis in order to provide reasonably updated data.

In our case, even though we have managed to work with several of the (as far as we know) largest samples of GS data that have been used for bibliometric analyses so far, the effort and time spent to extract them and process them has been considerable. During the extraction process, it was necessary to switch extraction methods in several occasions (distributing queries through a pool of IPs, automated and manual CAPTCHA-solving systems), because the throughput of data rapidly decreased over time as a result of Google strengthening its security measures (many people try to extract data from the general Google search engine, and these security measures also cover GS). This means that there is no reliable method to extract data from GS in bulk, because even if a method works for a period of time, it might stop working at any time. In the end, the best we could manage was a throughput of approximately 3,000 queries per hour. Depending on the type of query, this could yield between 3,000 bibliographic records (if each query only returns one record) and 60,000 records (if each query returns the maximum 20 results per page). During the process of data extraction, most of the time is spent solving CAPTCHAs manually, which sometimes have to be solved four of five times before queries can resume. Once the data was extracted, it was often necessary to clean and enrich it in order to make it ready for analysis. The cleaning process was sometimes necessary to remove important errors (for example, in the author profiles included in our prototype web applications), and the enriching process often involved carrying out a second round of queries to other services, such as CrossRef, or publisher APIs, further increasing the time and effort required to do the analyses. Ultimately, this means that regularly updating even the small applications developed for this thesis would require a considerable amount of time and effort for a small team such as ours.

An alternative approach to data extraction in GS that still has not been tested is to replace our current model of centralized extraction (we extract all the data) by a model where data extraction is highly distributed or crowdsourced (or a model where both methods are used complementarily). For example, in a platform that displays author profiles this could be implemented by making each author (or an authorized representative) responsible for collecting data from their GSC profile (a small quantity of information that would not require much time or effort to extract using specialized software) and importing it to another application where the data would be processed and added to a research information system that would offer functionalities that are not available in GSC profiles. Moving from an opt-out model (information is collected and displayed for all authors in a group, but can be removed at the request of the author) to an opt-in model (information about an author is only displayed if they decide to add their data to the platform) would greatly facilitate the task of updating the data in the platform (authors or their representatives could update the data themselves). On the other hand, the platform would probably face the “cold start” problem, that is, users would have to be convinced of the benefits of joining the new platform.

The road from proprietary to open research metadata

A relevant question with implications for research policy is whether there are cases in which investing in the extraction and processing of data available in freely accessible sources of metadata (including GS, but not limited to it) could be more cost-effective than investing in expensive licenses to be able to use the clean and rich (but in some areas biased) metadata sold by WoS, Scopus (or other commercial sources of citation data and research metadata in general).

This question, which could not have arisen fifteen years ago when all research metadata was proprietary, is beginning to cause arguments within the community. For example, in 2018 a formal complaint was sent to the European Ombudsman questioning the decision to select Elsevier as the sole subcontractor (in the capacity of data provider) for the European Open Science Monitor (Tennant, 2018).

Although this thesis cannot give a definitive answer to this increasingly important question, we note that the landscape of sources of research metadata is rapidly changing: there is a growing ecosystem of open sources of research metadata (open bibliographic data such that provided by CrossRef, open citation data brought by initiatives like OpenCitations and the Initiative for Open Citations (I4OC), or the open metadata on Open Access versions of articles provided by Unpaywall). More importantly, there is a growing agreement among the research community that research metadata must not be treated as a commodity that is only available from commercial companies. The potential of these data to provide insight into how scholarly communication works, how it develops over time, and how it might be improved is too great to be in the hands of just a few (International Society for Informetrics and Scientometrics, 2018). Rather, these data should be part of the commons (Shotton, 2013, 2018).

This message is finally getting across, and we might soon observe important changes in this regard: the recently released Guidance on the Implementation of Plan S (cOAlition S, 2018), a Plan to achieve full and immediate Open Access of publicly-funded research currently signed by 18 national and private funders (most of them from Europe), has already established that the release of cited references "in standard interoperable format, under CC0 public domain dedication" is a mandatory quality criteria for Plan S compliant journals, platforms, and other venues. Unless these criteria change as a consequence of the recent call for feedback, this will mean that research that is carried out with grants from funders that have joined cOAlition S will have to be published in a platform that makes citation data (as well as the rest of the metadata) openly available.

As a freely accessible search engine that does not offer a public API for bulk access and reuse of metadata (free, but not open), GS can be said to be halfway between the "old world" of proprietary and expensive research metadata (WoS, Scopus), and the new wave of open research metadata platforms (CrossRef, OpenCitations, Unpaywall). GS is entirely subsidized by Google, and therefore it does not rely on a business model based on selling metadata to customers. However, GS necessarily relies on the data contained in publisher websites to provide its service, and some of these publishers do have economic interests in the market of research metadata. In order to understand GS's position, it is important to remember the negotiations that GS had go through to get publishers on board, and that GS's primary objective has always been to help people find and access the research they need. Given their undeniable success in this primary goal, it would therefore be understandable if the GS team did not want to deviate from their main goal (facilitating content discovery) by engaging in a new activity (releasing citation data and bibliographic metadata in general) that will probably require renegotiating their agreements with publishers. It is well-known that, for the time being, some of the largest ones (Elsevier, ACS) are firmly against participating in this initiative (International Society for Informetrics and Scientometrics, 2019; Singh Chawla, 2019). It is also not clear whether the GS team is at all interested in becoming a source of open metadata, although Google has been known to provide API access to many of its products.

At the moment, even though initiatives to open research metadata are rapidly gaining momentum, currently available open citation graphs are still not comprehensive enough for use in real-life scenarios (Di Iorio, Peroni, & Poggi, 2019). The gaps in coverage in these citation graphs have already led some authors to propose the creation of a crowdsourced open citation graph (Heibi, Peroni, & Shotton, 2019) with data provided, for example, by scholars and publishers themselves. GS is therefore in a unique position to make an even greater contribution to the research community and the world in general by liberating its citation data. Opening up GS citation data for reuse and integration with other citation datasets could greatly advance the goal of open citations by avoiding large duplication of efforts: it could be years before open citation graphs are able to reach the level of comprehensiveness that GS has today.

Will Google Scholar become an ally of the Initiative for Open Citations and thus help accelerate its vision, or will they turn a deaf ear to this issue? Will Google Scholar take this unique chance to further assist the scholarly community, or will they decide to remain a "walled garden" of metadata? As usual, they remain silent.

References

cOAlition S. (2018). Guidance on the Implementation of Plan S. Retrieved from https://www.coalition-s.org/wp-content/uploads/271118_cOAlitionS_Guidance.pdf

Delgado López-Cózar, E., & Martín-Martín, A. (2018). Apagón digital de la producción científica española en Google Scholar. *Anuario ThinkEPI*, *12*, 265–276. https://doi.org/10.3145/thinkepi.2018.40

Di Iorio, A., Peroni, S., & Poggi, F. (2019). Open data to evaluate academic researchers: an experiment with the Italian Scientific Habilitation. Retrieved from http://arxiv.org/abs/1902.03287

Harzing, A. W., & Alakangas, S. (2017). Microsoft Academic is one year old: the Phoenix is ready to leave the nest. *Scientometrics*, *112*(3), 1887–1894. https://doi.org/10.1007/s11192-017-2454-3

Heibi, I., Peroni, S., & Shotton, D. (2019). Crowdsourcing open citations with CROCI - An analysis of the current status of open citations, and a proposal. Retrieved from http://arxiv.org/abs/1902.02534

International Society for Informetrics and Scientometrics. (2018). Open citations: A letter from the scientometric community to scholarly publishers. Retrieved from http://www.issi-society.org/open-citations-letter/

International Society for Informetrics and Scientometrics. (2019). Resignation of the editorial board of the Journal of Informetrics. Retrieved from http://issi-society.org/blog/posts/2019/january/resignation-of-the-editorial-board-of-the-journal-of-informetrics/

Martín-Martín, A., Ayllón, J. M., Delgado López-Cózar, E., & Orduna-Malea, E. (2015). Nature 's top 100 Re-revisited. *Journal of the Association for Information Science and Technology*, *66*(12), 2714–2714. https://doi.org/10.1002/asi.23570

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, *12*(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2016). *The counting house, measuring those who count: Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in GSC (Google Scholar Citations), ResearcherID, ResearchGate, Mendeley, & Twitter* (EC3 Working Papers No. 21). Retrieved from https://arxiv.org/abs/1602.02412

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2014). *Does Google Scholar contain all highly cited documents (1950-2013)?* (EC3 Working Papers No. 19). Retrieved from http://arxiv.org/abs/1410.8464

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentacion Cientifica*, *39*(4), e149. https://doi.org/10.3989/redc.2016.4.1405

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017). Journal Scholar Metrics: building an Arts, Humanities, and Social Sciences journal ranking with Google Scholar data. In *22nd International Conference on Science, Technology & Innovation Indicators (STI)*. Paris. https://doi.org/10.17605/OSF.IO/VXNW6

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. *Scientometrics*, *116*(3), 2175–2188. https://doi.org/10.1007/s11192-018-2820-9

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2017). Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors. *Revista Española de Documentación Científica*, *40*(4), e185. https://doi.org/10.3989/redc.2017.4.1500

Shotton, D. (2013). Publishing: Open citations. *Nature*, *502*(7471), 295–297. https://doi.org/10.1038/502295a

Shotton, D. (2018). Funders should mandate open citations. *Nature*.

Singh Chawla, D. (2019). Open-access row prompts editorial board of Elsevier journal to resign. *Nature*. https://doi.org/10.1038/d41586-019-00135-8

Tennant, J. (2018, July 5). Complaint to the European Ombudsman about Elsevier and the Open Science

Monitor. https://doi.org/10.5281/ZENODO.1305847

Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? *Journal of Informetrics*, *12*(2), 430–435. https://doi.org/10.1016/J.JOI.2018.03.006

Discusión y conclusiones finales

Los resultados de esta tesis muestran de manera consistente que los datos de GS, y en especial sus datos de citas, pueden ser útiles para llevar a cabo análisis bibliométricos. Sin embargo, a lo largo de todos estos análisis también ha quedado patente que existen importantes limitaciones que tienen que ser tenidas en cuenta a la hora de decidir si usar datos de esta fuente para fines bibliométricos. Muchas de estas limitaciones surgen del deseo de utilizar esta herramienta para un propósito que está fuera del ámbito para el que sus creadores lo diseñaron en un principio.

Fortalezas de Google Scholar como fuente de datos para análisis bibliométricos

Los estudios incluidos en esta tesis muestran que GS tiene una cobertura muy extensa de documentos académicos que incluye, pero no está limitada a la mayoría de los documentos cubiertos por los índices de citas multidisciplinares WoS y Scopus. Por ejemplo, GS tiene una cobertura considerablemente mayor que las otras bases de datos en las áreas de Arte, Humanidades, y Ciencias Sociales. GS cubre tipos documentales que han sido tradicionalmente excluídos de los análisis bibliométricos como tesis y disertaciones, libros, comunicaciones a congresos, y materiales no revisados por pares como informes, working papers, y preprints. También tiene una distribución de idiomas más diversa en sus fuentes. GS está mejor posicionado para funcionar en el escenario actual en el que los documentos son cada vez más entidades vivas que cambian a lo largo del tiempo (por ejemplo, preprints que sufren varias modificaciones hasta que acaban siendo publicados en revistas, o no), en vez de objetos estáticos que no cambian una vez son publicados por primera vez. Todo esto se evidencia en todas las muestras analizadas:

- De los 64.000 documentos altamente citados publicados entre 1950-2013 que se extrajeron de GS, solo el 51% de ellos estaban cubiertos por WoS. Al menos el 18% de estos documentos eran libros (Martín-Martín, Orduña-Malea, Ayllón, & Delgado-López-Cózar, 2014; Martín-Martín, Orduna-Malea, Ayllón, & Delgado López-Cózar, 2016).

- De las 9.188 revistas encontradas en GSM en las áreas de Arte, Humanidades y Ciencias Sociales (AHSS), casi 4.000 no estaban cubiertas por WoS o Scopus. La distribución de países de publicación e idiomas en estas revistas era más diversa que las distribuciones encontradas en WoS o Scopus (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2017).

- WoS y Scopus tienen una cobertura limitada en AHSS incluso cuando las muestras se cirscuncriben a documentos muy altamente citados en sus respectivas categorías (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018).

- De las 2,45 millones de citas encontradas en GS, WoS y Scopus a los documentos altamente citados de GSCP, GS fue capaz de encontrar el 94% (2,3 millones), mientras que WoS encontraba el 52%, y Scopus el 60%. Las citas provenientes de fuentes diferentes a revistas eran mucho más comunes entre el grupo de citas que solo encontraba GS, que entre el grupo de citas que también eran encontradas por WoS o Scopus (Martín-Martín, Orduna-Malea, Thelwall, & Delgado López-Cózar, 2018).

- De los 2,32 millones de artículos y revisiones con un DOI publicadas en 2009 o 2014 y cubiertas por WoS, 2,27 millones (el 97,6%) fueron satisfactoriamente encontradas en GS (Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018).

El grafo de citas de GS es probablemente uno de los más exhaustivos actualmente existentes (si no el más exhaustivo). Esto se ha observado en todos los casos de estudio incluídos en esta tesis (muestras de documentos de campos temáticos específicos, muestras de documentos altamente citados) y ha sido confirmado en el estudio más reciente, donde se lleva a cabo un estudio sistemático de las citas recibidas según GS, WoS, y Scopus por documentos de todas las áreas temáticas, usando una de las mayores muestras de datos de citas de GS usadas hasta la fecha (hasta donde nuestro conocimiento alcanza). Los datos de citas en GS se pueden considerar un superconjunto de los datos de citas de WoS y Scopus. De esto se pueden sacar dos conclusiones princpales: primero, que GS cubre la gran mayoría de los documentos fuente que WoS y Scopus cubren, además de un número significativo de otros documentos; y segundo, que el algoritmo de matching de citas de GS (su habilidad de detectar relaciones de citación entre documentos) es al menos tan válido como el de WoS y Scopus.

Quizás más interesante, los resultados muestran que a pesar de la gran cantidad de citas encontradas solo por GS (de tipos documentales no cubiertos por otras bases de datos, y en idiomas diferentes al inglés), y a pesar de los diferentes tipos de errores en los datos de citación que se pueden encontrar en GS (citas perdidas, duplicadas, o incorrectamente asignadas), las correlaciones Spearman entre citas de GS y WoS, y entre GS y Scopus son altas o muy altas en todas las muestras analizadas en esta tesis. Dependiendo de las características de la muestra, las correlaciones pueden llegar a 0,99, o ser un poco más bajas (la más baja encontrada fue 0,63). Las correlaciones más altas se han encontrado en las categorías temáticas donde la cobertura de las bases de datos tiene un mayor solapamiento (categorías STEM). Las correlaciones más bajas se han encontrado en muestras que tenían una o más de las siguientes características: muestras de documentos en las áreas de Humanidades y Ciencias Sociales, muestras donde la mayoría de los documentos estaban publicados en idiomas diferentes al inglés, y muestras que solo contenían documentos altamente citados.

Esto parece indicar que, al menos a nivel macro, los errores en los datos de citas de GS no parecen tener una gran influencia en los resultados análisis bibliométricos. Por supuesto, a nivel micro incluso un error individual importante en los datos de citas puede desembocar en comparaciones injustas. Por tanto, como no hay ninguna base de datos infalible, datos de varias fuentes deberían considerarse siempre que se lleven a cabo análisis a nivel micro. Esto nos condujo a la idea de “Scholar Mirrors” y al desarrollo de aplicaciones web que combinan indicadores bibliométricos de varias fuentes.

Las correlaciones también sugieren que los datos de citas de GS parecen ser tan válidos como los de WoS y Scopus para realizar análisis bibliométricos en los campos STEM, y significativamente más útiles cuando se necesita analizar a las Humanidades y las Ciencias Sociales, y cuando hay interés en analizar documentos académicos que no sean artículos científicos. En estos casos la cobertura de WoS y Scopus es menor y por tanto la realidad no puede reflejarse adecuadamente.

Finalmente, aunque en esta tesis los datos de GS se han comparado con los índices de citas más usados para llevar a cabo estudios bibliométricos (WoS y Scopus), en los últimos años han aparecido nuevas fuentes en este campo: Microsoft Academic (en febrero de 2016), Dimensions (en enero de 2018) y COCI (el índice OpenCitations de citas DOI-a-DOI extraído de CrossRef, lanzado en junio de 2018). Aunque estamos muy interesados en comparar a GS con estas nuevas fuentes, estos estudios no se han podido encajar en el marco de esta tesis, por razón de prioridades y falta de recursos. Sin embargo, algunos estudios ya han empezado a sugerir que en algunas disciplinas, Microsoft Academic tiene una cobertura documental y de citas similar a GS (Harzing & Alakangas, 2017), y los datos de citas en Dimensions parecen ser similares a los que se pueden encontrar en Scopus (Thelwall, 2018).

Limitaciones de Google Scholar como fuente de datos para análisis bibliométricos

En esta tesis definimos el concepto de limitación como las características de GS que no encajan bien con el propósito de utilizar esta fuente para llevar a cabo análisis bibliométricos. Las limitaciones más importantes que se han identificado en esta tesis son las siguientes:

- Falta de transparencia en lo que respecta al tamaño y cobertura de la base documental: GS no declara qué fuentes indiza (editoriales, revistas, repositorios, agregadores, páginas de instituciones académicas…). Toda la información que se conoce sobre su cobertura proviene de estudios empíricos realizados por investigadores no afiliados con GS. Esta limitación también significa que no es posible realizar una selección verdaderamente aleatoria de documentos (o revistas, o autores) en GS.

- Falta de opciones para llevar a cabo búsquedas y filtros avanzados: esto hace que sea difícil controlar exactamente que documentos se muestran en los resultados, y por tanto, introduce limitaciones en cómo los investigadores pueden seleccionar documentos para un análisis bibliométrico. En GS, los investigadores que quieran hacer una selección de documentos para un estudio bibliométrico tienen dos opciones:

    - pueden extraer los resultados que GS muestra para la consulta realizada: este método normalmente proporciona un número considerable de resultados respecto al esfuerzo invertido, pero la parte negativa es que los usuarios nunca saben completamente cómo el algoritmo de relevancia empleado por GS podría afectar a los resultados. Por tanto, no es posible realizar selecciones aleatorias de documentos con este método. El uso de palabras clave en combinación con operadores de búsqueda avanzados (OR, -, “”, intitle, source, site) y las opciones de filtrado por año pueden ayudar al usuario a especificar qué documentos quiere recuperar, pero esto no resuelve por completo el problema.

    - pueden llevar a cabo la selección de documentos de antemano en otra fuente de datos que no sea GS, y una vez seleccionados, buscar dichos documentos en GS: con este método se pueden realizar selecciones de documentos aleatorias, siempre y cuando se conozca el listado completo de documentos de la fuente en la que se seleccionan los documentos. Por otra parte, hacer consultas que devuelvan pocos resultados, o incluso consultas en las que se busca un solo documento, reduce drásticamente la velocidad a la que se extraen los datos. Además, seleccionar documentos en otras fuentes, especialmente si su cobertura es más limitada que la de GS, puede introducir limitaciones de cobertura en el estudio que no reflejan la cobertura real de GS.

- Cobertura dinámica: la base de documental de GS está cambiando constantemente. Se añaden documentos varias veces a la semana, pero al contrario que en WoS y Scopus, algunos documentos individuales e incluso dominios web completos pueden desaparecer de GS sin previo aviso. Esto ocurre cuando una página web que aloja un documento desaparece de la Web, o cuando GS detecta que hay algún problema con los metadatos proporcionados por la fuente. Un ejemplo de este último caso ocurrió en 2017: GS detectó que el agregador de metadatos bibliográficos Dialnet, especializado en publicaciones académicas españoles e iberoamericanas, tenía metadatos incorrectos en un grupo de registros antiguos. Debido a esto, la fuente al completo fue eliminada silenciosamente de GS. Esto tuvo la consecuencia de que un gran número de revistas AHSS que no tenían ninguna otra forma de exposición en la web desaparecieron de GS, lo cual se hizo aparente en la siguiente edición de GSM (2012-2016), en la que el número de revistas españolas disminuyó de 1.101 en la edición anterior, a 599 (Delgado López-Cózar & Martín-Martín, 2018). Este tipo de problemas tienen efectos en los indicadores bibliométricos proporcionados por GS, lo que se ha convertido en una fuente de frustración para los autores que prestan atención a sus perfiles en GSC. Combinado con la falta de transparencia, esta limitación significa que no es posible saber, al principio de un análisis, si alguna fuente importante ha desaparecido de GS por alguna razón técnica, o si esto podría afectar a los resultados del análisis. Este problema es difícil de detectar, porque para saber que una fuente ha desaparecido de GS, es necesario saber que antes sí estaba cubierta, para lo cual es necesario hacer un seguimiento continuado de la cobertura.

- Metadatos muy limitados: al contrario que WoS y Scopus, la información bibliográfica mostrada en GS para cualquier documento es muy limitada: título, (algunos) autores, (parte de) el nombre de la revista, y (no siempre) el año de publicación. Otra información necesaria para llevar a cabo análisis bibliométricos avanzados, como la afiliación institucional de los autores, el tipo documental, el idioma del documento, las instituciones financiadoras, el tipo de Acceso Abierto disponible… no es ofrecida por GS.
- Falta de opciones para exportar datos de manera masiva (API pública): GS nunca ha proporcionado, y aparentemente, no tiene intención de proporcionar, ningún sistema que permita a los usuarios extraer información de manera masiva de su base de datos. Por tanto, la única manera de exportar datos de GS es extraerlos de las páginas de resultados del buscador. En GS, cada página de resultados puede mostrar hasta 20 resultados por página, y para cualquier consulta, solo se muestran un máximo de 1.000 resultados. Además, GS, como parte de Google, ha implementado estrictas medidas de seguridad para detectar programas que intenten extraer datos de manera automática. Específicamente, a los usuarios se les pide resolver CAPTCHAs cuando el sistema detecta un volumen de consultas superior a lo normal. La barrera para disparar el CAPTCHA es muy baja, de manera que se dispara muchas veces incluso cuando se está haciendo un uso normal de la plataforma por un operador humano.
- Más abierto a manipulación que otras fuentes: al contrario que los índices de citas selectivos como WoS y Scopus, cualquier persona puede fabricar documentos falsos y colgarlos en dominios que GS indice regularmente (repositorios, páginas personales en dominios institucionales). Una vez estos documentos están indizados en GS, aumentarán el número de citas de los documentos que aparezcan en su lista de referencias (Delgado López-Cózar, Robinson-García, & Torres-Salinas, 2014). Aunque la fácil manipulación de datos puede ser considerada una limitación de GS, es importante resaltar que igualmente, los datos de citas de GS son públicos (cualquier persona puede comprobar el origen de las citas de un documento). El ingeniero jefe de GS considera que las potenciales consecuencias negativas para la carrera de los investigadores que realicen este tipo de prácticas (que él considera SPAM) son el mejor freno a las mismas (Van Noorden, 2014).

En esta tesis, algunas de las limitaciones descritas arriba se han superado, al menos en parte. Por ejemplo, la falta de metadatos ricos en GS se puede paliar al combinar los datos de GS con datos de otras fuentes accesibles gratuitamente, como CrossRef, los metadatos disponibles en las etiquetas meta del HTML de las páginas de las que GS extrajo los datos (páginas de editoriales, repositorios), así como APIs públicas. Sin embargo, otras limitaciones no se pueden resolver fácilmente, como la falta de métodos para exportar datos de manera masiva. Hasta ahora ha sido posible extraer pequeñas cantidades de datos de GS (sobre un autor, una revista, o una consulta por palabras clave) en un corto periodo de tiempo, pero extraer grandes cantidades de datos de una manera centralizada es una tarea que requiere mucho tiempo, porque requiere realizar muchas consultas, que disparan el sistema de CAPTCHAs. Además, GS podría decidir reforzar sus medidas de seguridad en cualquier momento, y hacer la extracción de datos incluso más difícil de lo que es ahora. Por tanto, en este momento GS no es una fuente de la que se puedan extraer grandes cantidades de datos de manera sostenible en el tiempo para alimentar aplicaciones bibliométricas.

Las limitaciones listadas arriba, en conjunción con los errores que GS comete en algunos casos (Martín-Martín, Ayllón, Delgado López-Cózar, & Orduna-Malea, 2015; Martín-Martín et al., 2014; Martín-Martín, Orduna-Malea, Ayllón, & Delgado-López-Cózar, 2016; Orduna-Malea, Martín-Martín, & Delgado López-

Cózar, 2017), hacen que sea difícil para un pequeño equipo de personas sin financiación (nuestro caso) desarrollar análisis a mediana y a gran escala con datos de GS. Además, en el caso de las aplicaciones web, estos análisis deberían poder ser repetidos regularmente para poder proporcionar datos razonablemente actualizados.

En nuestro caso, aunque hemos sido capaces de trabajar con varias de las (hasta donde nosotros sabemos) muestras más grandes de datos de GS utilizadas en estudios bibliométricos hasta el momento, el esfuerzo y tiempo dedicados a extraer y procesar estos datos ha sido considerable. Durante el proceso de extracción era necesario cambiar las técnicas en varias ocasiones (distribuir consultas a través de un conjunto de IPs, sistemas de resolución de CAPTCHA automáticos y manuales…), porque el caudal de datos podía disminuir en cualquier momento como consecuencia de un fortalecimiento de las medidas de seguridad de Google (mucha gente intenta extraer datos del buscador general de Google, y estas medidas también afectan a GS). Esto significa que no hay un método fiable para extraer datos de GS de manera masiva, porque incluso si un método funciona durante un tiempo, puede parar de hacerlo en cualquier momento. Lo mejor que pudimos conseguir era un flujo de aproximadamente 3.000 consultas por hora. Dependiendo del tipo de consulta, esto podía proporcionar entre 3.000 registros bibliográficos (cuando cada consulta devuelve un registro) y 60.000 registros (cuando cada consulta devuelve el máximo de 20 resultados por página). Durante el proceso de extracción de datos, la mayoría del tiempo se gasta en resolver CAPTCHAs manualmente, pues a veces era necesario resolverlo cuatro o cinco veces antes de que las consultas pudieran continuar. Una vez los datos habían sido extraídos, siempre era necesario limpiar y enriquecer los datos para que fueran aptos para el análisis. El proceso de limpiado de datos era a veces necesario para eliminar errores importantes (esto fue muy común en nuestras aplicaciones que proporcionan datos sobre perfiles de autor), y el proceso de enriquecimiento de metadatos a menudo requería llevar a cabo una segunda ronda de consultas a otros sevicios, como CrossRef, o las APIs de las editoriales, aumentando así el tiempo y esfuerzo necesario para realizar el análisis. En definitiva, esto significa que actualizar incluso las pequeñas aplicaciones desarrolladas para esta tesis requeriría una cantidad de esfuerzo y tiempo considerables para un equipo de nuestras características.

Un enfoque alternativo a la extracción de datos de GS que no ha sido puesto a prueba en esta tesis es sustituir el modelo de extracción centralizado (nosotros extraemos todos los datos) por un modelo en el que la extracción de datos esté altamente distribuida o crowdsourced (o al menos un modelo donde ambos métodos se utilicen de manera complementaria). Por ejemplo, en una plataforma que presente perfiles de autores esto se podría implementar estableciendo un flujo de trabajo en el que cada autor (o un representante autorizado) sea responsable de extraer los datos de su perfil GSC (lo cual representaría una pequeña cantidad de información que no requeriría mucho tiempo para ser extraída con un software especializado) y de importarlos en otra aplicación donde los datos serían procesados y añadidos a un sistema de información científica que ofrecería funcionalidades no disponibles en los perfiles de GSC. Cambiar de un modelo opt-out (en el que se recoge información para todos los autores de un grupo, pero se puede ocultar información a petición de un autor) a un modelo opt-in (solo se muestra información de los autores que decidan añadir su información a la plataforma) facilitaría en gran medida la tarea de actualizar los datos de la plataforma (los usuarios o sus representantes podrían actualizar los datos ellos mismos). Por otro lado, la plataforma seguramente se enfrentaría con el problema del “arranque en frío” (cold start), es decir, habría que convencer a los usuarios de los beneficios de unirse a la nueva plataforma, hasta que se consiguiera una masa critica que pusiera en funcionamiento un efecto llamada.

De los metadatos de investigación propietarios a los metadatos abiertos

Una cuestión relevante con implicaciones en política científica es si existen casos en los que invertir en la extracción y procesamiento de metadatos de investigación disponibles en fuentes gratuitas (incluyendo a GS, pero no limitándonos a él) podría ser más rentable que invertir en caras licencias para poder usar los metadatos limpios y ricos (pero en algunos casos sesgados) vendidos por WoS y Scopus (u otras fuentes comerciales de datos de citas y metadatos de investigación en general).

Esta cuestión, que hace tan solo 15 años no se podría haber presentado ya que la gran mayoría de los metadatos de investigación estaban en manos de empresas comerciales, está empezando a causar discusiones entre la comunidad científica. Por ejemplo, en 2018 se envió una queja formal al defensor europeo cuestionando la decisión de subcontratar únicamente a Elsevier para proporcionar datos que se utilicen para generar el European Open Science Monitor (Tennant, 2018).

Aunque esta tesis no puede dar una respuesta definitiva a esta importante pregunta, somos conscientes de que el panorama de las fuentes de metadatos de investigación está cambiando rápidamente: actualmente ya hay un ecosistema creciente de fuentes abiertas de metadatos de investigación (datos bibliográficos abiertos como los proporcionados por CrossRef, datos de citas abiertos proporcionados por iniciativas como OpenCitations y liberados tras la presión ejercida por la Initiative for Open Citations (I4OC), o metadatos sobre Acceso Abierto como los proporcionados por Unpaywall). Quizás más importante todavía, hay un creciente consenso entre la comunidad científica sobre que los metadatos de investigación no deberían ser tratados como mercancías que solo están disponibles de empresas comerciales. El potencial de estos datos para proporcionar conocimiento sobre cómo funciona la comunicación científica, cómo se desarrolla a lo largo del tiempo, y cómo podría ser mejorada, es demasiado grande para dejar estos datos en manos de solo unos pocos (International Society for Informetrics and Scientometrics, 2018). Al contrario, estos datos deberían formar parte del dominio público (Shotton, 2013, 2018).

Este mensaje está calando finalmente, y pronto podríamos ser testigos de importantes cambios en este ámbito, a juzgar por la recientemente publicada Guía sobre la Implementación del Plan S (cOAlition S, 2018). El Plan S pretende conseguir Acceso Abierto inmediato y total para toda la investigación realizada con fondos públicos y que actualmente ha sido firmado por 18 instituciones financiadoras públicas y privadas (principalmente de Europa). En la guía de implementación se establece que la liberación de las referencias citadas “en formatos standard e interoperables, bajo una licencia de dominio público CC0” es uno de los criterios obligatorios que deben cumplir todas las revistas y plataformas de publicación que quieran seguir publicando trabajos de investigadores con financiación de las entidades que han firmado este plan.

GS, en cuanto a su situación de motor de búsqueda de acceso gratuito pero que no ofrece una API que permite acceder y reutilizar sus metadatos (gratis, pero no abierto), se puede considerar que está a medio camino entre el “viejo mundo” de los metadatos caros y propietarios (WoS, Scopus), y la nueva ola de plataformas de metadatos de investigación abiertos (CrossRef, OpenCitations, Unpaywall). GS está completamente subvencionado por Google, y por tanto no depende de un modelo de negocio basado en vender metadatos a sus clientes. Sin embargo, GS necesariamente depende para ofrecer su servicio de los datos disponibles en las páginas web de las editoriales, y algunas de estas editoriales tienen intereses económicos en el mercado de los metadatos de investigación. Para entender la posición en la que se encuentra GS, es importante recordar las negociaciones que tuvo que realizar para conseguir el apoyo de las editoriales, así como que el objetivo principal de GS siempre ha sido ayudar a la gente a encontrar la información que necesita. Dado su éxito en este objetivo, es comprensible que el equipo que trabaja en GS no quiera desviarse de su objetivo principal (facilitar la búsqueda de información) para iniciarse en una nueva actividad (proporcionar metadatos bibliográficos y de citas) que probablemente requeriría renegociar sus acuerdos con las editoriales. Es bien conocido, además, que por ahora algunas de las grandes editoriales (Elsevier, ACS) no están dispuestas a participar en esta iniciativa (International Society for Informetrics and Scientometrics, 2019; Singh Chawla, 2019). Tampoco está claro que el equipo que trabaja en GS esté interesado en convertirse en una fuente de metadatos abiertos, aunque Google es conocido por proporcionar acceso mediante API a muchos de sus productos.

Por el momento, incluso aunque las iniciativas para abrir los metadatos de investigación están ganando tracción rápidamente, los grafos de citas abiertos disponibles actualmente todavía no son lo suficientemente completos para ser usados en situaciones reales (Di Iorio, Peroni, & Poggi, 2019). Los agujeros en la cobertura de estos grafos de citas han conducido a algunos autores a proponer la creación de grafos de citas alimentados colaborativamente (crowdsourced) (Heibi, Peroni, & Shotton, 2019) con datos proporcionados, por ejemplo, por los propios investigadores. GS está por tanto en una posición única para hacer una contribución a la comunidad científica y al mundo mayor de la que ya está realizando, si decidiera liberar sus datos de citas. Abrir los datos de citas de GS para su reutilización e integración con otros datasets de citas podría avanzar en gran medida el objetivo de las citas abiertas al reducir enormemente la necesidad de duplicación de esfuerzos. Podrían pasar años antes de que los grafos de citas abiertos lleguen al nivel de cobertura que GS tiene hoy día.

¿Se convertirá GS en un aliado de la Initiative for Open Citations y ayudará a acelerar su visión, o harán oídos sordos a este tema? ¿Tomará Google Scholar esta oportunidad de ayudar a la comunidad científica, o decidirán seguir siendo un "jardín vallado"? Por el momento, como es habitual, no han hablado.

Referencias

cOAlition S. (2018). Guidance on the Implementation of Plan S. Retrieved from https://www.coalition-s.org/wp-content/uploads/271118_cOAlitionS_Guidance.pdf

Delgado López-Cózar, E., & Martín-Martín, A. (2018). Apagón digital de la producción científica española en Google Scholar. *Anuario ThinkEPI*, *12*, 265–276. https://doi.org/10.3145/thinkepi.2018.40

Di Iorio, A., Peroni, S., & Poggi, F. (2019). Open data to evaluate academic researchers: an experiment with the Italian Scientific Habilitation. Retrieved from http://arxiv.org/abs/1902.03287

Harzing, A. W., & Alakangas, S. (2017). Microsoft Academic is one year old: the Phoenix is ready to leave the nest. *Scientometrics*, *112*(3), 1887–1894. https://doi.org/10.1007/s11192-017-2454-3

Heibi, I., Peroni, S., & Shotton, D. (2019). Crowdsourcing open citations with CROCI - An analysis of the current status of open citations, and a proposal. Retrieved from http://arxiv.org/abs/1902.02534

International Society for Informetrics and Scientometrics. (2018). Open citations: A letter from the scientometric community to scholarly publishers. Retrieved from http://www.issi-society.org/open-citations-letter/

International Society for Informetrics and Scientometrics. (2019). Resignation of the editorial board of the Journal of Informetrics. Retrieved from http://issi-society.org/blog/posts/2019/january/resignation-of-the-editorial-board-of-the-journal-of-informetrics/

Martín-Martín, A., Ayllón, J. M., Delgado López-Cózar, E., & Orduna-Malea, E. (2015). Nature 's top 100 Re-revisited. *Journal of the Association for Information Science and Technology*, *66*(12), 2714–2714. https://doi.org/10.1002/asi.23570

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, *12*(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2016). *The counting house, measuring those who count: Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in GSC (Google Scholar Citations), ResearcherID, ResearchGate, Mendeley, & Twitter* (EC3 Working Papers No. 21). Retrieved from https://arxiv.org/abs/1602.02412

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & Delgado-López-Cózar, E. (2014). *Does Google Scholar contain all highly cited documents (1950-2013)?* (EC3 Working Papers No. 19). Retrieved from http://arxiv.org/abs/1410.8464

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista*

Española de Documentacion Cientifica, *39*(4), e149. https://doi.org/10.3989/redc.2016.4.1405

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2017). Journal Scholar Metrics: building an Arts, Humanities, and Social Sciences journal ranking with Google Scholar data. In *22nd International Conference on Science, Technology & Innovation Indicators (STI)*. Paris. https://doi.org/10.17605/OSF.IO/VXNW6

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. *Scientometrics*, *116*(3), 2175–2188. https://doi.org/10.1007/s11192-018-2820-9

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2017). Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors. *Revista Española de Documentación Científica*, *40*(4), e185. https://doi.org/10.3989/redc.2017.4.1500

Shotton, D. (2013). Publishing: Open citations. *Nature*, *502*(7471), 295–297. https://doi.org/10.1038/502295a

Shotton, D. (2018). Funders should mandate open citations. *Nature*.

Singh Chawla, D. (2019). Open-access row prompts editorial board of Elsevier journal to resign. *Nature*. https://doi.org/10.1038/d41586-019-00135-8

Tennant, J. (2018, July 5). Complaint to the European Ombudsman about Elsevier and the Open Science Monitor. https://doi.org/10.5281/ZENODO.1305847

Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? *Journal of Informetrics*, *12*(2), 430–435. https://doi.org/10.1016/J.JOI.2018.03.006