UNIVERSIDAD DE GRANADA

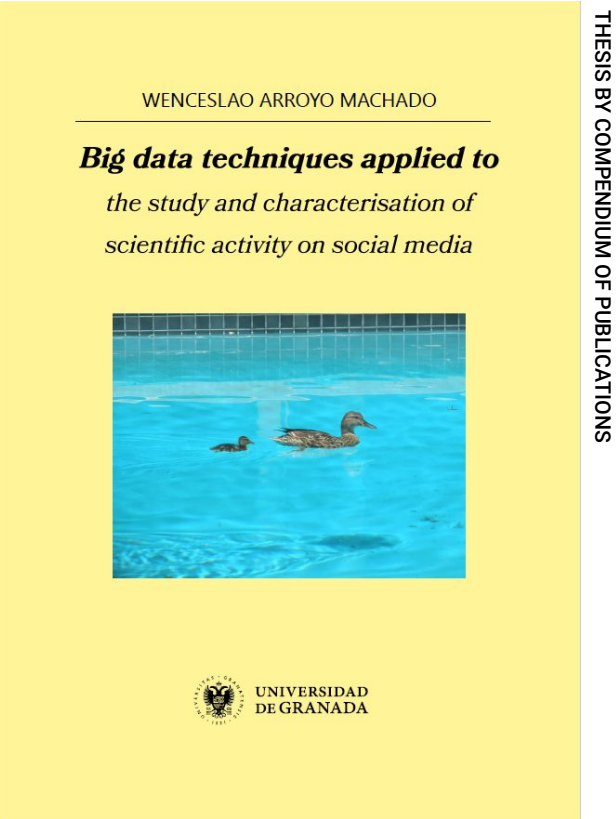
オルトメトリクス

BIG DATA TECHNIQUES APPLIED TO THE **STUDY** AND **CHARACTERISATION** OF **SCIENTIFIC ACTIVITY** ON **SOCIAL MEDIA**

by Wenceslao Arroyo Machado
PhD Advisors Enrique Herrera Viedma & Daniel Torres Salinas

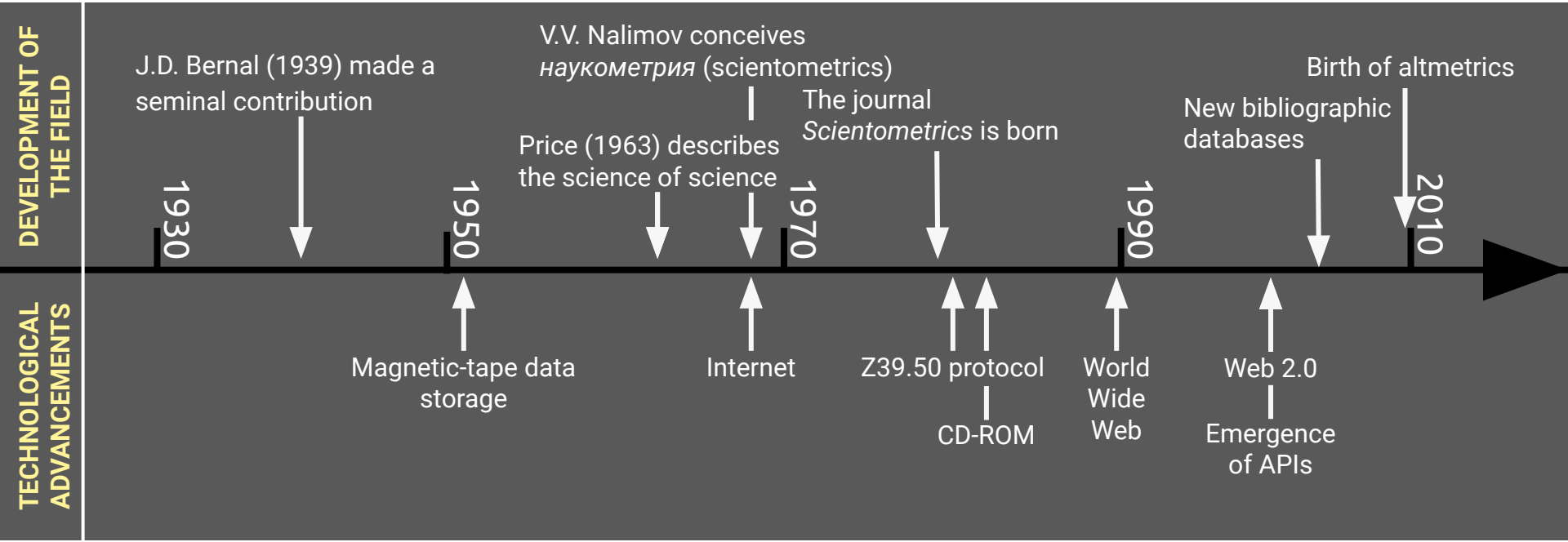
doi: 10.5281/zenodo.8403750

CONTENTS



THESIS BY COMPENDIUM OF PUBLICATIONS

The birth of scientometrics



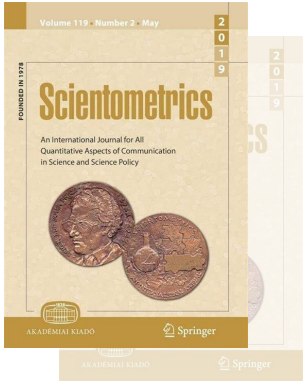
REFERENCES

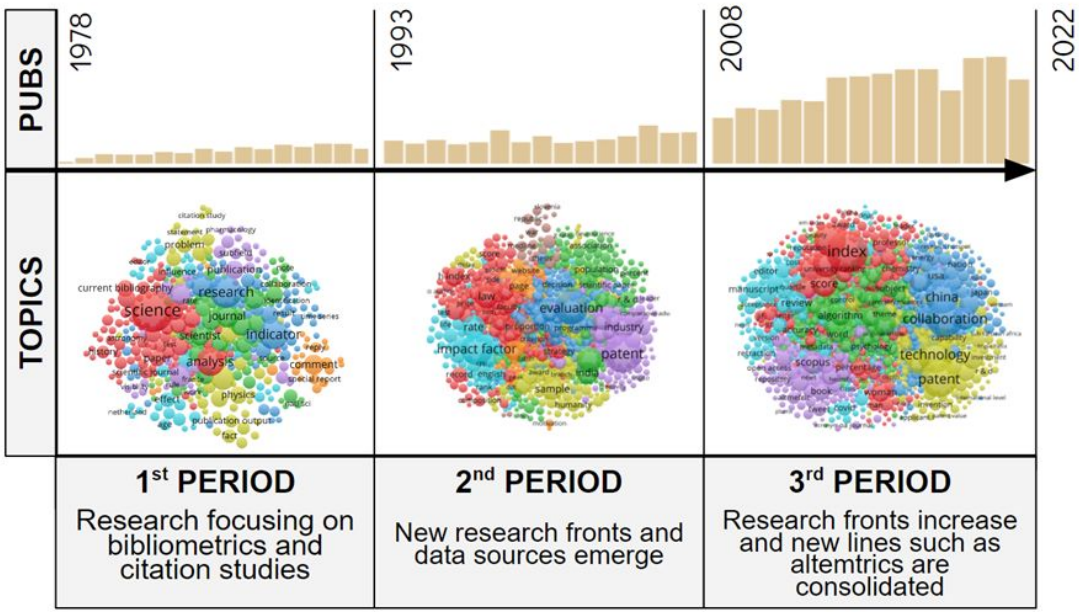
Bernal, J. D. (1939). The social function of science. *The Social Function of Science.*

Price, D. J. D. S. (1963). *Little Science, Big Science*. Columbia University Press. https://doi.org/10.7312/pric91844

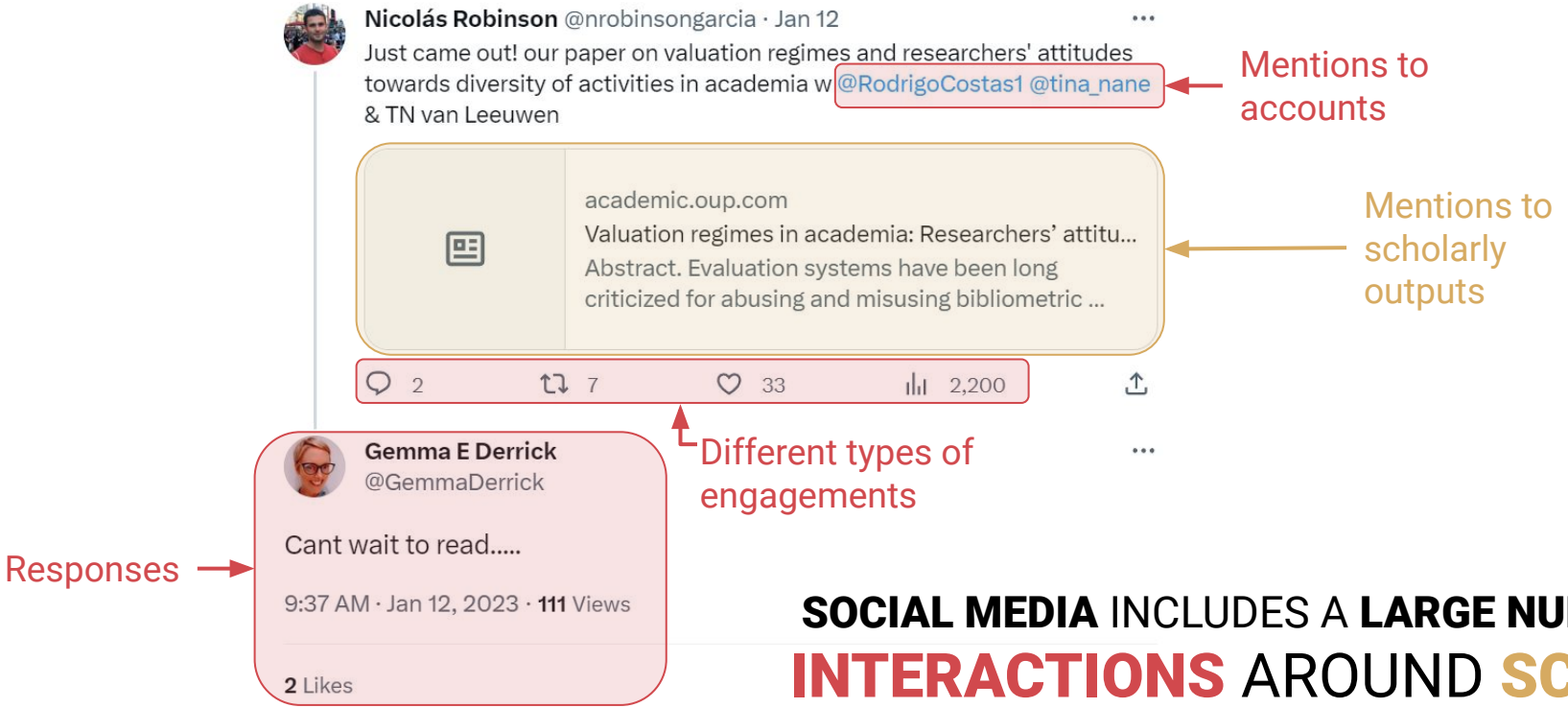
The development of scientometrics

The evolution of the journal ***Scientometrics*** reflects the evolution of the field



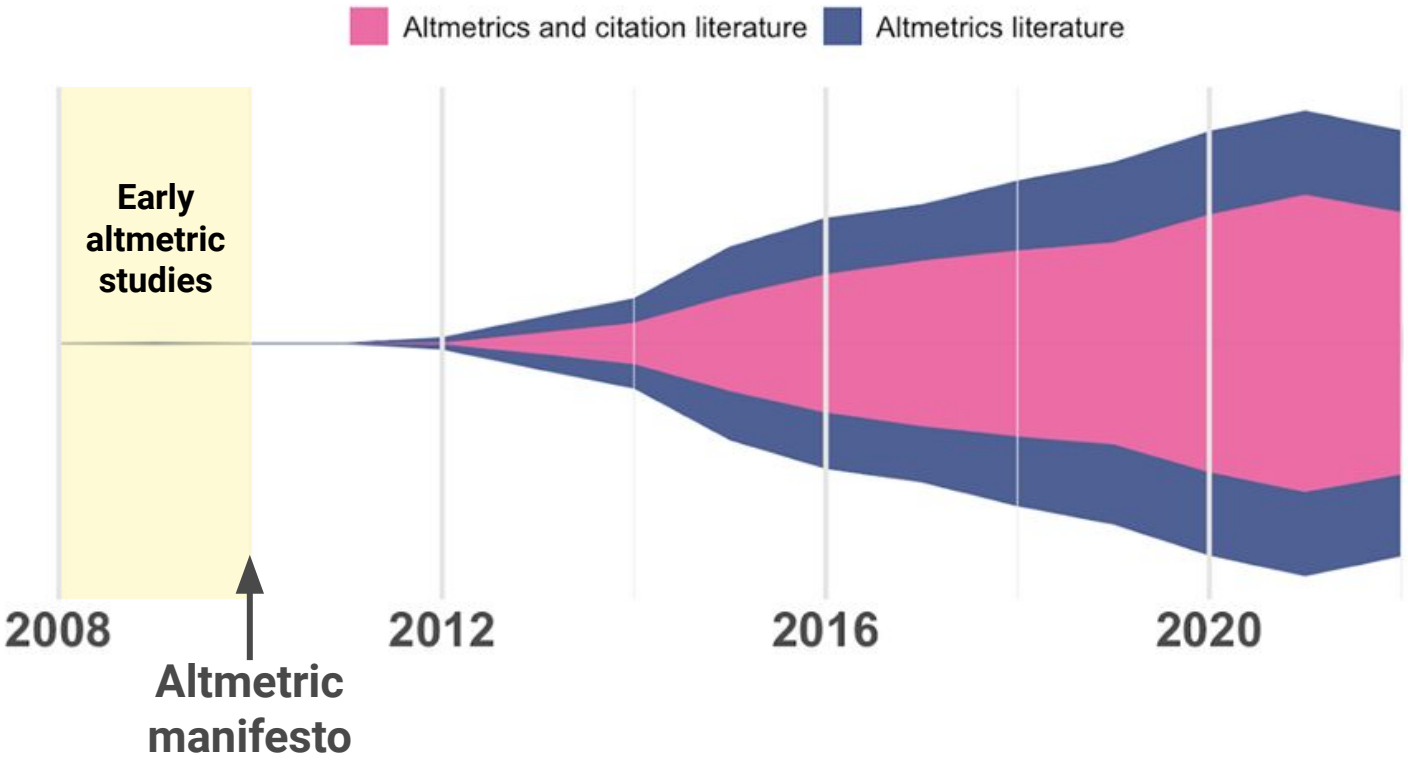


The birth of altmetrics

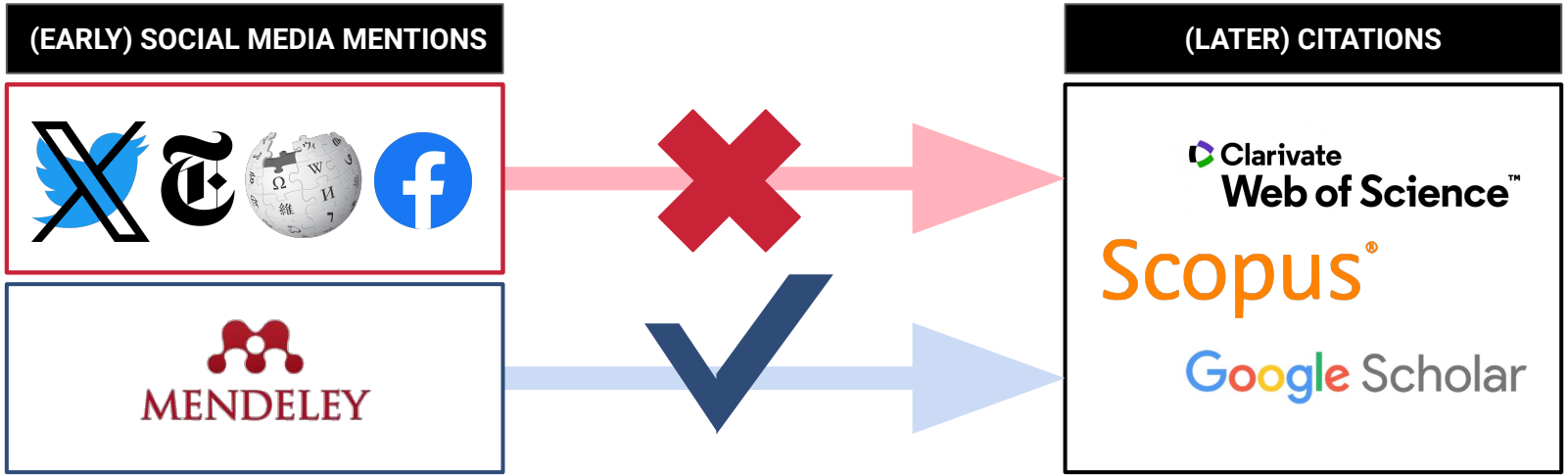


SOCIAL MEDIA INCLUDES A **LARGE NUMBER** OF **INTERACTIONS** AROUND **SCIENCE**

The birth of altmetrics



Towards a new generation of altmetrics



The **lack of correlation** between (almost all) altmetrics and citations gives rise to research focused on understanding this phenomenon and the context of mentions

Major **challenges** identified in **altmetric research**

1

A **large data** landscape of **social interactions** around science that remain unexplored

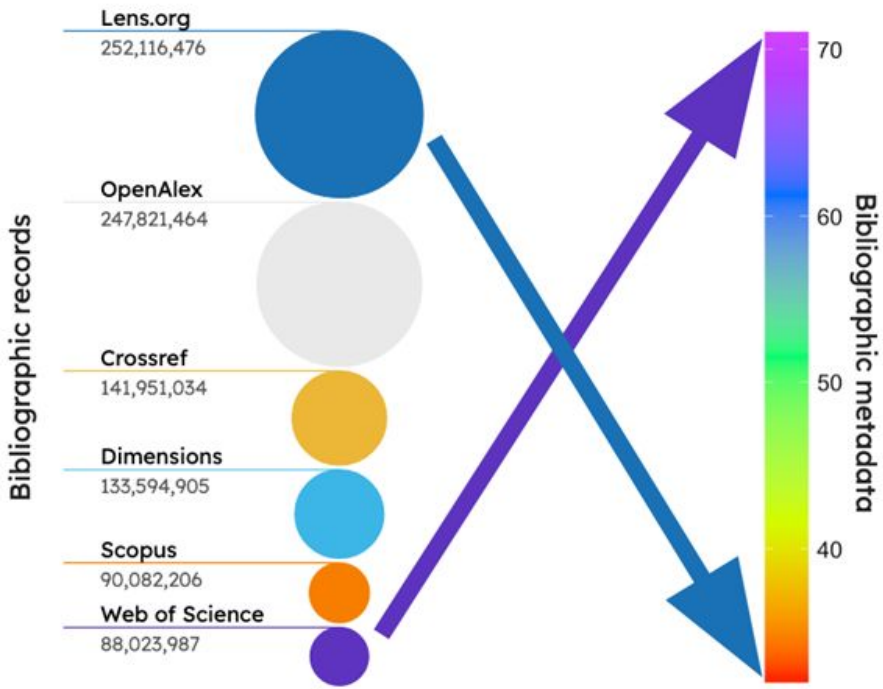
2

Lack of **scientometric methods** that exploit social media data

Major challenges in bibliographic data

The bibliographic universe has undergone an **avalanche** of databases, metadata and records

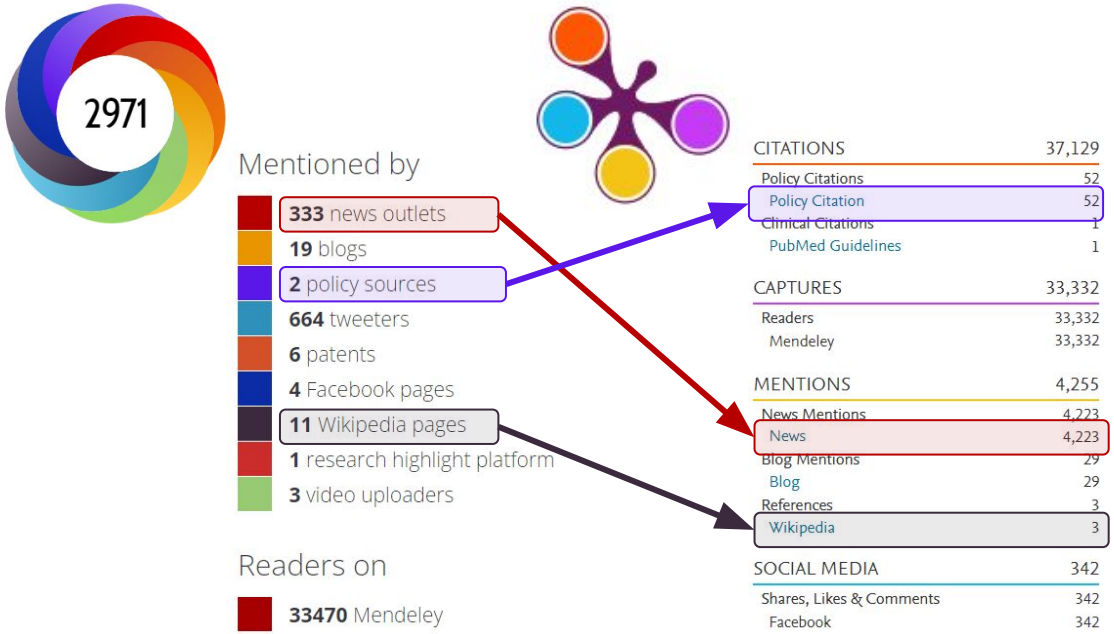
There is a noticeable gap between the **quantity** and **quality** of data



Major challenges in social media data

Global Cancer Statistics 2020

GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries



In altmetrics, **data aggregators** show important differences that can determine the direction of the research

Aggregators only include the main and most **popular metrics**

UNIVERSIDAD DE GRANADA

オルトメトリクス

MAIN CHALLENGES

Social media data **collecting** and **processing**

1. **Dependence** on data provided by aggregators
2. Lack of source **exploration**

Applying classic scientometrics

Scientometrics offers a wide range of methods that have demonstrated their **usefulness** in **quantitative analysis** of science through bibliographic records

Many are based on the analysis of **citation relationships** and **patterns**

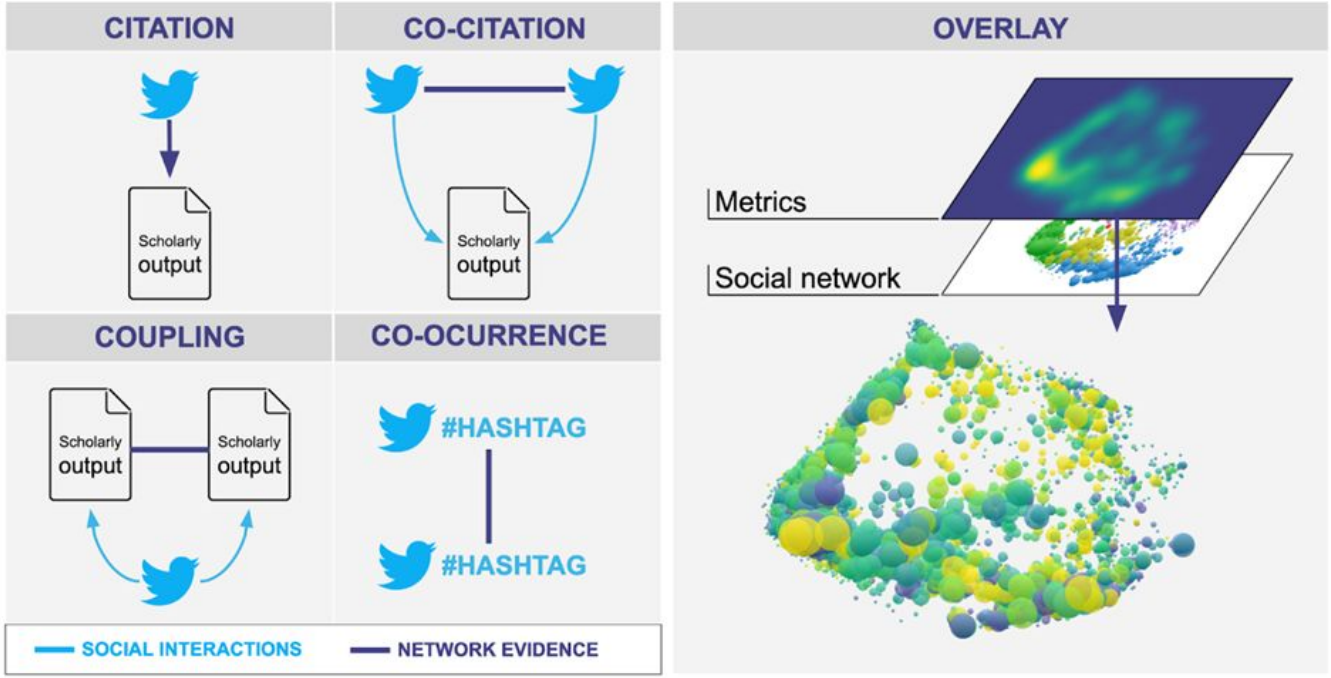
TABLE I

	Total	1921–1925	1916–1920	1911–1915	1906–1910	1901–1905	1896–1900	1891–1895	1886–1890	1881–1885	1876–1880	1871–1875
Ber.	686	78	30	67	115	79	64	60	56	53	44	33
J. Chem. Soc.	390	122	37	60	45	47	21	20	5	2	1	.....
Ann.	278	26	8	37	33	23	22	21	19	18	13	.....
Z. physik. Chem.	191	53	6	21	29	19	28	16	6	.....	.....	.....
Compt. rend.	126	26	3	23	15	23	15	21	7	9	8	.....
J. Phys. Chem.	93	42	13	13	5	1	1	.....	.....	.....	.....	.....
Ann. Physik	93	18	4	28	13	6	0	0	6	5	2	.....
J. Biol. Chem.	80	41	16	14	7	.....	.....	.....	.....	.....	.....	.....
Am. Chem. J.	70	.....	.....	9	21	20	14	8	4	2	1	.....
Z. anorg. Chem.	68	21	11	5	8	11	6	2	.....	.....	.....	.....
Ann. Chim.	68	5	0	6	9	7	3	5	1	8	4	2
Bull. Soc. Chim.	60	16	3	4	7	10	4	4	3	4	2	1
Proc. Roy. Soc.	55	30	5	4	8	5	1	0	1	.....	.....	.....
J. Ind. Eng. Chem.	53	33	10	5	1	.....	.....	.....	.....	.....	.....	.....
Z. Phys.	51	41	5	.....	.....	.....	.....	.....	.....	.....	.....	.....
Monatsch.	51	2	1	21	5	9	3	2	5	3	.....	.....
J. prakt. Chem.	50	6	1	2	2	6	3	12	6	6	2	2
Phil. Mag.	49	17	14	4	2	3	3	1	1	0	0	1
Gazz. chim. ital.	44	10	6	2	6	4	8	4	3	0	1	.....
Phys. Rev.	44	23	8	3	5	4	.....	.....	.....	.....	.....	.....
Physik. Zeit.	41	26	0	7	3	.....	.....	.....	.....	.....	.....	.....
Z. Elektrochem.	37	11	13	4	4	4	1	.....	.....	.....	.....	.....
Biochem. Z.	37	18	2	9	10	.....	.....	.....	.....	.....	.....	.....
Rec. trav. chim.	36	14	5	2	2	2	5	4	1	1	.....	.....
SCIENCE	27	22	3	.....	.....	.....	.....	.....	.....	.....	.....	.....
Trans. Far. Soc.	24	18	0	1	0	1	.....	.....	.....	.....	.....	.....
Proc. Nat'l Acad.	22	19	0	.....	.....	.....	.....	.....	.....	.....	.....	.....
Nature	21	13	5	1	.....	.....	.....	.....	.....	.....	.....	.....

The abbreviations used above and in the tables to follow are those accepted by *Chemical Abstracts* and may be found in their list of periodicals abstracted, issued October 20, 1926.

Gross, P. L. K., & Gross, E. M. (1927). College libraries and chemical education. Science, 66(1713), 385-389. https://doi.org/10.1126/science.66.1713.385

Applying classic scientometrics



Many of these classic methods have been successfully **adapted** to Twitter

UNIVERSIDAD DE GRANADA

オルトメトリクス

MAIN CHALLENGES

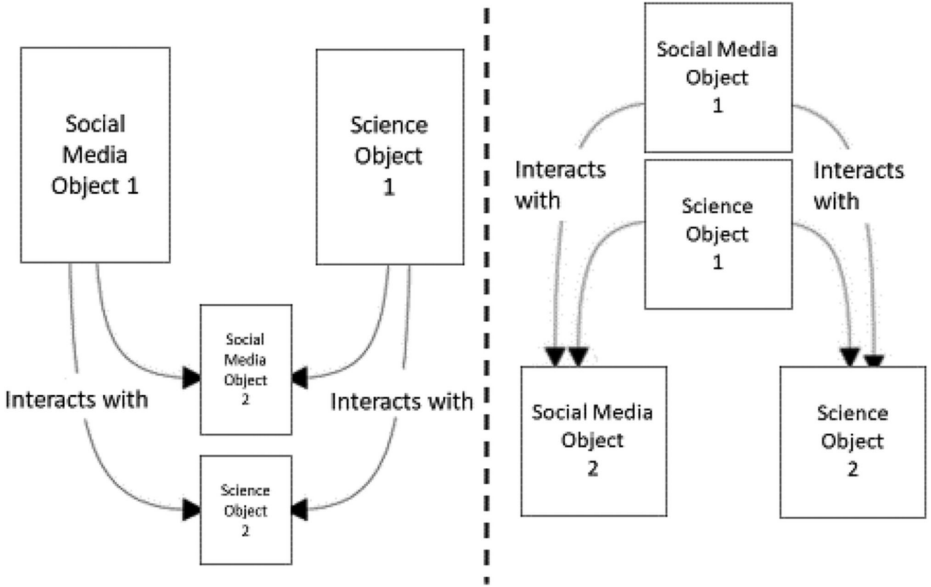
Adapting **standard scientometrics** methods in **altmetrics** research

1. Absence of approaches applying these methods to sources other than **Twitter**
2. Lack of **comparisons** between the social and scientific perspective of scientific knowledge

Towards New Horizons

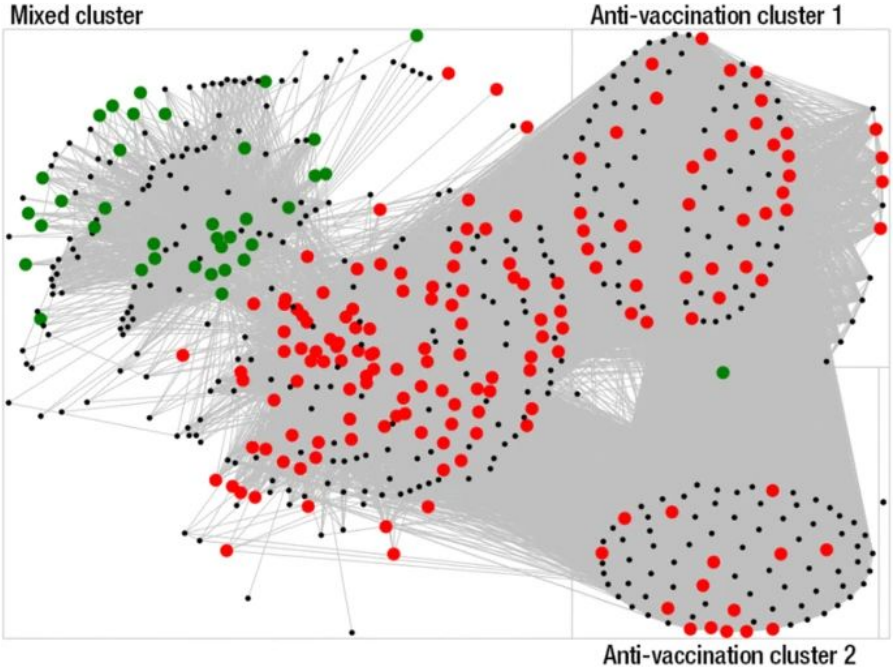
Beyond classic methods, **new methods** are needed that truly leverage the rich context and environment in which interactions around science occur

Evidence from different sources are also not combined



Costas, R., De Rijcke, S., & Marres, N. (2021). "Heterogeneous couplings": Operationalizing network perspectives to study science-society interactions through social media metrics. Journal of the Association for Information Science and Technology, 72(5), 595-610. https://doi.org/10.1002/asi.24427

Towards New Horizons



● *pro-science*
● *anti-vaccination*

Van Schalkwyk, F., Dudek, J., & Costas, R. (2020). Communities of shared interests and cognitive bridges: The case of the anti-vaccination movement on Twitter. Scientometrics, 125(2), 1499-1516. https://doi.org/10.1007/s11192-020-03551-0

There is great potential in studying engagement around academic objects to better understand **science-society interactions**

UNIVERSIDAD DE GRANADA

オルトメトリクス

MAIN CHALLENGES

Many **opportunities** and **possibilities**

1. Many unexplored opportunities in science maps and social network analysis

2. Updated research required due to the ever-changing social media landscape

Main objectives

OBJECTIVE 1

To explore challenges in processing large bibliographic and social media **data**, with a focus on combining them for altmetric studies

MAIN EXPECTED RESULT

Providing **tools** and **curated datasets** for altmetric research

OBJECTIVE 2

To adapt **scientometric methods** for social media, aiming to create science maps from Wikipedia that reflect social attention

MAIN EXPECTED RESULT

Map the science structure through the lens of Wikipedia

OBJECTIVE 3

To develop novel methodologies for scientific mapping by combining **social and semantic data** from diverse sources

MAIN EXPECTED RESULT

Implementation of **innovative methodologies** that integrate interactions and interests

Traditional scientific method

OBSERVATION 1

Exploration and mapping of the interaction between science and society on social media

HYPOTHESIS FORMULATION 2

Adaptation of traditional scientometric methods to fit social media environments

OBSERVATION GATHERING 3

Using results from applied methods to social media and validating with indicators from social network analysis

CONTRASTING THE HYPOTHESIS 4

Comparison of the acquired results with those published by other novel related proposals

HYPOTHESIS VALIDATION OR REFUSAL 5

Validation of the hypothesis through the conducted experiments and results

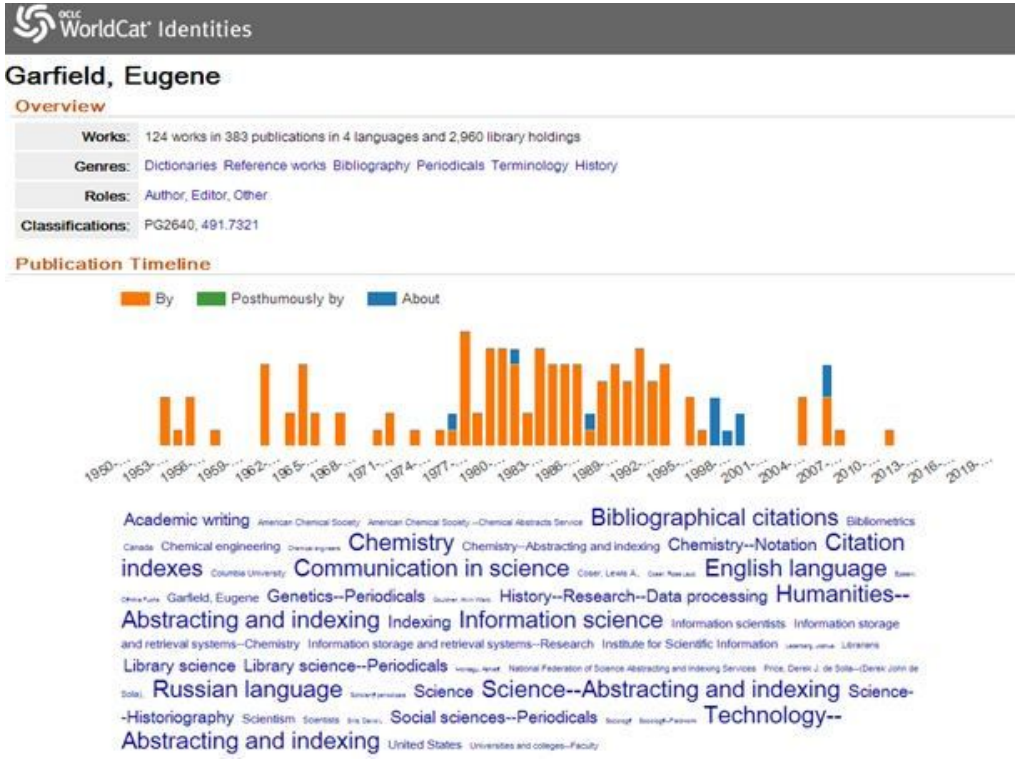
SCIENTIFIC THESIS 6

Extraction, redaction and acceptance of the conclusions

Main contributions

	OBJECTIVE DESCRIPTION	APPROACHES	RESULTING PAPERS
OBJECTIVE 1	Addressing **large-scale data** analysis from literature and social media	EXPLORING NEW SOURCES	*Scientometrics* SCIE – Q2
		PROVIDE CURATED DATA	*QSS* ESCI – Q1
OBJECTIVE 2	Adapting **scientometrics methods** to map science through the social media prism	ADAPTING CO-CITATION TO WIKIPEDIA	*J. of Informetrics* SCIE – Q1
		MAPPING WIKIPEDIA	*PLOS ONE* SCIE – Q2
OBJECTIVE 3	Proposing **scientific mapping** methods combining social and semantic information	ADAPTING OVERLAY MAPS	*FEMS Mic. Letter* SCIE – Q3
		SOCIO-SEMANTIC MAPS	*Scientometrics* SCIE – Q2

Exploring WorldCat Identities



CONTEXT

OCLC conducted an experimental project, **WorldCat Identities**, in which it generated **author profiles** from the WorldCat catalog with various **indicators**

OBJECTIVES

To **explore** WorldCat Identities as an information source and conduct a **case study** with scientometrics authors

Thesis objective 1 *Contribution* 1

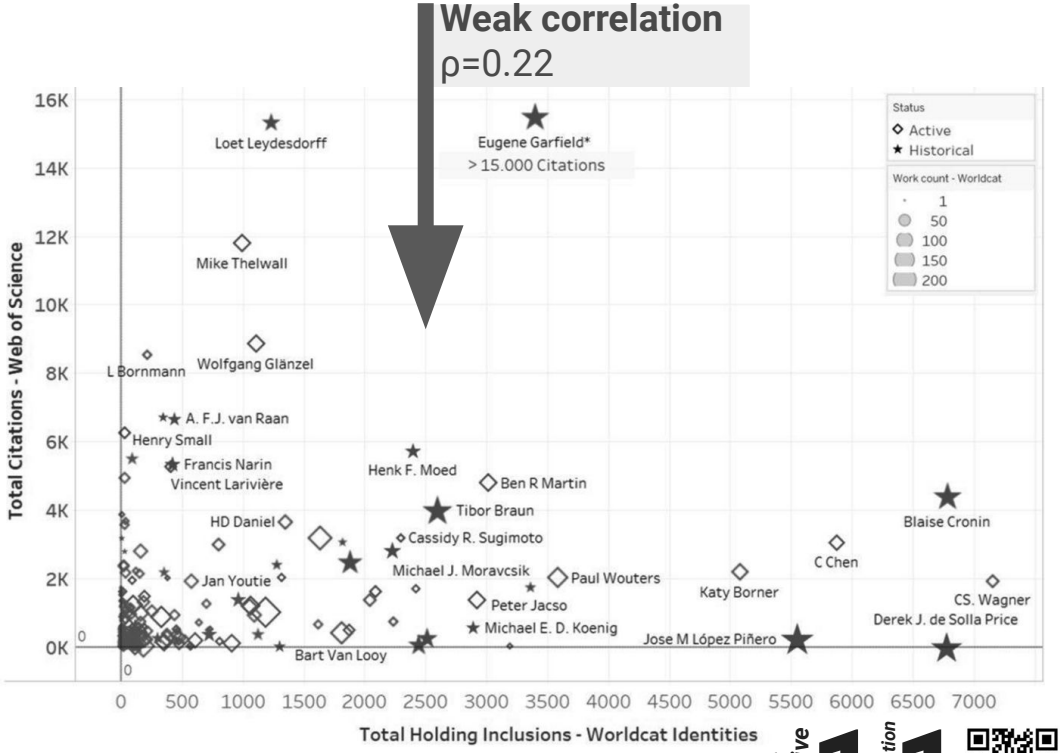
Exploring WorldCat Identities

A python™ **package** was built to retrieve author profiles and their metadata.
After retrieving data on scientometricians, we compared the **library holdings** with the citations

METHODS

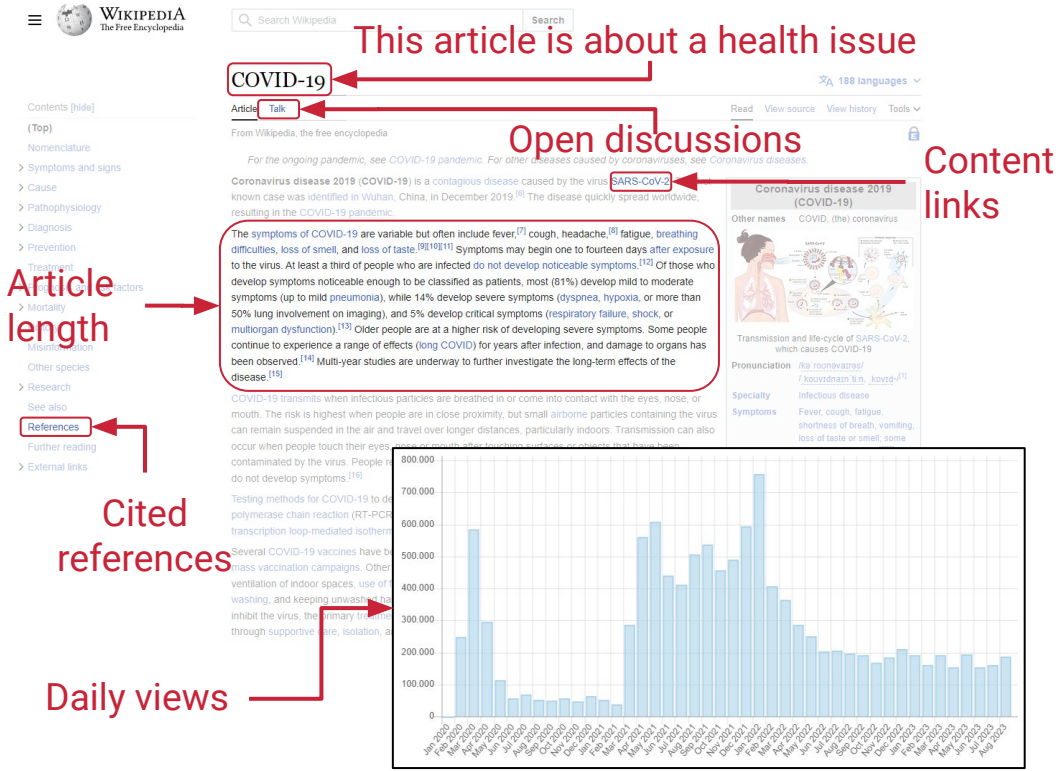
Despite the **disambiguation issues**, it proved useful as an altmetric tool, offering a **new dimension of influence** distinct from that of citations

FINDINGS



Thesis objective 1 Contribution 1

The analytical possibilities of Wikipedia



CONTEXT

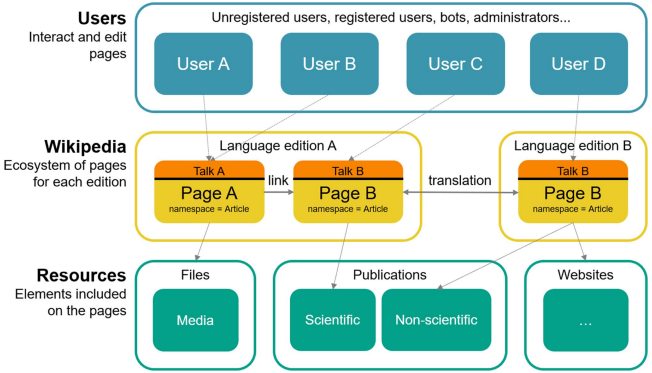
Wikipedia conceals an intricate, **unexplored infometric ecosystem** with vast potential to capture various social interactions

OBJECTIVES

To establish a **framework** for Wikipedia, create a large open knowledge graph based on it, and conduct a **descriptive quantitative study** of Wikipedia

Thesis objective 1 Contribution 2

The analytical possibilities of Wikipedia



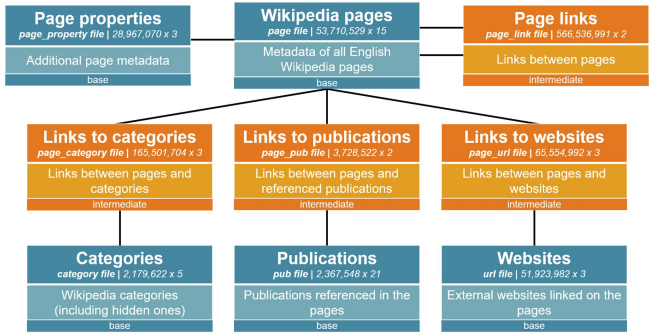
Wikimedia Downloads

+

Wikimedia API

+

zenodo



Wikipedia Knowledge Graph, dataset and description free at: **10.5281/zenodo.6346899**

Identification of elements involved in Wikipedia activity

Data processing

R python™

Developing of Open Wikipedia Knowledge Graph

Thesis objective 1 *Contribution* 2

The analytical possibilities of Wikipedia

METHODS

Using **data science** methods, heterogeneous data from the English Wikipedia was retrieved and processed to construct a **knowledge graph**

FINDINGS

Wikipedia has valuable metrics that capture different dimensions of **social attention** that contribute to contextualising **how science is consumed by society**

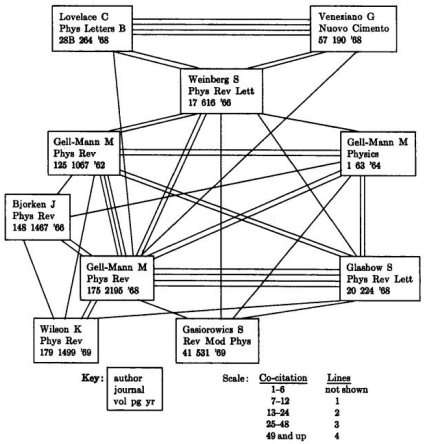
N. of articles → / *Wiki Metrics ↓*	All articles	Featured articles	Featured lists	A	Good	B	C	List	Start	Stub
	6,328,134	5945	3816	958	34,004	109,019	394,065	253,066	1,818,356	3,079,778
Editors	48.38	516.93	179.13	176.80	275.71	297.62	165.36	56.27	63.13	22.85
Edits	101.92	1491.35	593.61	564.91	724.13	705.41	369.89	159.80	129.52	40.23
Linked	80.53	725.25	175.84	202.01	330.18	417.00	234.08	107.34	93.03	55.70
Links	87.77	329.68	270.16	236.56	224.88	233.87	164.23	174.78	101.28	69.90
Age	9.59	14.33	11.52	12.74	12.06	12.47	10.92	9.13	10.45	9.20
Length	7844.68	61,248	51,549	43,329	39,444	35,009	21,676	18,202	10,033	3748
Talkers	5.38	66.17	16.62	27.90	29.64	28.16	15.03	4.98	6.56	3.64
Talks	9.19	258.40	42.36	92.21	88.56	88.35	35.32	9.07	9.69	4.32
Views	3345.07	64,801	26,685	16,011	29,229	30,359	15,829	3777	4094	710
References	4.6	53.95	55.49	31.76	38.87	26.51	15.40	9.20	5.79	1.84
Pub. Ref.	0.59	14.27	2.34	8.51	5.83	4.77	2.37	0.53	0.69	0.22
URLs	10.33	58.03	67.32	33.32	46.10	40.31	25.95	22.82	12.90	6.09

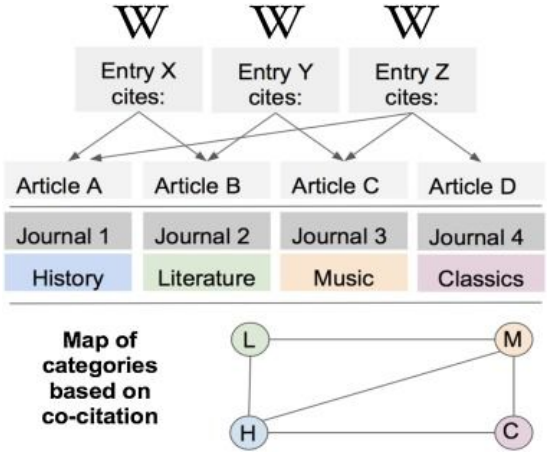
The **quality** of the articles is linked to metrics such as **page views** or **references**

Thesis objective 1 *Contribution* 2

Adapting co-citations to Wikipedia

Proof of concept





From **paper** co-citations

to **Wikipedia** co-citations

CONTEXT

Wikipedia articles engage with literature in a similar way to scientific publications, providing a good environment for adapting **classic scientometric techniques**

OBJECTIVES

To transfer co-citation methodology to Wikipedia and test the method by mapping the structure of the Humanities

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265-269. https://doi.org/10.1002/asi.4630240406

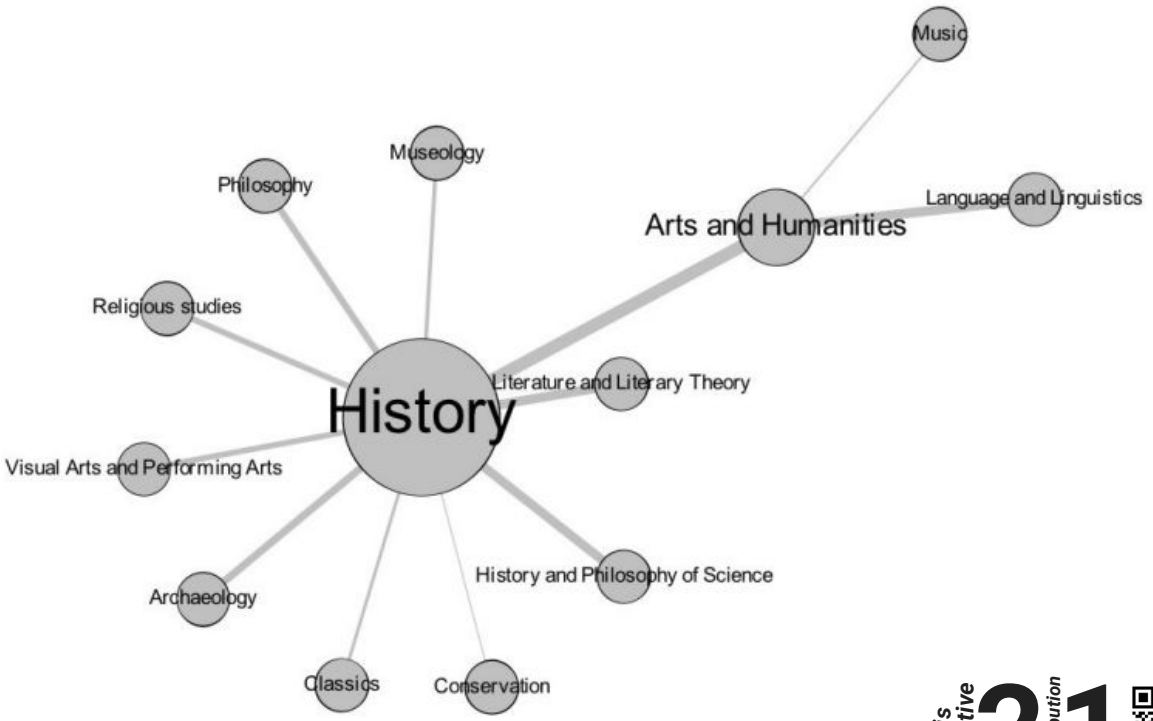
Thesis objective 2 Contribution 1

Adapting co-citations to Wikipedia

Proof of concept

Pathfinder networks (PFNETs)

Due to the characteristics of these networks, in which there is a high degree of connectivity between all the nodes in the network, it was decided to apply the **Pathfinder algorithm** to eliminate weak links



Thesis objective 2 Contribution 1

Adapting co-citations to Wikipedia

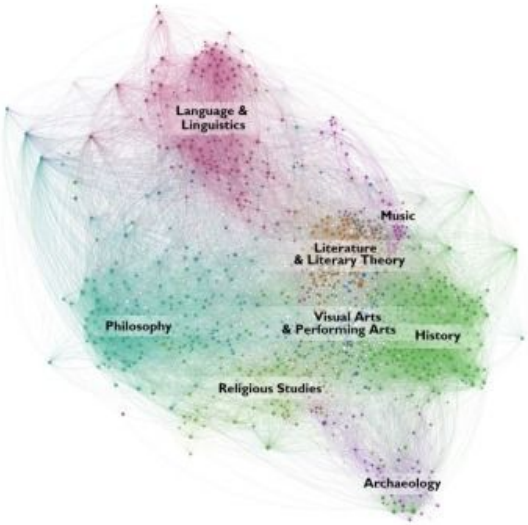
Proof of concept

METHODS

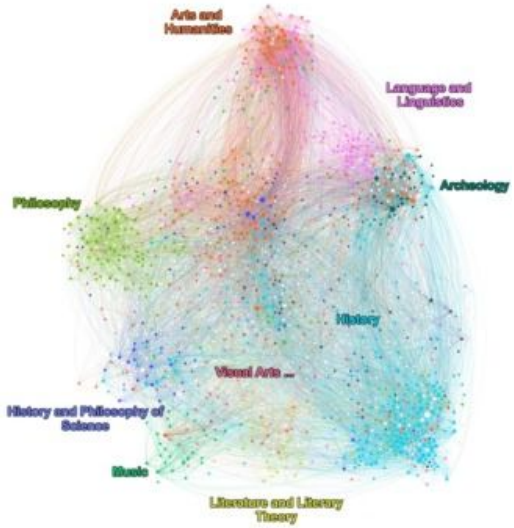
From Wikipedia references, **relationships between publications** are established that can be **aggregated** by journals and scientific disciplines

FINDINGS

The case study **validates** the adaptation of the co-citation method to Wikipedia and demonstrates its relevance by **exposing differences** between the academic and social realms



Scopus perspective



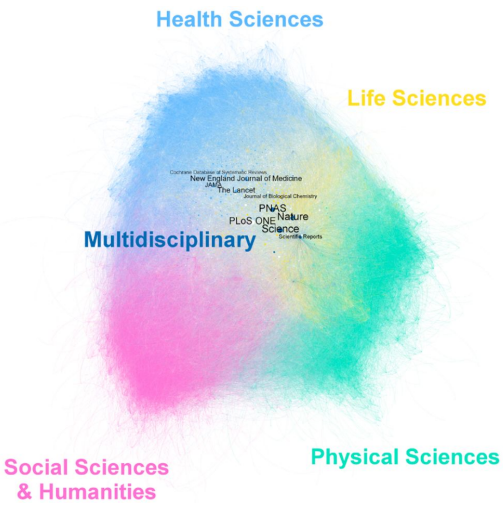
Wikipedia perspective

Richardson, M. (2013). Mapping the multidisciplinarity of the Arts & Humanities. Research Trends, 1(32), 5. https://www.researchtrends.com/cgi/viewcontent.cgi?article=1145&context=researchtrends

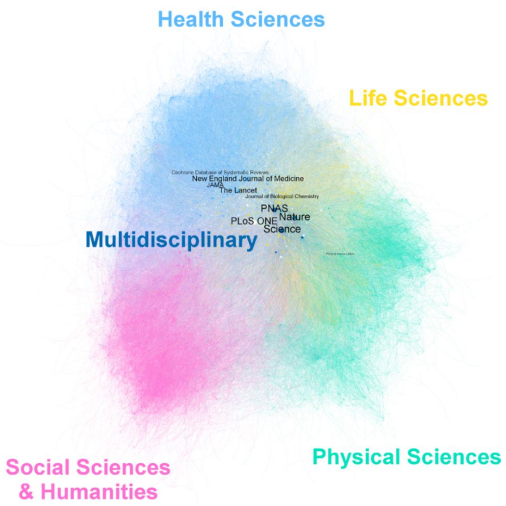
Thesis objective 2 Contribution 1

Adapting co-citations to Wikipedia

Large-scale mapping



Full network



Pathfinder

CONTEXT

After validating the co-citation method on Wikipedia, the ambition was to apply it on a **large scale** and to make a comprehensive analysis of the science

OBJECTIVES

Mapping the structure of science and offering a general **portrait** of science through the English Wikipedia

Thesis objective 2 Contribution 2

Adapting co-citations to Wikipedia

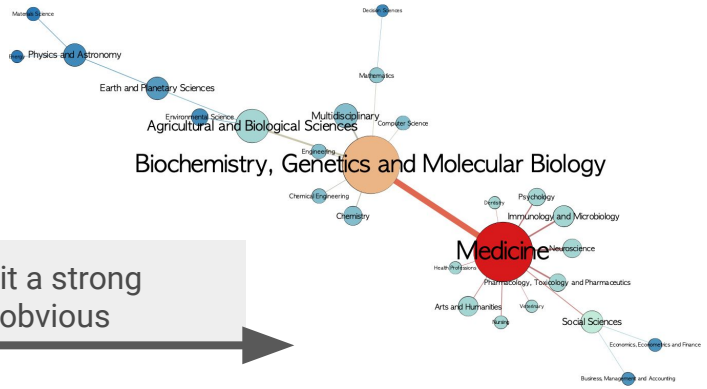
Large-scale mapping

METHODS

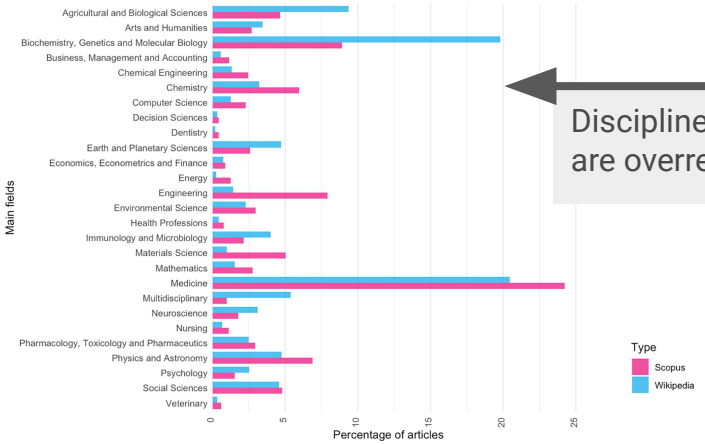
A **co-citation** methodology was adapted to generate **Pathfinder networks (PFNET)** from all references in the English Wikipedia articles

FINDINGS

Regarding the scientific realm, **discrepancies** were detected in the attention to certain disciplines and a shared interest in **high-impact** publications



Some disciplines exhibit a strong relationship that is not obvious

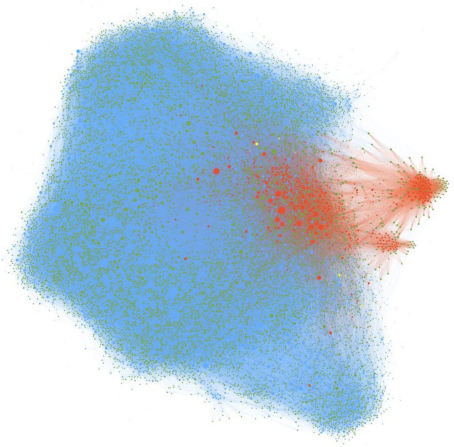


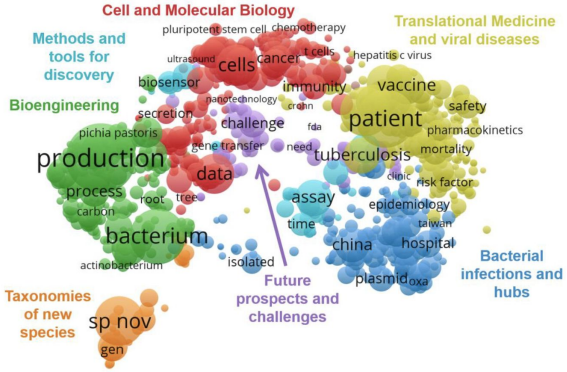
Disciplines such as Biochemistry are overrepresented in Wikipedia

Thesis objective 2 Contribution 2

Developing new methodologies

Shared interests





Social mentions
from Twitter, news outlets, and policy documents

Thematic landscape
from publications' titles

CONTEXT

Although the topics that capture the attention of a discipline in social media have been studied, potential differences in interests between different media have not been delimited

OBJECTIVES

To visualize **key social interest topics** in Microbiology identified via altmetric data

Thesis objective 3 Contribution 1

Developing new methodologies

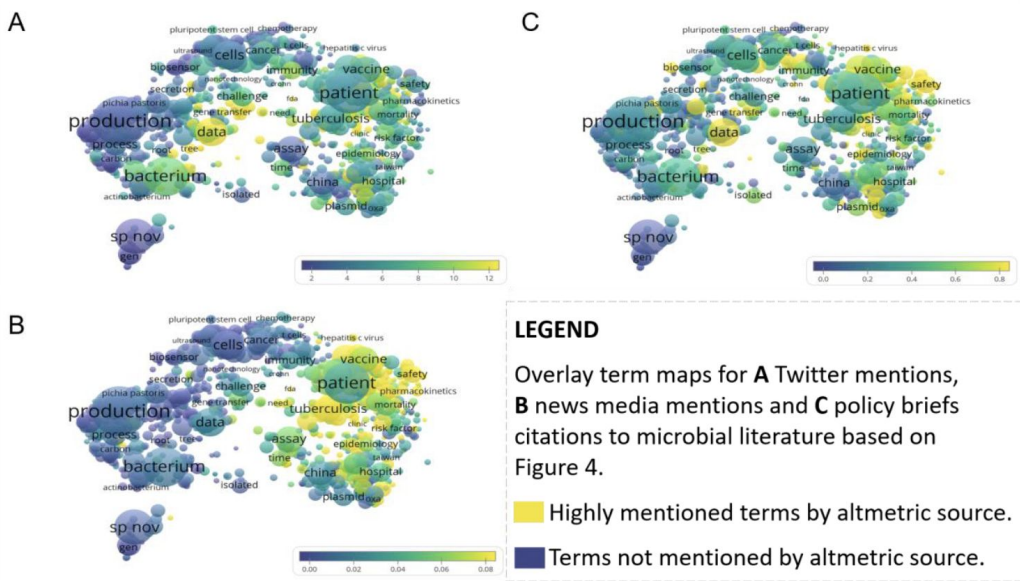
Shared interests

The **overlay maps** have been adapted to the altmetrics to identify the **topics of interest** of each source

METHODS

Not only have **differences in interests** among the various studied social media been highlighted, but also **peaks of attention** over time

FINDINGS



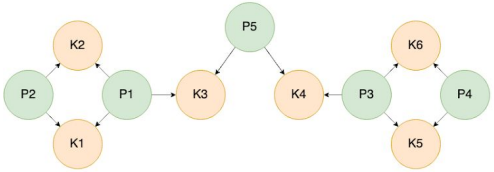
Thesis objective 3.1 Contribution

Developing new methodologies

Socio-semantic networks

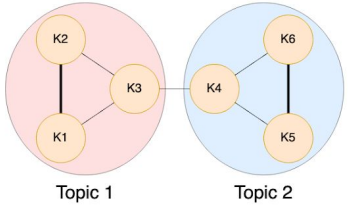
1. Publication-keyword network

Network of scientific publications (Pn) and their Web of Science author keywords (Kn).



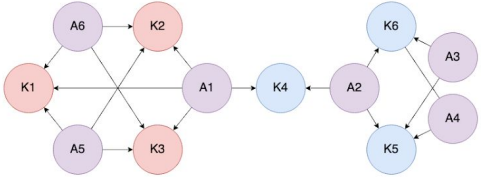
2. Semantic map

Network of author keyword co-occurrence. Topics are identified by community detection.



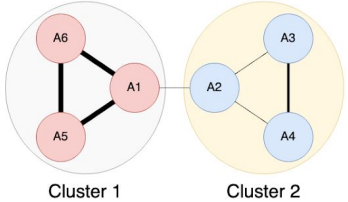
3. Actor-keyword network

Network of social actors (An) that mention keywords, based on the papers mentioned. Keywords belong to one of the topics identified in the semantic map.



4. Socio-semantic network

Network of co-occurrence of actors combined with the semantic map. Each actor belongs to a topic based on its keyword mention. Clusters of actors are identified by community detection.



CONTEXT

Altmetrics allow exploring and profiling social actors who discuss and share scientific literature, but it is a challenge to **identify** and **characterize communities**

OBJECTIVES

To develop and validate a new method for **profiling social media users** based on their interest on research topics

Thesis objective 3 Contribution 2

Developing new methodologies

Socio-semantic networks

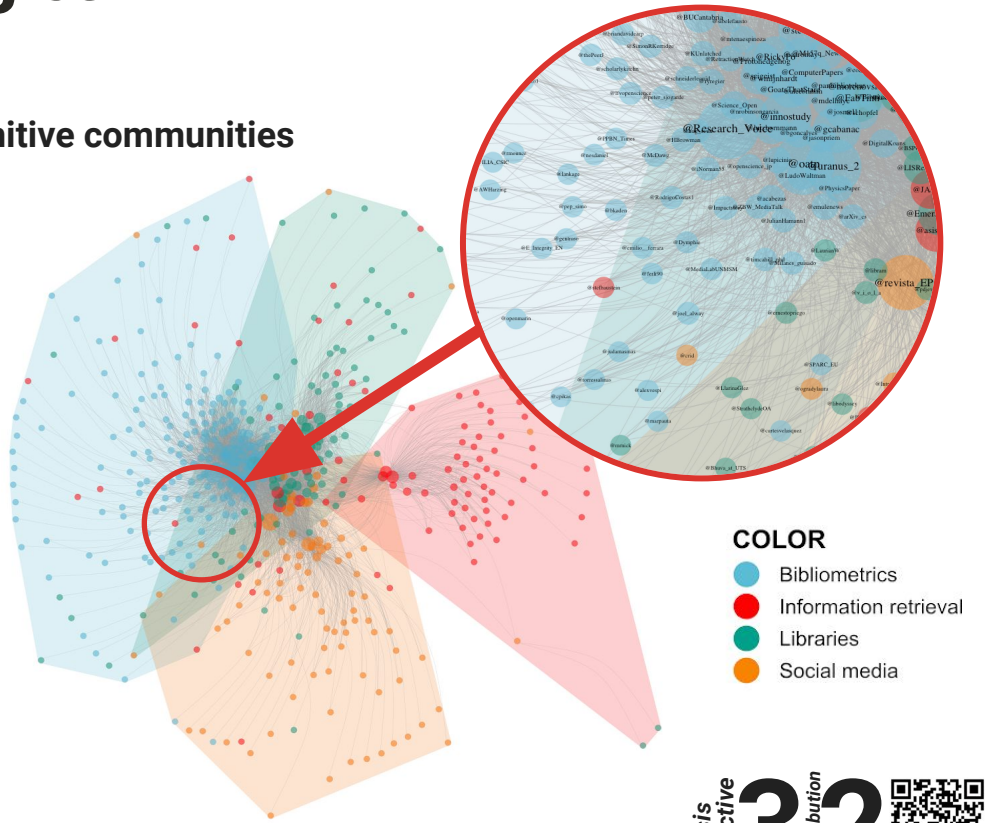
METHODS

A proposal for **socio-semantic networks** has been developed through a package in R and applied to IS&LS and Microbiology

FINDINGS

The proposed socio-semantic network can visually simplify social relations and thematic interests, illustrating whether they **align** or **differ**

Cognitive communities



Thesis objective 3 Contribution 2

Primary findings

OBJECTIVE 1

WorldCat Identities offers unique insights for author analyses but requires careful data validation due to various challenges

Wikipedia has untapped altmetric potential; our framework and metrics highlight its value but emphasize the need for intensive data processing

OBJECTIVE 2

Wikipedia's citation analysis in the Humanities reveals History as a dominant topic, but Humanities citations are just 5% of Wikipedia citations

A broad co-citation analysis mapped Wikipedia's perspectives on science, emphasizing areas like Biochemistry, and showing unique citation patterns, with only 13% to OA journals

OBJECTIVE 3

Overlay maps revealed distinct Microbiology interests across platforms, shedding light on how such attention is generated

Our socio-semantic approach identified genuine scientific interests on social media, hinting at its broader applicability

Social media & science interactions

1. Pioneering exploration of social media to study **science-society relations**
2. Applied **social data mining** for data extraction and processing in altmetrics
3. **Wikipedia**: a significant, yet underutilized, platform for altmetrics
4. **Adapted scientometric methods** to map science through social media's lens
5. Proposed **new methods** merging social and semantic data
6. Emphasized the vast and varied ways science and **society engage on social platforms**

This thesis reveals the transformative role of social data mining for science communication research

UNIVERSIDAD DE GRANADA

オルトメトリクス

BIG DATA TECHNIQUES APPLIED TO THE **STUDY** AND **CHARACTERISATION** OF **SCIENTIFIC ACTIVITY** ON **SOCIAL MEDIA**

by Wenceslao Arroyo Machado
PhD Advisors Enrique Herrera Viedma & Daniel Torres Salinas