OPERAS-D

DESIGN FOR OPEN ACCESS PUBLICATIONS IN EUROPEAN RESEARCH AREAS FOR SOCIAL SCIENCES AND HUMANITIES

WP 3.2 Use and Impact of Open Access Monographs

The Visibility of Open Access Monographs in a European Context: Full Report

30 January 2018

Cameron Neylon,[1,2] Lucy Montgomery,[1,2] Alkim Ozaygen,[1,2] Neil Saunders[2] and Frances Pinter[1,2]

1. Centre for Culture and Technology, Curtin University, Kent St, Bentley, Western Australia 2. Knowledge Unlatched Research, United Kingdom

The project has received funding from European Union's Horizon 2020 research and innovation programme under grant agreement 731031

Draft version

DESIGN FOR OPEN ACCESS PUBLICATIONS IN EUROPEAN RESEARCH AREAS FOR SOCIAL SCIENCES AND HUMANITIES

Deliverable 3.2

The Visibility of Open Access Monographs in a European Context

Grant Agreement number	:	731031
Project acronym	:	OPERAS-D
Project title	:	Design for Open Access Publications in European areas for Social Sciences and Humanities
Funding Scheme	:	INFRASUPP-03-2016
Project's coordinator Organization	:	CLEO-CNRS
E-mail address	:	pierre.mounier@openedition.org
Website	:	http://www.operas-eu.org
WP and tasks contributing	:	3.2
WP leader	:	Knowledge Unlatched CIC
Dissemination level	:	PU
Due date	:	30 January 2018
Delivery date	:	30 January 2018

Contents

I. WP 3.2 Objectives

This task addresses the challenges associated with tracking the use and impact of Open Access monographs across open global digital networks.
The task is broken into three parts:

- Mapping the digital visibility of OA monographs made available by the OPERAS network;
- Flagging technical challenges specific to the collection of metrics on usage and impact for OA monographs;
- Identifying opportunities for the more effective integration of information relating to the use of OA monographs into metrics and altmetrics ecosystems

II. Background

OPERAS is a distributed Research Infrastructure (RI) project for open scholarly communication. Its main goal is “to introduce the principle of Open Science and ensure effective dissemination and global access to research results in the Social Sciences and Humanities (SSH)”. The network includes a wide range of mainly European Open Access publishers and research institutions, and is in the process of engaging with a wider international network of potential partners.
The OPERAS Network includes a diversity of participants with differing interests, ranging from traditional publishers with a growing portfolio of Open Access content, through to OA only presses. It includes publishers as well as platforms, technology providers and research institutions. The diversity in OPERAS network participants makes available a range of different financial models, priorities, and technical concerns. The network also continues to grow over time, increasing in both numbers and types of stakeholder organisation. In particular 2017 brought the Latin American SciELO platform to OPERAS as an international partner alongside nine other new partners based in Europe.
OPERAS works in a range of areas. Through its seven working groups and two main H2020 projects its aim is to provide technical and social infrastructures that support Open Access publishing and optimising the use of scholarly content with a focus on Social Sciences and Humanities (SSH). While the network is not exclusively focussed on scholarly books, its focus on SSH means a greater emphasis on questions that relate to books than in many more Science, Technology, Engineering and Medicine (STEM) focussed projects and efforts.

A. The challenge of tracking scholarly books

While the modes and advantages of Open Access for journal articles are now broadly accepted, at least in STEM subjects, the funding models, technology, and most importantly, the advantages for Open Access books pose more of a challenge. Issues that are specific to SSH often combine with issues that are peculiar to book publishing and dissemination. In broad terms there are three areas where books pose a particular challenge compared to journal articles:

1. Digital books are not necessarily made available through a publisher controlled website and may be made available through multiple online platforms.
2. The technical infrastructure for cataloguing, indexing and discovering digital and online books is more recent than that for journal articles and is less consistent and reliable as a result. Dependence on intermediaries for the distribution of digital books means that monograph publishers and platforms also have less direct experience with these systems than tends to be the case for journal articles.
3. Traditionally, book publishers have focussed on the sale of print copies to intermediaries and have had less direct interactions with readers. Existing performance indicators are largely driven by measures of physical distribution. Print remains an important, and often parallel, part of book publishing.

When we consider Open Access books specifically this raises a number of issues. Firstly many of the platforms that exist for distributing books and bibliographic metadata were built with licensed content in mind. This leads to a range of assumptions about tracking of users, their institutions, and their usage that are not applicable to freely accessible Open Access books.

In comparison to journal articles, which made a transition to digital formats much earlier than has been the case for books, the challenges associated with making a shift towards open access are occurring in the context of an incomplete transition to digital distribution and funding models for HSS books.The diversity of HSS monograph publishers - which include many small publishers, as well as library-based and independent presses, adds an additional layer of complexity to the process of integrating OA digital books into digital landscapes of discoverability and use. Firstly publishers often do not host their own digital books on sites under their control but leave this to other platforms. Open Access platforms (such as OAPEN and OpenEdition Books) have developed in parallel with traditionally licensed platforms (such as JSTOR). Established platforms for traditionally licensed content, including JSTOR and Ingenta have also begun to create programs and infrastructure to support Open Access content. Some publishers have begun consciously making the same content available via a variety of distribution sites in order to maximise the visibility and use of digital monographs.[1]

[1] Examples of publishers making Open Access books available via several platforms include

The availability of services intended to help publishers to ensure that Open Access books are optimally integrated into pathways of discovery and use is increasing.

As platforms hosting open access books are maturing and systems for integrating OA content into digital landscapes become part of scholarly workflows, a second issue has emerged. An illustrative example of this is the challenge of applying the Crossref Digital Object Identifier (DOI) infrastructure, developed largely for journal articles, to books. DOIs serve two functions. They are both unique and persistable identifiers for scholarly works, and a *referral* mechanism by which a user may follow a link to arrive at a specific scholarly work. DOIs work well when applied to a single version of record of a journal article that can be found on a website under publisher control, particularly when the demand and use of print copies has been largely replaced by online discovery. DOIs are more problematic for books that might be found on multiple sites in digital form, where the repository is not under the control of the publisher[2]. Challenges of ensuring that correct redirection addresses are maintained in the absence for commercial incentives to ensure that OA content is easy to locate create additional resourcing challenges, particularly for the many smaller publishers operating in the OA monograph space.

The tangle of technical issues involved in identifying and discovering books, combined with a relative lack of investment by platforms in tracking the usage and conversations around books content leads to a reinforcement of a third challenge. Many publishers and presses remain focussed on traditional metrics and KPIs for monograph publishing. These are not focussed on the *usage* of books but on *distribution* through intermediaries - traditionally measured in terms of sales (which also assumes that all publishers make the same effort to sell their books equally). This in turn means a limited demand from presses for detailed information about the use of books, as well as limited capacity to influence the metrics and reporting services provided by platforms.

B. The importance of understanding digital visibility for Open Access books

With the shift towards Open Access, the question of visibility is crucial. It is perhaps a little

---

the four presses discussed in the study *Exploring the Uses of Open Access Books via the JSTOR platform,* available at: http://kuresearch.org/PDF/jstor_report.pdf

[2] It is worth noting that such multiple-location problems are increasing for journal articles with the increasing frequency of self archiving and preprint repositories. Solving this problem well for books may be of value in turn for the journal community. Crossref is currently piloting an approach for supporting multiple DOI for books with the intent of offering coordinated lookup.

harsh to describe traditional metrics as counting copies in warehouses. Nonetheless, even as a straw-person argument it illustrates the point that distribution based measures are simply not helpful for tracking the impact of freely accessible books with online distribution. This is particularly the case given the significantly greater per item investment for books compared to journal articles. Demonstrating the potential value of investing in Open Access, and identifying where that value is realised and the return on investment is greatest is critical to supporting the transition to a future where Open Access is the default for scholarly books. Another important aspect for books is the degree to which they will be accessible to entirely new, and perhaps unexpected, audiences. Scholarly books, much more so than journal articles, have potentially much wider audiences than they currently reach, particularly given the price of many scholarly monographs.

The question of visibility is therefore a complex one. It is clear that there is a need to track scholarly use, including citations and downloads within institutions, as well as the potential to track use and interest by wider publics. We can track the communities that discuss books and ask about how they discover and interact with these texts both online and in print. We can expect books to influence and impact society in ways that are very difficult to track and may not involve a visible trace of usage that we can measure.

The promise for Open Access scholarly books is immense, but the risks and the potential need for investment are also large. If we are to have an evidence-led conversation on strategies for investment, then we need to track the visibility, discoverability, and ultimately the use and impact of scholarly books. In turn, this evidence base will help to change the culture of publishing in HSS, leading perhaps to a greater concern with how an author and the support services in a press can help to shape a work so as to maximise its potential for use and impact.

III. Survey of OPERAS Partners

As part of the visibility project we surveyed OPERAS partners in order to understand how they engage with usage and other data relating to the titles that they publish or host. In particular we were interested in how partners saw the value of such data and how they were interacting with it. We had 18 responses to the questionnaire contributed by presses, platforms, and data and technology providers. The survey was not intended to be quantitative or representative but to provide a view into the thinking and needs of partners. We therefore do not report quantitative results but a qualitative interpretation and categorisation of the responses. The questionnaire rubric is available in Appendix X.

A. Findings

Partners are particular about how they describe themselves. While a range of options

were presented from which survey participants could choose (publisher, platform etc) many participants chose 'other' to provide a free text answer. Sometimes this was to provide greater specificity (e.g. "a library running a press") and sometimes to step outside the categories provided. This was particularly the case for contributors who were involved in funding OA books and other technical platforms.

This echoes the diversity of participants in the OPERAS network. It also suggests a heterogeneity in the ecosystem which we believe to be an important and distinguishing characteristic of book publishing and of scholarly publishing in SSH more generally.

OPERAS partners that are book publishers or book platforms are collecting a range of data. Every respondent who indicated that they were either a publisher or a platform, or both, stated that they (or their partners) were collecting usage data in some form. This ranged from simply collecting web analytics through a tool like Google Analytics or Piwik through to more sophisticated data collection and management pipelines.

Respondents generally showed a good awareness of the technical systems that were involved in collecting data, describing specific tools and systems, as well as standards, principally COUNTER. Named web analytics were fairly evenly split between Google Analytics, which provides a centralised and easily managed means of tracking web usage and Piwik, an open source tool that provides many of the same data collection functions but runs locally, meaning data is not transmitted to Google.

Respondents also showed an awareness of specific limitations in their systems, in several cases describing difficulties in obtaining data specifically on subsets of their collection. Distinctions were made between views and downloads in several cases, although there was limited evidence of that distinction being used in analysis. The two largest hosting platforms OAPEN and OpenEdition Books were the only two to specifically mention the COUNTER standard, with OAPEN passing data to IRUS-UK to generate COUNTER download counts.

The use, processing, and quality assurance of data is patchy. While the awareness of usage data was good, there were substantial differences in the way that data were being used, or indeed not being used. This was connected to differences in the sophistication of data processing and the existence of documented or automated processes. Several publishers and platforms used manual or ad hoc processes to collect data and in several cases there was an indication that data was being collected but not necessarily used.

While the wording of the question focused on 'processing' ('Do you have a process for gathering and managing usage data relating to your OA books?') we had hoped to elicit commentary on data management and quality assurance. However, while issues of data quality were implicit in some answers ("Download data is sent to IRUS-UK who create COUNTER compliant data", "PHP scripts calculate and produce COUNTER metrics...to COUNTER V4...V5 will be implemented [in]...2018") quality assurance processes, such as data validation or cross-checking procedures, re-use of data in internal systems were not

specifically mentioned.

The general lack of concern with quality assurance was consistent with the variety of uses that data was put to. In some cases the use of the data was explicitly limited (e.g. “The books we publish are selected on the basis of scholarly merit”, “Decisions are now based on print circulation, or number of e-books sold through commercial platforms”) to subsidiary and management issues. Others explicitly noted that usage was a key indicator of performance and important for reporting to stakeholders. This was particularly where a case was being made for Open Access, either to authors or to other stakeholders. Several respondents reported being unsure what it could be used for but nonetheless had a sense that it was, or would become, important, with plans for future work in development.

A desire for standards and consistency is in tension with a need for flexibility and contextualisation. Several respondents raised the issue of gathering and integrating data from multiple platforms as a challenge. Of these a number expressed a desire for simplified and standardised tools that could achieve this. At the same time respondents were concerned both about the advisability of combining data from multiple sources, their capacity for analysis of such complex data, and the uses and misuses it might be put to.

> *Analyzing usage data is difficult and can easily lead to wrong assumptions about the impact of a OA book. In our case this could be detrimental to our [authors institutions], which tend to compare their "success" to [other institutions]. This means that we clearly need to understand what the usage data is telling us before we have any use for it.*

A number of respondents expressed a desire for a “dashboard” or other visualisations that could bring multiple data sources together. The consequent need for data integration and standardisation to achieve this was mentioned in one or two responses but awareness of the challenges of comparison across sources appeared to be limited. There was some evidence of a conflation of visualisation with data integration.

Respondents are small organisations with limited capacity. There is a desire for coordination and shared services, infrastructures, standards. A common thread in the responses was that the publishers and platforms who are engaged in Open Access scholarly book publishing are relatively small. This is both a challenge and an opportunity. They have limited capacity to develop internal processes and systems are looking for shared services and platforms to assist in developing usage data capabilities.

> *It would be of great help if we could have a main service from where we could manage all the information related to statistical usage data.*
>
> *[To engage more effective with usage data we would like a]...consortium agreement with Google on how to gather and access usage data.*
>
> *We would like to see an usage aggregation service that consolidates usage data from different hosting partners into one standardised report in an automated way. In turn,*

this should translate into an usage dashboard that can be embedded into platforms and allows customers to use different filters to analyse usage by publisher, region, etc.

[one of our biggest challenges is...optimizing workflow, how to do more work with small resources.

What emerges overall is a picture in which platforms and publishers are implementing tools and approaches locally and using what they are provided with to some degree. There is generally a good technical awareness of the tools being deployed, but less apparent awareness of data curation and quality assurance issues.

Many of the challenges arise from issues of data integration and standardisation. Small, and even medium-sized, players have limited capacity to engage with detailed standards or technical development. Equally there are limitations on what capacity a small organisation can provide to investigate the meaning and context of the data being generated. The majority of data use seemed to be in promotion or advocacy rather than strategic decision making. Concerns were raised about the misuse of usage data or a lack of understanding of its limitations by downstream users.

V. Mapping the digital visibility of OA monographs made available by the OPERAS network

The idea of 'visibility' is not one that has been theorised in detail in existing library literature. Studies tend to focus on issues of information retrieval, addressing precision and recall for a specific information seeking task.[3] 'Visibility' as a concept also at least suggests a concern with serendipitous discovery or non-directed information seeking. In our case we are also concerned specifically with open access books, so 'visibility' presumably includes the clarity of information making about the availability of freely accessible copies of a work.

Ideally we would address the full range of information seeking behaviours, testing for instance the presence of a known book in specific catalogues, the likelihood of a book rising to the top of results for a well-crafted search query, and the potential for serendipitous discovery in a potential reader's regular work-flow. However, developing a well grounded taxonomy of visibility is beyond the scope of this report. We have therefore focussed on testing a range of information sources for the presence and quality of information on a specific set of identified books.

A. Identifying the target books

We developed a simple typology of OPERAS partners involved in the publication of OA monographs; and OPERAS partners involved in the hosting of OA monographs.

OPERAS partners involved in publishing OA monographs were contacted and basic information about their approach to the dissemination of OA books was requested. A metafile for the OA books published by each press was also requested.

In order to maximise the quality of our communications with publishers a personalised approach to email communications was chosen. This included sending an initial email explaining the purpose of our work package and requesting a metafile, as well as specific information needed in order to clarify technical points. Wherever possible we drew on information gathered in WP3.1.

There was substantial variation in the format and content of metadata provided by the

[3] The information retrieval literature focuses naturally on questions of precision and recall with visibility used as a non-technical term in many cases. Criticisms of web-based indicators often focus on the idea that they measure "mere visibility" without strictly defining it. Models that link discovery to usage with a sophisticated application of proxies are rare although see Haustein, Bowman and Costas (2016) in *Theories of Informetrics and Scholarly Communication*, Sugimoto (ed), De Gruyter, Berlin, and essays by Wouters and Cronin in the same volume.

various OPERAS partners. The provided files included Excel, XML, and OAI-PMH feeds. Some partners provided metadata feeds rather than a single output metadata file. These variations also reflected diversity within the partners in their activities as well as in their capacity and workflows. For example, IBL Pan is not a publisher of traditional monographs but involved in alternative approaches to OA books.

Publisher	Provided Metadata?	Format	Comments
UCL Press	Yes	ONIX	
IBL Pan	No		Not publishing traditional monographs
Coimbra University Press	No		Don't currently produce a single metafile as a standard process.
Göttingen University Press	Yes	OAI-PMH XML	
Open Book Publishers	Yes	They sent us an Excel xlsx file	
Ubiquity Press	Yes		Produces OAPEN compliant OAI-PMH
SHARE Press	Yes	OAI-PMH XML	

Table 1. Provision of metadata by OPERAS Partners

Ubiquity Press does not maintain a single meta datafile relating to published books but relies on OAPEN for onward metadata distribution (they are currently developing their own feeds for MARC records). In contrast, while UCL Press also uses OAPEN as a platform and generates OAI-PMH from their internal hosting platform. UCL Press maintains a separate metadata master file.

The metadata provided also showed some weaknesses in the handling of internal information by OPERAS partners. For instance, a small proportion of ISBNS (51 out of 11,000) provided by partners either did not validate via the internal check-sum or could not be automatically validated through a standard regular expression. This suggests that the metadata provided to this project is not generally re-used in internal systems where such errors would be discovered.

Overall, the initial findings in terms of the quality and availability of data from OPERAS partners was that it was inconsistent between partners, and of variable quality. As we will see this leads to a range of problems in information retrieval and visibility analysis.

B. Testing for 'visibility'

To address the question of visibility we conducted three broad kinds of survey:

1. Presence in relevant catalogues.
2. Visibility in web search.
3. Visibility in general information workflows.

The first approach was to survey whether the selected books could be identified within specific catalogues. The catalogues selected for examination were selected to cover common sources for books and open access content. These were WorldCat, BASE, Google Books, DOAB and OpenAIRE. We used their API by searching title and author, to check weather the titles were in their catalogue and to identify the repositories hosting most of these titles.

In each case a search was run using identifiers or titles, with the aim of exhaustively identifying all books that could be confirmed as being available in each catalogue. We used the WorldCat classification API to identify the subjects for each title using ISBN numbers.

We used Bielefeld Academic Search Engine (BASE) which harvests OAI metadata from institutional repositories and other academic digital libraries that implement OAI-PMH. We also checked the titles and their authors via the OpenAire API. As of November 2017, OpenAIRE contains around 23 million documents from 980 compatible data providers. The OpenAire system covers a higher proportion of titles from OAPEN and OpenEdition Books compared to BASE which covers the OBP corpus more completely. Both repositories support search via DOI but not by ISBN, and were designed primarily with journal articles, rather than books, in mind. We also used the Google Books API and compared its results with the DOAB metafile in order to identify whether ISBNs for individual titles were registered in both catalogues.

The second form of visibility was the presence of the book in web search. We used the Webometric Analyst 2.0 tool developed by the group of Thelwall et al.[4] to analyse both the

[4] Thelwall, M. (2009). Introduction to Webometrics: Quantitative Web Research for the

number of pages discovered with a search of the book's title and author's surname, and their top and second level domain names. This gives some indication of geographic location (via country TLDs) and of domain of interest (via TLDs and SLDs, e.g. '.ac.uk' or '.edu' vs '.com or '.com.au').

Finally, we examined a range of services for evidence of activity or presence that would support the visibility of books. We investigated the reported OA status of books with DOIs using the oaDOI service as well as the presence of ISBNs and DOIs relating to the target books in the ORCID 2017 public data dump. We additionally provided Altmetric.com with a complete list of DOIs and ISBNs which was used to interrogate their dataset for information on social and mainstream media that could be linked to one of the target books.

1. Visibility of Target Books in Specific Catalogues

Surprisingly, BASE shows relatively poor coverage overall. In most cases the general catalogues of content show fairly good coverage, but for BASE this is not the case. The visibility results are dominated by the large number of books from OpenEdition Books and from OAPEN. The aggregate results therefore hide some substantial differences between book sources. In particular it is the 29% representation of OpenEdition Books books in BASE, and about 50% coverage of OAPEN that drives the lower numbers for BASE overall.

Social Sciences. San Rafael, CA: Morgan & Claypool.

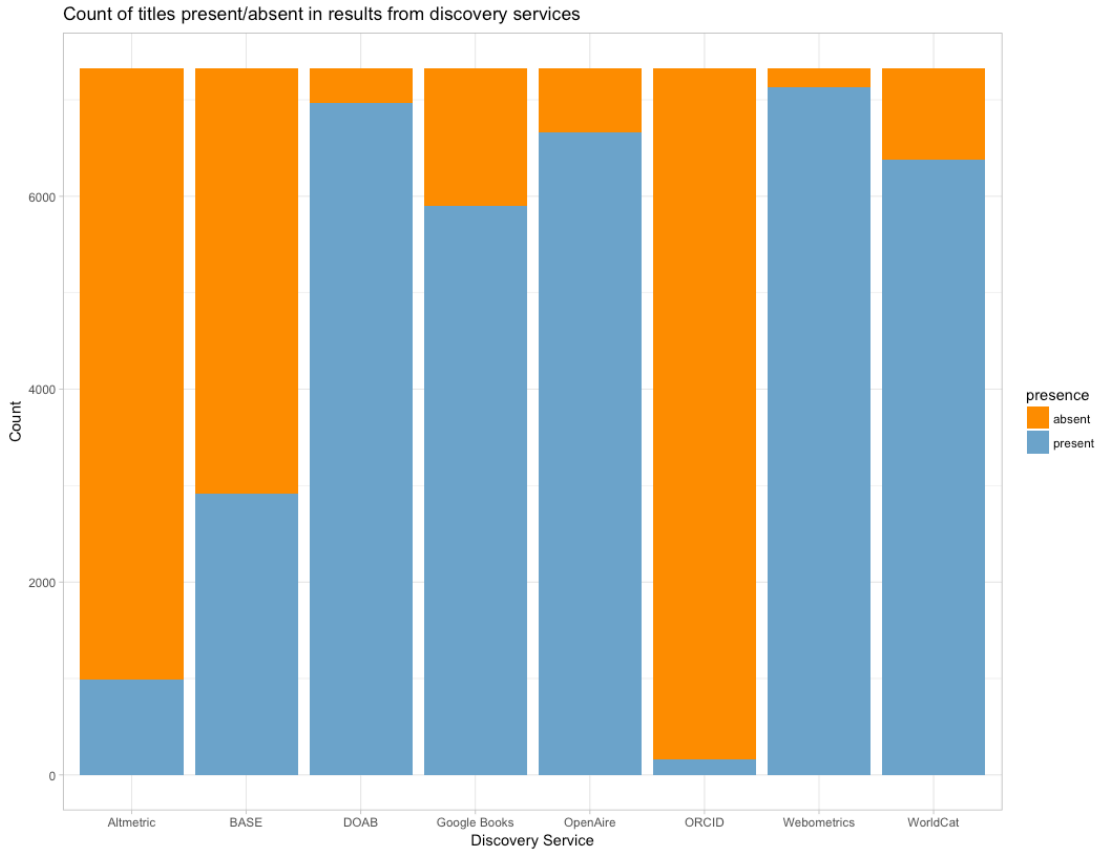


Figure 1. Shows the overall results for all the books in our set across the full range of 'discovery services'. Overall we see good coverage of the books in this set in DOAB, Google Books, OpenAIRE and WorldCat. There is also some form of web search results for most of the books. By contrast, presence in Altmetric results and in ORCID is much less comprehensive.

Coverage in DOAB is uniformly good across all sources of content, OpenAIRE coverage is generally good but weak for EKT, Gottingen, and Napoli University, and a similar pattern is seen for WorldCat, except that Gottingen has excellent WorldCat coverage. Overall the larger three sources (OAPEN, OBP, OpenEdition Books) show better visibility in these catalogues.

There are no obvious differences between catalogue visibility on the basis of language. The analysis here is challenging as a smaller number of European languages cover the majority of books and different content sources have differing language focus. Therefore the question of visibility by language is confounded with that of the visibility by source. Dutch books appear to be underrepresented in both DOAB (58% absent) and WorldCat (65% absent) but well represented in BASE (80%) and OpenAIRE (96%). This may be due to the fact that a significant number of books from the Netherlands in OAPEN do not have an open licence

and are therefore not in DOAB (which is in turn feeding WorldCat).

OPERAS Partner	Google Books	OpenAIRE	DOAB	BASE	World Cat
	(% present)	(% present)	(% present)	(% present)	(% present)
ekt	0	0	100	0	17
Gottingen University Press	89	42	98	39	96
Napoli University Federico II	44	28	97	34	28
OAPEN	73	91	92	49	85
Open Book Publishers	99	74	100	86	94
OpenEdition Books	89	93	99	29	90

Table 2. Visibility of OPERAS partner books in a range of catalogues.

2. Visibility of Target Books in Web Search

Web visibility was determined by running searches with the title and author's name. This provided a score as well a list of referring sites. Due to small numbers it is not possible to draw any comparative conclusions between platforms in terms of their web visibility.
In general terms each platform saw a similar pattern with a high variability in web presence across the collection i.e. some books show a significant web presence with many showing only a small presence. This is an expected pattern given the different level of interest expected across such a large corpus of books. As the corpus also includes older books some references may also not be to the online open access versions.

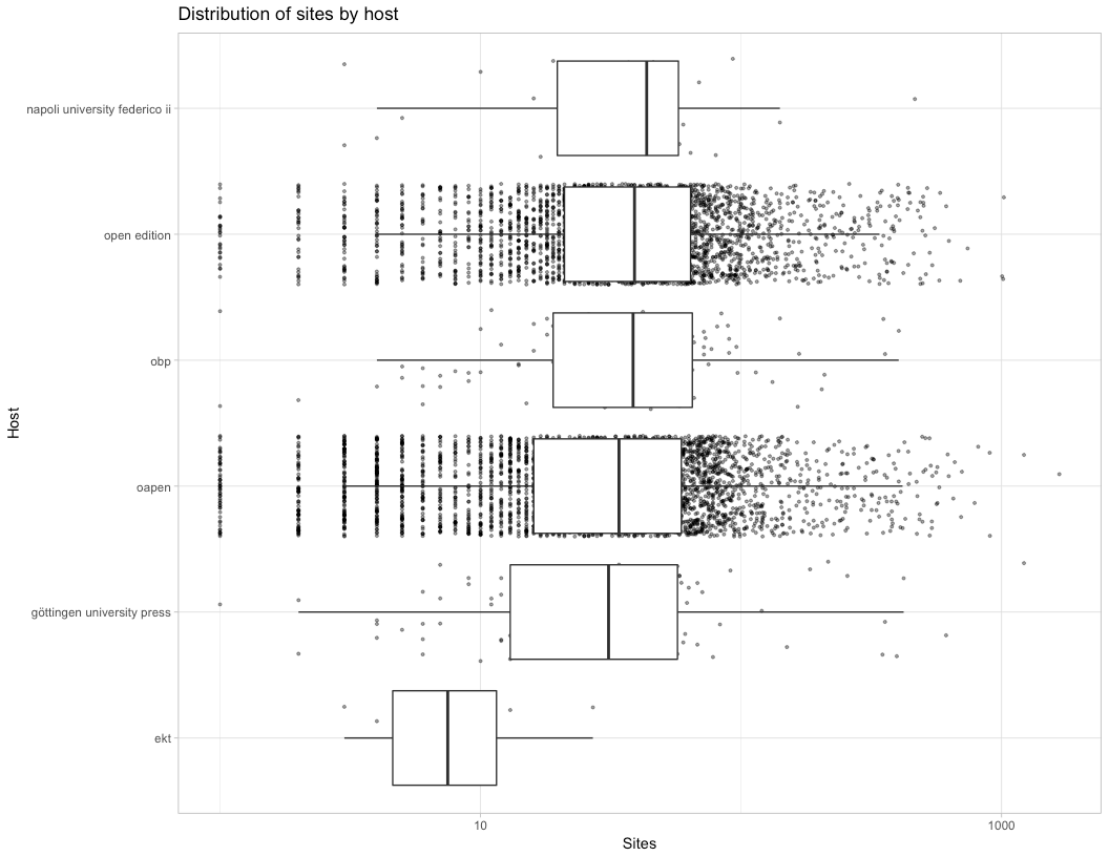


Figure 2. Box-plot showing the number of websites associated via web-search with each published book in the corpus. Each dot represents a single book. The box and line shows the mean and one standard deviation for each host platform.

This form of analysis may be of value in identifying both books with high web visibility and also those which would benefit from additional marketing activity. The analysis is relatively straightforward with the Webometrics tool and can provide quite rich information. As an example we look at how different languages feature in terms of their visibility. This analysis gives a sense of both the relative proportion of books in different languages as well as a comparative sense of visibility.

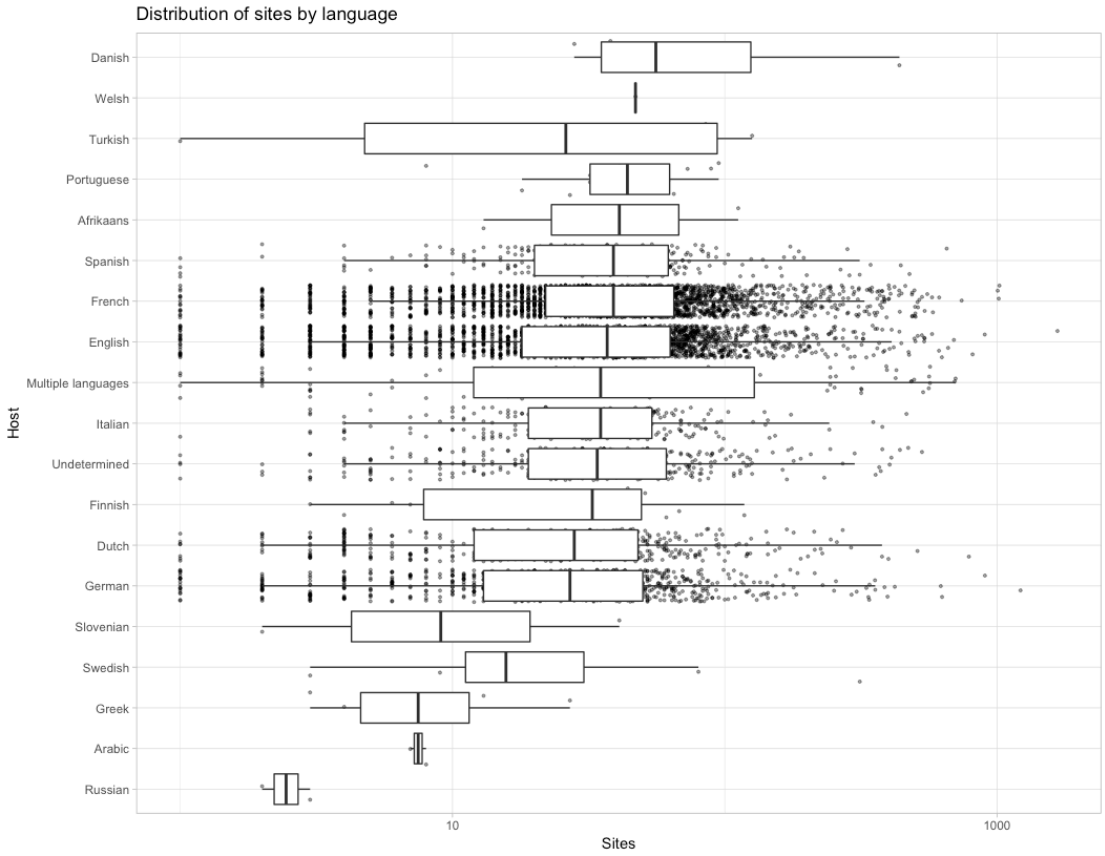


Figure 3. Distribution of web-presence by language of book. Languages are ordered by the mean number of linked websites. For most common languages, the means are within a single standard deviation of each other indicating no statistically significant difference.

In this case we see the dominance of French and English in this corpus (density of points) alongside German, Dutch, Spanish and Italian as other well represented languages. Overall we see no strong or significant difference between the web visibility of these books based on language. While a bias towards English might be expected this does not seem to be the case. This is at least in part due to the strong focus on French (and other non-english) language books by OpenEdition Books.

A different form of analysis is to look at Top Level Domain (i.e. country codes) in URLs referring to these books by the language of the book. This provides an interesting insight at an aggregate level as to the interest in books from different countries in different languages. Here we show the most represented language of book for each country top level domain. This reveals a logical pattern with Latin America showing a preference for Spanish books, with the exception of Suriname (Dutch, the official language), French Guiana and Brazil (French). Francophone and Anglophone Africa are quite clearly distinct and East Timor shows the expected preference for Portuguese. France, the Netherlands, Germany and Italy

all show a preference for their native language. There are apparently unexpected results which deserve more analysis on a larger corpus. Spain, Portugal, and Brazil all show a preference for French which is mostly likely due to the limited presence of Portuguese books in this corpus.

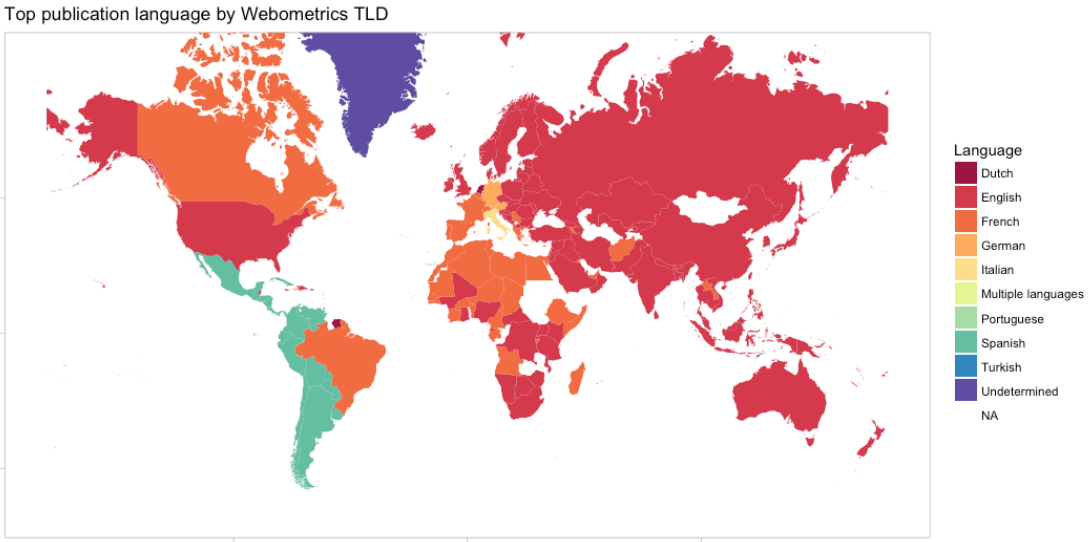


Figure 4. Top Publication Language by top-level domain. For each country code (e.g. '.uk.') the most visible book (the one referenced by the most search results) was identified and its language identified. Latin America has a higher visibility of spanish-language books and francophone and anglophone Africa are clearly visible.

3. Visibility in General Scholarly Information Workflows

To examine the visibility of OPERAS partner books in general scholarly workflows we examined a number of sources of information. The first of these is the oaDOI service that provides information on open access status of objects identified by Crossref DOIs. This service is being deployed in a range of library systems and within Web of Knowledge by Clarivate - so accurate information on open access books is of value.

The second source of visibility data was Altmetric.com, which provides data on mainstream and social media activity for scholarly works. Finally we searched the ORCID public data dump for 2017 for the presence of DOIs and ISBNs associated with OPERAS partner books. These would in most cases have been added by the authors to their profiles.

In all three cases we saw extremely poor visibility. Of the 636 DOIs that were available for this analysis within the OPERAS corpus only 41 were returned as Open Access by the oaDOI service. Only 31 were present in the ORCID data dump. The oaDOI service is limited to providing information on DOIs, which is only relevant for ~10% of the corpus, but the reasons for the poor results merit further investigation. It is likely to be a combination of a service that is focussed on journal articles and the general variability in quality of metadata

provided by OPERAS partners.

Only 160 ISBNs were identified in the ORCID data dump suggesting that overall there is little encouragement from either publishers, platforms or author's institutions to include information on book-length works in ORCID profiles. This may also represent a lack of support for the automated ingestion of book metadata to ORCID, which in turn would need to be supported by more consistent and complete metadata streams from publishers or platforms.

The data obtained from the Altmetric.com service is more interesting and also more informative. Nearly 1000 of the OPERAS books show some form of activity tracked by Altmetric.com, either mainstream or social media. The vast majority of these are on the OAPEN platform with a further contribution from OBP and OpenEdition Books. The dominance of OAPEN is possibly related to the presence of <meta> tags on OAPEN records.[5] Another 304 books are registered in the service but show no activity, again dominated by books from OAPEN followed by OBP. These are stub records that have been created for institutional customers of the Altmetric.com service where book authors are affiliated.

The Altmetric.com service was originally targeted at journal articles, with one primary location online at the publisher website. A large part of its value offering is a high quality aggregation of online references to articles that is achieved by tracking all the relevant URLs that refer to an article, rather than just DOIs as is common for some other services. This is much more challenging for books that often reside at multiple locations. Therefore the service works to actively track and aggregate URLs relevant to books that are of interest, particularly those published by authors based at institutions that are Altmetric.com service.

This is important because it illustrates how engagement with a downstream service can help motivate the gathering of relevant metadata to improve data aggregation and analysis. More generally it shows how the provision of good metadata, in this case a curated list of all the URLs where a book might be found, can prime a service to collect higher quality data. It is important to note that the responsibility for providing this kind of data, does not currently belong to anyone in the supply chain. Making a community decision about where to locate that responsibility and how partners might provide data is a role that OPERAS might take.

C. Findings

The metadata held and managed by OPERAS partners is inconsistent and variable in quality. Collecting and aggregating data from multiple OPERAS partners was a challenge due to inconsistency in bibliographic metadata processes and formats. Several partners were not explicitly included in the analysis because separate data was not available, and

[5] Euan Adie, Altmetric.com, personal communication

some analysis is limited by issues with the data provided. This includes ISBNs that appear to be incorrect.

These data quality issues create a number of downstream challenges. Firstly analysis is more challenging and involves more manual work, raising the cost and limiting the generalisability of findings. Secondly it creates a relative lack of interest amongst downstream data aggregators and providers in collecting data relating to books. Books offer particular challenges and the market remains focussed on journal articles. Nonetheless as we note below, there is interest in handling books better, which would be encouraged by the provision of more consistent and complete metadata.

The visibility of OPERAS partner books in catalogues varies by publisher. OPERAS partners have clearly focussed on different catalogues to optimise the visibility of their content. Given the heterogeneity of OPERAS partners this is not surprise. It is also evidence of a lack of crosstalk between catalogues. Again, the provision of standardised bibliographic metadata could aid both small and large publisher and platforms in gaining more visibility across all the relevant catalogues.

Evidence can be obtained that books relevant to specific regions gain interest and attention in that region. On aggregate we have shown evidence from the analysis of country top level domains that books are often more discussed and written about in countries where the language of the book is common. We have previously shown how web visibility and country-level usage analysis can demonstrate local usage of single books. This new analysis shows that similar information can be gained at a corpus level.

While we did not see an obvious visibility bias for languages that appear frequently in the OPERAS corpus, it may be the case that rarer languages do see a bias. It may also be the case that the lack of bias is due to strong representation of French work by OpenEdition Books. We did see less visibility for books in Greek, Arabic and Russian (i.e. in different scripts) however the small numbers here limit any statistical conclusions.

The variable quality of book metadata creates challenges in analysing visibility consistently. Throughout this analysis we have had challenges in comparing like with like due to the differences in metadata completeness and quality. Similarly this will create challenges within individual partners seeking to do similar analyses. Finding ways to maintain, use and deliver high quality metadata at low cost, probably through the development of shared platforms, offers multiple benefits for OPERAS partners including better internal information, greater ease in tracking and better engagement with downstream collectors and analysts of data.

The variable quality of book metadata creates challenges for downstream data aggregation and analysis providers. In discussion with a series of downstream data providers including oaDOI and Altmetric.com the issues of tracking information for books was raised. These downstream providers are aware that of limitations in their data collection

for books and have an interest in improving quality and completeness of the data they collect. In most cases they currently appear to be limited to manually updating data based on direct interactions with customers.
In general there is a question for those engaged in the production of books and open access books in particular as to who they want to design and implement solutions. By default the sector will get systems focussed on journal articles and STEM output processes. There is interest in engaging, but without a concerted effort from the providers of book content this is unlikely to be well integrated with book production.

VI. Digital Visibility Challenges and Opportunities for OPERAS Partners

The promise of Open Access scholarly monographs is multi-faceted. First it provides easier and more efficient access to scholarly work for scholars. Secondly it offers access to previously expensive content to broader communities of interest who either do not have access to, or would not think to use, an academic library. In particular the free distribution of content online offers to bring together communities of interest around a specific topic. These communities may be small as well as diverse and geographically distributed. Their engagement with, and ultimately their input into scholarship has the potential to strengthen public support and enrich and diversify its impact.
To achieve this promise it is not sufficient that open access monographs be available, they must also be visible and also accessible to these diverse audiences. OPERAS partners, funders, platforms, and publishers are already delivering on the issue of availability. Here we address the question of visibility. As has been discussed visibility is a complex issue. Visible to who? Under what circumstances? After what kinds of search? Mapping all the possible discovery pathways is a future challenge.
In this work we have taken a deliberately narrow scope. We start with the assumption that high quality and consistent bibliographic metadata at source is key to enabling the wide range of services and systems that will support discovery and visibility in diverse contexts. Our focus in these recommendations and issues is therefore on the way in which consistent metadata provision and dissemination through common channels provides a route towards visibility.

A. Challenge - The quality and consistency of OPERAS Partner metadata is variable

An early finding of the work package and consistent throughout the survey, the provided metadata, and the completeness of records in third party systems was variability in both the

format, completeness, and quality of metadata. In the survey there was qualitative evidence of differing degrees of concern and interest with specific issues, relevant to specific presses and platforms. In the metadata provided there were substantial inconsistencies in format, completeness and validity. For instance the small but significant presence of identifiers that were invalid (51 ISBNs that did not validate) was an issue.

Further downstream in the data and discovery process there was clear evidence of a lack of consistency in metadata delivery. As will be discussed below this at least in part a result of diversity in the mission and goals of specific OPERAS partners and their capacity to focus on internal metadata systems. It is also a function of existing discovery and metadata systems only recently grappling with the issues of books. However, in a distributed and global information ecosystem the provision of consistent, correct, and high quality metadata is a necessary condition of optimising for visibility and discovery.

B. Challenge - Diversity of gathering, cleaning, reporting usage data across OPERAS partners makes comparison difficult

Usage data was a focus of the survey work and previous work by KU Research has focussed on usage data collected by the OPERAS partner UCL Press[6] as well as for four presses using the JSTOR platform.[7] It was not part of the visibility mapping exercise, at least in part because the previous work and survey showed that a comparison is not feasible.

OPERAS Partners that host content collect data differently, clean that data differently, and report it differently. Even where a standard protocol is used, for instance where data is referred to as “COUNTER Compliant” or “COUNTER Protocol” there is evidence of substantial differences in collection, management, exclusions and reporting. In some cases this relates to differences in the definition of access status and in some just in differences in technical systems.

Details of internal operations tend to be sensitive as is the release of data, particularly where it is likely to be used for comparisons. Data quality issues currently mean that any comparison is likely invalid, but equally without an increase in transparency for data collection and reporting the development of best practice is unlikely. Legal, ethical and trust issues are also a significant challenge (see below).

In particular the small scale of many OPERAS partners means that they will not have the capacity to develop their own in-house expertise and systems. Adoption of good practice to generate high quality data will depend on sharing the burden of capacity building in some way. That in turn, cannot happen until there is a framework that provides sufficient trust to

[6] http://dx.doi.org/10.17613/M6H49K

[7] http://kuresearch.org/PDF/jstor_report.pdf

allow the sharing and comparison of data and its management.

C. Challenge - Application of existing systems is not always straightforward for books

Existing systems for digital and online research discovery and distribution have been largely built with journal articles in mind. The implicit assumption of a single Version of Record, hosted on a publisher-controlled website, that only rarely goes through any change is built into metadata creation, identifier systems, discovery and distribution channels. The dominant means of delivery for journal articles is now online with print a niche provision in many disciplines. In contrast for books, print still remains the focus for many publishers and the engagement with online and digital supply chains reflects that.

The confusion and inconsistency in coining and distributing Crossref DOIs and ISBNs is one example of this. Even though the set of OPERAS partners are focussed on online and digital as open-access focussed providers, there is confusion and inconsistency in the use of identifiers. Partner-provided metadata files referred to many different types or 'versions' of DOIs and ISBNs ('electronic', 'online', 'print', different file formats, platforms), in addition to the inconsistent provision of DOIs at the chapter level.

As noted elsewhere the scale of OPERAS partners and book providers in general means that the technical capacity is not necessarily available internally to engage with these issues and systems. In addition, as small players, OPERAS partners and others often do not have the levels of staff capacity to engage directly in community efforts to develop greater consistency in data practices.

The lack of applicability to books also plays out downstream. Systems such as Altmetric.com are able to exploit the (generally) single and predictable online location of journal articles to connect Crossref DOIs to URLs and aggregate mentions. For books Altmetric.com needs to undertake this work in a manual and directed fashion because there is no straightforward way to discover all the locations of a book online, and therefore to understand when social or mainstream media is linking to a copy. This challenge is also exacerbated by inconsistent practice and quality of metadata provided by publishers and platforms.

It is worth noting however that journal articles will start to face some of the same issues as green open access increases alongside preprint adoption. OPERAS partners could take a lead on developing best practice for identifying multiple locations online and take a leadership role in supporting the next generation of discovery and identifier infrastructures.

D. Challenge - Diversity of approaches, goals and definitions creates challenges for developing common platforms

As we have noted in several places in this report there is enormous diversity in the missions,

goals, and activities that different OPERAS partners undertake, even those that might be categorised together as “publishers” or as “platforms”. This plays out in many ways, in the different assumptions that various partners bring to engaging with external platforms, but also in the needs for reporting and the strategic goals that drive decision making.

One example of this is the different definitions of what constitutes “open access” amongst various OPERAS partners. OpenEdition Books and Open Book Publishers offer a set of freemium offerings where some formats of the book are free but others are charged for. Others deliver only one freely accessible online format. At the same time demonstrating the use of online content appears important for most partners. This leads to a situation where usage data is sensitive and potentially competitive but also not readily comparable.

In the longer term it will become necessary to address questions as to whether formats for screen reading (some of which may have restricted functionality) are more “visible” than epub and fully downloadable PDF, and how digital visibility relates to print sales. The diversity of OPERAS partners is a strength in providing offerings for different parts of the scholarly community. It will also be a challenge in divining how the investment in visibility supports different communities. The small scale and competitive nature of OPERAS partners means that finding ways to share information and best practice will be critical. The diversity of goals, funding streams and contexts will be a challenge in delivering that.

E. Challenge - A lack of engagement with data governance and ethics runs the risk of creating problems

While not a technical issue, the issue of data governance appears a substantial risk for OPERAS partners in two areas. Firstly there is significant variability in awareness of the implications of handling and analysing user logs. While some partners use Piwik as a local tool to collect logs many use Google Analytics. While Google Analytics (and other Google services) will presumably meet the standards being introduced under the General Data Protection Regulation in Europe there is a growing sense that they don’t meet the ethical expectations of the scholarly community.

Survey answers and parallel work in the HIRMEOS project suggests to us that while some partners are sensitive to these issues the majority are not. Further, it is not clear that the technical capacity exists to properly address issues of privacy that arise as the desire for more granular information on usage and visibility grows. Future work should address the legal liability issues that arise from holding such logs and the forms of analysis, data sharing, and data retention that are appropriate for our community.

A related issue is that of governance frameworks for data sharing. If the goal of OPERAS network is to support shared best practice and capacity building, then this will necessarily involve data transparency and sharing. As noted, usage data in particular can be highly sensitive, in addition to implicating privacy regulations. Building a framework in which trusted

parties can benefit from data and tool sharing will be crucial for achieving the goals of the OPERAS network.

F. Opportunity - OPERAS can act as a growing network for best practice and capacity building

A theme with many of the challenges is that of coordination and sharing the burden of developing technology and best practice. That in turn is a substantial opportunity for OPERAS to develop a network which can support partners in sharing the development and implementation of best practice. The ongoing growth of the OPERAS network is a positive sign in this sense.

OPERAS could benefit from building its own capacity to act as a hub for initiatives or even to act as a node for the coordination of resources. While it's current role as a focus for grant funded activities is a good step in this direction building up a long term capacity to deliver value for partners will support sustainability of the network as well as providing a focus for future activities.

The diversity of partners within OPERAS means that there already is both knowledge and existing best practice that could be shared from within the network. Building internal trust will be important, and this suggests that some of the issues raised above on governance arrangements should be tackled early. This will also need to develop a global focus to include other key players. If successful, OPERAS could play a key role in ensuring a continuing diversity of scholarly book publishing organisations in contrast to the continuing concentration of journal publishing and the issues that that brings.

G. Opportunity - Downstream suppliers and aggregators of data will respond positively to better and more consistent metadata provision

While we have focussed on the inconsistency of metadata provided by OPERAS partners, the deficiencies of downstream systems in handling books, and the consequent gap, we have also seen a desire to engage and improve these systems. In particular downstream systems face challenges in connecting identifiers to a complete set of online locations (URLs) and clarity on the use of metadata to signal access state and other issues.

If practice can be systematized and the overall quality of metadata improved, there are therefore significant opportunities to improve the visibility of open access books in these systems. There is also an opportunity to engage with these systems to ensure that the interests of OPERAS partners are served in implementation decisions that will need to be made.

There are unresolved questions of where in the supply and distribution system the responsibility for creating, managing, and distributing metadata lies. As noted elsewhere these decisions were largely made by default in the journal article system. For books with the complex relationships between publishers, aggregators, platforms and discovery tools these responsibilities are less clear. Who should register DOIs? Who is responsible for maintaining landing pages? For different editions and versions? How can multiple competing platforms work together to enable discovery? While the answers to these questions are beyond the scope of this report, working to resolve them is an opportunity for OPERAS to take a leadership role as well as to maximise the visibility and usage of OPERAS network books in ways that are appropriate and suitable for OPERAS partners.

VII. Appendix A - Survey Questions

Tracking the Uses of Open Access Books

As part of OPERAS-D Work Package 3.2 the Knowledge Unlatched Research team are gathering information about how OPERAS partners gather, manage and engage with usage data relating to open access books.

In order to ensure that we capture perspectives of OPERAS network members we would be grateful if you could spend a few minutes completing this survey.

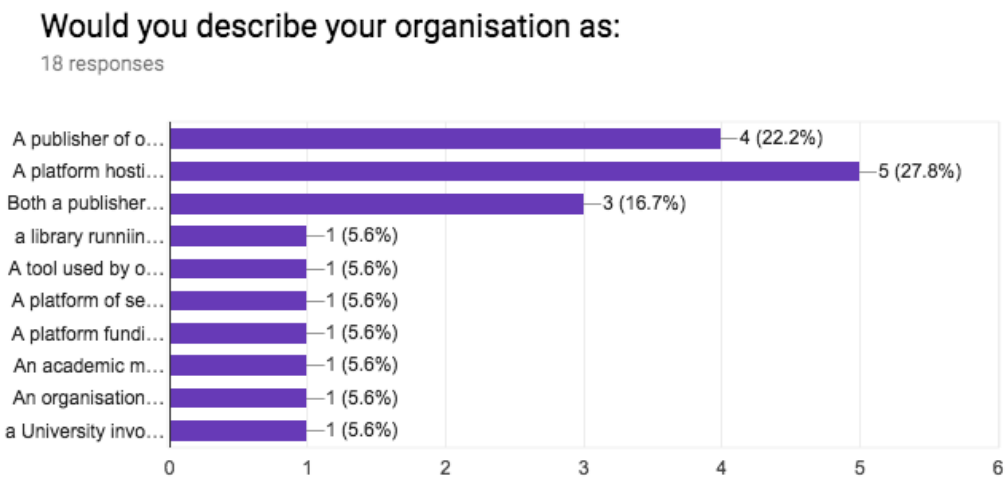
1. Email address *
2. Would you describe your organisation as: *Check all that apply.*

A publisher of open access monographs
A platform hosting open access monographs published by others
Both a publisher of open access monographs and a hosting platform
Other:

3. What kind of data about how your open access books are being used is available to your organisation?
4. Do you have a process for gathering and managing usage data relating to your OA books? If yes please describe it to us.
5. Who uses data about the usage of open access books within your organisation?
6. Does usage data impact on decisions made by editorial, marketing or other departments? If so, how?
7. What are the biggest challenges for your organisation when it comes to dealing with usage data about OA books?
8. What kinds of tools or services would help your organisation to engage more effectively with usage data?

VIII. Appendix B - Survey Responses

2. Would you describe your organisation as:



- A publisher of open access monographs (4)
- A platform hosting open access monographs published by others (5)
- Both a publisher of open access monographs and a hosting platform (3)
- a library running a press that publishes OA monographs
- A tool used by open access (and other types) publishers and platforms
- A platform of search allowing the access to digital data in SSH
- A platform funding open access monographs published by others
- An academic marketplace for journals that publish articles in open access.
- An organisation flipping subscription journals to Fair Open Access
- A publisher of open access monographs, a University involved in promoting Open Access

3. What kind of data about how your open access books are being used is available to your organisation?

- online editions & downloads: weblogs - with analytics (GA, COUNTER). We collect usage data from several other third party platforms - incl Google Books, OpenEdition Books, WorldReader, ...
- Download statistics, access log
- Piwik analysis, usage data supplied by repository software (DSpace), CrossRef analytics for our DOI, OAPEN usage statistics
- Download data, website usage
- We collect data on what is annotated. We've not looked into it, but we could run domain based queries on the platforms. Not sure if we could look by publisher, but maybe...
- We provide data about OA usage of our books to our organization. It is a crucial demonstration to them of the value of UCL Press
- Counter compliant downloads
- Available to Coimbra University Press (CUP) are mainly the data about the number of visualizations and downloads of the monographs, the rate of broken accesses and books accepted in indexing databases.
- We are using data generated by Piwik.
- Full-text PDF downloads by institution, PDF downloads by region (geolocation), views
- Not applicable
- For its books, OpenEdition Books (OE) gathers information both on access and usage.
    - Access metrics:
        - nb unique visitors (distinct IPs)
        - nb of views (distinct sessions)
        - nb of page views (distinct pages)
    - Usage metrics:
        - views/downloads of chapters (sorted by books, publisher's collection, authors, referrers)
    - COUNTER metrics:
        - BR1: books (PDF and epub downloads only)
        - BR2: chapters
        - BR3: books or chapters (unauthorized access)
- The books of the Presses universitaires de Liège in open access are available on OpenEdition Books Freemium. Data are coming from this platform.
- We are a publisher of monographs and use both an external repository (http://rcin.org.pl/ibl/dlibra) to deposit books and journals, as well as our own platform (http://nplp.pl/) for extended monographs.
- Available tools:
    - Repository: Google Analytics, WebLog Expert number of downloads
    - NPLP: Google Analytics
    - Alas only stats available for repository are for all books (incl. other institutions), so we are not able to monitor the usage of our books only,
        - Data: (Standard GA data): visits, unique users, bots, pages displayed (all data for different time periods: daily, weekly, monthly)
        - User access data, monograph views, monograph downloads, Google Analytics
- Available data is related to loading books on the platform, to download users, to visualizing the series main page, the monograph abstract page and press main page

4. Do you have a process for gathering and managing usage data relating to your OA books? If yes please describe it to us.

- This the heart of HIRMEOS WP6 - we have drivers collecting data from all third party sites. We store in database which we query for specific usage questions. Which data is aggregated depends on the question.
- The download statistics are collected by means of a wordpress plugin.
- not really yet, we do ad hoc analysis if requested by authors
- Download data is sent to IRUS-UK who create COUNTER compliant data. Based on that data, OAPEN creates reports for publishers.
- Not at the moment.
- We gather data from our institutional repository manually and put it into a spreadsheet (weekly). We are sent data from other platforms that host our books and we enter that manually into the spreadsheet (varies from twice a year to monthly)
- Yes, gathering through third party IRUS-UK, and creating periodic usage reports for customers
- Usage data is provided by page view count, Crossref DOI and Sushi Counter protocol.
- We don't manage usage data much at the moment. We will soon change our whole system (perspectivia.net) and use MyCoRe. After the final migration of perspectivia.net, we will see how to manage usage data and if we need to develop a process for managing usage data.
- Yes, (as you know ;))by working together with OAPEN we receive quarterly reports on the usage data for the titles we fund. We analyse the quarterly reports and upload this as spreadsheet to our platform to inform our customers about usage. However, this is a rather manual process and would benefit significantly from more automation.
- N.a.
- OE metrics are based on Awstats results with further processing made by PHP scripts.
- Access logs contain the raw access data
- Awstats delivers the access metrics based on the logs
- PHP scripts calculate and produce the usage metrics from access metrics
- PHP scripts calculate and produce COUNTER metrics (BR1, BR2, BR3)
- COUNTER metrics are at the moment produced according to COUNTER V4; COUNTER V5 will be implemented on January 1st 2018.
- Not especialy
- We do not monitor the usage regularly, but we would like to develop procedures to do so.
- EKT has developed an online application that gathers and manages usage data
- Within the platform there’s a tool for importing /exporting data in different format (onyx, xml etc.), for creating usage statistics and custom reports

5. Who uses data about the usage of open access books within your organisation?

- Displayed on our website. Regular reports made to authors and to libraries.
- The download statistics are publicly displayed after each item
- all five staff members
- Publishers that are members of OAPEN
- N/A
- It is used by a number of different departments such as Research and Global Engagement, to demonstrate the reach and impact of the open access books we publish.
- All
- We are a section of the University of Coimbra and the main usage of the data that is collected and gathered is made by the Administration of the University and some specific Investigation Centers that are responsible for that analysis and data management.
- The editorial staff perspectivia.net and partly the scientific editorial boards of our institutes.
- Sales & marketing team
- N.a.
- Usage metrics are used primarily by the publishers, who have specific access to the display page and its searching features.
- Usage metrics exposed through COUNTER are also used by the libraries who subscribed one of the Freemium OE's offer.
- Within OE, usage metrics are used by:
    - the OE Books team
    - the IT team
- the management team
- Me as director of the Presses universitaires
- Director of the Institution and publisher management
- Service owners and collaborating publishers (upon request)
- The staff working at open access publishing

6. Does usage data impact on decisions made by editorial, marketing or other departments? If so, how?

- No (3)
- Marketing - we monitor sources of traffic, responses to marketing activities etc. Reports back to authors etc and in general 'marketing' of impact of OA publication process.
- we take it into account and plan to identify some KPIs, but as the press' rationale is serving Göttingen Campus with OA publishing options, usage data plays up to now only a minor role in the sense of understanding how well the service works overall, to gain narratives out of unexpected success stories or analyse the reach of our platforms
- N/A (not a publisher)
- It doesn't. The books we publish are selected on the basis of scholarly merit. Our marketing strategy is undertaken based on our assessment of the potential size of the audience. Our dissemination strategy is affected by usage data - we choose where to have our OA books hosted depending on the level of readership that we see.
- Indirectly. They are part of our performance as platform, and of our service to customers
- Yes, it has impact on CUP's editorial team. These data are important because they show us the impact and propagation of the several thematic areas of our contents. They allow us to define new strategic paths and directions, either for the contents that are being more distributed because of the open access politics, or for the ones that started to have more exposure and impact after the same open access benefits. At the same time that information gives us a method of action regarding the financial investments that could be taken or followed for the future.
- No, it doesn't impact any decisions.
- Yes, we are using it for the targeting of our sales and marketing to institutions. It supports the fundraising process of our titles.
- N.a.
- No specific information available on this aspect.
- To promote open access for new authors who are sometimes afraid of open access, I use data to explain to them the benefits in terms of visibility that open access offers.
- No. Decisions are now based on print circulation, or number of e-books sold through commercial platforms.
- Yes, It does. We are a Not-For-Profit Open Access Publisher. Usage data are important for us to assess the value of a collection and, consequently, to develop our promotional strategy to reach the largest number of readers.

7. What are the biggest challenges for your organisation when it comes to dealing with usage data about OA books?

- Accessing usage data from numerous platforms where the book is available. Aggregating usage data statistics collected in different ways by different platforms.
- Collecting and keeping data according the FAIR principles.
- I. So far, we lack a robust -- normalised, cleaned from robots, based on COUNTER -- data basis that would a) compile usage data from different sources, b) allow implementation of dashboards or data warehousing efforts and c) justify momentuous management decisions based on such data. II. Interaction with Google Books maintains to be a challenge, as Google repeatedly changes their interaction and access parameters. Although the data from Google is very valuable, we can't really use it as we often have data gaps and not enough time to close them let alone track why they happened in the first place.
- Optimizing workflow, how to do more work with small resources
- N/A No difference from other annotation data.
- Manual collation of stats that is only increasing with the number of books we publish and the length of time we have been in existence.
    a. Disseminating metadata to other platforms to increase usage
    b. Aggregating data from various platforms
    c. Technological/financial: to be able to provide usage data automated, online, in real time
- To decide which areas or collections should have a faster and direct intervention when it comes to display those contents in open access, and manage that effort with the need and vision of the financial investment that should embrace the whole process. To improve the quality and the external legibility and interconnectivity of the data provided.
- Since we (almost exclusively) host publications of our institutes abroad, we need to rise awareness why and how usage data is important to them and their researchers. Analyzing usage data is difficult and can easily lead to wrong assumptions about the impact of a OA book. In our case this could be detrimental to our institutes, which tend to compare their "success" to the other institutes. This means that we clearly need to understand what the usage data is telling us before we have any use for it. We also face strong restrictions concerning data security as we are a public entity (part of the German ministry for education).
- Different usage reporting formatting across hosting partners (JSTOR, OAPEN and Hathitrust all report different). Then it is also a challenge to obtain all this information, which consumes a lot of time right now. Finally, generating unified reports is done manual, which is not scalable.
- Regarding the books, the main challenge is the paradox of finding, on one hand, a measure that is standardized enough to allow for comparison and, on the other hand, a measure that is specific enough to provide useful information to our different partners. Therefore, alongside with the COUNTER metrics which give a good standardized measure mainly for libraries, OE has provided a usage metrics interface which relies on the documentary unit and its publication environment (authors, collections, etc.).
- We are a small structure. Data about OA books are only used in an empirical way.
- The construction of our repository's reporting system which makes it difficult to bowse daa about our content.
- No particular challenges have been identified yet
- It is a big challenge for our organisation to aggregate usage data about Open Access Books published under a Creative Commons Attribution Licence 4.0 and actively disseminated on different online platforms, such as our website, the institutional and disciplinary repositories, the social media etc.

8. What kinds of tools or services would help your organisation to engage more effectively with usage data?

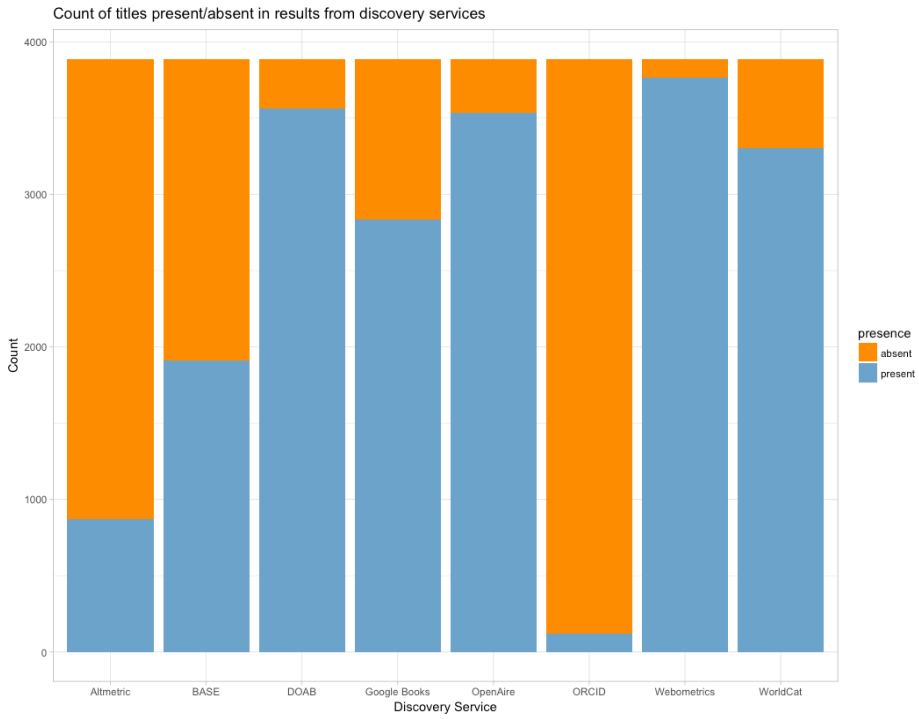
- A range of visualisation tools that query the underlying database
- We mean usage data as a service to the public: every (open source) tool enabling us to offer a richer and more informative representation of them would be welcome.
- To I. Import scripts to gather usage data from different sources via API, dashboard-like applications as in Piwik analysis to allow customised reports and chron jobs. To II. A consortium agreement with Google on how to gather and access usage data. Maybe this could be a smaller funded project.
- Automated reporting tools, including publishing results online.
- We use Metabase currently.
- We are currently working with KU Research to customize Tableau to provide an automated ingest tool and a dashboard to help us view the data in a simpler way, and to analyse it in new ways to better understand reader behavior.
- Aggregating from different platforms, displaying data (dashboard), automated reference extraction from pdf files (for submission to CrossRef)
- Although we know that some services can provide the usage data we want, such as Google Analytics or similar but that we still don't/can't use, what we really wish is a kind of data that are relevant respecting the origin/provenance/country of the visualizations and downloads of our contents. It would be of great help if we could have a main service from where we could manage all the information related to statistical usage data. It would also be extremely useful to be able to compare our data with those of other publishers, especially academic presses, and make a clearer contribution to typical institutional processes, like assessments/rankings and the displaying of big data.
- Right now we can only gather our isolated usage data. It would be great to be able to compare this with other usage data on similar topics or from similar platforms, maybe even from libraries.
- We would like to see an usage aggregation service that consolidates usage data from different hosting partners into one standardised report in an automated way. In turn, this should translate into an usage dashboard that can be embedded into platforms and allows customers to use different filters to analyse usage by publisher, region, etc.
- N.a.
- The existing services provided by OE seem for now sufficient to deliver detailed and accurate information on the metrics. The main challenge is the first level of raw data gathering: the tool used to collect connection metrics has to be able to distinguish thoroughly human connections from robots' connections. Awstats appears to be weaker than Piwik from this point of view but the counterpart is that Piwik produces large amounts of data uneasy to process.
- Tools and statistics from OpenEdition Books are very useful.
- A dissemination platform with stable reporting system (that is why we aim to publish our content through OpenEdition Books)
- We would like intelligent tools for automatic information integration of usage data extracted from different sources, such as publisher's website, Open Access repositories, social media platforms.

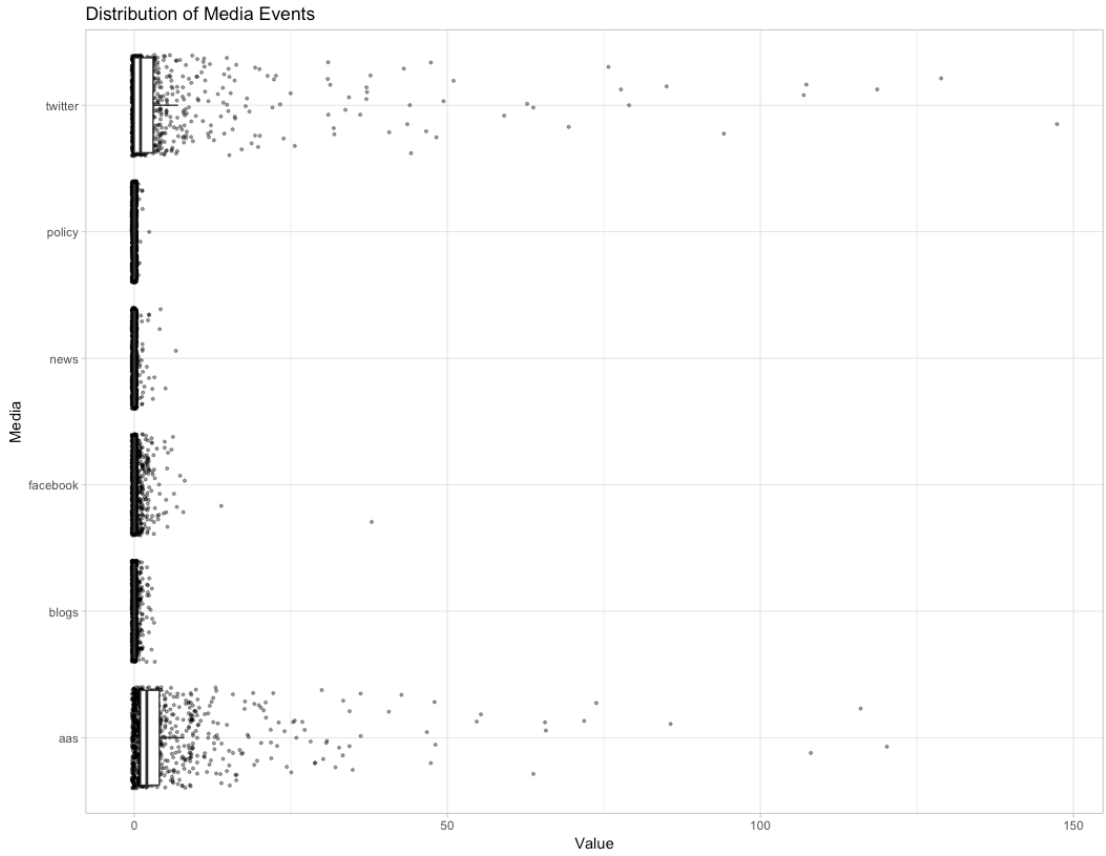
IX. Appendix C - Analysis by platform/publisher

A. OAPEN

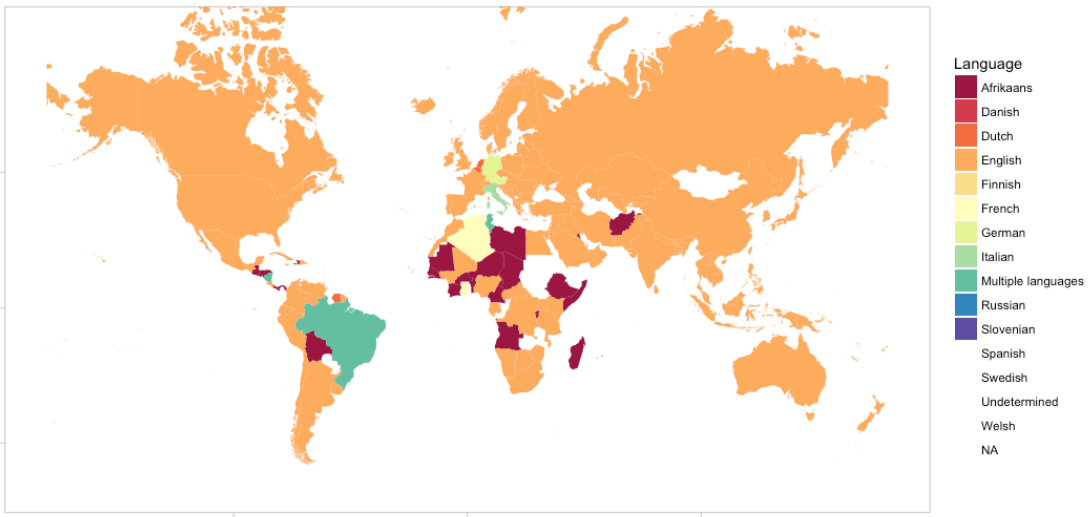
The OAPEN Library contains freely accessible academic books, mainly in the area of humanities and social sciences. OAPEN works with publishers to build a quality controlled collection of open access books, and provides services for publishers, libraries and research funders in the areas of deposit, quality assurance, dissemination, and digital preservation.
Books in the OAPEN Library are available for download in PDF format of the entire book (rather than individual book chapters).
In September 2017 the OAPEN Library metafile contained 3,888 books from 181 publishers - 2,231 of which were in English, 601 in German, 503 in Dutch.

Parameter	Number	%
Number of Books	3,888	100
Books with ISBNs	3,562	92
Books with DOIs	548	14




Distribution of Media Events
twitter
policy
news
facebook
blogs
aas
Media
0
50
100
150
Value

Top publication language by Webometrics TLD

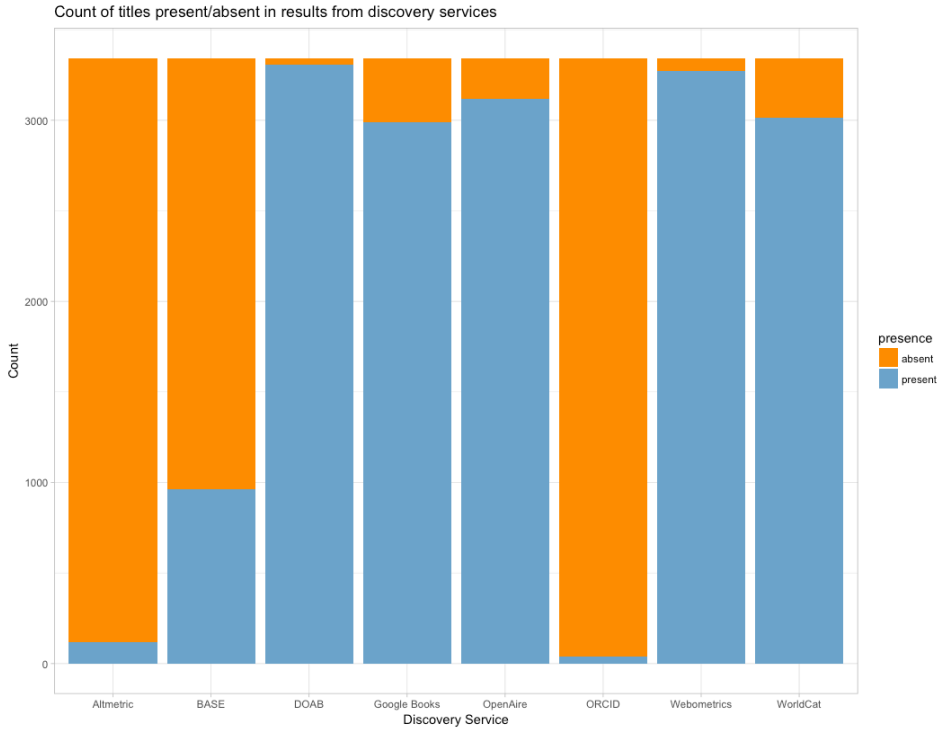


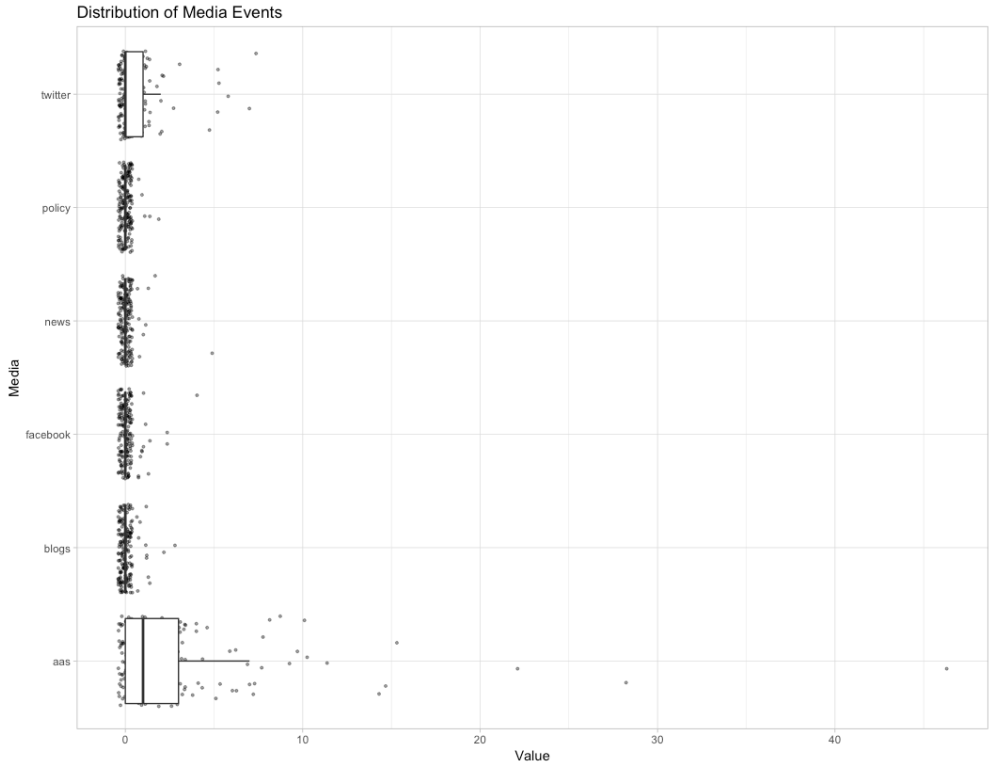
B. OpenEdition Books

OpenEdition is a web platform for books, journals, blogs and events in the humanities and social sciences. OpenEdition Books is run by the Centre for open electronic publishing (Cléo – UMS 3287), a unit that brings together the Centre National de la Recherche Scientifique (CNRS), the université d'Aix-Marseille, the École des Hautes Études en Sciences Sociales (EHESS) and the Université d'Avignon et des Pays de Vaucluse.
In September 2017 OpenEdition Books' metafile included 3,343 books from 69 publishers - including 2,610 in French, 303 in English and 215 in Spanish. More than half of the books hosted by OpenEdition Books are available in Open Access - generally via HTML. OpenEdition Books makes additional services available to libraries and institutions on a subscription basis.

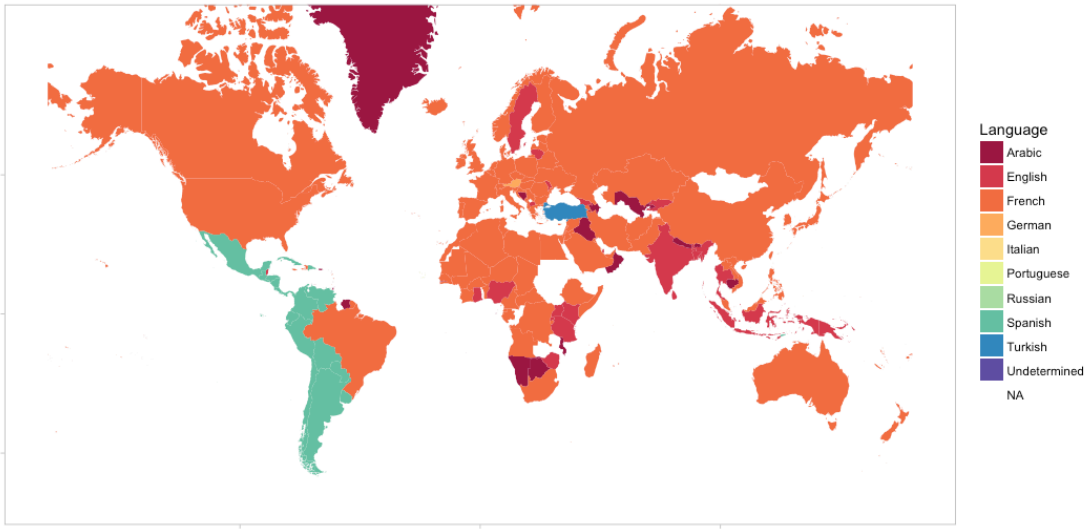
Parameter	Number	%
Number of Books	3,343	100
Books with ISBNs	3,305	99
Books with DOIs	2,945	88[8]

[8] The original dataset from OpenEdition Books did not contain DOIs and that was used for the rest of the analysis. This updated figure was calculated from a new metafile provided by Open Edition Books in January 2018 which included DOIs. The remaining analysis is unchanged.


Count of titles present/absent in results from discovery services
Count
3000
2000
1000
0
Altmetric
BASE
DOAB
Google Books
OpenAire
ORCID
Webometrics
WorldCat
Discovery Service
presence
absent
present


Distribution of Media Events
Media
twitter
policy
news
facebook
blogs
aas
0
10
20
30
40
Value

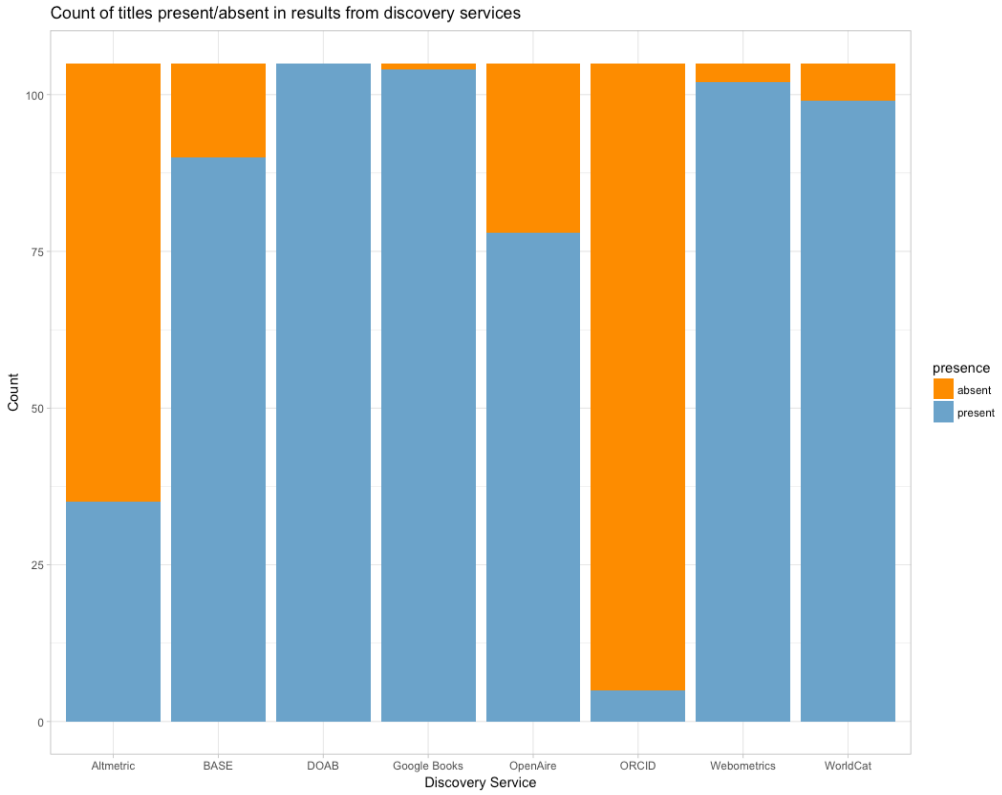
Top publication language by Webometrics TLD

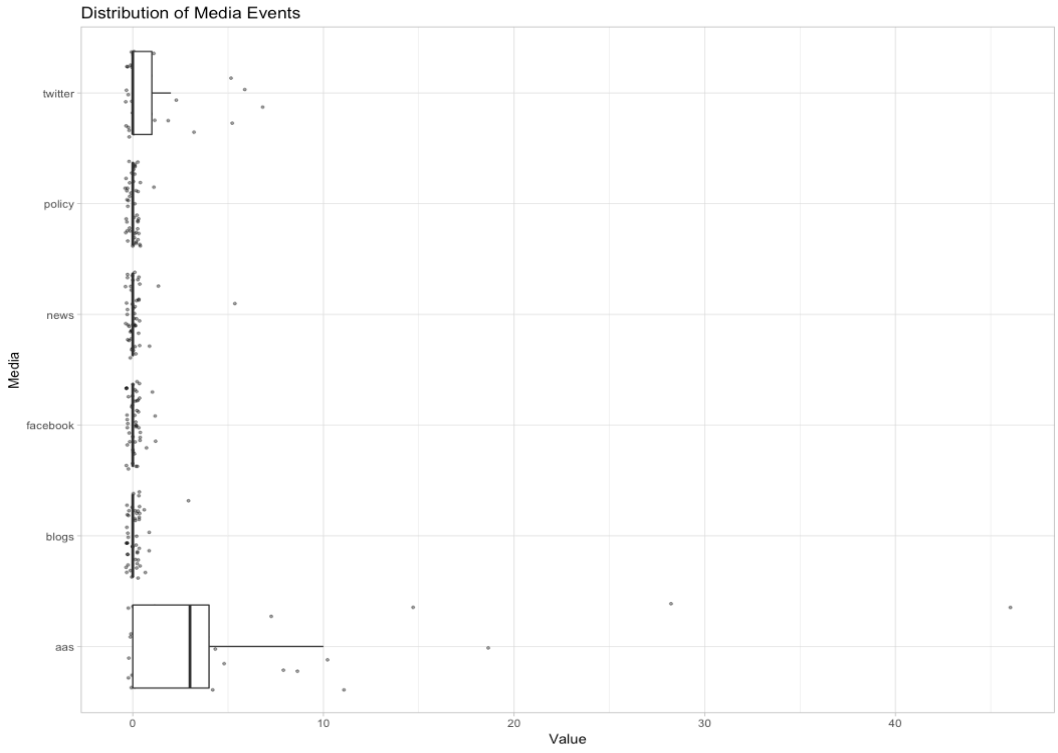


C. Open Book Publishers

Open Book Publishers is an independent publisher of Open Access scholarly books based in the United Kingdom. Open Book Publishers makes books available in hardback, paperback and ebook editions, as well as in Open Access. Some of OBP's books are available for free HTML on-screen reading. Others are available in Open Access as fully downloadable PDFs or ePUBs. Open Book Publishers hosts books via its own servers. Open Book Publishers titles are also hosted by OAPEN, JSTOR and OpenEdition Books.
In September 2017 the Open Book Publishers metafile included 105 books, 103 of which are in English, 2 in French.

Parameter	Number	%
Number of Books	105	100
Books with ISBNs	105	100
Books with DOIs	105	100


Count of titles present/absent in results from discovery services
Count
100
75
50
25
0
Altmetric
BASE
DOAB
Google Books
OpenAire
ORCID
Webometrics
WorldCat
Discovery Service
presence
absent
present

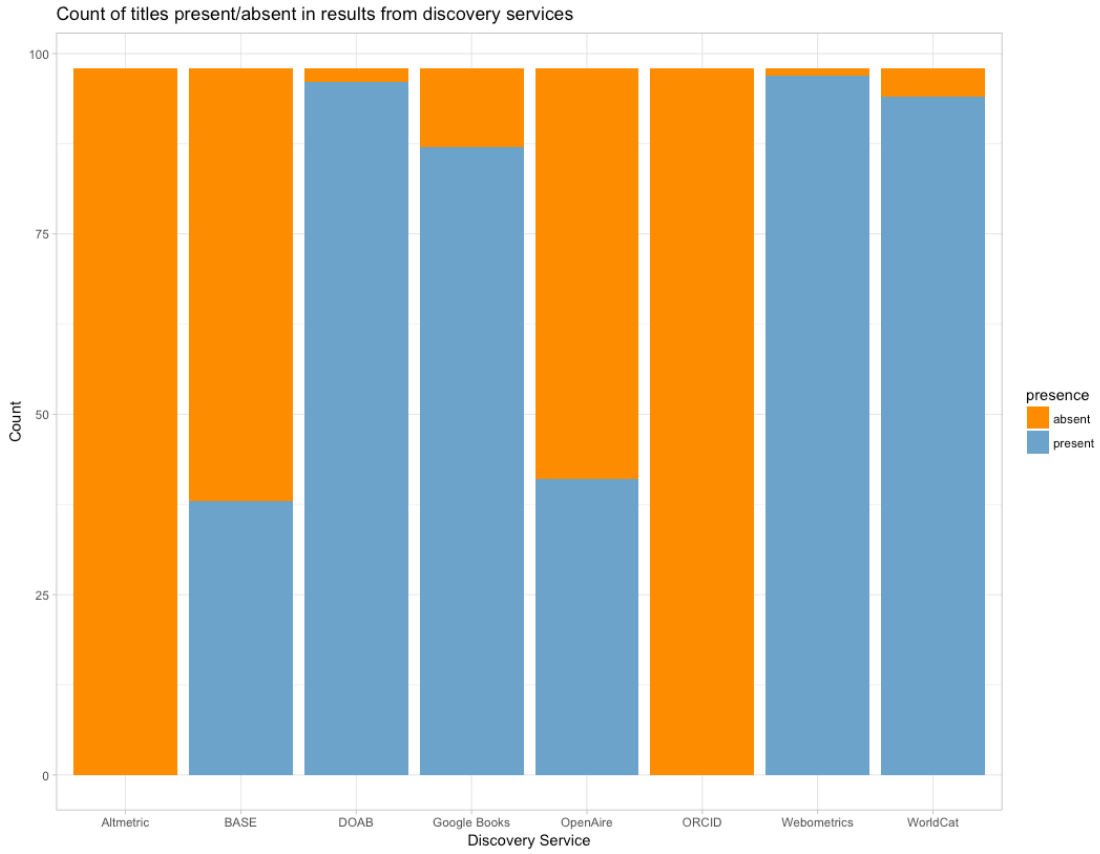

Distribution of Media Events
Media
twitter
policy
news
facebook
blogs
aas
0
10
20
30
40
Value

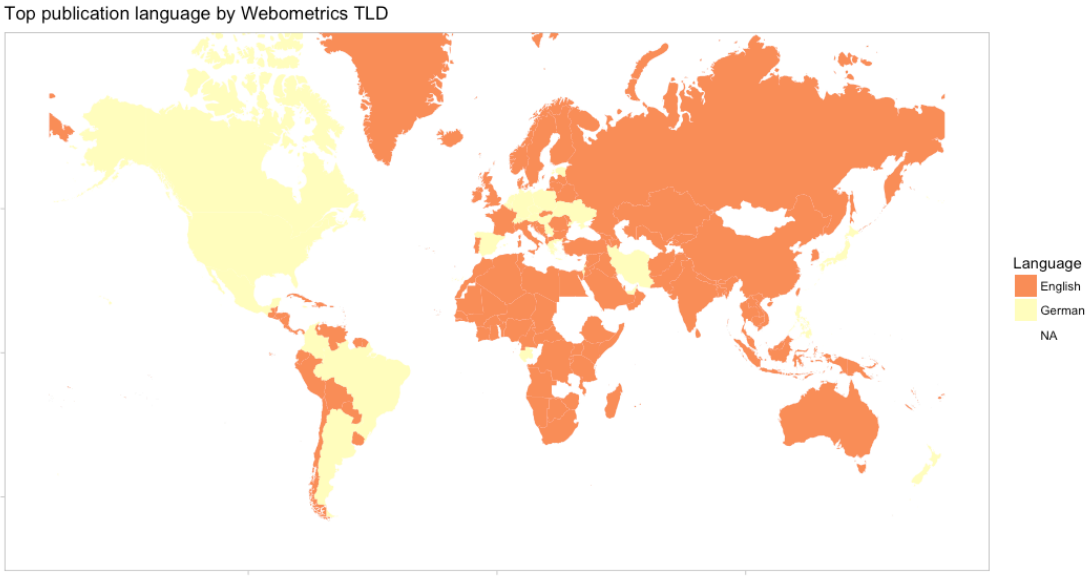
Top publication language by Webometrics TLD

D. Göttingen University Press

Göttingen University Press is the publishing house of Göttingen University and has published scholarly texts by researchers affiliated with the university since 2003. The Press is strongly committed to Open Access publishing and makes use of Göttingen University's Open Access DSpace Archive. In addition to Open Access publishing services, Göttingen University Press also makes titles available in print-on-demand formats.
In August 2017 the Göttingen University Press metafile included 98 books, 70 of which were in German and 25 of which were in English.

Parameter	Number	%
Number of Books	98	100
Books with ISBNs	96	98
Books with DOIs	94	96

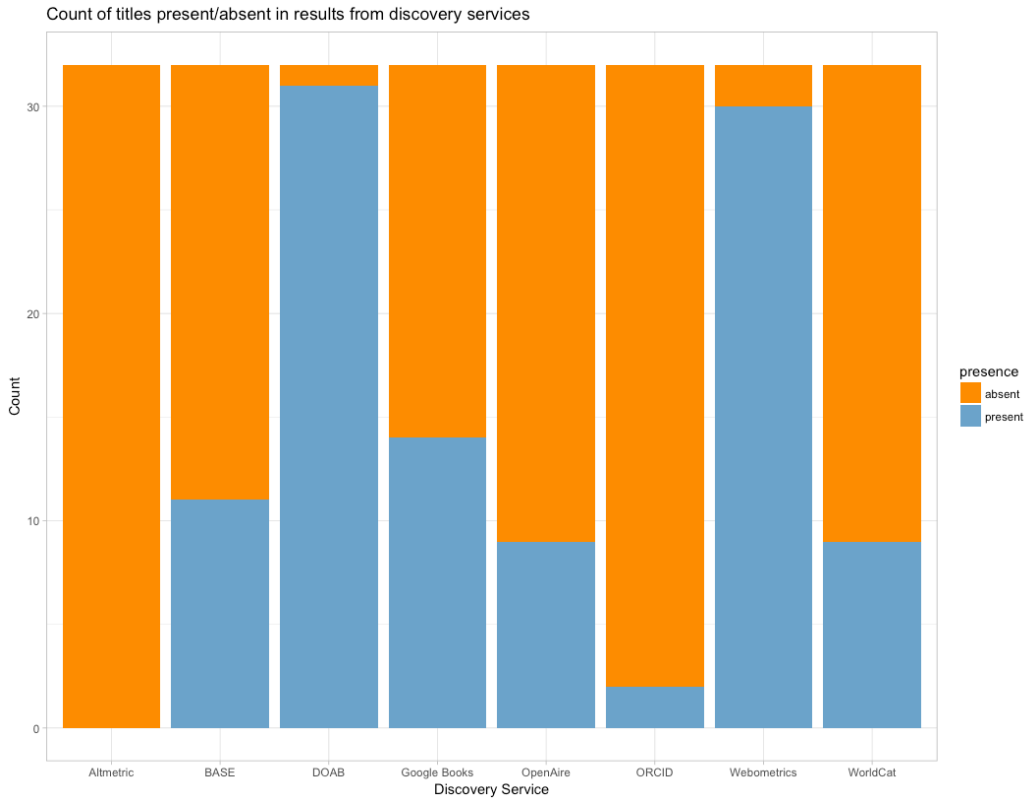

Count of titles present/absent in results from discovery services
100
75
50
25
0
Count
Altmetric
BASE
DOAB
Google Books
OpenAire
ORCID
Webometrics
WorldCat
Discovery Service
presence
absent
present


Top publication language by Webometrics TLD
Language
English
German
NA

E. SHARE Press

SHARE Press is a not-for-profit open access publisher that operates as a collaboration between the Universities of Naples (Federico II, Istituto Orientale, Parthenope), Salerno, Sannio and Basilicata. As well as books, SHARE Press also publishes journals, research data and historical documentation. SHARE Press books are hosted via the University of Naples Federico II institutional repository.
In August 2017 the SHARE Press metafile included 32 books, all of which were in Italian.

Parameter	Number	%
Number of Books	32	100
Books with ISBNs	31	97
Books with DOIs	32	100



F. EKT

EKT is the National Documentation Centre of Greece, located at the National Hellenic Research Foundation in Athens. EKT operates as national infrastructure: seeking to collect, organise, and preserve the entire Greek scientific, research and cultural output (content and data), while making it available at both a national and global level via their own repository. In August 2017 the EKT metafile included 6 books, all of which were in Greek.

Parameter	Number	%
Number of Books	6	100
Books with ISBNs	6	100
Books with DOIs	6	100