Open access books through open data sources: Assessing prevalence, providers, and preservation

Mikael Laakso

ORCID 0000-0003-3951-7990

Mikael.laakso@hanken.fi

Department of Management and Organisation, Hanken School of Economics, Helsinki, Finland

Abstract

Science policy and practice for open access (OA) books is a rapidly evolving area in the scholarly domain. However, there is much that remains unknown. Utilizing open bibliometric data sources, this study aims to answer three questions: 1) How prevalent are OA books (data sources: Directory of Open Access Books, OpenAIRE, OpenAlex, Scielo Books, The Lens, WorldCat), 2) what web domains are responsible for offering full-text access to these OA books, and 3) to what degree can OA books be verified to be archived in trusted preservation services (data sources: Cariniana Network, CLOCKSS, Global LOCKSS Network, Portico). 396 995 unique records were identified from the OA book bibliometric sources, of which 19% were found to be included in at least one of the preservation services. The results suggest reason for concern for the long tail of OA books distributed at thousands of different web domains as these include volatile cloud storage or sometimes no longer contained the files at all. Data quality issues, varying definitions of OA across services, and inconsistent implementation of unique identifiers were discovered as key challenges. The study includes recommendations for publishers, libraries, data providers, and preservation services for improving monitoring and practices for OA book preservation.

Acknowledgements

The author is grateful to Alicia Wise and Ronald Snijder for assisting in the identification of available datasets and valuable feedback throughout the study.

Funding statement

This research was commissioned by CLOCKSS, DOAB, and OAPEN.

Data availability statement

The research data is made available as open data through Zenodo and can be downloaded from https://doi.org/10.5281/zenodo.7305477

Introduction

Making academic content openly available for everyone using the web has never been easier from a technical and financial standpoint. The maturity and widespread adoption of web and document standards take care of a lot of challenges that were creating friction in the past. Web services that facilitate content upload and open distribution of academic monographs, book chapters, individual article manuscripts, and entire journals are spiraling up at an unprecedented pace which has led to a rapid increase in the volume of academic content available out in the open. While these dissemination practices provide open access (OA) to the content for the moment, the practices for ensuring preservation and long-term access to OA book content are in their infancy. The number of OA books preserved is largely unknown, and practices are still developing. Based on evidence from recent interviews and workshops on OA book preservation with key stakeholders, many of the central questions related to best practices of preservation are still evolving and there is a need to gain more information about current practices and work towards robust preservation solutions (Bell 2020; Barnes, Bell, Cole et al 2022).

A recent study gauged the degree to which content from OA journals had vanished from the web since the year 2000, finding that at least 174 OA journals had vanished from the active web and had lacking preservation coverage for their published materials (Laakso, Matthias & Jahn 2021). Partly inspired by the findings of this study Project JASPER (JournAlS are Preserved forevER) was initiated (DOAJ.org 2021) which is a collaboration between CLOCKSS, DOAJ, The Internet Archive, The Keepers Registry, and PKP. There is currently no similar overview of materials lost, or at risk of being lost, for OA books. As there is growing momentum by science policy makers to require OA for academic books it would be important to scope the landscape through systemic studies to map the current preservation status of published materials.

The focus of this study was to conduct a data-driven mapping of the current landscape of prevalence, content providers, and preservation within the content domain of OA books. The focus of the study was on academic monographs and edited books that are, or have been, available OA. The aim was to filter out and exclude non-published theses and dissertations, reports, and individual book chapters to the degree possible. Outside the scope of this study were issues related to specific file formats for preservation, but rather whether a title is included in the archive of a recognized preservation service. The three specific research questions of this study were:

1. What is the current prevalence of OA books?
2. What web domains offer full-text access to OA books?
3. To what degree is this content able to be verified to be included in the coverage of content preservation services?

Previous research

The context of this study relates to two broader fields of research: 1) E-books and their preservation, 2) the context of OA book publishing. This section reviews and summarizes the key advances that have been made in both fields, with a focus on findings that are of relevance for the design and interpretation of the results of the study documented in this paper.

E-books and their preservation

The preservation challenges related to e-books have existed roughly as long as the medium has had any significant volumes of content published. It is around two decades since Frank Romano authored an article titled "E-books and the Challenge of Preservation" that identifies three related challenges to the preservation of e-books: "1. The location of the stored information, 2. The organization storing the information and its long-term viability and commitment to preservation, 3. Technical issues involving coded and recorded format, interfacing, and rights management.". (Romano 2003) When it comes to the context of OA books, we can today argue that the first two are still largely unresolved and a motivation for initiating this study, while the third could be argued to be partially resolved through the mature standardization of the most commonly used formats for representing static print digitally (PDF, EPUB) and the availability of reuse rights and the lack of digital rights management considerations for OA content. A central theme in Romano´s paper is the uncertainty of the different responsibilities between publishers and libraries when it comes to content in the purely digital domain. Therefore preservation organizations such as Portico and CLOCKSS, owned and governed by publishers and libraries together, have been set up and contribute to bridging this gap for scholarly journals and books. An earlier version of Romano´s text was included as part of a report commissioned by the Library of Congress and the Council on Library and Information Resources in the United States titled "Building a national strategy for digital preservation: Issues in digital media archiving" that included similar chapters that related to other mediums where digital preservation needs were emerging (e.g. periodicals, websites, sound and video)(CLIR and Library of Congress 2002). A report titled "Preserving eBooks" and published by the Digital Preservation Coalition efficiently summarizes the key challenges when it comes to the preservation of e-books when the specific context of OA content does not have to be considered (Kirchhof & Morrissey 2014). Hurley (2019) provides a comprehensive history, description, and comparison between the largest preservation service providers in the e-book space: Portico, CLOCKSS, and the Global LOCKSS Network.

In order for something digital to be uniquely identified there needs to be a widely adopted standardized system for how to assign identifiers to objects. In the paper-based world, ISBNs functioned reasonably well for this purpose, but as Scott & Orlikowski´s (2021) study on the digital transformation of the book industry reveals, the ISBN system has started to show some serious limitations with lacking adoption and differing practices for its use among publishers for e-books.

The context of OA book publishing

When it comes to the relationship between libraries and publishers in the context of e-book preservation the popular and professional press has published several items where the connection has been depicted as adversarial (see e.g. Robertson 2014; Kelley 2014; McGarry 2020). While titillating, these are often individual examples and not representative of the industry landscape as such, where publishers are also supportive of the need for preservation if it is efficient, affordable and does not compromise business models. One could put forward the argument that the particular context of scholarly OA books is unique and why a more harmonious relationship is likely to be of benefit to all involved parties. Most importantly there are few reserved rights and no digital rights management to be concerned about.

During the last decade there have been several national and European projects that have supported the building of infrastructures for OA books. For someone new to this space the number of projects and acronyms can seem confusing. Stern (2021) provides helpful narrative for the origins of OPERAS (open scholarly communication in the European research area for social sciences and humanities) and how many of the related initiatives like OAPEN (Open Access Publishing in European Networks), DOAB (Directory of Open Access Books), and OpenEdition are connected in this context. A recent project with significant focus on preservation is the COPIM project (Community-led Open Publication Infrastructures for Monographs) where a work package is dedicated to archiving and digital preservation. At the time of writing the project has produced a scoping report which covers a brief overview of existing technical methods for digital preservation together with findings from interviews and workshops, where a main initial conclusion is that diverse solutions are needed and that it is unlikely that any single model will provide a solution for everyone when it comes to preservation (Barnes, Bell, Cole et al 2022).

A key study in the context of scholarly OA books is Neylon, Montgomery, Ozaygen et al (2018) which presents an overview of their visibility and integration in the digital landscape of scholarly works. The authors conducted both a web survey and analysis of the OA book content and associated metadata for 7 publishers (including indexing inclusion in various web services), all partners of the European OPERAS network. While preservation is not directly dealt with in the report, there are many identified aspects upstream and in the overall technical landscape that influence whether and how preservation can later take place. The study highlighted several key challenges that OA books have in comparison to OA journal articles. Books are often distributed through multiple online platforms and the publisher's website might or might not be one, and there is no persistent identifier for the overarching work which strains the use of persistent identifiers such as DOIs and ISBNs for the various manifestations of that work. There is currently no comprehensive collection of usage data as there is with journals via COUNTER. The open systems used for cataloguing, indexing and discovering OA books are much younger than they are for OA journals, which shows in their lack of consistency and reliability. While journals have strongly shifted to online only, there is still a larger demand and practice for books to be printed, suggesting that the processes for digital and print will remain parallel at least for some time. The 7 studied publishers were small organizations with limited resources and capacity, calling out for coordination, shared services, infrastructures, and standards in their survey responses. The publishers delivered their metadata in various formats and levels of quality, from various file types to APIs, demonstrating the diversity in managing and making data available of published works. A particular challenge in the metadata

was the inconsistent use of persistent identifiers, where multiple ISBNs could be reported for different manifestations (e.g. editions and formats) of a book, in addition to a potential DOI that could also be inconsistently reported in cases where individual chapters also had their own DOIs. Only around 10% of all OA books from the publishers were associated with a DOI. The authors argued the variable quality of book metadata created challenges for reliably studying their presence and indexation in various web services as the study compared publisher-provided records and that of various external services and indices (WorldCat, BASE, Google Books, DOAB, OpenAIRE). Most of these resources were at least 80% comprehensive, however, BASE had a low inclusion rate of around 40%. The study found no major difference in the degree to which content in different languages was included in the various services.

Building on these findings, a substantial part of the COPIM project was dedicated to developing a minimum metadata requirement for OA books based on the needs of key stakeholders. This work was part of a more encompassing Open Dissemination System, which also included guidelines for standardized use of persistent identifiers (Stone, Gatti, van Gerven Oei et al 2020). In the COPIM project, Barnes, Bell & Cole et al (2022) found, through their interviews with stakeholders, that there are some publishers that upload published content to local repositories. The role that repositories could potentially fill in the context of preservation is still largely unexplored and undefined. One challenge is the heterogenous landscape of repositories, operating with varying organizational backing and technical expertise for ensuring long-term access to content. The way through which repositories perceive themselves as responsible for long-term preservation of the outputs of an institution varies and is not in any way an underlying requirement for running a repository (Francke, Gamalielsson & Lundell 2017). A recent literature review of 21 studies dealing with long-term preservation in institutional repositories showed concerning findings where the review "…has not found clear evidence about how institutional repositories are implementing digital preservation" suggesting that more clarity into the roles and responsibilities of repositories is needed (Barrueco & Termens 2022:161). Another recent study found that about a quarter of repositories registered in widely used repository indexing services gave an erroneous response and were inaccessible when an attempt was made to visit the registered URL to the repository (Mannocci, Baglioni, & Manghi 2022), a finding which is very concerning. There is an active discussion ongoing about what kind of information would be required to evaluate which repositories can be trusted for long-term preservation (Lin, Crabtree, Dillo et al 2020), but so far there is no wider adoption of any practical standards outside of CoreTrust Seal which currently has certified only around 190 repositories and data archives (coretrustseal.org, n.d.).

From the reports and studies done on the landscape of organizations involved in OA book publishing it is known that many actors are small. In such contexts it is important that IT systems automate and guide as much of routine workflow steps as possible. One practical example of this is given as part of a case study of a university press where Taylor (2019) notes the following related to the selection of a particular publication management system: "Features which particularly appealed to us included the automatic registration of digital object identifiers (DOIs), the ability to send content to a preservation service at the click of a button…" (p.6). Comprehensive and standardized metadata for facilitating aggregation into external services is also important and should be supported and automated by the publishing platform. In a study of how users access books from the OAPEN Library website 73% directly accessed the full-

text file without opening the actual website, suggesting access by other means such as aggregators, search engines, and libraries (Snijder 2019).

Momentum is building for libraries getting more seriously involved in the structural funding of scholarly OA books (see e.g. doabooks.org 2022) which brings the content closer to libraries and their potential preservation processes. Recently UKRI (UK Research and Innovation) commissioned a gap analysis of the infrastructures for OA scholarly books, where preservation was included in the comprehensive scope of analysis (Ferwerda, Mosterd, Snijder et al 2021). Regarding preservation the authors identified a gap as "Technical challenge of preservation and ambiguity concerning who is responsible for the preservation of OA books." (p.7) with a recommended action to collaborate with UK legal deposit libraries and international partners.

The strategies and processes for how national libraries are approaching e-book preservation have appeared in the academic literature, with some prominent examples being China (Wei, Ji & Dong 2014), and France (Derrot & Clément 2014; Derrot, Moreux, Oury et al 2014), and the UK (Gooding, Terras & Berube 2021), but none of these mention or deal with OA materials specifically. One way that many national libraries have for centuries tried to preserve works published in the country is through legal deposit, where publishers are required by law to submit copies (be it digital or printed) to the national library upon publication. Muir (2001) is one of the early seminal works studying how libraries approach the deposit of digital publications. More recently Roudik, Buchanan, Ahmad et al (2018) provided a review and comparison of how digital legal deposit is implemented in 15 countries. The International Internet Preservation Consortium (IIPC) also maintains a list of countries with Legal Deposit laws covering web archiving IIPC (n.d). Unfortunately, most access to these archives is restricted to either on-premise access or based on evaluation of individual applications. This approach to long-term preservation therefore does not facilitate sustained and uninterrupted access to OA or other published works. An exception to this is an initiative focusing on OA books specifically that is run by the Library of Congress, where they are ingesting titles for which they already have print holdings directly from DOAB, both for preservation purposes and for making them available through their own website (Cassidy-Amstutz, Darby, Holdzkom et al 2022). Implemented at a larger scale, involving more libraries and titles, these kinds of actions could help contribute a robust layer of resilience to the preservation of open materials.

From this overview of previous research, it can be concluded that broad and systematic studies on the three main focus areas of this study (prevalence, providers, and preservation) for OA books has not really been attempted before. Overall, the literature related to these topics is dominantly populated with project reports and conference proceedings, suggesting that there is potential for making substantial contributions to the academic literature through empirical studies.

Methods

Already from the outset it was known that the data collection circumstances for OA book content differ significantly from that of scholarly journals. It is possible to identify journals, assess which have potentially vanished from the web, and verify their preservation status using the Keepers Registry and Internet Archive snapshots of the last known URL (Laakso, Matthias & Jahn 2021). For OA books the situation is much more fragmented and challenging because there is no centralized registry for archival holdings by preservation service providers.

For this study two datasets needed to be put together and compared: one for OA books and the second for preservation coverage of books. For both, open data sources were utilized in order to enable assessment of the quantity and quality of the data, and to enable the study to be easily replicated by anyone. A third methodological component of the study focused on retrieving the web domain information for the OA books with an assigned DOI. All collected data is available as an open dataset (Laakso 2022).

OA book data

Not all books on the web are of key interest to this study, where focus is on non-fiction academic books. Most bibliometric databases provide filtering to either "Book" and/or "Monograph" with very few offering further ways to reliably narrow the scope down from there. There is no widely used tag for "peer-reviewed" or similar that would make it possible to filter the large quantity of entries down, leaving it up to the inclusion criteria/data harvesting methods of each service provider to what is included and what is not. Further, as categories are so wide there can be theses, reports, and individual book chapters sprinkled in among the search results which are hard to identify and separate in any automated way. This is not only a factor that concerns metadata, but also overall transparency and knowledge about what kind of editorial processes are behind published works. For the sake of replicability it is not viable to manually edit the data without clear criteria. Ambiguity is also introduced by the concept of OA, as some sources allow filtering to content available in full text for free (potentially without reuse rights in perpetuity and therefore not an OA type), some do not have OA filtering at all, and some have very granular metadata concerning OA metadata.

As described earlier, the information environment concerning OA books is heterogenous and an appropriate study design should take this into account to avoid drawing conclusions based on the circumstances of books included in just one of many information sources. As there is no single data source that would comprehensively list all currently available OA books or their metadata, sampling books records from multiple sources is an efficient way to get a good grasp of the characteristics of preservation coverage for materials listed or stored across different services.

In Table 1 an overview of the bibliometric sources containing records of OA books is provided, with columns describing URL to the service, the search criteria used and the number of results it generated, the prevalence of ISBNs and DOIs among the results, and the point in time when the service was accessed.

Table 1 Overview of bibliometric sources containing records of OA books

Service	URL	Search criteria and volume of results	Unique identifier availability	Method and time of access
The Lens	https://www.lens.org	348 678 records under "Open Access" and "Book" published between year 0 and 2050	ISBN = 0% DOI = 99%	Web service queried 06052022
OpenAIRE	https://explore.openaire.eu/	211 749 records under "Open Access" and "Books" after removal of content labeled as chapters, thesis, reports, and preprints	ISBN = 0% DOI = 99%	JSON dataset by Baglioni, Bardi, Atzori et al (2022). Data based on OpenAire dump published 23122021
OpenAlex	https://openalex.org/	134 718 records of type "Book", or "Monograph" and OA type Gold, Hybrid or Bronze	ISBN = 0% DOI = 100%	API queried 04072022
DOAB	https://www.doabooks.org/	52 002 scholarly peer-reviewed books, all OA, 49 600 after removing items tagged as chapter	ISBN = 82% DOI = 83%	CSV dataset accessed 06052022
WorldCat (OCLC)	https://www.worldcat.org/	4485 non-fiction e-books tagged as OA	ISBN = 100% DOI = 21%	Website queried 28042022
Scielo Books	https://books.scielo.org/	1006 OA book records	ISBN = 100% DOI = 93%	Website queried 06052022

Data was aggregated from a variety of sources, scoped both narrow and wide, and some having exclusively OA book content while others are general purpose bibliometric databases. OpenAlex (Priem, Piwowar, Orr 2022), The Lens, and OpenAire (Baglioni, Bardi, Atzori et al 2022) can all be considered to be broad scholarly bibliometric databases that require a lot of filtering in order to narrow down to the specific information relating to the target group of OA books. The Lens and OpenAire offer no way to select specific OA mechanisms to be included/excluded, but in OpenAlex the option to exclude Green OA/repository copies was utilized to primarily obtain content that had been published OA directly by the publisher. WorldCat is also a broad database but offered quick ways to filter down to the relatively small number of records that related to this study directly from the webpage. The data from DOAB and Scielo Books was imported wholesale since they contain only data relevant for this study, with the exception of excluding individual chapters for a small part of the DOAB data.

Though the volume and quality of openly available metadata concerning OA books is better than it has ever been and is constantly improving, there are some obstacles for straightforward duplication checking when data is aggregated from several complementary data sources. There is varying use of unique identifiers for books, where DOIs are mostly the key identifier used among bibliometric data providers included here. Matching only by title is not optimal due to even small differences in spelling, format and punctuation leading to incorrect matches.

Through deduplication utilizing DOIs, ISBNs, and Titles the total unique record count was 396 995.

Preservation data

The challenges mentioned so far have concerned creating a comprehensive dataset of OA books, but none of the data so far can provide an indication for which titles are reported to be preserved through a trusted preservation service. CLOCKSS (2022), Portico (2022), Global LOCKSS Network (2022), Cariniana Network (Márdero Arellano, Abbud Grácio2021) all provide open datasets that describe which books they have included in their holdings. None of these four provide DOIs for their records, only ISBNs which is not optimal as most of the major bibliometric service providers focused on OA book content rely on DOIs.

Table 2 Overview of data sources containing preservation information of books

Service	URLs	Coverage	Unique identifier availability	Time accessed
CLOCKSS	https://reports.clockss.org/keepers/keepers-CLOCKSS-books-report.csv	389 820 books	ISBN = 100% DOI = 0%	Date downloaded 01062022, File dated 23052022
Portico	https://api.portico.org/holdings/ebooks/e-books-part1.xlsx https://api.portico.org/holdings/ebooks/e-books-part2.xlsx	1 945 233 books	ISBN = 100% DOI = 0%	Date downloaded 01062022, File date UNKNOWN
Global LOCKSS Network	https://reports.lockss.org/keepers/keepers-LOCKSS-books-report.csv	21 260 books	ISBN = 100% DOI = 0%	Date downloaded 01062022, File dated 23052022
Cariniana Network	https://livroaberto.ibict.br/browse?type=title&sort_by=1&order=ASC&rpp=500&etal=0&submit_browse=Update	461 books	ISBN = 100% DOI = 0%	Date downloaded 01062022

National libraries have good data within them but programmatic access from outside is still limited. Barnes, Bell & Cole et al (2022) found that some OA monograph publishers deposit copies into national library holdings, something which would be very interesting to obtain more information about on a larger scale. However, the holdings of national libraries around the world are not easy to query programmatically from the outside.

Determining web domains for OA books

This study explored what domains host the OA book content, by checking which URLs their DOIs direct to when queried. This is not a way of verifying the preservation status of the books, but such an exploration can shed light on the nature and capacity of the long tail for OA book providers. An automated process of querying was set up in the Octoparse software application, simply recording the URL that was received as a response (if any response was received) when a web browser queried the DOI web address. This process was performed individually for the

DOIs of the records in four of the six largest OA book data sources. In the case of Scielo Books all books are hosted on the platform itself, and in the case of WorldCat the number of DOIs found among records was low. In total 745 535 DOIs were queried during June, July and August of 2022. The reason for the high number of DOIs in comparison to the overall number of unique records in the final OA books dataset (396 995) was that queries were performed for all records per-data source and in parallel with the deduplication and analysis process. The results for this part of the study are also presented by data source to better describe the content distribution for individual databases. In some individual cases, likely due to blocking frequent queries from the same IP address, the DOIs would not automatically be resolved. In such cases the same web domain was recorded as for other records with the same DOI registrar prefix.

Matching OA book data to preservation data

In order to establish which records from the OA book dataset were also represented in the preservation dataset the common denominators were ISBN and book title, both of which were used to find matches in the datasets. Due to the lack of DOI data in the preservation datasets matching had to be performed on only these two data points, were a match on either would be considered to indicate that the title was included in the holdings of a trusted preservation service. Some books were recorded with multiple ISBNs in both the book data and the preservation data, and matching was performed on all these ISBNs in all possible combinations.

Limitations and considerations

There is a need for further study into the correctness and data quality of the bibliometric data offered by the sources utilized in this study. For particularly the broader datasets, it is apparent that not all records classified as “book” or “monograph” are actually that in reality. This would also extend to verification of the OA status and classification of content. Due to the size of the dataset and emphasis on replicability, this study relies on the classifications given by the data providers with all the uncertainty that entails.

Books, similar to journals, can also be available in physical print format. For OA books the specific practice of print-on-demand is also characteristic, meaning that there can be small quantities of a printed OA book in circulation. What this study cannot establish is the print preservation status of OA books that are, or have at some point been, available to purchase in print form. It would be useful to gain more insight into the current adoption of print media for books that are available OA, however establishing this knowledge reliably is best suited for a dedicated study. Another useful study would be on the inter-relationship of print and digital preservation practices in the library world. Digital preservation is needed in addition to print preservation if functionality and reader experience are to be preserved as well as the content itself.

One of the challenges that e-books introduce in addition to their storage and distribution media is the potential to integrate dynamic multimedia content rather than just static information. While it would be interesting to look more closely at the unique preservation challenges and risks associated with the content of such books, a wide-scale study like this is not suited to go deeply into these issues.

The Venn diagrams produced to visualize the data distributions were produced using DeepVenn (Hulsen, de Vlieg & Alkema 2008). They are calculated to be area-proportional, however, due to the complexity of several overlapping datasets and the limitations of circular shapes that is not always completely possible. As such we advise the use of these visualizations to be approximations of the distributions. Please consult the exact percentages and numbers provided as Appendix 1 for any exact calculations.

Results

This section is divided into three sections, each corresponding to one of the three research questions.

What is the current prevalence of OA books?

To answer this question descriptive statistics for the query results from the six bibliometric data sources are presented (i.e. The Lens, OpenAIRE, OpenAlex, DOAB, WorldCat, Scielo Books), both individually and together as a deduplicated dataset. The breakdown of content identified through the individual databases was presented in the methods section, with the result of deduplication being 396 995 records.

Since all data sources provided a metadata field for year of publication, an analysis of what age the content provided was from was performed. The results can be seen in Figure 1, where it is clear that a considerable share of content provided through The Lens (130 413 records), and to a lesser degree OpenAlex (30 979 records), had been published before 1975. There is only a relatively minimal amount of content published between 1975-1999 with a consistent growth trend starting from the year 2000 onwards. It warrants a mention that the data is not consistently capable of telling when a specific piece of content was made OA, only when the original work was published.

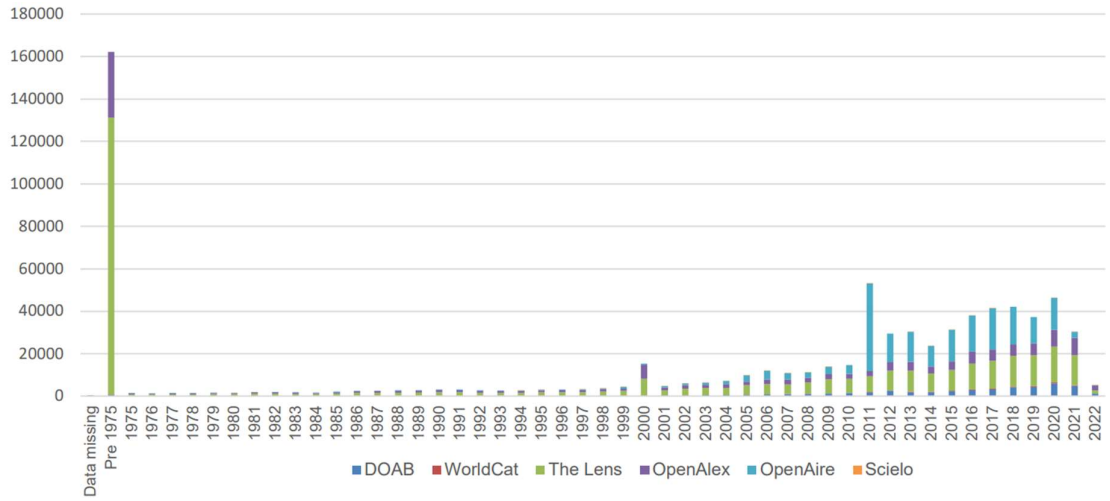

Figure 1 Publication years of OA book content from the six bibliometric sources

The next step for understanding OA book prevalence, and for informing future studies in the area, is to see how content is distributed across the different bibliometric datasets. Table 1 in the methods section already presented how many records each individual data source provided but that presentation did not analyze for overlap. Figure 2 presents the content distribution across the six bibliometric data sources with the overlap record counts across all data sources being available in Appendix 1.

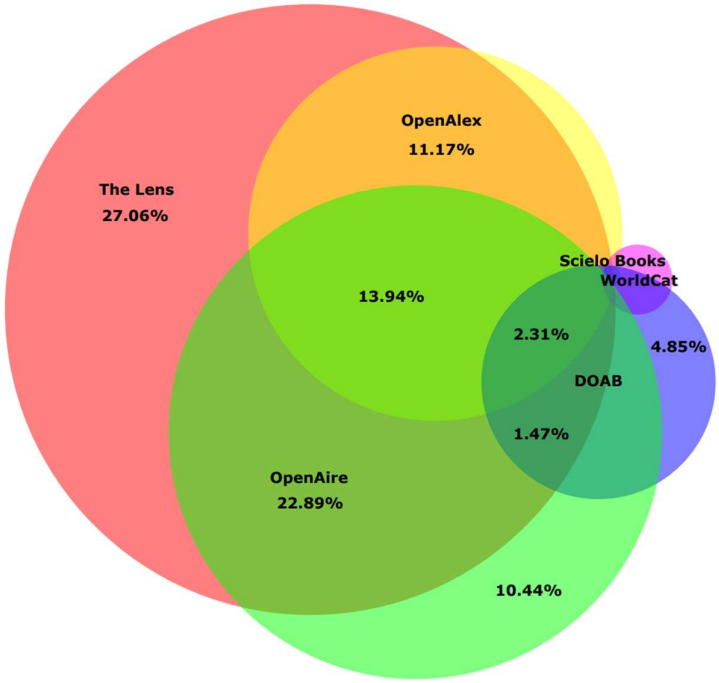

Figure 2 Content distribution across the six bibliometric data sources. Shares under 1% are not marked with a numeric label.

From the results of this analysis it is clear that there is a lot of overlap in content between the sources, which explains why almost half of all records were merged in the deduplication process (from 750 236 to 396 995). As Figure 2 illustrates, most unique records were contributed by The Lens (27% of the deduplicated dataset) followed by OpenAire (10%). The Lens shares substantial overlap with both OpenAlex and OpenAire, and 41% of records were found in all three. OpenAlex and The Lens are almost completely overlapping with under 1% (3280 records) of the final dataset being unique to OpenAlex. What is perhaps a bit surprising is that despite the DOAB data being considerably smaller than the larger data sources included in the study, it still contributed 5% of the unique records for the final dataset.

What web domains offer full-text access to OA books?

The next step in the inquiry was to find out what web domains the DOIs of the records point to when queried. The methods section describes the approach used (web scraping the 745 535

DOIs found in the four largest data sources used). The results are visualized as treemaps in Figure 3 with a summarizing breakdown of top domains provided as Table 3.

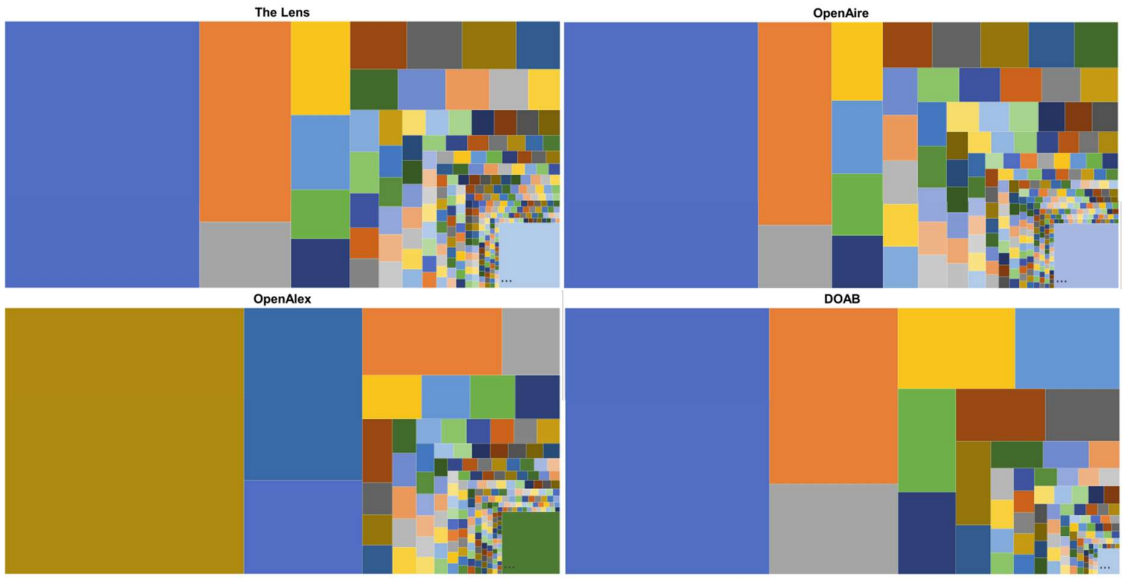

Figure 3 Visual treemap representation of the distribution of content on unique web domains for records with DOIs retrieved from these sources.

While not providing exact numbers, Figure 3 conveys a general content distribution overview for the four data sources. Despite the difference in volume of records they share the general trait that around half of the content, or somewhat over in the case of OpenAlex and DOAB, is hosted by three individual web domains. As Table 3 indicates, link.springer.com and biodiversitylibrary.org can be found in the top 3 web domains for three of the four data sources. In the bottom row of Table 3 is a summary of the remaining web domains that did not fit the table, giving an indication of the "long tail" of content provided by these domains. DOAB is exceptionally clustered with only 7% of content from 188 domains that did not have domains listed in the top frequency list. Content from the other three data sources was found to be much more widely distributed with 23%-32% of content being held on 1453-1816 web domains that did not fit into the frequency top list.

Due to the sheer volume of web traffic already created by querying the almost 750 000 DOIs content analysis of page content or download of full-text copies was not performed, so this study is not capable of providing information about actual download capability/availability on the pages that are directed to. Though it would warrant more detailed dedicated investigation, the long tail of web domains contains some clearly volatile services (e.g. Dropbox, Google Drive, organizational subpages etc) as well as some DOI addresses giving HTTP 404 errors indicating that the web page is no longer accessible.

DOAB	count	OpenAire	count	OpenAlex	count	The Lens	count
library.oapen.org	18889	biodiversitylibrary.org	74133	biodiversitylibrary.org	23347	biodiversitylibrary.org	121956
books.openedition.org	7889	link.springer.com	21466	afghandata.org	19840	link.springer.com	43261
mdpi.com	4023	elibrary.worldbank.org	6666	link.springer.com	8174	law.acku.edu.af	14183
mts.intechopen.com	3274	degruyter.com	5747	law.acku.edu.af	5562	afghandata.org	12944
frontiersin.org	2930	cambridge.org	5356	books.openedition.org	4566	onlinelibrary.wiley.com	10323
intechopen.com	2092	classiques.uqac.ca	4583	library.si.edu	4237	elibrary.worldbank.org	6910
degruyter.com	1652	books.openedition.org	3856	classiques.uqac.ca	4193	classiques.uqac.ca	6759
ksp.kit.edu	1647	taylorfrancis.com	3239	repository.usta.edu.co	4031	ieeexplore.ieee.org	6416
media.fupress.com	1371	library.si.edu	3175	openknowledge.worldbank.org	2036	degruyter.com	6241
books.scielo.org	1009	apps.crossref.org	3168	constellation.uqac.ca	1950	dl.acm.org	6078
omp.zrc-sazu.si	591	mr.crossref.org	3002	vr-elibrary.de	1837	journals.openedition.org	4922
ucdigitalis.uc.pt	494	repository.usta.edu.co	2854	darchive.mblwhoilibrary.org	1721	taylorfrancis.com	4569
nomos-elibrary.de	429	vr-elibrary.de	2350	press.umich.edu	1692	apps.crossref.org	4518
edp-open.org	288	oxford.universitypressscholarship.com	2277	apps.crossref.org	1634	repository.si.edu	4193
link.springer.com	252	press.umich.edu	2203	mohrsiebeck.com	1445	repository.usta.edu.co	3702
ledizioni.it	228	rand.org	2109	books.fupress.com	1353	mdpi.com	3007
bloomsburycollections.com	193	worldscientific.com	2077	liu.diva-portal.org	1294	jstor.org	2855
e-archivo.uc3m.es	170	darchive.mblwhoilibrary.org	2028	jstor.org	1279	deepblue.lib.umich.edu	2819
api.intechopen.com	162	constellation.uqac.ca	2026	rand.org	1230	academic.oup.com	2410
188 more domains containing the remaining 7% of items		1453 more domains containing the remaining 28% of items		1470 more domains containing the remaining 32% of items		1816 more domains containing the remaining 23% of items	

Table 3. Web domains for content with DOIs included in the four largest data sources of the study.

To what degree is this content able to be verified to be included in the coverage of content preservation services?

This step of the study cross-matched the OA book records found from the six bibliometric databases with the data retrieved from the four preservation service providers. Table 4 presents a breakdown of how the content records retrieved from each OA book data source was represented in the various preservation services based on ISBN and/or book title matching.

Table 4 illustrates that OA book content listed in DOAB is covered to the highest degree by at least one of the services (46% of all DOAB records) with WorldCat (33%), OpenAire (25%), OpenAlex (13%), The Lens (10%), and Scielo Books (9%) following in descending order. Among the preservation service providers Portico provides the overall highest numbers with 31% of coverage for both DOAB and WorldCat, and also the highest numbers for any service for the rest. Matches were only minimally found to records in the Global LOCKSS Network and the Cariniana Network preservation data, ranging between 0% and 1% percent depending on OA book data source.

Table 4 Preservation coverage analysis for OA book content derived from the six bibliometric databases.

	DOAB	WorldCat	The Lens	OpenAlex	OpenAire	Scielo
CLOCKSS	22 %	7 %	3 %	4 %	8 %	0 %
Portico	31 %	31 %	9 %	11 %	22 %	9 %
Global LOCKSS Network	0 %	1 %	0 %	0 %	0 %	0 %
Cariniana Network	0 %	0 %	0 %	0 %	0 %	0 %
Found in at least one of the above	**46 %**	**33 %**	**10 %**	**13 %**	**25 %**	**9 %**

Note: There is overlap in the coverage between the different preservation service providers leading to the bottom row being less than the direct sum of the rows above.

As a secondary perspective on the preservation coverage Figure 4 presents a visualization of the uniqueness and overlap of preservation coverage for the deduplicated dataset of 396 995 OA book records. Portico provides 14% of preservation coverage uniquely, with CLOCKSS having 2% and sharing a 3% total coverage overlap with Portico. Overall this analysis also provides the total coverage for preservation for the deduplicated dataset based on the data sources utilized and compared to each other, which is 19%.

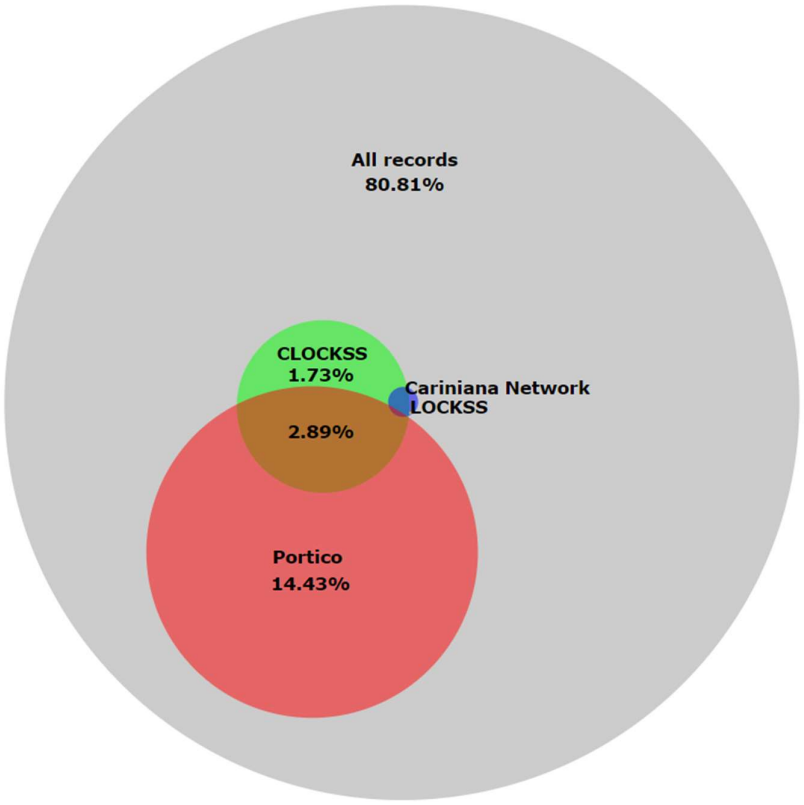

Figure 4 Preservation coverage for the 396 995 OA book records.

Discussion

The main finding of the study is that, based on aggregation of data from various widely used open bibliometric databases, one can identify 396 995 OA book records, of which 19% were found to be archived by one of the four preservation service providers which open data was used for this study. While these are exact numbers, the experiences garnered by executing this study raised many flags of uncertainty when it comes to making an exact science of preservation coverage with the current data availability and data quality that there is for both content and preservation. As such the results of this study should be considered an estimate rather than absolute and comprehensive. This is because the definitions and practices in the landscape are still emerging, something which the many caveats of this study illustrate. It should also be noted that based on best-practice, content should be preserved through more than one provider, some have argued three different trusted long-term archives to be safe (blog.dshr.org 2022). One archive is good but should not be considered great.

Though the issues of ambiguity in the definitions were known already at the outset through the findings of Neylon, Montgomery, Ozaygen et al (2018), the varying ways through which content providers and aggregators classify scholarly books and OA to content based on their own non-transparent criteria or erroneous automated classification presented a larger challenge than expected. There is a need for more standardization in how metadata can reliably indicate e.g. peer-review status of content in a reliable way, as well commonly adopted definitions for the different content types (e.g. monograph, edited book) that would reduce the amount of

obviously non-book content that shows up among search results with the most suitable criteria available today. While it could be argued that many of these services are primarily intended for discovery of relevant content rather than comprehensive bibliometric research, having these two data points enhanced would likely also cater to more relevant content being offered users when querying the growing amount of content that gets indexed in these services.

The gaps left by the varying practices for usage of unique identifiers for content is something that would need to be remedied in order for data matching to be more reliable. Currently there is a lot of room for error for studies that extend beyond one single data source when there is reliance on book title matching. Data sources that include book materials should strive to include both ISBNs and DOIs in the metadata when they are available since that makes matching to preservation data much more reliable. Preservation service reports are still dominantly ISBN-based at least when it comes to public book preservation data, an expansion into also including DOIs would be beneficial for many purposes.

Recommendations

How should collaboration evolve among major stakeholders (e.g. publishers, libraries, preservation services providers) develop in order to establish higher coverage and flexible workflows?

It could be argued that OA content would benefit from OA status information for preservation, i.e. that there would be practices and data in place that would make it easy to both deposit and verify where specific pieces of openly available content are properly preserved. Concerning preservation data national libraries could on their own or through collaboration make available open machine-readable data concerning which books are preserved in their digital holdings. A service similar to The Keepers Registry that the ISSN International Centre maintains for journals would be very helpful for books as well, so preservation service providers could automatically report which titles they include in their holdings.

For future research the open data produced by this research should help facilitate extended and deepened data-driven inquiries into the landscape of OA books. The study also functions as a detailed snapshot of the current situation on the entire spectrum, opening up for comparative studies in the future. The study lays an empirical foundation to develop theoretical connections between preservation and the concepts of time and temporality within library and information science (Haider, Johansson & Hammarfelt 2021). Preservation is through its inherent actions preparing for a future state in time, that in a best case scenario will not have to be utilized, and scholarly explorations in this domain would likely prove fruitful.

References

Baglioni, M., Bardi, A., Atzori, C., & Manghi, P. (2022). Books from the OpenAIRE Research Graph (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6619395

Barnes, M., Bell, E., Cole, G., Fry, J., Gatti, R., & Stone, G. (2022). WP7 Scoping Report on Archiving and Preserving OA Monographs (1.0). Zenodo. https://doi.org/10.5281/zenodo.6725309

Barrueco & Termens (2022). Digital preservation in institutional repositories: A systematic literature review. Digital Library Perspectives, 38(2), 161–174. https://doi.org/10.1108/DLP-02-2021-0011

Bell, E. (2020). COPIM Archiving and Preservation Workshop, September 2020. *COPIM*. https://doi.org/10.21428/785a6451.0e666456

blog.dshr.org (2022). Where Did The Number 3 Come From? David Rosenthals Blog. https://web.archive.org/web/20221031174833/https://blog.dshr.org/2022/06/where-did-number-3-come-from.html

Cassidy-Amstutz, A., Darby, K., Holdzkom, E., Salas, C., Seroka, L. (2022). Creating Workflows to Scale Out Open Access E-book Acquisitions at the Library of Congress Andrew. Paper presented at the International Conference on Digital Preservation (iPres 2022). Glasgow, Scotland 12-16.9.2022. https://web.archive.org/web/20221006131808/https://az659834.vo.msecnd.net/eventsairwesteuprod/production-inconference-public/3f4dd08cbb3842739c82ccac5a422de0

Márdero Arellano, M.A., Abbud Grácio, J.C. (2021). The Cariniana Network for Digital Preservation. The Digital Preservation Coalition. https://web.archive.org/web/20220127073614/https://www.dpconline.org/blog/wdpd/cariniana-wdpd21. Accessed on the 8th of October 2022.

CLIR and Library of Congress (2002). Building a national strategy for digital preservation: Issues in digital media archiving. Council on Library and Information Resources and Library of Congress. https://web.archive.org/web/20180107015602/https://www.clir.org/wp-content/uploads/sites/6/2016/09/pub106.pdf

CLOCKSS (2022) https://reports.clockss.org/keepers/keepers-CLOCKSS-books-report.csv

coretrustseal.org (n.d). CoreTrust Seal. https://www.coretrustseal.org/.

Derrot, S., Moreux, J-P., Oury, C., Reecht, S. (2014). Preservation of ebooks:from digitized to born-digital. 11th International Conference on Digital Preservation (iPRES),Oct 2014, Melbourne, Australia. Proceedings of the International Conference on DigitalPreservation (iPRES). https://hal-bnf.archives-ouvertes.fr/hal-01088755

Derrot, S., Clément, O. (2014) *Ebooks: rather electronic or book? Extending legal deposit to ebooks at the Bibliothèque nationale de France*. Paper presented at: IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 87 - Information Technology with Preservation and Conservation and National Libraries. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France. https://web.archive.org/web/20220121090010/https://library.ifla.org/id/eprint/830/1/087-derrot-en.pdf

Doabooks.org (2022). Building stronger infrastructures to support open access books: LYRASIS, DOAB and OAPEN. https://web.archive.org/web/20220401080150/https://doabooks.org/en/article/building-stronger-infrastructures-to-support-open-access-books-lyrasis-doab-and-oapen

DOAJ.org (2021). Project JASPER - Open access journals must be preserved forever. https://web.archive.org/web/20210916132815/https://doaj.org/preservation/

Ferwerda, E., Mosterd, T., Snijder, R., & Mounier, P. (2021). UKRI Gap Analysis of Open Access Monographs Infrastructure. Zenodo. https://doi.org/10.5281/zenodo.5771945

Francke, H., Gamalielsson, J., & Lundell, B. (2017). Institutional repositories as infrastructures for long-term preservation. Information research, 22(2).

Global LOCKSS Network (2022). https://reports.lockss.org/keepers/keepers-LOCKSS-books-report.csvGooding, P., Terras, M., & Berube, L. (2021). Identifying the future direction of legal deposit in the United Kingdom: The Digital Library Futures approach. Journal of Documentation, 77(5), 1154–1172. https://doi.org/10.1108/jd-09-2020-0159

Haider, J., Johansson, V., & Hammarfelt, B. (2021). Time and temporality in library and information science. Journal of Documentation, 78(1), 1–17. https://doi.org/10.1108/jd-09-2021-0171

Hulsen, T., de Vlieg, J. & Alkema, W. (2008). BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **9**, 488. https://doi.org/10.1186/1471-2164-9-488

IIPC (n.d). International Internet Preservation Consortium – Legal Deposit. https://web.archive.org/web/20221007095441/https://netpreserve.org/web-archiving/legal-deposit/. Accessed on the 7th of October 2022.

Kelley, M. (2014). How Libraries Preserve E-books - How do we keep e-books from being lost in a deepening digital memory hole? https://web.archive.org/web/20210419015521/https://www.publishersweekly.com/pw/by-topic/industry-news/libraries/article/64271-check-it-out-with-michael-kelley-how-libraries-preserve-e-books.html

Kirchhof, M., Morrissey, S. (2014) Preserving eBooks. DPC Technology Watch Report 14-01 June 2014. Digital Preservation Coalition. https://web.archive.org/web/20211010182422/https://www.dpconline.org/docs/technology-watch-reports/1230-dpctw14-01/file. Video recording: https://www.dpconline.org/events/past-events/preserving-ebooks

Laakso, M. (2022). Dataset for "Open access books through open data sources: Assessing prevalence, providers, and preservation". Version 1.0. Zenodo. https://doi.org/10.5281/zenodo.7305477

Laakso, M., Matthias, L., Jahn, N. (2021). Open is not forever: A study of vanished open access journals. *J Assoc Inf Sci Technol*. 2021; 72: 1099– 1112. https://doi.org/10.1002/asi.24460

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., ... & Westbrook, J. (2020). The TRUST Principles for digital repositories. *Scientific Data*, *7*(1), 1-5. https://doi.org/10.1038/s41597-020-0486-7

Mannocci, A., Baglioni, M., & Manghi, P. (2022). „Knock knock! Who's there?" A study on scholarly repositories' availability. http://arxiv.org/abs/2207.12879

McGarry C. (2020). Libraries Could Preserve Ebooks Forever, But Greedy Publishers Won't Let Them. Gizmodo. 2nd of March 2020. https://web.archive.org/web/20211017110515/https://gizmodo.com/libraries-could-preserve-ebooks-forever-but-greedy-pub-1841922375

Muir, A. (2001). Legal deposit and preservation of digital publications: a review of research and development activity. Journal of Documentation, 57(5), 652–682. https://doi.org/10.1108/eum0000000007097

Neylon, C., Montgomery, L., Ozaygen, A., Saunders, N. & Pinter, F. (2018). The Visibility of Open Access Monographs in a European Context: Full Report. Zenodo. https://doi.org/10.5281/zenodo.1230342

Portico (2022) https://api.portico.org/holdings/ebooks/e-books-part1.xlsx and https://api.portico.org/holdings/ebooks/e-books-part2.xlsx

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*. https://doi.org/10.48550/arXiv.2205.01833

Robertson, A. (2014). The fight to save endangered ebooks. The Verge. May 9th 2014. https://web.archive.org/web/20210205082656/https://www.theverge.com/2014/5/9/5688146/the-fight-to-save-endangered-ebooks

Roudik, P., Buchanan, K., Ahmad, T. T., Zhang, L., Isajanyan, N., Boring, N., Gesley, J., Levush, R., Figueroa, D., Umeda, S., Hofverberg, E., Rodriguez-Ferrand, G., & Feikert-Ahalt, C. (2018). *Digital Legal Deposit in Selected Jurisdictions: Australia, Canada, China, Estonia, France, Germany, Israel, Italy, Japan, Netherlands, New Zealand, Norway, South Korea, Spain, United Kingdom* (p. 78). Law Library of (United States) Congress, Global Legal Research Center, July 2018. https://www.loc.gov/law/help/digital-legal-deposit/digital-legal-deposit.pdf.

Romano, F. (2003). *E-books and the Challenge of Preservation. Microform & Imaging Review, 32(1).* https://doi.org/10.1515/mfir.2003.13

Scott, S. & Orlikowski, W. (2021). The Digital Undertow: How the Corollary Effects of Digital Transformation Affect Industry Standards. Information Systems Research 33(1):311-336. https://doi.org/10.1287/isre.2021.1056

Snijder, R. (2019). The deliverance of open access books: Examining usage and dissemination. Amsterdam University Press. https://doi.org/10.26530/OAPEN_1004809

Stern, N. (2021). A Brief Saga about Open Access Books. *Nordic Perspectives on Open Science*, March 2021. https://doi.org/10.7557/11.5751

Stone, G., Gatti, R., van Gerven Oei, V. W. J., Arias, J., Steiner, T., & Ferwerda, E. (2020). WP5 Scoping Report: Building an Open Dissemination System. Zenodo. https://doi.org/10.5281/zenodo.3961564

Taylor, M. (2019). Mapping the Publishing Challenges for an Open Access University Press. Publications, 7(4), 63. https://doi.org/10.3390/publications704006

Wei, D., Ji, S., & Dong, X. (2014). The Preservation Practice of EBooks in the National Library of China. Paper presented at: IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies:

Confluence for Knowledge in Session 87 - Information Technology with Preservation and Conservation and National Libraries. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France. https://web.archive.org/web/20200319233833/http://library.ifla.org/859/1/087-wei-en.pdf

Appendix 1

Numbers relating to Figure 2

Groups	Number of titles
The Lens	107444
The Lens ∩ OpenAire	90876
The Lens ∩ OpenAire ∩ DOAB	5855
The Lens ∩ OpenAire ∩ DOAB ∩ OpenAlex	9184
The Lens ∩ OpenAire ∩ DOAB ∩ OpenAlex ∩ WorldCat	228
The Lens ∩ OpenAire ∩ DOAB ∩ OpenAlex ∩ WorldCat ∩ Scielo	1
The Lens ∩ OpenAire ∩ DOAB ∩ OpenAlex ∩ Scielo	158
The Lens ∩ OpenAire ∩ DOAB ∩ WorldCat	174
The Lens ∩ OpenAire ∩ DOAB ∩ Scielo	16
The Lens ∩ OpenAire ∩ OpenAlex	55355
The Lens ∩ OpenAire ∩ OpenAlex ∩ WorldCat	125
The Lens ∩ OpenAire ∩ OpenAlex ∩ Scielo	3
The Lens ∩ OpenAire ∩ WorldCat	31
The Lens ∩ OpenAire ∩ Scielo	2
The Lens ∩ DOAB	5576
The Lens ∩ DOAB ∩ OpenAlex	2567
The Lens ∩ DOAB ∩ OpenAlex ∩ WorldCat	40
The Lens ∩ DOAB ∩ OpenAlex ∩ Scielo	84
The Lens ∩ DOAB ∩ WorldCat	55
The Lens ∩ DOAB ∩ Scielo	27
The Lens ∩ OpenAlex	44 363
The Lens ∩ OpenAlex ∩ WorldCat	86
The Lens ∩ OpenAlex ∩ Scielo	4
The Lens ∩ WorldCat	38
OpenAire	41 457
OpenAire ∩ DOAB	1668
OpenAire ∩ DOAB ∩ OpenAlex	333
OpenAire ∩ DOAB ∩ OpenAlex ∩ WorldCat	17
OpenAire ∩ DOAB ∩ OpenAlex ∩ WorldCat	12
OpenAire ∩ DOAB ∩ WorldCat	66
OpenAire ∩ DOAB ∩ WorldCat	1
OpenAire ∩ OpenAlex	5005
OpenAire ∩ OpenAlex ∩ WorldCat	8
OpenAire ∩ WorldCat	39
DOAB	19263
DOAB ∩ OpenAlex	142
DOAB ∩ OpenAlex ∩ WorldCat	3
DOAB ∩ OpenAlex ∩ WorldCat	7
DOAB ∩ WorldCat	1853
DOAB ∩ WorldCat	81
OpenAlex	3280
OpenAlex ∩ WorldCat	30
OpenAlex ∩ WorldCat	1
WorldCat	1434
Scielo	1

Numbers relating to Figure 4

Groups	Number of titles
Portico ∩ CLOCKSS ∩ LOCKSS ∩ Cariniana Network	1
Portico ∩ CLOCKSS ∩ LOCKSS	134
Portico ∩ CLOCKSS	11493
Portico ∩ LOCKSS	139
Portico	57286
CLOCKSS ∩ LOCKSS	237
CLOCKSS	6853
LOCKSS	55
Cariniana Network	4
Not included in any service	320793