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

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

Received 2 February 2023
Revised 23 May 2023
Accepted 29 May 2023

Abstract

Purpose – 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, including how many OA books there are and to what degree they are included in preservation coverage. The purpose of this study is to contribute towards filling this knowledge gap in order to advance both research and practice in the domain of OA books.

Design/methodology/approach – This study utilized open bibliometric data sources to aggregate a harmonized dataset of metadata records for OA books (data sources: the Directory of Open Access Books, OpenAIRE, OpenAlex, Scielo Books, The Lens, and WorldCat). This dataset was then cross-matched based on unique identifiers and book titles to openly available content listings of trusted preservation services (data sources: Cariniana Network, CLOCKSS, Global LOCKSS Network, and Portico). The web domains of the OA books were determined by querying the web addresses or digital object identifiers provided in the metadata of the bibliometric database entries.

Findings – In total, 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.

Research limitations/implications – 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.

Originality/value – This study provides methodological and empirical findings for advancing the practices of OA book publishing, preservation and research.

Keywords Metadata, Preservation, Publishing, Indexing, Monographs, Open access

Paper type Article

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 the content upload and open distribution of academic monographs, book

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: 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

Journal of Documentation
Emerald Publishing Limited
0022-0418
DOI 10.1108/JD-02-2023-0016

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 the 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 toward robust preservation solutions (Bell, 2020; Barnes *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 *et al.*, 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 policymakers to work toward OA for academic books (see e.g. cordis.europa.eu,2023; coalition-s.org, 2021), 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 and 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 USA 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 non-OA e-books (Kirchhof and Morrissey, 2014). A central finding of the report was that e-books outside of the OA context are substantially different in comparison to traditional books in that readers and libraries usually only gain access to a limited license to access the contents of the book rather than owning the book itself. This foundational difference combined with the challenges of still evolving technical format standardization, competing systems for digital-rights management and emerging business models introduced a substantially new environment for preservation service providers, 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 since the standard was created in the late 1960s, but as Scott and Orlikowski's (2021) study on the digital transformation of the book industry reveals, the ISBN system has started to show some serious limitations with not all books getting indexed in the system, and differing practices for its use among publishers for e-books. The study points out that the ISBN standard commonly requires that a separate ISBN be issued for every edition and format for a book, which means that PDF files, EPUB files and other formats all need their own ISBNs to be compliant. Digital files under different licensing or digital rights management restrictions also need their own individual ISBNs. The exception to all this necessary ISBN assignment is if an edition of a book, or part of a book, is available through one single source only and does not need to be processed into the broader book supply chain. The authors point out the fact that Amazon does not require books to have been assigned an ISBN to be sold on the Kindle book store to be a notable shift in the practice of ISBNs use in the e-book context. The next section will delve more particularly into how digital object identifiers (DOIs) work in this space, as they have grown into use in parallel with the emergence of digital publishing, further challenging the perception of ISBNs when it comes to unique identifiers for books.

The context of OA book publishing

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 a 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 a 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 *et al.*, 2022).

A key study in the context of scholarly OA books is Neylon *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 seven 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 seven 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 and 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.

Echoing many of the same observations with the challenges of OA book metadata as Neylon *et al.* (2018), the 2019 report "The State of Open Monographs" by Grimme *et al.* (2019) provides a helpful review of the historical and technical origins for why the data landscape for OA books is so fragmented. The report highlights the basic difficulty of even exactly knowing how many OA books are published, using a multi-method approach to come up with an approximation of around 80,000 for the year 2013.

Based on the earlier findings of a lack of standard practices for metadata for OA books, 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 the standardized use of persistent identifiers (Stone *et al.*, 2020). In the COPIM project, Barnes *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 the long-term preservation of the outputs of an institution varies and is not in any way an underlying requirement for running a repository (Francke *et al.*, 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 and Termens, 2022, p. 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 *et al.*, 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 *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 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 to get 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 *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 *et al.*, 2014), France (Derrot and Clément, 2014; Derrot *et al.*, 2014) and UK (Gooding *et al.*, 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 *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 the 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 *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 have 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 *et al.*, 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 book 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 the 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.

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 *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

Source(s): Created by authors

Table 1. Overview of bibliometric sources containing records of OA books

Data were 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 *et al.*, 2022), The Lens and OpenAire (Baglioni *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, which means that they include all types of OA under one single content metadata tag which is not optimal for research purposes like the one currently at hand. This study is primarily interested in OA books made available directly through the publisher or similar primary distributor, as that has arguably a connection to having a potential preservation coverage compared to, e.g. self-archived green OA copies available on the web. Not being able to filter out green OA is also a drawback in attempting to minimize the occurrence of thesis 'and dissertations ending up in the dataset since they are often hosted in repositories which are a location type almost exclusive to green OA. Nevertheless, the methodological choice was made to include the OA book content from The Lens and OpenAire since they are such substantial data sources, relying rather on post-hoc filtering out content based on content metadata tags identifying them as thesis and dissertation works. OpenAlex has the option to exclude Green OA/repository copies which were 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 to 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 are better than it has ever been and is constantly improving, there are some obstacles to 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, hence why that should only be considered as a last resort after primarily using unique identifier data. For performing the data cleaning and analysis Microsoft Excel version 16 was used. The exception was the OpenAire and OpenAlex datasets that first had to be imported into OpenRefine version 3.4 in order to parse the JSON data into spreadsheet format to become comparable to the rest of the data sources.

After the extraction of all data from the data sources, there were 750,136 items in total. Of which, 99% had a DOI/URL in the metadata; 6% had an ISBN number in the metadata; 0.2% had neither ISBN or DOI/URL. As the first step, all items with an identical DOI/URL or ISBN were deduplicated, which resulted in 430,745 entries remaining. The next step was deduplicating based on the exact title of the book, considering the first 250 characters of each title as per the restrictions in Microsoft Excel. This process resulted in a total unique record count of 396,995. As such the title matching was not a very impactful deduplication method relative to the unique identifier approach used in the previous step, but still offering some additional matches that would have otherwise remained in the dataset.

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 of which titles are reported to be preserved through a trusted preservation service. CLOCKSS (2022), Portico (2022), Global LOCKSS Network (2022) and Cariniana Network (Márdero Arellano and Abbud Grácio 2021) 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 provides an overview retrieved preservation data.

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

Source(s): Created by authors

National libraries have good data within them but programmatic access from outside is still limited. Barnes *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, 745,525 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 the 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, where 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 the verification of the OA status and classification of content. Due to the size of the dataset and emphasis on the 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 printed OA books 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 *et al.*, 2008). They are calculated to be area-proportional, however, due to the complexity of several overlapping datasets and the limitations of circular shapes that are 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 in Appendix for any exact calculations.

Results

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

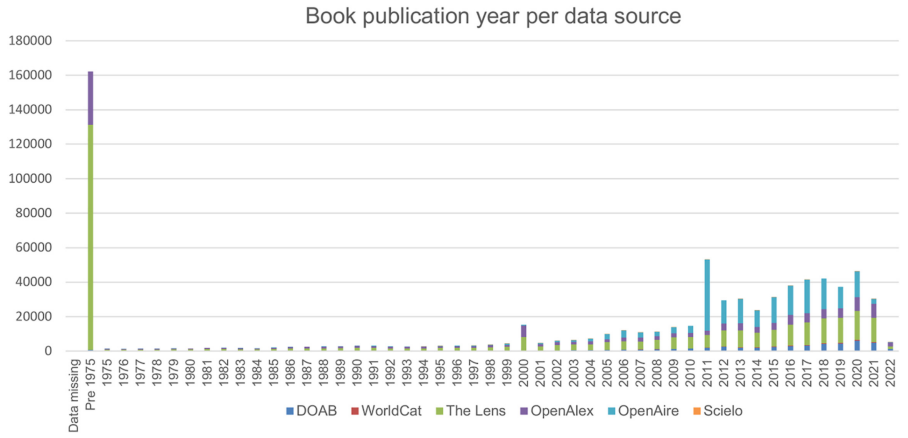
What is the current prevalence of OA books?

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 and 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 the 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 and 1999 with a consistent growth trend starting from the year 2000 onwards. Since the amount of older digitized materials (published earlier than 1975) is so large in proportion to the latter individual years, Figure 2 presents a zoomed-in view of publication years 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.

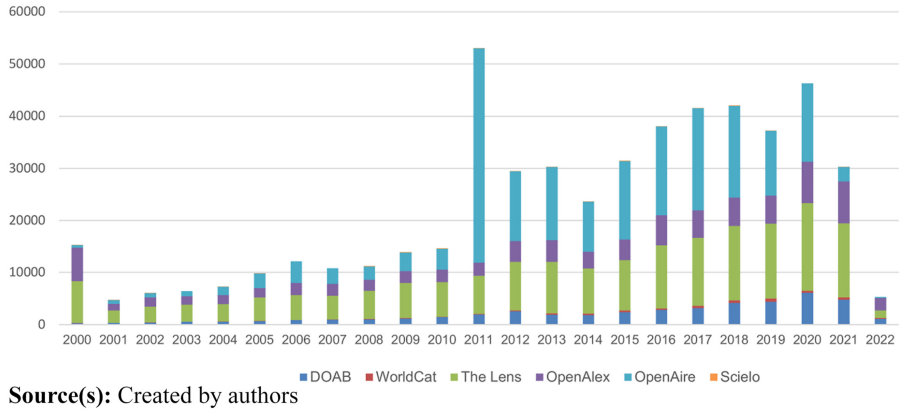
The next step for understanding OA book prevalence, and for informing future studies in the area, is to see how content is distributed across the different bibliometric datasets. Table 1 in the methods section already presented how many records each individual data source provided but that presentation did not analyze for overlap. Figure 3 presents the content distribution across the six bibliometric data sources with the overlap record counts across all data sources being available in Appendix.

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 750,136 to 396,995). As Figure 3 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% (3,280 records) of the final dataset being unique to OpenAlex. What is perhaps a bit surprising is that despite the DOAB data being considerably smaller



Source(s): Created by authors

Figure 1.
Publication years of OA book content from the six bibliometric sources (all years)



Source(s): Created by authors

Figure 2.
Publication years of OA book content from the six bibliometric sources (only year 2000 and onwards)

than the larger data sources included in the study, it still contributed 5% of the unique records for the final dataset. Based on this distribution, it would be hard to argue for selecting just one or two data sources as the basis of future studies in this space if maximum coverage of unique records is the target.

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 745,525 DOIs found in the four largest data sources used). The results are visualized as treemaps in Figure 4 with a summarizing breakdown of the top domains provided in Table 3.

While not providing exact numbers, Figure 4 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 three web domains for three of the four

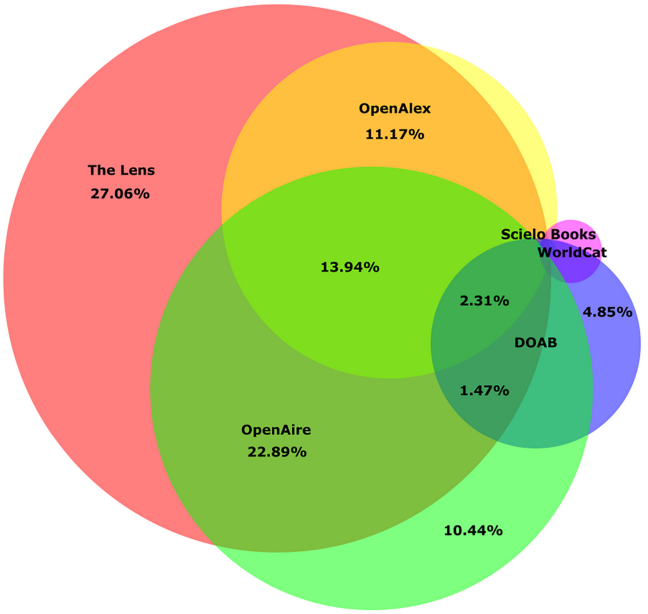


Figure 3.
Content distribution across the six bibliometric data sources

Note(s): Shares under 1% are not marked with a numeric label
Source(s): Created by authors

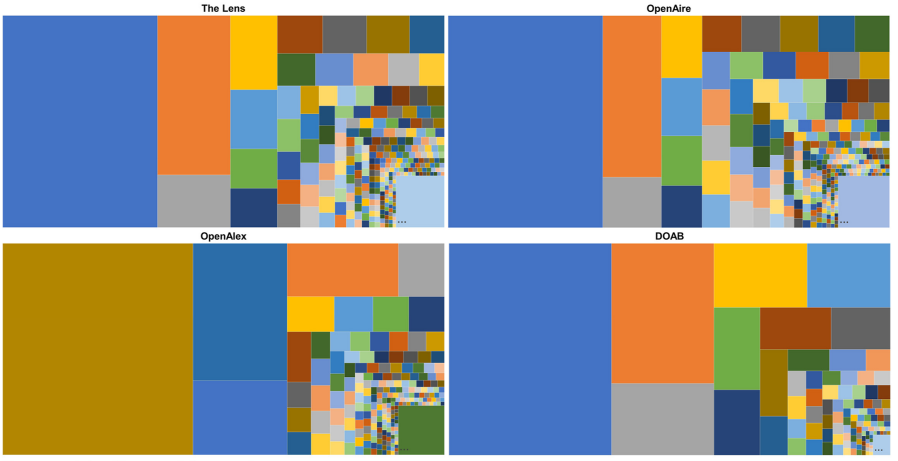


Figure 4.
Visual treemap representation of the distribution of content on unique web domains for records with DOIs retrieved from these sources

Source(s): Created by authors

data sources. biodiversitylibrary.org contains a large volume of digitized materials, with a lot of it being rich in images as it focuses on natural history literature and archival material. The prominence of link.springer.com in particularly The Lens is both due to Springer being one of the larger book and OA book publishers in the world, but also because some of the materials

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

DOAB	Count	OpenAire	Count	OpenAlex	Count	The lens	Count
library.oapen.org	18,889	biodiversitylibrary.org	74,133	biodiversitylibrary.org	23,347	biodiversitylibrary.org	121,956
books.openedition.org	7,889	link.springer.com	21,466	afghandata.org	19,840	link.springer.com	43,261
mdpi.com	4,023	elibrary.worldbank.org	6,666	link.springer.com	8,174	law.acku.edu.af	14,183
mts.intechopen.com	3,274	degruyter.com	5,747	law.acku.edu.af	5,562	afghandata.org	12,944
frontiersin.org	2,930	cambridge.org	5,356	books.openedition.org	4,566	onlinelibrary.wiley.com	10,323
intechopen.com	2,092	classiques.uqac.ca	4,583	library.si.edu	4,237	elibrary.worldbank.org	6,910
degruyter.com	1,652	books.openedition.org	3,856	classiques.uqac.ca	4,193	classiques.uqac.ca	6,759
ksp.kit.edu	1,647	taylorfrancis.com	3,239	repository.usta.edu.co	4,031	ieeexplore.ieee.org	6,416
media.fupress.com	1,371	library.si.edu	3,175	openknowledge.worldbank.org	2,036	degruyter.com	6,241
books.scielo.org	1,009	apps.crossref.org	3,168	constellation.uqac.ca	1,950	dl.acm.org	6,078
omp.zrc-sazu.si	591	mr.crossref.org	3,002	vr-elibrary.de	1,837	journals.openedition.org	4,922
ucdigitalis.uc.pt	494	repository.usta.edu.co	2,854	darchive.mblwhoilibrary.org	1,721	taylorfrancis.com	4,569
nomos-elibrary.de	429	vr-elibrary.de	2,350	press.umich.edu	1,692	apps.crossref.org	4,518
edp-open.org	288	Oxford.universitypressscholarship.com	2,277	apps.crossref.org	1,634	repository.si.edu	4,193
link.springer.com	252	press.umich.edu	2,203	mohrsiebeck.com	1,445	repository.usta.edu.co	3,702
ledizioni.it	228	rand.org	2,109	books.fupress.com	1,353	mdpi.com	3,007
bloomsburycollections.com	193	worldscientific.com	2,077	liu.diva-portal.org	1,294	jstor.org	2,855
e-archivo.uc3m.es	170	darchive.mblwhoilibrary.org	2,028	jstor.org	1,279	deepblue.lib.umich.edu	2,819
api.intechopen.com	162	constellation.uqac.ca	2,026	rand.org	1,230	academic.oup.com	2,410
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	

Source(s): Created by authors

are separated as individual chapters since in some cases every chapter has been assigned a separate DOI, but at still picked up as entire books by the search function in The Lens. 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 the 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 1,453–1,816 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 a 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.

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 were represented in the various preservation services based on ISBN and/or book title matching.

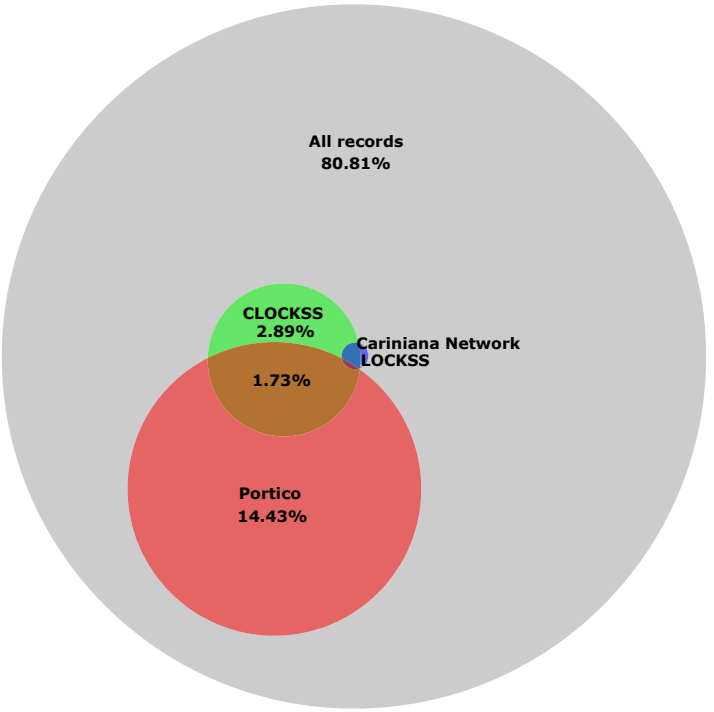
Table 4 illustrates that OA book content listed in DOAB is covered to the highest degree by at least one of the services (46% of all DOAB records) with WorldCat (33%), OpenAire (25%), OpenAlex (13%), The Lens (10%) and Scielo Books (9%) following in descending order. Among the preservation service providers, 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 in records in the Global LOCKSS Network and the Cariniana Network preservation data, ranging between 0% and 1% depending on the OA book data source. The high share of preservation coverage in DOAB is largely influenced by the collaboration between OAPEN and Portico, since the web domain analysis revealed that OAPEN (library.oapen.org) hosted 38% of all OA books in DOAB.

As a secondary perspective on the preservation coverage, Figure 5 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 the preservation of the deduplicated dataset based on the data sources utilized and compared to each other, which is 19%.

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(s): 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
Source(s): Created by authors



Source(s): Created by authors

Figure 5.
Preservation coverage for the 396,995 OA book records

Discussion

The main finding of the study is that, based on an 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. What came as something of a surprise was the quite prominent division in the population of OA books having a publication date earlier than 1975, and the rest of the higher annual volumes of materials being published from the year 2000 onwards and in particular from 2010 onwards. The circumstances for creating and preserving older digitized materials are arguably different to that of born-digital materials published directly and only as OA on the web, which suggests that future inquiries into this space should take this finding into account when designing data collection methods and drawing conclusions from them. Among the used data sources The Lens was particularly strong in providing unique records for materials published earlier than 1975, while OpenAire showed its particular strength with regard to unique materials from 2010 onwards.

Among individual commercial publishers, Springer Nature stood out as the largest publisher of OA books, followed by De Gruyter, MDPI and Frontiers. These are familiar names that have also had a strong presence in the rapidly growing market of OA publishing in the context of scholarly journals (see e.g. Rodrigues *et al.*, 2020). It bears noting that the figures for some of these large publishers in particular would warrant dedicated investigation in the future to better understand their actual strength of presence in the OA book space. From checking out some random entries in the dataset MDPI and Frontiers book DOIs often lead to edited research topics and special issues of their journals, something that many might

perceive challenge the definition of a book. Whether the fault is at Springer Nature or the content aggregators misclassifying the content is unknown, but there seems to be a number of individual chapters and conference proceedings inflating the actual number of books for this particular publisher. Table 3 which displayed the frequency of domains, where OA books are hosted revealed that also university presses rank relatively high, which aligns well with them having been an early-mover in the OA book space. No national library showed up among the top frequencies of domains. The earlier literature review revealed that several national libraries around the world are also including OA books in their collections; however, these findings suggest that the low discoverability and limited access to these items outside library premises or networks are still not having a notable influence on the broader picture of how much content is visible on the open searchable web.

As summarized earlier in the article, Romano (2003) pointed out two decades ago that there was an ongoing uncertainty with regard to the roles and responsibilities publishers and libraries should play when it comes to e-book preservation. While this seems to still be something not fully resolved, the solutions that have emerged and evolved since then have worked towards closing this gap as digital materials have become commonplace items. What this study can contribute to this space is the consideration of what role intermediary content aggregators and bibliometric databases could take on or help with to even further minimize the cracks that content might otherwise fall into. Neither individual publishers nor individual libraries alone are as focused in their main tasks on discovering and indexing content made available openly on the web as these intermediaries, and the information they maintain and grow would likely be helpful for whatever collaboration constellations emerge to preserve such materials bottom-up based on what is already out there. Such an approach would be quite a close comparison to what is currently ongoing with Project JASPER for the preservation of scholarly journals (DOAJ.org 2021).

While 396,995 is an exact number, 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 are 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 *et al.* (2018) and Grimme *et al.* (2019), 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 as 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 the 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 to users when querying the growing amount of content that gets indexed in these services.

The gaps left by the varying practices for the usage of unique identifiers for content are something that would need to be remedied in order for data matching to be more comprehensive and 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, where a match is

not found due to a missing or additional character in the book title. 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. Overall, what the future holds for the realm of unique identifiers in the context of OA books is not clear at the moment. Drawing on Scott and Orlikowski's (2021) study on the challenges that the ISBN system has faced as book publishing has been introduced to digital publishing possibilities, OA books present a particular challenge as they can also exist as print books (necessitating ISBNs, and separate identifiers for all editions) but can just as well be digital only with a single license at the publishers webpage in a single format (thus having ISBN issuing as something optional to implement). With both multiple ISBNs and no ISBN being valid scenarios for OA books bibliometric researchers as well as preservation providers will have to continue to figure out different ways to overcome this methodological challenge.

Recommendations

How should collaboration evolve among major stakeholders (e.g. publishers, libraries and 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 *et al.*, 2021). Preservation is through its inherent actions preparing for a future state in time, that in the 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. and Manghi, P. (2022), *Books from the OpenAIRE Research Graph (1.0) [Data Set]*, Zenodo, doi: 10.5281/zenodo.6619395.

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

Barrueco and Termens (2022), "Digital preservation in institutional repositories: a systematic literature review", *Digital Library Perspectives*, Vol. 38 No. 2, pp. 161-174, doi: 10.1108/DLP-02-2021-0011.

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

blog.dshr.org (2022), "Where did the number 3 come from? David rosenthals blog", available at: 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. and 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, available at: https://web.archive.org/web/20221006131808/https://az659834.vo.msecnd.net/eventsairwesteuprod/production-inconference-public/3f4dd08cbb3842739c82ccac5a422de0

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, available at: https://web.archive.org/web/20180107015602/https://www.clir.org/wp-content/uploads/sites/6/2016/09/pub106.pdf

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

coalition-s.org (2021), "cOAlition S statement on Open Access for academic books", available at: https://www.coalition-s.org/coalition-s-statement-on-open-access-for-academic-books/

cordis.europa.eu (2023), "PALOMERA - policy alignment of open access monographs in the European research area", available at: https://cordis.europa.eu/project/id/101094270

coretrustseal.org (n.d), "CoreTrust seal", available at: https://www.coretrustseal.org/

Derrot, S. and 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*, Lyon, 16-22 August 2014, available at: https://web.archive.org/web/20220121090010/https://library.ifla.org/id/eprint/830/1/087-derrot-en.pdf

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

Doabooks.org (2022), "Building stronger infrastructures to support open access books: LYRASIS, DOAB and OAPEN", available at: 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", available at: https://web.archive.org/web/20210916132815/https://doaj.org/preservation/

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

Francke, H., Gamalielsson, J. and Lundell, B. (2017), "Institutional repositories as infrastructures for long-term preservation", *Information Research*, Vol. 22 No. 2, 757.

Global LOCKSS Network (2022), available at: https://reports.lockss.org/keepers/keepers-LOCKSS-books-report.csv

Gooding, P., Terras, M. and Berube, L. (2021), "Identifying the future direction of legal deposit in the United Kingdom: the Digital Library Futures approach", *Journal of Documentation*, Vol. 77 No. 5, pp. 1154-1172, doi: 10.1108/jd-09-2020-0159.

Grimme, S., Taylor, M., Elliott, M.A., Holland, C., Potter, P. and Watkinson, C. (2019), "The state of open monographs (version 4)", *Digital Science*. doi: 10.6084/m9.figshare.8197625.v4.

Haider, J., Johansson, V. and Hammarfelt, B. (2021), "Time and temporality in library and information science", *Journal of Documentation*, Vol. 78 No. 1, pp. 1-17, doi: 10.1108/jd-09-2021-0171.

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

IIPC (n.d), "International Internet preservation Consortium – legal deposit", available at: https://web.archive.org/web/20221007095441/https://netpreserve.org/web-archiving/legal-deposit/ (accessed 7 October 2022).

Kirchhof, M., Morrissey, S. (2014), "Preserving eBooks. DPC technology watch report 14-01 June 2014. Digital preservation coalition", available at: https://web.archive.org/web/20211010182422/https://www.dpconline.org/docs/technology-watch-reports/1230-dpctw14-01/file. Video recording, available at: 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" (1.0) [Data set], Zenodo, doi: 10.5281/zenodo.7305477.

Laakso, M., Matthias, L. and Jahn, N. (2021), "Open is not forever: a study of vanished open access journals", *Journal of the Association for Information Science and Technology*, Vol. 72, pp. 1099-1112, 2021, doi: 10.1002/asi.24460.

Lin, D., Crabtree, J., Dillo, I., Downs, R.R., Edmunds, R., Giaretta, D., . . . and Westbrook, J. (2020), "The TRUST Principles for digital repositories", *Scientific Data*, Vol. 7 No. 1, pp. 1-5, doi: 10.1038/s41597-020-0486-7.

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

Márdero Arellano, M.A. and Abbud Grácio, J.C. (2021), "The cariniana network for digital preservation", The Digital Preservation Coalition, available at: https://web.archive.org/web/20220127073614/https://www.dpconline.org/blog/wdpd/cariniana-wdpd21 (accessed 8 October 2022).

Muir, A. (2001), "Legal deposit and preservation of digital publications: a review of research and development activity", *Journal of Documentation*, Vol. 57 No. 5, pp. 652-682, doi: 10.1108/eum0000000007097.

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

Portico (2022), available at: 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. and Orr, R. (2022), "OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts", *arXiv Preprint arXiv:2205.01833*. doi: 10.48550/arXiv.2205.01833.

Rodrigues, R.S., Abadal, E. and de Araújo, B.K.H. (2020), "Open access publishers: the new players", *PLOS ONE*, Vol. 15 No. 6, e0233432, doi: 10.1371/journal.pone.02334.

Romano, F. (2003), "E-Books and the challenge of preservation", *Microform and Imaging Review*, Vol. 32 No. 1, pp. 13-25, doi: 10.1515/mfir.2003.13.

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. and 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*, Law Library of (United States) Congress, Global Legal Research Center, Spain, p. 78, July 2018, available at: https://www.loc.gov/law/help/digital-legal-deposit/digital-legal-deposit.pdf

Scott, S. and Orlikowski, W. (2021), "The digital undertow: how the corollary effects of digital transformation affect industry standards", *Information Systems Research*, Vol. 33 No. 1, pp. 311-336, doi: 10.1287/isre.2021.1056.

Snijder, R. (2019), *The Deliverance of Open Access Books: Examining Usage and Dissemination*, Amsterdam University Press, doi: 10.26530/OAPEN_1004809.

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

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

Taylor, M. (2019), "Mapping the", *Publishing Challenges for an Open Access University Press. Publications*, Vol. 7 No. 4, p. 63, doi: 10.3390/publications704006.

Wei, D., Ji, S. and 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, Lyon, 16-22 August 2014, available at: https://web.archive.org/web/20200319233833/http://library.ifla.org/859/1/087-wei-en.pdf

Appendix

Table A1.
Numbers relating to Figure 3

Groups	Number of titles
The Lens	107,444
The Lens ∩ OpenAire	90,876
The Lens ∩ OpenAire ∩ DOAB	5,855
The Lens ∩ OpenAire ∩ DOAB ∩ OpenAlex	9,184
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	55,355
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	5,576
The Lens ∩ DOAB ∩ OpenAlex	2,567
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	1,668
OpenAire ∩ DOAB ∩ OpenAlex	333
OpenAire ∩ DOAB ∩ OpenAlex ∩ WorldCat	17
OpenAire ∩ DOAB ∩ OpenAlex ∩ WorldCat	12
OpenAire ∩ DOAB ∩ WorldCat	66
OpenAire ∩ DOAB ∩ WorldCat	1
OpenAire ∩ OpenAlex	5,005
OpenAire ∩ OpenAlex ∩ WorldCat	8
OpenAire ∩ WorldCat	39
DOAB	19,263
DOAB ∩ OpenAlex	142
DOAB ∩ OpenAlex ∩ WorldCat	3
DOAB ∩ OpenAlex ∩ WorldCat	7
DOAB ∩ WorldCat	1,853
DOAB ∩ WorldCat	81
OpenAlex	3,280
OpenAlex ∩ WorldCat	30
OpenAlex ∩ WorldCat	1
WorldCat	1,434
Scielo	1

Groups	Number of titles
Portico ∩ CLOCKSS ∩ LOCKSS ∩ Cariniana Network	1
Portico ∩ CLOCKSS ∩ LOCKSS	134
Portico ∩ CLOCKSS	11,493
Portico ∩ LOCKSS	139
Portico	57,286
CLOCKSS ∩ LOCKSS	237
CLOCKSS	6,853
LOCKSS	55
Cariniana Network	4
Not included in any service	320,793

Table A2.
Numbers relating to Figure 5

Corresponding author
Mikael Laakso can be contacted at: mikael.laakso@hanken.fi