Peering into the jungle

Challenges in determining preservation status of open access books

|  |  |  |
| --- | --- | --- |
| **Mikael Laakso** | **Alicia Wise** | **Ronald Snijder** |
| *Hanken School of Economics*  *Helsinki, Finland*  *mikael.laakso@hanken.fi*  *0000-0003-3951-7990* | *CLOCKSS*  *California, USA*  *awise@clockss.org*  *0000-0002-7898-3428* | *OAPEN Foundation*  *The Hague, The Netherlands*  *r.snijder@oapen.org*  *0000-0001-9260-4941* |
|  |  |  |

**Abstract – This paper reports on some of the initial observations from an ongoing study focused on determining the preservation status of academic open access books. The central challenges discussed revolve around lack of common definitions, metadata, and established practices for openly recording preservation status for books.**

**Keywords – open access, books, monographs, preservation**

**Conference Topic – Resilience**

# Introduction

Making academic content openly available for everyone using the web has never been easier from a technical and financial cost 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 works like monographs, book chapters, individual article manuscripts, and entire journals are sprawling up at an unprecedented pace which has led to a rapidly increasing volume of academic content available out in the open. While the act of making something openly available provides open access (OA) to the content for the moment, the practices for ensuring preservation to such content for the long-term are still developing, and to a degree unknown. 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[1,2].

A recent study gauged the degree to which content from OA journals has 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 [3]. Partly inspired by the findings of this study Project JASPER (JournAlS are Preserved forevER) was initiated which is a collaboration between CLOCKSS, DOAJ, The Internet Archive, The Keepers Registry, and PKP [4]. There is currently no similar overview of materials lost or at risk of being lost due to lacking preservation coverage concerning OA books. As there is growing momentum for advancing OA to academic books through science policy it would be important to scope the landscape through a systemic study to map the current preservation status of published materials.

# Towards better knowledge about preservation coverage

The aim of an ongoing study is to conduct a data-driven mapping of the current landscape of preservation within the content domain of OA books. The focus of the study is on academic monographs and edited books that are or have been available OA. Excluded are non-published theses and dissertations, and individual book chapters. The definition of at-risk materials is lack of preservation inclusion in a preservation service e.g. Portico, CLOCKSS, or other similar recognized infrastructure. This study is not focused on issues related to specific file formats of preservation, merely that an indication of some preservation exists for a specific title.

Already from the outset it was known that the data collection circumstances for vanished and currently online OA book content differs significantly from that of scholarly journals. Laakso, Matthias & Jahn (2021) utilized mainly past and present journal lists provided by journal indexing services to identify potentially vanished journals, and verified the preservation status through information from the Keepers Registry and Internet Archive snapshots of the last known URL. For OA books the situation is more fragmented due to the lack of comprehensive international services for content indexation, and for registering preservation inclusion across service providers.

An additional component in this ongoing study will be to figure out what domains host the OA book content, by checking which URLs their DOIs or full text links point to. This is not a way of verifying preservation, but such an exploration can shed light on what the long tail for content providers looks like and potentially what type of organizations are running them if it can be derived from the domain names.

# Challenges observed so far

## Definitions: When is a book an academic book, and when is it open access?

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 is often a lot of thesis´, 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 only metadata, but also overall transparency and knowledge available about what kind of editorial processes are behind published works.

Ambiguity is also introduced by the concept of OA, as some sources allow filtering to content available in full text for free (without any distinction between OA types), some do not have OA filtering at all, and some have very granular metadata concerning OA metadata. The circumstances for preservation are different if the content is available in a document repository in manuscript form compared to the publisher’s website to which also the publications DOI also points to. Barnes, Bell & Cole et al [2] found out through their interviews with stakeholders in the landscape there are some publishers that upload their published content to local repositories, but if that archival counts as preservation depends on the policies and precautions of the institution running the service, which makes gauging the viability of such practices hard at scale. There has been a lot of progress in this area but there is still work to be done with it comes to reliably filtering OA content across key services.

## *Data management: Physical extraction of metadata to represent the “global bookshelf” of academic OA books*

The amount and quality of freely available metadata describing publications has never been as good as it is now. However, the growing size of increasingly detailed and comprehensive metadata comes at the price of data size (and to some degree data precision, as some other mentioned challenges point out).

Slicing out book metadata from some of the widest openly available bibliometric datasets in the world (e.g. Crossref, OpenAlex, OpenAIRE) requires either downloading the entire datasets which are often in the 100s of gigabytes uncompressed, mapping the JSON files to a database, and designing queries to extract the wanted data concerning books contained in the data. Many bibliometric datasets are becoming challenging to process locally even on a modern desktop computer since they do not fit to be processed into available computer working memory. What is of interest for the purposes of book preservation information is relatively small, but extracting it often requires dealing with the entire dataset at the outset which limits accessibility.

The services mentioned above also offer API access which means that they can be queried programmatically for extraction of specific records. This requires some familiarity with programming or setting up scripts to send requests for multiple chained queries as only a limited number of records are given as response per request. Unfortunately, some API services like OpenAIRE do not allow queries to be filtered to only books, making that path unviable for book-related queries. Crossref on the other hand has no reliable way to extract only items available OA.

## Unique identifiers: Taming the wilderness of identifier metadata describing OA books

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. Matching by title or author is not reliable due to even small differences in spelling, format and punctuation leading to incorrect matches. There is varying use of unique identifiers for books, where ISBNs might be the only value available in one dataset (e.g. WorldCat) or but not available for any entries in another service (e.g. Lense) where DOIs are the key identifier used. Many services also have their own unique identifiers for entries but these are of little use when the data is to be aggregated with data from elsewhere. Table 1 provides a coarse overview of some key data sources, their estimated volume of OA books together with unique identifier availability in the metadata of the records.

TABLE 1

Overview of bibliometric sources containing records of OA books

|  |  |  |
| --- | --- | --- |
| Service | Scope of OA book content | Unique identifier availability in metadata |
| OpenAlex | 4 545 046 indexed objects of type “Book”, 203 857 objects of type “Monograph” + OA | ISBN = 0%  DOI = Not yet measured, but high share |
| Crossref | **Works of types book or monograph** 328 098 that have license information and link to full-text (not necessarily OA) | ISBN = Not yet measured, but high share  DOI = Not yet measured, but high share |
| WorldCat (OCLC) | **4 597 non-fiction e-books tagged as OA** | ISBN = 100%  DOI = 0% |
| DOAB | 55 723 academic peer-reviewed books, all OA | ISBN = 86%  DOI = 83% |
| Scielo Books | 1564 complete titles of which 963 are OA | ISBN = 100%  DOI = 93% |
| Lense | 348 267 records under “Open Access” and “Book” published between year 0 and 2050. | ISBN = 0%  DOI = 99% |
| OpenAIRE | 211 749 records under “Open Access” and “Books” after removal of individual chapters, thesis, reports, and preprints. | ISBN = 0%  DOI = 99% |

## Openly available preservation data: Preservation data is scarce for all but the largest service providers, and even their datasets could be improved

The challenges mentioned so far have concerned creating a comprehensive dataset of OA books, but none of the data so far is capable of providing indication for which titles are reported to be preserved through some service. CLOCKSS [5], Portico [6], and Global LOCKSS Network [7] all provide open datasets that describe which books they have included in their coverage. None of these three provide DOI´s for their records, only ISBNs which is not optimal as most of the major bibliometric service providers focusing on OA book content rely on DOIs.

National libraries have good data within them but programmatic access from outside is still limited. Barnes, Bell & Cole et al [2] 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 libraries around the world are not easy to query programmatically from the outside.

## Building a path forward

With all these intertwining challenges present, pinning down the status for preservation of OA books is not a straightforward process and will even under optimal circumstances be an estimate rather than absolute and comprehensive as the definitions and practices in the landscape are still emerging. Below are some observations that could help shape the path forward for a more transparent preservation landscape for OA books.

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. Early experiments have shown promise in fuzzy matching of book titles to preservation records based on combinations of author information, book title, and publisher. However, the approach needs to be assessed more extensively but in cases where direct matching does not garner results such an approach might show utility as long as the number of false positive matches can be contained.

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 much needed for books as well, where preservation service providers could automatically report which titles they include in their coverage.

# References

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

[2] 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

[3] 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>

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

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

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

[7] Global LOCKSS Network (2022) https://reports.lockss.org/keepers/keepers-LOCKSS-books-report.csv