Researching Research

Mikael Laakso, D.Sc. (Econ.)
Associate Professor, Information Systems Science
Hanken School of Economics, Helsinki, Finland
@mikaellaakso


**Background**

- Research for the last 15 years has been on various aspects of open science, and in particular open access publishing
- Chair of the Association for Scholarly Publishing in Finland
- Have taken part in expert groups by the EC and ERCEA
- Member of the steering group for the Finnish national library consortia, FinElib
- Was member of the Finnish national open science steering group when the first national OA policy was set in 2020
A disclaimer

» My opinions and statements are not representative of any particular group of researchers or any organisation, they are solely my own.
Agenda

1. The interesting world of meta-research
2. Bibliometrics and Scientometrics as Data Science
3. The evolving data environment of meta-research
4. Specific software available for supporting workflows
5. Some examples taken from my own work

» Questions & answers
» Recommended readings
1. The interesting world of meta-research
Scholarly publishing in numbers

» The global scholarly publishing market annual revenue is currently valued to around 26 billion euro annually

» There are over 70 000 academic journals publishing millions of articles annually

» Growth in journal article content around 4%-5% annually

» Most revenue is generated in the United States, but China has risen to be the most prolific producer of publishable research output

The relationship between research and money is super interesting

https://www.nature.com/articles/495426a: Brendan Monroe
Open APC – One window into the money flows supporting academic publishing

https://treemaps.openapc.net/apcdataset/combined/#publisher/period=2022
The big publishers have constantly gotten bigger

Larivière et al. (2015) [https://doi.org/10.1371/journal.pone.0127502]
2. Bibliometrics and Scientometrics as Data Science
The term was first used in 1969, then referring to “Statistical Bibliography”, which had been an emerging field since the 1920s where the relationships among scientific papers, numbers of patents, amounts of experts, and other quantities had been explored.

For a long time information about publications and their cross-citations were not centrally available anywhere, but Eugene Garfield founded the Science Citation Index in 1964 which opened up a new era in research on research.

Initially intended as a tool for libraries to better keep track of what to subscribe to, but has become something else over time.

Broadus (1987) https://doi.org/10.1007/BF02016680
Garfield (1964) https://doi.org/10.1126/science.144.3619.649
“Scientometrics is the study of the quantitative aspects of the process of science as a communication system. It is centrally, but not only, concerned with the analysis of citations in the academic literature. In recent years it has come to play a major role in the measurement and evaluation of research performance.“

Basically a broader term than Bibliometrics, that also includes Bibliometrics as one element.

In the beginning when this term was firstly coined (1971) the possibilities for expansion were still limited, but the digital environment has opened up so many new possibilities for inquiry.

European scholarly journals from small- and mid-size publishers: mapping journals and public funding mechanisms

Mikael Laakso1,2 and Anna-Maja Mutiaa

Abstract
This study investigates the relationship between scholarly journal publishing and public funding, specifically concerning the context of small- and mid-size publishers (SMPs). As part of the scholarly communication landscape, SMPs publish a variety of research outputs, and their role is increasingly acknowledged. SMPs are often located in North- and South-East Asia, and the study explores the main sources of public funding for their journals. The results show that SMEs are not only dependent on public funding, but also that the type of funding varies across different countries. The study also highlights the importance of specific funding mechanisms that are unique to particular regions, which can provide valuable insights for those interested in scholarly communication.

Keywords: scholarly journals, public funding, scholarly communication

1. Introduction
The scholarly journal publishing sector has faced several important changes over the last few decades. The first of these is the proliferation of digital and digital content delivery, which is changing the way journals and publishers are able to communicate and share information. The second change is related to the number of publishers, which has increased from small and mid-sized publishers (SMPs) to large publishers. This has led to an increased emphasis on the role of journals in the scholarly communication landscape and how they are funded.

2. Methodology
This study examines the relationship between scholarly journal publishing and public funding, specifically concerning the context of small- and mid-size publishers. The study focuses on journals published in China, Japan, South Korea, Taiwan, and Singapore. These countries are chosen because they are home to some of the largest SMPs in the world. The study examines the main sources of public funding for their journals and how these sources vary across different regions.

3. Results
The results show that SMEs are not only dependent on public funding, but also that the type of funding varies across different countries. The study also highlights the importance of specific funding mechanisms that are unique to particular regions, which can provide valuable insights for those interested in scholarly communication.

4. Conclusion
This study provides valuable insights for those interested in scholarly communication and how it is funded. The results show that SMEs are not only dependent on public funding, but also that the type of funding varies across different countries. The study also highlights the importance of specific funding mechanisms that are unique to particular regions, which can provide valuable insights for those interested in scholarly communication.

References


An increasing number of journals also make article processing history and peer-review reports open and public

Peer Review reports

<table>
<thead>
<tr>
<th>Original Submission</th>
<th>25 Apr 2020</th>
<th>Submitted</th>
<th>Original manuscript</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15 May 2020</td>
<td>Reviewed</td>
<td>Reviewer Report</td>
</tr>
<tr>
<td></td>
<td>22 May 2020</td>
<td>Reviewed</td>
<td>Reviewer Report</td>
</tr>
<tr>
<td></td>
<td>27 May 2020</td>
<td>Author responded</td>
<td>Author comments - Feixiong Cheng</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resubmission - Version 2</th>
<th>27 May 2020</th>
<th>Submitted</th>
<th>Manuscript version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>29 May 2020</td>
<td>Reviewed</td>
<td>Reviewer Report</td>
</tr>
<tr>
<td></td>
<td>15 Jun 2020</td>
<td>Reviewed</td>
<td>Reviewer Report</td>
</tr>
<tr>
<td></td>
<td>18 Jun 2020</td>
<td>Author responded</td>
<td>Author comments - Yuan Hou</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resubmission - Version 3</th>
<th>18 Jun 2020</th>
<th>Submitted</th>
<th>Manuscript version 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>22 Jun 2020</td>
<td>Reviewed</td>
<td>Reviewer Report</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resubmission - Version 4</th>
<th>Submitted</th>
<th>Manuscript version 4</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Publishing</th>
<th>22 Jun 2020</th>
<th>Editorialy accepted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15 Jul 2020</td>
<td>Article published</td>
</tr>
</tbody>
</table>

You can find further information about peer review here.

3. The evolving data environment of meta-research
In addition to recording citations the digital environment has enabled that other types of activity around a published articles is also tracked, including:

- **Views** (HTML views and PDF downloads)
- **Discussions** (mentions in the news, social media, wikipedia etc)
- **Bookmarks** (how often the content has been bookmarked on various social media for researchers)
- **Recommendations** (how often the content has been recommended on various social media for researchers)
Despite a lot of advancement, there are still a lot of gaps and problems with the information environment.

» Readily available data about scholarly publishing is not of just relevance to bibliometric research – it would help many actors in their tasks.

» Despite journals being dominantly digital and web-based, comprehensive record keeping and monitoring of outlets and their outputs still leaves room for improvement.
Three persistent obstacles

1. Commercial dominance

- Access to the most comprehensive commercial databases is limited, and datasets created on the basis of such proprietary data can rarely be freely redistributed in their most usable form.
Three key obstacles in current journal indexing services (cont.)

2. Amnesia

- Current bibliometric databases focus primarily on snapshots of results, **they are not designed to deliver time-series data that would account for classification and status changes of individual journal/article metadata.**
Three key obstacles in current journal indexing services (cont.)

3. Selective coverage

Each bibliometric database comes with its own biases and limitations in how comprehensively journals across disciplines, countries, and languages are selected for inclusion.
Various indexes/databases to choose from, all with different implications

» Scopus
» Web of Science
» Dimensions
» (Microsoft Academic)
» The Lens
» Ulrichsweb
» Crossref/DOI

» DOAJ
» ROAD
» Google Scholar
» National research databases
Open Access is constantly evolving

Open Science

The Needs of Research
Technology Development
Financial Aspects
Science Policy
Open Access

“Open access (OA) literature is digital, online, free of charge, and free of most copyright and licensing restrictions.”

(Peter Suber, 2012:4)

Gold OA

Open Access made available by journals themselves (either in full or part). Free for everyone or enabled by author-side payment.

Green OA

Open Access elsewhere on the web. Often manuscript-versions of published journal articles. Free to authors.
What open access looks like to most web users

Google search for "fish consumption"

Scholar search results:

1. Fish consumption, fish oils, and cardiovascular events: still waiting for definitive evidence
   PM Ridker, L Ailaitre, C Couture, F Leclerc, M Charest, A Marin, J Lepine, T Tailbot, D Tchernof, A Lamarche. A randomized, crossover, head-to-head comparison of eicosapentaenoic acid and docosahexaenoic acid supplementation to reduce inflammation markers in men and women.

Related articles
- All 2 versions
- Cite
- Save

Trends in blood mercury concentrations and fish consumption among US women of reproductive age, NHANES, 1999–2010
   RJ Birch, J Bigler, JW Rogers, Y Zhuang, ..., - Environmental ..., 2014 - Elsevier
   Background Consumption of finfish and shellfish is the primary exposure pathway of methylmercury (MeHg) in the US. MeHg exposure in utero is associated with neurodevelopmental and motor function deficits. Regulations and fish advisories may be needed to protect vulnerable populations.

Cited by 26
- Related articles
- All 9 versions
- Cite
- Save

No association between fish consumption and risk of stroke in the Spanish cohort of the European Prospective Investigation into Cancer and Nutrition (EPIC-Spain): a ..., P Amiano, S Chamosa, N Estebanezt ..., - Public health ..., 2016 - Cambridge Univ Press
   Objective To prospectively assess the associations between lean fish, fatty fish and total fish intake and risk of stroke in the Spanish cohort of the European Prospective Investigation into Cancer and Nutrition (EPIC-Spain). Design Fish intake was estimated from a validated food-frequency questionnaire.

Cited by 25
- Related articles
- All 6 versions
- Cite
- Save

Regular fish consumption and age-related brain gray matter loss
   CA Raji, KJ Eriksson, OL Lopez, LH Kuller, ..., - American journal of ..., 2014 - Elsevier
   Background Brain health may be affected by modifiable lifestyle factors; consuming fish and antioxidative omega-3 fatty acids may reduce brain structural abnormality risk. Purpose To determine whether dietary fish consumption is related to brain structural integrity among older adults.

Cited by 34
- Related articles
- All 10 versions
- Cite
- Save

[HTML] infona.pl

[PDF] cambridge.org

[HTML] nih.gov
Web services built upon and enhanced by more open metadata APIs and/or open access
Some things keeping me up at night

What is considered open access?

» Strict definition (incl.) license requirement
» Basic requirement of free access?
» Available by any means?
» How to consider or adjust for embargos?
Journal vs Article perspectives

- A complicated relationship
- Partial openness of journals
- Journals can and have disappeared, merged, changed OA model, some articles might still be available online elsewhere.
Different types of data is created throughout the research process.
Crossref – The “master data” of research publications

Non-profit founded in the early 2000s that has become the largest and most authoritative curator of metadata concerning scholarly publications.

Largest issuer of “DOIs” (Digital Object Identifiers).

Has both free and premium APIs that one can use to query publication records both for research and creation of ancillary services.

Since September 2023 also includes metadata about articles being retracted.

https://www.crossref.org/
https://prep.labs.crossref.org/
OpenAlex – The most comprehensive database that augments and expands upon Crossref data

“Inspired by the ancient Library of Alexandria, OpenAlex is an index of hundreds of millions of interconnected entities across the global research system. We're 100% free and open source, and offer access via a web interface, API, and database snapshot.”

- **246M** Works
  - 50M Open Access works
  - 27M from the Global South
  - 3M datasets

- **93M** Authors
  - 5M with ORCID IDs
  - 12M from the Global South

- **248K** Sources
  - 42K that are Open Access

- **10K** Publishers

- **32K** Funders

- **107K** Institutions

https://openalex.org/
Timeline of key data sources and main methodologies for studying open access publishing

- **Anecdotal**
- **Limited**
- **Manual sampling**
- **Automated sampling**
- **Real-time**

Anecdotal methods were used before 1998. From 1999 to 2000, limited sampling methods were employed. Manual sampling methods were used from 2001 to 2004, followed by automated sampling methods from 2005 to 2011. Real-time methods were introduced from 2012.

- **DOAJ**
  - Curated collection of active full OA journals fulfilling certain criteria: growth from 300 to over 20 000

- **Crossref**
  - Registry of article-level metadata, DOI registration for journals and articles

- **GOOgle Scholar**
  - Bottom-up identification of individual OA articles (and versions) on the web

- **OpenAlex**
  - Bottom-up DOI-based OA article location database

- **Registry of journal identifiers and publisher information**

- **RingRing of article metadata, DOI registration for journals and articles**

- **The Initiative for Open Abstracts I4OA**
4. Specific software available for supporting workflows
Python is a core language for data science in relation to scientometrics

- **PyAlex** is a Python library for OpenAlex. PyAlex is a lightweight and thin Python interface to the OpenAlex API. PyAlex tries to stay as close as possible to the design of the original service. https://github.com/J535D165/pyalex

- **Crossref API Client** is a Python library with functions to iterate through the Crossref API. https://github.com/fabiobatalha/crossrefapi

- **pyBibX** is a bibliometric and scientometric python library that uses the raw files generated by Scopus (.bib files), WOS (Web of Science) (.bib files), and PubMed (.txt files) scientific databases. https://pypi.org/project/pyBibX/
OpenRefine is a powerful free, open source tool for working with 
messy data: cleaning it, transforming it from one format into 
another; and extending it with web services and external data.

Main features

Faceting
Drill through large datasets using facets and apply operations on filtered views of your dataset.

Clustering
Fix inconsistencies by merging similar values thanks to powerful heuristics.

Reconciliation
Match your dataset to external databases via reconciliation services.

Infinite undo/redo
Rewind to any previous state of your dataset and replay your operation history on a new version of it.

Privacy
Your data is cleaned on your machine, not in some dubious data laundering cloud.

Wikibase
Contribute to Wikidata, the free knowledge base anyone can edit, and other Wikibase instances.

https://openrefine.org/
OpenRefine is a very accessible tool that can act as a middle ground between a spreadsheet program and a more complex database/dataframe.

Really easy to create dynamic queries to any REST API based on the values found in any column of your data.

OpenRefine

Add column by fetching URLs based on column ISBN

New column name: VoyageData

On error: set to blank

Throttle delay: 300 milliseconds

HTTP headers to use when fetching URLs: none

Formulate the URLs to fetch:

Expression: https://libapp.library.yale.edu/VoySearch/GetBibitem?isbn=

No syntax error.

Preview

<table>
<thead>
<tr>
<th>row</th>
<th>value</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9780415704953</td>
<td><a href="https://libapp.library.yale.edu/VoySearch/GetBibitem?isbn=9780415704953">https://libapp.library.yale.edu/VoySearch/GetBibitem?isbn=9780415704953</a></td>
</tr>
<tr>
<td>2</td>
<td>9782745327338</td>
<td><a href="https://libapp.library.yale.edu/VoySearch/GetBibitem?isbn=9782745327338">https://libapp.library.yale.edu/VoySearch/GetBibitem?isbn=9782745327338</a></td>
</tr>
<tr>
<td>3</td>
<td>9781780232652</td>
<td><a href="https://libapp.library.yale.edu/VoySearch/GetBibitem?isbn=9781780232652">https://libapp.library.yale.edu/VoySearch/GetBibitem?isbn=9781780232652</a></td>
</tr>
<tr>
<td>4</td>
<td>9788492865895</td>
<td><a href="https://libapp.library.yale.edu/VoySearch/GetBibitem?isbn=9788492865895">https://libapp.library.yale.edu/VoySearch/GetBibitem?isbn=9788492865895</a></td>
</tr>
</tbody>
</table>

OK Cancel

“VOSviewer is a software tool for constructing and visualizing bibliometric networks. These networks may for instance include journals, researchers, or individual publications, and they can be constructed based on citation, bibliographic coupling, co-citation, or co-authorship relations. VOSviewer also offers text mining functionality that can be used to construct and visualize co-occurrence networks of important terms extracted from a body of scientific literature.”
5. Some examples taken from my own work
2 recent studies I would like to talk about

Open is not forever: A study of vanished open access journals
Mikael Laakso, Lisa Matthias, Najko Jahn
First published: 21 February 2021
https://doi.org/10.1002/asi.24460

Open access books through open data sources: assessing prevalence, providers, and preservation
Mikael Laakso
Journal of Documentation
ISSN: 0022-0418
Article publication date: 13 June 2023
https://doi.org/10.1108/JD-02-2023-0016
Study 1
Open is not forever: a study of vanished open access journals

» Digital-only content is fragile, even though it is available openly on the web does not mean that anyone has made comprehensive backups that will be made available if the initial copies disappear.

» We were interested in taking a first systematic look at if, and if so, how much already published scholarly articles vanish from the web for various reasons.

1. How many OA journals have vanished from the web?
2. When did the OA journals vanish from the web?
3. What are the characteristics of vanished OA journals?
Data collection process

» A main challenge is the lack of any changes being recorded in the most recent dataset made available, and them only including active journals and silently removing those that have been found to be inactive

» We collected many old datasets containing listings of journals through 2000-2019, comparing these old lists to the currently active ones, looking out for which had been removed over time

» Manually visit each last known URL, search for the website if not active

» Each journal website was also traversed with the wayback machine to confirm that it had once existed and it had been publishing open access at some point

Data analysis done in R, data and scripts available at https://github.com/njahn82/vanished_journals
Results

» We were able to verify 174 OA journals that have vanished from the web. In many cases, the journals first transitioned to an inactive state for several years before eventually disappearing.

» This should be considered as a lower-bound count and that the number of vanished journals is likely to be much greater, but identifying and verifying additional cases would require a different methodological approach.
Publication history of vanished OA journals

Vanished OA Journals

Publication activity


Last year of publication

2015
2010
2005
2000
Period between the last journal publication and vanishing in years
Lifespan distribution of vanished journals across subject domains, in years
What could we learn from this?

» Internet archeology is quite exciting, though it is not getting all that much attention.

» With quite modest data collection practices and tools it is possible to derive important new knowledge that can influence practice.

» The research ignited a comprehensive preservation effort among key actors in the landscape that is still ongoing, attempting to increase the coverage of preservation service to the long tail of scholarly journals [https://doaj.org/preservation/](https://doaj.org/preservation/)

» It would be interesting to do a similar study based on vanished articles, but that is much more complicated for many reasons.
Aim of the research

1. To create a dataset of “all” currently known OA books from openly available sources

2. To assess enrolment of these OA books in international preservation services (e.g. CLOCKSS, Portico, Global LOCKSS Network)

3. Explore URL domains the DOIs of these books resolve to, in order to assess the distribution of content and technical environment surrounding them.
Data collection

- Already from the outset, it was known that the data collection circumstances for OA book content differ significantly from that of scholarly journals.

- For this study, two datasets needed to be put together and compared: one for academic OA books and the second for preservation coverage of books.

- Very mixed approaches to extract data from six bibliometric databases.
  - Directory of Open Access Books (CSV file)
  - WorldCat (Automated web scraping using Octoparse)
  - OpenAlex (self-written Python script to query JSON files, import into OpenRefine)
  - Scielo Books (Automated web scraping using Octoparse)
  - The Lens (CSV file)
  - OpenAire (Database dump imported into OpenRefine)
Data sources and their content.
Deduplication not straightforward

<table>
<thead>
<tr>
<th>Open Access Books</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOAB</td>
<td>49,600</td>
</tr>
<tr>
<td>WorldCat (OCLC)</td>
<td>4,385</td>
</tr>
<tr>
<td>OpenAlex</td>
<td>134,718</td>
</tr>
<tr>
<td>Scielo Books</td>
<td>1,006</td>
</tr>
<tr>
<td>The Lens</td>
<td>348,678</td>
</tr>
<tr>
<td>OpenAIRE</td>
<td>211,749</td>
</tr>
<tr>
<td><strong>~396,995 unique items</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Preservation services (book content only)</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLOCKSS</td>
<td>389,818</td>
</tr>
<tr>
<td>Portico</td>
<td>1,945,254</td>
</tr>
<tr>
<td>LOCKSS</td>
<td>21,258</td>
</tr>
<tr>
<td>Livro Aberto</td>
<td>461</td>
</tr>
</tbody>
</table>
Content distribution over major sources
Challenges

» **Definitions** - When is a book an academic book, and when is it open access?
  » Different indexing services have different levels of data quality, despite narrowing down searches as much as possible still a lot of “noise” in the data.
  » Interpretation/tagging of Open Access status also varies across services

» **Data management** - Physical extraction of metadata to represent the “global bookshelf” of academic OA books
  » Putting together the dataset a varied mix of REST APIs, JSON dumps, CSV files.
  » Many datasets needlessly large for investigations of this kind, hopefully more varied ways to access data like this in the future.

» **Unique identifiers** - Taming the wilderness of identifier metadata describing OA books
  » A single book can be assigned an ISBN (or nine ISBNs), a DOI, both or none.
  » Preservation services still dominantly ISBN-based at least when it comes to public book preservation data, an expansion into also including DOIs would for many purposes be beneficial.
Where are OA books hosted?

<table>
<thead>
<tr>
<th>DOAB</th>
<th>count</th>
<th>OpenAire</th>
<th>count</th>
<th>OpenAlex</th>
<th>count</th>
<th>The Lens</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>library.oapen.org</td>
<td>18889</td>
<td>biodiversitylibrary.org</td>
<td>74133</td>
<td>biodiversitylibrary.org</td>
<td>23347</td>
<td>biodiversitylibrary.org</td>
<td>121956</td>
</tr>
<tr>
<td>books.openedition.org</td>
<td>7889</td>
<td>link.springer.com</td>
<td>21466</td>
<td>afghandata.org</td>
<td>19840</td>
<td>link.springer.com</td>
<td>43261</td>
</tr>
<tr>
<td>mdpi.com</td>
<td>4023</td>
<td>elibrary.worldbank.org</td>
<td>6666</td>
<td>link.springer.com</td>
<td>8174</td>
<td>law.acu.edu.af</td>
<td>14183</td>
</tr>
<tr>
<td>mts.intechopen.com</td>
<td>3274</td>
<td>degruyter.com</td>
<td>5747</td>
<td>law.acu.edu.af</td>
<td>5562</td>
<td>afghandata.org</td>
<td>12944</td>
</tr>
<tr>
<td>frontiersin.org</td>
<td>2930</td>
<td>cambridge.com</td>
<td>5356</td>
<td>books.openedition.org</td>
<td>4566</td>
<td>online.library.wiley.com</td>
<td>10323</td>
</tr>
<tr>
<td>intechopen.com</td>
<td>2092</td>
<td>classiques.uqac.ca</td>
<td>4583</td>
<td>library.si.edu</td>
<td>4237</td>
<td>elibrary.worldbank.org</td>
<td>6910</td>
</tr>
<tr>
<td>degruyter.com</td>
<td>1652</td>
<td>books.openedition.org</td>
<td>3856</td>
<td>classiques.uqac.ca</td>
<td>4193</td>
<td>classiques.uqac.ca</td>
<td>6759</td>
</tr>
<tr>
<td>ksp.kit.edu</td>
<td>1647</td>
<td>taylorfrancis.com</td>
<td>3239</td>
<td>repository.usta.edu.co</td>
<td>4031</td>
<td>ieeexplore.ieee.org</td>
<td>6416</td>
</tr>
<tr>
<td>media.fupress.com</td>
<td>1371</td>
<td>library.si.edu</td>
<td>3175</td>
<td>openknowledge.wolrdbank.org</td>
<td>2036</td>
<td>degruyter.com</td>
<td>6241</td>
</tr>
<tr>
<td>books.scielo.org</td>
<td>1009</td>
<td>apps.crossref.org</td>
<td>3198</td>
<td>constellation.uqac.ca</td>
<td>1950</td>
<td>dl.acm.org</td>
<td>6078</td>
</tr>
<tr>
<td>omp.zrc-sazu.si</td>
<td>591</td>
<td>mr.crossref.org</td>
<td>3002</td>
<td>constellation.uqac.ca</td>
<td>1950</td>
<td>dl.acm.org</td>
<td>6078</td>
</tr>
<tr>
<td>ucdigitalis.uc.pt</td>
<td>494</td>
<td>repository.usta.edu.co</td>
<td>2854</td>
<td>darchive.mblwhollibrary.org</td>
<td>1721</td>
<td>taylorfrancis.com</td>
<td>4569</td>
</tr>
<tr>
<td>nomos-elibrary.de</td>
<td>429</td>
<td>vr-elibrary.de</td>
<td>2350</td>
<td>press.umich.edu</td>
<td>1692</td>
<td>apps.crossref.org</td>
<td>4518</td>
</tr>
<tr>
<td>edp-open.org</td>
<td>288</td>
<td>oxford.universitypressscholarship.com</td>
<td>2277</td>
<td>apps.crossref.org</td>
<td>1634</td>
<td>repository.si.edu</td>
<td>4193</td>
</tr>
<tr>
<td>link.springer.com</td>
<td>252</td>
<td>press.umich.edu</td>
<td>2203</td>
<td>mohrsiebeck.com</td>
<td>1445</td>
<td>repository.usta.edu.co</td>
<td>3702</td>
</tr>
<tr>
<td>ledizioni.it</td>
<td>228</td>
<td>rand.org</td>
<td>2109</td>
<td>books.fupress.com</td>
<td>1353</td>
<td>mdpi.com</td>
<td>3007</td>
</tr>
<tr>
<td>bloomsburycollections.com</td>
<td>193</td>
<td>worldscientific.com</td>
<td>2077</td>
<td>liu.diva-portal.org</td>
<td>1294</td>
<td>jstor.org</td>
<td>2855</td>
</tr>
<tr>
<td>e-archivo.uc3m.es</td>
<td>170</td>
<td>darchive.mblwhollibrary.org</td>
<td>2028</td>
<td>jstor.org</td>
<td>1279</td>
<td>deepblue.lib.umich.edu</td>
<td>2819</td>
</tr>
<tr>
<td>api.intechopen.com</td>
<td>162</td>
<td>constellation.uqac.ca</td>
<td>2026</td>
<td>rand.org</td>
<td>1230</td>
<td>academic.oup.com</td>
<td>2410</td>
</tr>
</tbody>
</table>

...and 188 more domains containing the remaining 7% of items.

...and 1453 more domains containing the remaining 28% of items.

...and 1470 more domains containing the remaining 32% of items.

...and 1816 more domains containing the remaining 23% of items.
### Existing coverage of content in international preservation services
*(match either by ISBN or exact book title match)*

<table>
<thead>
<tr>
<th>Service</th>
<th>DOAB</th>
<th>WorldCat</th>
<th>The Lens</th>
<th>OpenAlex</th>
<th>OpenAire</th>
<th>Scielo</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLOCKSS</td>
<td>22.5%</td>
<td>6.6%</td>
<td>3.3%</td>
<td>4.2%</td>
<td>8.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Portico</td>
<td>31.2%</td>
<td>31.3%</td>
<td>8.6%</td>
<td>11.3%</td>
<td>22.3%</td>
<td>8.7%</td>
</tr>
<tr>
<td>LOCKSS</td>
<td>22.5%</td>
<td>1.2%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.2%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Livro Aberto</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Found in at least one of the above preservation services</td>
<td>45.7%</td>
<td>32.6%</td>
<td>9.7%</td>
<td>12.5%</td>
<td>24.7%</td>
<td>8.7%</td>
</tr>
</tbody>
</table>
Some thoughts

» A number of DOIs resolve to error pages or very volatile hosting services. It is likely that we have already started losing content that has once been published.

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

» How best to capture the current, and likely growing, long tail of OA book content into preservation coverage?
Better metadata and use of identifiers is key to data improvement

» There needs to be added transparency and data concerning key entities of relevance to the scholarly publishing landscape.

» Actors (individuals), affiliated organisations, journals, funders etc.

» Most parts are moving and can appear in various configurations and combinations.

» ORCID is one step towards better data, but affiliation data and organisational identifiers need to be further enforced and standardised.
Key takeaways

» There has been rapid increase in the **openness** of data describing scholarly journal publishing and open access specifically. But more can be done!

» Whatever metadata standards and databases are developed, and existing ones expanded, they need to be **sustainable** in their approach.

» A lot of methodological options for defining and researching open access publishing. **Reproducibility** and comparability between measurements has so far been low, though things are improving.

» Better automatic, **longitudinal data are needed**, the world of scholarly journal publishing moves fast and good data and tools are needed to keep up!
Q & A
Reccomended readings


Thank You!