



Researching Research

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

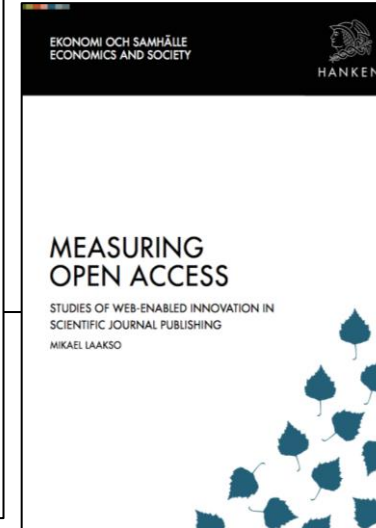
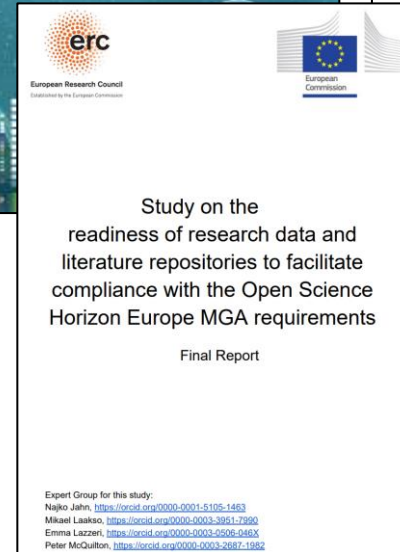
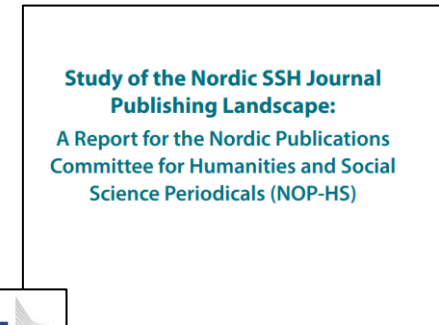


Background



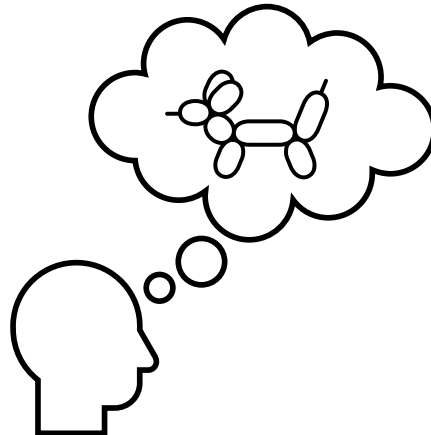
HANKEN

- » 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

https://www.stm-assoc.org/2022_08_24_STM_White_Report_a4_v15.pdf

<https://direct.mit.edu/qss/article/3/4/912/114119/Recalibrating-the-scope-of-scholarly-publishing-A>

The relationship between research and money is super interesting



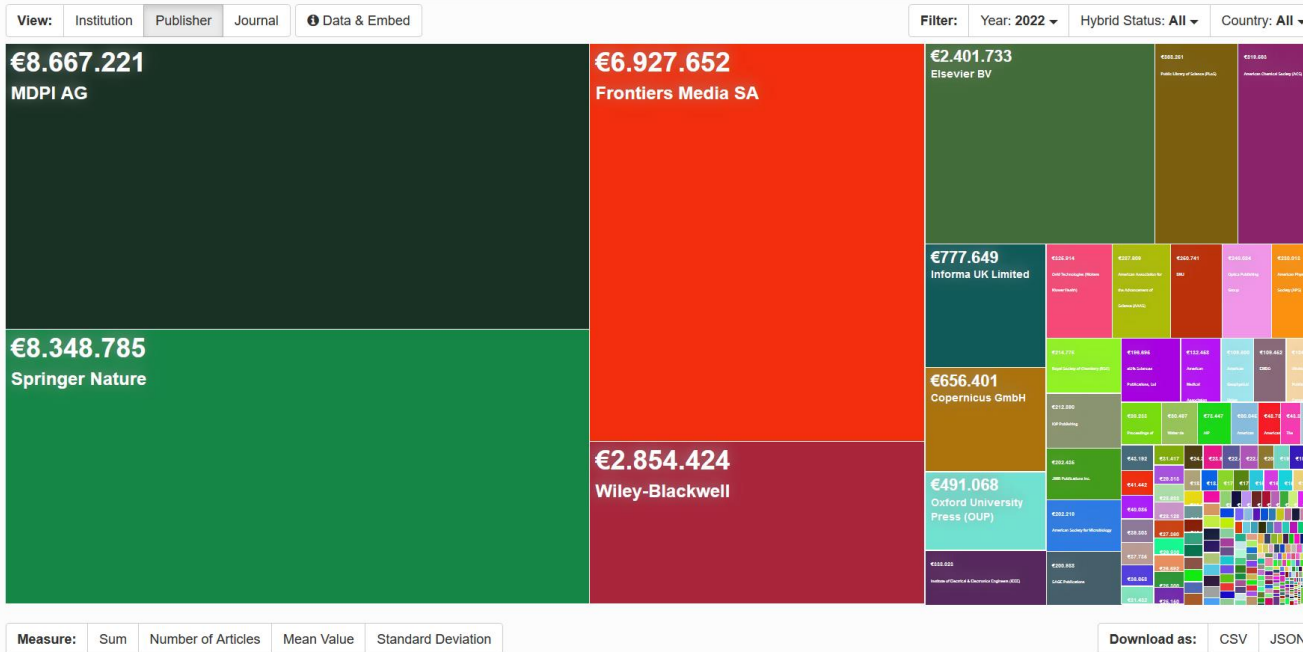
Open APC – One window into the money flows supporting academic publishing



OPEN@APC

ABOUT OLAP SERVER GITHUB OPENAPC

COMBINED (COST DATA FROM OPENAPC AND TRANSFORMATIVE AGREEMENTS DATA SETS)

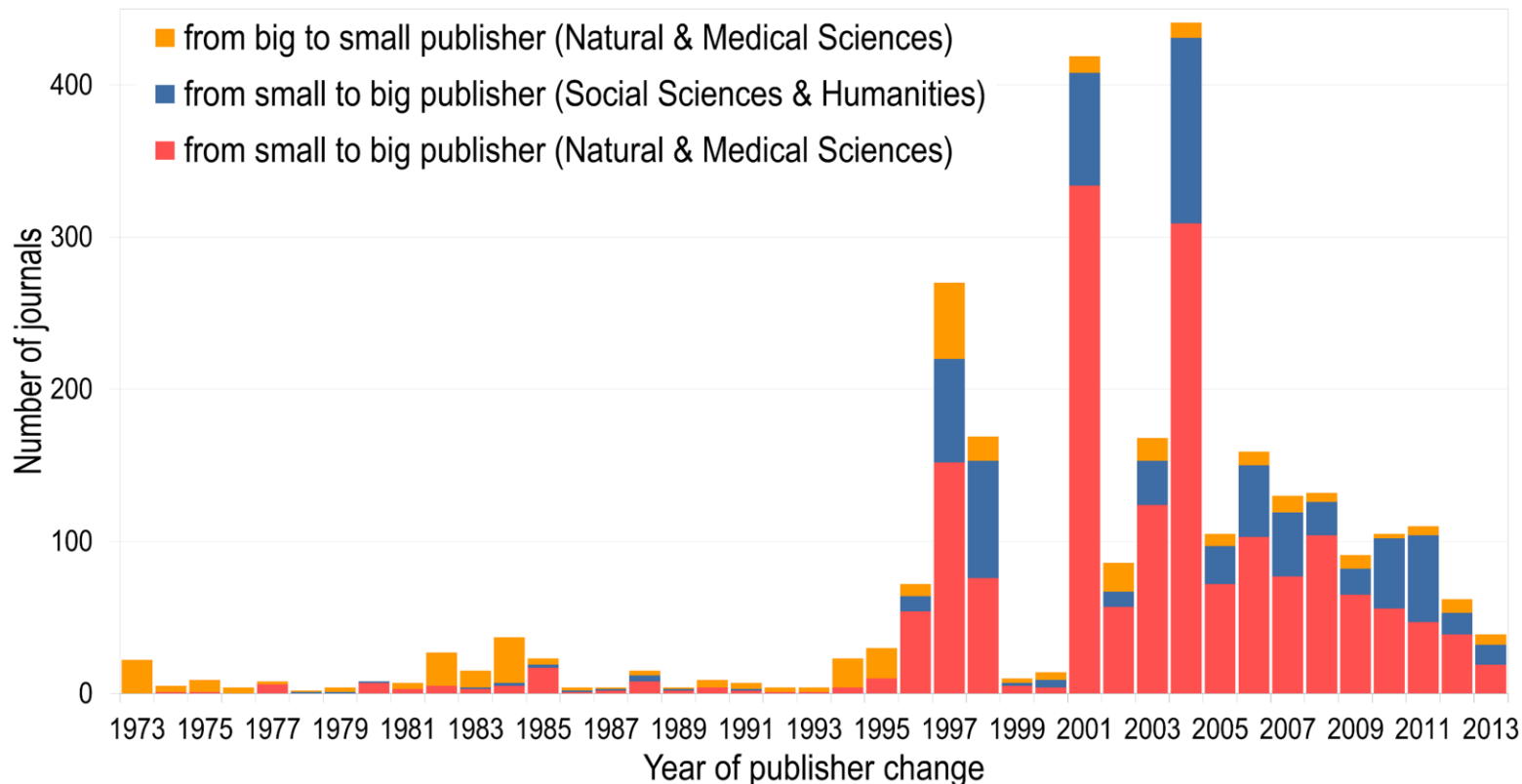


<https://treemaps.openapc.net/apcdata/combined/#publisher/period=2022>

The big publishers have constantly gotten bigger



HANKEN



2. Bibliometrics and Scientometrics as Data Science

Bibliometrics



- » The term was first used in 1969, then referring to “Statstical 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 (1955) <https://doi.org/10.1093/ije/dyl189>

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 first coined (1971) the possibilities for expansion were still limited, but the digital environment has opened up so many new possibilities for inquiry.

Some basics to start with regarding the components of academic articles



HANKEN

Science and Public Policy, 2023, 50, 445–456
DOI: <https://doi.org/10.1093/sppol/scz045>

Advance Access Publication Date: 17 February 2023

Article

OXFORD

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

Mikael Laakso^{1,*} and Anna-Maija Multas²

¹Information Systems Science, Hanken School of Economics, Arkadiankatu 22, Helsinki 00100, Finland and ²Information Studies, Faculty of Humanities, University of Oulu, P. O. Box 1000, Oulu 90014, Finland

*Corresponding author. E-mail: mikael.laakso@hanken.fi

Abstract

This study investigates the relationship between scholarly journal publishing and public funding, specifically concerning the context of small- and mid-sized journal publishers in European countries. As part of the movement towards open science, an increasing number of journals globally are free to both read and publish in, which increases the need for journals to seek other resources instead of subscription income. The study includes two separate data components, collecting data separately for each European country (including transcontinental states): (1) the volume and key bibliometric characteristics of small- and mid-sized journal publishers and (2) information about country-level public funding mechanisms for scholarly journals. The study found that there are 16,367 journals from small- and mid-sized publishers being published in European countries, of which 26 per cent are already publishing open access. There is a large diversity in how countries reserve and distribute funds to journals, ranging from continuous inclusive subsidies to competitive grant funding or nothing at all.

Key words: journals; funding; national; science policy.

1. Introduction

The scholarly journal publishing sector has faced three intertwined and impactful changes during the last three decades. The first one of these is proliferation of digitisation and digital content delivery, which in the beginning posed challenges as individual journals and smaller publishers were not able to invest in and fully exploit it. The second change is related to the pattern of large publishers becoming even larger by acquiring smaller publishers and individual titles into their portfolios (Larivière et al. 2015). Publisher oligopolisation together with digitisation fuelled the ‘big-deal’ business model. The third change is the growth of open access (OA) that has disrupted the sector in many ways as access can be provided through journals directly as well as authors indirectly. In the late 1980s and the early 1990s, OA started to gain momentum as a largely community-driven bottom-up movement but has since been shaped strongly by commercial interests and science policy (Moore 2020; Schöpfel 2015).

When compared with paywalled subscription-based access, OA fundamentally changes the operating circumstances for journals as subscription income significantly decreases or disappears and journals are required to acquire other forms of funding or support to continue their activities. The largest international publishers have adjusted their offerings and business models to accommodate the growing demand for OA. This has often been done by introducing, e.g., transformative agreements in which case the customer institutions buy pre-paid quotas for affiliated authors to publish OA in

the publishers’ journals (ESAC-initiative.org 2021). Overall, OA has not posed an immediate financial threat to large publishers who, on the contrary, have been able to monetise the science policy pressure placed on its growth. For small and mid-sized publishers, which act outside the realm of institutional agreements with substantial leverage in contract negotiations, operational circumstances can appear very different.

Regardless of the publication model, scholarly journals need resources to run and persist. Such resources can come from many different directions and in many different forms (e.g. monetary, volunteer work, and shared infrastructures). However, without sufficient resources, a scholarly journal cannot continue to exist in the long run. Insufficiently resourced journals can also pose a risk to the integrity of the scholarly record if technical precautions for preservation are not adequately taken care of (Laakso et al. 2021). Based on the size of the primary audience, the potential for gathering resources is higher for English language, internationally-oriented journals than non-English journals that have a narrower geographical focus. It is here where journals’ national-level funding instruments often offer the key resources to support non-profit publication outlets, which could otherwise fail to survive. The existence of financial support for journals brings with it the need to deliberate on both how such instruments should be designed and how such mechanisms should evolve over time as the scholarly journal and scholarly communication landscape changes.

Science and Public Policy

that many globally-oriented journals might not publish. In cases where communication in local languages is a high priority, it makes sense to have dedicated funds directed at outlets that take this aim further rather than mixing such funds into institutional funding schemes with the assumption that some of the resources would go towards publication-related practices.

The financial and contractual knowledge base for international publishers is well developed through advances made as part of the collaborative Efficiency and Standards for Article Charges initiative in which the consortia and libraries share the terms of their contracts often together with cost breakdowns (ESAC-initiative.org 2021). However, the same cannot be said of information concerning public funding directed to local journals. This happens even though such information is theoretically easier to make public as commercial non-disclosure agreements do not hinder what can be made public information, and there is an ideological ground to make use of public funds as transparent as possible for citizens. Hence, one of our practical recommendations would be for national actors to collaborate internationally on designing and implementing practices through which non-profit journals can most efficiently be supported with public funds. This would enable learning from each other and making the endeavours compatible with the circumstances of OA publishing. Such actions would also likely lower the threshold for collaboration on other fronts, such as on common investments into further development of open-source publishing platforms.

For future research, it could prove interesting to take this initial charting of the landscape and paint a fuller picture by zooming in on various aspects of interest. One suggestion would be to take a closer look at the publishing organisations that are responsible for one or a handful of journals and to examine how are they working and what steps could be taken to facilitate their activities. Another study could enhance the level of analysis to also include article counts for journals, an element that was now missing due to the lack of such data in the Ulrichweb database. By aggregating data from multiple sources and conducting manual data collection where needed, one could get an enhanced perspective that would consider the size differences among journals. Other research could conduct an evaluation of the strengths and weaknesses of the different funding models available by consulting journals that have experience in utilising them. This would likely provide valuable input into future policy-making.

6. Conclusions

We consider that, as the push towards more OA publishing increases, the aspect of public funding for journals is something that would warrant more systemic global attention. Due to the reduction and eventual ease of subscription income, journals must find alternative funding streams to cover costs or alternatively seek a publishing agreement with an international commercial publisher to gain financial stability and predictability. The problem with such arrangements is that multilingualism is often compromised in favour of English. This may lead to the journals’ scopes becoming broader to attract a global audience of both readers and authors, something that undesirably reduces the local relevance of the journal. Ultimately, in a such scenario, it is likely that public-sector funds are still used to a high degree, just

funneling through large international companies that require their own share of the transaction. This makes it more expensive compared to direct public subsidies to the journal. A well-designed public funding instrument is likely to enable the existence and diversity of scholarly publication outlets, which are of high relevance to more specific audiences than just the generic universal global target audience.

Supplementary material

Supplementary material is available at *Science and Public Policy* online.

Data availability

The data concerning the identified funding mechanisms are made openly available as supplementary data to this article. The bibliometric journal data as been made available as open data (Laakso and Multas, 2022b).

Funding

This research was funded by the Finnish Association for Scholarly Publishing.

Conflict of interest statement. None declared.

Acknowledgements

The authors wish to express their gratitude to the journal editors, who took the time to respond to our survey, and the OpenAIRE contact persons for giving helpful information on the public funding circumstances in many of the included countries.

References

- Biok, B.-C. (2017) ‘Journal Portals—An Important Infrastructure for Non-commercial Scholarly Open Access Publishing’, *Online Information Review*, 41: 643–54.
- Björnsom, A., Tsoukala, V., Barbaresco, E., et al. (2022) *Access to and Preservation of Scientific Information in Europe. Report on the Implementation of Commission Recommendation C(2012)2375 Final* <<https://op.europa.eu/en/publication-detail/publication/665718ef-6179-11ea-991b-01aa75ed71a1>> accessed: 25 Jan 2022.
- BOAI (2002) *Budapest Open Access Initiative* <<http://web.archive.org/web/20220125142553/https://www.budapestopenaccessinitiative.org/reads>> accessed: 25 Jan 2022.
- Boutman, J., Frankvis, J. E., Kramer, B., et al. (2021) ‘OA Diamond Journals Study. Part I: Findings’, *Zenodo*, accessed: 9 Mar 2021.
- Brewer, J. D. (2013) *The Public Value of the Social Sciences: An Interpretive Essay*. Bloomsbury Academic.
- Bruno, A., Leske, C., Schmidt, C., et al. (2020a) *ISSN-Matching of Gold OA Journals (ISSN-GOLD-OA) 4.0*. Bielefeld University.
- Bruno, A., Rimmert, C., and Taubert, N. (2020b) ‘Who Pays? Comparing Cost Sharing Models for a Gold Open Access Publications Environment’, *Journal of Library Administration*, 60: 853–74.
- Brysbært, M. (2021) ‘The Role of Learned Societies and Grant-Funding Agencies in Fostering a Culture of Open Science’, *PsyArXiv*, accessed: 8 Jun 2021.
- CNN (2018) *Hungary’s PM Bans Gender Study at Colleges Saying ‘People Are Born Either Male or Female’* (published online: 3 October 2018). <<https://web.archive.org/web/20220118101918263/https://edition.cnn.com/2018/10/17/europe/hungary-bans-gender-study-at-colleges-trnd/index.html>> accessed: 19 Oct 2018.

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



HANKEN

Peer Review reports

From: [New insights into genetic susceptibility of COVID-19: an ACE2 and TMPRSS2 polymorphism analysis](#)

Original Submission		
25 Apr 2020	Submitted	Original manuscript
15 May 2020	Reviewed	Reviewer Report
22 May 2020	Reviewed	Reviewer Report
27 May 2020	Author responded	Author comments - Feixiong Cheng
Resubmission - Version 2		
27 May 2020	Submitted	Manuscript version 2
29 May 2020	Reviewed	Reviewer Report
15 Jun 2020	Reviewed	Reviewer Report
18 Jun 2020	Author responded	Author comments - Yuan Hou
Resubmission - Version 3		
18 Jun 2020	Submitted	Manuscript version 3
22 Jun 2020	Reviewed	Reviewer Report
Resubmission - Version 4		
	Submitted	Manuscript version 4
Publishing		
22 Jun 2020	Editorially accepted	
15 Jul 2020	Article published	10.1186/s12916-020-01673-z

You can find [further information about peer review here](#).

[Back to article page >](#)

3. The evolving data environment of meta-research

Alternative metrics/altmetrics

- » 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)
 - » **Reccomendations** (how often the content has been reccomended 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

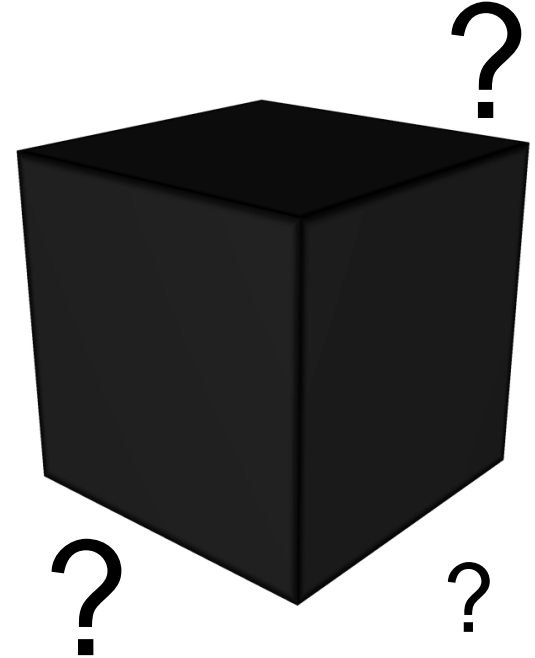
» Access to the most comprehensive commercial databases is limited, and **datasets created on the basis of such proprietary data can rarely be freely redistributed in their most usable form.**



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.**

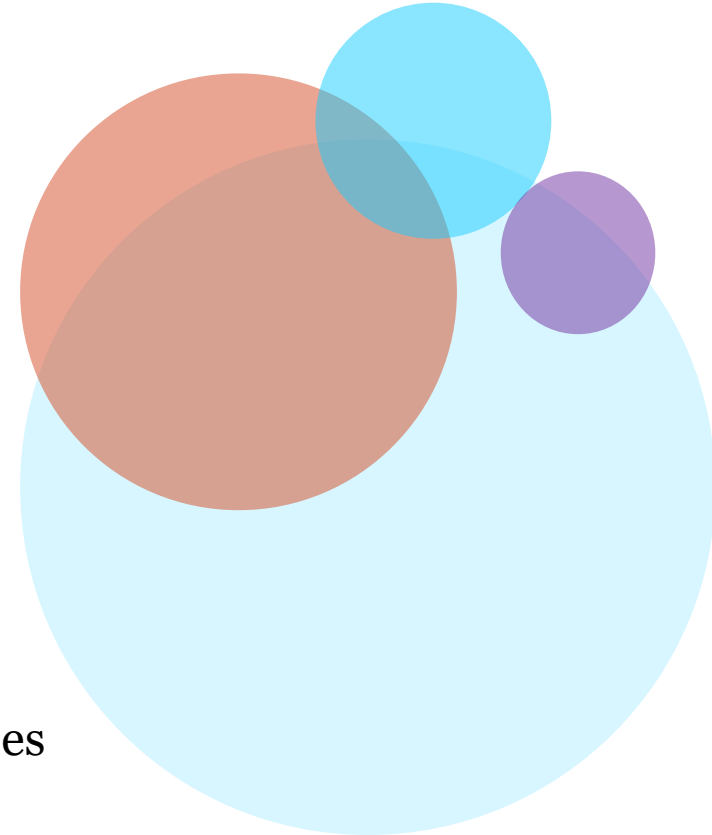


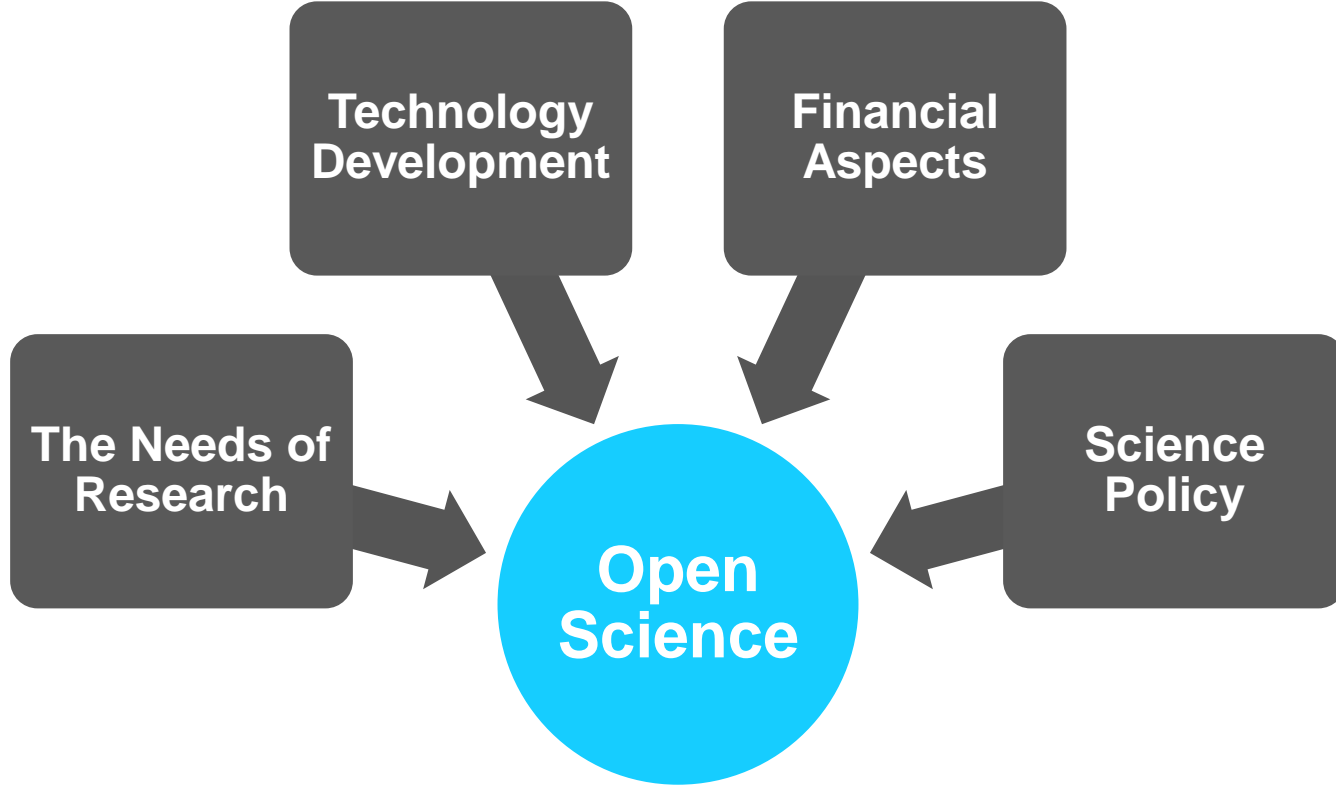
Various indexes/databases to choose from, all with different implications



HANKEN

- » Scopus
- » Web of Science
- » Dimensions
- » (Microsoft Academic)
- » The Lens
- » Ulrichsweb
- » Crossref/DOI
- » DOAJ
- » ROAD
- » Google Scholar
- » National research databases







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



HANKEN

Google

Scholar About 126,000 results (0.12 sec)

Articles

Case law

My library

Any time

Since 2017

Since 2016

Since 2013

Custom range...

Sort by relevance

Sort by date

include patents

include citations

Create alert

Fish consumption, fish oils, and cardiovascular events: still waiting for definitive evidence
[PM Ridker](#) - [The American Journal of Clinical Nutrition](#), 2016 - [Am Soc Nutrition](#)
↩ 1 Allaire J, Couture P, Leclerc M, Charest A, Marin J, Lépine MC, Talbot D, Tcherno A, Lamarche B. A randomized, crossover, head-to-head comparison of eicosapentaenoic acid and docosahexaenoic acid supplementation to reduce inflammation markers in men and
[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
[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 Etxezarreta](#)... - [Public health ...](#), 2016 - [Cambridge Univ Press](#)
Objective To prospectively assess the associations between lean **fish**, fatty **fish** and total **fish** intakes 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
[Related articles](#) [All 6 versions](#) [Cite](#) [Save](#)

Regular fish consumption and age-related brain gray matter loss
[CA Raji](#), [KI Erickson](#), [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
[Cited by 34](#) [Related articles](#) [All 10 versions](#) [Cite](#) [Save](#)

[HTML] infona.pl

[PDF] cambridge.org

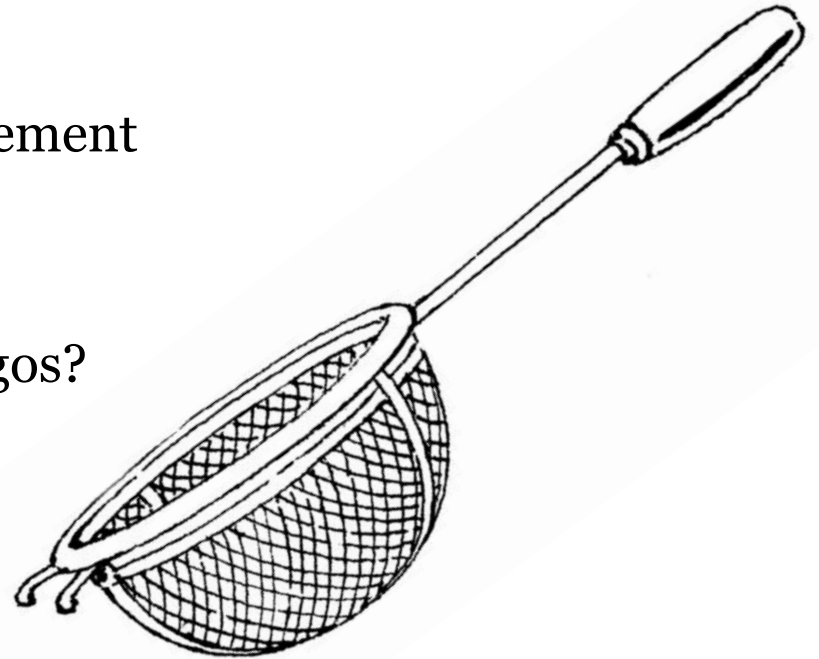
[HTML] nih.gov

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?

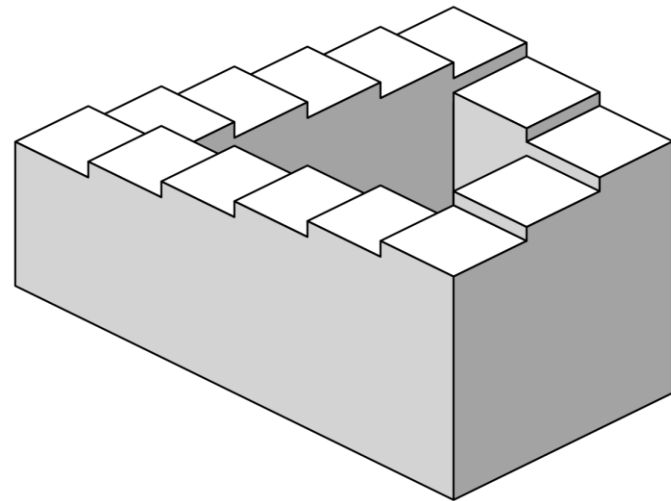


Journal vs Article perspectives



HANKEN

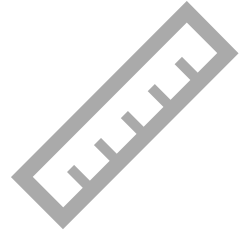
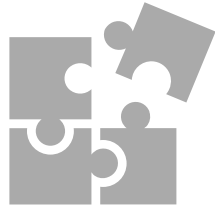
- » **A complicated relationship**
- » Partial openness of journals
- » Journals can and have disappeared, merged, changed OA model, some articles might still be available online elsewhere.



Different types of data is created throughout the research process



HANKEN



Discovery

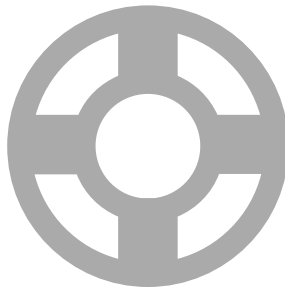
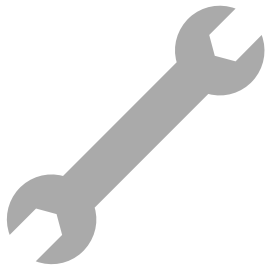
Analysis

Writing

Publication

Outreach

Assessment



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



HANKEN

- » “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 ORCIDs

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

< 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020



Registry of journal identifiers and publisher information



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

The Initiative for Open Citations I4OC
The Initiative for Open Abstracts I4OA



DOAJ

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



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



OpenAlex

Bottom-up
DOI-based OA
article location
database

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

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.



Download

Main features



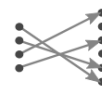
Faceting

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.

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



HANKEN

OpenRefine clipboard

1001 rows

Extensions: Wikidata

Using facets and filters

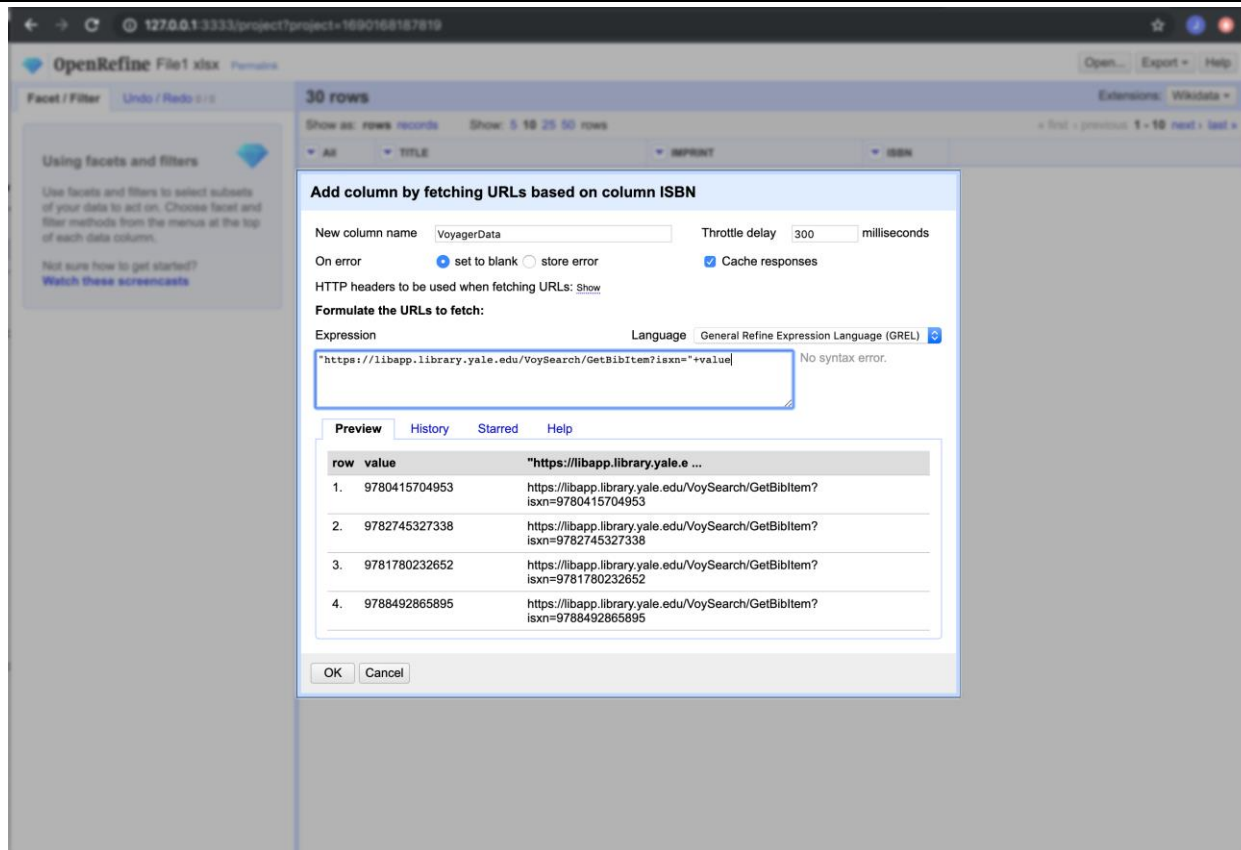
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

IDNs	handle	Publisher	Citation	License
1099-4300	http://hdl.handle.net/10079/1001208085	MDPI AG	Entropy, Vol 17, Iss 12, Pp 3945-3948 (2015)	CC BY
2077-0472	http://hdl.handle.net/10079/10012093072	MDPI AG	Agriculture (Basel), Vol 5, Iss 4, Pp 1172-1182 (2015)	CC BY
1422-0067	http://hdl.handle.net/10079/10012030226	MDPI AG	International Journal of Molecular Sciences, Vol 16, Iss 12, Pp 28265-28266 (2015)	CC BY
2304-4740		MDPI AG	Inorganics (Basel), Vol 3, Iss 4, Pp 634-635 (2015)	CC BY
2306-5338		MDPI AG	Hydrology, Vol 2, Iss 4,	CC BY

- Facet
- Text filter
- Edit cells
- Edit column
 - Split into several columns...
 - Add column based on this column...
 - Add column by fetching URLs...
 - Add columns from reconciled values...
- Transpose
- Sort...
- View
 - Add columns from reconciled values...
- Reconcile
 - Rename this column
 - Remove this column
 - Move column to beginning
 - Move column to end
 - Move column left
 - Move column right

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



The screenshot shows the OpenRefine web interface. A dialog box titled "Add column by fetching URLs based on column ISBN" is open. The dialog contains the following fields and options:

- New column name:
- Throttle delay: milliseconds
- On error: set to blank store error
- Cache responses
- HTTP headers to be used when fetching URLs: [show](#)
- Formulate the URLs to fetch:
 - Expression:
 - Language:

Below the input fields is a preview table:

row	value	"https://libapp.library.yale.e ..."
1.	9780415704953	https://libapp.library.yale.edu/VoySearch/GetBibItem?isxn=9780415704953
2.	9782745327338	https://libapp.library.yale.edu/VoySearch/GetBibItem?isxn=9782745327338
3.	9781780232652	https://libapp.library.yale.edu/VoySearch/GetBibItem?isxn=9781780232652
4.	9788492865895	https://libapp.library.yale.edu/VoySearch/GetBibItem?isxn=9788492865895

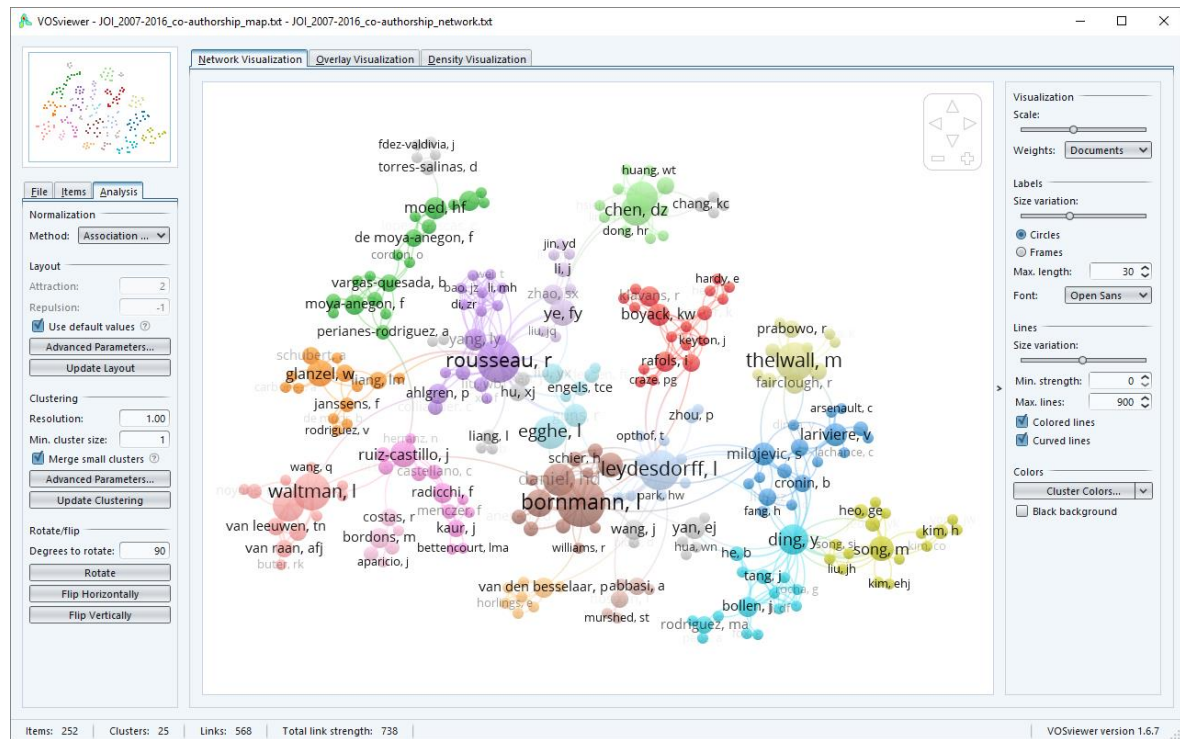
At the bottom of the dialog are "OK" and "Cancel" buttons.

VOSviewer



HANKEN

“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.”



<https://www.vosviewer.com/>

Web-browser based version: <https://app.vosviewer.com/>

5. Some examples taken from my own work

2 recent studies I would like to talk about



HANKEN

JASIST
JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY

asis&t
Association for Information Science and Technology

RESEARCH ARTICLE | Open Access |

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

DOWNLOADS



ALTMETRICS




<https://doi.org/10.1108/JD-02-2023-0016>

- » 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



HANKEN

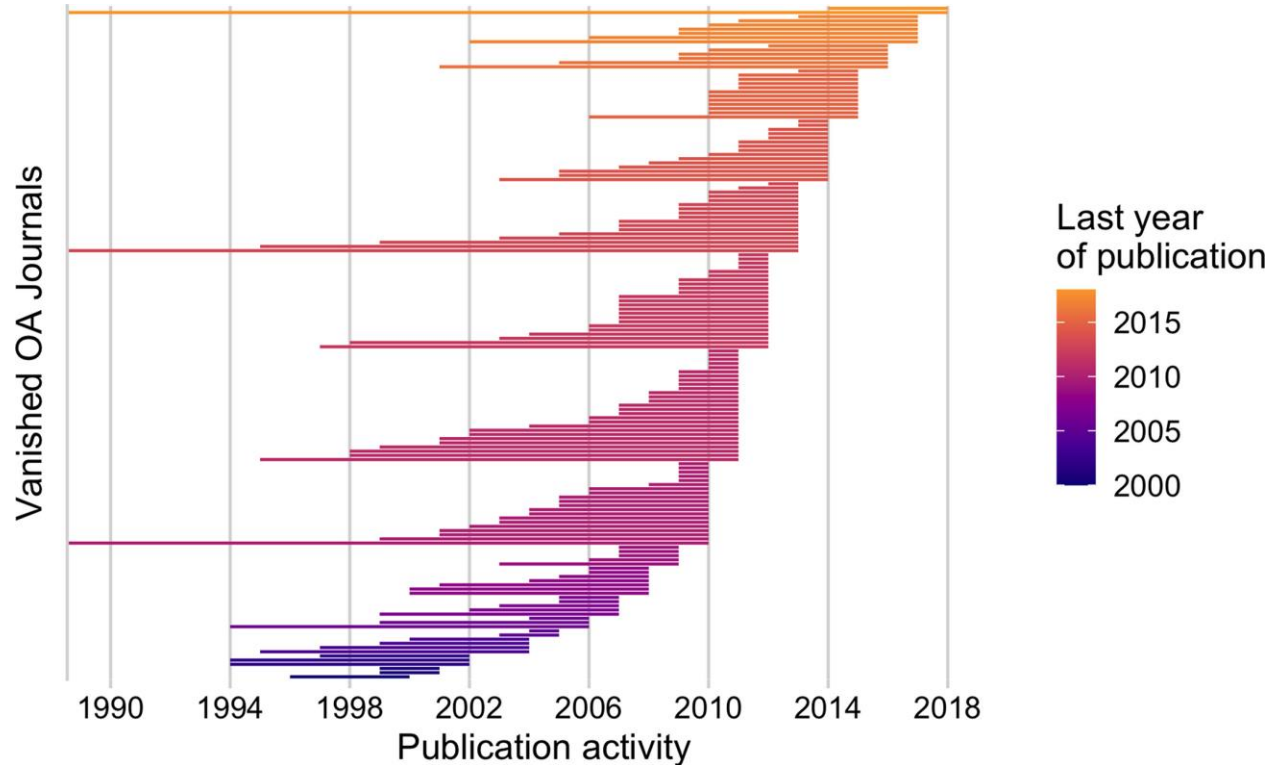
- » A main challenge is the lack of any changes being recorded in the most recent dataset made available, and them only including active journals and silently removing those that have been found to be inactive
- » We collected many old datasets containing listings of journals through 2000-2019, comparing 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  to confirm that it had once existed and it had been publishing open access at some point



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

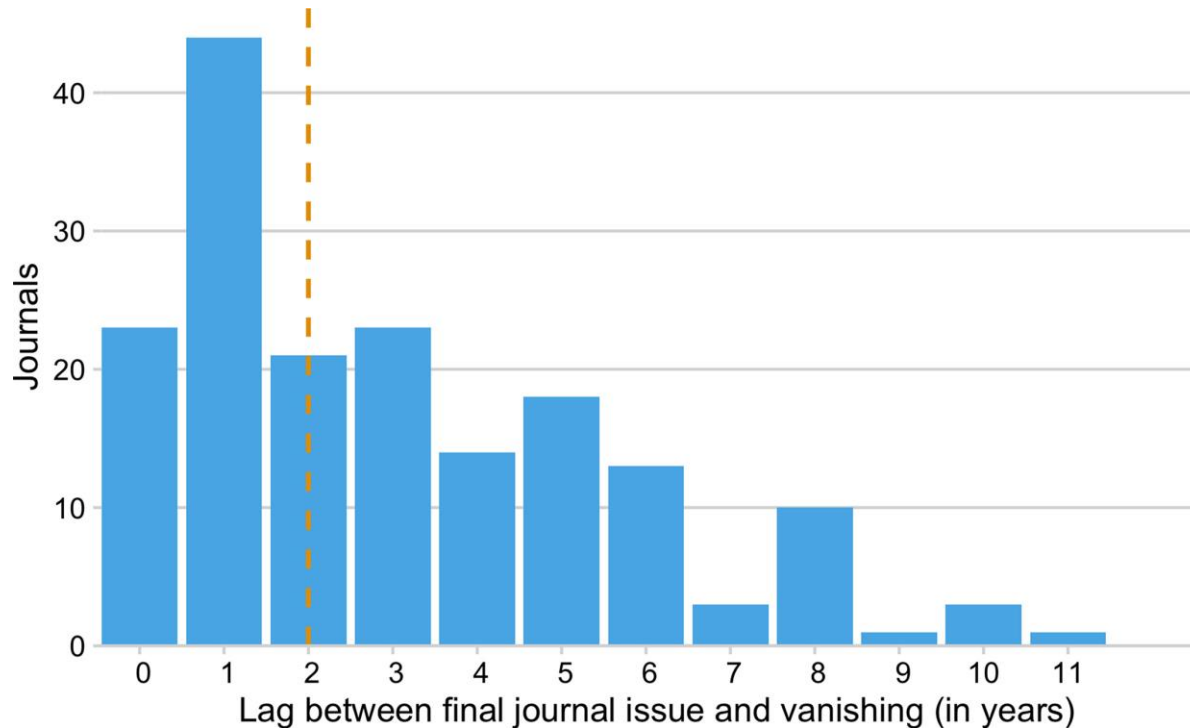
Publication history of vanished OA journals



Period between the last journal publication and vanishing in years



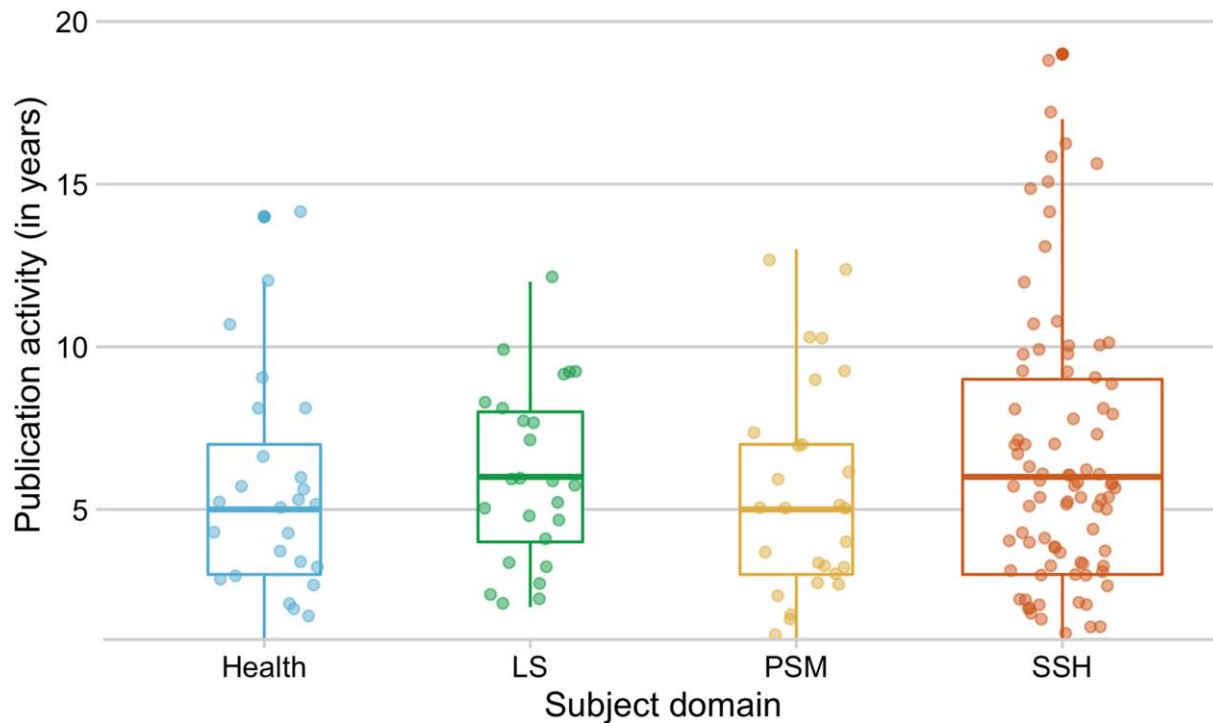
HANKEN



Lifespan distribution of vanished journals across subject domains, in years



HANKEN



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/>
- » 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

Open Access Books

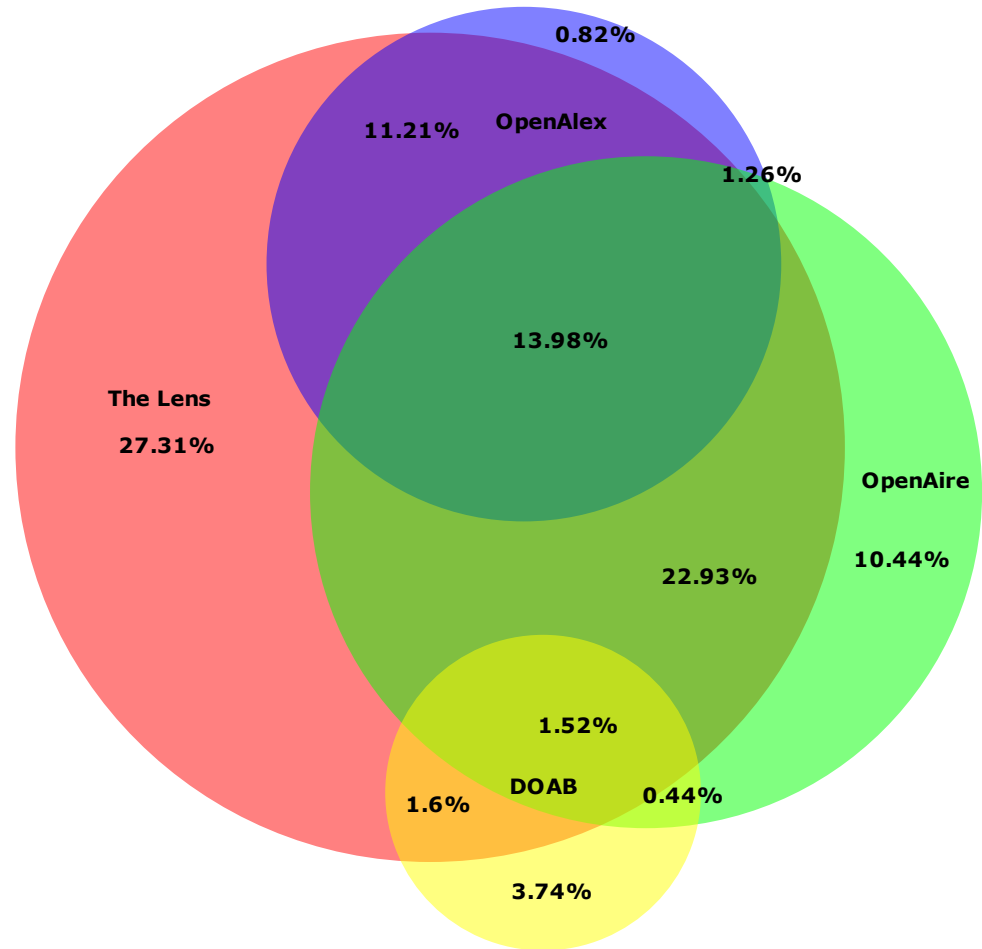
Service	Items
DOAB	49 600
WorldCat (OCLC)	4 385
OpenAlex	134 718
Scielo Books	1 006
The Lens	348 678
OpenAIRE	211 749

~396 995 unique items

Preservation services (book content only)

Service	Items
CLOCKSS	389 818
Portico	1 945 254
LOCKSS	21 258
Livro Aberto	461

*Content
distribution over
major sources*



Challenges



HANKEN

- » *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?



HANKEN

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

...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)



HANKEN

	DOAB	WorldCat	The Lens	OpenAlex	OpenAire	Scielo
CLOCKSS	22,5 %	6,6 %	3,3 %	4,2 %	8,1 %	0,0 %
Portico	31,2 %	31,3 %	8,6 %	11,3 %	22,3 %	8,7 %
LOCKSS	22,5 %	1,2 %	0,1 %	0,1 %	0,2 %	0,0 %
Livro Aberto	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %
Found in at least one of the above preservation services	45,7 %	32,6 %	9,7 %	12,5 %	24,7 %	8,7 %



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



HANKEN

- » Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, Farley A, West J, Haustein S. (2018). **The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles**. PeerJ 6:e4375 <https://doi.org/10.7717/peerj.4375>
- » Lamers, W.S., Boyack, K., Larivière, V., Sugimoto, C.R., van Eck, N.J., Waltman, L., Murray, D. (2021). **Meta-Research: Investigating disagreement in the scientific literature** eLife 10:e72737. <https://doi.org/10.7554/eLife.72737>

Thank You!