Infrastructure for data on open access: openness, sustainability, reproducibility

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

Keynote presentation at the 2019 OA Tage 2.10.2019
@mikaellaakso

*It is not only open access that is in the dark,
this also concerns scholarly publishing more broadly.*

But through open access it is possible to remedy the problem of lacking data in many ways.

I do not have the answers, but I have identified limitations of the current data environment.

Some relevant questions which have yet to have good answers due to lack of data

HANKEN

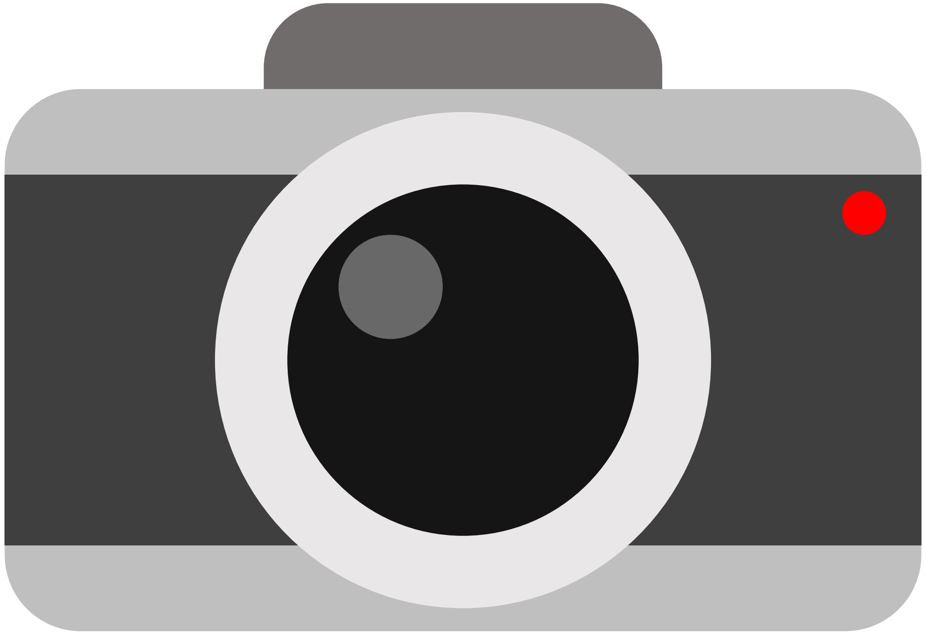
How much of OA growth is down to just improved indexing and detection?

How has the open access publishing landscape changed since the introduction of a specific policy intervention (e.g. Plan S) ?

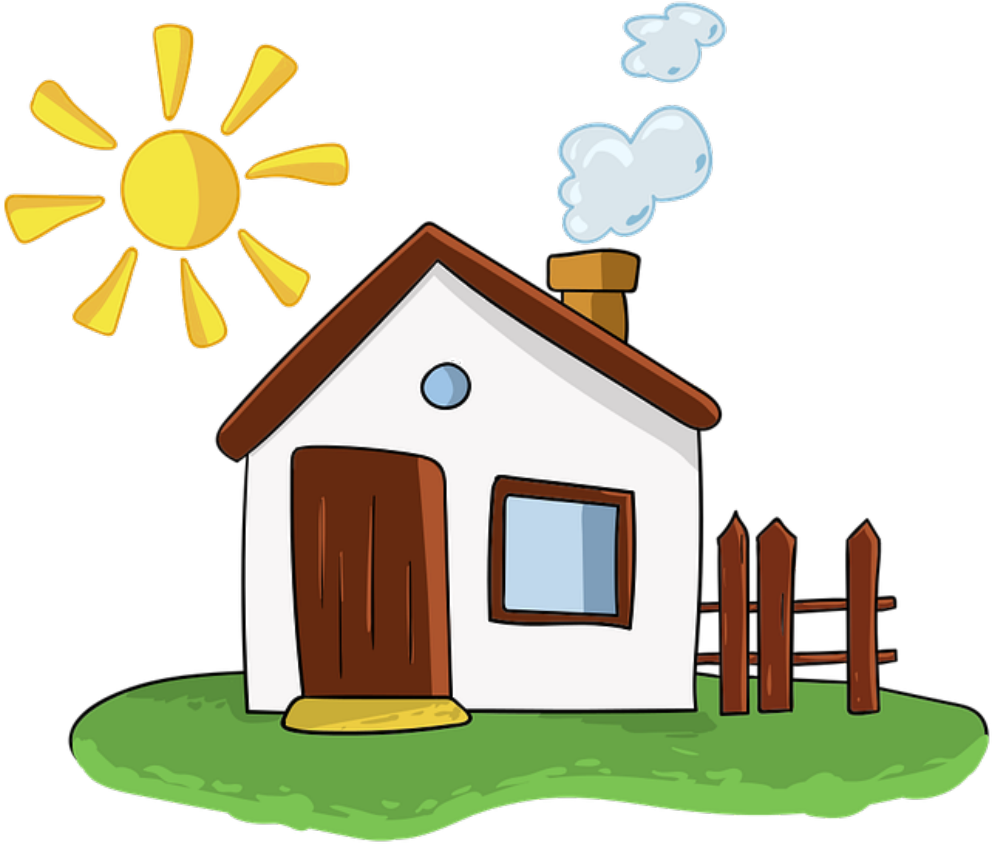
How is open access publishing growing in comparison to the overall growth of science, in terms of number of journals and number of articles?

How have APCs developed over time? Is this development uniform across research disciplines/ publishers/ countries?

Currently OA measuring can be considered to be done by "Camera", taking snapshots.



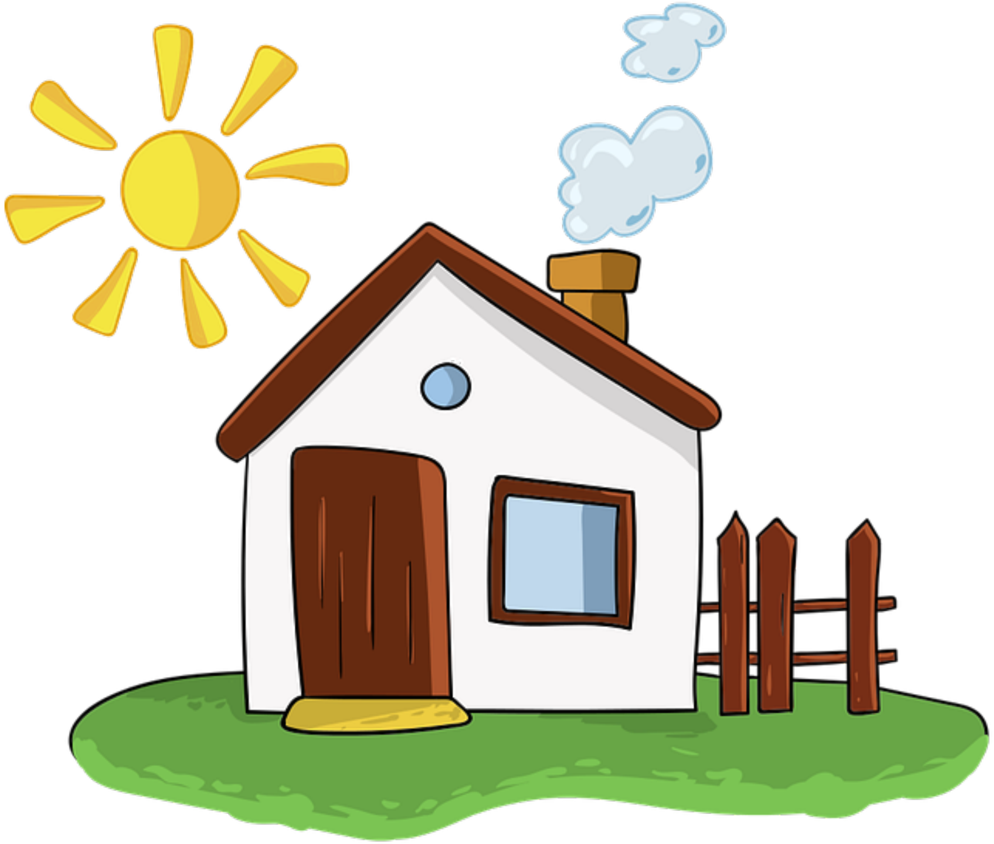
Comparing different photographs can be difficult, so much variation in method, focus, and equipment



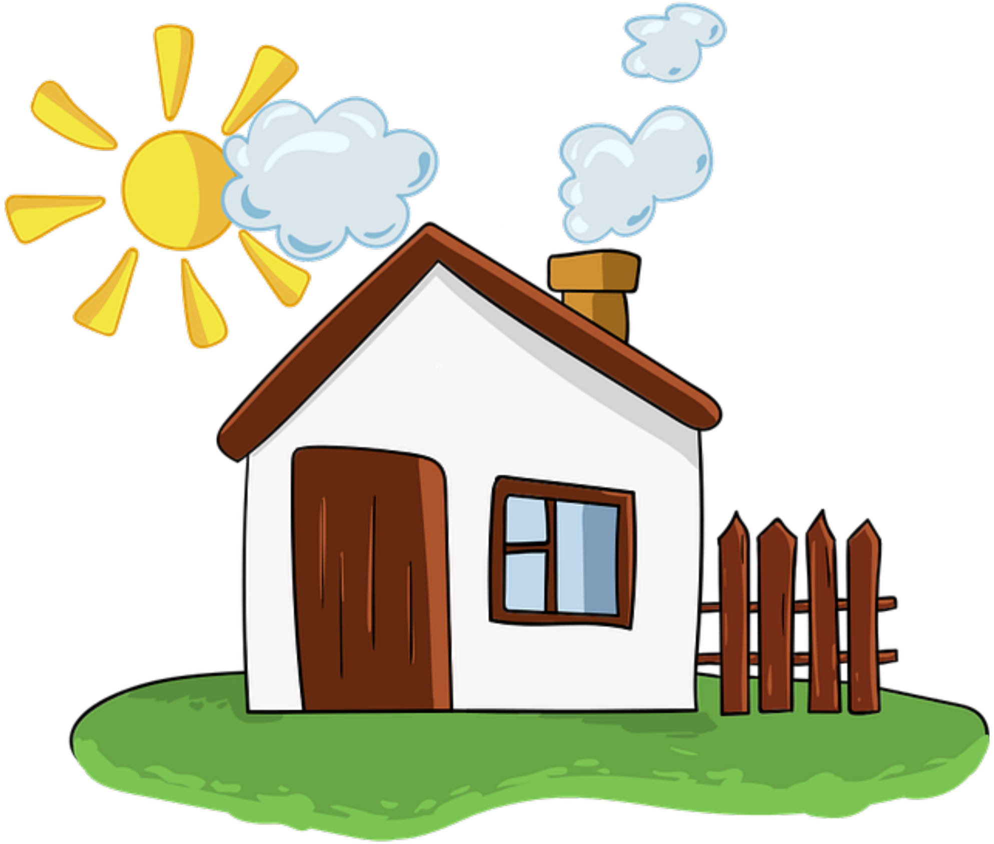
2014

2019

Comparing two photographs with compareable method and eqipment you can observe differences, but further knowledge is limited

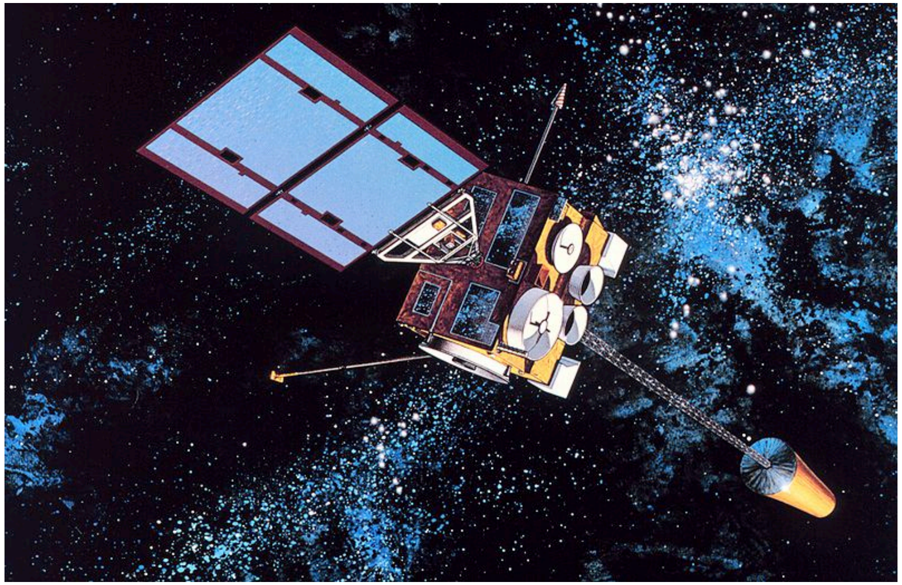


2014



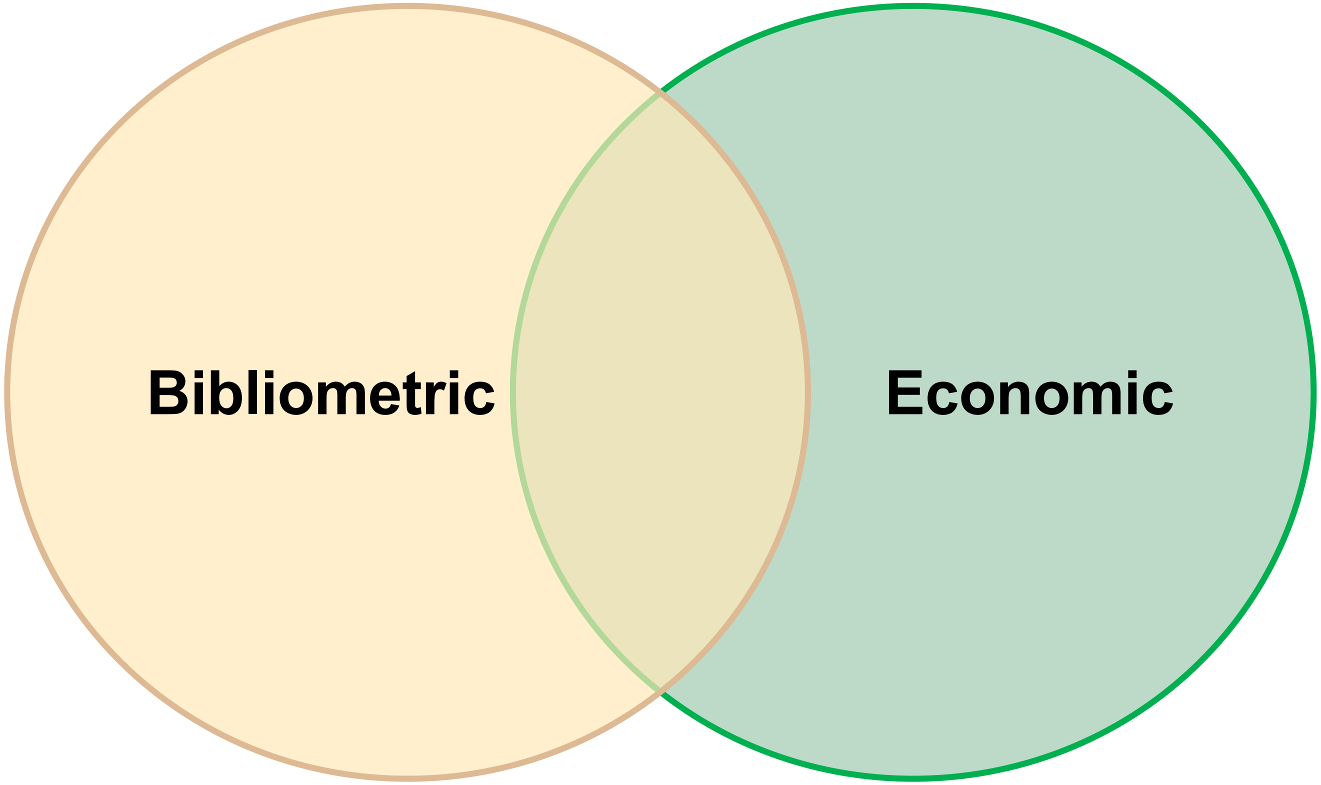
2019

So what kind of indexing/data collection method would need to be designed to improve the OA data environment?



My perspective & structure for my talk: Necessary components of OA data

HANKEN



1. Bibliometric

What? Why is this important?

- Readily available bibliometric data about scholarly publishing and open access 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 key obstacles in current journal indexing services

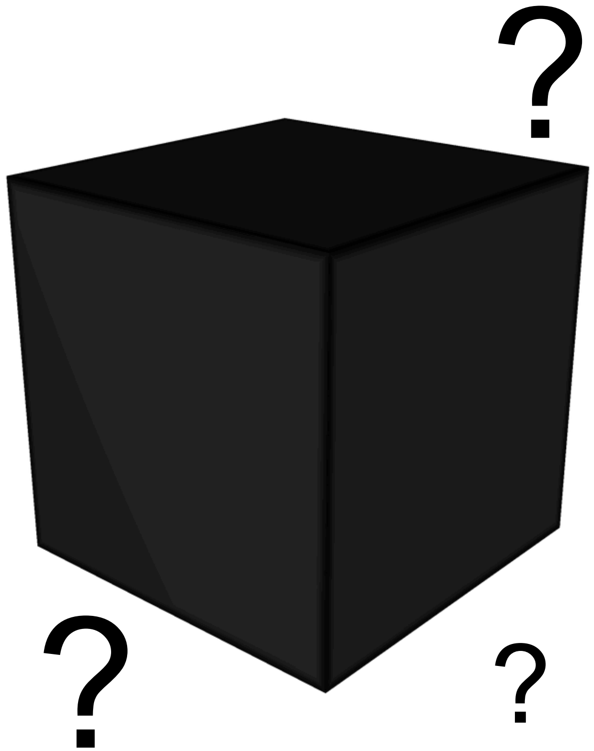
- **1. Commercial dominance**

  - Access to the most comprehensive databases, e.g. Web of Science, Scopus, and Ulrichsweb 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.)

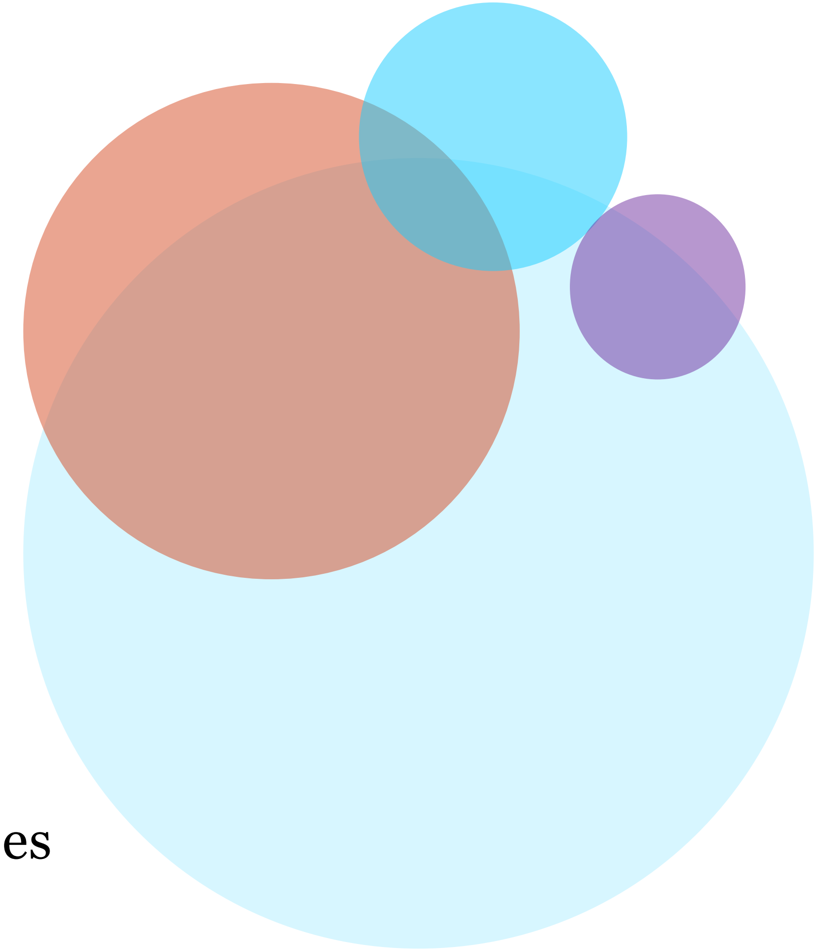
HANKEN

» **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.**

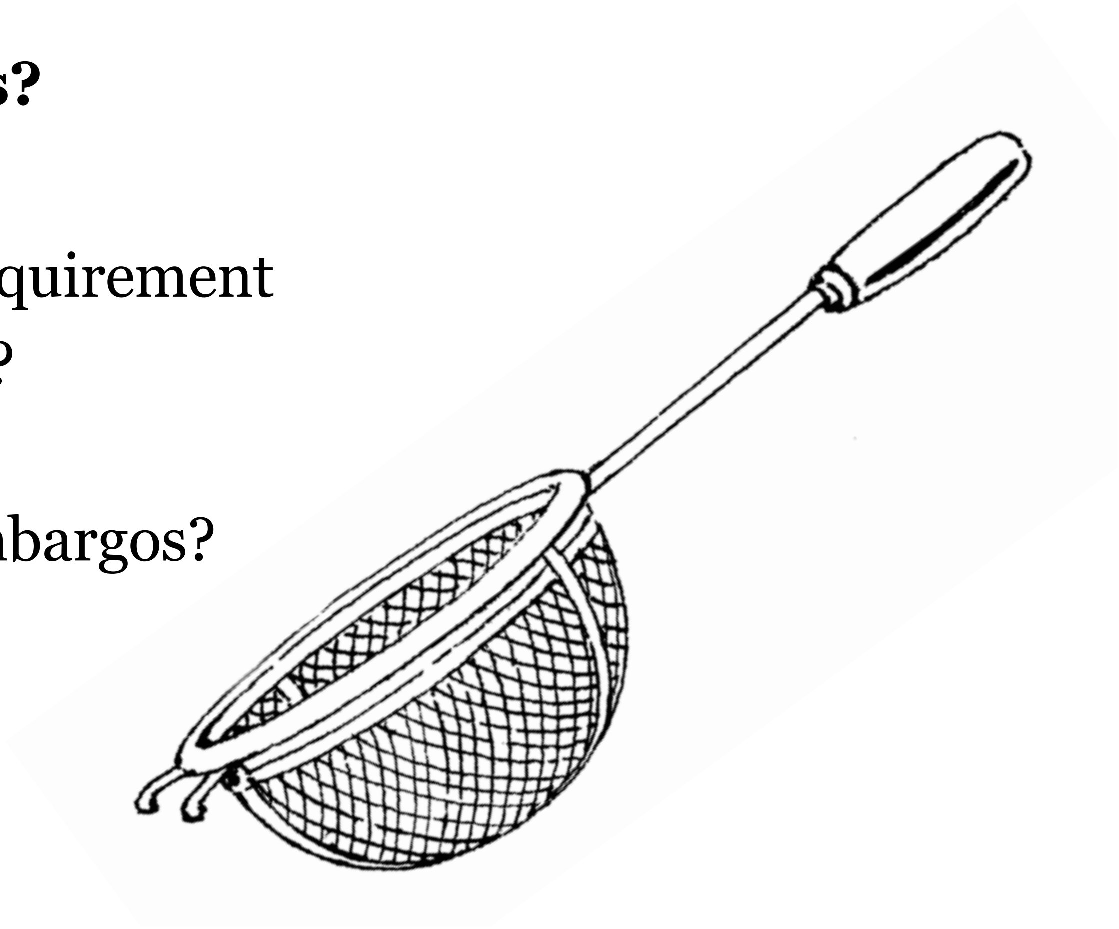
Various indexes/databases to choose from, all with different implications

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



Central questions for data on open access

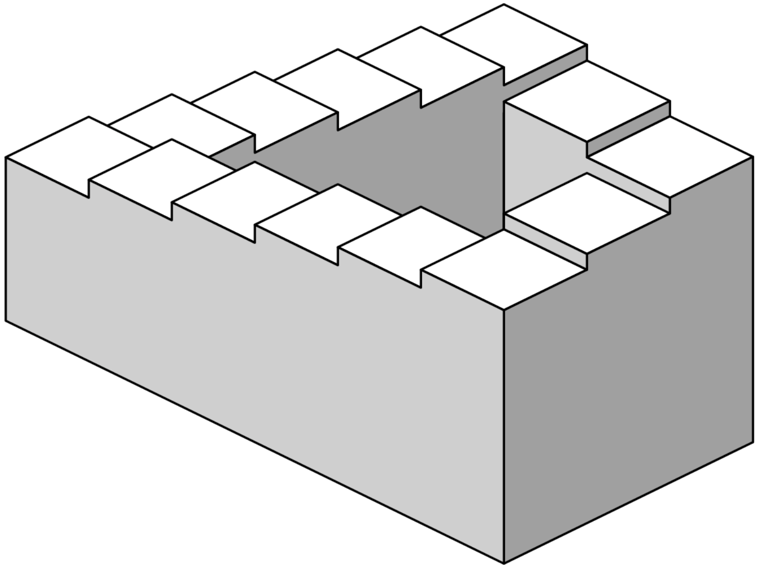
- **What is considered open access?**
  - Strict definition (incl.) license requirement
  - Basic requirement of free access?
  - Available by any means?
  - How to consider or adjust for embargos?



***Journal** vs **Article** perspectives to determining OA status*

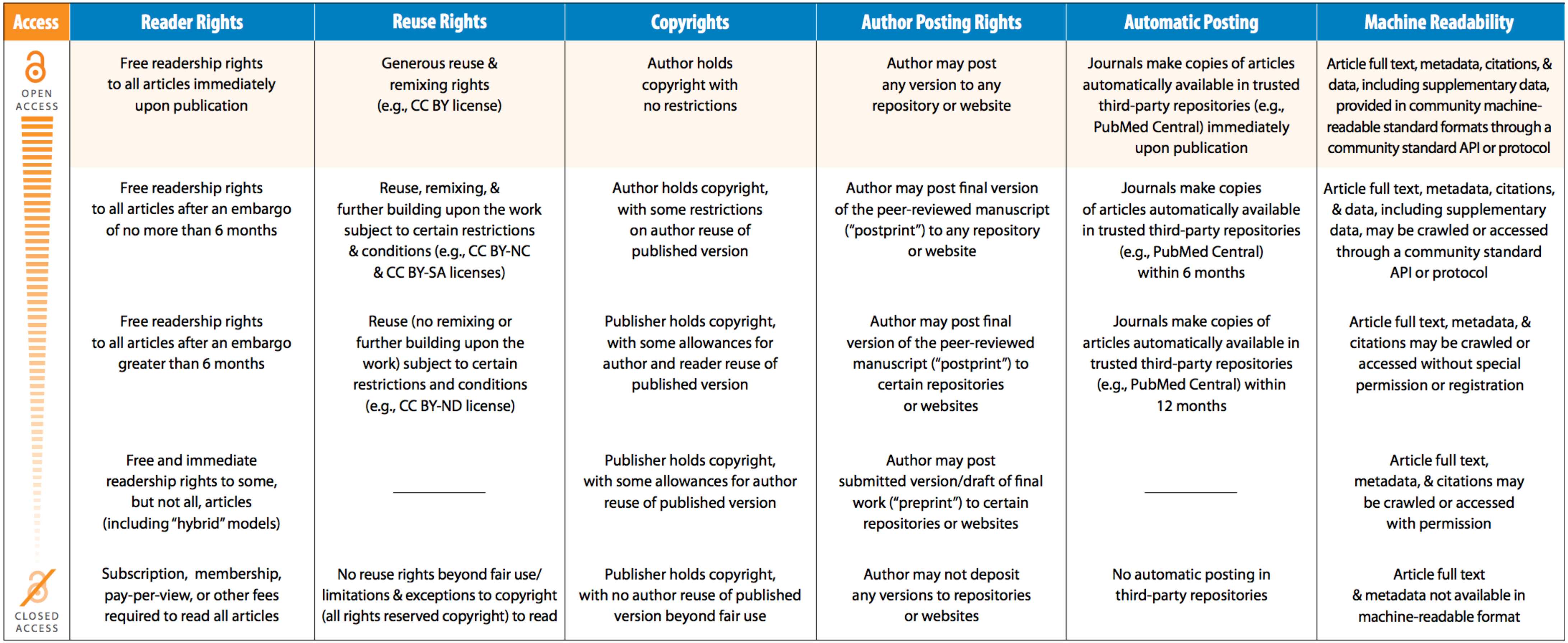
HANKEN

- **A complicated relationship**
- Partial openness of journals (e.g. Hybrid OA), green OA.
- Journals can and have dissapeared, merged, changed OA model, some articles might still be available online elsewhere.



The Open Acess Spectrum (OAS)

HANKEN

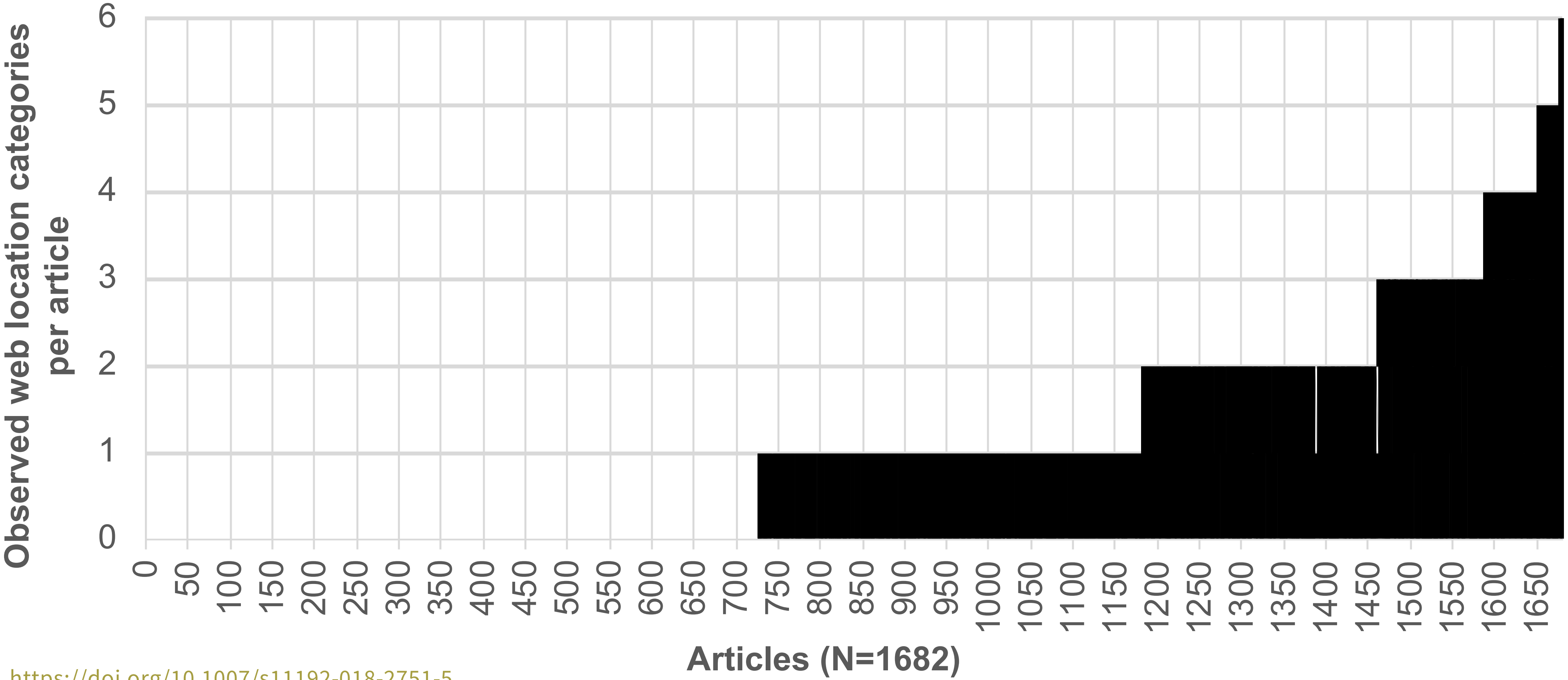


Access	Reader Rights	Reuse Rights	Copyrights	Author Posting Rights	Automatic Posting	Machine Readability
OPEN ACCESS	Free readership rights to all articles immediately upon publication	Generous reuse & remixing rights (e.g., CC BY license)	Author holds copyright with no restrictions	Author may post any version to any repository or website	Journals make copies of articles automatically available in trusted third-party repositories (e.g., PubMed Central) immediately upon publication	Article full text, metadata, citations, & data, including supplementary data, provided in community machine-readable standard formats through a community standard API or protocol
	Free readership rights to all articles after an embargo of no more than 6 months	Reuse, remixing, & further building upon the work subject to certain restrictions & conditions (e.g., CC BY-NC & CC BY-SA licenses)	Author holds copyright, with some restrictions on author reuse of published version	Author may post final version of the peer-reviewed manuscript ("postprint") to any repository or website	Journals make copies of articles automatically available in trusted third-party repositories (e.g., PubMed Central) within 6 months	Article full text, metadata, citations, & data, including supplementary data, may be crawled or accessed through a community standard API or protocol
	Free readership rights to all articles after an embargo greater than 6 months	Reuse (no remixing or further building upon the work) subject to certain restrictions and conditions (e.g., CC BY-ND license)	Publisher holds copyright, with some allowances for author and reader reuse of published version	Author may post final version of the peer-reviewed manuscript ("postprint") to certain repositories or websites	Journals make copies of articles automatically available in trusted third-party repositories (e.g., PubMed Central) within 12 months	Article full text, metadata, & citations may be crawled or accessed without special permission or registration
	Free and immediate readership rights to some, but not all, articles (including "hybrid" models)	————	Publisher holds copyright, with some allowances for author reuse of published version	Author may post submitted version/draft of final work ("preprint") to certain repositories or websites	————	Article full text, metadata, & citations may be crawled or accessed with permission
CLOSED ACCESS	Subscription, membership, pay-per-view, or other fees required to read all articles	No reuse rights beyond fair use/ limitations & exceptions to copyright (all rights reserved copyright) to read	Publisher holds copyright, with no author reuse of published version beyond fair use	Author may not deposit any versions to repositories or websites	No automatic posting in third-party repositories	Article full text & metadata not available in machine-readable format

http://sparcopen.org/our-work/howopenisit/

Chen and Olijhoek (2016)

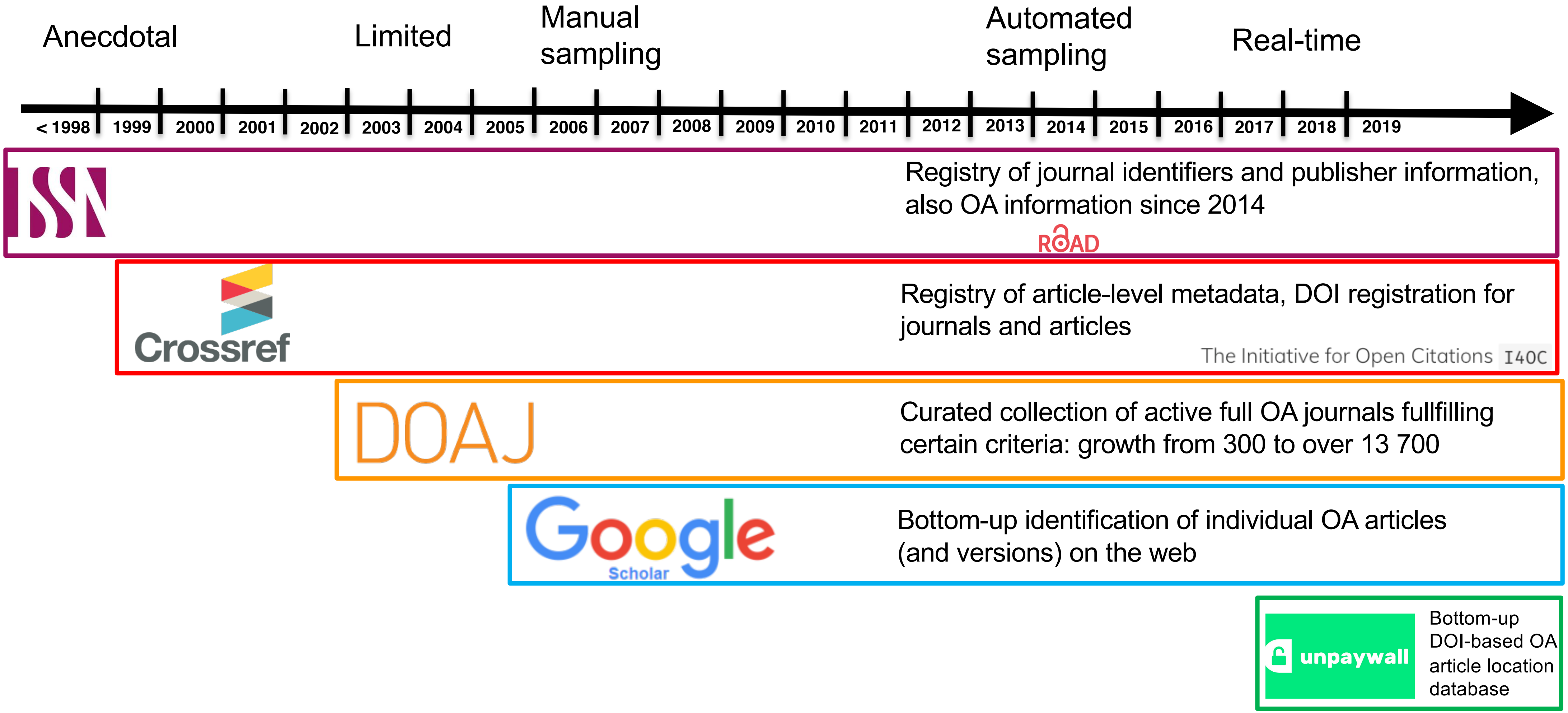
Considerable overlap in OA mechanisms, what should be registered?



https://doi.org/10.1007/s11192-018-2751-5

Timeline of key OA data sources and main methodoligies of published studies

HANKEN



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

ISSN

Registry of journal identifiers and publisher information, also OA information since 2014

ROAD

Crossref

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

The Initiative for Open Citations I4OC

DOAJ

Curated collection of active full OA journals fullfilling certain criteria: growth from 300 to over 13 700

Google Scholar

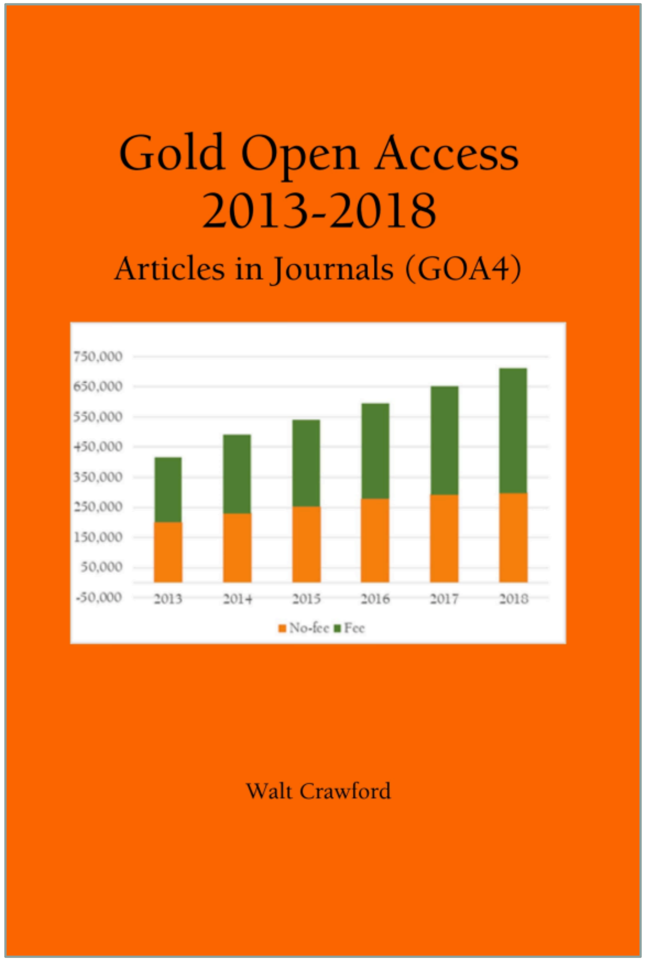
Bottom-up identification of individual OA articles (and versions) on the web

unpaywall

Bottom-up DOI-based OA article location database

State-of-the-art insight on OA journals

HANKEN



Gold Open Access 2013-2018

Articles in Journals (GOA4)

Walt Crawford

The most comprehensive mapping of open access journals has been put together manually by visiting over 13 000 journal websites and counting the number of articles published.

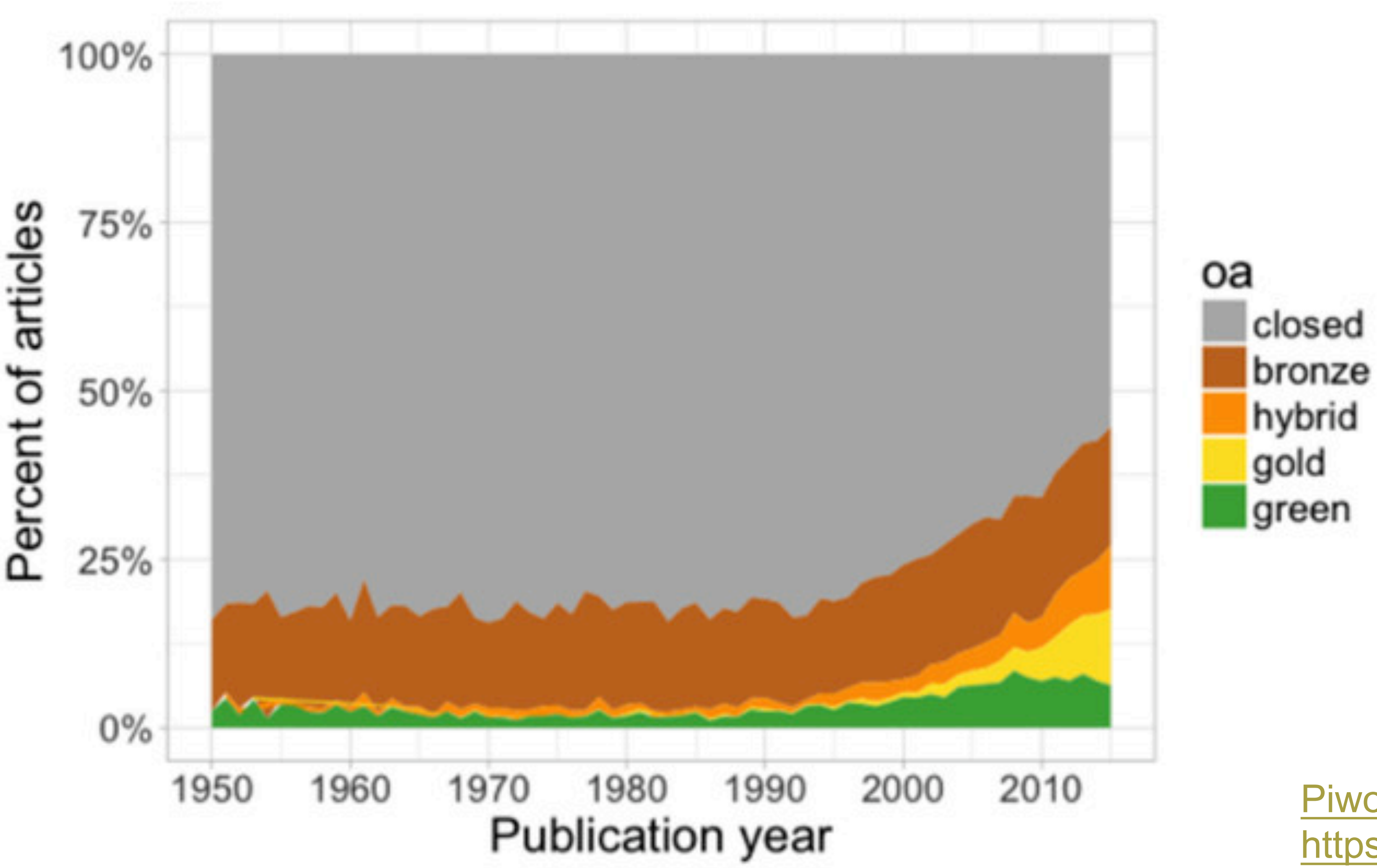
While a tremendous effort and resource, should getting data on open access be reliant on manual data collection?

DOAJ is also just a subset of all OA journals.

https://waltcrawford.name/goa4.pdf

State-of-the-art of insight on article-level OA

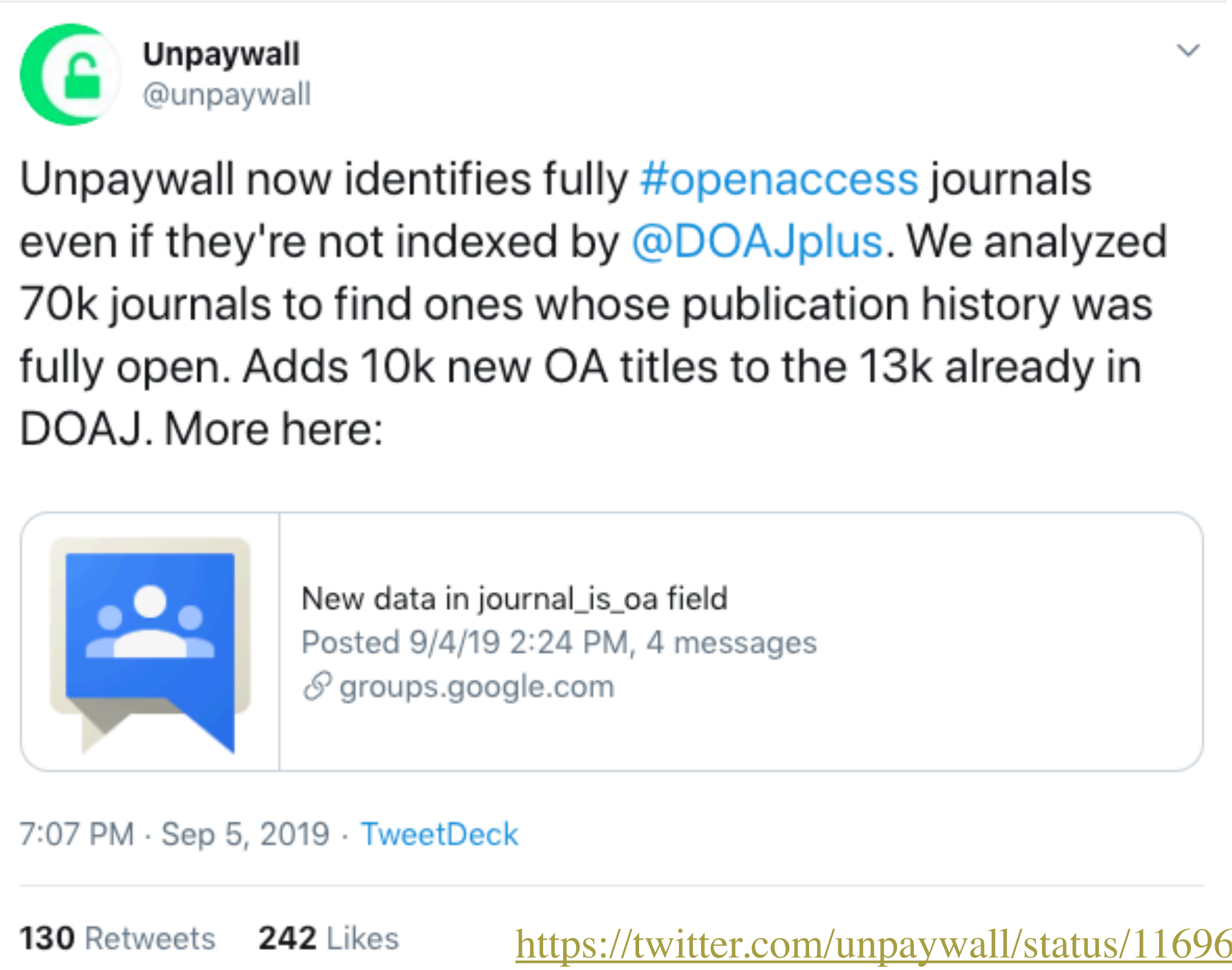
HANKEN



unpaywall

Piwowar et al (2018)
https://doi.org/10.7717/peerj.4375

The future looks bright for more complete identification of full OA journals



Unpaywall
@unpaywall

Unpaywall now identifies fully #openaccess journals even if they're not indexed by @DOAJplus. We analyzed 70k journals to find ones whose publication history was fully open. Adds 10k new OA titles to the 13k already in DOAJ. More here:

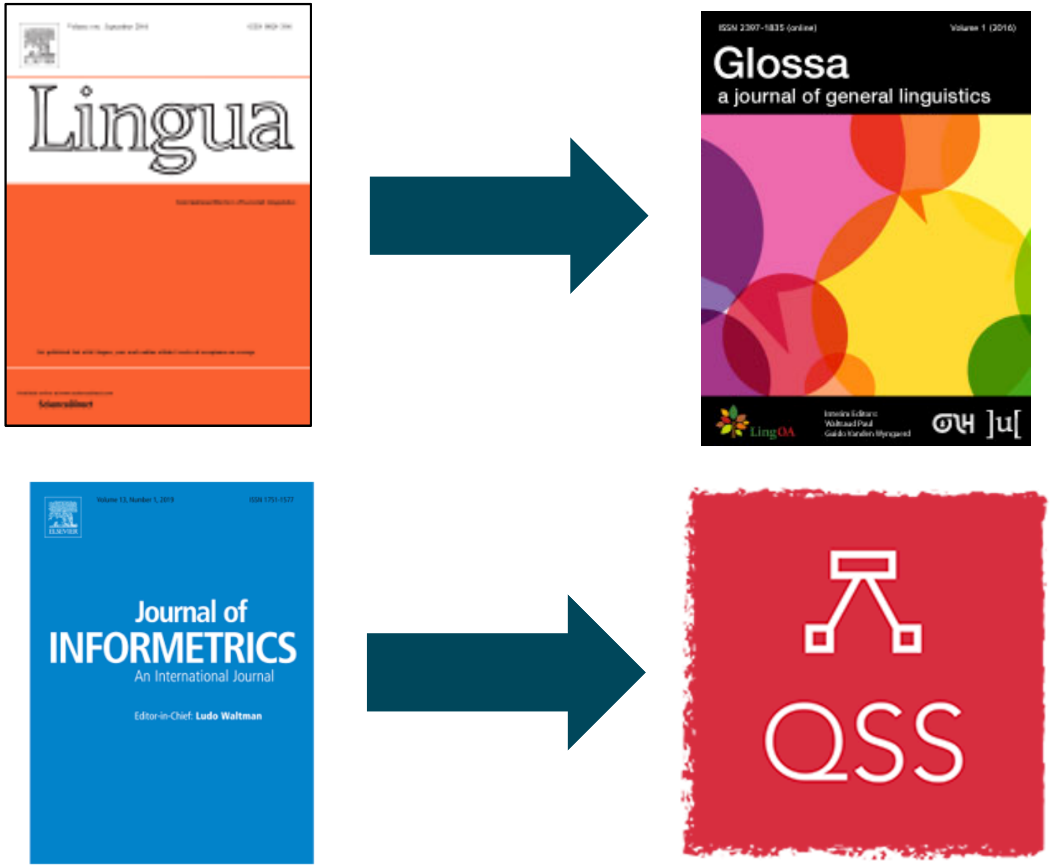
New data in journal_is_oa field
Posted 9/4/19 2:24 PM, 4 messages
groups.google.com

7:07 PM · Sep 5, 2019 · TweetDeck

130 Retweets **242** Likes

https://twitter.com/unpaywall/status/1169643265966137348?s=20

But how to represent in data e.g. journal editorial board transitions?

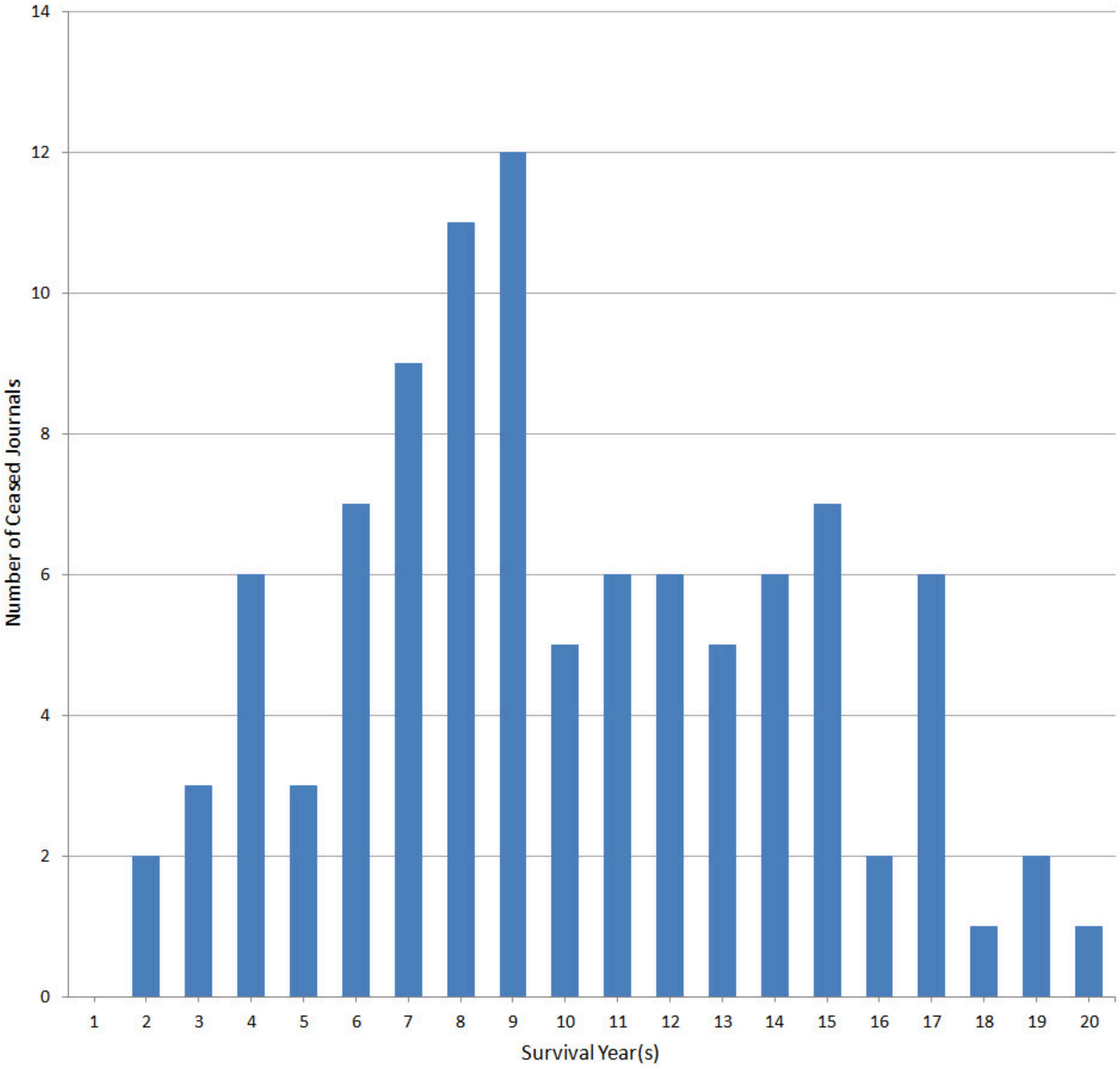


Journals come and go, but who keeps track?

- Not just an issue for preservation, but for understanding how the ground is shifting.

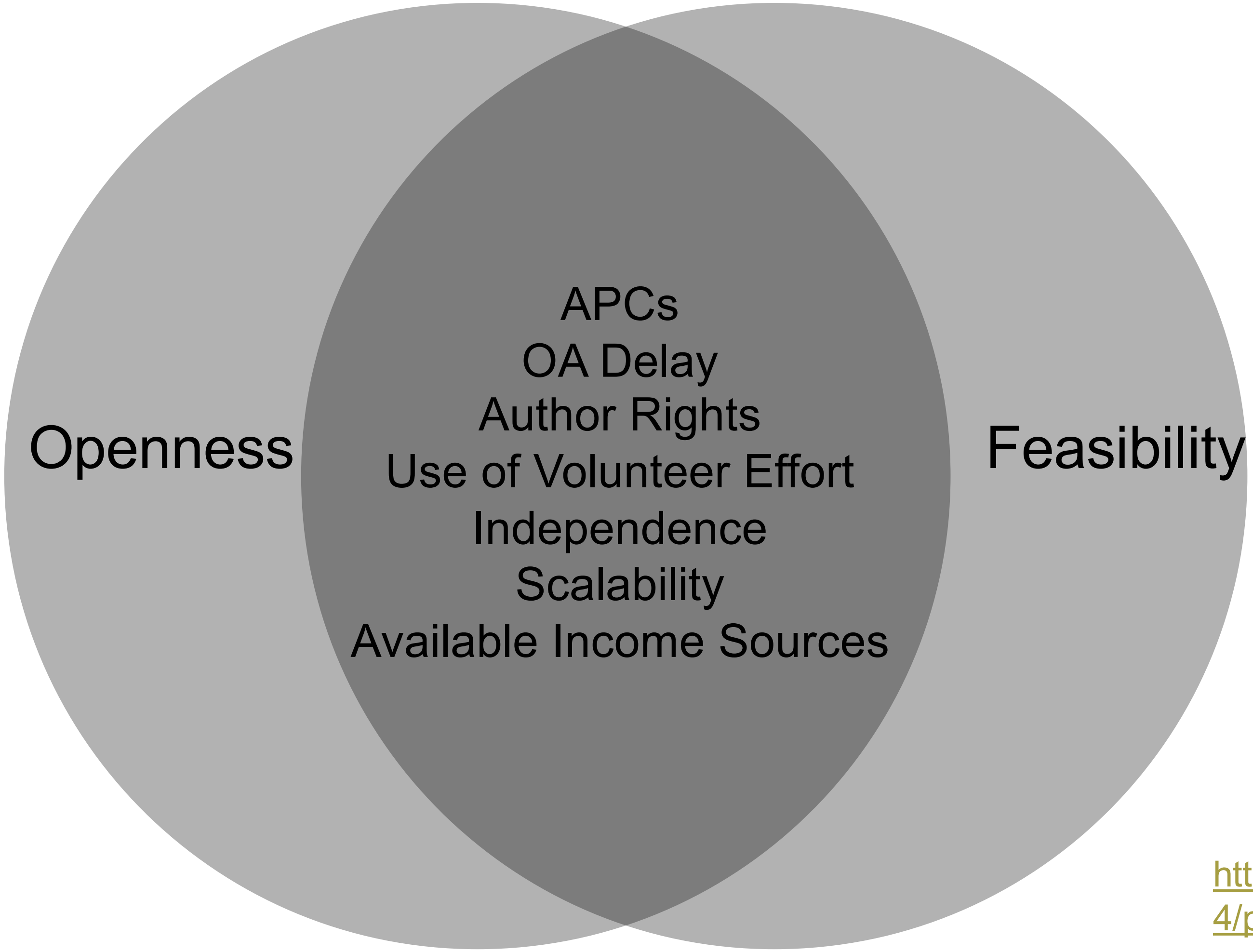
- The figure shows the mortality of 250 OA journals started prior to 2002, 51% were still active in 2014

https://doi.org/10.7717/peerj.1990



Open access is often weighed against other factors in journal publishing

HANKEN



Openness

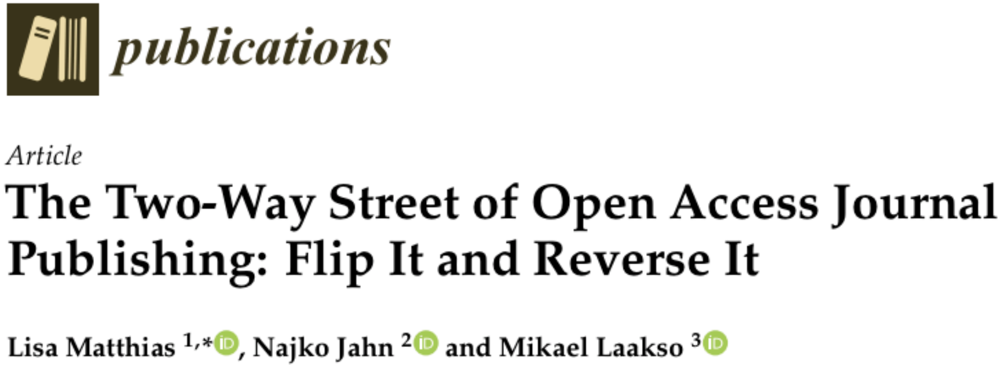
APCs
OA Delay
Author Rights
Use of Volunteer Effort
Independence
Scalability
Available Income Sources

Feasibility

http://www.informationr.net/ir/22-4/paper773.html

Journals also reverse-flip

HANKEN



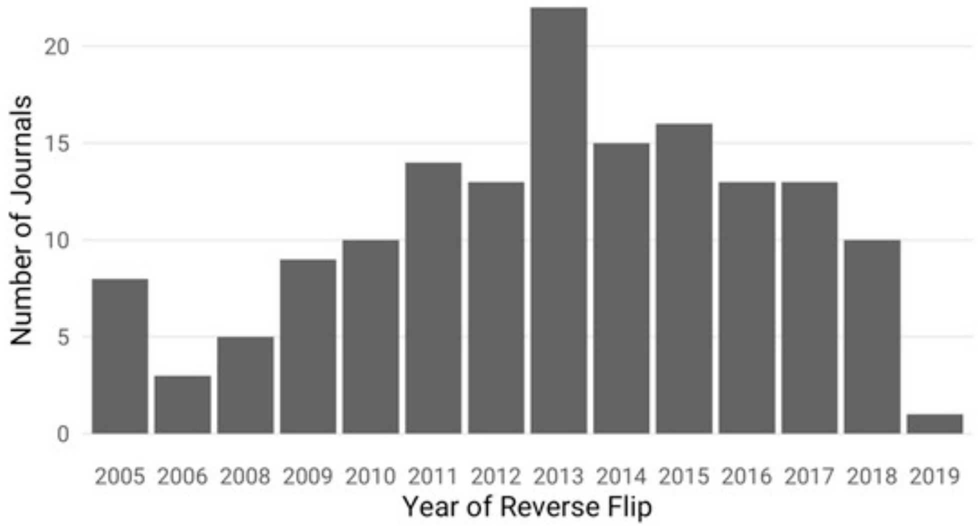
publications

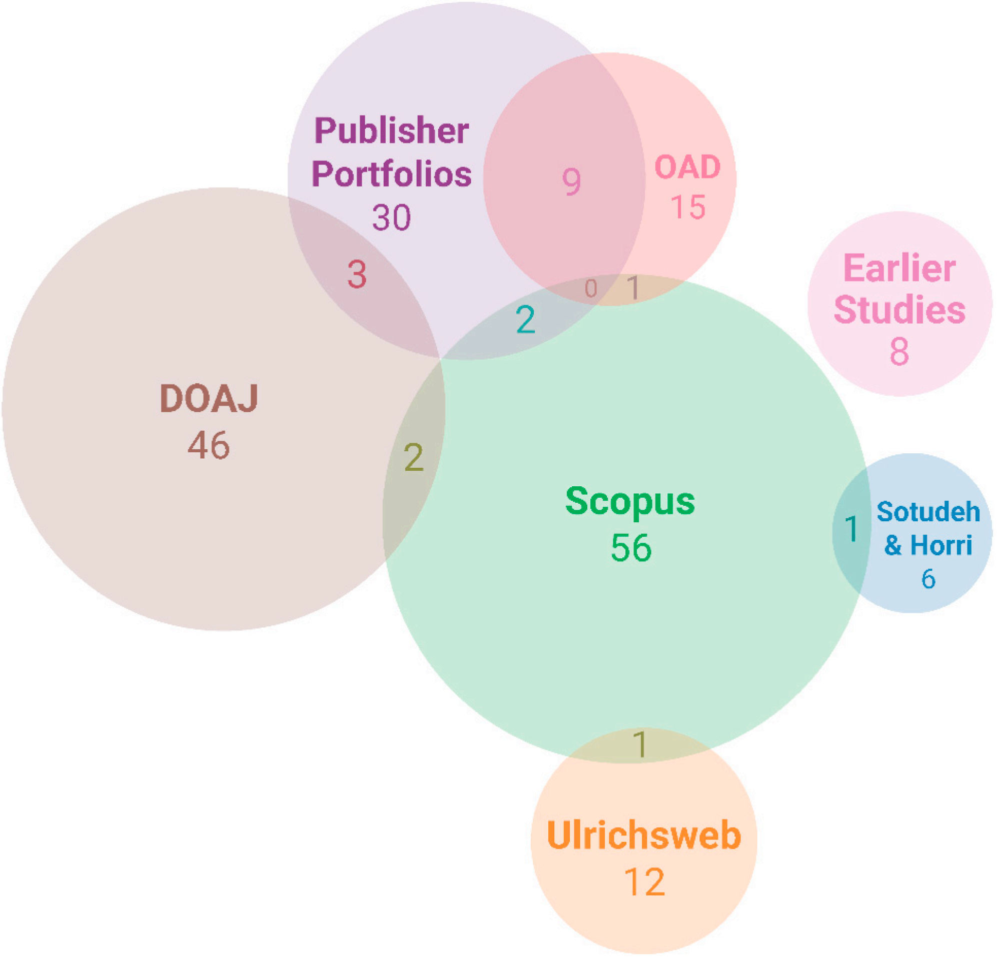
Article

The Two-Way Street of Open Access Journal Publishing: Flip It and Reverse It

Lisa Matthias [1,*], Najko Jahn [2] and Mikael Laakso [3]

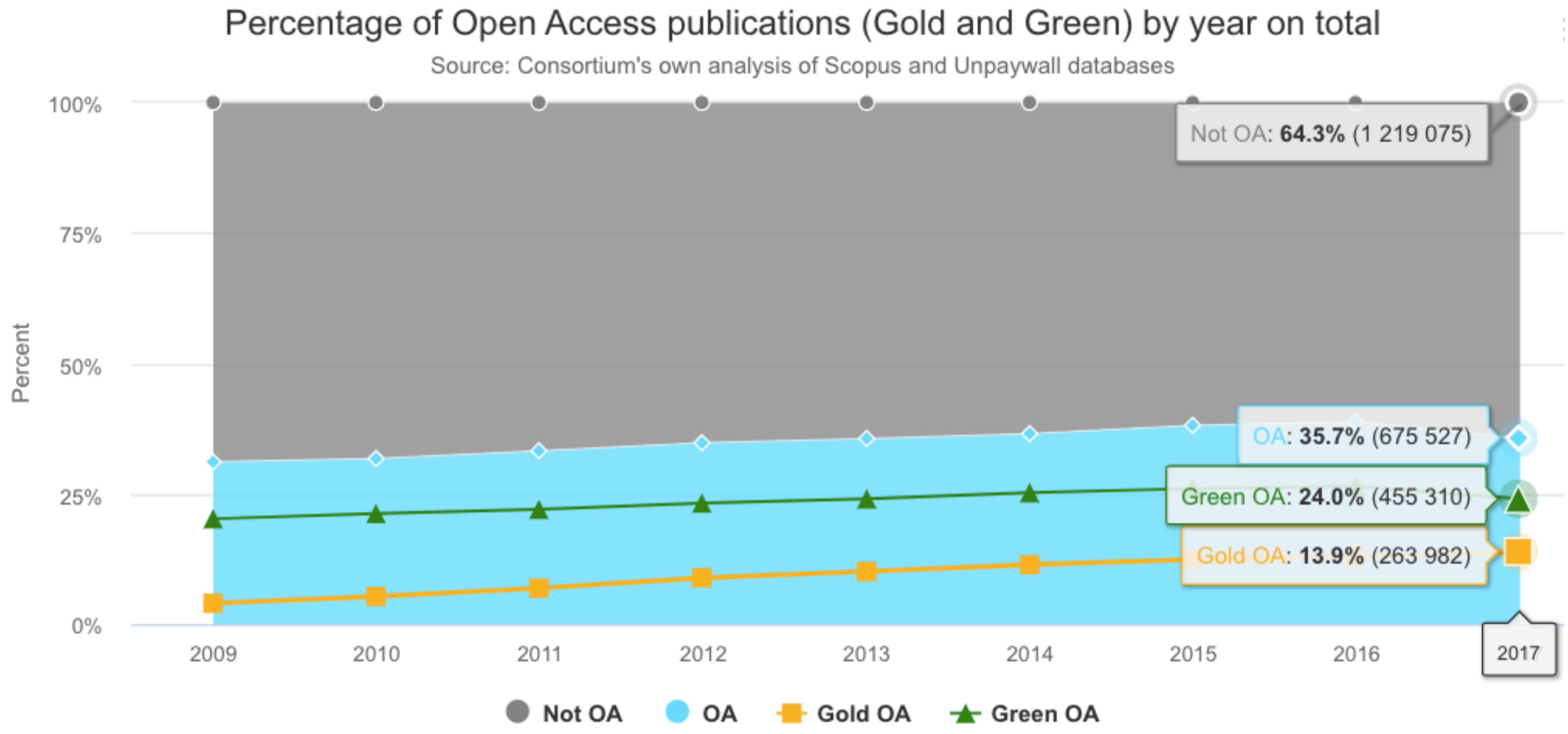
Publications **2019** https://doi.org/10.3390/publications7020023





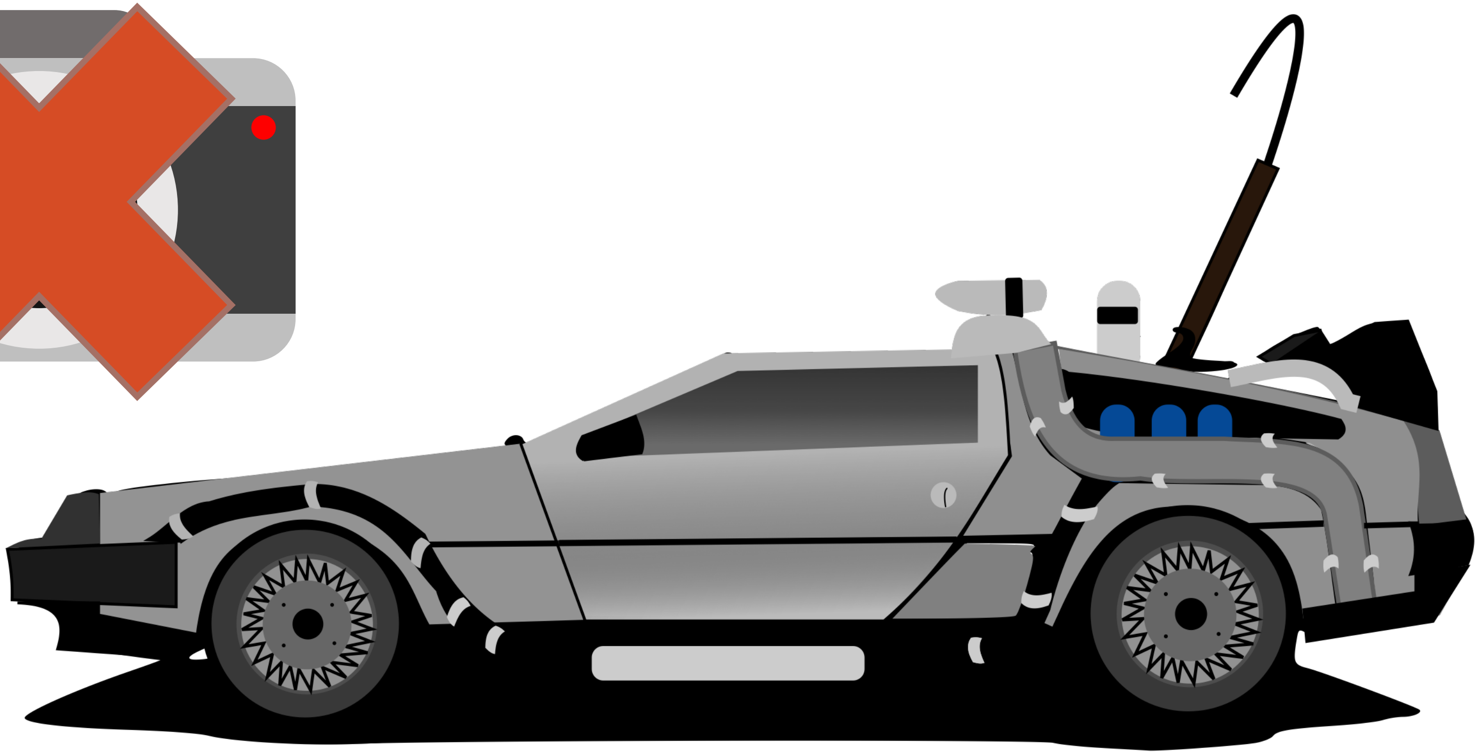
High-level open access monitoring can currently only tell us so much

HANKEN



https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/trends-open-access-publications_en

In lack of comprehensive data, what is the most viable way to gain information about OA development of the past?

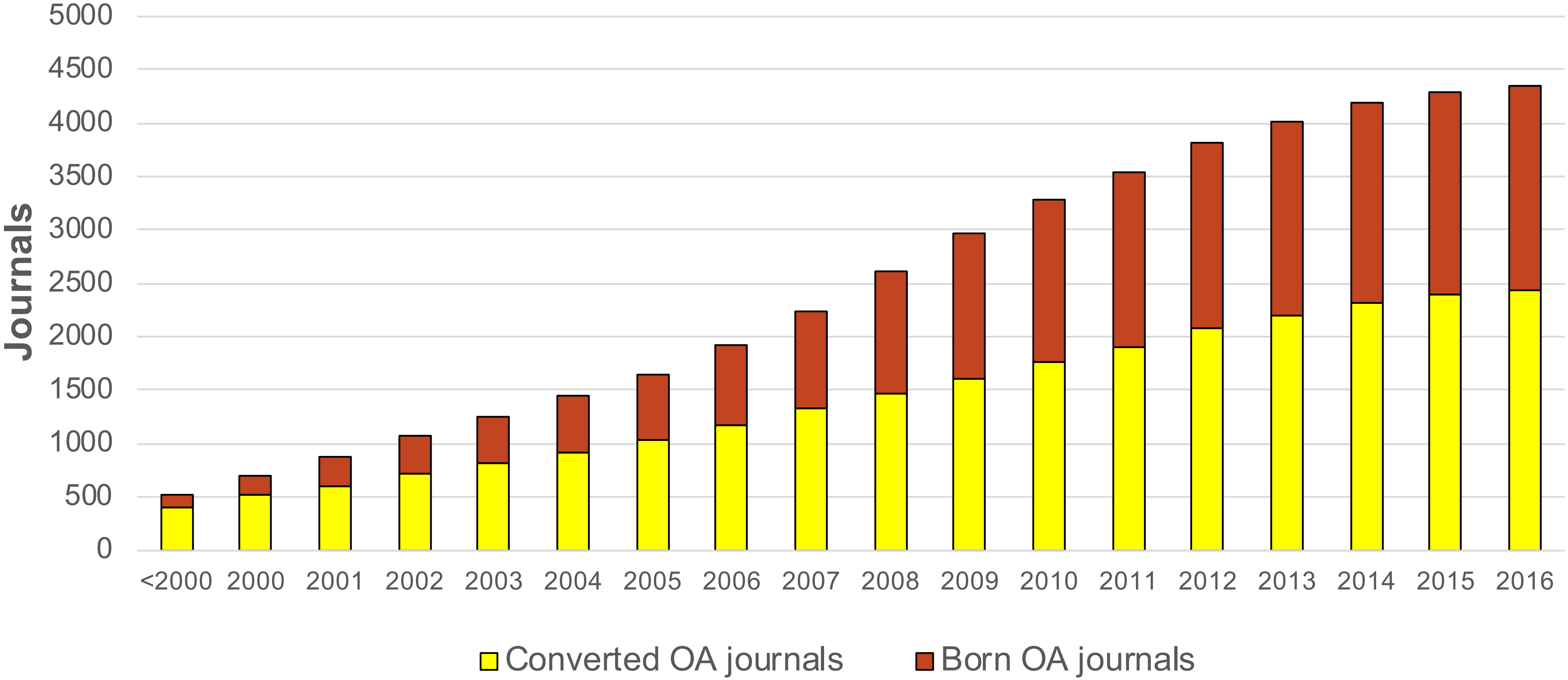


INTERNET ARCHIVE

WayBackMachine

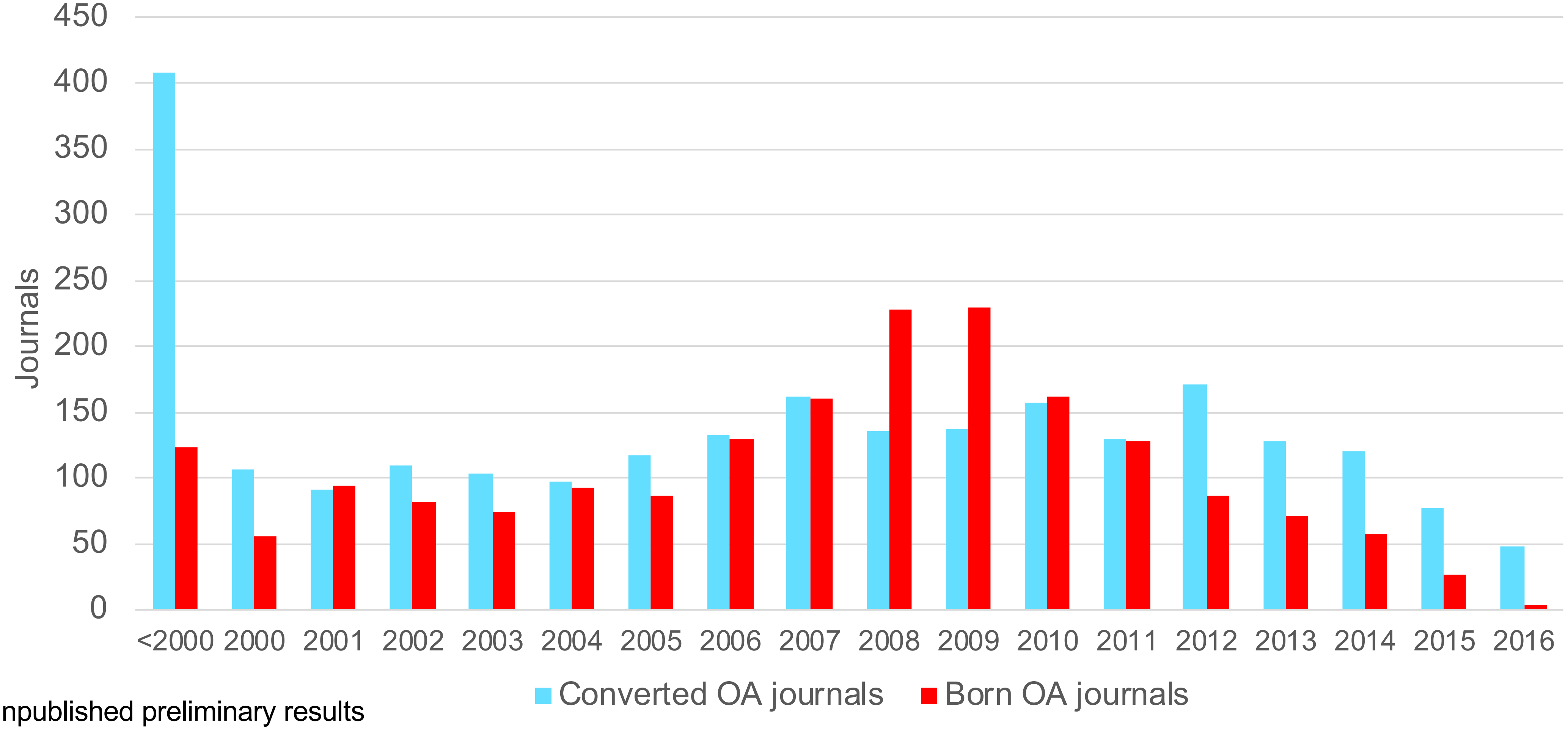
Is the journal landscape shifting or is it just growing? (Scopus OA journals)

HANKEN



When did journals start OA publishing? (Scopus OA journals)

HANKEN



Unpublished preliminary results

Economic

What? Why is this important?

HANKEN

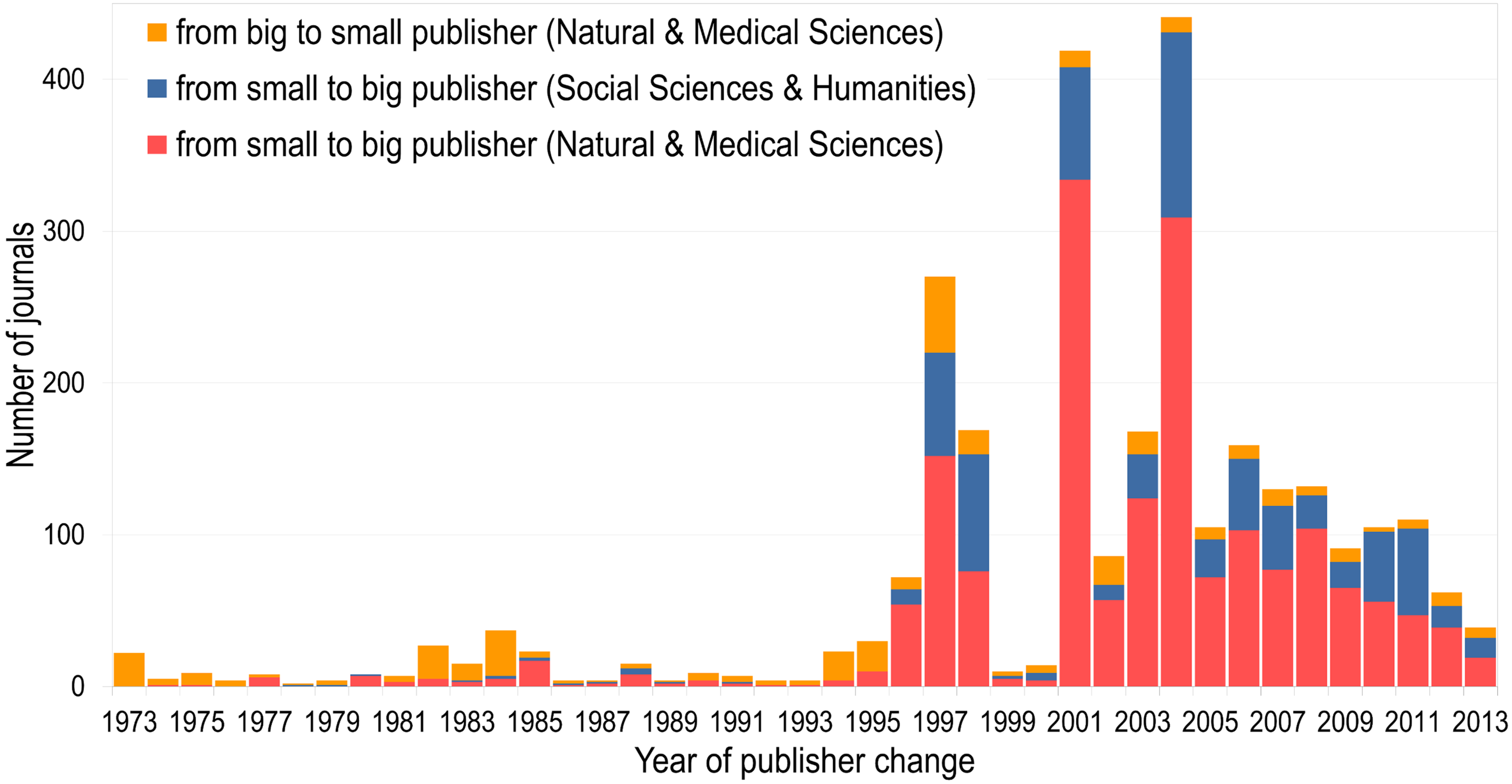
- The issue of money has been intimately tied to OA from early on, yet there is only limited knowledge and experience about how to align the two.
- Price and cost transparency is of benefit to everyone one else other than publishers who seek financial gain by not making such information readily available.
- Gain added perception of **cost vs price**, thus making additional value added by providers more observable.

A hard fact

- **Commercial companies,** particularly publicly traded, **are out to increase profits and seek growth.**
- That is what makes shareholders happy and the leadership of the companies keep their jobs.
- This growth can come from expanding business into new areas, or it can come from increasing market share and/or prices in existing segments.
- There is evidence of both strategies happening.

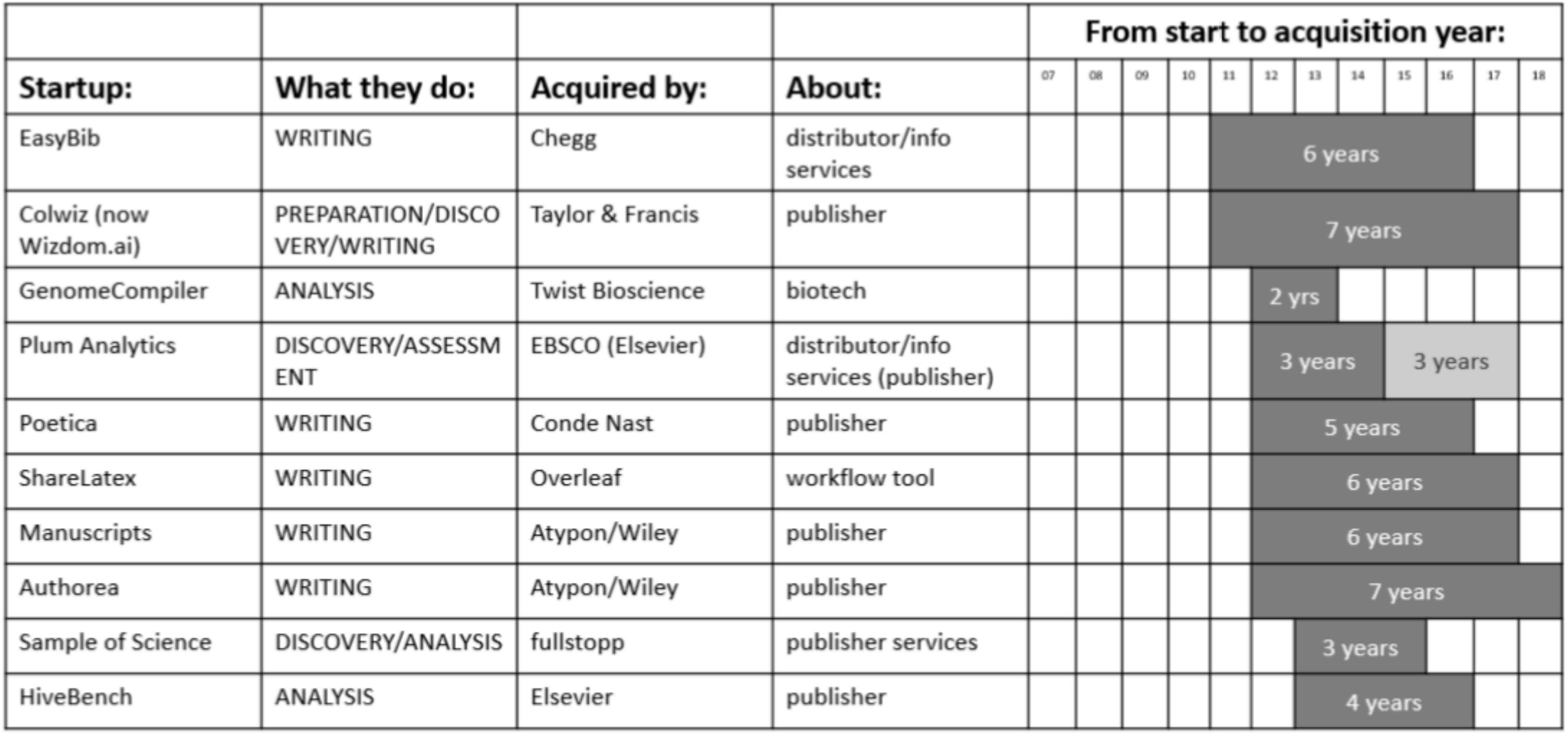
The big have gotten bigger

HANKEN



Larivière et al. (2015) https://doi.org/10.1371/journal.pone.0127502

Many new startups in scholarly communications are aquired by commercial publishers

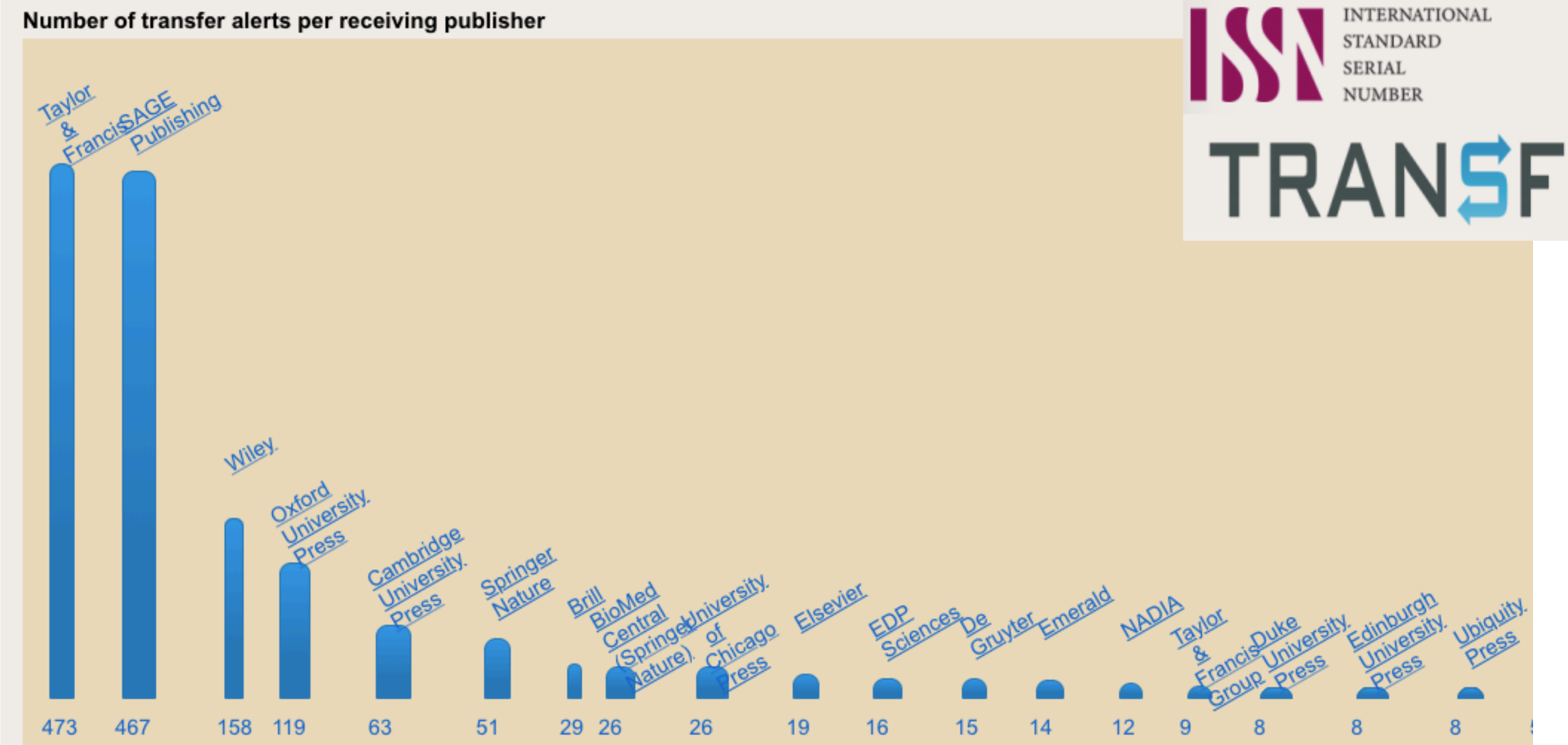


				From start to acquisition year:											
Startup:	**What they do:**	**Acquired by:**	**About:**	07	08	09	10	11	12	13	14	15	16	17	18
EasyBib	WRITING	Chegg	distributor/info services					6 years							
Colwiz (now Wizdom.ai)	PREPARATION/DISCOVERY/WRITING	Taylor & Francis	publisher					7 years							
GenomeCompiler	ANALYSIS	Twist Bioscience	biotech						2 yrs						
Plum Analytics	DISCOVERY/ASSESSMENT	EBSCO (Elsevier)	distributor/info services (publisher)						3 years			3 years			
Poetica	WRITING	Conde Nast	publisher						5 years						
ShareLatex	WRITING	Overleaf	workflow tool						6 years						
Manuscripts	WRITING	Atypon/Wiley	publisher						6 years						
Authorea	WRITING	Atypon/Wiley	publisher						7 years						
Sample of Science	DISCOVERY/ANALYSIS	fullstopp	publisher services							3 years					
HiveBench	ANALYSIS	Elsevier	publisher							4 years					

Campfens (2019) https://doi.org/10.31219/osf.io/a78zj

…and the general trend concerning journals seems to continue

HANKEN


Number of transfer alerts per receiving publisher

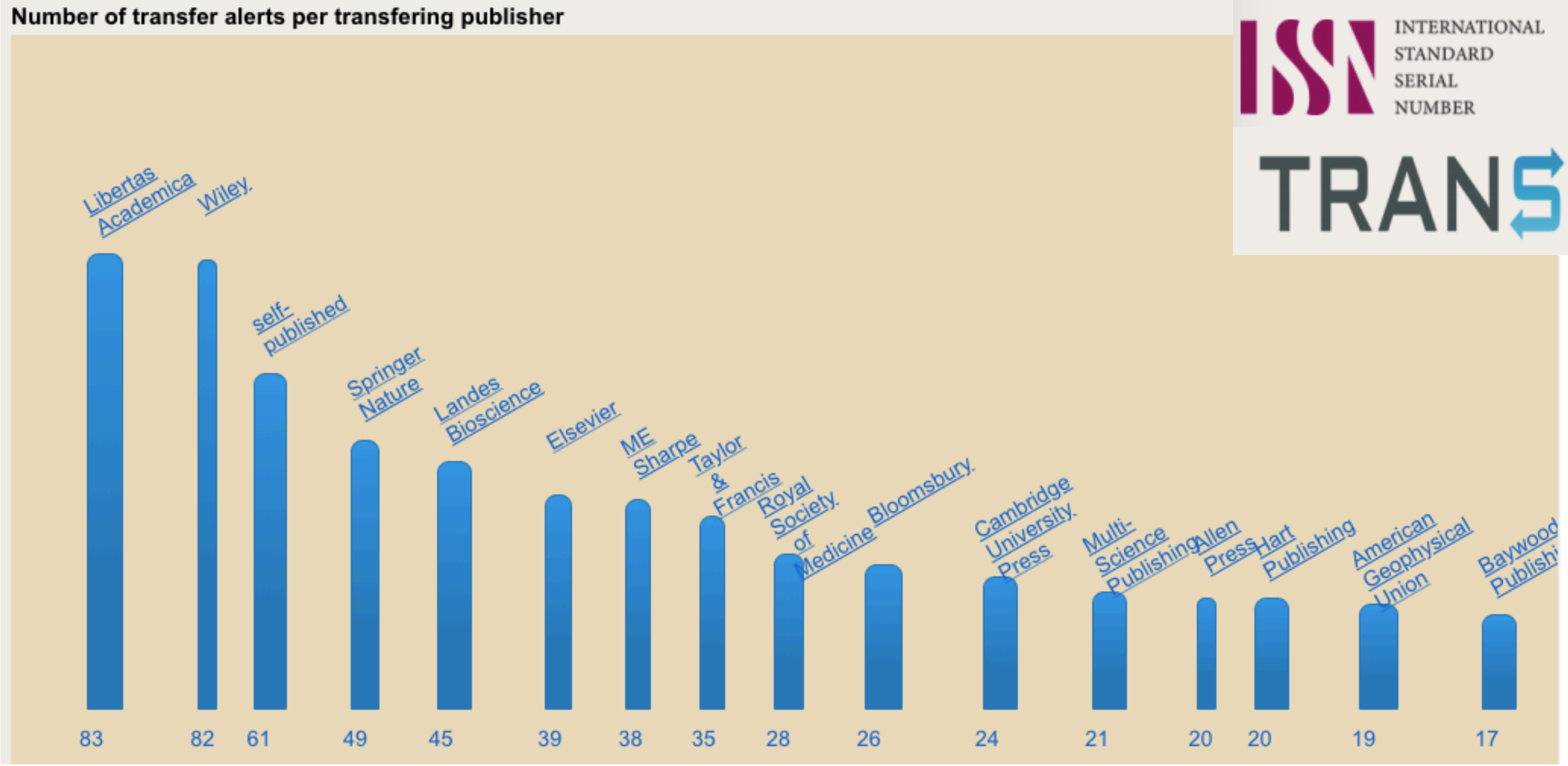
ISSN INTERNATIONAL STANDARD SERIAL NUMBER

TRANSFER

https://journaltransfer.issn.org/statistics

…and the general trend concerning journals seems to continue (cont.)

HANKEN



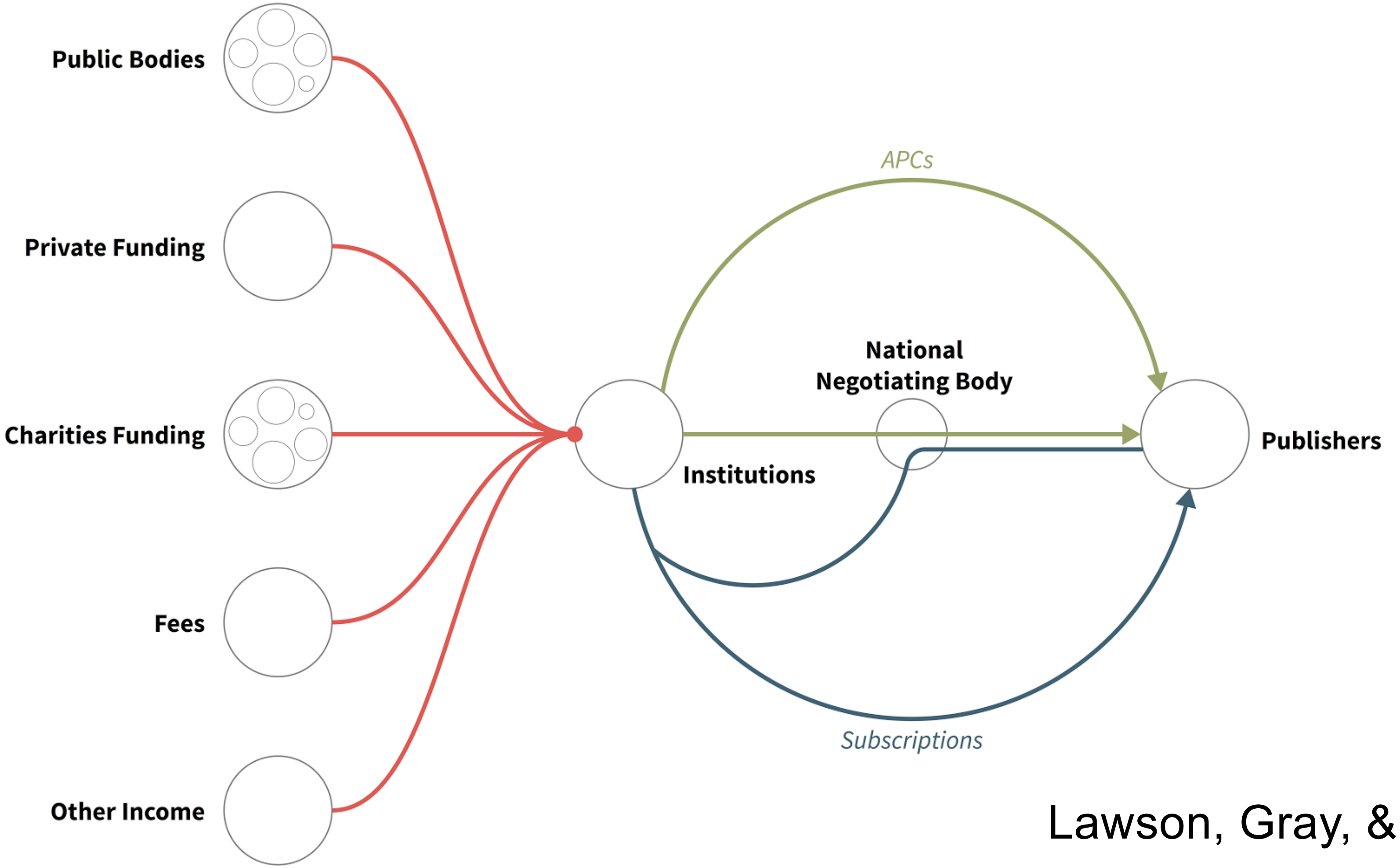
Number of transfer alerts per transfering publisher

ISSN INTERNATIONAL STANDARD SERIAL NUMBER

TRANSFER

https://journaltransfer.issn.org/statistics

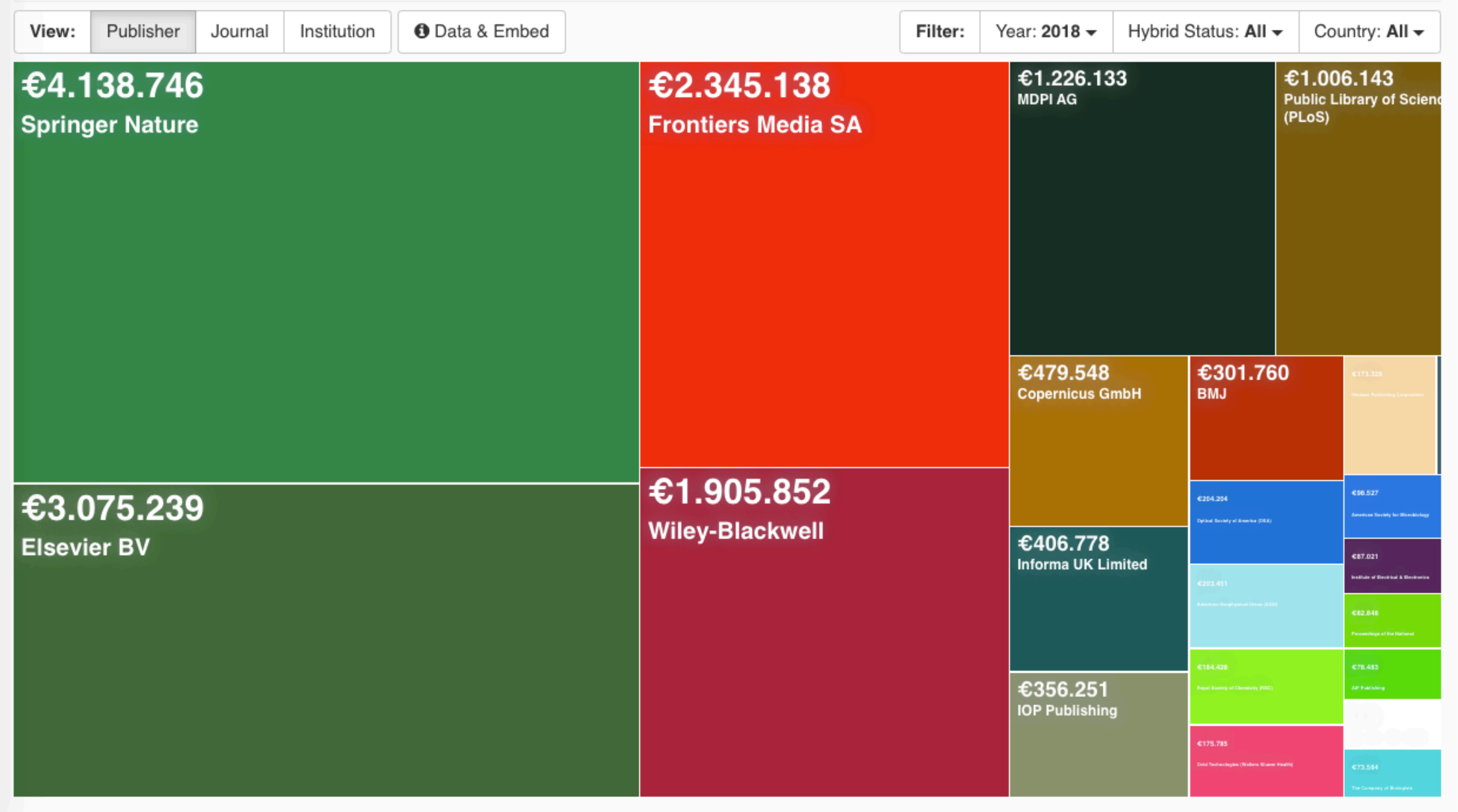
Model of financial flows in scholarly publishing



Lawson, Gray, & Mauri (2016)

Open APC – A great foundation

HANKEN



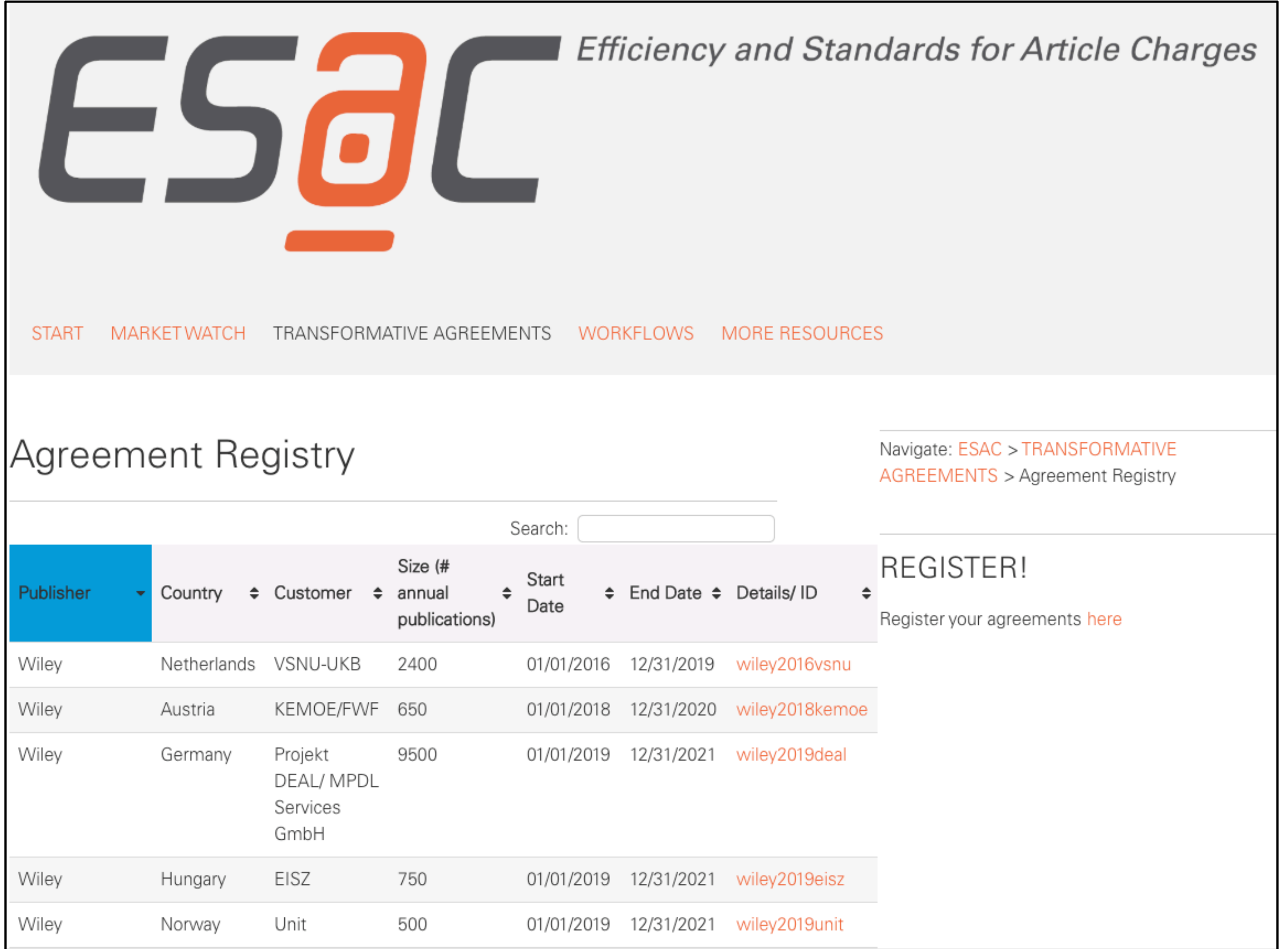
Year: 2018

Hybrid Status: All

Country: All

https://treemaps.intact-project.org/apcdata/combined/#publisher/period=2018

Agreements with publishers also increasingly transparent and cross compareable


ESAC
Efficiency and Standards for Article Charges
START
MARKET WATCH
TRANSFORMATIVE AGREEMENTS
WORKFLOWS
MORE RESOURCES
Search:

Agreement Registry

Navigate: ESAC > TRANSFORMATIVE AGREEMENTS > Agreement Registry

Publisher	Country	Customer	Size (# annual publications)	Start Date	End Date	Details/ ID
Wiley	Netherlands	VSNU-UKB	2400	01/01/2016	12/31/2019	wiley2016vsnu
Wiley	Austria	KEMOE/FWF	650	01/01/2018	12/31/2020	wiley2018kemoe
Wiley	Germany	Projekt DEAL/ MPDL Services GmbH	9500	01/01/2019	12/31/2021	wiley2019deal
Wiley	Hungary	EISZ	750	01/01/2019	12/31/2021	wiley2019eisz
Wiley	Norway	Unit	500	01/01/2019	12/31/2021	wiley2019unit

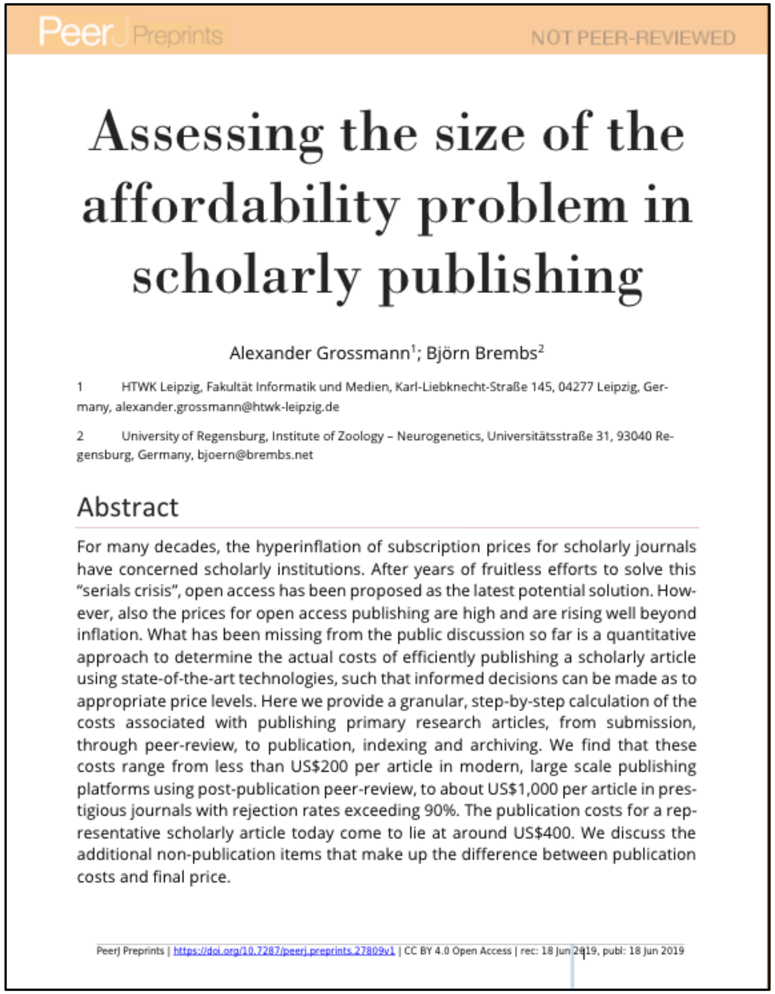
REGISTER!

Register your agreements here

https://esac-initiative.org/about/transformative-agreements/agreement-registry/

Could better data on actual costs be collected? What does publishing actually cost?

HANKEN



PeerJ Preprints NOT PEER-REVIEWED

Assessing the size of the affordability problem in scholarly publishing

Alexander Grossmann[1]; Björn Brembs[2]

1 HTWK Leipzig, Fakultät Informatik und Medien, Karl-Liebknecht-Straße 145, 04277 Leipzig, Germany, alexander.grossmann@htwk-leipzig.de

2 University of Regensburg, Institute of Zoology – Neurogenetics, Universitätsstraße 31, 93040 Regensburg, Germany, bjoern@brembs.net

Abstract

For many decades, the hyperinflation of subscription prices for scholarly journals have concerned scholarly institutions. After years of fruitless efforts to solve this “serials crisis”, open access has been proposed as the latest potential solution. However, also the prices for open access publishing are high and are rising well beyond inflation. What has been missing from the public discussion so far is a quantitative approach to determine the actual costs of efficiently publishing a scholarly article using state-of-the-art technologies, such that informed decisions can be made as to appropriate price levels. Here we provide a granular, step-by-step calculation of the costs associated with publishing primary research articles, from submission, through peer-review, to publication, indexing and archiving. We find that these costs range from less than US$200 per article in modern, large scale publishing platforms using post-publication peer-review, to about US$1,000 per article in prestigious journals with rejection rates exceeding 90%. The publication costs for a representative scholarly article today come to lie at around US$400. We discuss the additional non-publication items that make up the difference between publication costs and final price.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27809v1 | CC BY 4.0 Open Access | rec: 18 Jun 2019, publ: 18 Jun 2019

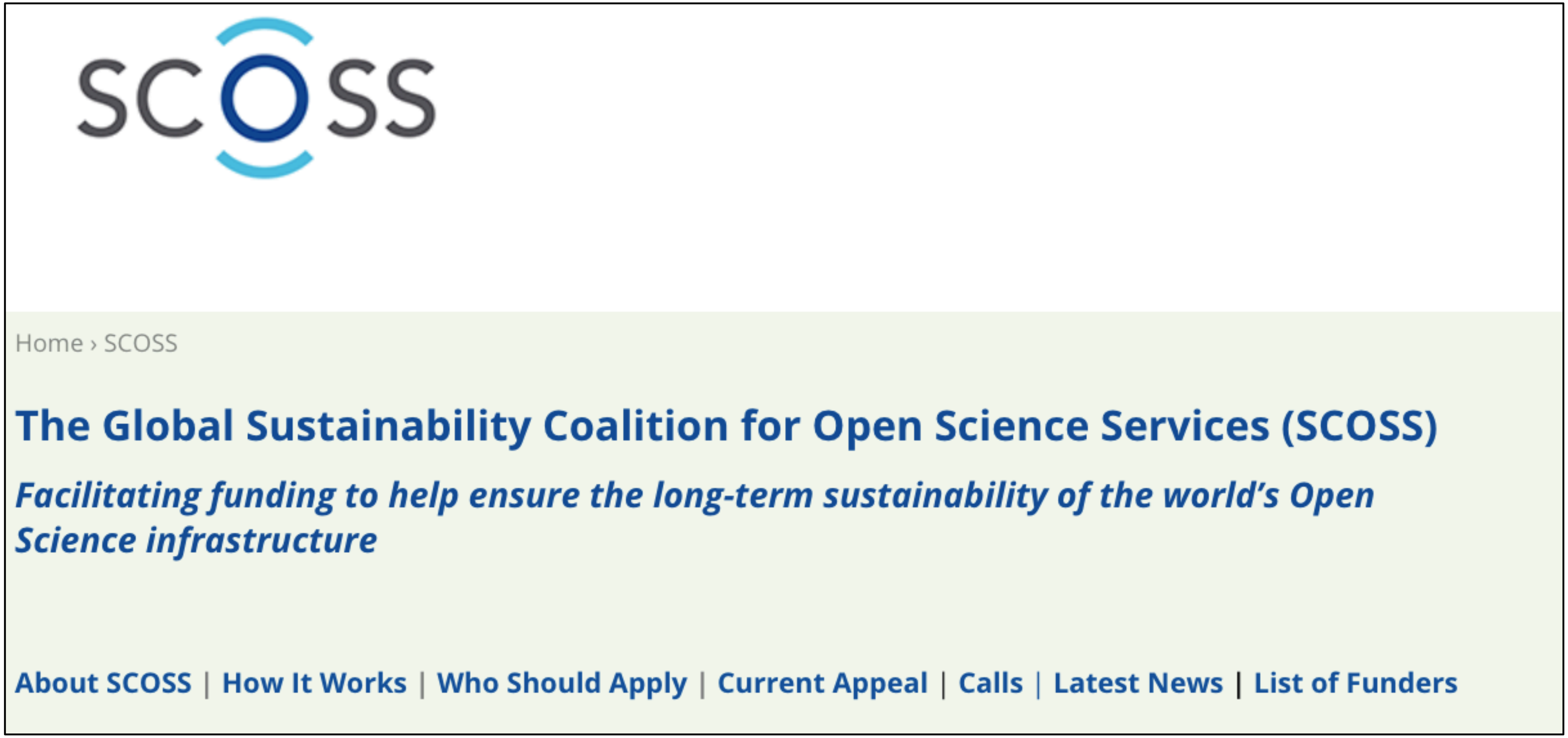
”[...] we provide a granular, step-by-step calculation of the costs associated with publishing primary research articles, from submission, through peer-review, to publication, indexing and archiving.”

”The publication costs for a representative scholarly article today come to lie at around **US$400**.”

https://doi.org/10.7287/peerj.preprints.27809v1

Grossmann & Brembs (2019)

There are mechanisms to fund open science infrastructure, but how to scale up in contributions and scope?



SCOSS

Home › SCOSS

The Global Sustainability Coalition for Open Science Services (SCOSS)

Facilitating funding to help ensure the long-term sustainability of the world's Open Science infrastructure

About SCOSS | **How It Works** | **Who Should Apply** | **Current Appeal** | **Calls** | **Latest News** | **List of Funders**

http://scoss.org/

Need for collective action – nationally and internationally

HANKEN

- **The Dilemma of Collective Action** (Wenzler 2017)
  - "For academic libraries to continue to achieve their traditional role of storing, organizing, preserving, and providing access to the scholarly record, they increasingly will have to take responsibility for the entire cycle of scholarly communication from publishing and editing through preservation, but it is unlikely that they will succeed in doing so through the uncoordinated actions of individual institutions and will require new experiments in cooperation and coordination."

- **The 2.5% Commitment** (Lewis 2017)
  - "…every academic library should commit to contribute 2.5% of its total budget to support the common infrastructure needed to create the open scholarly commons."
  - "…if we don't collectively invest in the infrastructure we need for the open scholarly commons, it will not get built or it will only be haphazardly half built. "

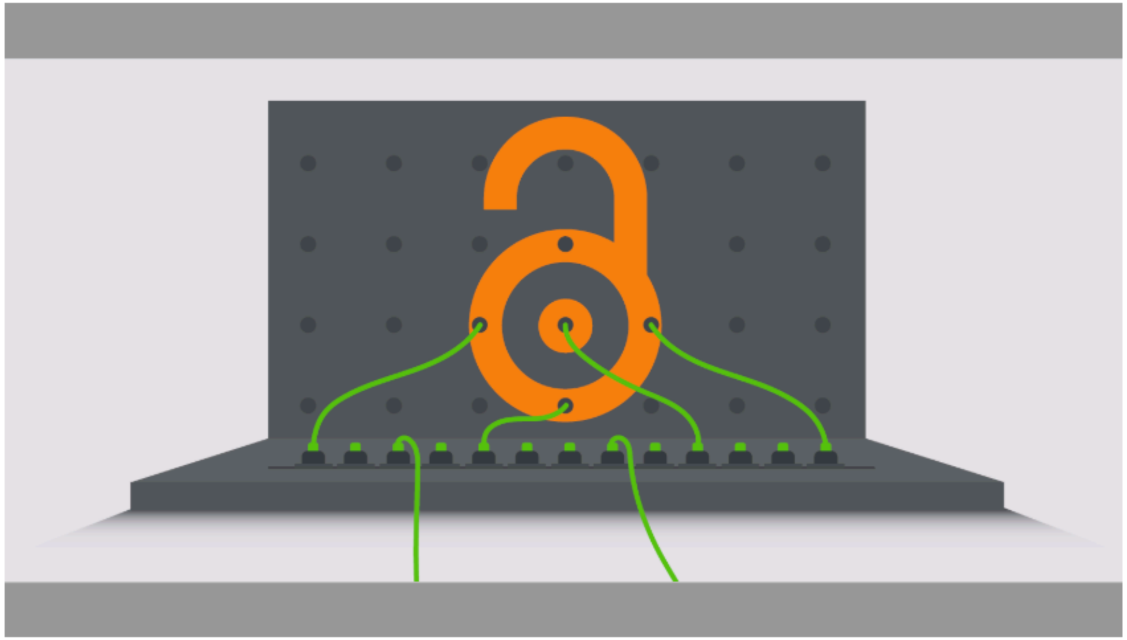
http://doi.org/10.5860/crl.78.2.183 http://hdl.handle.net/1805/14063

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.

Interesting new development in this regard: The OA Switchboard

OASPA Open Access Scholarly Publishers Association



https://oaspa.org/oa-switchboard/

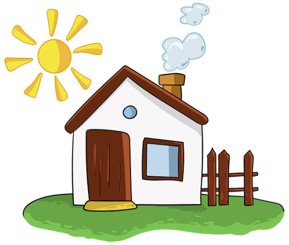
"...the OA Switchboard is designed to enable publishers, academic institutions, and research funders to seamlessly communicate information about open access publications, without trying to serve as an intermediary for any payments...”

Metadata-driven approach

Targeted to authors, but is designed to facilitate workflow management and reporting at institutions

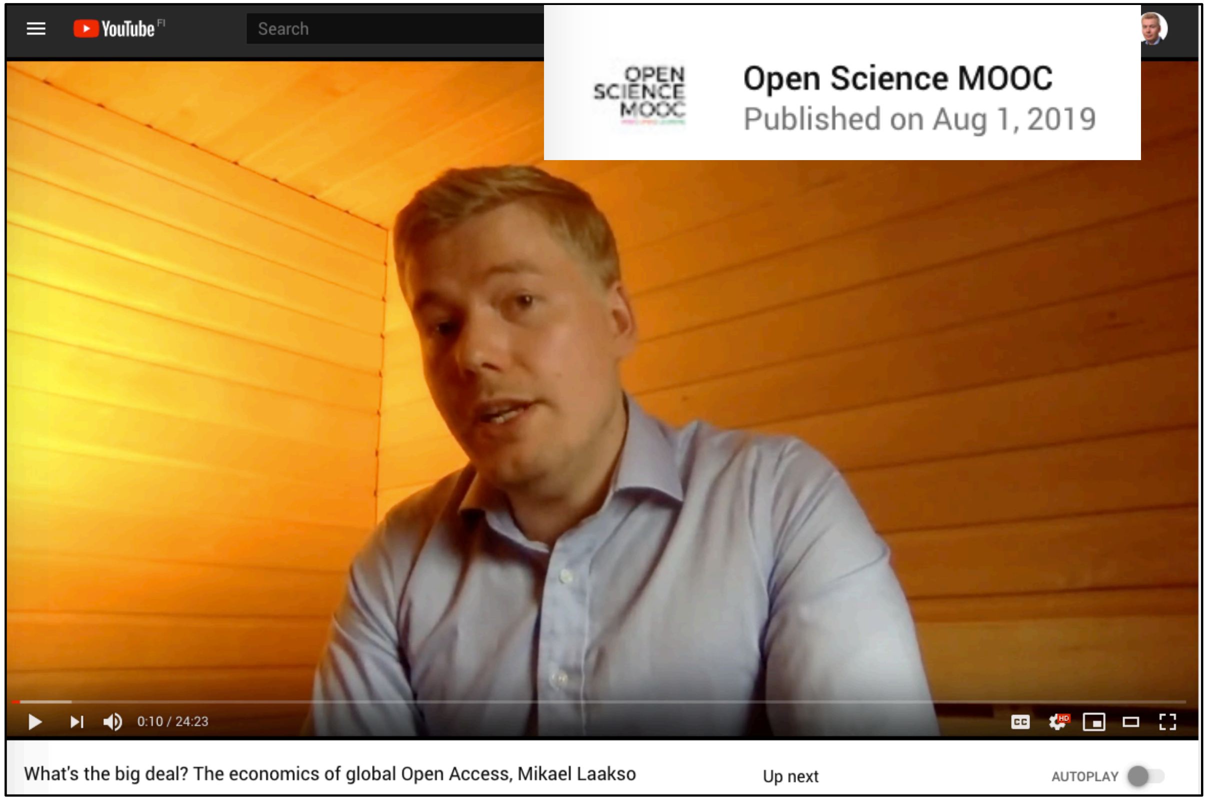
Key takeaways

- 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!



Still want more?

HANKEN



Open Science MOOC
Published on Aug 1, 2019

What's the big deal? The economics of global Open Access, Mikael Laakso

https://youtu.be/3rmbeWGgrWE

ELEPHANT

IN THE LAB

Laakso, M. (2019). **Why we need a public infrastructure for data on open access**. *Elephant in the Lab*. https://doi.org/10.5281/zenodo.2540472why-we-need-a-public-infrastructure-for-data-on-open-access/

Danke!