Doing bibliometrics responsibly
Openness of bibliometric metadata

Ludo Waltman
Centre for Science and Technology Studies, Leiden University

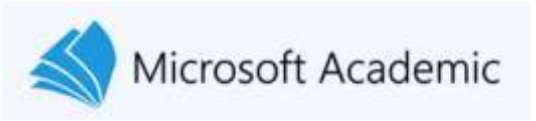
ISKO UK Research Observatory
April 27, 2022

Outline

- Proprietary bibliometric data sources
- Toward openness of bibliometric metadata
- Comparison of bibliometric data sources

CWTS

Scientometric data sources



Some key questions

How do we know what's happening in the research system?

Who gets to decide what counts and what doesn't?

How can we make the research system more open and transparent, more equitable, and ultimately more democratic?

CWTS

Proprietary bibliometric data sources

Proprietary bibliometric data sources

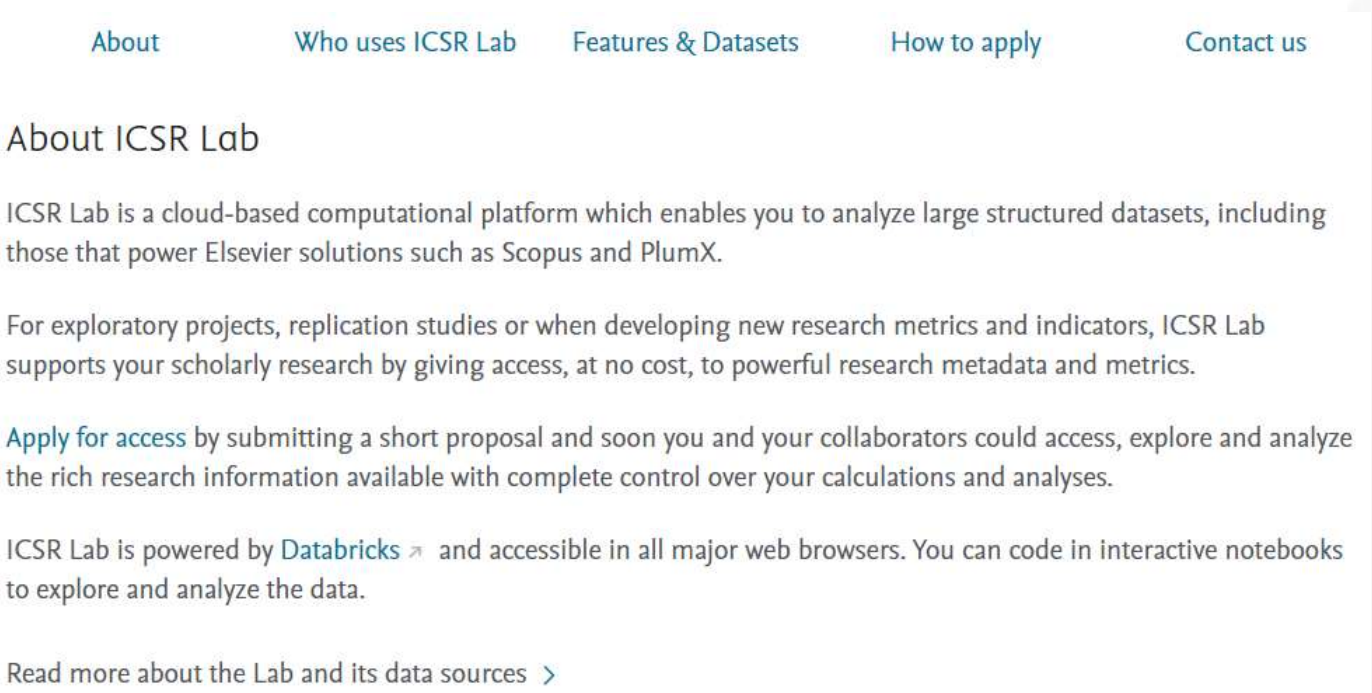
- Web of Science:
  - Launched in 1964 by the Institute of Scientific Information (ISI) as the Science Citation Index (SCI)
  - Nowadays owned by Clarivate Analytics
- Scopus:
  - Launched in 2004 by Elsevier
- Dimensions:
  - Launched in 2018 by Digital Science
  - Limited version freely accessible
  - Large-scale data access (BigQuery)

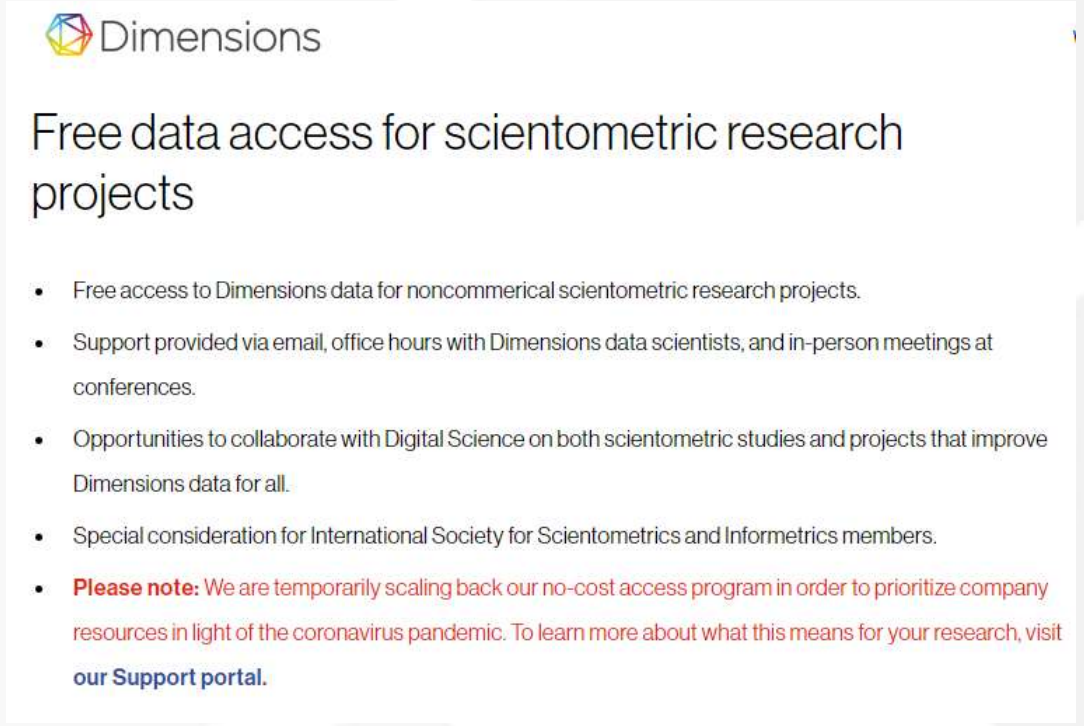
CWTS

Content selection policies

- Web of Science:
  - Focus on selectivity
  - Content selection by internal Editorial Development team
- Scopus:
  - Balance between selectivity and comprehensiveness; claims to deliver “the most comprehensive overview of the world’s research output”
  - Content selection by Content Selection and Advisory Board
- Dimensions:
  - Focus on comprehensiveness: “The database should not be selective but rather should be open to encompassing all scholarly content that is available for inclusion … The community should then be able to choose the filter that they wish to apply to explore the data according to their use case.”

Free access to Scopus and Dimensions data for research purposes





CWTS

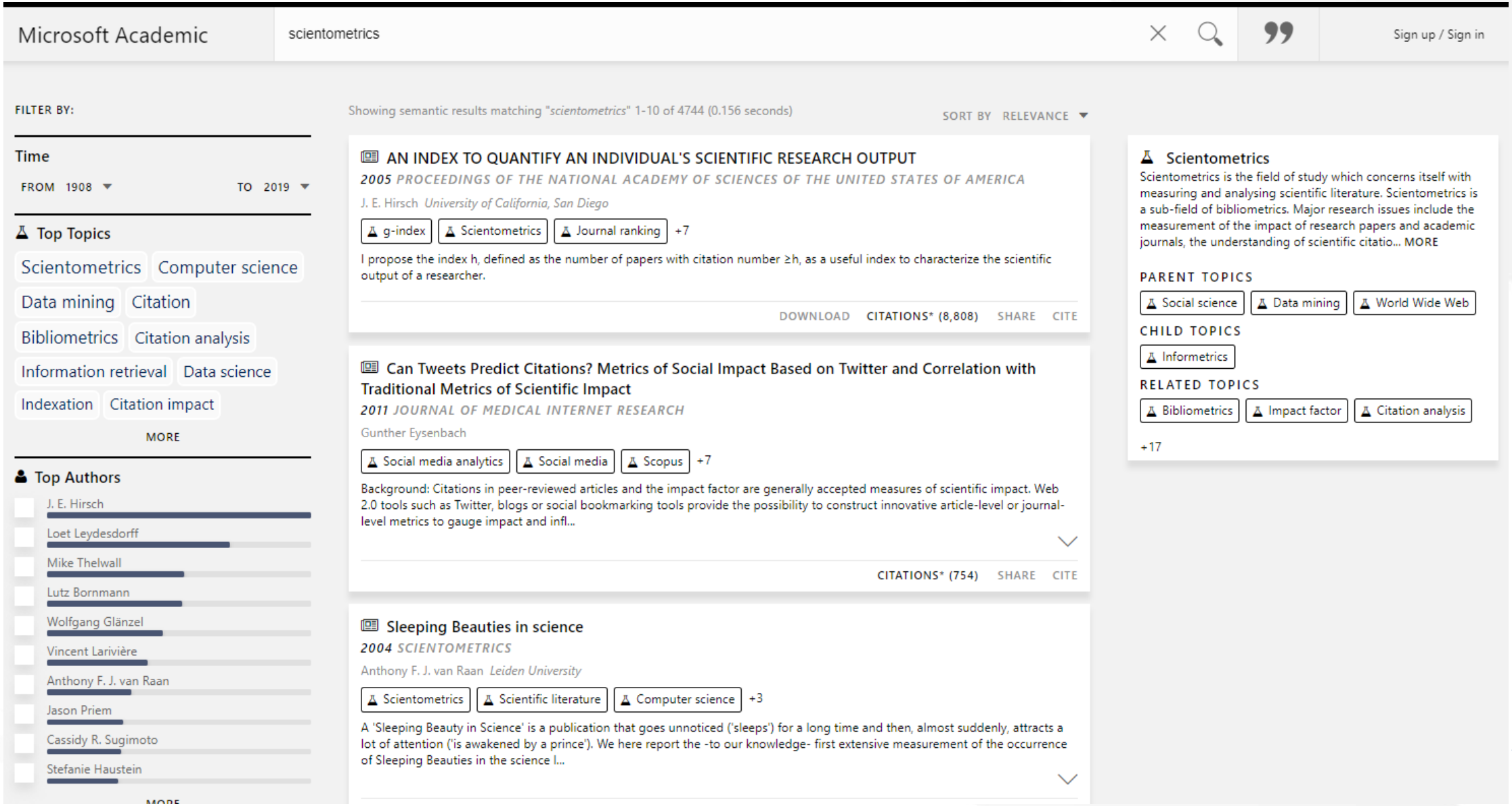
Toward openness of bibliometric metadata

Open bibliometric data sources

- PubMed:
  - Launched in 1996 by NLM/NIH
  - Not a citation index
- Crossref:
  - Large-scale data access
  - Major gaps due to publishers depositing incomplete data
- Microsoft Academic:
  - Launched in 2016 by Microsoft; discontinued in 2021
- OpenAlex:
  - Launched in 2022 by OurResearch as a successor of Microsoft Academic

CWTS

Microsoft Academic



Comprehensiveness of Microsoft Academic

Thursday 26 November 2015

WEDDING IN VALENCIA: ALEXANDRA + PABLO

POSTED IN WEDDING DAY

Alexandra and Paulo celebrated their wedding this summer in the Gardens of Monforte, the party was held at the **Golf Club Scorpion** Valencia. As a wedding photographer I love to work with people from different countries. Alexandra is from Russia and Paulo was born in Valencia. Fate brought them together by chance and they found out to be made one for each other.

The wedding of Alexandra and Paul was an **intimate ceremony with family and friends**. The civil ceremony was held in the Gardens of Monforte, an elegant and cozy place in Valencia. Grooms uttered the "Yes I do" in an atmosphere of happiness and joy. After the ceremony we made some group photos and portraits on the grounds. Then the couple and the guests moved to the place of the banquet: the Golf Club "Scorpion" a few kilometers from the city.

CWTS

OpenAlex

OurResearch blog

News from the OurResearch team

MAG replacement update: meet OpenAlex!

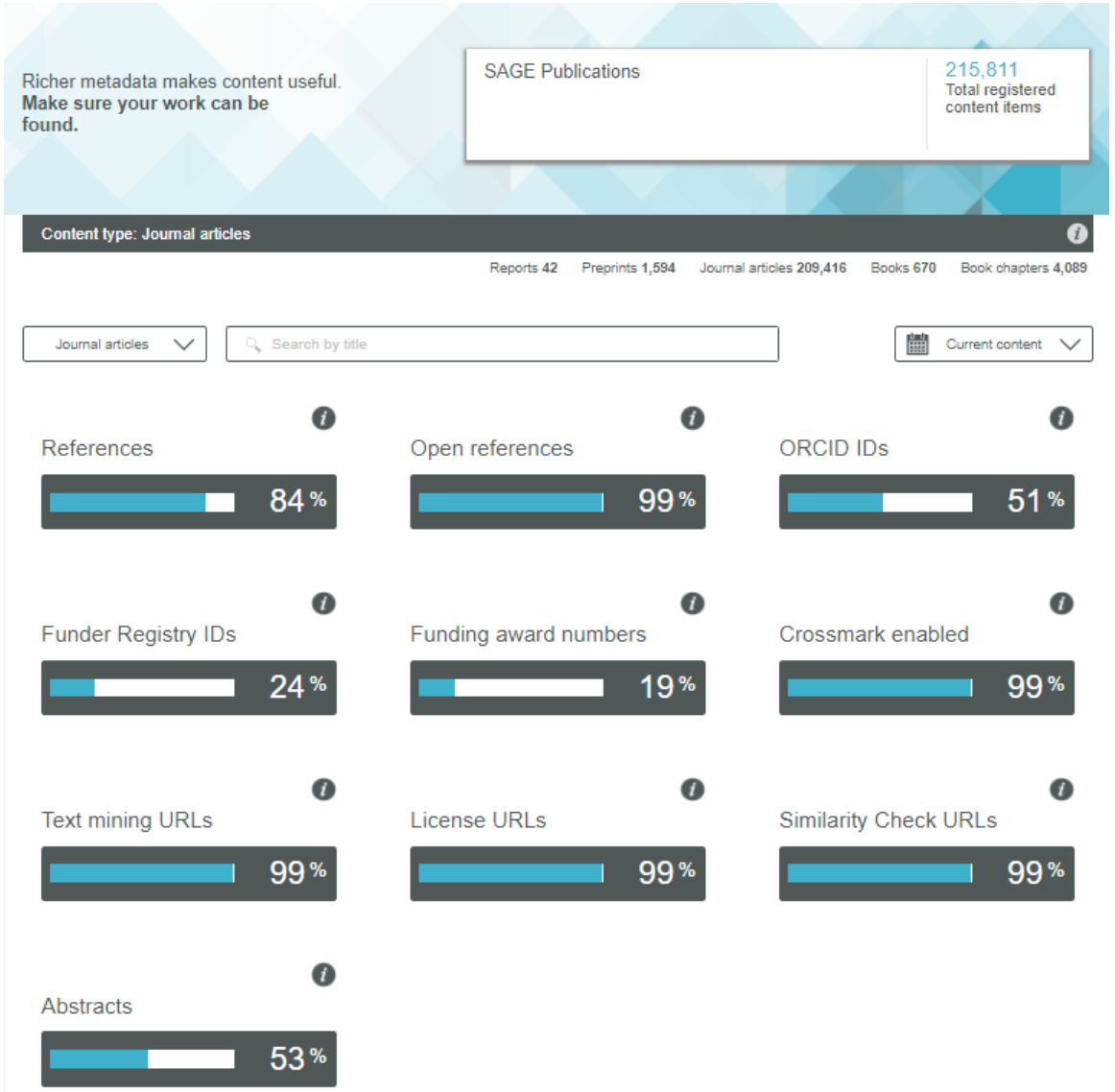
Last month, we announced that we're launching a replacement for Microsoft Academic Graph (MAG) this December–just before MAG itself will be discontinued. We've heard from a *lot* of current MAG users since then. All of them have offered their support and encouragement (which we really appreciate), and all have also all been curious to learn more. So: here's more! It's a snapshot of what we know right now. As the project progresses, we'll have more details to share, keeping everyone as up-to-date as we can.

Name

We've now got a name for this project: **OpenAlex**. We like that it (a) emphasizes Open, and (b) is inspired by the ancient Library of Alexandria — like that fabled institution, OpenAlex will strive to create a comprehensive map of the global scholarly conversation. We'll start with MAG data, and we'll expand over time. Along with the name, we've got the beginnings of a webpage at openalex.org, and a Twitter account at @OpenAlex_org.

CWTS

Crossref


Richer metadata makes content useful.
Make sure your work can be found.
SAGE Publications
215,811
Total registered content items
Content type: Journal articles
Reports 42
Preprints 1,594
Journal articles 209,416
Books 670
Book chapters 4,089
Journal articles
Search by title
Current content
References
84 %
Open references
99 %
ORCID IDs
51 %
Funder Registry IDs
24 %
Funding award numbers
19 %
Crossmark enabled
99 %
Text mining URLs
99 %
License URLs
99 %
Similarity Check URLs
99 %
Abstracts
53 %


Richer metadata makes content useful.
Make sure your work can be found.
American Psychological Association (APA)
27,945
Total registered content items
Content type: Journal articles
Journal articles 19,823
Datasets 6,468
Books 133
Book chapters 1,521
Journal articles
Search by title
Current content
References
1 %
Open references
0 %
ORCID IDs
66 %
Funder Registry IDs
36 %
Funding award numbers
27 %
Crossmark enabled
1 %
Text mining URLs
1 %
License URLs
18 %
Similarity Check URLs
75 %
Abstracts
0 %

Initiative for Open Citations (I4OC)

I4OC

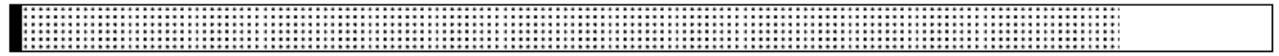
I4OC About Goals Publishers Stakeholders Founders FAQ News Press

An initiative to open up citation data

The aim of this initiative is to promote the availability of data on citations that are **structured**, **separable**, and **open**.

Structured means the data representing each publication and each citation instance are expressed in common, machine-readable formats, and that these data can be accessed programmatically. **Separable** means the citation instances can be accessed and analyzed without the need to access the source bibliographic products (such as journal articles and books) in which the citations are created. **Open** means the data are freely accessible and reusable.

How many citations are open today?



1% 87% 12%

As of October 2021, the fraction of publications with open references has grown from 1% to 88% out of 56.9 million articles with references deposited with Crossref.

We encourage all other scholarly publishers to follow the example of these trail-blazing publishers by making their reference metadata publicly available. Please contact Crossref Support (support@crossref.org) for more information, or to let them know that you are ready to open up your reference metadata now. See also our list of responses to frequently asked questions.

CWTS

Initiative for Open Abstracts (I4OA)

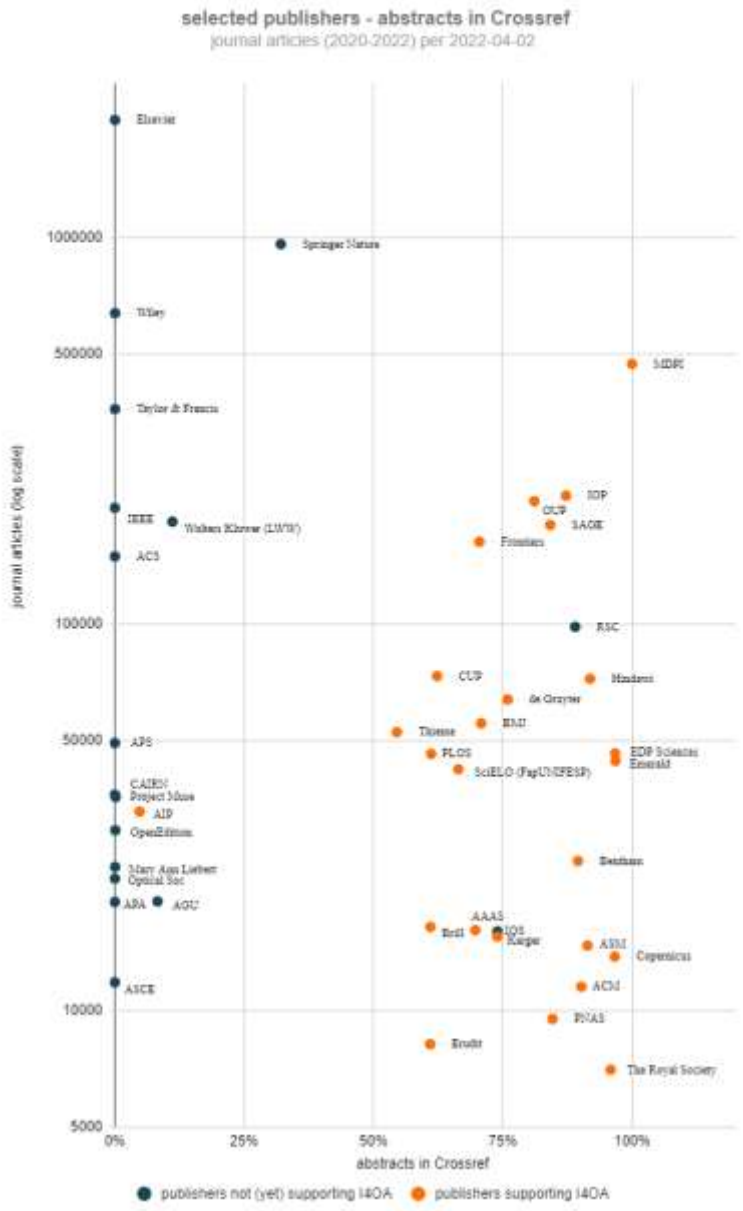
ABS TRA CTS

Initiative for Open Abstracts

The Initiative for Open Abstracts (I4OA) is a collaboration between scholarly publishers, infrastructure organizations, librarians, researchers and other interested parties to advocate and promote the unrestricted availability of the abstracts of the world's scholarly publications, particularly journal articles and book chapters, in trusted repositories where they are open and machine-accessible. I4OA calls on all scholarly publishers to open the abstracts of their published works, and where possible to submit them to Crossref.

CWTS

Initiative for Open Abstracts (I4OA)



Blog

Home > Blog > **Open Abstracts: Where are we?**

6 minute read.

Open Abstracts: Where are we?

Ludo Waltman, Bianca Kramer, Ginny Hendricks, Bryan Vickery – 2020 September 25
In Metadata, Content Registration, Collaboration, Community

The Initiative for Open Abstracts (I4OA) launched this week. The initiative calls on scholarly publishers to make the abstracts of their publications openly available. More specifically, publishers that work with Crossref to register DOIs for their publications are requested to include abstracts in the metadata they deposit in Crossref. These abstracts will then be made openly available by Crossref. 39 publishers have already agreed to join I4OA and to open their abstracts.

Where are we at the moment in terms of openness of abstracts? For an individual publisher working with Crossref, the percentage of the publisher's content for which an abstract is available in Crossref can be found in Crossref's Participation Reports. The chart presented below gives the overall picture (as of September 1, 2020) for medium-sized and large publishers working with Crossref. The vertical axis shows the number of journal articles of a publisher in the period 2018-2020. Because of the large differences between publishers in the number of articles they publish, this axis has a logarithmic scale. The horizontal axis shows the percentage of the articles of a publisher for which an abstract is available in Crossref. The orange dots represent publishers that have agreed to join I4OA. The publishers colored in blue have not yet agreed to join the initiative.

FAIRness of publications

Comparison of bibliometric data sources

Comparison of bibliographic data sources

April 08 2021

Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic

Martijn Visser, Nees Jan van Eck, Ludo Waltman

Quantitative Science Studies (2021) 2 (1): 20–41.

https://doi.org/10.1162/qss_a_00112 **Article history**

Abstract

We present a large-scale comparison of five multidisciplinary bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. The comparison considers scientific documents from the period 2008–2017 covered by these data sources. Scopus is compared in a pairwise manner with each of the other data sources. We first analyze differences between the data sources in the coverage of documents, focusing for instance on differences over time, differences per document type, and differences per discipline. We then study differences in the completeness and accuracy of citation links. Based on our analysis, we discuss the strengths and weaknesses of the different data sources. We emphasize the importance of combining a comprehensive coverage of the scientific literature with a flexible set of filters for making selections of the literature.

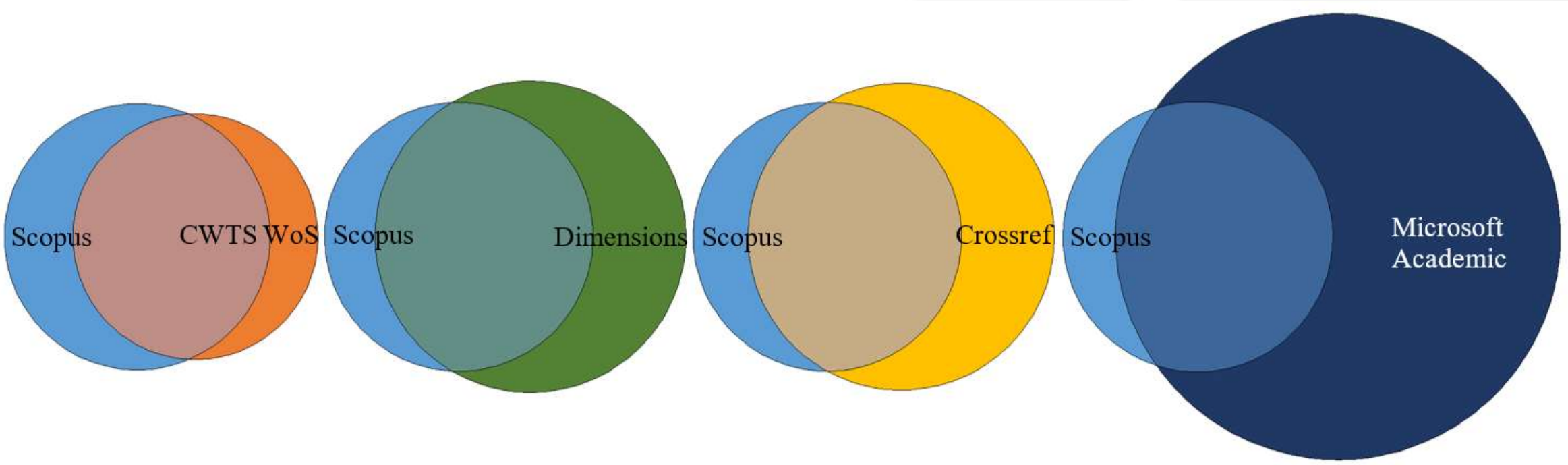
Keywords: bibliographic data source, Crossref, Dimensions, Microsoft Academic, Scopus, Web of Science

CWTS

Comparison of bibliographic data sources

- Bibliographic data sources:
  - Web of Science (SCIE, SSCI, AHCI, and CPCI)
  - Scopus
  - Dimensions
  - Microsoft Academic
  - Crossref
- For practical reasons, Scopus is used as baseline
- Document-level matching of data sources
- Period of analysis: 2008–2017

Comparison of coverage of documents



Scopus: 27.0M
CWTS WoS: 22.9M
Overlap: 17.7M

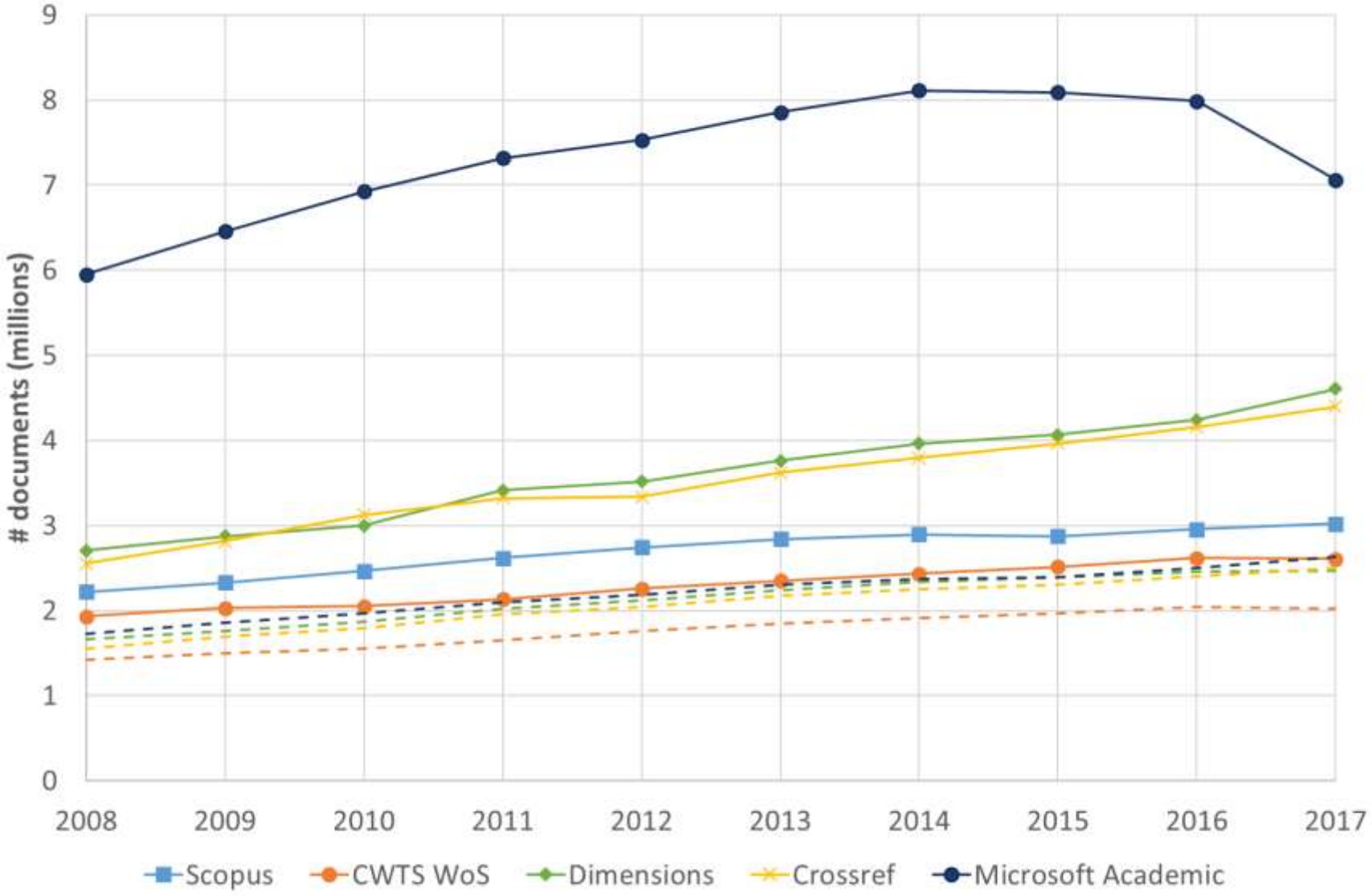
Scopus: 27.0M
Dimensions: 36.1M
Overlap: 21.3M

Scopus: 27.0M
Crossref: 35.1M
Overlap: 20.7M

Scopus: 27.0M
Microsoft Academic: 73.3M
Overlap: 22.0M

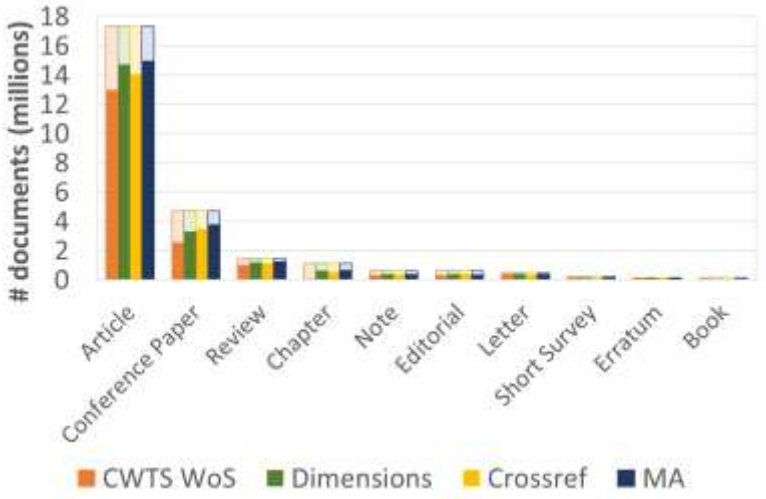
CWTS

Differences in coverage by publication year

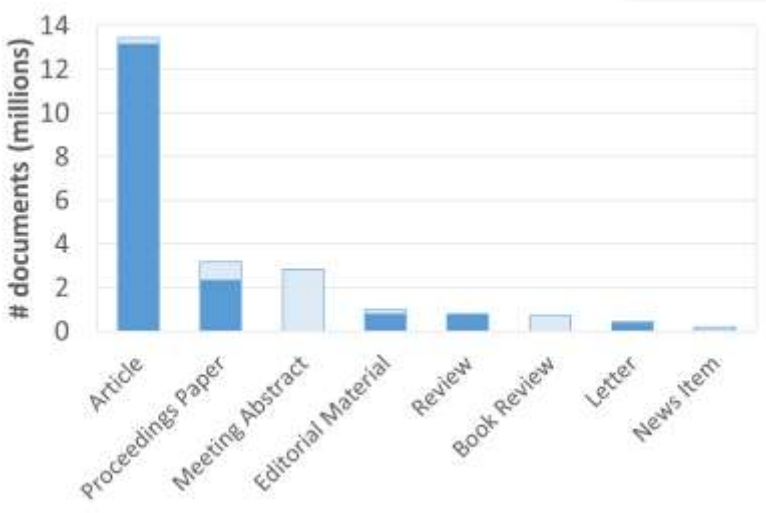


Differences in coverage by document type

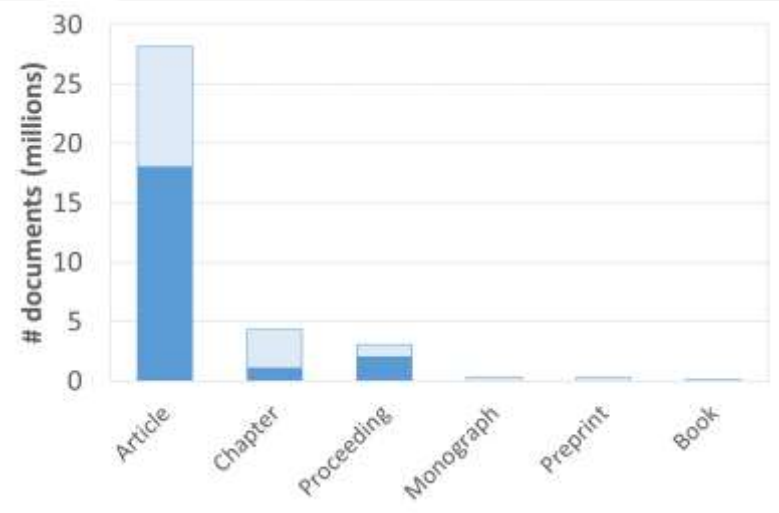
Other data sources from Scopus perspective



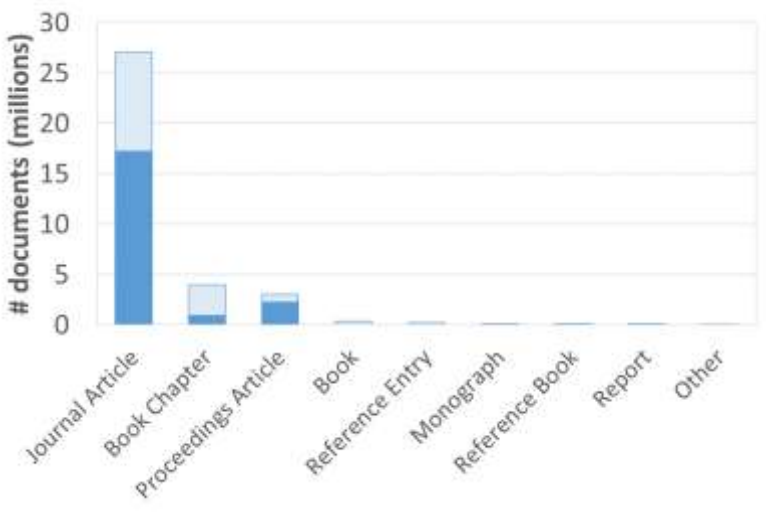
Scopus from Web of Science perspective



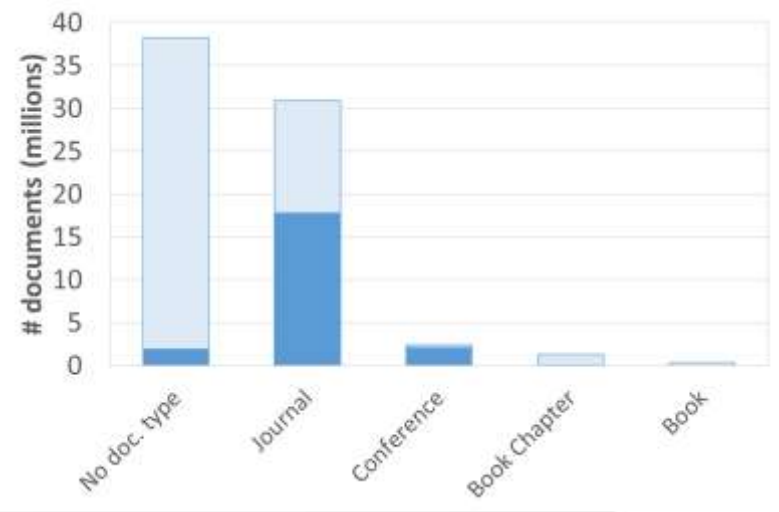
Scopus from Dimensions perspective



Scopus from Crossref perspective

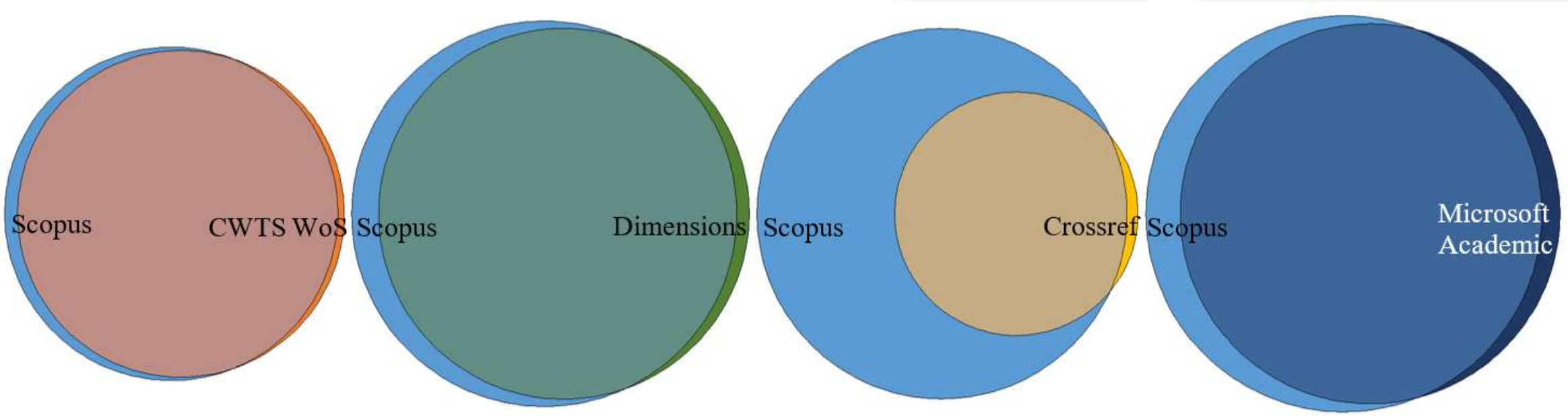


Scopus from Microsoft Academic perspective



CWTS

Comparison of completeness and accuracy of citation links



Scopus: 176.5M
CWTS WoS: 169.9M
Overlap: 166.4M

Scopus: 205.0M
Dimensions: 189.7M
Overlap: 183.3M

Scopus: 199.9M
Crossref: 86.3M
Overlap: 84.1M

Scopus: 209.7M
Microsoft Academic: 193.0M
Overlap: 183.1M

CWTS

Conclusions

Some key questions

How do we know what's happening in the research system?

Who gets to decide what counts and what doesn't?

How can we make the research system more open and transparent, more equitable, and ultimately more democratic?

Why does openness of bibliometric metadata matter?

- Bibliometric metadata plays a crucial role in finding and accessing scientific literature
  - Openness will improve discoverability

- Bibliometric metadata plays a crucial role in the way we value scientific work and assess researchers and research teams
  - Openness ensures that the scientific community will be able to decide itself how it wants scientific work to be valued and researchers to be assessed

- Openness creates an equal playing field in which everyone can decide themselves what should be counted and what shouldn't
- Openness enables research metrics to be made fully transparent and their merits and shortcomings to be discussed in a democratic way

Thank you for your attention!