



Alternative data sources for bibliometric analyses pros and cons, benefits and caveats

Martijn Visser

Centre for Science and Technology Studies, Leiden University

Wednesday 9 December 2020, 12:30-14:00

RISIS Research Seminar



**Universiteit
Leiden**

Criteria for evaluating bibliographic data sources

- Coverage of scientific literature
- Completeness and accuracy of metadata
- Data provider enhancements
- Accessibility (User interface, licensing, costs)

Multidisciplinary bibliographic databases suitable for citation analysis

- 1964: Web of Science
- 2004: Scopus
- 2004: Google Scholar
- 2016: Microsoft Academic
- 2018: Dimensions
- Crossref

Recent studies comparing bibliographic data sources

- Coverage of the publication output of 15 universities
- WoS Core Collection, Scopus, Microsoft Academic

Huang, C.-K et al. (2020) Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies*. https://doi.org/10.1162/qss_a_00031.

- Coverage of literature citing over 2,500 highly cited documents
- WoS Core Collection, Scopus, Dimensions, COCI, Google Scholar and Microsoft Academic

Martín-Martín, A. *et al.* (2020). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics* <https://doi.org/10.1007/s11192-020-03690-4>

New elements in our study

- Comprehensive analysis involving all scientific documents
- Comparison at the document level
- Sophisticated matching procedure

Comparison of publication coverage

- Bibliographic data sources:

- Web of Science (SCIE, SSCI, AHCI, and CPCI)
- Scopus
- Dimensions
- Crossref
- Microsoft Academic

Jan 2019

May 2019

June 2019

August 2018

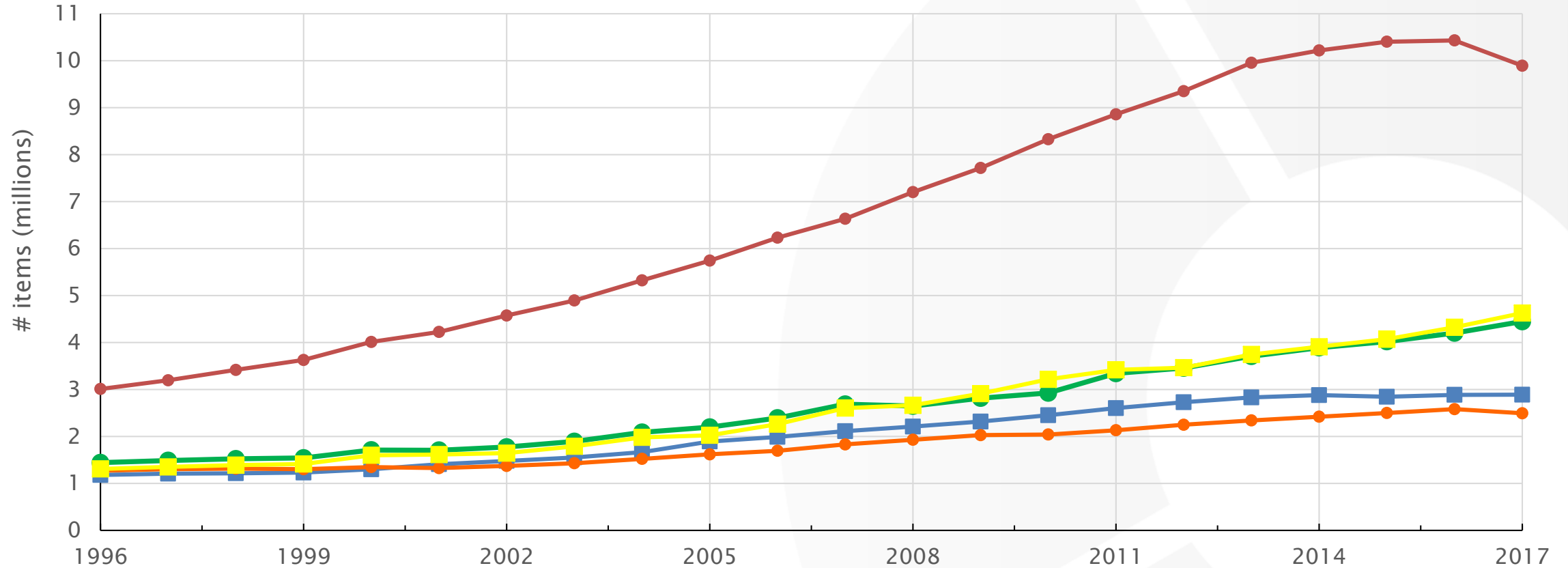
March 2019

- Comparison based on document-level matching between data sources
- For practical reasons, Scopus is used as a reference

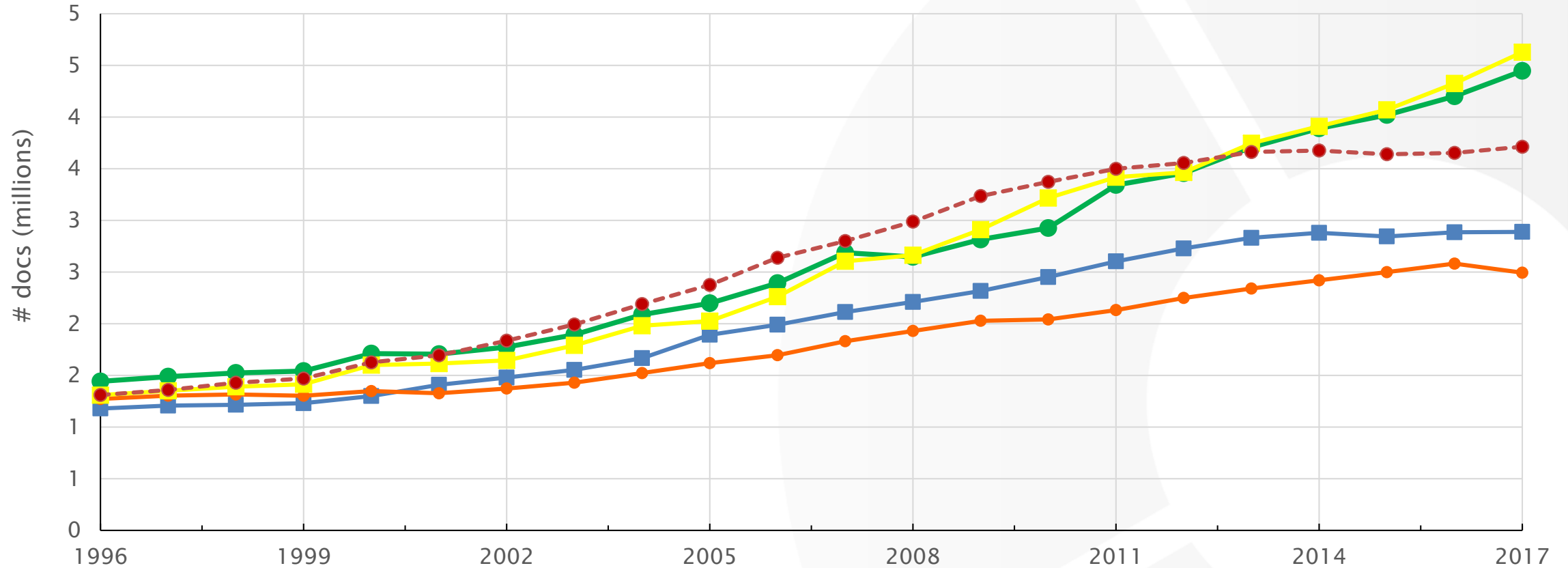
Web of Science: Different citation indices

- **Web of Science Core Collection:**
 - Science Citation Index Expanded
 - Social Sciences Citation Index
 - Arts & Humanities Citation Index
 - Emerging Sources Citation Index
 - Book Citation Index
 - Conference Proceedings Citation Index
- **Regional Collection:**
 - Chinese Science Citation Database
 - Russian Science Citation Index
 - KCI Korean Journal Database
 - SciELO Citation Index
- **Specialist Collection**
- **Data Citation Index**
- **Derwent Innovations Index**

Number of items indexed 1996 - 2017



Number of **documents** indexed 1996 - 2017



Match on a paper by paper basis

1. Preprocessing data of bibliographic elements
2. Retrieving pairs of possible matches based on different search criteria
3. Calculating similarity for each pair based on many different fields (doi, title, first author, volume, issue, first page etc.)
4. Determining optimal scores and thresholds

Comparison of publication coverage

- Bibliographic data sources:
 - Web of Science (SCIE, SSCI, AHCI, and CPCI)
 - Scopus
 - Dimensions
 - Crossref
 - Microsoft Academic
- Comparison based on document-level matching between data sources
- For practical reasons, Scopus is used as a reference
- Feedback from Scopus and Dimensions

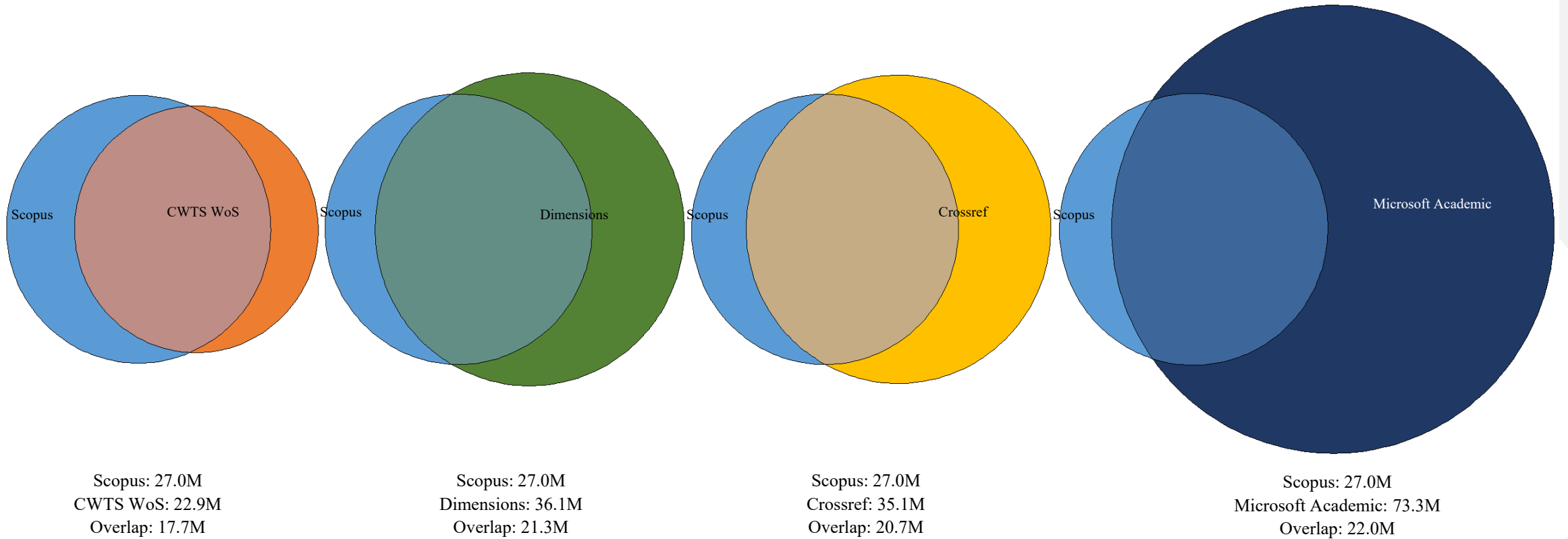
Content selection policies (1)

- WoS:
 - Focus on selectivity
 - Content selection by internal Editorial Development team
- Scopus:
 - Focus on comprehensiveness; Scopus claims to be “the largest abstract and citation database of peer-reviewed literature”
 - Content selection by Content Selection and Advisory Board
- Dimensions:
 - “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.” (Hook et al., 2018)

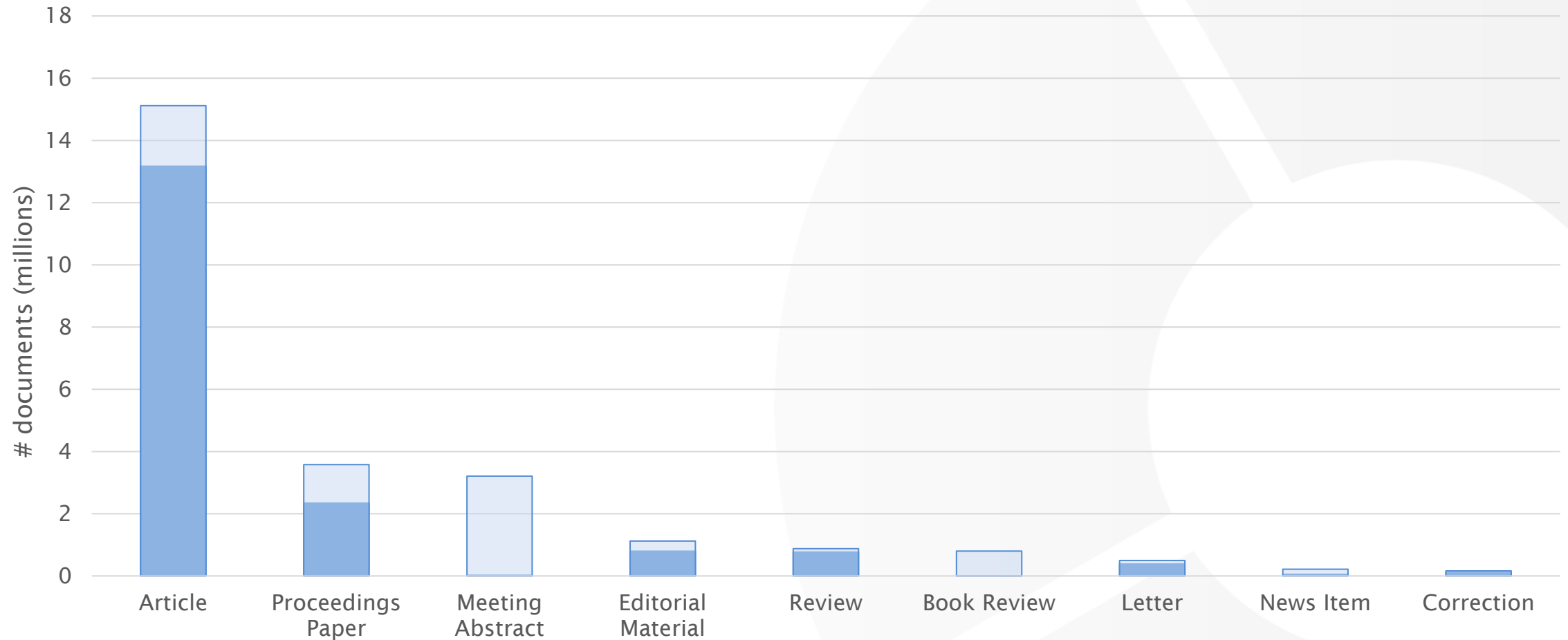
Content selection policies (2)

- Crossref:
 - Content selection is not done by Crossref, but by members registering for DOIs and depositing metadata
- Microsoft Academic :
 - Collects content from the web through Bing and publisher feeds
 - Unclear which filters they apply to identify academic content

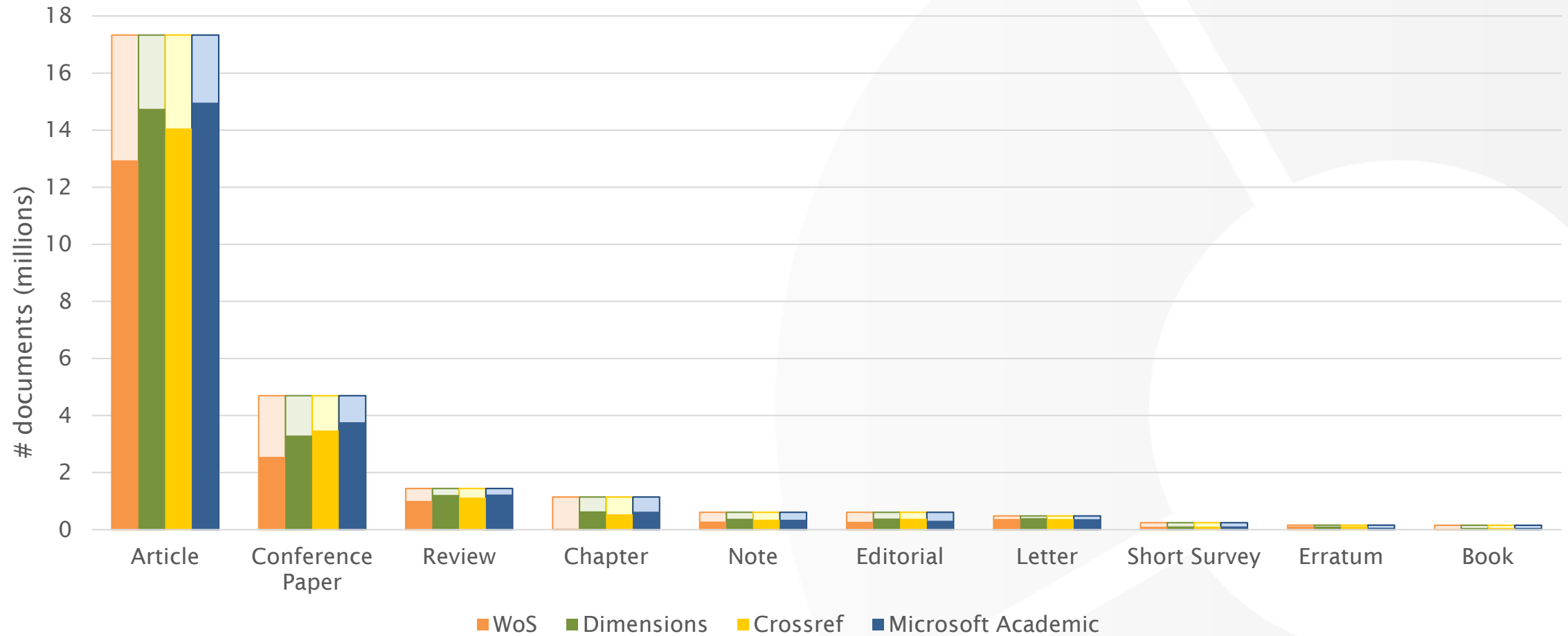
Comparison of coverage of documents 2008 - 2017



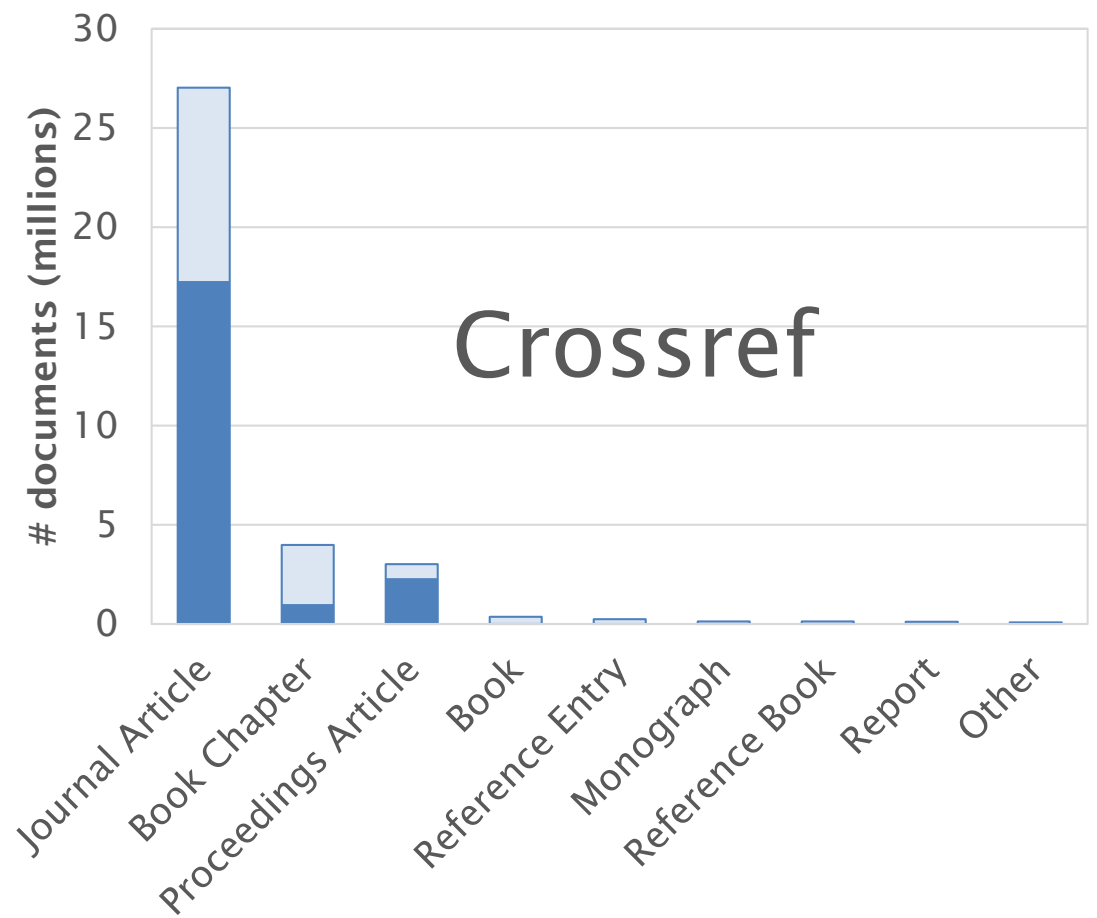
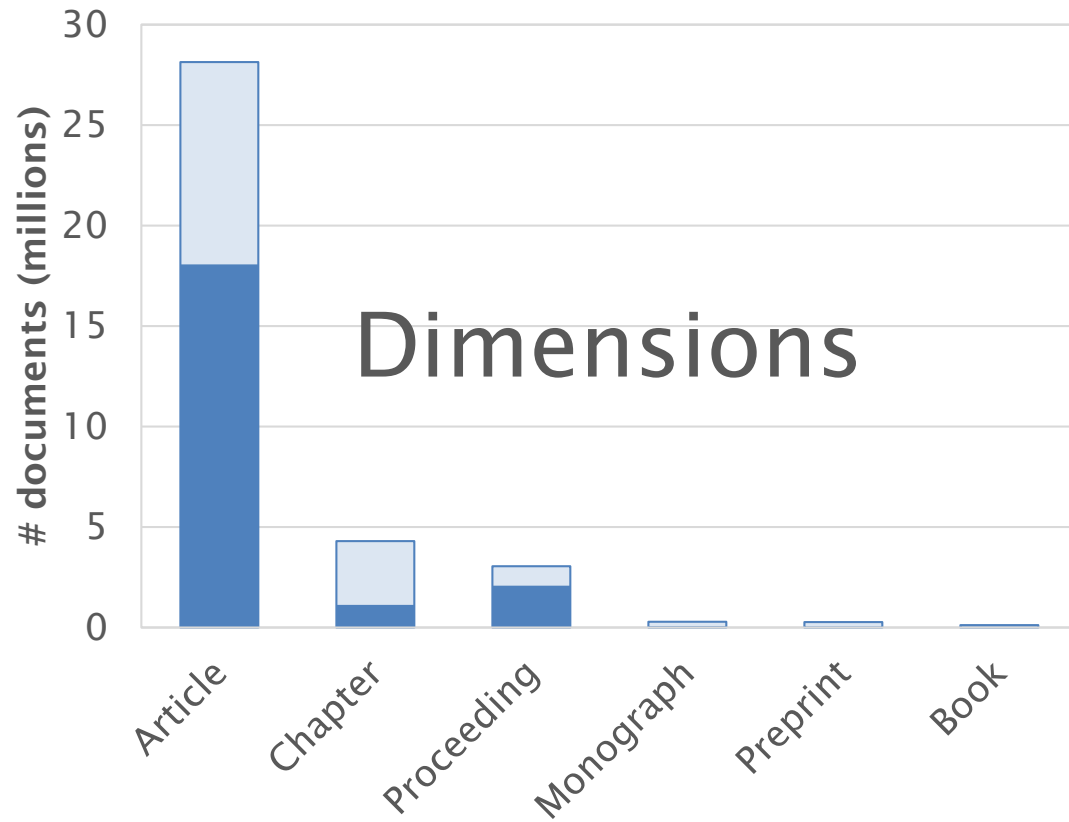
Coverage of CWTS WoS document types 2008-2017



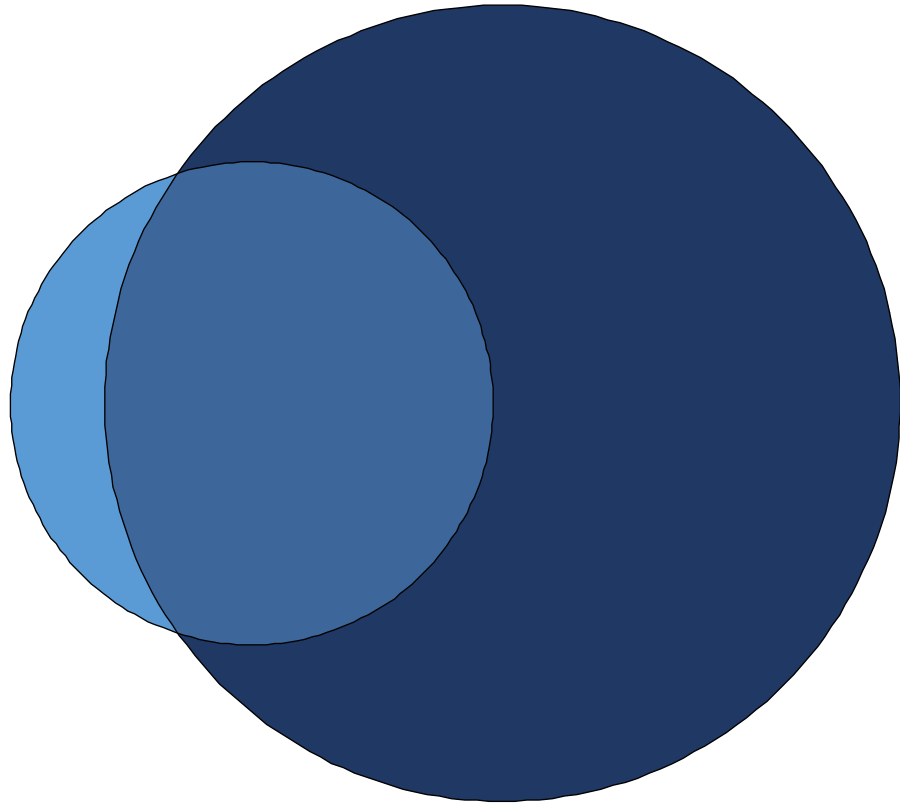
Coverage of Scopus document types 2008-2017



Dimensions and Crossref document types

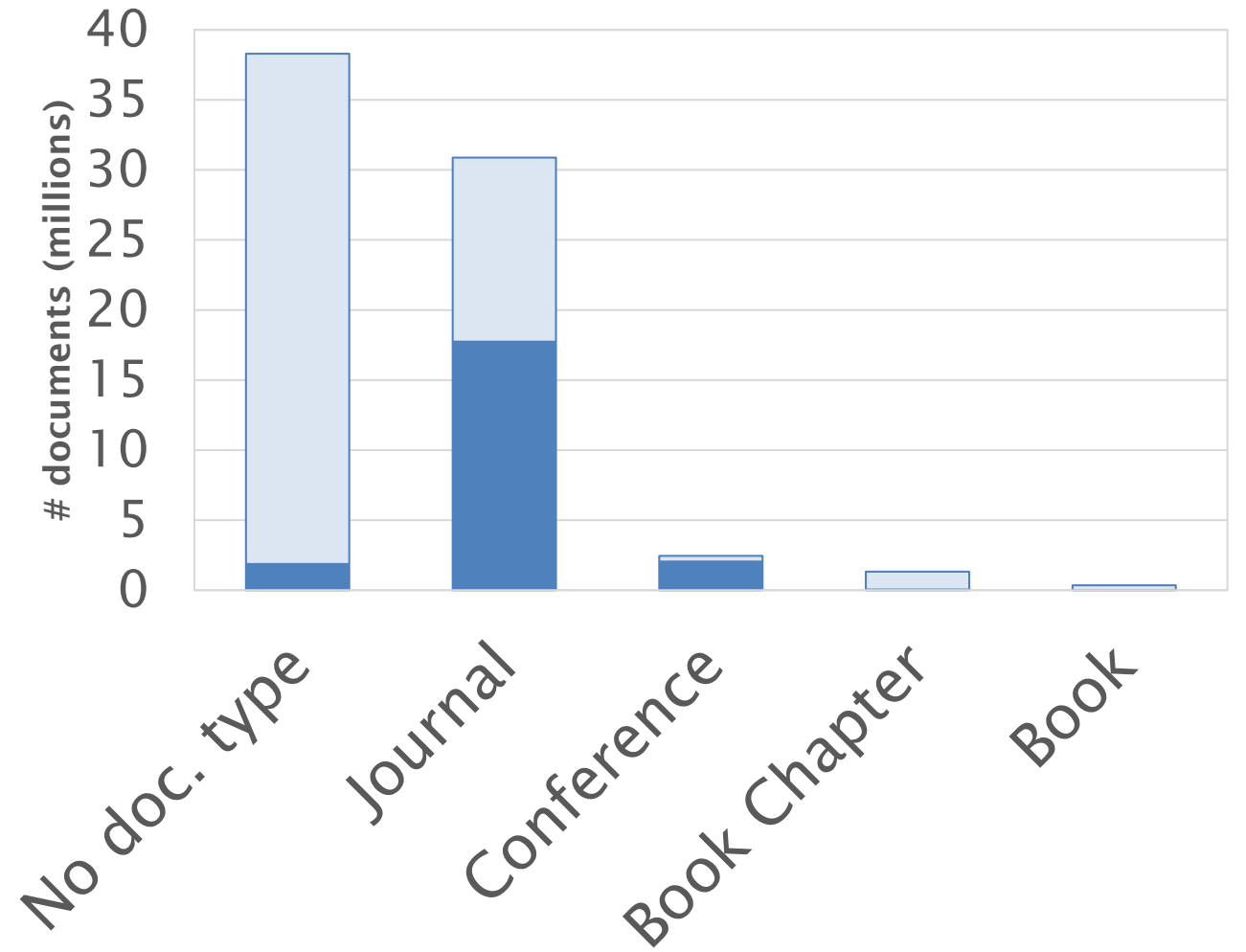


Microsoft Academic document types 2008 - 2017



Scopus 9.7M


Microsoft Academic
40,9M



Friday 11 July 2014


DESTINATION WEDDING IN ROME: TANIA + FELIX


POSTED IN DESTINATION WEDDINGS

 Destination Wedding in Rome: Tania + Felix

2014

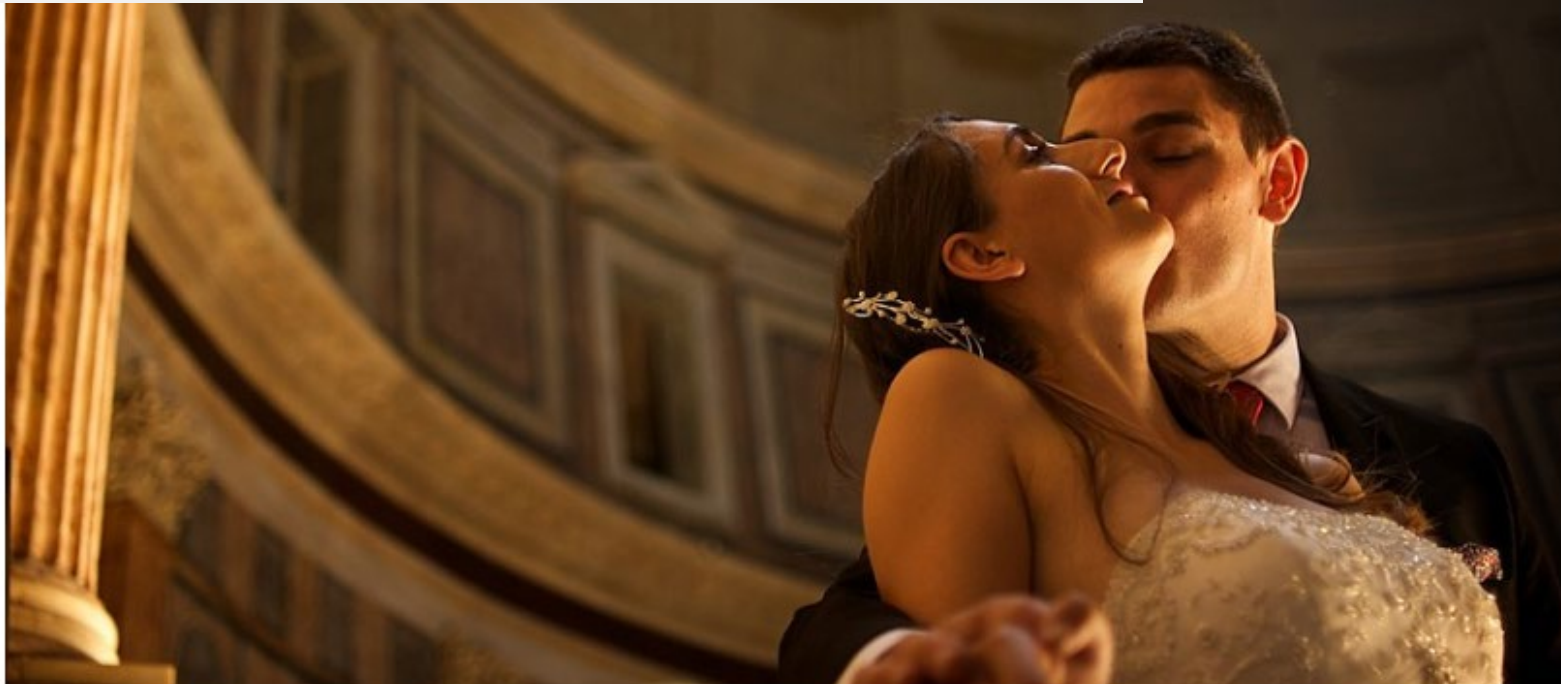
Martino Buzzi

 Cartography

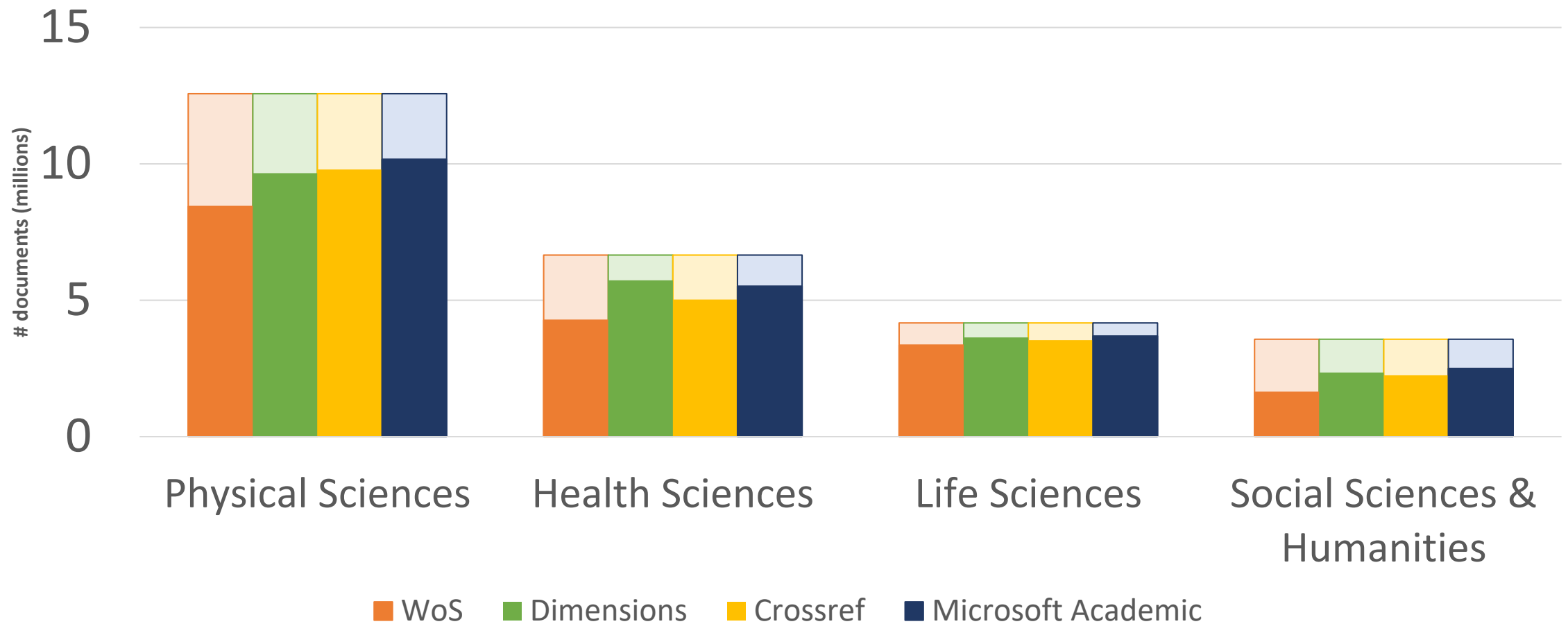
 Art history

 Art

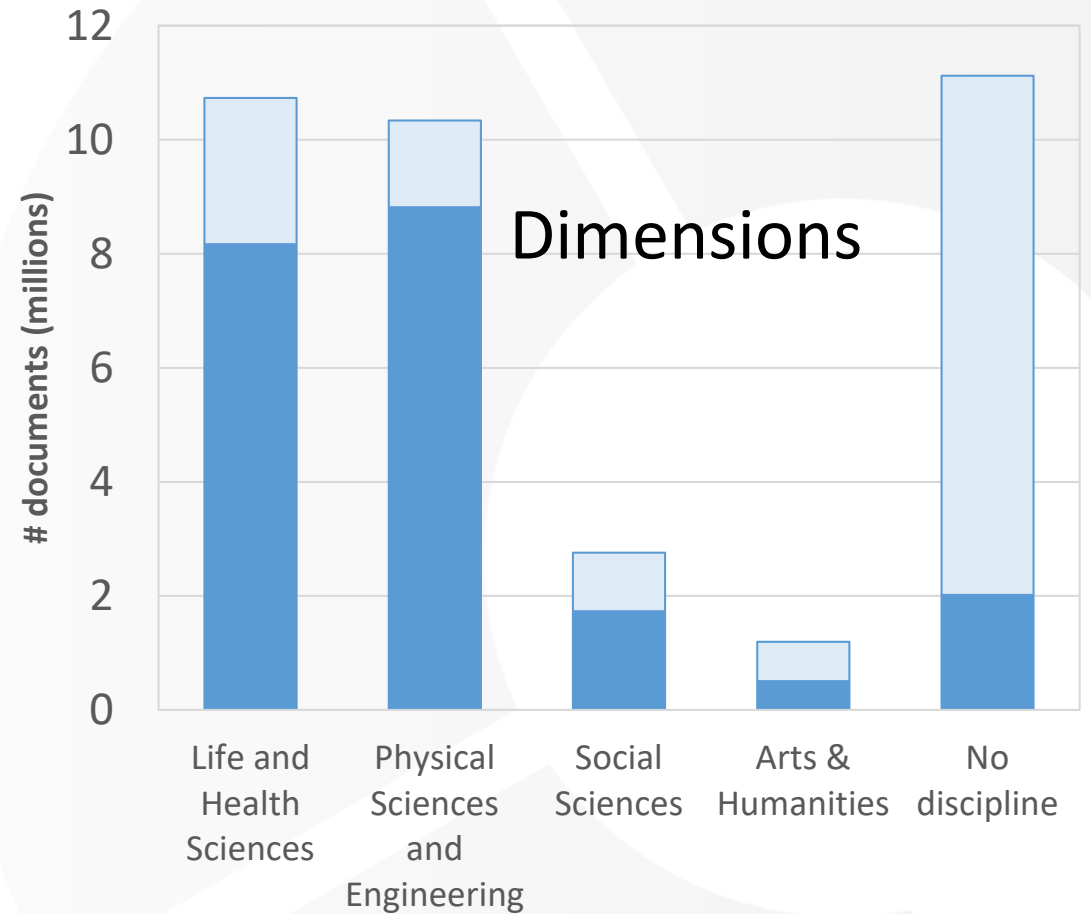
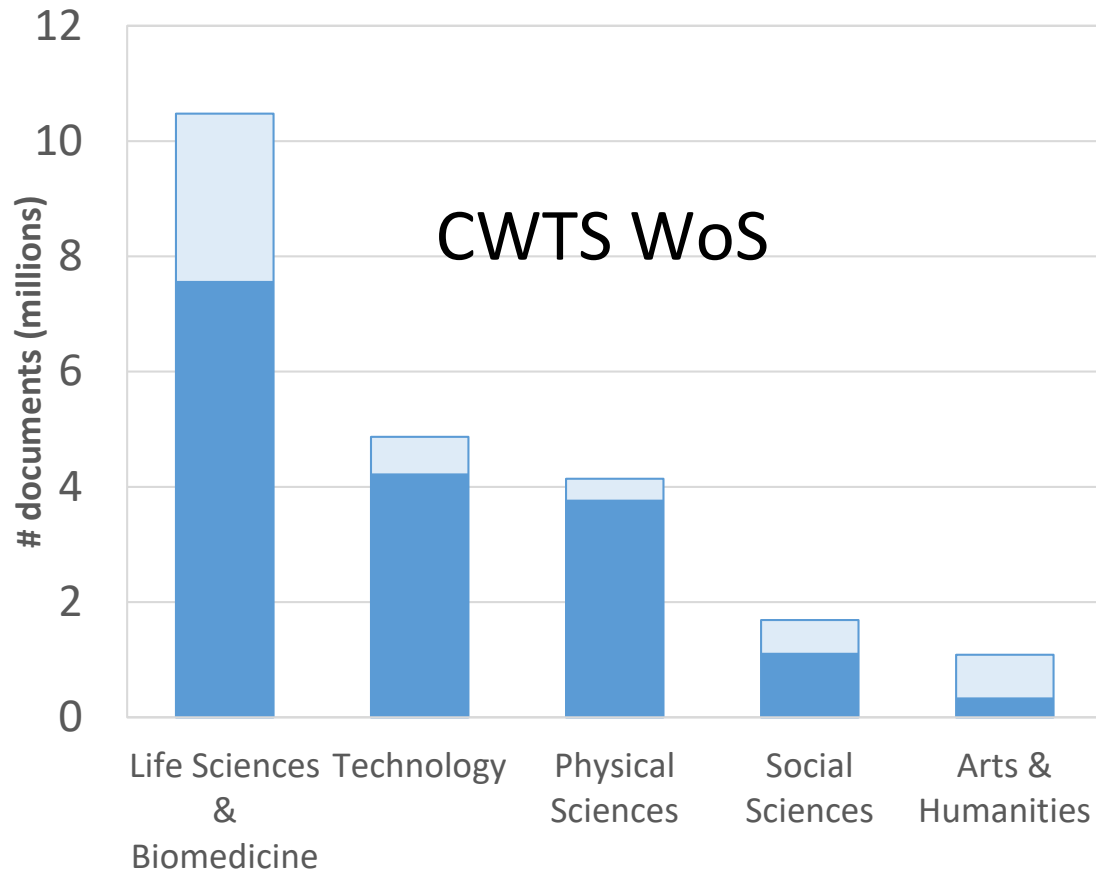
CITATIONS* (0)   



Differences in coverage by discipline

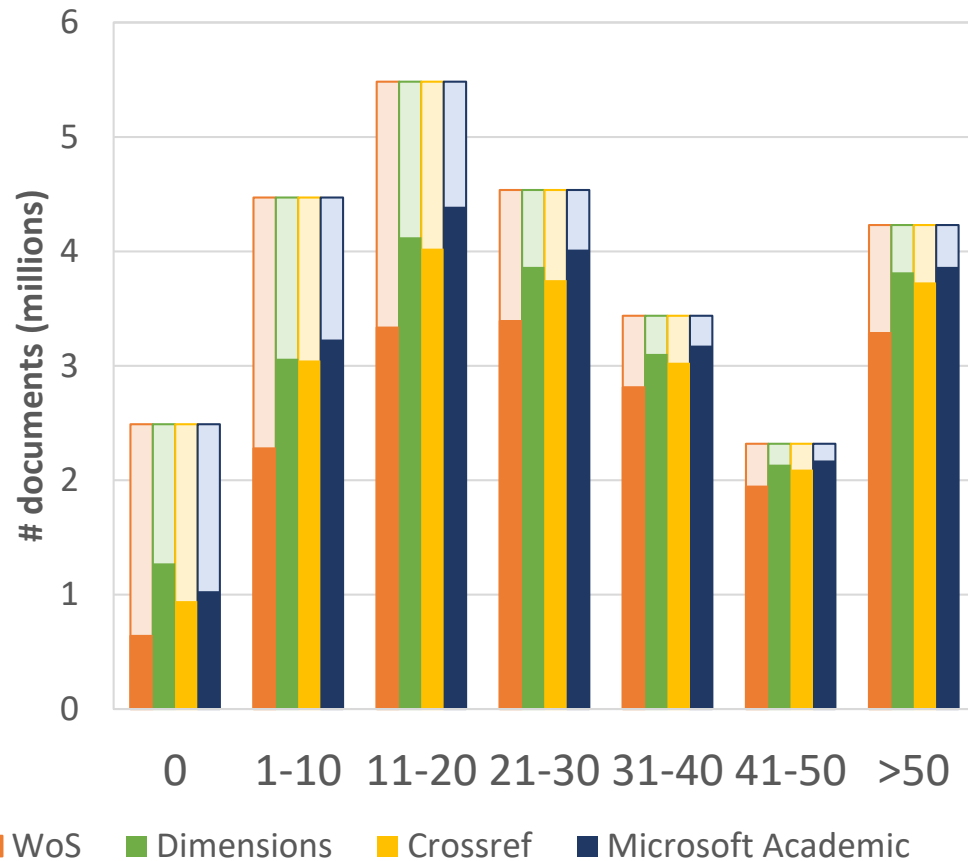


Difference in coverage of fields

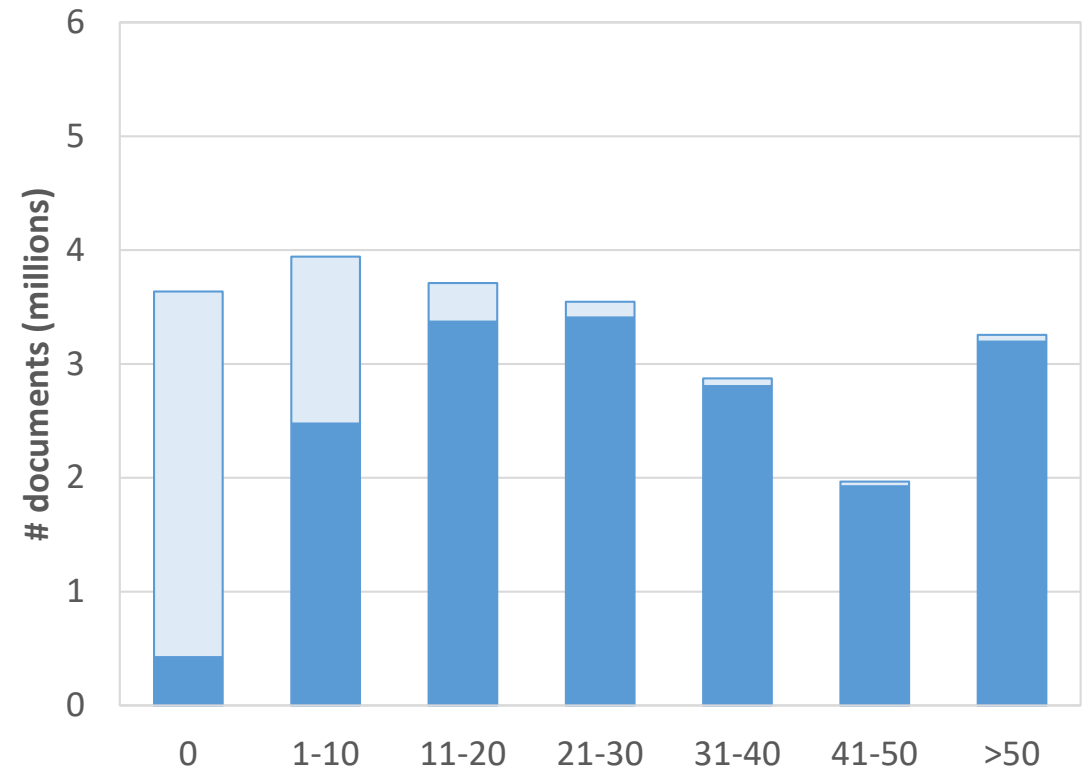


Differences in coverage by number of references

Other data sources from Scopus perspective

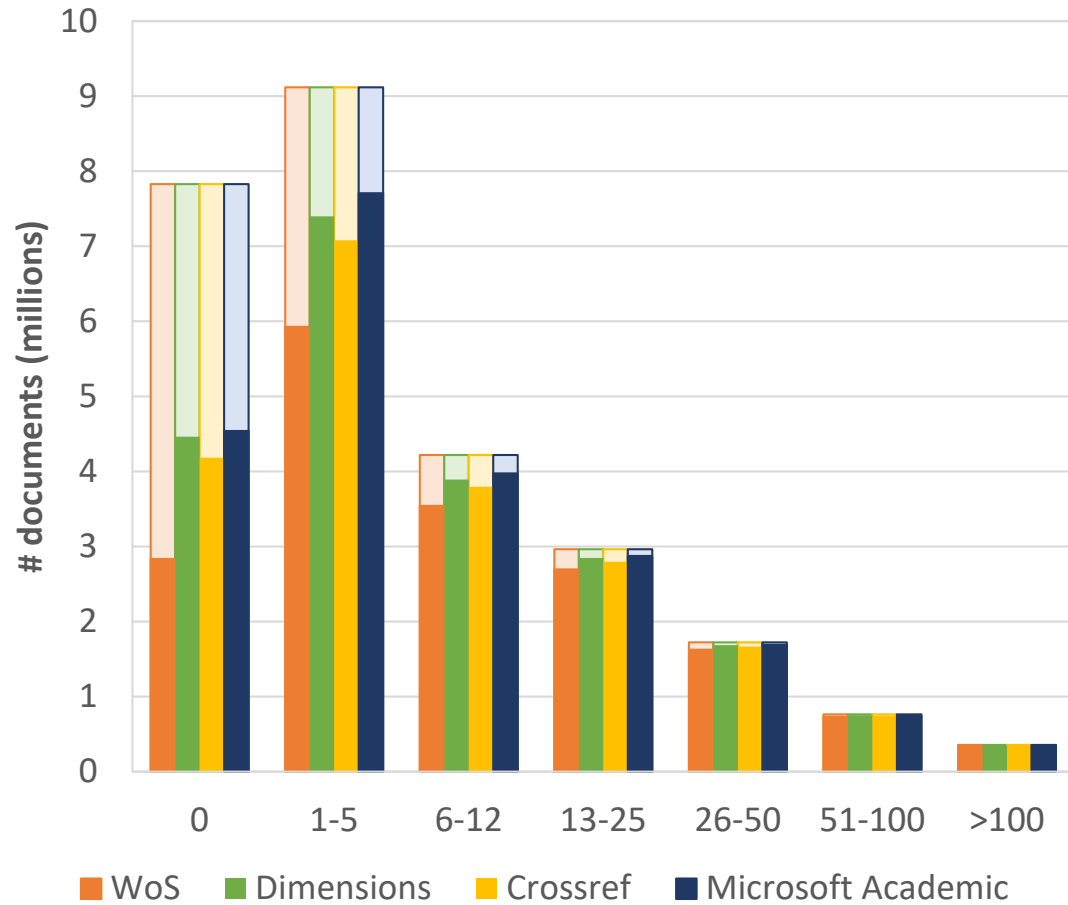


Scopus from CWTS WoS perspective

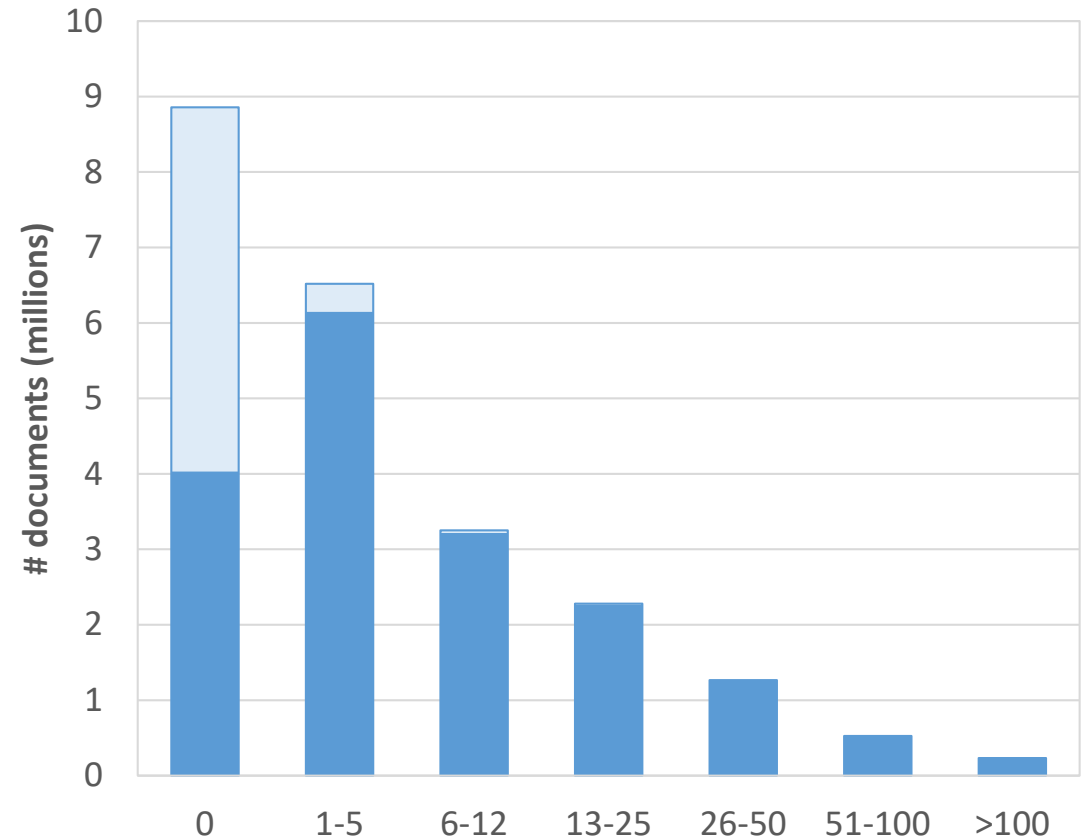


Differences in coverage by number of citations received (1)

Other data sources from Scopus perspective

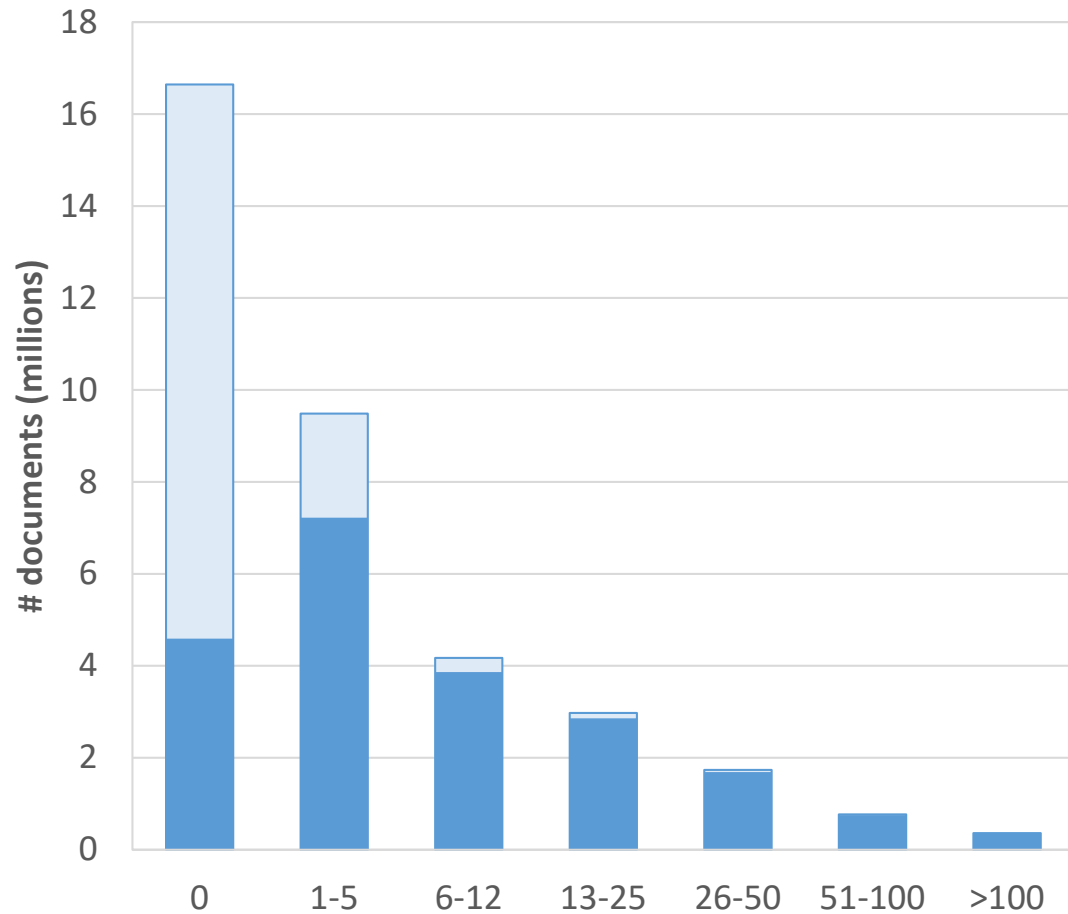


Scopus from CWTS WoS perspective

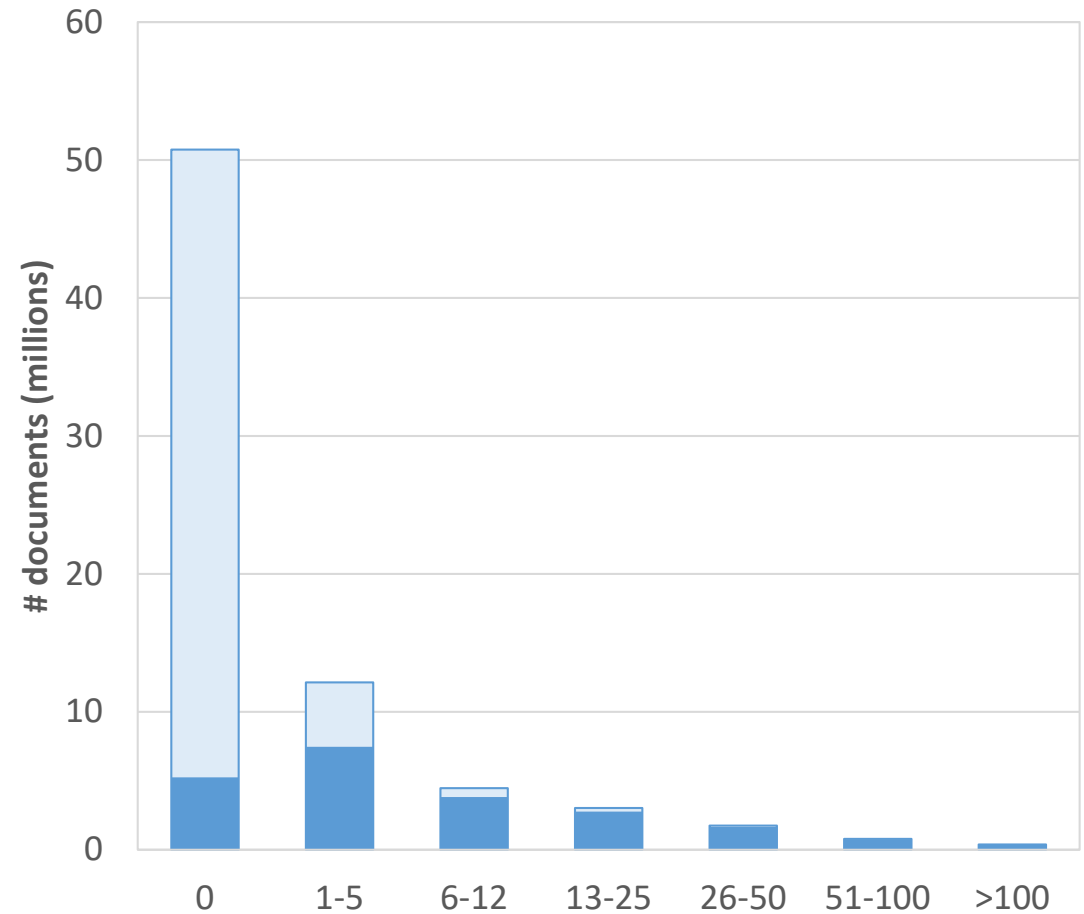


Differences in coverage by number of citations received (2)

Scopus from Dimensions perspective



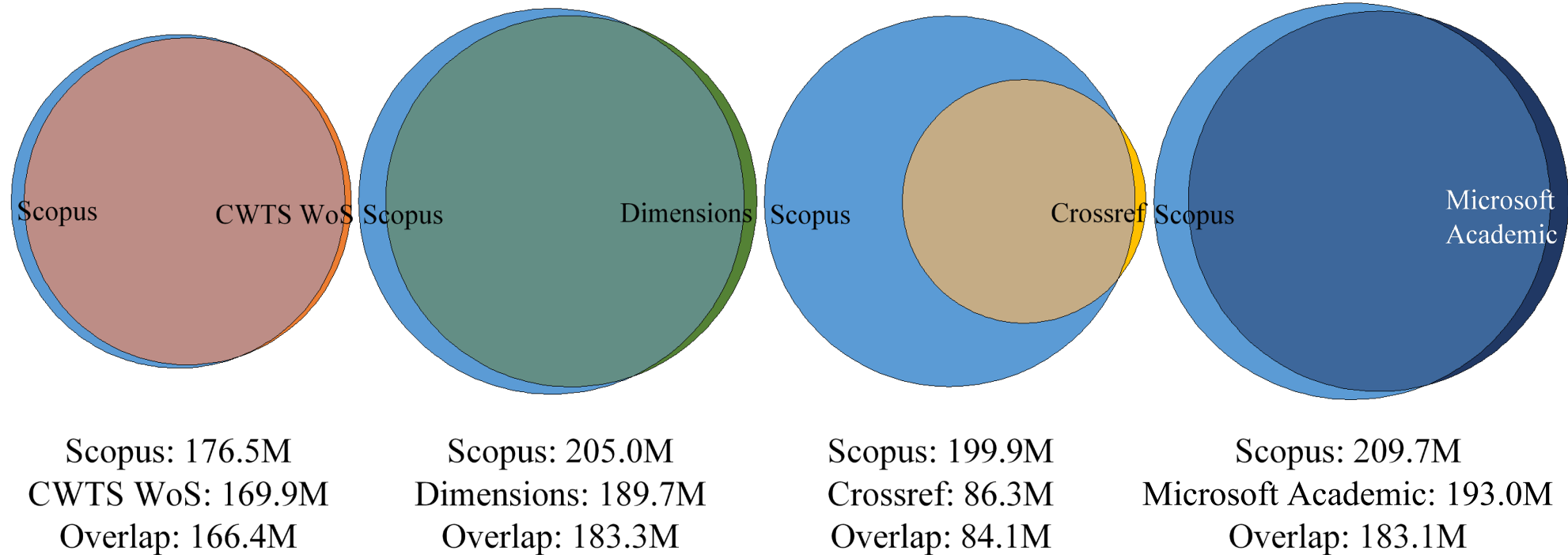
Scopus from Microsoft Academic perspective



Comparison of citation links in bibliographic databases



Comparison of the presence of citation links



Comparison of completeness and accuracy of citation links

- Web of Science has problems with missing and incorrect references
- Scopus has problems with references that incorrectly have not been matched
- Dimensions has problems in distinguishing between different versions of a cited document
- Dimensions and Crossref have problems with missing reference lists

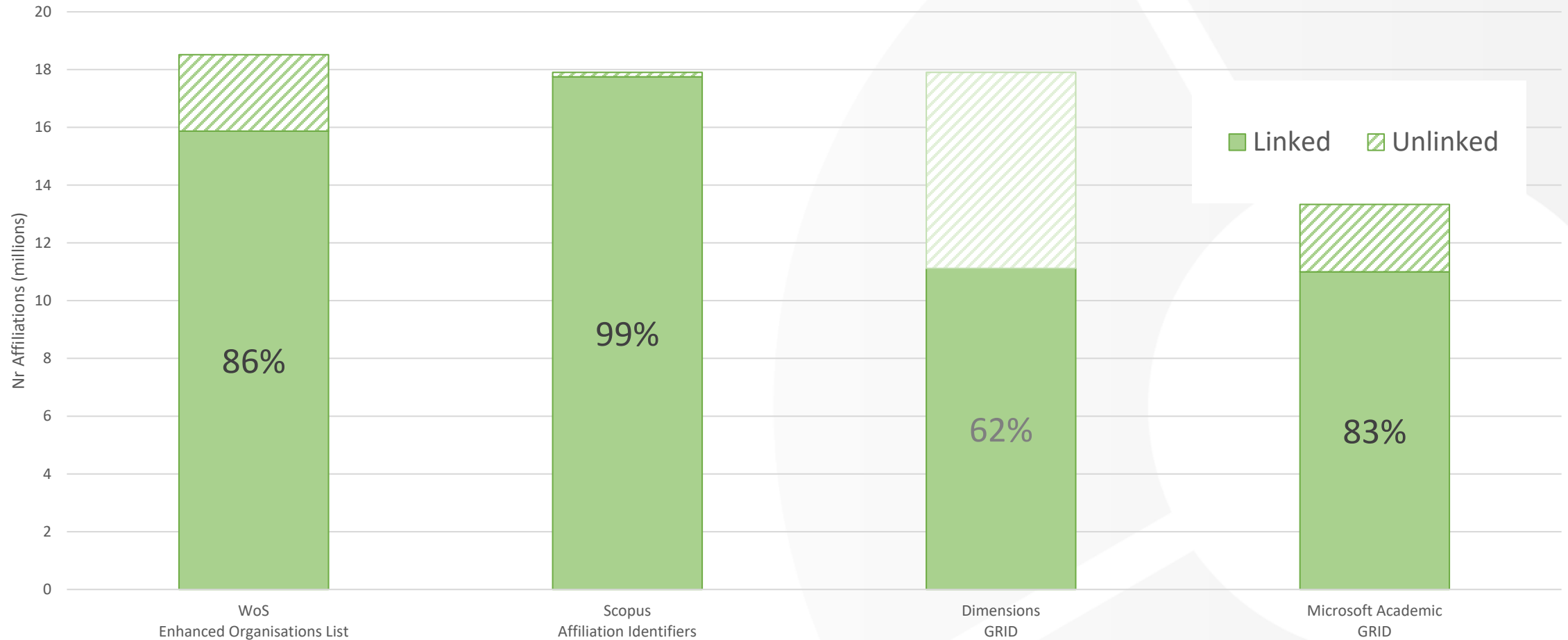
Phantom references in Web of Science

<i>Reference in WoS</i>	<i>Reference in original publication</i>
WANG J, 2006, CHINESE CHEM LETT, V17, P49	J. Wang, J.K. Carson, M.F. North, D.J. Cleland, Int. J. Heat Mass Transfer 49 (17) (2006) 3075–3083.
KANBER B, 2013, CEREBROVASC DIS S2, V35, P21	Kanber B, Hartshorne TC, Horsfield MA, Naylor AR, Robinson TG, Ramnarine KV. Dynamic variations in the ultrasound gray-scale median of carotid artery plaques. Cardiovasc Ultrasound 2013a;11:21.
ZHANG K, 2014, IEEE T PATTERN ANAL, V1, P1	K. Zhang, H. Chen, G. Wu, K. Chen, H. Yang, High expression of SPHK1 in sacral chordoma and association with patients' poor prognosis, Med. Oncol. 31 (11) (2014) 247.

Comparison of author affiliations links in bibliographic databases

An abstract graphic consisting of several overlapping blue shapes. On the right side, there is a large circle. To its left, a larger, semi-transparent circular arc overlaps it. Several thick blue lines radiate from the right side of the circle towards the left, creating a sense of movement or connection. The background is a solid blue color on the left, transitioning to white on the right.

Percentage of linked affiliations in bibliographic databases



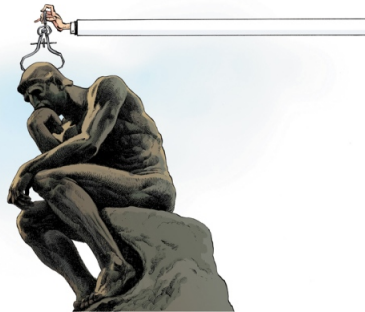
Concluding remarks:

- Important shortcomings regarding:
 - reference lists
 - field assignment
 - document type classification

Concluding remarks: Is more always better?

More is better

3) Protect excellence in locally relevant research. In many parts of the world, research excellence is equated with English-language publication. Spanish law, for example, states the desirability of Spanish scholars publishing in high-impact journals. The impact factor is calculated for journals indexed in the US-based and still mostly English-language Web of Science. These biases are particularly problematic in the social sciences and humanities, in which research is more regionally and nationally engaged. Many other fields have a national or regional dimension — for instance, HIV epidemiology in sub-Saharan Africa.



More need not be better

