WHAT DO WE LOSE WITH MICROSOFT ACADEMIC?

WHAT WE HAVE, WHAT WE NEED, AND HOW TO MAINTAIN IT?

BIANCA KRAMER & CAMERON NEYLON - STI2021

Utrecht University

CC KI CURTIN OPEN KNOWLEDGE INITIATIVE

WHAT IS THE ISSUE HERE?

CURTIN OPEN KNOWLEDGE INITIATIVE

[Microsoft Academic Graph]...will be supported until the end of calendar year 2021, upon which time MAS will be retired

GOOD NEWS AND BAD NEWS

- Microsoft made the data available under an open (ODC-By) license which offers opportunities to develop new tools and resources, particularly based on machine learning approaches.
- Much of the technology is open source and can be adapted or rebuilt to provide a replacement. Several groups are working on this.
- Some aspects of what made MAG so useful were dependent on licenses to content that will be difficult or impossible to renegotiate. Some elements of the workflows are dependent on the Bing infrastructure.
- A replacement will need to work smarter in some ways, especially if we want to reach beyond the core of the academic record (e.g. beyond DOIs)
- Efforts to replace MAG largely focus on inferring or scraping metadata and are therefore dependent on access to semi-structured metadata or full text
- Structured metadata resources are also moving forward in leaps and bounds. Increasing the upstream provision of structured metadata (i.e. by publishers) is a complementary strategy route towards a rich open metadata environment

Can we rebuild it? Yes...but...

...how would we know?

...and what content do we have (or not) ?

A QUICK SEGUE ON THE DATA

THE CURTIN OPEN KNOWLEDGE INITIATIVE

@COKIPROJECT - HTTP://OPENKNOWLEDGE.COMMUNITY

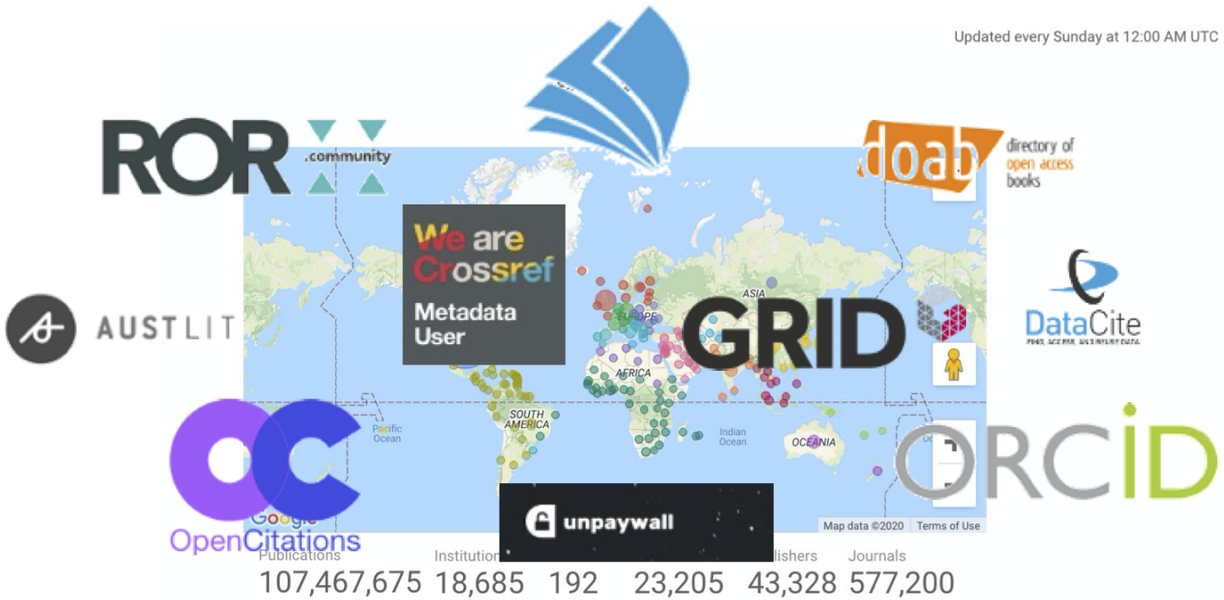
CURTIN OPEN KNOWLEDGE INITIATIVE

"Our goal is to **change the stories that universities tell about themselves**, placing open knowledge at the heart of that narrative"

Utrecht University

CURTIN OPEN KNOWLEDGE INITIATIVE

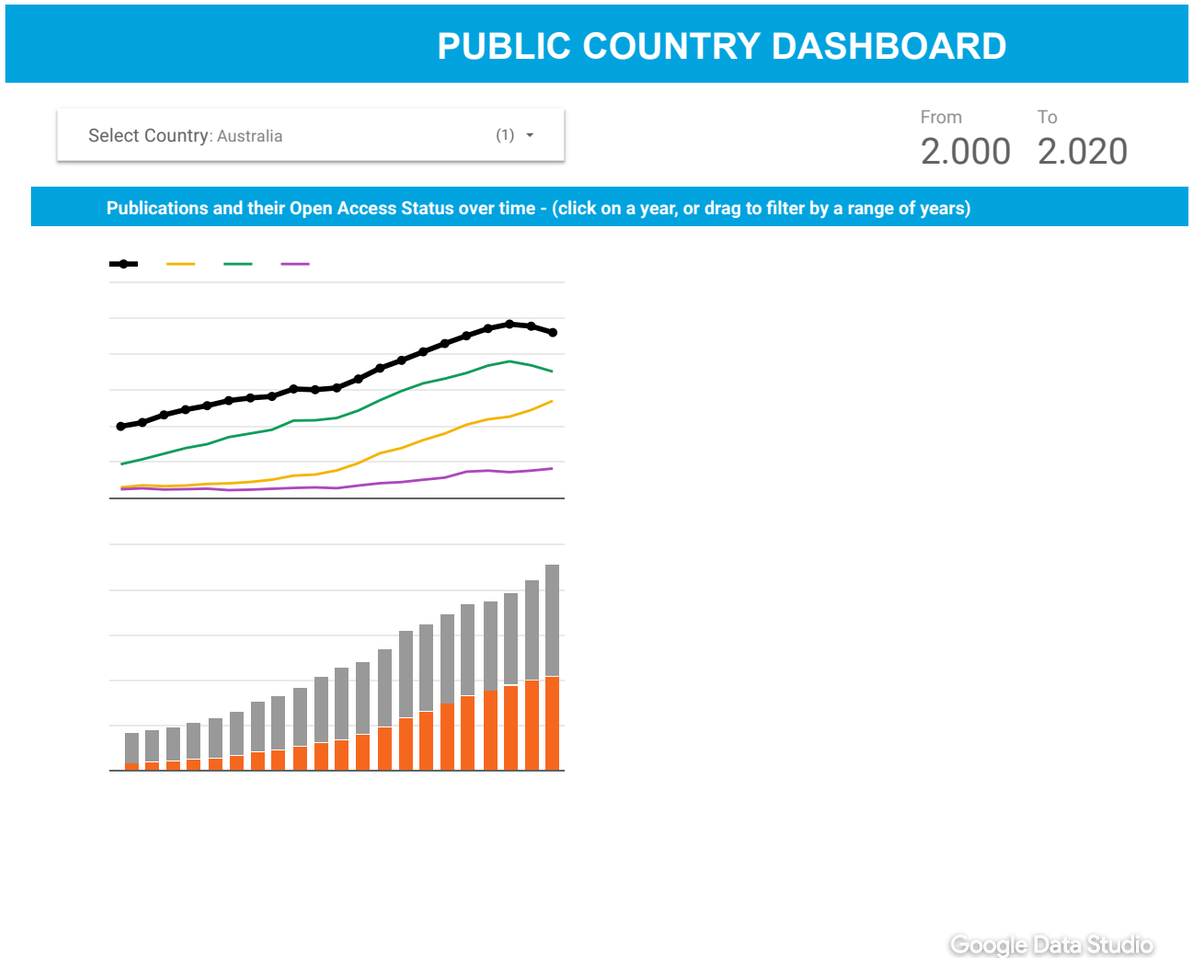
THE DATA



Explore Different Views of the Data

Institution View | Country View | Funder View | Publisher View | Group View

INFORMATION



http://openknowledge.community/dashboards/

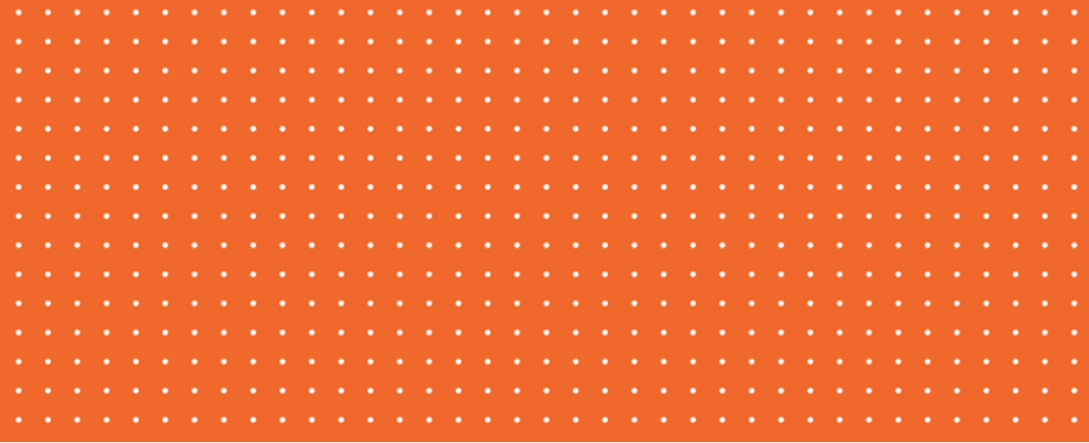
CURTIN OPEN KNOWLEDGE INITIATIVE

Utrecht University

...but because we focus on provenance and transparency, it's also good for...

COMPARING MAG WITH CROSSREF METADATA

WHO HAS WHAT, AND WHAT DO WE LOSE?



CURTIN OPEN KNOWLEDGE INITIATIVE

THE ANALYSIS

Data is derived from the following sources

- Crossref - weekly dump via Metadata Plus program
- *Microsoft Academic - Affiliation and authorship data via biweekly dump*
- GRID - Information on organisations via regular data dump

Data is integrated and processed via Observatory Platform, an open source workflow system developed within COKI to integrate data related to scholarly communications. The code is available on Github

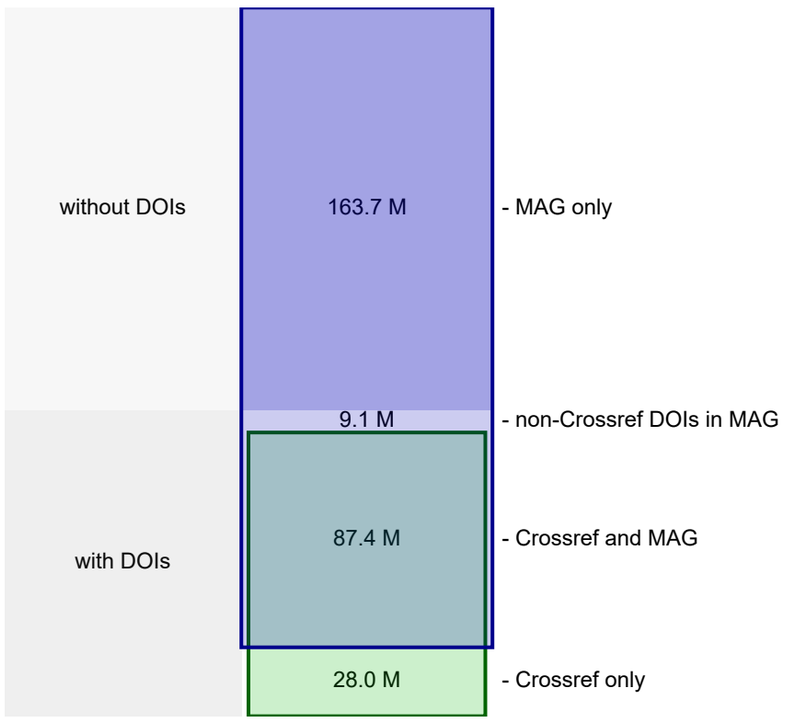
For this analysis we largely use the "DOI Table" which is an aggregation of multiple data sources that provide information on the outputs identified by Crossref DOIs. To supplement this we use a de-normalised version of the MAG database.

Code for the analysis, including the queries and processing and a local copy of the derived data are available at the presentation Github repository.

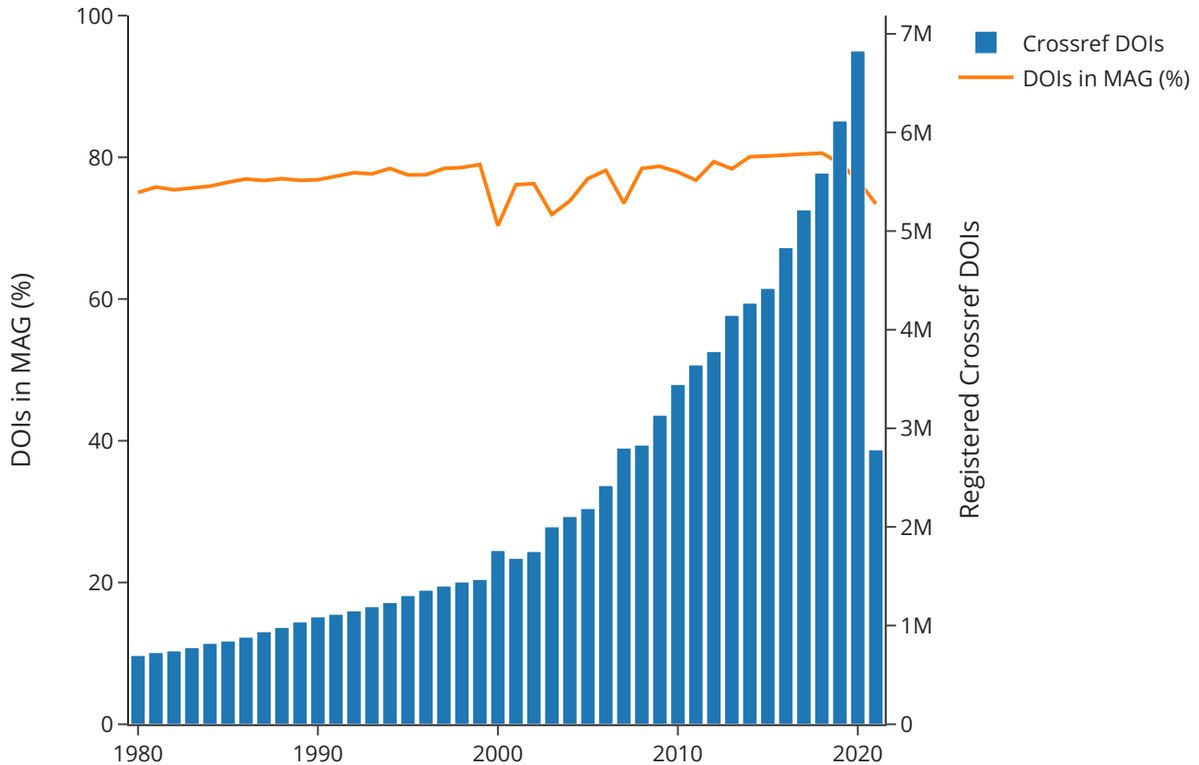
THE ANALYSIS

- Additional metadata for Crossref DOIs / metadata in MAG for non-DOIs
- Metadata on:
  - Affiliations
  - Abstracts
  - Citations
  - (Open) References
  - Subjects
- Split by publication type, year of publication
- All time and for Crossref "current" (2019-21)
- This analysis uses data snapshots from 18 July 2021

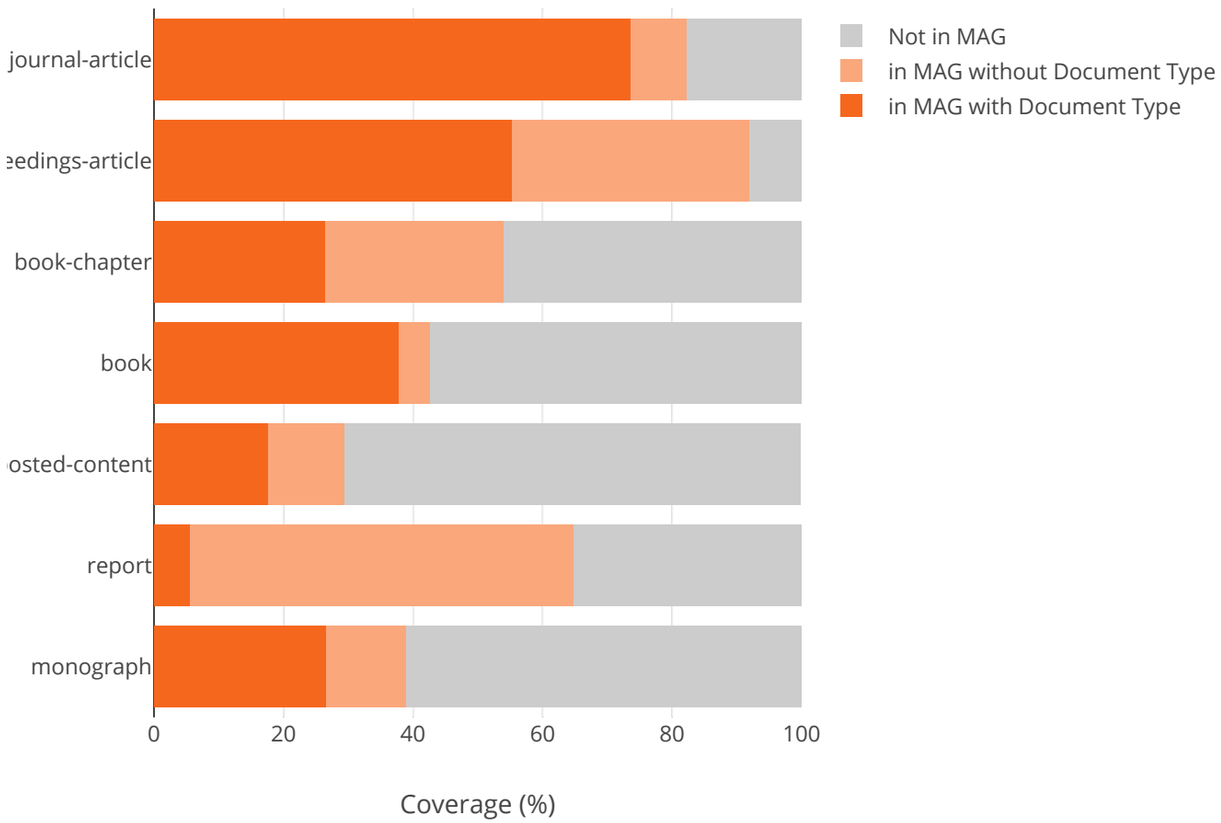
MAG VS CROSSREF COVERAGE



CROSSREF RECORDS IN MAG



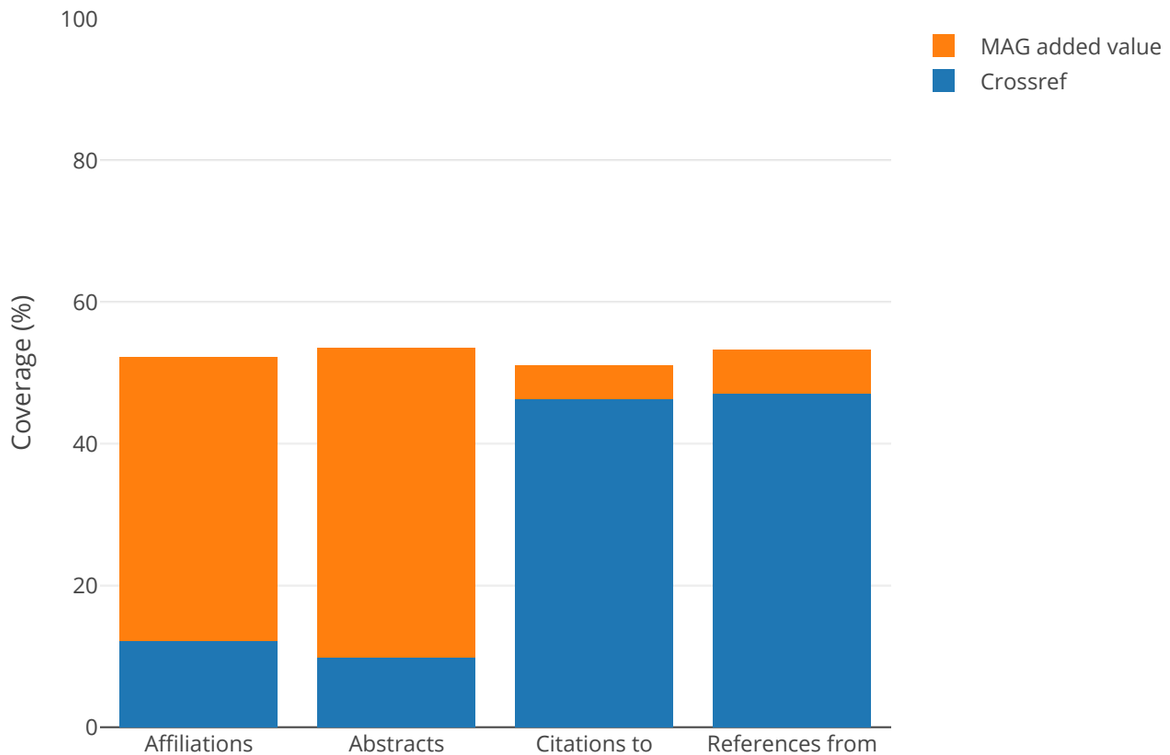
CROSSREF RECORDS IN MAG - BY CROSSREF TYPE



CURTIN OPEN KNOWLEDGE INITIATIVE

Utrecht University

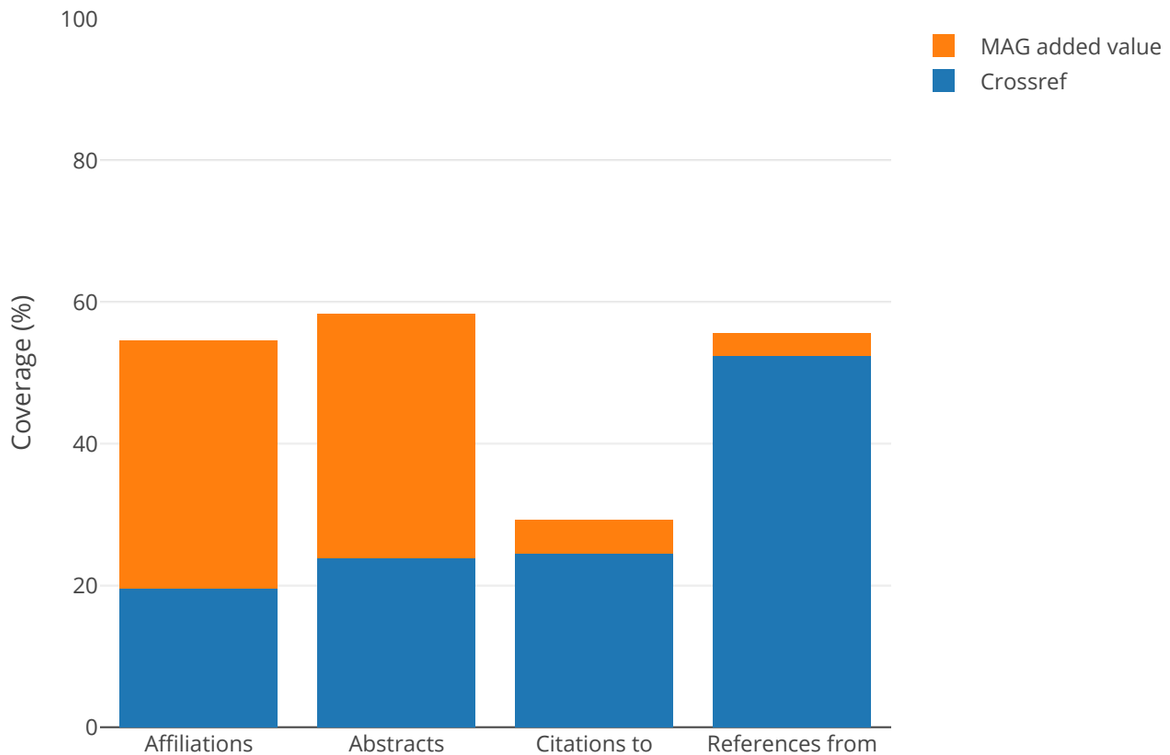
CROSSREF RECORDS: MAG ADDED VALUE



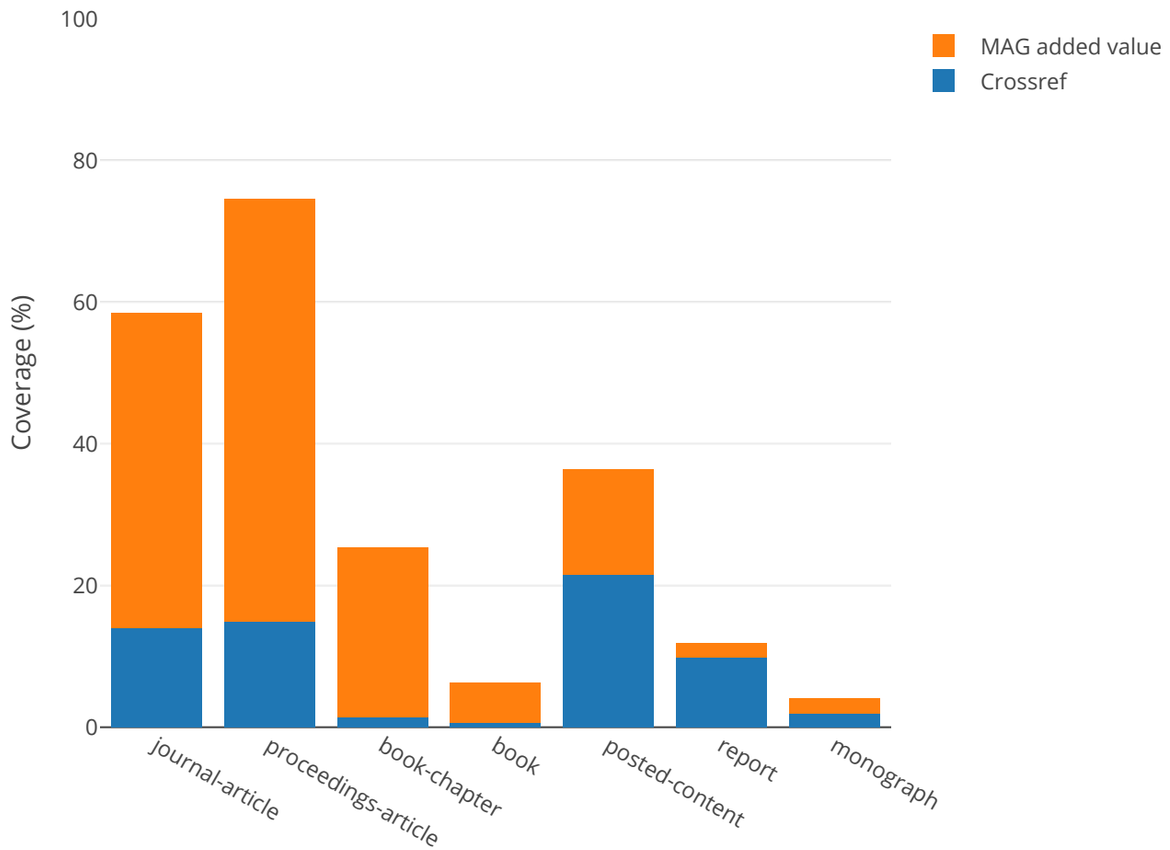
CURTIN OPEN KNOWLEDGE INITIATIVE

Utrecht University

CROSSREF RECORDS: MAG ADDED VALUE (CURRENT)



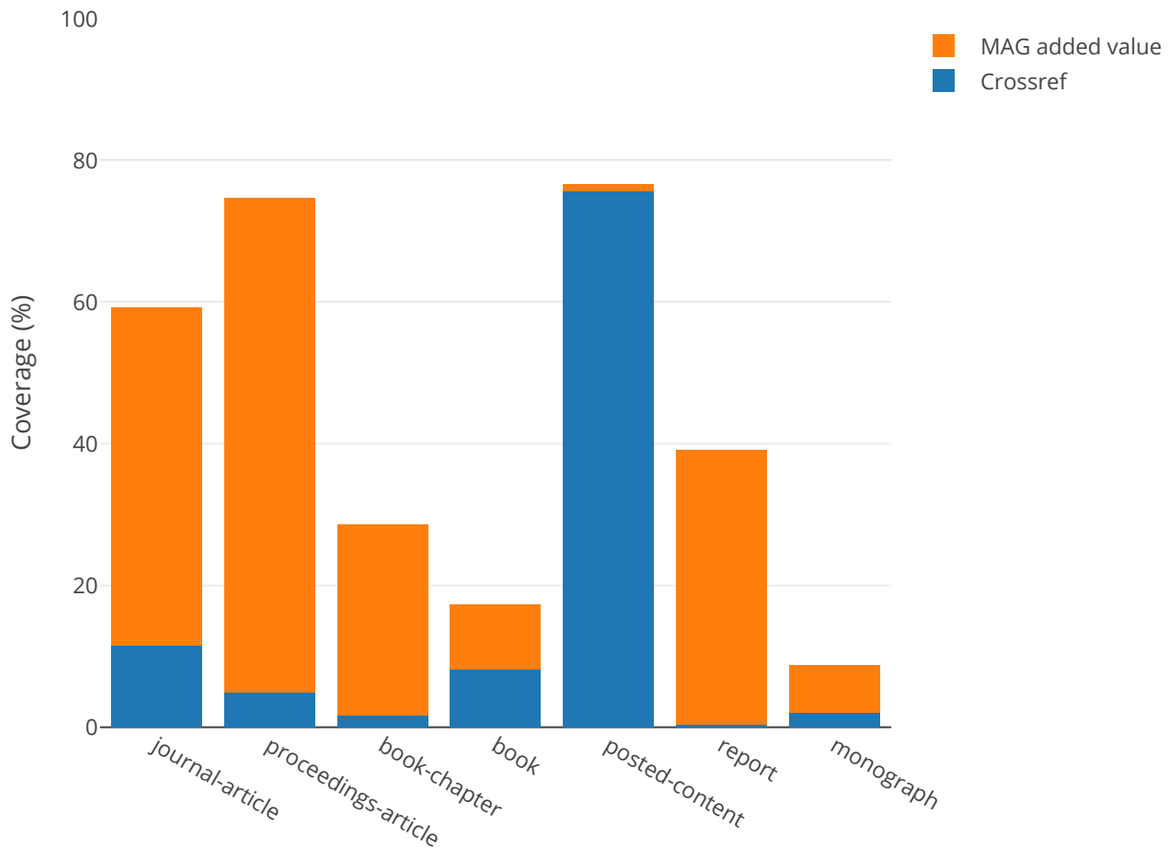
MAG ADDED VALUE BY CROSSREF TYPE - AFFILIATIONS



CURTIN OPEN KNOWLEDGE INITIATIVE

Utrecht University

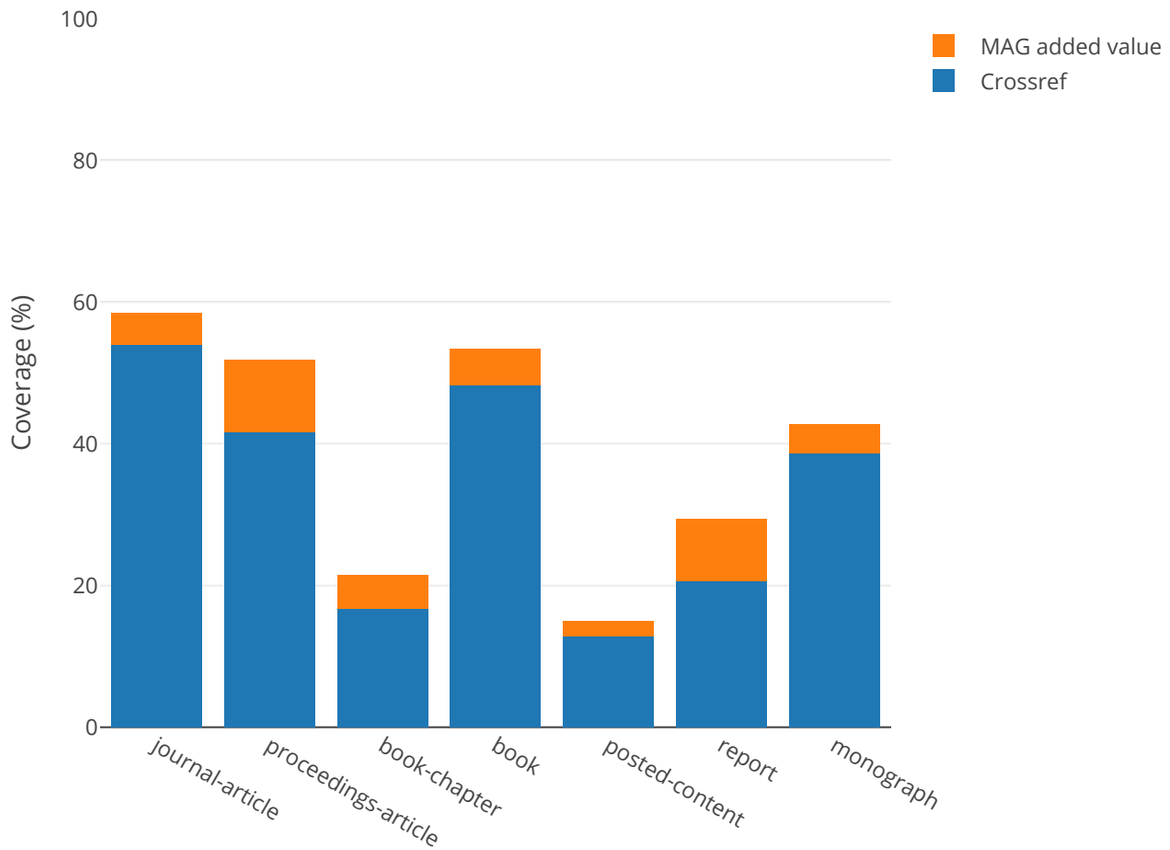
MAG ADDED VALUE BY CROSSREF TYPE - ABSTRACTS



CURTIN OPEN KNOWLEDGE INITIATIVE

Utrecht University

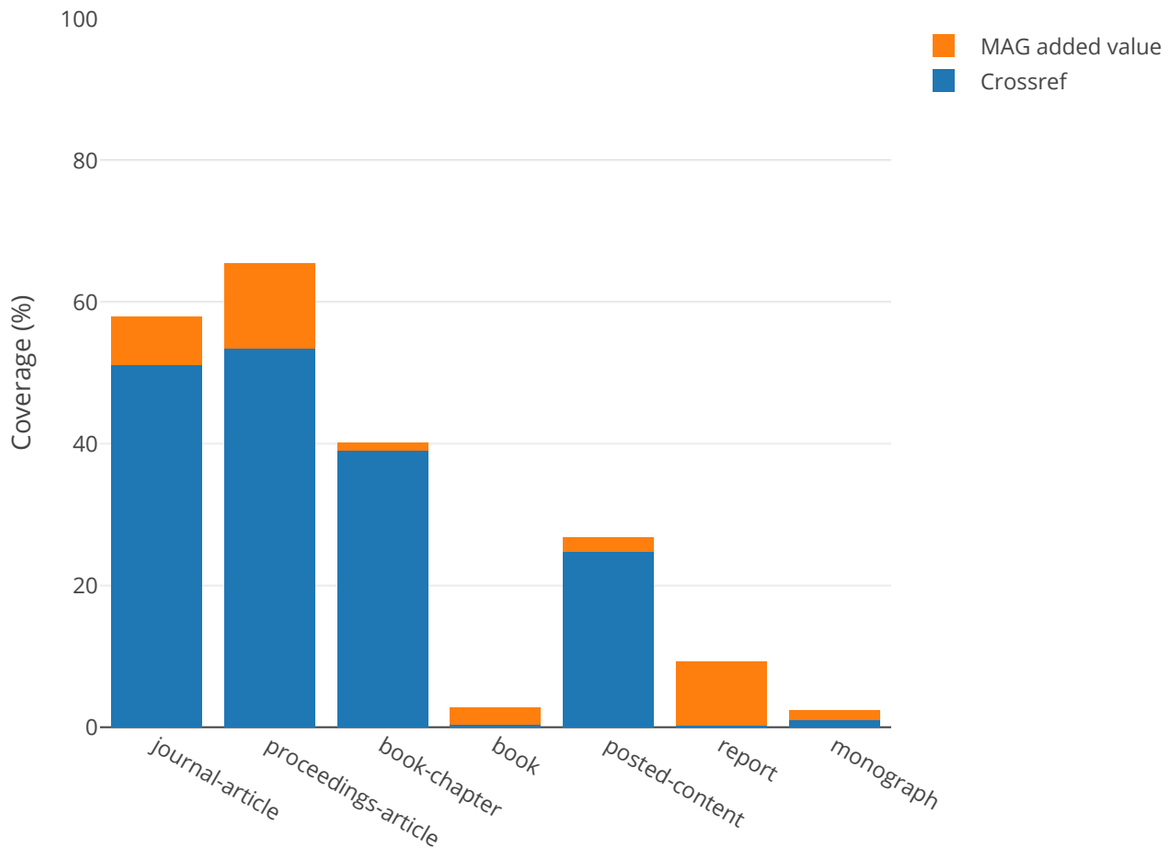
MAG ADDED VALUE BY CROSSREF TYPE - CITATIONS TO



CURTIN OPEN KNOWLEDGE INITIATIVE

Utrecht University

MAG ADDED VALUE BY CROSSREF TYPE - CITATIONS FROM



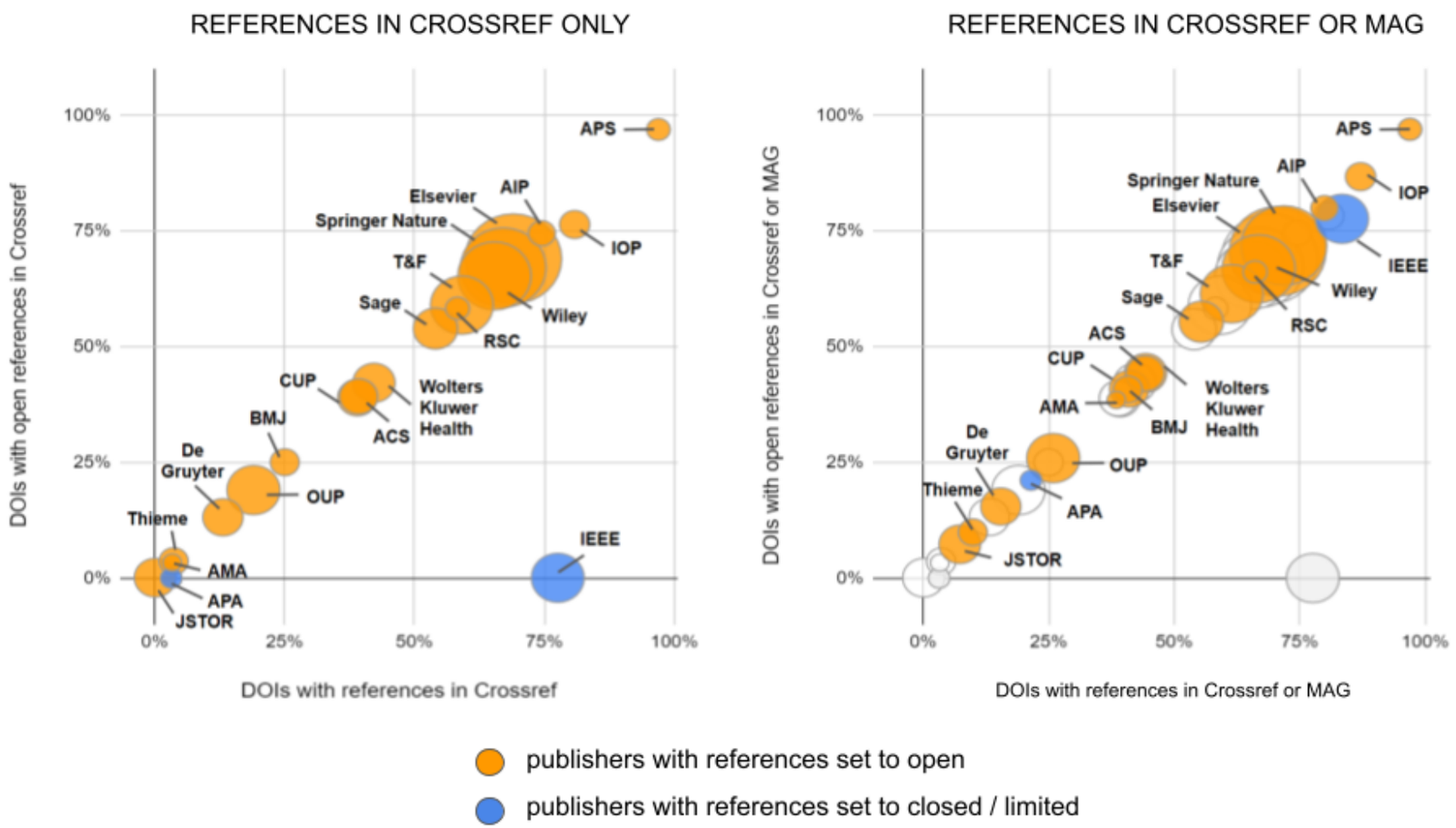
CURTIN OPEN KNOWLEDGE INITIATIVE

Utrecht University

CASE STUDIES

CURTIN OPEN KNOWLEDGE INITIATIVE

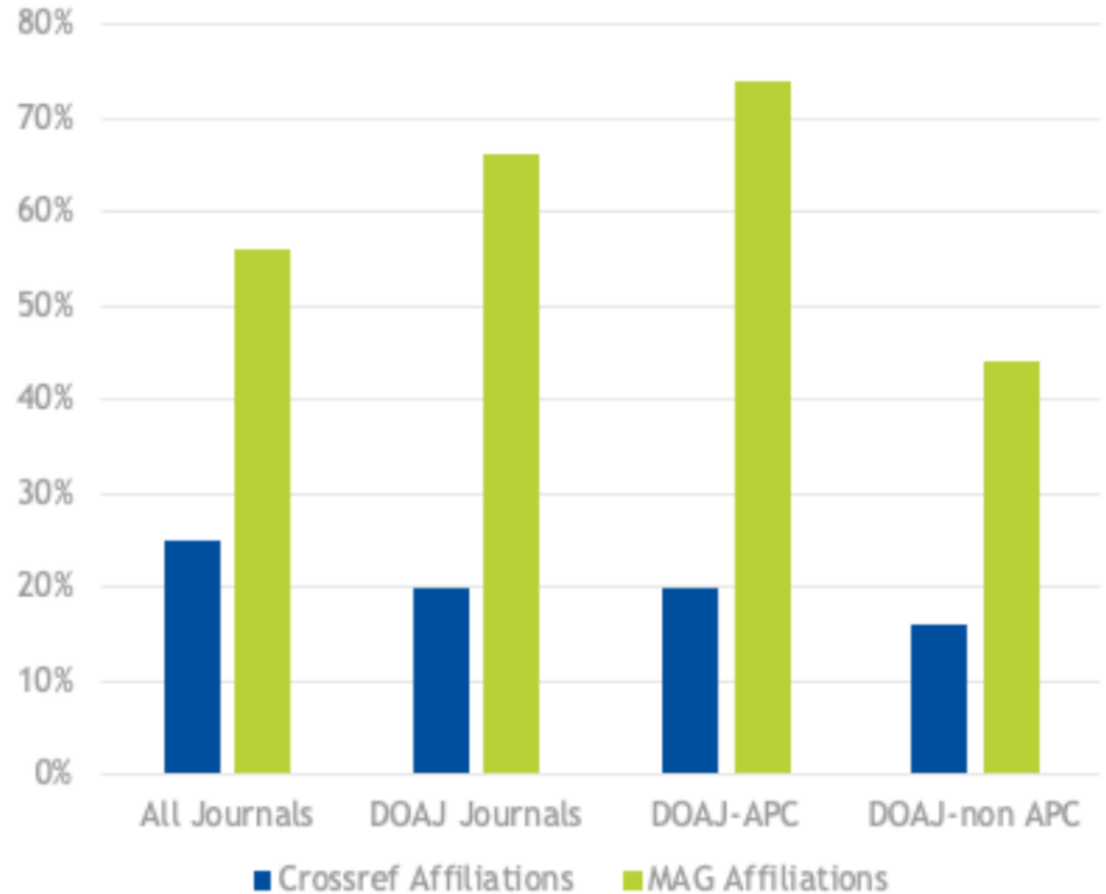
REFERENCES IN CROSSREF - ADDED VALUE MAG (BY PUBLISHER)



CURTIN OPEN KNOWLEDGE INITIATIVE

Utrecht University

COVERAGE OF AFFILIATIONS BY JOURNAL CATEGORY



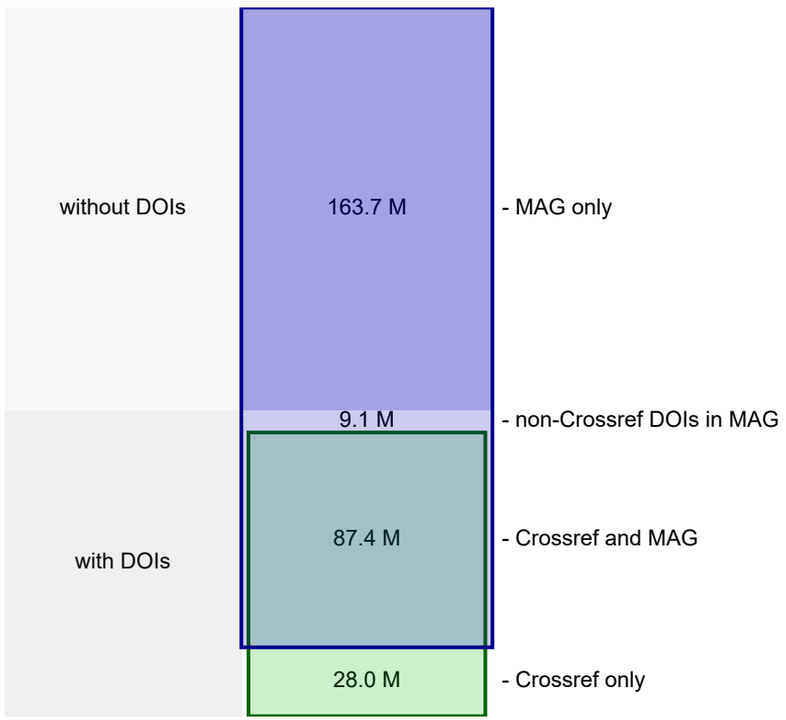
CURTIN OPEN KNOWLEDGE INITIATIVE

Utrecht University

WHAT IS IN MAG BUT NOT IN CROSSREF?

CURTIN OPEN KNOWLEDGE INITIATIVE

MAG VS CROSSREF COVERAGE



COVERAGE OF NON-DOIS IN MAG

MAG document type	count	% DOI
Journal	88,4 M	75.6%
NULL	82,5 M	22.6%
Patent	64,9 M	0.0%
Thesis	5,6 M	5.8%
Conference	5,1 M	87.3%
Repository	4,7 M	40.4%
Book	4,5 M	10.7%
BookChapter	3,9 M	89.8%
Dataset	0,1 M	58.5%
Total	260,0 M	37.0%

CONCLUSIONS

CURTIN OPEN KNOWLEDGE INITIATIVE

CONCLUSIONS

- MAG was a great resource, and its openness will enable us to build back better
- There are challenges in systematically gaining access to underlying information to replicate some aspects of MAG (eg subject classifications)
- Improvements to the provision of structured metadata by publishers (ROR, ORCID, I4OC, I4OA etc) have a great potential to improve metadata coverage
- But gaps will still need to be filled (and back-filled). Access to content, including abstracts and ideally full-text is critical to make this happen beyond the low hanging fruit
- There are big gaps and under-represented areas, including content beyond journal articles and smaller (and often non-APC open access) journals

To safeguard transparent data collection, provenance and sustainability ...

...we believe services seeking to replace MAG should demonstrate commitment to ...

… the Principles of Open Scholarly Infrastructure (POSI)

- TRANSPARENCY
- COMMUNITY GOVERNANCE
- INSURANCE PLAN FOR LONG-TERM AVAILABILITY

COLOPHON

- Code, data and slides on Github: https://github.com/Curtin-Open-Knowledge-Initiative/what_do_we_lose_mag
- PDF slides generated with: https://github.com/astefanutti/decktape

COKI TEAM

Centre for Culture and Technology

- Cameron Neylon
- Lucy Montgomery
- Katie Wilson
- Chun-Kai (Karl) Huang
- Chloe-Brookes Kenworthy
- Tim Winkler

Curtin Institute for Computation

- Richard Hosking
- Rebecca Handcock
- Aniek Roelofs
- Jamie Diprose
- Tuan Chien

Educopia Foundation

- Katherine Skinner
- Rebecca Meyerson

Funding from

- Research Office at Curtin
- Faculty of Humanities
- School of Media, Creative Arts and Social Enquiry
- Andrew W. Mellon Foundation
- Arcadia