Now is the time to work together toward open infrastructures for scholarly metadata

As part of Open Access Week 2021, **Ginny Hendricks**, **Bianca Kramer**, **Catriona J. Maccallum**, **Paolo Manghi**, **Cameron Neylon**, **Silvio Peroni**, **David Shotton**, **Aaron Tay**, *and* **Ludo Waltman** *make the case for community action toward open infrastructures for scholarly metadata. Discussing the impending loss of Microsoft Academic, the need for more sustainable infrastructures and the contributions these can make to research equity, they outline how stakeholders across the scholarly communications ecosystem can contribute to making open metadata a reality now.*

Recent events highlight the importance and the fragility of infrastructures for open scholarly communication. The $4.5 million grant awarded to OurResearch, the $3.47 million grant awarded to Invest in Open Infrastructure (IOI), alongside the development of IOI's new strategic plan are all welcome. The news in May 2021 that Microsoft Academic was closing down, less so.

Microsoft Academic has been one of the key players providing metadata about scholarly publications. Metadata on authors, affiliations, abstracts, citations, subject fields, etc. is of crucial importance for scholarly literature search and research assessment. Microsoft did a great job by making this data openly available (even though the license conditions put some limits on the kinds of re-use allowed). It's heartening that OurResearch (of Unpaywall fame) has announced, through OpenAlex, to continue producing open metadata about scholarly publications, partly building on the work done by Microsoft. However, the forthcoming closure of Microsoft Academic demonstrates the fragility of infrastructures that do not meet appropriate standards of governance, whether they are provided by large financially secure commercial companies, or small grant-funded academic initiatives.

> The very reason that IOI and related initiatives such as SCOSS exist, is an acknowledgement of the current vulnerability of open infrastructures

One response to this issue has been greater clarification of what appropriate standards could be, for instance those specified in the Principles of Open Scholarly Infrastructure (POSI). Indeed, the POSI principles on governance, sustainability, and insurance of open infrastructures are topics that IOI hopes to research and address as part of its new strategic plan. The very reason that IOI and related initiatives such as SCOSS exist, is an acknowledgement of the current vulnerability of open infrastructures and the inadequacy of the funding mechanisms available for such infrastructures. Crossref recently committed to adopting the POSI principles. In its strategic agenda to 2025 it set out practical ways to implement the principles, such as opening its code, broadening governance, providing transparent information about all processes and policies, and clarifying licenses for the metadata and services it provides. Organisations like DataCite, Dryad, JOSS, OA Switchboard, OpenCitations, OurResearch, and ROR are taking similar steps.

What do we lose with the closure of Microsoft Academic?

Microsoft Academic is remarkable for the broad coverage of its metadata for scholarly publications. As shown in Figure 1, it provides metadata for over 170 million records that are not covered by Crossref, including over 9 million with non-Crossref DOIs and over 163 million without DOIs. Conversely, Crossref holds metadata on more than 28 million records with DOIs that are not covered by Microsoft Academic, while about 82 million DOI-bearing records are common to both platforms.

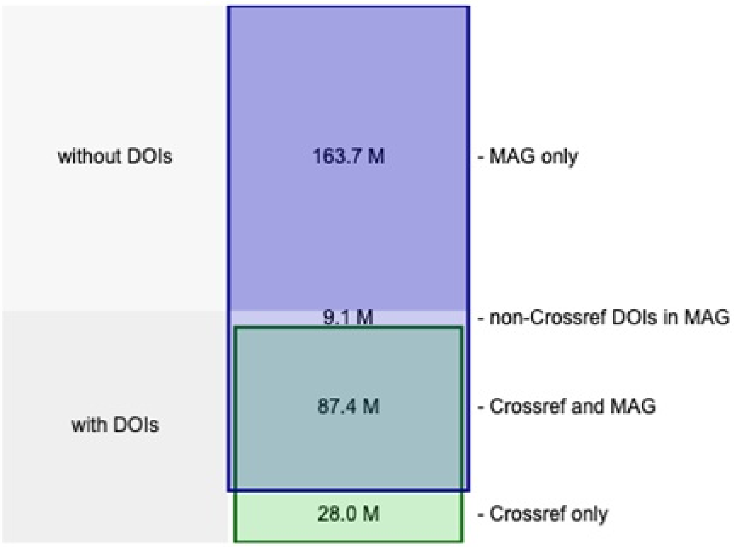

Figure 1. Comparison of coverage of Microsoft Academic (MAG) and Crossref

For the latter records, open metadata can be obtained either from Microsoft Academic or from Crossref. However, as shown in Figure 2, the metadata from Microsoft Academic is often more complete, since many publishers fail to submit complete metadata to Crossref. Thanks to the Initiative for Open Citations (I4OC), the open availability of citation metadata in Crossref has improved greatly over the last four years. In contrast, the availability of abstracts in Crossref is still limited, in particular because of the lack of support for the Initiative for Open Abstracts (I4OA) from a number of large publishers. Other metadata elements, such as affiliations, are also often missing in the metadata submitted by publishers to Crossref.

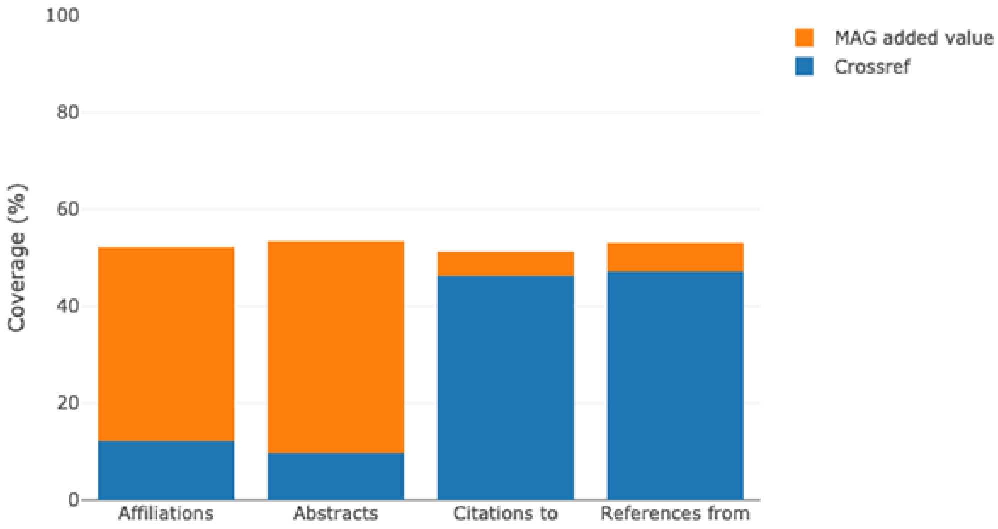

Figure 2. Added value of metadata in Microsoft Academic (MAG) for Crossref records

A unique feature of Microsoft Academic is the metadata (e.g. on subject fields, authors, affiliations, and citations) it infers at a large scale from the full text of published articles, accessed thanks to special agreements Microsoft has with publishers. While other open infrastructures, such as OpenAIRE, are also mining full texts, they lack the privileged full text access of Microsoft Academic.

The closure of Microsoft Academic data means that search engines (e.g. Lens, Semantic Scholar) and other discovery tools (e.g. Inciteful, ResearchRabbit) based on these data will give less visibility to certain types of scholarly content, such as non-DOI outputs, grey literature, and global south literature. For research assessments it will become more difficult to obtain bibliometric statistics that provide a comprehensive and inclusive perspective on the scholarly literature. In this way the closure of Microsoft Academic represents a step backwards in efforts to build structural equity in scholarly communication.

Lessons learned

The case of Microsoft Academic presents four lessons for developing open infrastructures:

1. Infrastructures for metadata of scholarly publications and other scientific outputs need to be organised according to the Principles of Open Scholarly Infrastructure (POSI).

It is encouraging that more infrastructures are making commitments to follow POSI practices. Meanwhile, organisations such as SCOSS and Invest in Open Infrastructure are striving to develop a culture where infrastructures are sustainably supported by the scholarly community. However, many more stakeholders are still needed to commit to providing financial support before this becomes established practice.

2. Organisations providing and using open metadata need to collaborate.

To prevent silos, both in datasets and their usage, collaboration is required. At surface level, this involves standards and open licenses. At a deeper level, it involves sustainable funding and exploration of collaborations. Metadata for scholarly publications should be as complete as possible, including abstracts, author identities, affiliation identities, bibliographic references, funder identities and so on. Publishers and other scholarly communication organisations need to work together with researcher funders, research institutions and individual researchers to make such metadata available in open infrastructures such as Crossref and DataCite.

Collaboration is also essential to establish mechanisms for the collective updating, curation, validation, and correction of scholarly metadata, so that it becomes superior in scope, depth, and accuracy to that currently available. This is challenging, because, with the possible exception of Crossref, we currently have *no* mechanisms for achieving this.

3. There is an urgent need for a systematic effort to develop policies that require openness of all metadata associated with scholarly publications.

Research funders and institutions need to extend their open access and open science policies by requiring publications to be FAIR (findable, accessible, interoperable and reusable). Openness of metadata is crucial to make publications findable and interoperable. The same policies can also point to international best practices (e.g. Crossref schema, DataCite schema, and OpenAIRE guidelines) to which such open metadata needs to adhere.

In addition, the provision of open metadata needs to be a formal requirement in procurement and tender processes for infrastructure platforms and publishing services, as discussed on a recent Metadata 20/20 panel. It should be a required element in transformative agreements and other deals between research institutions or research funders and publishers.

4. Metadata is always incomplete, open access to full texts is needed to fill the gaps.

Metadata is provided via different platforms, it is frequently not up-to-date with the most recent standards, ingested with varying degrees of curation, and the information needed by ever evolving applications is often not yet described by suitable metadata formats. Open access to full texts is the only way to enable metadata to be enriched algorithmically.

Towards an open ecosystem for scholarly metadata – A call to action

The closing down of Microsoft Academic will be a great loss, but it also provides a golden opportunity to create something even better. This requires concerted action across the scholarly communications ecosystem.

Publishers should commit to making complete metadata for all their works, including references and abstracts, available in a suitable open infrastructures such as Crossref or DataCite. Publishers that do not yet support the Initiative for Open Citations and the Initiative for Open Abstracts should join these initiatives. Likewise, **preprint servers and institutional repositories** should also deposit their metadata to open infrastructures.

To ensure optimal dissemination of their work, **academic researchers** should choose to report their work only in journals whose publishers support not only open access to the full text of an article, but also open, complete, and validated metadata.

Funding agencies should mandate that metadata be made open, following the lead of an increasing number of funders that now mandate that the full text of publications arising from research they fund is made open access. We support the work of the Open Research Funders Group in furthering these aims.

Research institutions and academic libraries should ensure that contracts, deals and service-level agreements made with publishers include the requirement for publishers to make complete metadata for all their works available in a suitable open infrastructure.

Infrastructure organisations and other service providers should work closely together with publishers to streamline workflows for making metadata openly available. Crossref, DataCite and other infrastructure organisations should simplify the deposition of open metadata as much as possible, in particular for smaller publishers with limited resources and technical expertise. **Publication and submission platforms** should facilitate the creation of complete metadata as each work is prepared for publication, and automated submission of this metadata to Crossref or an equivalent infrastructure. Presently, there are big gaps and under-represented areas, including content beyond journal articles and smaller (often non-APC open access) journals.

Disseminators of scholarly metadata such as BASE, CORE, Lens, OpenAIRE, OpenCitations, and others, which undertake the curation, combination, correction, enrichment, and sharing of metadata from various sources, should ensure their enriched metadata is openly available with full provenance records, using common standards and open licenses. This should not prevent them from providing their own unique services on top of the metadata they have collected and enriched.

Finally, while increasing the upstream provision of structured metadata from publishers is an important route towards an open metadata environment, gaps will still need to be filled. Access to full text content is critical to make this happen. For this reason, as well as for all other well-documented benefits of open access, institutions, funders, governments, and publishers should continue to work to make full open access a reality, with licenses that allow unrestricted text and data mining.

It is currently Open Access Week 2021, much focus is on open access to the full text of scholarly publications, but we should carefully consider that full unrestricted open access also requires open availability of rich metadata. At present, infrastructures for open metadata are way behind. We therefore appeal to all like-minded parties to work together to achieve the goal of open scholarly metadata through POSI-compliant infrastructures. Only through such infrastructures will we be able to contribute to building structural equity, the theme of this year's Open Access Week. We also encourage your involvement with initiatives such as Metadata 20/20, which aims to increase advocacy efforts by joining up with other related initiatives and organisations. Finally, we invite you to assess how the foundational infrastructures you rely on measure up to the POSI principles.

We are fully aware that there are other clusters of discussions and plans underway in the community. We would like to be part of your conversations and your work; if your aims are similar, we invite you to contact any one of us to express your thoughts and desire to collaborate to achieve this end.

The time to act as a community is now.

Note: This article gives the views of the author, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

Image Credit: Igor Starkov via Pexels.