Looking Ahead: The Research Nexus and the State of Metadata in 2050

Ed Pentz, Martyn Rittman, and Dominika Tkaczyk

As research itself changes, an increasing variety of research outputs are available. *Metadata*—including persistent identifiers (PIDs)—describes the research objects that are essential for discovery, citation, provenance, and trust. In addition to research outputs, the people doing the research, and the organizations funding and supporting the research need to be transparently identified, through, for example, ORCID iDs[1] and ROR IDs.[2] It is also essential to capture the relationships between these outputs, people, and organizations in an open and dynamic way. Before the digital age, the focus was primarily on the published paper. Now there is open access to datasets, code, materials, equipment, funders, supporting institutions, preprints, posters, and so much more that result from a single project. Each of these components can be reused, repurposed, or discussed as part of a different project. At Crossref, we use the term *the research nexus* to refer to this complex, evolving network of objects, along with descriptions of how they relate.

We see the development and description of the research nexus as key to communicating science in the next 25 years. It is much bigger than Crossref, and a number of organizations are pursuing similar goals from different perspectives. Our contribution is to collect, maintain, and make available identifiers and metadata from the organizations that publish research outputs.[3] We also seek to supplement this metadata with other community sources,[4] and to run automated enrichment strategies at scale to provide additional metadata and relationships that were not captured earlier.[5] There is potential for us and others to develop all of these areas in the coming years and decades. We still see a key role for Crossref to gather diverse metadata from the community that can be used and enhanced by others.

Ed Pentz (https://orcid.org/0000-0002-5993-8592) is Executive Director, Crossref. Martyn Rittman (https://orcid.org/0000-0001-9327-3734) is Program Lead, Crossref. Dominika Tkaczyk (https://orcid.org/0000-0001-5055-7876) is Director of Data Science, Crossref.

Opinions expressed are those of the authors and do not necessarily reflect the opinions or policies of their employers, the Council of Science Editors, or the Editorial Board of Science Editor.

https://doi.org/10.36591/SE-4801-13

Looking Back to Look Forward

Looking back to the state of metadata in 2000 can provide lessons to make predictions about the state of metadata in 2050. Metadata is a very broad area, so the focus here is on scholarly metadata for digital and online resources and Crossref's experience. Crossref was founded in 2000, so how metadata has changed in the last 25 years is reflected in how metadata has changed for Crossref both in what we collect and its purpose. The changes in metadata have reflected the changes in scholarly research and publishing, and this will continue into the future.

The growth of the internet in the 1990s, particularly the World Wide Web, created an urgent need for standardized ways to identify and describe online resources. This led to initiatives like the Dublin Core Metadata Initiative[6] in 1995, which aimed to create a core set of metadata elements for describing web resources; and the Digital Object Identifier (DOI) System to create a system for persistent identifiers that are also persistent links.

The DOI-X prototype that led to Crossref started in 1999 and created a system for linking journal reference lists.[7] A persistent identifier and standardized metadata were needed to accomplish this, so the DOI-X project was designed to test out the DOI System, along with basic journal metadata—including only the journal title, first author last name, volume, issue, first page, and article title (which was optional). It was flat, fixed, and covered one type of research output. There were no relationships, and the only other primary identifier included was ISSN.

Over 25 years, Crossref's metadata has expanded to cover 30 research output types, including journal articles, books, book chapters, reference works, conference proceedings papers, datasets/supplementary material,

CONTINUED

dissertations/theses, grants, peer review reports, preprints, working papers, reports, and standards. There is also a richer set of metadata, including licenses, references, abstracts (stretching the bounds of "metadata"), and retractions.

Over time, we have observed a growing need to identify other types of objects within the scholarly record. New types of persistent identifiers emerged, most notably ORCID iD for identifying people and ROR ID for research organisations. As a result, it became possible to capture relationships between objects. We started with citation relationships between research objects, over time expanding to contributor relationships between research objects and people, affiliation relationships between people and organisations, funding relationships between research objects and organisations, relationships between journal articles and preprints, articles and reviews, and many more. We have also moved from seeing the metadata records as static, to more dynamic with updates to the status of an item (e.g., corrections and retractions).

So the story has been one of moving from very flat XML records with minimal metadata for a limited set of traditional scholarly outputs to a rich set of records capturing a broad range of relationships for a much wider variety of outputs and other objects. Crossref refers to this as the research nexus and believes the development and description of the research nexus as key to communicating science in the next 25 years. This reflects how research has been changing, with big data, software, reproducibility, and research integrity all as major concerns. All this open, foundational, scholarly metadata drives discovery services, analytics, and supports open research, which, in the end, increases human knowledge.

There are some things that have been consistent over the last 25 years and will be for the next 25. Metadata acts as trust signals, and so provenance is critical—who created and registered it, who maintains it, and is it open or subject to copyright or licensing terms? Persistent identifiers are also a critical element of metadata—can you link to the research output, or information about it even if it changes location or a different organization takes responsibility for it? With artificial intelligence (AI) chatbots driven by large language models (LLMs), provenance and persistent identifiers are more important than ever because LLMs are statistical abstractions with no concept of citing sources or even providing information that exists (it is common for fake citations and nonexistent identifiers to be generated[8]). Improvements have been made on this front, but it is a problem inherent to how LLMs function, so metadata and persistent identifiers can help solve this challenge.

Research Nexus in 2050

Research practices and outputs will change over the next 5 years. While journal articles will still be important, we expect that what is considered the scholarly record will expand, and therefore new metadata, identifiers, and relationships will be needed. For example, there will be an increasing need to identify and capture relationships between software code, computational notebooks, virtual/augmented reality experiences, brain–computer interface recordings, AI and machine learning–assisted research, and forms of scholarly communication we have not yet imagined. As a result, new types of persistent identifiers and relationships might be emerging, and the scholarly infrastructure will have to be adapted to handle them.

In the coming decades, we would expect more of a focus on reproducibility and reliability in research outputs. This could mean more of an emphasis on publishing complete results sets, including associated code and data. It is also likely to lead to changes in incentives for researchers: Rather than the traditional publication and citation counts, they may be assessed on the standard of their research practices, broader impact assessments, and activities that take place alongside research (such as advocacy and public engagement). The broadening of assessment approaches will mean a broadening of the need to track a more diverse set of research outputs. Here, the research nexus, and the metadata, identifiers, and relationships that are its foundation, has a key role to play.

The challenge of metadata quality will likely shift from basic accuracy and completeness to capturing nuanced and dynamic context and relationships. In such a complex and dynamic scenario, the scholarly community will increasingly rely on machine learning systems to help identify all relationships both early in the publishing workflows and further downstream. At the same time, we hope automated strategies enriching the scholarly record will be used responsibly—with sufficient quality control, transparency, keeping provenance, and considering the carbon footprint of using resource-intensive approaches. Human expertise will hopefully remain crucial for curating the relationships and controlling what the machines are doing, and that it is done in an open and transparent way. This is especially crucial where the scholarly metadata and relationships are used to make key decisions about research and people. Community involvement and input will also be important in ensuring metadata quality and what policies apply to how the metadata is used and interpreted.

Another challenge will be the globalization of research outputs. Many more regions of the world now generate large volumes of scholarship, and in a wide variety of languages. It is necessary to capture metadata in multiple languages, but also essential that the systems that collect and disseminate metadata are accessible to those whose main language is not English—the current lingua franca. We need to invest in documentation, support structures, and

CONTINUED

knowledge sharing that is adapted to different linguistic and cultural situations, to ensure that there is no regionalisation and fragmentation of the knowledge-sharing infrastructure. For Crossref, a large part of this is listening to the needs of our current, highly diverse membership, as well as reaching out to those who are not yet fully part of our community.

Challenges in Getting There

Cultural change is hard, and in order for the vision of the research nexus to come to fruition, we have to work on open data and open research becoming the default and change incentive structures for how research is assessed. Publishing an article in a high Impact Factor journal is not sufficient. Another challenge is financial—research and scholarly publishing require significant resources, as does the creation and maintenance of high-quality metadata. This all needs support from, and collaboration between, government, funding bodies, research institutions, researchers, open infrastructure providers, scholarly societies, and commercial companies.

A key underpinning for our vision of the future are the Principles of Open Scholarly Infrastructure (POSI).[9] These are 16 principles covering open data, sustainability, and inclusive governance that are essential for metadata and will continue to be as relevant in 2050 as they are now.

Supporting and embracing technological innovation in a measured way and being globally inclusive are also very important. More work is needed to expand the scholarly record to more fully include the Global South and expand the scholarly record to cover areas such as grey literature and Indigenous Knowledge.

All the elements are in place for ensuring that in 2050, we will have overcome the current challenges so that metadata supports a fully open and dynamic global research ecosystem.

References and Links

1. https://orcid.org/
2. https://ror.org/
3. Hendricks G, Tkaczyk D, Lin J, Feeney P. Crossref: the sustainable source of community-owned scholarly metadata. Quant Sci Stud. 2020;1:414–427. https://doi.org/10.1162/qss_a_00022.
4. Rittman M. Retraction Watch retractions now in the Crossref API. Crossref Blog. January 29, 2025. https://doi.org/10.13003/692016.
5. Tkaczyk D, Buttrick A. Metadata matching 101: what is it and why do we need it? Crossref Blog. May 16, 2024. https://doi.org/10.13003/aewi1cai.
6. https://www.dublincore.org/
7. Atkins H, Lyons C, Ratner H, Risher C, Shillum C, Sidman D, Stevens A. Reference linking with DOI: a case study. D-Lib Magazine 2000;6. https://doi.org/10.1045/february2000-risher.
8. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. J Med Internet Res 2024;26:e52935. https://doi.org/10.2196/52935.
9. Bilder G, Lin J, Neylon C. The principles of open scholarly infrastructure. 2020. https://doi.org/10.24343/C34W2H.