Check for updates

Article

Building open scholarly infrastructure: A journey of collaboration and diplomacy

Information Services and Use
2024, Vol. 44(4) 285–290
© The Author(s) 2024

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/18758789241296761
journals.sagepub.com/home/isu

Edward Pentz[1]

Abstract

This article expands on the Miles Conrad Award lecture delivered at the NISO Plus 2024 conference. Drawing from over three decades in scholarly publishing, including twenty-four years at Crossref, Edward Pentz, the Executive Director of Crossref, explores the critical role of collaboration and diplomacy in developing open scholarly infrastructure. The piece examines key inflection points in scholarly communication, lessons learned from collaborative initiatives, and future challenges and opportunities in the field. It also reflects on the importance of diversity, equity, and inclusion in shaping the future of open scholarly infrastructure.

Keywords

Miles Conrad Lecture 2024, Edward Pentz, Crossref, Principles of Open Scholarly Infrastructure, POSI, Research Organization Registry, ROR

Received: 9 October 2024; accepted: 10 October 2024

Introduction

Receiving the Miles Conrad Award is a real honor and looking back at the previous recipients is humbling. Miles Conrad's legacy, embodied in the founding of NFAIS (now part of NISO), centered on bringing together diverse information industry stakeholders to work collaboratively. An impetus for the founding of NFAIS was the launch of Sputnik, the ensuing U.S. space program, and the need to advance U.S. science—it was an inflection point. I will be talking about inflection points and how collaborative efforts—and diplomacy—are critical to taking advantage of them.

I have spent most of my career working at, and with, collaborative associations and projects, so I am very glad to get this award. I did want to note that since this award was started in 1968 about fifty-six people have received it—forty-four men (including me) and thirteen women. Since 2013 the mix has gotten better—there have been six women and five men. So, I wanted to note my privilege as a university-educated, straight white man—it has helped my career. Another big help was that when we had children, my wife, Fiona Kelbrick, and I agreed that she would take a career break to look after them. This enabled me to focus more on my job and travel a lot—so, a big thank you to her.

Being aware of these issues of inequality means that at Crossref we have focused on creating an environment that promotes a positive work/life balance and diversity, equity, and inclusion, and we have a range of policies to achieve this. Some highlights are the following: we offer unlimited sick time and 12 weeks of paid family or medical leave. Flexible, remote work has definitely helped with work/life balance and diversity—Crossref is fully distributed with forty-six employees in ten different countries. Over the last 5 or 6 years, we have overhauled our recruitment practices to ensure that we get a diverse pool of applicants, for example, keeping jobs open for longer, advertising in a wide-range of places, carefully crafting the language in job descriptions to remove any biases, and including a salary range, but there is still a lot to do and it is an ongoing process.

[1]Crossref, Lynnfield, MA, USA

Corresponding author:
Edward Pentz, Crossref, 50 Salem Street, Lynnfield, MA 01940, USA.
Email: epentz@crossref.org

As I look back on my journey from a commercial, subscription-based publishing environment to leading a non-profit provider of open scholarly infrastructure, I am struck by the parallels with Crossref's own evolution. What began as a solution to a narrow problem involving closed metadata from large scientific, technical, and medical publishers has grown into an essential component of open scholarly infrastructure serving over twenty-thousand members from more than one hundred and fifty countries.

Last year's Miles Conrad lecture by Dr. Safiya Noble on Decolonizing Standards was inspiring.[1] A couple of things that resonated with me were her points about rethinking "who gets to control knowledge and information" and that during the Cold War and the National Science Foundation's early years the approach was to keep knowledge away from scientists in nations of color. 2021's recipient, Heather Joseph, also highlighted the importance of Open Knowledge and open access to knowledge citing Article 27 of the UN Universal Declaration of Human Rights:[2]

1. *Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts, and to share in scientific advancement and its benefits*.
2. *Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary, or artistic production of which he is the author*.[3]

She also cited the UN's Sustainable Development Goals[4] (SDGs)—a blueprint and action plan through 2030 for ensuring a better future for all citizens of the world. Open access to knowledge is called out in goal sixteen but supports all the goals. Scholarly publishers are signing up to the STM Association's SDG Publisher Compact explicitly tying their publishing activities to helping achieve the SDGs.

My take on this is that foundational, open scholarly infrastructure that supports Open Research is vitally important to these larger goals, which are moral imperatives and critical to advancing human knowledge and improving the human condition. And like Heather, I did not start out my career thinking about these larger goals, but I have been on a journey and I have since come to views these larger goals as essential.

The power of collaboration and diplomacy

Looking back at the organizations and projects I have been involved with there are important themes:

- Solving collective problems takes collaboration and diplomacy to bring together a group of stakeholders, balance their different concerns, build trust, and hopefully reach consensus on a way forward.
- Patience and focus are very important. The gestation period for projects can be quite long, and there are many discussions and meetings. The results are not always perfect, but if things are set up properly they can evolve over time.
- Being non-profit has been essential for foundational, open scholarly infrastructure initiatives and services.
- Long-term sustainability is also essential. Initiatives and organizations need sustainability models (aka business models) that involve mission aligned income generation and a surplus for investing in the service, investing in new initiatives, and building financial reserves for security and to weather downturns and unexpected events.

Inflection points

There are moments when something catalyzes action and big changes are possible and other times where change can be a struggle. I will be mentioning a number of key turning points where big changes happened during my career in scholarly communications.

I have been in scholarly publishing for about thirty-two years, and February 1, 2024 was my twenty-fourth anniversary at Crossref. I was lucky early in my career because I started in publishing in the early 1990s just before a major inflection point—the arrival of the World Wide Web and the rapid expansion of electronic publishing—in particular online publishing on the Web.

Tim Berners-Lee wrote a proposal in 1990 for a distributed hypertext system,[5] which was the basis for the World Wide Web. Things really took off with the launch of the National Center for Supercomputing Applications (NCSA) Mosaic web browser in 1993[6] followed by the Netscape browser in 1994.

Around this time, I had an editorial role working on some life science journals and we were working on a project to get the table of contents listed on BUBL (BUlletin Board for Libraries)—a Gopher service.[7] This got me interested in online things and I started reading *Wired* magazine. I was reading it 1 day at my desk at lunch and Chris Gibson—a senior person in

production at the time—saw me reading it and stopped for a chat and soon after that offered me a job as an electronic publishing assistant in the brand-new electronic publishing department.

Being involved in electronic publishing in the mid-90s meant experiments and prototypes and getting journals online on the World Wide Web in HTML and PDF—and a key benefit of online journals to researchers was linking references. Every publisher had their own domain names and URL naming schemes which frequently changed, so it quickly became a nightmare. In addition, publishers signed bi-lateral linking agreements—actual legal agreements to link to one another and agree to share data. This was not scalable and benefited no one except the lawyers.

The solution to this problem was collective action because no individual publisher could solve it on their own. Working through the Association of American Publishers Enabling Technologies committee, a small group of publishers collaborated on a prototype system using Digital Object Identifiers (DOIs) and metadata in XML to enable cross-publisher reference matching and linking. The project was called DOI-X, and it created a proof-of-concept system announced at the STM meeting at the 1999 Frankfurt Book Fair.[8] But it was not just a technical problem—it took ongoing collaboration, so Crossref was founded in 2000 to ensure trust and reciprocity.

Luckily, Crossref was established as a non-profit with one member, one vote with a broad mission. The Articles of Incorporation say the organization's mission is "To promote the development and cooperative use of new and innovative technologies to speed and facilitate scientific and other scholarly research." It is important to note that it does not mention identifiers, DOIs, metadata, or publishers. Crossref is not a "PID Provider." This has given a lot of room to evolve and expand Crossref's services. At the start, Crossref metadata was limited and closed. Some publishers even objected to providing all author names and journal article titles! Crossref did not mandate more metadata. It was a diplomatic compromise to have a minimal set of metadata. Insisting on more metadata would have limited participation, but Crossref has evolved and now provides fully open metadata (including references and abstracts) for more than one hundred and sixty million research outputs.

Some key lessons from the founding of Crossref are as follows:

- Focus on solving a particular problem and use the right technology to do it (not the other way around). Crossref did not start out to assign persistent identifiers or collect metadata.
- Collaboration is essential.
- Quickly develop proof of concepts and prototypes.
- Build trust by agreeing to a set of principles and getting the governance right.
- Develop a sustainability model and be prepared to adapt quickly.
- Compromise within the framework of the solution and the principles.

There is also another factor in the founding of Crossref and that is fear. In 1999, NIH Director Harold Varmus released his E-Biomed proposal[9] to create a central, NIH-run, web-based archive for biomedical research articles. It called for a preprint service alongside published, peer-reviewed articles, all openly available with no fees. The big society and commercial publishers, and many journal editors, did not like the proposal and in 2000, after a lot of push back, the project was scaled back and became PubMed Central—which now has 9.5 million full-text articles. The E-Biomed proposal helped push the big society and commercial publishers into action because they realized they better improve their online services and linking, and this played a part in the founding of Crossref, and also led to PubMed Central, a very valuable resource.

An early critical juncture for Crossref came with a challenge around Crossref's sustainability model. Along with membership and content registration fees there was a fee for matching DOIs—in effect, a fee for looking up DOIs to enable reference linking. This was a problem and was holding back uptake—it was very expensive for members to do a lot of reference linking, which is what we were trying to get them to do more of. The fee was reviewed, and a committee prepared a proposal for the board to remove it. In a tense vote, the board narrowly approved removing the fee by a vote of 8-to-7. This was a momentous decision that could have gone either way but ultimately proved crucial for Crossref's growth.

The importance of open infrastructure and sustainability

The development of ORCID is a testament to the power of collaboration across different stakeholder groups. Once Crossref was established and proved what was possible, discussions about "author DOIs," or researcher identifiers, began in 2007/2008, but it became clear that a broader coalition was needed, including researchers, funders, universities, and research institutions. There was an important decision to move the discussions out of Crossref—although we continued to be closely involved. After a couple of years of many meetings and discussions, ORCID was incorporated in 2010 and the ORCID system launched in October 2012 and reached break-even in 2019, more than ten years after the initial discussions of the project. One of the reasons ORCID was successful was the stakeholders agreed on a set of principles—which ORCID still

follows today—before doing anything else. This was essential to build trust among the different stakeholders. The patience and persistence of everyone involved also helped.

Moving on to organizations

The most recent part of the open scholarly infrastructure is ROR (the Research Organization Registry)—an open registry of organization identifiers and basic metadata. What was the use case for this? Well, research outputs were being addressed—mainly through Crossref, DataCite, and other DOI Registration Agencies. There was ORCID for researchers, but the missing piece of the puzzle was uniquely identifying organizations—specifically, for identifying affiliations in research outputs—articles, books, data, and software—and in grants and awards. Serious discussion about organization identifiers started in 2016 and 2017. There was a large group of stakeholders involved and it was very frustrating at times—progress was slow. Conversations, meetings, and working groups took place over a few years. We stuck with it and reached consensus on a way forward. Not everyone was happy—but most of those involved were.

A key aspect of ROR is that we decided not to create another new organization. People had started to feel "organizational fatigue" because it was not scalable to create new organizations for every new identifier. So, ROR is jointly managed by Crossref, DataCite, and the California Digital Library and supported by a very active community. The ROR registry launched in 2019 and just celebrated its fifth anniversary.

The principles of open scholarly infrastructure

A major turning point for Crossref was the adoption by the board of the Principles of Open Scholarly Infrastructure (POSI) in November 2020. This was a seminal moment for Crossref and was the culmination of many years of effort, planning, and discussions. Looking back, this is the most important development at Crossref since its founding and the early fees change in 2003.

Initially, POSI was written by Jennifer Lin, Cameron Neylon, and Geoffrey Bilder.[10] They distilled lessons learned from Crossref, ORCID, and other initiatives into a set of principles that now guide many open scholarly organizations and initiatives. There are sixteen principles in three areas: governance, sustainability, and insurance. The principles help build trust with the community, ensure sustainability of organizations and initiatives, and provide insurance if things go wrong, for example, open data, Open Source, so that services can be forked or restarted by others in the community. Organizations that adopt POSI do self-assessments against the principles to show where they meet them, where they do not, and what their plans are to work toward meeting all of them. So POSI provides a framework for building trust and ensuring the long-term viability of open scholarly infrastructure.

Crossref adopting POSI was part of Crossref's journey to providing fully open metadata. At its founding, Crossref metadata was minimal and only available to members and paying subscribers. Crossref's REST API,[11] launched in beta in 2012 in production 2014, marked a turning point by providing open metadata except for references for the first time. Following the adoption of POSI, Crossref opened all references in 2022, a move that I thought would never happen. There are now open references available for more than sixty-five million articles. Continuing this trend, Crossref acquired and opened the Retraction Watch[12] data and we continue to support its ongoing curation and updating.

Looking forward

We are still working toward fulfilling Crossref's vision of "a rich and reusable open network of relationships connecting research organizations, people, things, and actions; a scholarly record that the global community can build on forever, for the benefit of society." There are currently too many gaps in the metadata that is available and the relationships between grant funding, research outputs, authors, organizations, data, and software that make up the scholarly record. We need more integrations in many more systems and need to have much more metadata, in particular, identifiers for grants via the Crossref Grant Linking System, ROR IDs for affiliations, abstracts, corrections, and retractions. One important goal of all of this is to make things easier for researchers so that they do not have to re-enter metadata in different systems and can focus on research and teaching. An important 2022 editorial in *Nature* highlighted the value of open metadata and said, "Depositing all relevant metadata in Crossref should become the norm in scholarly publishing."[13]

Expanding the scholarly record to be more diverse, equitable, and inclusive is another critical challenge. We need to lower barriers to participation and change the scholarly record to reflect how research is changing globally. The two aspects of this are to lower barriers to participation in Crossref. In 2023, Crossref launched its Global Equitable Members (GEM) program[14] which waives all fees for members in the least economically advantaged countries in the world. More than four hundred members are currently benefiting from this program. We are also underway with a review of all Crossref fees with

the goal of making them more equitable. The other aspect of a more diverse scholarly record is the type of content that Crossref registers, so we are looking at adding Indigenous Knowledge to our services. The Global North doesn't have all the answers and has much to learn from other perspectives.

Research integrity and future challenges

To finish I will take a look at some other more general issues confronting scholarly research and communications. Research integrity is becoming an increasingly critical issue in scholarly communication. We face major challenges from paper mills, bad actors corrupting the editorial process, misuse of large language models (LLMs), and manipulated data and images. Addressing these issues requires collaborative action, and there are numerous initiatives underway, including those by COPE,[15] the Directory of Open Access Journals (DOAJ),[16] United2Act,[17] NISO CREC,[18] and PubPeer.[19]

Publishers put a lot of resources into research integrity—both humans and machines. One of the challenges for Crossref is that this means there are fewer resources available to improve metadata and establish relationships, so Crossref is working on doing more matching using Machine Learning (ML).

An interesting recent phenomenon is the rise of "science sleuths" or "data sleuths"—researchers conducting *ad hoc* investigations—often in their spare time—into research integrity issues such as image manipulation, fraudulent data, sneaked references, and citation cartels. The science sleuths use open metadata from Crossref and other sources to do their work highlighting another reason why open scholarly infrastructure is so important. This emerging discipline will likely grow, and there may be a need for formal training in this area.

A quick word on Artificial Intelligence (AI), specifically LLMs and chatbots. There are both opportunities and challenges. While we are currently at peak hype (remember things such as Second Life and Blockchain were going to revolutionize research), I believe many useful applications will emerge. Maybe not as separate applications, but built into many systems. AI can assist humans in various ways, from helping non-native speakers with translation to generating ideas and summarizing information. AI can also be used to fabricate data, images, and text very easily. In this environment, provenance and citation become even more critical, and maintaining trust is paramount. Metadata, identifiers, and relationships critical for proper provenance and citation—the integrity of the scholarly record is an important part of research integrity.

Lessons learned and future directions

So, to wrap things up, what are some of the things I have learned? Here are some things:

- Focus on solving specific problems using appropriate technology.
- Collaboration is essential for addressing collective challenges.
- Build trust by agreeing on principles upfront and establishing proper governance.
- Develop sustainable models that can evolve over time.
- Be prepared to compromise within the framework of agreed principles.
- Recognize the importance of fear as a motivator, but channel it constructively.
- Prioritize diversity, equity, and inclusion in all aspects of the scholarly infrastructure.
- Be clear about who the stakeholders are and what problem(s) you are trying to solve.
- Balance the concerns of stakeholders—build trust by agreeing on principles up front.
- Stay attuned to inflection points and be ready to act when opportunities arise.

Inflection points are easier to see in hindsight and not everything works, but stay curious and engaged in what is happening in the community and globally to spot trends and get input from a range of stakeholders.

As I reflect on my career and this award, I am reminded that progress in scholarly communication is a collective endeavor and does not happen quickly. The challenges we face—from ensuring the integrity of the scholarly record to making research more open and accessible—require continued collaboration, diplomacy, and innovation.

I want to emphasize that while identifiers and PIDs are important, they are not ends in themselves. We must focus on the broader ecosystem of metadata, relationships, and services that make scholarly communication more efficient, transparent, and impactful. By continuing to collaborate, innovate, and adapt, we can build a scholarly infrastructure that truly serves the global research community and society at large.

The journey of building open scholarly infrastructure is ongoing. It requires us to stay curious, engage with global trends, and remain committed to the larger goals of advancing human knowledge and improving the human condition. As we move

forward, let us remember that our collective efforts today shape the foundation upon which future generations of researchers will build.

In closing, a few hopes and dreams for open scholarly infrastructure in the future are to have closer collaboration and integration with other organizations and to have a global scholarly record open and accessible to all. Leaders in this field, including me, must strive for more diversity in the development and governance of open scholarly infrastructure.

Statements and declarations

Conflicting of interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Edward Pentz https://orcid.org/0000-0002-5993-8592

Notes

1. Noble, S. U. (2023), "Decolonizing standards," *Information Services & Use* (Vol. 43, Issues 3–4, pp. 327–333). IOS Press. https://doi.org/10.3233/isu-230214, accessed September 16, 2024.
2. Joseph, H. (2021). 2021 Miles Conrad Award Lecture: Heather Joseph. *Information Services & Use (*Vol. 41, Issues 1–2, pp. 81–90). IOS Press. https://doi.org/10.3233/isu-210116, accessed September 16, 2024.
3. United Nations. (1948). Universal Declaration of Human Rights https://www.un.org/en/about-us/universal-declaration-of-human-rights, accessed September 16, 2024.
4. https://sdgs.un.org/goals, accessed September 16, 2024.
5. Berners-Lee, T.J., 1989. Information management: A proposal (No. CERN-DD-89-001-OC). https://cds.cern.ch/record/369245/files/dd-89-001.pdf, accessed September 16, 2024.
6. "NCSA Mosaic." Wikipedia, The Free Encyclopedia, Wikimedia Foundation, 17 June 2024 last updated https://en.wikipedia.org/wiki/NCSA_Mosaic, accessed September 16, 2014.
7. Gold, J., (1996), "The BUBL Information Service," *The Serials Librarian*, 29(3–4), 165–174. https://doi.org/10.1300/J123v29n03_14, accessed September 16, 2024.
8. Atkins, H., Lyons, C., Ratner, H., Risher, C., Shillum, C., Sidman, D., & Stevens, A. (2000). "Reference Linking with DOI," *D-Lib Magazine* (Vol. 6, Issue 2). CNRI. https://doi.org/10.1045/february2000-risher, accessed September 16, 2024.
9. Varmus, H., (1999) Original Proposal for E-Biomed (Draft and Addendum). E-BIOMED: A Proposal for Electronic Publications in the Biomedical Sciences. https://collections.nlm.nih.gov/catalog/nlm:nlmuid-101584926X356-doc, accessed September 16, 2024.
10. Bilder G, Lin J, Neylon C (2020), The Principles of Open Scholarly Infrastructure, retrieved [date], https://doi.org/10.24343/C34W2H, accessed September 16, 2024.
11. See: https://www.crossref.org/documentation/retrieve-metadata/rest-api/, accessed September 16, 2024.
12. See: https://www.crossref.org/documentation/retrieve-metadata/retraction-watch/, accessed September 16, 2024.
13. "Citation data are now open, but that's far from enough." (2022) *Nature,* Vol. 609, Issue 7927, pp. 441–441. https://doi.org/10.1038/d41586-022-02915-1, accessed September 16, 2024.
14. See: https://www.crossref.org/gem/ accessed September 16, 2024.
15. See: https://publicationethics.org, accessed September 16, 2024.
16. See: https://doaj.org, accessed September 16, 2024.
17. See: https://publicationethics.org/about/press/paper-mills, accessed September 16, 2024.
18. See: https://www.niso.org/publications/rp-45-2024-crec, accessed September 16, 2024.
19. See: https://en.wikipedia.org/wiki/PubPeer, accessed September 16, 2024.