A Compass for the Scholarly Publishing Journey

Recorded at London Book Fair 2023
Research & Scholarly Publishing Forum

For podcast release
Monday, May 1, 2023

KENNEALLY: A watch is handy for meeting trains and catching planes, but a compass will get you where you're going. Like the Earth itself, scholarly publishing has a true north and a magnetic pole. The axis of our scientific world is firmly fixed on research. The pole is variable and in our time has shifted to open. The open publishing compass for knowing the direction of our travel is metadata, the stuff about data. This instrument uses persistent IDs to recognize authors, funders, and institutions – details that are necessary for managing workflow and that are critical in a pressurized publishing environment that includes funder mandates, data policies, and expectations for interoperability.

Dan Shanahan, publishing director at Public Library of Science, welcome. Two weeks ago, PLOS released another six months of data for a public dataset identifying and quantifying open science practices, including about preprints, data sharing, and code sharing. Where do the open science indicators tell us we are, and how good is the data?

SHANAHAN: How good is the data? It's for what it's intended to be done for. Damian earlier for eLife was saying they looked to transform science communication, and actually that's PLOS's goal as well. To do that, we need to understand if what we're doing works, and it's recognizing that the communities are different, the challenges they face are different, the barriers they face are different, and what matters to them is different. So it can't just be one size fits all. That's what the data is about. It's about understanding at a more granular level what challenges communities face and if our individual products, if our services, if our policies are working, or if they're perhaps not. For those purposes, the data is pretty good.

The data for our own journals – we have all of it, because we've got the end-to-end systems, and we're able to assess it and come up with quite robust data. So we can see in a longitudinal way if what we've tried has improved things or made it worse. Now, the comparator dataset is much lighter touch, because we don't have as much access. Therefore, I must emphasize the OSI – the Open Science Indicator, OSI – is not intended to be a benchmark. It's not intended to sit there and rate how good all the journals are in

terms of open science. It is about sitting there, going is what we're doing moving the needle? Is what we're doing going to advance it? And for that purpose, it is, for where we are, as good as it can be. But we are still hit by all the limitations that affect the industry as a whole. We can do it with the data we have, but we know that capturing information around authors, around affiliation, around funding is still very limited.

KENNEALLY: As a goal for PLOS and perhaps for the whole ecosystem, it's about moving beyond the article, and the metadata allows you to do that.

SHANAHAN: Our argument and our view is that open science practices allow you to really use it and take these articles and these data further. This is about understanding if we've set that road to make it possible. It doesn't as yet – you know, we haven't followed up and extended this into proper impact data. So the aim for it is the open science allows people to pick it up and use it in policy and use it in further research. We want to be able to identify how the impact of this data and these evidence and this research is picked up, is used, is implemented in practice, so that we can see beyond just the contents of the article to the wider real-world implications and actually moving society, moving what matters.

KENNEALLY: We heard earlier today about concerns around equity in publishing. Talk about how metadata can play a role in enabling greater equity within the system.

SHANAHAN: A lot of what we talk about when we talk about open access is still gold APCs and all of this, and we still have people talking about waivers, which is, in my opinion, a really inequitable way of doing it, because you're putting the emphasis on someone to come to you and say, please, sir, I'd like a free article, which is absolutely farcical. We know that the affordability varies by country. It varies by people's background. And we don't want to put all the onus and the emphasis on the author to go cap in hand.

Metadata enables us to identify this in a more systematic way. We know the affordability in different countries is different, and if we have good metadata, we can identify them up front, and you can vary the APC according to affordability. You can end up doing this like Research4Life without asking people. You can just say, no, it's free for you. There is no charge. Because that is more equitable. That is more about access. And it is enabling people to sit there and identify – you might have seen, actually, we've got a policy on parachute research, which is when you've got a list of authors, and it goes to one person, normally from America, with a list of authors from a different country. We can identify that up front and potentially have a check and check this is all equitable, this is all fair.

KENNEALLY: And it's a window into the whole world, not just a piece of the world. It's about breaking down silos, bringing together the research, the authors, the entire ecosystem.

SHANAHAN: Yes, very much so, because again, it's coming back to how you discover research. It's how you access it. It's how you discover aspects of research. And again, talking to the DEI aspect, it's about sitting there and being able to then contextualize it, understanding what part of it is coming from different areas – what is based on the context, what is sort of a more fundamental method-based region.

KENNEALLY: Dan Shanahan, thank you. I want to turn to Dr. José Salm. He's a professor at Santa Catarina State University in Brazil. Welcome. *Bom dia*. Your expertise, Dr. Salm, is in knowledge engineering. You've worked for a variety of government and science organizations – the Pan American Health Organization, the National Institutes of Health, and the National Science Foundation. Nearly 25 years ago in Brazil, Brazil began development of the Lattes platform, a data-management platform named for physicist César Lattes. Today, Lattes has over 8 million users in Brazil and other countries. You work on that platform to expand and deepen the use of metadata. Can you tell us about the ways that persistent identifiers tell you about the direction that you're going. How has that information changed the way science and research are conducted in Brazil?

We should perhaps explain for people that the system in Brazil is rather like in the US. It's a federal government with a variety of states. And the challenge really was to ensure that there was an equitable distribution of funds throughout the country, rather than resources always flowing in the same direction.

SALM: Exactly. In the end of the 1990s, we saw a lot of concentration of funding going to certain regions in Brazil, mainly São Paulo and Rio. What this platform helped to do was to show first this difference in research and researchers' profiles and exactly how people were doing good research and not receiving the fair amount of funding. So once this platform starts running, you see this funding map change.

Regarding PIDs, there's a lot of issues on quality data. Dan was talking about using metadata and the importance of this. What we had – we had these real strange situations where we had politicians lose their jobs because of this Lattes platform. Ministers of education cited that they did specific research or had their master's degree or PhD degrees, and the information was incorrect. So the science community came to the government and said this information is wrong, and this person is not who he or she says that they are regarding research. Two ministers of education had to step down because of this.

It's an open platform. Universities can extract data once they relay it to the funding agencies. And now we have – last year, we published what is called the Lattes Data Platform. That's a Dataverse open science initiative that's now connecting these 8 million researchers' profiles to open science and open data initiatives.

KENNEALLY: And it's important to point out, as I say, that this was begun in the late 1990s, a very different world – a different world in Brazil, a different world everywhere. Tell us about how it started out and what the state of metadata was at that point.

SALM: Well, in different government initiatives, we had like 10 different systems to collect data for the same purpose. So the discussion was we gathered 600 researchers from various fields and discussed what would be the best way to collect data. We then went to the universities, public and private, and we discussed how they could use this information. And we then start building – at that time, it was a Windows application, but it did interesting things like connect your research to others – and cite co-authors and tell you who are the researchers that you publish most, things like building a web page for the researcher. We're talking in the early 2000s, so this was something fairly new for them. After that, we started to then build some analyses, and using persistent identifiers, connect to other databases to validate information, mostly, and try to scale the data into a better quality-assurance model.

KENNEALLY: That's a system that began at a national level, but now we live in a global publishing ecosystem. So you've probably had to begin to map your identifiers and work with the various other standards that exist. Is that a challenge?

SALM: Yes, a huge challenge. We started to work with different groups connecting DOIs, connecting ORCID numbers, connecting different persistent identifiers, and we started to understand the differences of how people inform in these specific databases the information and how they relate to their curriculum and their past activities. So now, what we're building – and this is fairly new research – is building narratives, using large language models to build stories about the curriculum of that researcher. So not just adding metadata about that specific research project or paper, but helping to build a story on how that benefited his or her research and the role that the researcher had on that specific project. Those are the things that you don't usually see in a bio sketch or a short bio. For funding agencies and other organizations that do funding, this is important to understand the potential and the capacities that the lead researcher has in developing that project – how he works with teams or how she works with the research team.

KENNEALLY: We'll come back to the connection with AI and metadata in just a second, so thank you, Dr. Salm. Matt Cannon, head of open research with Taylor & Francis Group, welcome to you as well. Your colleague, Michaela Atherley, is unfortunately unable to join us today, and you've stepped in. I really appreciate it.

In 2022, you co-authored a paper for the *Learned Publishing* journal about the impact of introducing a data-sharing policy that was open and FAIR. FAIR is findable, accessible,

interoperable, and reusable. This was to apply to six T&F earth and environmental science journals. How does a FAIR data compass help researchers navigate on this journey to open science?

CANNON: Thanks. And thanks for the opportunity to be here and talk with you all. Yeah, the paper that we wrote was really T&F, kind of following what Dan was saying, trying to move the needle and trying to take some positive action to move things forward in kind of an open science, open research world. And the journals that we selected in earth and environmental science was kind of an obvious choice, given that it's more of a community initiative to have more open data in those subject areas and given the fact that climate change and Earth observation data is such a key part of measuring and evaluating that, and also because actually the risk in sharing that data is lower than in some other areas, because there's no personal GDPR, those sorts of things. There was already kind of a community movement towards having data sharing, open data tied to manuscripts published in those sorts of journals.

Taylor & Francis had had a data-sharing policy since 2018, but nearly all of our journals at that point were on the lightest touch, with encouragements to share rather than any other requirements. So we launched this pilot to move a selection of these journals to our most open policy, which has quite strict requirements around including data availability statements and having open datasets stored in a repository that sit behind that research.

And in terms of metadata, one of the things that we've had to do as a result of this project and this case study is really think about the links between the articles that we were publishing and the data that sits in those repositories and making sure that the references between those two objects are both useful for humans, who can read a paper and say, oh, yeah, there's a dataset in PANGAEA or in Dryad, and they can go and look there, but also that there's a machine-readable link that all the software and other tools in our ecosystem can use to make sure that there are these links and that it's kind of visible and discoverable. There's lots of benefits for that around increasing transparency, one of the main aims of open science, but also increasingly we're seeing lots of bad actors and paper mills and lots of stress on the publishing industry from that sort of thing. And increasing the transparency and the number of objects and the machine-readable links between them is a really good way of shining a light on that and tackling some of these things both from a research integrity angle, but also from an open research angle.

KENNEALLY: Dan Shanahan was talking about breaking down silos. Yet it seems whatever silos we're talking about, publishers always end up in the middle with all of this. For authors, talk about their perceptions around their role in this. They submit a manuscript, and they think they're done, and metadata be darned.

CANNON: Dan and I were talking about it as we came in. I think one of the things that maybe authors don't always understand is that the process of submitting a manuscript is – they're obviously putting forward their Word document or LaTeX file or whatever for peer review, but a lot of the kind of forms that they have to fill out and a lot of the information they're adding isn't just duplication. It's not just for fun that we're asking for these things. A lot of that is being used to build all the metadata and all the links and the funding information and the institutional information. All of that is being used to build that metadata that sits behind the paper. And adding the data-sharing requirement is another kind of metadata that we're asking for. We're also asking for the PID for the data that they've shared in a repository, hopefully, and making sure that we've got that in the right format that we can bake that into the metadata of the paper as well. But yeah, I'm not always sure how great a job we do of explaining that that is exactly what we're going to be using this for, because we still see plenty of errors and mistakes in some of those fields that end up being used for metadata.

KENNEALLY: To that point, on Tuesday this week, CCC launched The State of Scholarly Metadata: 2023, an interactive visual report that's based on a study of metadata management across the research lifecycle, drawn from interviews with dozens of industry stakeholders. The report depicts the economic and social impact of all these challenges, all these fragmented pieces in the supply chain for metadata. Dan, I want to ask you – Dan Shanahan with PLOS – you say the fault with data doesn't lie with the north star, the compass, but with ourselves. We're responsible for the challenges we have.

SHANAHAN: I would agree with what Matt said, that there's an element of collecting the data, but I also think our infrastructure is wholly inadequate for this. We've had this already. Publishing is now an online, digital world, and we still treat it as if you're sending your article in in a paper form. And it does lead to us having to request the same data multiple times. It does lead to the fact that the data doesn't transfer from one place to another. It does actually mean that we're recreating. And any time you're doing this, errors creep in.

This, for me, is the biggest challenge we've got, in that we do need better data, but it's not accurate to just point fingers at authors. It's not accurate to point fingers at people further down the line. As an industry and as everyone, we need to actually come to grips with this and recognize that it's not just the article that matters. The science is of utmost importance, but it's not the only thing – or the book or the what have you. We've got to actually come up with a way of really capturing all these data, making sure it's accurate, making sure it's accessible.

I would argue that a fundamental role of publishers going forward is going to be just that. When we talk about curation, I don't now think of it as just checking the article – is X good to be published? I mean making sure it is then discoverable, making sure the data is

there. That means going both backwards and forwards down the line to ensure that it's standardized, it's accurate, it's collected, and that everybody's on board and you're making it possible. It's too easy to delegate responsibility to the authors and say you got it wrong. But we didn't make it easy for them to get it right.

KENNEALLY: Matt Cannon, how precise does the data compass need to be? I understand there are new PIDs emerging all the time. Is the compass in danger of getting overcrowded?

CANNON: I think there's still plenty of room to add more PIDs and more granularity. I think we're a way away from being overcrowded at the moment. I think as the other speakers have said, we now have good clarity around authors via ORCID. FundRef provides really accurate funder metadata. But there's new PIDs emerging all the time. RAiD, the Research Activity ID, is getting a lot of activity and movement coming out of Australia, which is about having a PID for the whole project you're working on. You might use that to have all of the constituent parts. You might have your data, which would have also a PID in the code, and all the other objects you could all put into this kind of project ID.

I know there's also talk in some subject areas about having PIDs for instruments, so you can imagine large telescopes or other pieces of really specialized equipment having a PID, so you know exactly which pieces of equipment have been used to generate a certain dataset, for example, because that's going to be really useful when you're thinking about downstream uses of this stuff, like either checking – people trying to take a transparency angle and check, did they do what they say they did? But also from a reproducibility angle, like can I take this, can I reproduce it in my own lab? Do I need to use a specific instrument and then kind of build on that and accelerate, push forward, and make societal changes?

KENNEALLY: You used an important word, reproducibility. This is going to play a role in greater reproducibility for all of the work we do.

CANNON: Yeah, definitely. I realize it's not applicable for all subject areas. There's definitely arts and humanities areas where reproducibility isn't going to be an endgame. But I think there are plenty of areas where reproducibility in some way will be a factor and should be something that people are pushing towards. And I think we're seeing more appetite for that around people using new methodologies, like registered reports or other things, to kind of check and try and make use of what's already been shared and is available to drive progress on key topics.

KENNEALLY: Dr. José Salm, you've already brought up the subject that's being talked about everywhere, which is AI, of course. And you believe that metadata can play a role in

addressing one of the real challenges, which is when these chatbots hallucinate, when they give you answers that are extrapolations of the data that lead you in misleading or misinformed directions. Talk about the vision you have for the way that metadata can begin to address some of those challenges.

SALM: We're seeing different services now that use scientific literature and other references, that use AI and large language models. AI is a field, so there's various sub-areas. Specifically on language models that enable us to speak and to chat with the computer, we see some of these companies that they publish their reports showing that there are some very specific situations where they have encountered hallucinations. This is where the model predicts a chain of words, and this prediction is only based on their creation and their combination. They are not based on facts or references. These hallucinations are common.

What we talked about using metadata is what we see – and this is happening here in the UK. There are a group in Alan Turing Institute that they've done some work on this using metadata to retrain these large language models to make them more specific on their response, so they're more truthful to the reference that they use. And if you go on these services – it's called Science, or something like that, Elect – there's also some open-source services, nonprofit services, that index papers and let you question them. What they do is they use GPT-4 or LaMDA – so they use a large language model and retrain it, sometimes using a technique called reinforcement learning with human feedback, where they get some specialists to evaluate and retrain the model so the model response is in a more accurate and truthful manner.

This is something that I've been studying, because as I mentioned, we want to use this to build narratives. This links to project, and as you were saying, once the project ID infrastructure is set up, we can also index this and have it be trained, so people that are doing funding evaluation can look at the submission and the papers that are linked to that submission and have a glimpse of what that really means as intent and research methodology and in outcomes.

So I encourage who's listening here – if you haven't used these services on the web, they're very interesting and very scary sometimes, because you can ask them to compare the method used on that paper with similar papers. This will have a huge impact in academia in general and in publishing as well. We can go to the limit and say, OK, you can just talk to a book, right? You can say, what is the best sense of chapter three? What does this mean in real life? And then have this large language model response and not hallucinate – based on fact, give you a good response.

KENNEALLY: I'll point out that Douglas Adams, author of *The Hitchhiker's Guide to the Galaxy*, said that we are stuck with technology, when what we really want is just stuff that works. The future that Adams imagined of talking digital books and media is here. The next giant leap will be to train those books and journals to really listen to us.

I want to thank my panel today. Matt Cannon, head of open research with Taylor & Francis Group, thank you very much. Dr. José Salm, professor at Santa Catarina State University in Brazil, thank you very much. And Dan Shanahan, publishing director at Public Library of Science, PLOS, thank you very much.

I'll mention that CCC's interactive visual report, The State of Scholarly Metadata: 2023, is online at stateofmetadata.com. We invite you to have a look at it, interact with it, give us your feedback, let us know the problems and the challenges that you have around metadata. Again, that's at stateofmetadata.com. I'm Christopher Kenneally for CCC. Thank you very much.

(applause)

END OF FILE