The Data Quality Imperative

Recorded at Frankfurt Book Fair
Thursday, October 21, 2022

KENNEALLY: Welcome to the Data Quality Imperative – Improving the Scholarly Publishing Ecosystem. I'm Christopher Kenneally with Copyright Clearance Center, and I host our podcast series, Velocity of Content.

Thomas Redman, who goes by the Data Doc on Twitter, wrote in the *Harvard Business Review* that you can't do anything important in your company without high-quality data. Redman advises us to get data right from the get-go. This new approach and the changes needed to make it happen must be step one, he wrote, for any leader serious about cultivating a data-driven mindset across the company.

In publishing, open access is transforming the scholarly journal. In the laboratory and at the university, open science is remaking research. For this new open environment, best practices with data are those that strive to be efficient, transparent, and FAIR – that is, findable, accessible, interoperable, and reusable.

My guests this afternoon will share their insights on the impact that a strategic commitment to data quality can have on the people, processes, and technology in your publishing organization, as well as the entire research system, including authors, libraries, research institutions, and funders.

I want to introduce who we have to speak today. On the far end from me is Sybille Geisenheyner. Sybille, welcome. Sybille is director of open science strategy and licensing for the American Chemical Society, one of the world's largest scientific organizations, with membership of over 151,000 people in 140 countries.

In the center here is my colleague Laura Cox. Laura, welcome. Laura is senior director, publishing industry data, for CCC. In May, CCC acquired Ringgold, a provider of persistent organization identifiers widely used by the scholarly communications industry. Laura Cox is a board member and treasurer of ISNI, an ISO standard in use by numerous libraries, publishers, databases, and rights management organizations around the world.

And immediately to my right is Dr. Jo Havemann. Jo is the founder and coordinating director at AfricaArXiv, the African open access portal, with the mission to increase the discoverability of research achievements from and about Africa. Dr. Havemann holds a PhD in evolution and developmental biology and served in Nairobi, Kenya, with the UN

Environmental Program Climate Change Adaptation Unit. She's currently a trainer and consultant in open science communication and digital science project management for access to perspectives.

I'd like to turn first to Sybille Geisenheyner. When we're thinking about data in the scholarly publishing ecosystem, research is first, but it seems that data is a very close second, and it's much more a part of the publisher's job than ever before.

GEISENHEYNER: Yes. And I think the big question mark is what does data mean in general? You have research data, you have publishing data, and a lot of data in the house as well. I think at least from a publisher's perspective, some of us or many of us just realized here in the last couple of years how much data we actually have and what we can do with it and how we can approach, customize – if it is the researcher, the institution, or a funder – with the data we have in house.

KENNEALLY: And it's really about creating a more efficient, more effective workflow, isn't it? So all those different touch points along the way from the submission to the publication and after that.

GEISENHEYNER: Yes. As a scientific society, for us, the researcher is always at the center – the researcher as a reader, so as a consumer, or as an author, like a producer. It's really in the center of what we are doing. And to create workflows for the researcher to make it as easy as possible for him to either create content or consume content is really, yeah, simply a key thing to do.

KENNEALLY: In the world of transformative agreements, it's also about communication. You have to be able to work with the researcher as well, right? It's a two-way process.

GEISENHEYNER: Yes, absolutely. Those deals – and maybe some people have seen the presentation before about transformative agreements – those arrangements between institutions and publishers are getting more and more complex by adding a publishing element to a former reading or subscription agreement, and this is really complex in what we are doing. It starts by really – yeah, when a researcher submits an article to really see, is that researcher covered or affiliated with a read-and-publish institution, so to really give him the right direction right from the beginning of the process – this is your way to go. Or if they are not in an agreement, if they have a certain funder, for example, with a certain mandate applied, to really direct them through the right spot in the system – this is the way for you to be compliant with your mandate, and this is the right way to check out of the system when your article is accepted. For that, you really need big data resources.

KENNEALLY: And the obligation is on the publisher to collect all this data, but there's also an obligation for the researcher, too, to be sure from the get-go that they are delivering the right kind of information. It will have an impact on their work, but it will also have an impact on their audience.

GEISENHEYNER: Yes, absolutely. This is why you need standardized registries, like Ringgold, for example, where people know out of a list of institutions, this is mine. This is the one that I need to pick to be recognized as affiliated with that institution – and not free-flow and put something in which really ends up with a list of different name variations. So to have those implemented – those quality checkpoints – more or less right from the beginning is really key.

KENNEALLY: Right. So it's important to the researcher, to the reader, but it's also important to ACS, because you are watching this data. It's informing you about these agreements, how well they're working, whether they're achieving the objectives that they have. You need to have high-quality data at that side, too, because it's going to matter in negotiations and the way you plan for future agreements.

GEISENHEYNER: Yeah, absolutely. For example, the most recent deal we did in the second quarter of this year with the California consortias – it's three consortias, actually, combining there – is not just to capture the correct data. It's also triggering the right communication, because it is a complete different setting of having possible funder involvement and all that. And to combine all this, you need really to have systems who are capable and agile enough, more or less, to really have that fine tuning on the individual situation of an agreement, because this is definitely something we learned over the last couple of years – that all these deals, read or read-and-publish agreements, have a different flavor from one to the other. But this is something the researcher shouldn't worry about. They should have the same author experience coming through the system, knowing that in the background, people take care that they are really doing the right thing at the right moment.

KENNEALLY: To your point about the variety of transformative agreements, I'm fond of saying that transformative agreements are like marriages and snowflakes. There are no two alike.

GEISENHEYNER: Yeah.

KENNEALLY: So for ACS, that's a real challenge.

GEISENHEYNER: Yeah, not just for us. I think that's for everyone. And also from an institution perspective, they have 10 agreements with 10 publishers, and all of them have

different variations. So also that data to keep track of that – what am I allowed with this agreement, or what is the demand from that agreement? This is also from our perspective – from a publisher perspective – but as well from an institutional perspective key.

KENNEALLY: Sybille Geisenheyner, finally, about the real interest that we have more than ever to make the research ecosystem a global ecosystem – truly a global one – and be able to include researchers and communities from all parts of the world, does high-quality data play a role in ACS's interactions with the various regions of the planet?

GEISENHEYNER: Yeah, from what we can read out of the agreements we have in place, that is absolutely key, and we would want to include and not exclude. That's really, yeah, something we are looking for. That's not always possible, but that's definitely something we're trying to do. And to have, really, the funder involvement there much more on the front allows us to be much more global in a way, because funders act global in many ways, not just fund in a certain country. Research is a global endeavor, so you have a lot of different authors on a paper from very different institutions, from all parts of the globe, and to enable them to be compliant with all they ask for is really key. This is why we need those data. This is why we need those connections. Yeah, it's a constantly evolving environment.

KENNEALLY: Sybille Geisenheyner with ACS, thank you. I want to turn to my colleague Laura Cox at CCC. CCC has a longstanding interest in quality data. It's important to our rights licensing programs as well as RightsLink, our open access solution. So does the acquisition of Ringgold signal that CCC believes this issue of data quality is really more important than ever?

COX: Absolutely. We're moving into, as Sybille says, quite rightly, a much more complex environment as we gradually transition to open access and open science. So the issues around data – the quality of the data, the accuracy of the data, the ability to link data – become more and more important. So as we're moving into read-and-publish deals, as Sybille says, both sides of the negotiation need to have the right data. The publisher needs to know which components of an institution are included in the read part as existing licensees and which are publishing open access, possibly through an APC model, and who's paying for that APC? Because if it's the funder paying, then should it be part of the construct for the library? There'll be a conversation about that.

So actually understanding who the author is with a persistent ID like an ORCID or an ISNI, where they're affiliated in the hierarchy of an institution to know whether it's included in any transformative deal, with a Ringgold ID or a ROR or an (inaudible) is really important. And then we're tying all of that information together to produce the transformative deals that further this effort to move to open access, which everyone is

behind. But we also need some of the data to start being less burden on the researcher, because they're still having to input a lot of this information, and it's still difficult. And it really needs to start earlier in the process – in the funding process, in the grant application process – and for persistent IDs particularly to flow with that downstream and throughout the cycle, and then you can pull that to analytics.

KENNEALLY: So talk, Laura Cox, about the role that high-quality data can play in facilitating collaboration. Because as you say, the farther up the stream we move to really the very beginning of the process really means that relationships become, I would imagine, more fluid, right? So that's going to create opportunities for collaboration.

COX: Well, there's all sorts of opportunities in different settings as well. So it's collaborative, but it's also improvement of efficiencies. There's a European national funder who have integrated persistent IDs of all flavors into their national curriculum system, their CV system, their (inaudible) system, and then filter that down to the local institutions' (inaudible) systems, and it's saved 90,000 hours of administration across that country a year in research. That's pretty phenomenal. And then that data, if it filters through to the publisher, saves a lot of the pain points where we're experiencing data missing because of some kind of piece of workflow where it's not passed through or where something has been misused, put in the wrong place, and then data is having to be rebuilt. So we're having to add the persistent IDs back into the system, which really should just flow through continuously throughout the lifecycle.

KENNEALLY: And publishers with these various agreements, they're under pressure, because the funders are expecting compliance. They're expecting compliance from the authors, from the researchers. But that responsibility does sort of bleed into the publishing world.

COX: It does, and it's complicated, because funders have a variety of mandates, and they use different terminology and language to mean similar things in their mandates. We haven't got any standards set around how we communicate that through into systems to enable authors to make the right choices about the journal that they're selecting. Or in the cases where they have multiple funding sources, those funding mandates can actually be quite different and occasionally conflict. So actually passing your information into the user experience of the system to enable researchers to select journals, to ensure that they're following the mandate, and for all of that information to then go back to the funder to say, yeah, this is all where you wanted it, and the dataset is posted, and we have the protocols and all of the other pieces.

KENNEALLY: You mentioned the case of the European university consortium or organization and the achievement of saving 90,000 hours of what could be research time or time better spent otherwise at the university, but this also has a commercial benefit as well. Sybille's

mentioned some of that. But tell us a little bit more about the role that quality data plays in this ambition to have organizations be data-driven and to be making decisions around data.

COX: You've described to me, anyway, Chris, in the past as data is being a bit blurry, so we add metadata to it so we have descriptions of that data. And then when we add persistent IDs, we bring it into focus. Those persistent IDs are interoperable. That's the key word is that we're creating something that is a permanent piece of information that can be linked from one thing to another. So we can look at collaborations. How often does Author A collaborate with Author B? Which institutions are that, what are they working on, and who's funding them? This is all information that plays back into the system and drives – it enables decision-making rather than things being slightly obfuscated.

KENNEALLY: Finally, Laura Cox, the publishing ecosystem, the word I'm fond of using today, is less and less focused only on the journal or the journal article. It's interested in all sorts of parts of the research lifecycle. Does quality data matter there as well?

COX: Yes. I think it's all part of the same process, and we really need to start with – you know, we've got the grant applications. We've got the process we're going through. We're undertaking search or creating datasets. We have methodology where we're generating preprints. Then there's submission processes. There's access to those articles. There's usage. There's citation. And ultimately, analytics that help us understand impact and whether we're meeting diversity and inclusion goals or Sustainable Development Goals and look at those sorts of things that genuinely benefit society throughout the system.

KENNEALLY: That's an important point, isn't it – impact? That's what funders want to know. That's what the researchers want to know. Everybody involved really wants to know what difference they're making, right? Laura Cox, thank you.

I want to turn finally to Jo Havemann. Tell us about AfricaArXiv and particularly how data really matters in this objective you have to bring African research and African researchers to the world.

HAVEMANN: Thanks so much. I want to acknowledge also all my colleagues in Africa who cannot be here with us today – first and foremost, also, Joy Owango, who is a dear colleague and also a partner institution to AfricaArXiv, which is also our umbrella organization, the Training Centre for Communication in Nairobi, Kenya.

AfricaArXiv was founded in 2018 to increase discoverability of African research output, because what we see in the (inaudible) indexing databases is only what's discoverable through the persistent identifiers, and (inaudible) have been built in a Western context. That's historically explainable and nothing to criticize. History has played how it's played.

Now, our approach is, OK, what infrastructure exists that can also be utilized by researchers from around the world, because it has a global mandate, as you also said the Chemical Society has. So our idea was to use existing infrastructure to encourage African scholars to disseminate their work as preprints, post-prints, whatever research outputs they have – datasets – anything they can share. Have a DOI assigned, like the first persistent identifier. Then preferably also sign up with an ORCID ID, the other one, and also have an institutional affiliation identifier. Now, you've partnered with Ringgold. There's also ROR. There's some complementary aspects between the two. So the idea with – there's also other researcher identifiers other than ORCID, and they all have their reason for existence and can complement each other nicely.

So by using these identifiers and infrastructures and technology behind, we saw a chance, and that's also nicely playing out now in its fifth year, how we can boost almost overnight the discoverability of, in our case, African research output and thereby change the narrative from, oh, what we see what might only be less than 0.1%, but it's actually more, because a lot is still in print. A lot is not discoverable, because it's sitting on institutional repositories without internet connection, without the persistent identifiers and other standards to make it discoverable to the international databases and algorithms.

This is basically where we are plugging in to acknowledge and build on the existing infrastructure in the region to map. We've done a mapping exercise over three years which is continuously growing to map what research output is being archived – where in the world and where across the continent. How discoverable is it already to our well known and not-so well known discovery databases? So there's also a variety thereof – not only the usual suspect go-to references. We're informing about all of that, really, and talking to all kinds of stakeholders on the continent, internationally – Africa is international, with 54 countries – but also on a global scale to see where we can learn from other world regions. Latin America, there's a huge open access experience over many decades. They basically pivoted and piloted and pioneered the whole open access movement ever since, I think, the last 30 years or so or longer, and now also are kind of experiencing the same challenges of everywhere else.

But I think we're on a good pathway here. As I want to acknowledge, I appreciate what you said in terms of – it has to be simple and easy for the researcher, because they are there to do research and not have to boggle around with all kinds of technologies. Of course, they're also responsible to add the content-specific metadata, and that's what I'm also teaching with Access 2 Perspectives – data management, research project management, FAIR data, CARE data, also considering things like who's responsible? Who owns the data before, during, and after the project? Who's responsible for maintenance and to inform the dataset with the required metadata? But I should maybe take a break here.

KENNEALLY: Jo Havemann, we've been hearing from the others about the importance of impact. You tell that story, and I have to imagine that it's having a tremendous impact for the African researchers themselves, for the institutions. Tell us a bit on the ground. What's it like for them to have that resource to then become part of other resources, other databases?

HAVEMANN: Well, what I said in the past two minutes sounds really exciting. We thought, oh, this is easy to adopt as a concept. It turns out there's a lot of reservation when it comes to preprints, post-prints, sharing your research output anywhere else than a peer-reviewed publisher journal. That's an experience we're not only making across Africa, but the whole researcher community has that reservation. I can only tell you, as we also said before when we spoke, preprints are here to stay, and they're being established as part of the publishing workflow as we speak. The publishers are already on it, and the funders are already demanding it – increasingly so.

It's just that people like myself and Joy with the Training Centre – what's missing or what we need more of to inform also now the practicing researchers of the opportunities that preprint publishing and following preprint (inaudible), community-based peer review, and also editorial through journal publishing and layouting and typesetting and curation – I think that's the most important task for the journals and publishing industry – the curation of the methods of data that are being produced across all disciplines now. We've never seen so much, and that needs to be curated.

KENNEALLY: And I think maybe a good way to end is on that very point about – as we evolve, we are seeing many more relationships created. Interrelationships become more important. Interdisciplinary research becomes more important. Is that how you see it?

HAVEMANN: Totally. AfricaArXiv is interdisciplinary by its nature, because we chose to focus on the region, like a regional focus, and it enables interdisciplinary research. Also cross-regional. We're also multilingual by nature. We accept submissions from any language of the world. Also, thankfully, we've observed many publishers also adopt that concept to make it not only interdisciplinary, but also breaking the language barriers in many ways and very efficiently so. Much of that can be seen here in the second floor if you come and join us on the scientific publishing level. Many publishers and service providers are facilitating that. Yes, it's crucial for impact. When we talk about impact, there is unfortunate a misconception for a high-impact-factor journal –

KENNEALLY: The impact factor. Is that what you're talking about?

HAVEMANN: The impact factor for the journals is I think also not what you were referring to, but what societal research can research actually have? Like why are we doing research in

the first place? Yes, to acquire knowledge, but then what? We have urgent challenges to solve here. Researchers hold the answers, and we just need to make sense of the research that the researchers have generated – where we come back to the persistent identifiers and the curation and – yeah, so we're on a good path, I would say.

KENNEALLY: Well, Jo Havemann, thank you for that background on AfricaArXiv. We learned a lot, and I appreciate the contributions of my panel – Sybille Geisenheyner with the American Chemical Society, Laura Cox with CCC, Jo Havemann with AfricaArXiv and Access 2 Perspectives.

I appreciate you joining us today. I'll mention that a Nobel laureate once said that if you torture your data long enough, it'll tell you anything. But my advice is to treat your data well. It will return the favor. My name's Chris Kenneally with CCC. Thanks for being here.

(applause)

END OF FILE