For the scholarly communications community to extract the most value from scientific research, as well as to successfully move from subscription to open access publishing models, it is essential to have ‘clean’ metadata and a robust infrastructure to build the required workflows, processes and systems to support effective use of that metadata. The obstacles to achieving a supporting infrastructure and implementing effective workflows are many, negatively impacting every stakeholder group and every aspect of the research and publishing process.

**Keywords**

bibliographic metadata; metadata; open access; data quality

**Introduction**

For the scholarly communications community to extract the most value from scientific research, as well as to successfully move from subscription to open access publishing models, it is essential to have clean metadata and a robust infrastructure to build the required workflows, processes and systems to support effective use of that metadata. To support these efforts, Copyright Clearance Center (CCC) and Media Growth Strategies (MGS) created a visual representation of the metadata challenges that market stakeholders face, an infographic that is accessible to the entire industry to provide input and help update it as new initiatives and solutions emerge and are implemented. The analysis comprises a deep dive into the metadata management aspect of the ecosystem, focusing on challenges faced by each group of stakeholders throughout the research life cycle.

CCC has a uniquely broad perspective on the complexities as an intermediary supporting the industry, including in its move to open access, being able to identify where the value is being delivered across stakeholder groups and where the breakages are. A large body of research exists outlining and clearly supporting...
the need for robust metadata, persistent identifiers and both open and proprietary infrastructure – for example, the literature review of scholarly communications metadata,\textsuperscript{1} the subsequent article in \textit{Quantitative Science Studies}\textsuperscript{2} and the MoreBrains Cost Benefit Analysis commissioned by Jisc in early 2021,\textsuperscript{3} in addition to many other articles and blog posts. The research commissioned by CCC complements and builds on that work with the goal of advancing and enriching these types of initiatives around metadata currently under way.

**The project**

MGS and CCC reached out to more than 50 industry stakeholders: librarians, researchers, publishers, institutional repository managers, service providers, funders, grants managers, industry consultants and standards organizations. Interviews were conducted with dozens of stakeholders, see Figure 1, to get a better understanding of what they believe to be the biggest hurdles to achieving, and greatest benefits deriving from, consistent quality metadata implementation and deployment. The results of the interviews are laid out in a visual depiction\textsuperscript{4} of the various ‘swim lanes’ of stakeholder groups and include metadata management challenges faced along the scholarly communications path by each of those groups.

Figure 1. Stakeholder group representation

The stakeholder groups represented in the visual map include:

- researchers/authors
- institutional librarians, heads of sponsored programs, research offices
- funders
- publishers
- institutional repository managers, data curators.

The publishing cycle represented horizontally includes:

- **Idea development** of a research project, including collaboration and exchange of ideas
- **Preparation** of a project, including data management, technology required to support research, ethical considerations
- **Proposal submission** writing, including researching available grants and projects previously funded by relevant funders
• **Researching and authoring** the project, including data collection, capture and analysis, as well as submission to preprint servers in some cases

• **Publication** of research findings, including submitting the article, choosing the license, peer review, paying article processing charges (APCs) (if applicable) and indexing

• **Preservation** of research findings in repositories or article and data archives

• **Reuse and measurement** of the impact of the article and its content, including impact on the researcher, funder, publisher and institution.

**Key findings**

Interviews with representatives from each stakeholder group (in most cases more than one representative) revealed the following top-of-mind challenges facing the industry around metadata in those phases of the scientific research life cycle. The results of the interviews are available from the link in the data accessibility statement at the end of this article.

**Idea development**

• Under-utilization of Open Researcher and Contributor Identifiers (ORCIDs) by institutions and researchers, as well as a lack of accessibility in emerging economies.

• Hindered search and discovery due to inconsistency in identifying content users and enabling appropriate access to published research.

**Proposal submission**

• Inconsistent metadata capture across grant application processes/systems, resulting in possible loss of the metadata necessary to determine open access (OA) funding entitlements at a later stage, e.g. institutional affiliations, tax IDs, etc.

• Lack of interoperability between systems with which researchers interact (e.g. current research information system (CRIS) systems, grant management systems, curriculum management systems, ORCIDs, etc.), leaving room for gaps in metadata and persistent identifiers (PIDs).

• Low adoption of standardized PIDs due to limitations of legacy systems or lack of awareness, hindering conflict-of-interest management and later-stage funding, tracking and analysis of research output.

• Low-quality data resulting in confusion with grant numbers or proposal numbers versus grant identifiers (IDs), leading to inaccurate data entry during later stages, the lack of linking of grant IDs to particular research outputs and the inability to query funding and award IDs.

• Metadata gaps where datasets are edited for confidentiality purposes, leading to loss of metadata during the review and funding management process.

**Research and authoring**

• Poor metadata quality creates access barriers to research from under-represented researchers, including lack of or improper use of digital object identifiers (DOIs) (e.g. uniform resource locators (URLs) are often used in place of DOIs). There is inconsistency in identifying users and enabling appropriate access to research, as well as inequitable access to search and discovery services or certain content by contributors from under-represented areas.
• Difficulty managing references occurs due to integration with different citation tools using different PIDs.

• Variable quality of preprint metadata at submission, especially the use of PIDs versus free text fields.

• Risk of researchers appearing to not comply with funder mandates because of missing metadata.

Publication

• Too much manual data entry, eating up valuable resources or leading to missing or erroneous metadata.

• Missed funding opportunities due to:
  o under-utilization of metadata validation services
  o use of inaccurate or out-of-date researcher profiles (e.g. organizational affiliation)
  o inconsistency between journals’ policies and metadata procedures
  o lack of funding information captured at submission
  o multiplicity of standard identifiers with little interoperability.

• Difficulty flagging conflicts of interest due to inconsistent collection or provision of affiliation information and other metadata within a submission system, thereby creating issues with monitoring compliance with mandates and sanctions.

• Inconsistent and incomplete metadata leading to dropped or omitted data.

• Complex institutional organizational structures, complicating the determination of accurate affiliation for funding eligibility.

• Missing links to funded research leading to an inability to track contracts or co-operative agreements with government funders.

• Inability to link multiple grant awards.

• Inability to systematically provide information on funding requirements to help support authors comply with myriad and sometimes conflicting mandates.

• Variability of affiliations as they change during or after the research process, impacting publication rights retention by authors and uncertainty around applicable licenses.

• Funding entitlements negatively impacted by using email address for affiliation, use of abbreviations or nicknames when funder information is manually input and standardized name is not used, and changes in grant IDs between submission and publication.

• Loss of data influencing the outcome of what entity funds OA APCs.

• Interoperability issues caused by lack of connection between submission and production systems.

• Manual entry by publishers of PIDs prior to registering DOIs for more complete publication records, leading to errors and inconsistencies.

Preservation

• Institutions spending unbudgeted financial and human resources to educate researchers on how and why to create or utilize metadata.
Reuse and measurement

- Lack of consistent affiliation and funding data making modeling future agreements difficult for publishers and institutions.
- Datasets without DOIs, making them difficult to find, access and reproduce, which creates a barrier to open science.
- Difficulty tracking funder and research impact resulting from a lack of adoption of metadata standards.

Observations

The interviews that produced the above key findings also revealed noteworthy insights into the thinking around robust metadata in the various stakeholder groups.

From an institutional consortium:

‘Consortia have many read-and-publish contracts and because there isn’t a single definition of a good contract, every deal is different, and every process behind it is different, and behind the processes are the metadata. It’s very fragmented, with different publishers having different internal infrastructure and legacy systems. It makes it hard to have metadata standards in the industry because it takes time to develop them, but we need to speed things up … they can be related to what the author is experiencing in the workflow and with OA there are different challenges in the read-and-publish environment. It’s about article types, when does it start and end (the deal), there are caps, there are a lot of challenges within a single data point related to metadata in the system.’

From a funder:

‘From a metadata standpoint, the fact is that there are very few standards that people will adhere to, unless it’s something like PubMed or Crossref, in which case they’ll meet requirements but it’s only as good as what goes in and how it’s reused. If you aren’t going to reuse, you won’t adhere to standards. If the ecosystem sets a bar that’s too low, that’s what people will use. When you need to add additional information or say there’s a new piece of information we want to track on an article, that’s almost impossible to add in a consistent way. From a funder perspective, any funder or grant information included is very difficult. The list goes on. If it isn’t a traditional article package it isn’t considered to be managed as well.’

From a standards organization:

‘In the flip to OA there are several metadata problems. Starting with a general issue, which is that people presume metadata is always the same and it never changes. It changes. That’s one of the big problems people fail to understand in their metadata management. We have spent decades – probably eight or nine – developing a metadata structure for the order processing of subscriptions which now needs to get turned on its head in an author-pays model, for which we have no infrastructure. That is one key problem. Another is discovery and use by the end users. For example, what is vetted content? What is the difference between an author accepted manuscript (AAM) and a version of record (VOR)? What am I as a user reading? What is the vetting process it has gone through? One of the benefits of open content and an open content ecosystem is you can take it and move it anywhere else. How do you know what you’ve taken from where and its validity? This is core to the trust in scholarly communications and that can be lost without associated metadata. This continues in the workflow – if it was retracted, how do you know if it isn’t stated at the publisher’s site that it was, or if there’s no connectivity? So, you have the publishing process which has totally different …’
metadata, you have the discovery process of what you want to find and what you have found, and the third challenge around use, assessment, connectivity with other resources, to provide metrics by which you can measure the performance or success of outputs. If everything is diverse, you have no way to measure the end results. That’s an overarching landscape view.’

From an institution:

‘Part of the issue is many stakeholders don’t know what metadata is, so documenting things in general isn’t something researchers do a lot of. If you look at specialized roles like librarians, that awareness is much higher, but it’s mainly technical people that understand it while others have a low-level understanding (if any). At the meta level, the problem is the costs and benefits are mismatched. The person who has to create metadata isn’t the one who reaps the benefits of doing that. So, the researcher who wants to share data has to make it discoverable but they don’t get credit for creating metadata: people aren’t incentivized or rewarded for creating good quality metadata and when they do others benefit so it’s almost altruistic. And where there’s enough of an imperative, like with ORCID and Crossref, they are not always well-funded, which can be more problematic.’

From a standards provider:

‘The challenges in the flip to OA are in many ways the same as other metadata challenges: different recipients of the metadata have different requirements or options for what they can and should include so it might be different from data sent to OCLC or libraries, for example. Another is the vendors or technologies they depend on may or may not fully capture the information we want to include or should include, or do it in a way that makes the workflow easier instead of harder. Often when something new is introduced, there’s a change that manuscript submission systems have to accommodate, and factors might affect how quickly they can include that information and everyone has to update workflows. A challenge is keeping track of what they have, what they don’t have, and what they’re sending to different sources. In that sense, the metadata supply chain issues are very much the same, whether OA or otherwise.’

From a researcher:

‘There are many stakeholders. Understanding data is critical. Missing metadata is a huge challenge. I have to take time to track down the authors and ask them for information and then they have to ask whoever was in their lab, many of whom are long gone. And you may or may not get an answer.’

From a publisher:

‘The need for metadata is increasing all the time. Office of Science and Technology Policy (OSTP) guidelines have requirements around funding, there are diversity, equity, and inclusion (DEI) initiatives and the need to capture author and researcher information, where submissions come from, publications, etc. The challenge is ever increasing and to truly get an understanding of what’s going on across the board we need standards, and they need to be consistently applied.’

Summary

The issues brought to light in the interviews merely scratch the surface of the many challenges facing the industry in its quest to standardize and disseminate the use of robust metadata and the valuable linking of critical information. The benefits to the industry when
standards, PIDs and tools are implemented to support the reliable use of various types of metadata are unquestionable. Industry discussions and projects continue to move the development of those infrastructure elements forward slowly but surely, but progress will truly be made when industry stakeholders actually implement and enforce metadata standards. It is incumbent upon each stakeholder group to make incremental changes against the challenges outlined in this article, including implementation of standards, to achieve that progress. With that will come increased efficiency, effectiveness – and sustainability – in the dissemination of scientific research.

**Data accessibility statement**
The interview questions with representatives from each stakeholder group are available here: [https://figshare.com/s/b43b58371b55f9c80dea](https://figshare.com/s/b43b58371b55f9c80dea).

**Abbreviations and Acronyms**
A list of the abbreviations and acronyms used in this and other Insight articles can be accessed here – click on the URL below and then select the ‘full list of industry A&As’ link: [http://www.uksg.org/publications#aa](http://www.uksg.org/publications#aa).

**Competing interests**
The authors have declared no competing interests.

**References**