Metadata mapping and vocabulary: consistency for all in scholarly communication via the Metadata2020 initiative

Rachael Lammey, Head of Community Outreach, Crossref

Metadata 2020 is a collaboration of publishers, librarians, service providers and providers of platforms and tools, data publishers and repositories, researchers and funders. They have the shared mission of improving the richness, connectedness, reusability and openness of metadata for all research outputs.

Richer, accurate, interoperable metadata is certainly a worthwhile goal, and one that would benefit many communities within the research space. There are many separate stakeholders who are working towards improving their metadata, however many of these efforts remained unconnected. Metadata2020 was launched to try to bring the work that these groups were doing together, and to make metadata improvements a top priority in areas where it may have languished as a 'nice to have' in the past.

The initiative was launched in 2017. It was conceived by Ginny Hendricks from Crossref. However, Metadata2020 needed to pull from the expertise of lots of different groups, so work started by bringing on board interested parties from the different communities. These parties formed interest groups in September 2017, and met to define the core problems for each of their communities. Today, around 120 people are currently involved between the different groups, which have now formed cross-community collaborative projects to address some of the core issues needed for metadata improvements.

These projects include:

- metadata mapping
- making metadata vocabulary more consistent
- forming best practice statements
- creating metadata evaluation tools
- forming and communicating business cases to enrich metadata
- communicating the importance of metadata to researchers

These projects are being undertaken by individuals with metadata experience from across scholarly communications, by individuals acting collectively to make improvements. The key to the group work is that they have been formed for a limited period of time to complete the work, and will then be reassessed to move the outputs of the groups forward, be that either via publicising the work that the group has done to the wider community, driving adoption or putting certain communities together via workshops, webinars or other means yet to be decided!

To come at the project from a Crossref perspective, we know that there are issues with publisher metadata management. It is costly for them, a huge challenge to implement for previously published material, they work with vendor systems whose requirements they have limited control over, and the community lacks an effective metadata distribution model. Publishers end up sending different subsets of metadata to lots of different end-points rather than being able to use one standard output. They often find it difficult to make the case to authors as to why good metadata (and supplying comprehensive information up front) is valuable to them.

Similar issues occur for service providers, platforms and tools. Metadata creators make assumptions that don't travel with the data, there is a lack of consistency in and between metadata schema, records can disappear without transparency as to where or why, there are no community standards about metadata vocabulary, and finally, metadata can easily get out of date if updates aren't provided by the metadata creators.

But there are opportunities too – the possibility of collaborating to find a consistent metadata vocabulary, working together to improve system workflows, developing business cases for improved metadata, mapping metadata to improve interoperability between services, service providers and publishers. Crossref has also launched Participation Reports for publishers so that they can easily see what metadata they're providing to Crossref, as sometimes this can be obscured by relationships with intermediaries or sit within one specific department. For example, publishers may be collecting references and ORCID iDs, but ensuring that information is in the metadata they distribute can be key to downstream discoverability and efficiency gains for researchers.

Those are just two communities struggling with common issues that they need to work together to solve, with metadata issues ranging far beyond their immediate constituencies. Hence the idea of forming groups and tasking them with finding solutions based on their shared knowledge.

At the CILIP Cataloguing and Indexing Group (CIG) conference in September 2018, the focus was on surfacing some of the key developments made in the metadata mapping and metadata vocabulary projects.

The role of the Metadata Recommendations and Element Mappings group is to converge communities and publishers towards a shared set of recommended metadata concepts with related mappings between those recommended concepts and elements in important dialects. It is a challenge for the community that there are many different ways that metadata is created, vetted, used and distributed; and the complexity of this makes finding new efficiencies and systems implementation difficult. Most groups face interoperability challenges with systems and processes, and there are silos within organizations themselves, making communications challenging.

The group was tasked with exploring solutions such as; identifying concepts included in relevant community and publisher recommendations, identifying concepts shared across recommendations and mapping across schemas to identify inefficiencies, breakages, etc. which may be leading to interoperability problems.

The group's work thus far has involved compiling an index of metadata schemas that different parts of the industry use. This list stretches to 36 different schemas from Dublin Core to BIBFRAME through PubMed and KBART. It's useful to have such a list to look at the scale of the task, but the decision was made not to try to map elements across all 36 schema, but just across the most popular/frequently used ones. Then, within these schema, the group listed out 34 elements and looked at how these are represented across each of the schemas.

Things that this exercise highlighted are that there are of course differences in how elements are named e.g. authors/contributors/creators, license_ref/rights/permissions/legal constraints/copyright permissions. The term 'abstract' was pretty consistent across most schemas, but absent in some, and there are lots of gaps! It's useful for this information to be exposed so that schemas and schema creators can identify these pitfalls, see where information might be going missing and adopt existing terms if adding to their own schemas rather than creating new ones.

Another project has been tasked with Defining the Terms We Use About Metadata. In order to communicate effectively about anything, a common language must be acknowledged, tacitly or purposefully. In the metadata space, there is not agreement on what words like property, term, concept, schema, title refer to, and different groups, librarians, publishers, researchers all use different vocabulary. This project specifically aims to create a glossary of words associated with metadata, both for core concepts and disciplinary areas, which should help accuracy when it comes to passing metadata though different workflows.

The first thing the group worked on was to list current definitions to find the most common uses of those definitions across the board. They are also working on a core metadata glossary of terms e.g. "concept", "schema", "title".

The end goal is to evolve a consistent core metadata glossary to speak to different research fields. It's important to acknowledge that there are other activities that are similar to this, for example CASRAI, the W3C group dealing with linked data and DCMI, all of whom the project is engaging or hoping to engage with.

The final project highlighted in the meeting is around creating shared best practice and principles around using metadata across the scholarly communication cycle, in order to facilitate interoperability and easier exchange of information and data across the stakeholders in the process. At the moment, there is a lack of central core principles, best practice and guidance, so lots of different stakeholders are creating their own (which don't necessarily match up).

The project group working on this have been working to define core principles for metadata around scholarly communications, created and disseminated in easily digestible ways for different groups. This also includes looking at issues around metadata ownership and governance.

The list is not meant to be comprehensive. If you think a resource is missing from this list, please let us know! Email us at info@metadata2020.org

Name (BPs = Best Practices)	Year (if known)	Primary Audience	Source type	General or type-specific?	Note
CHORUS Publisher Implementation Guide	2016	Publishers	Community Group	Journals/Conference Proceedings; Detailed	Provides recommendations for the data publishers submit about funding information for publications supported by publicly funded research. 2018 revisions expected
Cornell University Library Repository Principles and Strategies Handbook	2018	Libraries/Repositories	Library	Repositories	An organized and accessible description of different types of metadata created and used in library and archival settings
Crossref Books Advisory Group BPs	2016	Publishers	Community Group	Books; Broad	Provides descriptions of what the Crossref metadata repository requires to provide access and linking between books and journals. Some 2018 revisions expected
Crossref: Best Practices for Depositing Funding Data	2015	Publishers	Community Group	General; Detailed	General information for providing data for FundRef, the part of CrossRef that is a repository of funding sources for published research
Crossref: Depositing Reuse License Information		Publishers	Community Group	Journals/Conference Proceedings; Detailed	Describes CrossRef's policy for creating URIs with license information about access and use of a journal article or book; provides text and data mining users with a clear way of determining what they are permitted to do with content identified by a CrossRef DOI (digital object identifier)
D-Lib article: OpenDOAR Repositories and Metadata Practices	2015	Libraries/Repositories	Article	General; Broad	A study of how academic institutions create metadata for institutional repositories and archival systems; the scope of this study is restricted to DOAR (Directory of Open Access Repositories)
DateONE BPs		Environmental Sciences Researchers	Community Group	Data; Detailed; Environmental sciences	DataONE Best Practices database provides individuals with recommendations on how to effectively work with their data through all stages of the data lifecycle.
Emory University's Core Metadata Guidelines	2014	Libraries/Repositories	Library	General	Emory University Library's Core metadata items that represent the 18 minimum items are required to support search and discovery.
Journal Article Versions (JAV): Recommendations of the NISO/ALPSP JAV Technical Working Group	2008	Publishers/Libraries	Standards Body	Journals/Conference Proceedings; Detailed	These NISO/ALPSP Journal Article Versions (JAV) Technical Working Group recommendations provide a simple, practical way of describing the versions of scholarly journal articles that typically appear online before, during, and after formal journal publication.

Existing best practice documentation, available at:
http://www.metadata2020.org/resources/metadata-best-practices/

In late 2018, Metadata2020 ran two workshops, one in New York and one in London with participants from both the interest groups and the project groups. These workshops focused on working towards building a metadata flow diagram, to look at where metadata 'falls out' of the system and all of the different places it travels to. Work was also done to sketch out the 'big benefits' of metadata so that this can be used to communicate to businesses, researchers and other parties why it's worth investing time and effort in improving how we work with this valuable information.

Finally, the groups worked on best practice, breaking down the parties who work with metadata into a number of roles, depending on what they do (creator, curator, custodian and consumer) and look at what good practice would be at each point e.g. creators should provide the best possible metadata for the use case, curators should have quality control, custodians should make the metadata openly available - where possible - and metadata consumers should show the provenance of where they got their metadata from, especially if they are integrating it into their own tools and services. Note, an organisation can play more than one of these roles. This output is being shared with the wider group for discussion, and will then be released to the community more widely for feedback.

Now that the workshops are complete, the outputs will be taken back to the project groups to see what progress has been made and what the next steps should be. This might take the form of outreach around the group outputs, refocusing the group tasks or adding groups or individuals in whose voices are missing. It's important to note that the groups all consist of people volunteering their time to the project - it won't run forever, and the aim of the groups is to work for a short period of time on set projects to produce outputs - but sometimes if people are short on time, things will take a bit longer to come together, and that's ok.

There are lots of ways you can get involved in Metadata2020 if you're interested. If you'd like to contribute to a project in your area of expertise (even just for a set period of time), you can email info@metadata2020.org for details. You can also help promote the initiative to the wider community through your organizations, word of mouth, and social media - Twitter, Facebook, LinkedIn, and at metadata2020.org. Get involved and help us improve the quality of metadata for research!

Biography

Rachael has worked at Crossref in various roles since 2012. She is a participant in the Metadata2020 Researcher Communications group and advocate of richer metadata. She can be contacted at: rlammey@crossref.org/@rachaellammey. This article is based on Rachael's presentation at the CILIP Cataloguing and Indexing Group (CIG) conference in Edinburgh, September 2018.

ORCID: https://orcid.org/0000-0001-5800-1434