Proceedings of the 2013 CAMP-4-DATA Workshop

Jane Greenberg
University of North Carolina, USA
janeg@email.unc.edu

Alex Ball
DCC/UKOLN Informatics, University of Bath, UK
a.ball@ukoln.ac.uk

Keith Jeffery
Keith G Jeffery Consultants, UK
keith.jeffery@keithgjefferyconsultants.co.uk

Jian Qin
Syracuse University, USA
jqin@syr.edu

Rebecca Koskela
DataONE DataNet, USA
rkoskela@unm.edu

1. Introduction and overview

The Cyberinfrastructure and Metadata Protocols for Data (CAMP-4-DATA) workshop was held on Friday, September 6, 2013 at the joint Dublin Core/iPRES conference in Lisbon, Portugal. A joint Dublin Core Science and Metadata (DC-SAM) and Research Data Alliance (RDA) Metadata Working Group/Interest Group undertaking, the workshop's central aims were to examine and discuss issues surrounding metadata standards—specifically infrastructure design challenges and opportunities, applications, and policies. The Data Observation Network for Earth (DataONE) DataNet, the Metadata Research Center (MRC), University of North Carolina at Chapel Hill, the UK Digital Curation Centre (DCC), and the euroCRIS organization were also chief supporters of the workshop.

Factors motivating the workshop included

1. increased attention toward open data interconnected with national and international data sharing policies;
2. a range of efforts for registering or listing metadata standards, and particular interest in the DCC Disciplinary Directory and its applicability for metadata stakeholders;
3. the World Wide Web Consortium's (W3C) and Schema.org's attention toward data infrastructure and metadata for data access via the web (e.g., DCAT, DataCube, and VoID and Schema.org's data profile); and
4. the anticipated positive synergy from bringing together members of the DC-SAM, RDA, and iPRES community—all of whom have with overlapping and complementary interests.

The workshop was organized around presentations, discussion, and a series of breakout group activities. The day had an extremely positive and enthusiastic tenor, drawing 28 participants, spanning four continents, and representing 13 countries (China, Georgia, France, Germany, Italy, Luxembourg, the Netherlands, Portugal, Singapore, South Korea, Switzerland, United Kingdom, United States). Participants were mainly from informatics disciplines and working in metadata and data curation areas, and connected to specific data initiative or repositories, research centers, digital libraries, or university library systems. A handful of participants were also disciplinary scientists associated with a specific domain (e.g., geology, ecology, and so forth), bringing forward additional, valuable perspectives.

2. Presentations

The first half of the workshop consisted of a series of presentations. Following the initial introduction were two invited, foundational presentations that set the stage for the remainder of

the day. The next ten presentations had been selected based on submissions to the workshop. These submissions were a mixture of

- short papers that clarified current challenges, reported on research, or proposed solutions;
- position statements intended to provoke discussion; and
- abstracts that described metadata tools and technologies.

One final presentation was delivered as a lightning talk during the discussion session.

The full list of presentations follows. The asterisk symbol (*) identifies workshop presenters for collaborative works. The slides accompanying the presentations may be downloaded from the website of the International Conference on Dublin Core and Metadata Applications 2013;[1] abstracts and author biographies can be found on the Dublin Core Science and Metadata (DC-SAM) Community wiki.[2]

Section 1: Introduction and Foundations

1. Introduction and logistics (Jane Greenberg)
2. "The Metadata Zoo" (Rebecca Koskela)
3. "DCC Scheme Directory"(Alex Ball)

Section 2: Infrastructure Models and Frameworks

4. "A 3-Layer Model for Metadata" (*Keith Jeffery, Anne Asserson, Nikos Houssos and Brigitte Joerg)
5. "Cross-Domain Metadata Interoperability: Lessons Learnt in INSPIRE" (*Andrea Perego, presented on collaborative work with Michael Lutz, Max Craglia and Silvia Dalla Costa)

Section 3: Usage and Tracking

6. "Usage data for metadata properties to support open data registries and semantic wikis" (*Muriel Foulonneau, Sébastien Martin, Jacques Ducloy, Thierry Daunois and Slim Turki)
7. "Provenance Central: More Mileage from Provenance Metadata" (Bertram Ludaescher and *Paolo Missier)

Section 5: PIDs (Persistent Identifiers)

8. "Persistent Identifiers for Terms in a Crowd-Sourced Vocabulary" (John Kunze, Greg Janee, Christopher Patton)
9. "Separation of Concerns: PID Information Types and Domain Metadata" (Tobias Weigel and Timothy Dilauro)

Section 6: Applications

10. "Ontology-Enabled Metadata Schema Generator: The Design Approach" (*Jian Qin, Xiaozhong Liu, and Miao Chen)
11. "Metadictionary: Advocating for a Community-driven Metadata Vocabulary Application" (*Jane Greenberg, Angela Murillo, John Kunze, Sarah Callahan, Robert Guralnick, Greg Janee, Nassib Nassar, Christopher Patton, and Karthik Ram.)

---

[1] CAMP-4-DATA page on the DC-2013 website: http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/184

[2] CAMP-4-Data page on the DC-SAM wiki: http://wiki.dublincore.org/index.php/DC_2013_SAM_Science_and_Metadata_CAMP_4_DATA_AGENDA

12. "CLEPSYDRA Data Aggregation and Enrichment Framework" (Cezary Mazurek, *Marcin Mielnicki, Aleksandra Nowak, Krzysztof Sielski, Maciej Stroinski, Marcin Werla and Jan Węglarz)
13. "Useful Data: Metadata For Context and Reuse in a Multipurpose, multidisciplinary Repository" (Grace Agnew and Mary Beth Weber)
14. "OpenAIREplus: supporting interoperability through guidelines" (Najla Rettberg, *Pedro Príncipe, and Eloy Rodrigues)

Questions, Provocations, and Consensus

The second half of the workshop engaged all participants in discussion around three theme areas specific to metadata for scientific data:

1. amount of metadata and number of metadata standards;
2. necessity and feasibility of universal metadata approaches;
3. sustainability of a metadata directory.

Area three was a chief motivator for the CAMP-4-DATA, given the objectives of the Research Data Alliance Metadata Standards Directory Working Group and success of the DCC Disciplinary Directory of Metadata Standards. Participants divided into two groups, with group A starting with theme area one, and working toward area three; and group B starting with area three, and working in reverse. Following two periods of discussion, each group brought their observations and ideas to a general, open discussion.

Below we list the original guiding questions/provocations and a synthesis of the discussion that resulted.

Amount of metadata and number of metadata standards

- *Too much metadata: Should we be creating metadata proactively? Is it wise or foolish to be generating metadata via the "what if" scenario? Is there too much metadata?*
- *Schema [standard] overkill: Is there schema overkill? If so, who are the culprits? How can or should we harness the energy of those who want to generate yet another scheme? Should these efforts/energies be directed to other more crucial tasks and priorities?*

The discussion focused on the second of these two issues, that is, the number of metadata standards. It was not one that elicited a straightforward answer. While the consensus was that there are indeed too many standards that did not mean that the standards in existence satisfied all requirements. The challenge, it was felt, is two-fold: to de-duplicate effort across existing standards, while finding a way to satisfy more metadata requirements using existing standards.

On the latter point, application profiles were identified as a key technique. Not only do they allow elements from several standards to be mixed together as needed, they also allow generic metadata elements to be specialized to satisfy highly specific domain needs. Indeed, one breakout group felt that as a general rule, no new metadata elements should be created unless they explicitly specialize an existing element.

When discussing the former point, participants decided that mapping between metadata standards and between data structures could reduce the complexity of the space. Participants noted that standards with a larger number of mappings tend to be used more. Data mapping was seen as the more difficult of the two types, as data structures are more tightly coupled to applications. Even though equivalences can be codified using Resource Description Framework (RDF) relations such as *owl:sameAs*, it is not always possible to make robust inferences using them due to subtle differences in semantics and syntax.

A further difficulty with mapping lies in the scale of the task. In order to direct effort more efficiently, it was suggested that we should concentrate on providing partial mappings involving

the highest priority elements in the first instance, and incrementally improve them over time in response to demand.

Lastly, participants considered how to handle the versioning of data and metadata, and the place of persistent identifiers in creating robust links to particular datasets.

Universal schemes/models

- *Is it necessary, feasible or desirable to have a universal metadata format for accessing research datasets?*
- *Is there a threshold or limitation for when a universal scheme is possibly desirable?*
- *How can we [participants], as informaticians and scientists, collaborate and determine universal scheme desirability or other via related needs?*
- *How plausible is it for an ontological approach?*

These questions proved to be the most divisive. One breakout group decided that a universal scheme would be neither feasible nor desirable. The other, meanwhile, enthusiastically explored ways in which universality of a sort might be achieved, drawing on the efforts presented in the morning session.

Participants recognized that it would not be possible to enforce use of a single scheme, so a modular approach would be needed, such as that employed by CERIF (as described in Jeffery, et al.'s presentation). CERIF's use of layers—a general top level, with richer and more granular lower levels—was felt to be particularly useful. The generic layer of metadata is good for serendipitous discovery, the richer levels for evaluating the data's suitability for a particular purpose.

Suggestions for advancing toward universality included

- supporting approaches similar to natural language, where all metadata properties are included in one 'pot' (such as the Metadictionary approach, reported on by both Kunze, et al., and Greenberg, et al.);
- developing nano-standards for terms to advance richer semantic meaning;
- using links as the *de facto* way to relate terms from different schema, and using ontologies to define those links;
- developing a sense of trust in and democratic ownership of standards (participants recognized that people create new standards because they cannot control the existing ones), and
- providing a central authority or directory to manage standards (as discussed in Ball's presentation).

Enthusiasm aside, participants unanimously agreed that the available funding would not be sufficient to realize the perceived solutions.

Metadata Directory Sustainability

- *How can we motivate scientists, curators, and others engaged in the metadata process to participate in an open directory project?*
- *Are there models of successful, open and participatory directories?*
- *Are there models of successful, collaborative ownership that might inform the design of an open directory?*

Both breakout groups emphasized usability as an important factor in sustainability. There was agreement that low entry barriers facilitate access and participation. Metadata applications supporting clear, navigable workflows will, at the very least, not dissuade a user and may potentially aid motivation. One example given of this was the EuroGEOSS Broker, which provides a unified set of services over several data providers without requiring them to implement

any particular interoperability technologies or make significant changes to their tools and standards.

A major use case of a metadata standards directory would be researchers looking to see how to document their data properly. Their motivation for doing this might come from several sources.

One possibility would be institutional officers providing encouragement. One participant provided anecdotal evidence of consultants, employed to perform metadata work on a project, who ensured researchers provided the required metadata by working closely (and persistently) with them.

Another might be funders requiring researchers to provide high quality metadata with their shared data. Participants felt, though, that researchers would respond to this at best half-heartedly. In order to get high quality metadata, researchers must themselves want to provide it.

If researchers could see tangible benefits to providing good quality metadata, increasing as the quality improves, this would provide a strong motivation. It was noted that website authors have been keen to use Schema.org metadata because doing so improves how their sites are displayed in search engine results. A similar effect would result to making deficiencies in metadata more visible.

It was also argued that researchers would respond to evidence (perhaps from alt metrics) that well-documented data is more likely to be reused, if it was coupled with, for example,

- a reward system for funding and tenure that recognizes the impact of shared data;
- lists of the most used, most downloaded or most trusted datasets, on which the researcher might aspire to appear.

EBay and Amazon were identified as good models for mass participation, crowdsourcing, and trust. On the other hand, the Unified Digital Format Registry (UDFR)—a project funded by the Library of Congress (LC) that was developed for specific, internal needs—was felt to be a less than ideal model, because it was not adaptable to communities beyond the LC. Participants felt that Data.gov was worth looking at, although it is not yet clear how successful it is. Questions were raised about how to deal with participants voted down in a crowd-sourced or community-driven environment.

Participants agreed there is a need for metadata training. Education should address the value of metadata, and this could in turn raise metadata awareness.

Open discussion

The period of open discussion began with a set of clarifications about the RDA metadata work, in particular the relationship between the Metadata Standards Directory Working Group (MSDWG), the Metadata Interest Group, and the other groups with a strong interest in metadata.

It was explained that RDA Working Groups are expected to have a defined task that they complete in a period of 12–18 months. The Interests Groups, meanwhile, are intended as longer-term forums for discussing particular issues. The MSDWG could be seen as an offshoot of the Interest Group, dealing with the specific issue of improving access to and the discoverability of metadata standards. Other Working Groups could emerge from the Metadata Interest Group, but that might depend on the success of the MSDWG.

In RDA, no one appears to be pushing for a single metadata model. Indeed, there are a variety of groups interested in metadata—for example, the Data in Context Interest Group, and disciplinary groups such as the Agricultural Data Interoperability Interest Group and Wheat Data Interoperability Working Group—and it is likely that new metadata standards might emerge from their work. This reinforces the need for a metadata standards directory.

Discussion then turned to matters of interoperability. Much work has already been done to bring together discipline-specific data into unified systems: examples include NERC's Data Catalogue Service, which cross-searches various environmental data centers in the UK;

DataONE, which provides a unified infrastructure for earth observation data in the US and elsewhere; the EarthCube system for sharing US geoscience data; and the European Plate Observing System. These systems take different approaches. In US, systems tend to perform pairwise mapping, while European systems tend to map to a common core metadata catalog using standards such as CERIF. The CODATA World Data System has been interoperating for decades using community-agreed metadata standards, but the component repositories only interoperate within their own communities.

Following the lightning talk by Rettberg, Príncipe, and Rodrigues on OpenAIREplus, discussion turned to interoperability beyond disciplinary boundaries. One approach might be to take inspiration from Dublin Core and agree on a base set of elements that could be applied to any data set. This resonated with the idea of typed information associated with persistent identifiers, presented earlier in the day by Weigel and Dilauro.

It was noted that taking metadata elements out of context can be dangerous. Context plays an important part when mapping between metadata schemes, the process of which can be summarized as follows:

1. Ensure that the terms in the schemes are defined in a machine-interpretable way, taking into account the semantic context and the links to other terms.
2. Find matching terms in the two schemes.
3. Create a mapping between the matching terms, being careful to distinguish exact equivalence and cases where transformations need to be made.
4. Generate a converter.

The SeaIce Metadictionary[3] (as reported on by both Kunze, et al., and Greenberg, et al.) takes a novel approach to preserving context, which many of the participants found appealing.

Conclusions

Jane Greenberg concluded the workshop with a brief summation of the key points. One of the conclusions she drew was that, when dealing with metadata, it is important to keep focused on the objectives one is aiming to meet. She underscored that while we aim to maintain as clean a separation as possible between aspects of the metadata, we need to be cognizant of the fact that a single metadata property can be multi-functional, thus serving a series of purposes.

Among participants there was broad approval of the opportunities presented by the workshop discussions. There was an overall consensus that a future CAMP-4-DATA would be a worthwhile undertaking.

[3] The SeaIce project is also known as YAMZ (Yet Another Metadata Zoo). This change was instituted following the CAMP-4-DATA.