IDCC14 | *Practice Paper*

Building a disciplinary metadata standards directory

Alex Ball
DCC/UKOLN Informatics
University of Bath

Sean Chen
School of Law
Duke University

Jane Greenberg, Cristina Perez
University of North Carolina, Chapel Hill

Keith Jeffery
EuroCRIS

Rebecca Koskela
University of New Mexico

Abstract

The Research Data Alliance (RDA) Metadata Standards Directory Working Group (MSDWG) is building a directory of descriptive, discipline-specific metadata standards. The purpose of the directory is to promote the discovery, access and use of such standards, thereby improving the state of research data interoperability and reducing duplicative standards development work.

This work builds upon the UK Digital Curation Centre's Disciplinary Metadata Catalogue, a resource created with much the same aim in mind. The first stage of the MSDWG's work was to update and extend the information contained in the catalogue. In the current, second stage, a new platform is being developed in order to extend the functionality of the directory beyond that of the catalogue, and to make it easier to maintain and sustain. Future work will include making the directory more amenable to use by automated tools.

Submitted 24 October 2013

Correspondence should be addressed to Alex Ball, UKOLN Informatics, University of Bath, Claverton Down, Bath BA2 7AY, UK. Email: a.ball@ukoln.ac.uk

The 9th International Digital Curation Conference will take place on 24–27 February 2014 in San Francisco. Please ensure you use the guidance in this template to produce your paper. Please submit your paper in one of the following formats: Microsoft Word (.doc, .docx), Open Document Format (.odt) or Rich Text (.rtf). http://www.dcc.ac.uk/events/idcc14/submissions

Introduction

There are many barriers that need to be overcome in order for the full benefit of data sharing to be realized. The Research Data Alliance (RDA) – an international initiative supported by the European Commission and the US and Australian governments – aims to break down these barriers and thereby develop a global data infrastructure (Parsons, 2013; Showstack, 2012). The RDA focuses on bottom-up, collaborative activity; its Working Groups, for example, are proposed by researchers themselves and aim to implement tools, standards or best practices across multiple institutions in the space of 12 to 18 months.

The Metadata Standards Directory Working Group (MSDWG) was set up in 2013 with the aim of implementing a protoype wiki-based directory of metadata standards relevant to research data (Greenberg, Jeffery, & Koskela, 2013). The idea of listing metadata standards was not a new one; several lists and directories had already been compiled, including the following:

- Science Data Literacy Project list of metadata standards (Qin, Small, & D'Ignazio, 2008);
- *Seeing standards: A visualization of the metadata universe* (Riley & Becker, 2010);
- BioSharing's list of metadata standards;[1]
- the Global Earth Observation System of Systems Standards and Interoperability Registry;[2]
- the Marine Metadata Interoperability Project list of references to content standards.[3]

None of these, however, had all the qualities desired for the Metadata Standards Directory: some were static and unable to be curated by the community, while the others concentrated on a particular group of disciplines.

In parallel with the establishment of the MSDWG, the UK Digital Curation Centre (DCC) had independently developed its own Disciplinary Metadata Catalogue;[4] this was launched in January 2013. The MSDWG evaluated the resource and found that it aligned closely with its own ideals. It therefore entered into a collaboration with the DCC, using the Disciplinary Metadata Catalogue as a starting point for the RDA Metadata Standards Directory.

In the following sections we will explore the motivation behind the MSDWG and DCC efforts, and report on how the Disciplinary Metadata Catalogue was first developed by the DCC and subsequently expanded by the MSDWG. We will then discuss how this work will be taken forward, and the Working Group's future plans for the Metadata Standards Directory.

---

1 BioSharing Standards: http://biosharing.org/standards

2 GEOSS Standards and Interoperability Registry: http://seabass.ieee.org/groups/geoss/

3 Marine Metadata Interoperability Content Standard References: https://marinemetadata.org/conventions/content-standards

4 DCC Disciplinary Metadata: http://www.dcc.ac.uk/resources/metadata-standards

Motivation

The common use of standards alleviates many difficulties one might encounter when sharing data. Standard protocols allow different systems to communicate, while standard file formats allow different software to work with the same files. Standard metadata allows data to be processed, searched, preserved, recombined and reused across many different contexts.

It is worth stressing, though, that these benefits only come about when multiple parties adopt the same standard. They cannot be realised where no standards exist, nor where there are so many standards that none can achieve universal adoption (Willis, Greenberg, & White, 2012). As Tanenbaum (1998, p. 254) points out, 'the nice thing about standards is that you have so many to choose from,' and indeed there has been a proliferation of discipline-specific metadata standards in some areas. The greatest problems occur, though, where standards compete directly and in the case of metadata standards this is rarer than might first be apparent.

Partly this is because metadata is often specific to a particular purpose. Some standards are designed to support discovery services such as search engines or directories. Some are designed to support preservation activities, others to support packaging and transmission. Still others provide the contextual metadata needed to support a full range of administrative tasks, and clarify how a resource may be used. But perhaps the greatest variety exists among standards aimed at making data reusable in highly specific contexts, such as microarray experiments or materials testing.

Another way metadata standards avoid direct competition with one another is by eschewing independence in favour of a more linked approach. They may borrow elements wholesale from other standards; they may reuse elements with a more restricted choice of encoding or vocabulary, perhaps with narrower semantics; or they may define new elements that are explicit specializations of existing ones. Such approaches allow one to develop metadata profiles that are highly specific to one application while remaining intelligible to a wide range of others (Heery & Patel, 2000).

This should not encourage complacency, because the proliferation of incompatible standards is always a danger, and this is a concern even when direct competition is not involved. While the applications to which metadata standards are tuned will have a different overall character, there are often points of correspondence where the same or similar metadata techniques or elements could be applied. This would both save development and maintenance effort and provide a 'bridge' should metadata records using the respective standards ever need to be merged, perhaps in the course of interdisciplinary research.

The technique of producing application profiles has mitigated one of the drivers for the proliferation of new and ad hoc standards, and for duplicative standards work: that of existing standards not being quite suited to a given specific context. The largest remaining driver is ignorance that suitable (or partly suitable) metadata solutions already exist. By perceiving a gap that is not really there, potential standards developers are distracted from either engaging with relevant standards or tackling the genuine gaps that remain.

Today, many researchers are encountering metadata issues for the first time due to incoming data management plan requirements, and there is a wave of higher education institutions setting up new generalist data repositories. If the datasets populating these repositories are to be properly documented, it is key that researchers and data librarians alike are fully aware of the metadata standards that can be employed for the task.

DCC Disciplinary Metadata Catalogue

The DCC Disciplinary Metadata Catalogue was conceived as a resource that institutional data curators could consult when advising researchers on how they should document their data.[5] The initial proposal for the catalogue suggested that such curators would first want to know what standards are in use within the discipline in question. If there were none or very few, they might want to know about broader standards that could be adapted. For any given standard, they might be interested in

- the specification for the metadata standard;
- vocabularies or taxonomies commonly used in conjunction with the standard;
- any profiles that tailor the standard to a particular application context;
- any tools that are available for working with the standard;
- any examples of the standard being used by repositories or data portals: these might be useful as sources of practical advice on using the standard, and also indicate the level of adoption among researchers in the area.

As the effort available to develop the catalogue was limited, the catalogue had to be given a tight scope. The selection policy for standards to include was therefore kept narrow:

- Standards that define what information to collect about data were included, while standards that only specify how to structure, serialize or transmit data or metadata were excluded.
- Standards for detailed, descriptive metadata were included, while standards that focus on administration, preservation or the wider context were excluded.
- Standards for documenting tabular data were included, while standards describing publications, learning objects, audiovisual files or narrative text (e.g. interview transcripts) were excluded.

The scope was not restricted by discipline, though, as the aim was to help data curators to support as wide a range of researchers as possible.

The catalogue was developed on this basis by research consultant Liz Bedford over the course of seven months and published in January 2013. The standards and repositories chosen for inclusion were predominantly drawn from existing publications (Ball, 2009; Riley & Becker, 2010) and web resources.[6,7,8]

Figure 1 shows an example catalogue entry. It begins with a short description of the standard, indicating how it is intended to be used and something of its provenance. This is followed by a table of links to key resources associated with the standard, such as

- mappings from that standard to other metadata standards;
- vocabularies that could or should be used with it;
- the specification for the standard, and its website or home page.

---

5 The proposal for the catalogue was written by Liz Bedford in May 2012. Ball (2013) provides further details.

6 Application Profiles Support Project: http://www.ukoln.ac.uk/projects/ap/

7 Databib: http://databib.org/

8 DCC DIFFUSE Standards Frameworks: http://www.dcc.ac.uk/resources/standards/diffuse/

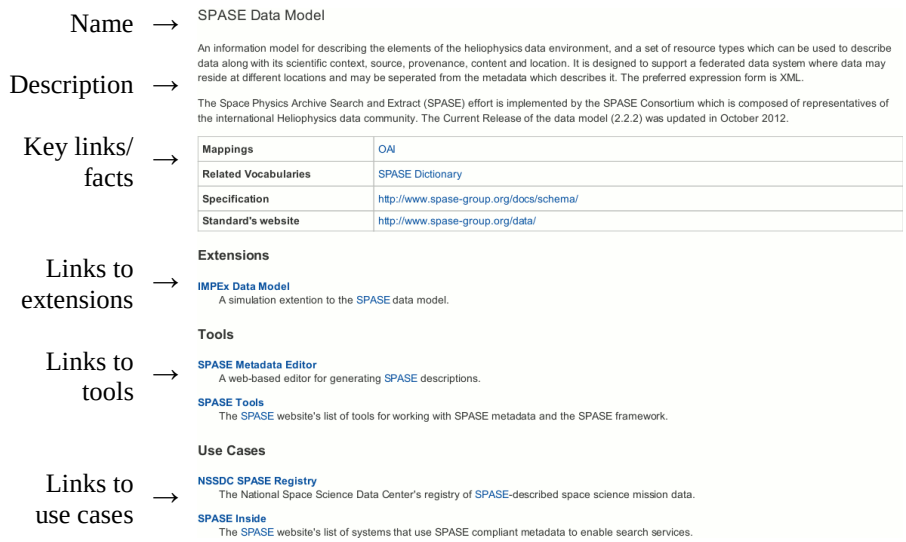

Figure 1. The Disciplinary Metadata Catalogue page for the SPASE Data Model

The table may also contain details of the organization that develops the standard ('sponsor') and an indication of its currency. Beneath that are three lists of annotated links. The first is of application profiles ('extensions') that either refine the standard or borrow significantly from it. The second is of services and software ('tools') available for working with the standard, such as metadata editors or extractors. The third is of repositories, catalogues and services where the standard is being used actively ('use cases').

The catalogue is intended to be browsed rather than searched. The front page provides a link to an alphabetical list of all the standards in the catalogue, and also links to three similar lists of all the extensions, tools and use cases (respectively) that have been included. Probably more useful, however, are the links to the subject area pages. There of five of these, relating respectively to Biology, Earth Science, Physical Science and Social Science & Humanities, with the fifth one reserved for discipline-agnostic metadata. The latter is intended to support multidisciplinary research, or disciplines without specialist metadata standards.

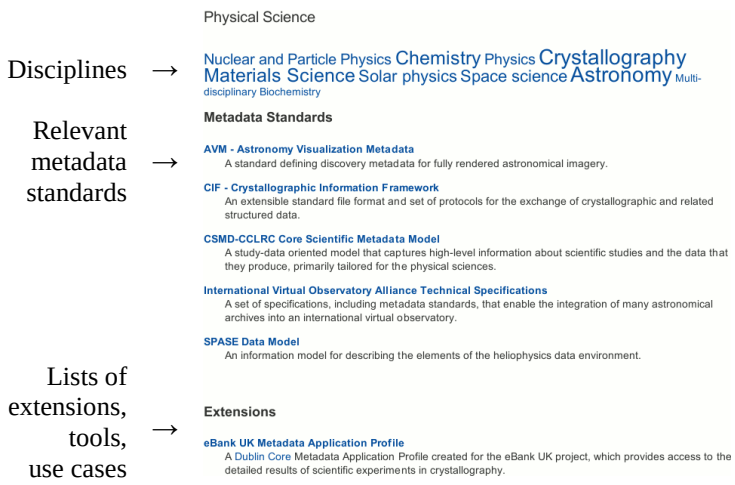

Figure 2. The subject area page for Physical Science

Figure 2 shows an example of a subject area page. The page contains four lists. The first is a list of catalogue records for the standards relevant to the subject area. The other three are lists, respectively, of the extensions, tools and use cases associated with the standards in the first list. At the top of the page is a tag cloud that allows the lists to be filtered further by specific discipline. The taxonomy used for the disciplines is the one used for classifying degree courses in the UK (HESA & UCAS, n.d.); the motivation for using this over other similar taxonomies was that it would be familiar to the primary intended audience for the catalogue, the UK higher education sector.

At its launch, the catalogue contained records for 19 metadata standards. Acknowledging that this did not represent a comprehensive set, Bedford (2013) issued an open invitation for researchers to suggest additional standards to include. Over the course of the following months, five new standards and their associated resources were added to the catalogue as a result.

Extending the DCC Catalogue

When the MSDWG evaluated the catalogue, it found it to be a highly promising resource, but noted ways in which it would need to be further developed in order to fulfil the requirements for the Metadata Standards Directory:

- The catalogue, understandably, had a bias towards standards of interest to UK researchers. The directory would need to avoid any geographical bias.
- Even though anyone could contribute standards to the catalogue, the invitation to do so had become poorly visible over time, no particular structure was provided, and since the effort available to process suggestions was limited, there were delays in adding new entries. The directory would need a more transparent, structured and rapid submission procedure.
- The functionality of the catalogue and administrative access to it was limited, due to it being tightly integrated into a much larger, pre-existing website. The directory would ideally have its own hosting platform on which it would be easier to innovate.
- As a resource hosted by a single organization, the sustainability of the catalogue was questionable. The directory would need to be supported by multiple organisations in order to be resilient.

The MSDWG and DCC jointly considered the options for collaboration, and agreed on a phased development programme. The first phase would involve the MSDWG expanding and updating the catalogue using information gathered from a survey of RDA members and other interest groups. Subsequent phases would involve migrating the information from the catalogue to a newly developed system and performing further development there.

The work of updating the catalogue was performed by two students within the School of Information and Library Science at the University of North Carolina, Chapel Hill, under the supervision of the chairs of the MSDWG and with technical assistance from the DCC (Perez, 2013). The students devised a survey form with which to collect information on disciplinary metadata standards and associated resources. The form was hosted on Google Docs and sent in the first instance to the MSDWG chairs and five other individuals. Feedback from this exercise led to improvements to the wording of

the form and the addition of a section asking respondents if they would consent to being identified as a contributor.

On October 8, 2013, Once these revisions were made, the students sent out an invitation to complete the form to the following mailing lists and groups:

- RDA (all);
- RDA MSDWG;
- RDA Metadata Interest Group;
- EuroCRIS;
- European Plate Observing System (EPOS);
- Dublin Core Science and Metadata Community;
- Association for Information Science and Technology (ASIS&T) Research Data Access and Preservation (RDAP) summit series;
- DataONE;
- Earth Science Information Partners (ESIP);
- UK Science and Technology Facilities Council (STFC);
- attendees of the MSDWG session at the RDA Second Plenary.

Recipients were asked to respond within two weeks. In that period the survey attracted 32 responses (of which 28 contained sufficient useful information) from Australia, Europe and North America, covering a wide range of disciplines.

The responses were transferred to a new spreadsheet; where responses discussed the same standard they were merged. They were then compared to the existing records in the catalogue. Where records already existed, the details they contained were compared against those provided by the survey responses, and any new or updated information noted. Where records did not exist, they were drafted using the information from the responses, supplemented by desk research.

As a result, 11 new standards were added to the catalogue, along with 5 extensions, 7 tools and 19 use cases. Updates were made to 4 standards, 5 extensions, 5 tools and 4 use cases. These changes did not precipitate any major changes to the structure of the catalogue: all the new standards fitted within the broad subject categories already in use. It was notable, though, that several of the standards suggested were outside the original scope of the catalogue.

The survey was left open for further responses, initially to allow US federal employees ample opportunity to contribute – the US government had shut down for the majority of the survey period – but as it had worked well as a more visible and structured input mechanism for the catalogue, the MSDWG decided to keep it open indefinitely.[9] A further 9 responses were received in the remainder of 2013.

[9] MSDWG/DCC survey form: http://bit.ly/1fToaqd

Future Plans

Short Term

Work has already begun on developing a new platform and interface for the Metadata Standards Directory. Among the desiderata for the new interface in its first iteration are the following:

1. Ability for community members to add and edit the entries with the ease of a wiki-based system such as Wikipedia.
2. Version control, both to protect entries from vandalism and for long-term historical interest.
3. Ability for community members to interact with the entries by adding their own annotations, discussions, star ratings and so on.
4. Ability to share entries via major social networks such as Facebook, Twitter and LinkedIn.
5. A more flexible data model for the entries, so that for example tools can be associated with particular extensions directly, instead of via a top-level standard.

The Drupal system used for the DCC catalogue (and indeed the rest of the DCC website) provides version control, a page widget for sharing via social networks, and a comment facility. The MSDWG is therefore considering using a new instance of Drupal as the platform for the directory, as it would simplify the transfer of information from the catalogue. Wiki systems such as MediaWiki, and version control services that include lightweight wikis, such as GitHub, are also being considered.

The issue of sustainability is currently being addressed through partnerships with DataONE and the Dublin Core Science and Metadata Community.

Longer term

The catalogue was designed with a single use case in mind: a data curator browsing for metadata standards and resources that might be useful for a particular researcher or project. The MSDWG plans to extend the utility of the directory to other use cases, including making the information easier to search, and making it easier for automated tools to query and process.

On the latter point, there are again several levels through which the directory might progress. In the first instance, the information from the catalogue would be more amenable to automated access if given an RDF representation, whether embedded in the human-oriented web pages or provided separately through content negotiation. At the next level, the directory could provide Linked Open Data about the elements defined by each of the metadata standards. This would provide the basis for tools with which users could search for and incorporate metadata elements into their own application profiles. It would also be an ideal reference for software developers and (in due course) tools looking up how to interface with conformant metadata records. The feasibility of such plans has already been explored in several projects (Hillman & Phipps, 2007; Tonkin & Strelnikov, 2009).

As the size of the directory increases, greater care will need to be taken to ensure users can still discover the standards and resources of most interest. One way the MSDWG plans to do this is by categorizing metadata standards by how they would

normally be used: for example, for discovery, for enabling reuse by third parties, or for enabling reuse across multiple systems.

Conclusions

The effort to build a Metadata Standards Directory is a timely one. Researchers are under increasing pressure to document and share their data, but if they do so in an ad hoc manner this places an additional barrier in the way of anyone attempting to reuse the data. The DCC Disciplinary Metadata Catalogue and other similar efforts are helping to guide researchers towards existing metadata standards that they can use or adapt. In time this will both aid interoperability and avoid effort being wasted on the development of unnecessary new standards. The transformation of the catalogue into an open and collaborative directory will help ensure that the information contained therein remains current, useful and visible long into the future.

Acknowledgements

The Research Data Alliance is supported by the European Commission and the US and Australian governments. The Digital Curation Centre receives support from Jisc.

References

[Report] Ball, A. (2009). *Scientific data application profile scoping study*. Retrieved from University of Bath, UKOLN website: http://www.ukoln.ac.uk/projects/sdapss/

[Conference paper] Ball, A. (2013). *The DCC Disciplinary Metadata Catalogue*. Paper presented at the CAMP-4-DATA Workshop, International Conference on Dublin Core and Metadata Applications 2013, Lisbon, Portugal. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/203

[Blog post] Bedford, E. (2013). New DCC resource: Disciplinary metadata [Web log post]. Retrieved from: http://www.dcc.ac.uk/news/new-dcc-resource-disciplinary-metadata

[Informal publication] Greenberg, J., Jeffery, K., & Koskela, R. (2013). *Metadata Standards Directory Working Group*. Retrieved from Research Data Alliance website: https://www.rd-alliance.org/filedepot_download/418/100

[Web page] HESA & UCAS. (n.d.). Joint Academic Coding System (JACS) version 3.0. Retrieved from Higher Education Statistics Agency website: http://www.hesa.ac.uk/content/view/1776/649/

[Journal article] Heery, R., & Patel, M. (2000). Application profiles: Mixing and matching metadata schemas. *Ariadne*, *25*. Retrieved from http://www.ariadne.ac.uk/issue25/app-profiles

[Conference paper] Hillman, D. I., & Phipps, J. (2007). Application Profiles: Exposing and Enforcing Metadata Quality. Paper presented at the International Conference on Dublin Core and Metadata Applications 2007, Singapore. Retrieved from http://hdl.handle.net/1813/9371

[Journal article] Parsons, M. A. (2013). Building global partnerships: Second Plenary Meeting of the Research Data Alliance. *D-Lib Magazine 19*(11/12). doi:10.1045/november2013-parsons

[Unpublished dissertation] Perez, C. I. (2013). *The RDA's Metadata Standards Directory: Information gathering* (Unpublished master's paper, University of North Carolina, Chapel Hill).

[Web page] Qin, J., Small, R., & D'Ignazio, J. (2008). Metadata standards. Retrieved from Syracuse University, Science Data Literacy Project website: http://sdl.syr.edu/?page_id=32

[Informal publication] Riley, J., & Becker, D. (2010). *Glossary of Metadata Standards.* Retrieved from http://www.dlib.indiana.edu/~jenlrile/metadatamap/seeingstandards_glossary_pamphlet.pdf

[Journal article] Showstack, R. (2012). Initiative to establish Research Data Alliance moves forward. *Eos 93*(37), 354. doi:10.1029/2012EO370002

[Book] Tanenbaum, A. S. (1988). *Computer networks* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

[Journal article] Tonkin, E., & Strelnikov, A. (2009). Spinning a Semantic Web for metadata: Developments in the IEMSR. *Ariadne 59.* Retrieved from http://www.ariadne.ac.uk/issue59/tonkin-strelnikov/

[Journal article] Willis, C., Greenberg, J., & White, H. (2012). Analysis and synthesis of metadata goals for scientific data. *Journal of the American Society for Information Science and Technology*, *63*(8), 1505–1520. doi:10.1002/asi.22683