Metadata Standards Directory

Alex Ball

University of Bath

RDA Europe Webinar, 16 February 2016

Abstract

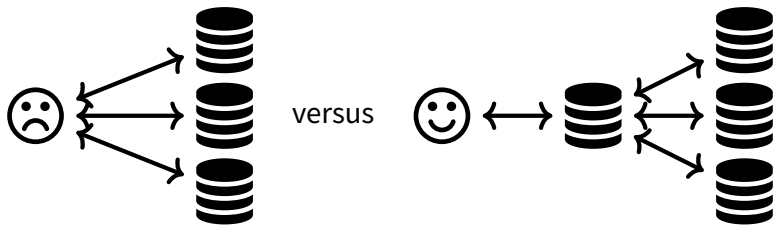
The Research Data Alliance Metadata Standards Directory Working Group (MSDWG) ran from August 2013 to March 2015, with the aim of building a directory to promote the discovery, access and use of metadata standards relevant for research data. The work was conducted in three stages. First, the group examined existing metadata standard discovery tools and compared them against its own list of requirements. Second, the group updated and extended the information contained in the UK Digital Curation Centre's Disciplinary Metadata Catalogue. Third, the group migrated the information in the DCC directory to a new directory hosted on GitHub. In parallel with this, the group collected case studies that will inform future development of the directory under the auspices of the Metadata Standards Catalog Working Group (MSCWG).

Contents

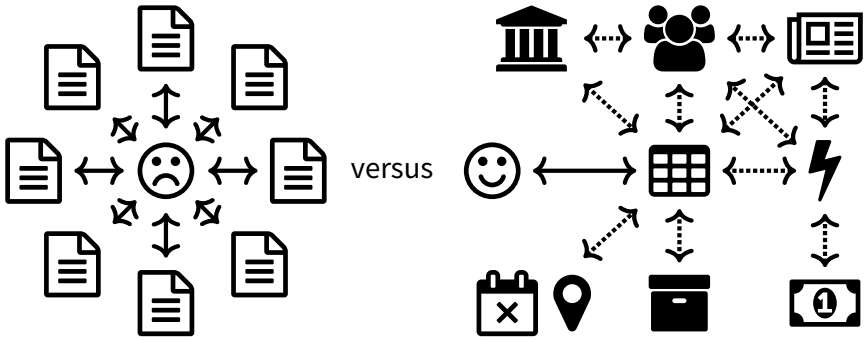
1 Motivation: Why use metadata standards, and why many don't

The motivation for our group comes from a firm belief that metadata standards are a good thing. I don't expect that's too contentious an opinion among RDA folk, but it takes an investment of time and effort to create standardised metadata, so for some it might fall under the 'ideally, if I have time' part of their to-do list. If they were to ask, 'Why should I use a metadata standard?' the answer would be fundamentally about interoperability, but that's something only those of us on the infrastructure side really care about. What *they* want to know is what the interoperation of metadata-powered systems can do for them.

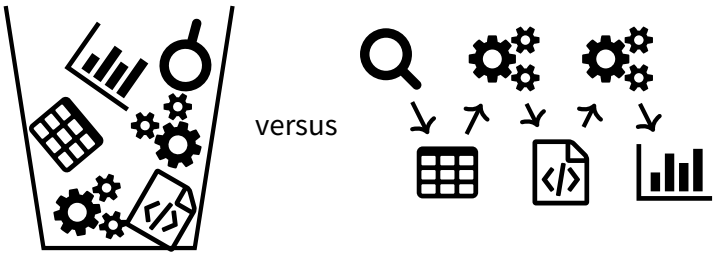
¶ First, it makes *discovering* existing data easier. If lots of *datasets* use a common metadata standard, you can build catalogues that search across that information in useful ways. If many *catalogues* use a common standard, you can build aggregators that mean you can run the same search across all of them at once.



¶ Second, it means putting the data in context is easier. If lots of different entities are documented in a reliable way you can build systems for traversing the relationships between them, instead of having to look things up in different ways, over and over again. And the more entities that are documented in a common way, the more comprehensive a network you can provide.



¶ Third, if a dataset is documented in a reliable way, it saves the next person a lot of time and effort coming to understand the data, meaning they can verify, build on and integrate the data without having to do a lot of detective work first.



¶ Quite apart from systems interoperability, there are a number of side benefits that accrue from people concentrating on a single standard for a single purpose.

If people have a standard to work from, they don't have to work everything out from scratch each time, which means they are less likely to miss out useful information. After using the same standard a number of times, by dint of the practice, they get better and quicker at it.

When many people are using the standard, they can help each other when they get stuck by answering questions and sharing notes. And over time this coalesces into better documentation. Common problems are easier to identify, and if a lot of people care about it, it is easier to get the effort together to fix it. When you have a critical mass of people who would benefit from it, you begin to see tools emerge that take away some of the pain of creating the metadata. And so on and so forth.

¶ So if metadata standards are so great, that leads us to the obvious question: So why doesn't everyone use a metadata standard? *(Audience participation.)*

¶ Well, I might not use a standard because I don't know of any that are suitable for my field of research. Here (Figure 1) is a statistic from a study published in 2011, showing that out of a sample of twelve hundred scientists, over 78% of respondents used either no metadata at all or a home-grown metadata solution. The authors of the study blamed this on a lack of training: researchers either didn't know or didn't care about metadata standards. This points to a need for a resource for discovering relevant standards.

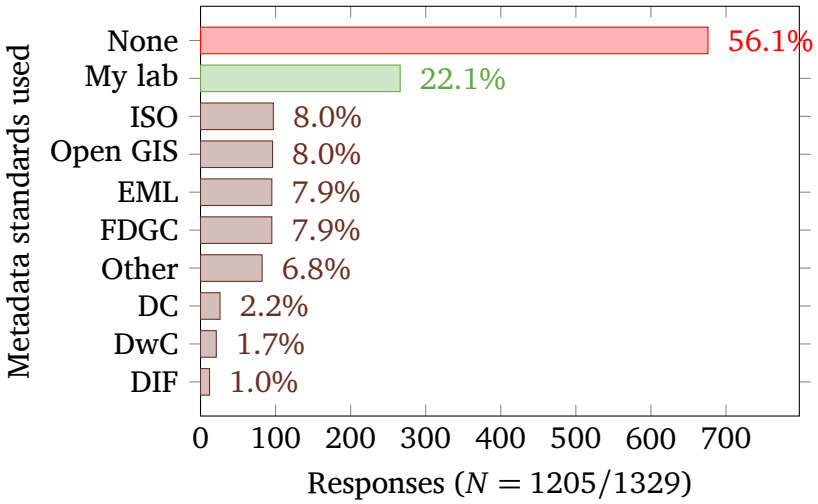
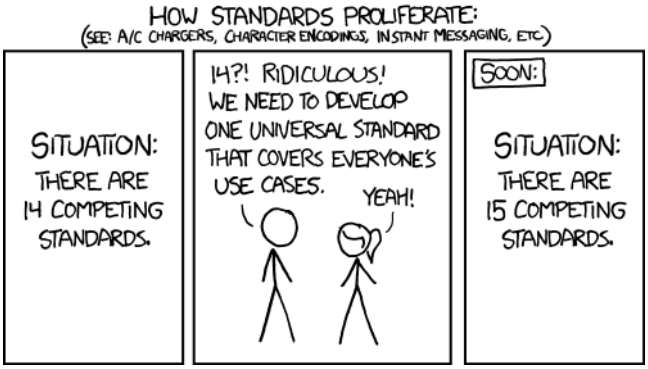


Figure 1: Metadata for scientific data (Source: Tenopir et al. 2011).

¶ Another possibility is that there are just too many standards to choose from (Figure 2). This gives rise to a massive duplication of effort, and dilution of the effort available to go into tools, documentation, training, and future development of the standards themselves.



(Source: Randall Munroe)

'The nice thing about standards is that you have so many to choose from'
— Tanenbaum (1988)

Figure 2: Too many standards?

We can blame some of this on broken communities and inflexible specifications, but I think some of this duplication is down to ignorance: ignorance of the standards that are already out there, being used, and ignorance of the profiling techniques that exist for adapting common standards for local needs.

¶ One other reason I can think of is that it is not always as simple as, 'Use this standard'. Just giving someone an XML specification and telling them to get on with it, well, it's not going to win friends and influence people, is it? At the very least you want to give them a stack of examples they can adapt. Ideally you want to be able to hide all the complexity behind a simple Web form with half of the fields already filled in with the right values. Oh, and give them a phone number of someone they can call if they're stuck.

The point is that it is not just the standards themselves that are important, but also the ecosystem that has grown up around them.

2 Prior work: Finding a perfect match

So we've established that there's a need for some sort of discovery service out there for metadata standards and the tools and agencies that support them.

What has been done to fill that need? There have been several reports published:

- The Science Data Literacy Project published a list of metadata standards relevant for research in 2008 (Qin, Small, and D'Ignazio 2008).

- I wrote a scoping study for Jisc in 2009 looking at the feasibility of a generic scientific data application profile for UK institutional repositories (Ball 2009). In it I looked at the metadata standards used by subject-specific data centres and institutional data repositories in operation at the time.

- Riley and Becker (2010) produced a visualisation of the community, domain, function, and purpose of 105 metadata standards of relevance to the cultural heritage sector.

These are great resources, but they're static. They aren't being updated or expanded. They also don't address the ecosystem question very well. What's really needed are dynamic directories that keep up to date and get fixed when gaps are identified. Happily there are some of these around as well:

- BioSharing is a directory of databases, standards and policies relevant to the life sciences.[1] These include not only what I'd call metadata standards but also vocabularies and file formats, and the search facilities are well matched to its target community.

- The Marine Metadata Interoperability Project maintains a directory of Content Standard References applicable to marine science.[2] By 'content standard' they mean a list or hierarchy of metadata elements making up a metadata record.

- GEOSS, the Group on Earth Observations System of Systems, has to think hard about interoperability, as the whole point of it is to link together a large number of autonomous systems. To help them, they maintain a registry of the standards they support or will consider supporting.[3] There are quite a few metadata standards in there, though there are also technical protocols and file formats.

- EarthCube, the cyberinfrastructure initiative for the geosciences, maintains CINERGI, the Community Inventory of EarthCube Resources for Geosciences Interoperability.[4] This is a diverse collection of resources, including many metadata standards.

These are, again, great resources, but they are specific to particular communities. If you don't happen to fall into one of these groups, or are working between disciplines, you might get a bit stuck.

Incidentally, if you know of any other discipline-specific lists of metadata standards we'd love to know about them.

So we identified a gap for a directory that could be used by any researcher, and could be kept current and comprehensive by the community. It was to fill that gap that we formed the Metadata Standards Directory Working Group.

---

1 https://biosharing.org/standards/
2 https://marinemetadata.org/conventions/content-standards
3 https://www.earthobservations.org/gci_sr.shtml
4 http://earthcube.org/group/cinergi

3 Methodology: The Working Group

1. *The purpose of the group was to* Develop an *RDA Metadata Standards Directory* listing standards relevant for research data *that would be*
    - Comprehensive*, covering all disciplines and most generic applications, and*
    - Easy for anyone to contribute *to* or update
2. *The group also decided to* Define and develop *use cases* for research metadata*, to help with organizing standards within the directory, and to*
3. Develop a plan for *the* long-term growth and maintenance of the directory

¶ *We could see that many different* Stakeholders *would have an interest in this directory:*

- researchers
- data managers
- data scientists
- research support staff
- tool developers
- repositories
- funders
- publishers

…in fact most of the Research Data Alliance's growing community.

We have engaged with them in various ways…

- *Two years ago at the Dublin Core Conference in Lisbon we held* CAMP-4-DATA: *This event attracted* 26 participants from 15 countries *and covered a range of scientific metadata issues, from infrastructure models and frameworks, through usage and tracking of metadata, to common fields to associate with persistent IDs.*
- *I gave a presentation at the* International Digital Curation Conference 2014, *in* San Francisco
- *And at about the same time, my colleague Jane Greenberg presented our work at the* RDA-EU Working Group Core Meeting, 2014, *in* Garching, *near* Munich
- *We also made use of our* Working Group mailing list*, which is accessible from the RDA website. At the last count that put us in touch with 140 RDA members.*

4 Results: The Metadata Standards Directory

Before I get around to talking about the directory we developed, there is one more resource I should tell you about, and that is the Disciplinary Metadata Catalogue developed by the UK Digital Curation Centre (DCC).[5] This was launched in January 2013, at about the same time that the Working Group was being planned.

[5] http://www.dcc.ac.uk/resources/metadata-standards

It was conceived as a resource that institutional data curators could consult when advising researchers on how they should document their data. The thinking was 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, as 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.

It was almost exactly the kind of thing the Working Group was looking for. Almost, but not quite. But on the same principle as before, that it is better to develop and improve what is already there than to start again from scratch, the Working Group and the DCC entered into collaboration to use the catalogue as a starting point for the Metadata Standards Directory.

¶ There were some specific issues to address:

- UK selection bias – *as it had been compiled with a UK audience in mind, certain profiles and standards specific to other countries had been missed.*
- Incompleteness – *it had been compiled mainly by Liz Bedford and myself, and we couldn't claim deep knowledge of every discipline, so there were some standards we had missed.*
  - → *§ To address those issues, the Working Group decided to Conduct a worldwide survey of researchers to fill in the gaps. That was not all.*
- *§* Process for maintenance slow and opaque – *if anyone wanted to add in a change, they would have to email the DCC Helpdesk, the Helpdesk would pass it on to me, and I would make the change when I had time. And because the catalogue was just another set of pages on the DCC Website, it wasn't all that simple to bring in help from outside due to the security implications for the site.*
- *For the same reason, there was* Little scope for future development – *we could only do what Drupal allowed us to do, and only if it would not compromise the rest of the site.*
  - → *§ So, with an eye on the longer term, we planned to migrate the service and content to a new platform to*
    - » *make it easier and more transparent for anyone to contribute to it,*
    - » *support a larger team of volunteer editors, and*
    - » *allow for future development.*

¶ The survey was performed by two students, Sean Chen and Cristina Perez, within the School of Information and Library Science at the University of North Carolina, Chapel Hill, under the supervision of the chairs of the Working Group and with technical assistance from the DCC (Perez 2013). Following an initial pilot, the survey was circulated to stakeholder mailing lists around the globe in October 2013.[6] We had intended to keep it open for only two weeks, but unfortunately the US Government was shut down for most of that time, so we agreed to keep the survey running as our primary way of collecting contributions until we could set up a better solution.

Our last contribution via the form was in April 2014. In that time we had received a total of 41 responses, giving us 50 new entries and 18 updates for the DCC catalogue.

¶ The next task was of course to migrate the content to a new platform. In early 2014, Sean Chen set up a prototype for us on GitHub Pages that mirrored the DCC data structures but was generated from a set of really simple text files. I'll show you what I mean in a moment. It was polished further by Kate Anne Alderete, our DataONE intern, between May and August 2014,[7] and finally made production ready by Adrian Ogletree, Dustin Allen, Sean and myself between January and February 2015. We launched it at the RDA Fifth Plenary.

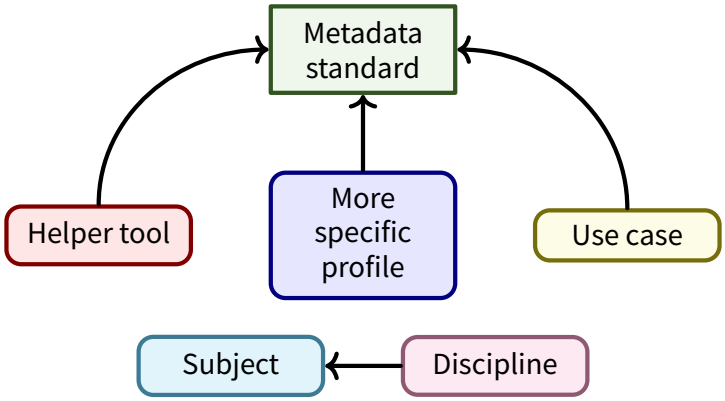
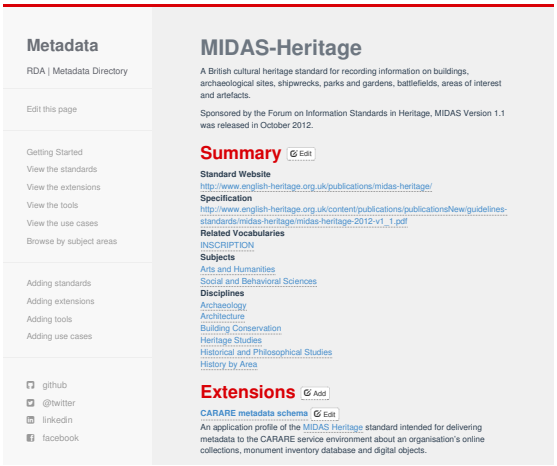


Figure 3: Data model for the Metadata Standards Directory.

¶ The Metadata Standards Directory has a six element data model (Figure 3). The records for the metadata standards themselves are the most detailed, of course, but there are also much simpler records for § profiles of those standards, § tools for working with them, § and examples of repositories and services that use them. All of these records are classified by § the broad subject area and § the specific disciplines to which they relate. We also have special classifications for general and discipline-agnostic standards.

¶ So now let me show you what a record for a metadata standard looks like (Figure 4, left side). At the top we have the name and a short description that tells you what the standard is used for. We then have a summary table with additional information, such as a link to the standard's home page, a direct link to its specification document, links to vocabularies commonly used with the standard, lists of known mappings to other standards, and so on. The last two parts of the summary show how the

[6] http://bit.ly/1fToaqd
[7] https://notebooks.dataone.org/metadata-standards/




```
---
title: MIDAS-Heritage
slug: midas-heritage
description: <p>A British cultural heritage standard for
  recording information on buildings, archaeological ...
  MIDAS Version 1.1 was released in October 2012.</p>
website: http://www.english-heritage.org.uk/publicatio...
subjects:
- arts-and-humanities
- social-and-behavioral-sciences
disciplines:
- history-area
- heritage-studies
- building-conservation
- historical-and-philosophical-studies
- architecture
- archaeology
specification_url: http://www.english-heritage.org.uk/...
related_vocabularies:
- name: INSCRIPTION
  url: http://fishforum.weebly.com
---
```

Figure 4: A record for a metadata standard, (*left*) as rendered, and (*right*) as YAML source code.

standard has been classified according to its relevance to broad subject areas and specific disciplines. You can click on the names to look up what other standards have been classified with the same terms.

Below that are the lists of profiles or extensions, tools and use cases that have been tagged as relevant to this standard.

You will notice there are little buttons marked ‘edit’ and ‘add’. These take you through to the GitHub online text editor where you can, respectively, edit the entry you see or add a new one. Let me show you what the source for this page looks like (Figure 4, right side). § As you can see, it’s about as simple as you can get. It’s in a format called YAML, which is a recursive acronym for ‘YAML Ain’t Markup Language’. Whether that name is accurate is debatable, but the syntax is clear: you have key, colon, value, with some extra notation for multiple values and hierarchies. This means it can be parsed easily by software as well as by humans. We provide templates with all the keys we use, and where we use controlled values, they relate to filenames in the system.

¶ You can browse through the directory in several ways (Figure 5). We have an index that list the names and descriptions of all the standards, arranged by broad subject area. The names link through to the record pages. There are similar indexes for profiles, tools and use cases, but for those entries the names link through directly to their home pages rather than a local record. That may change if we get a demand for more detailed entries for them.

§ Alternatively, you can view a page for a subject area or discipline and see all the entries that relate to it. On the slide is the page for Archaeology, and you can see it has the four different lists.

§ Those pages are all accessible from our index of disciplines, again subdivided by broad subject area.

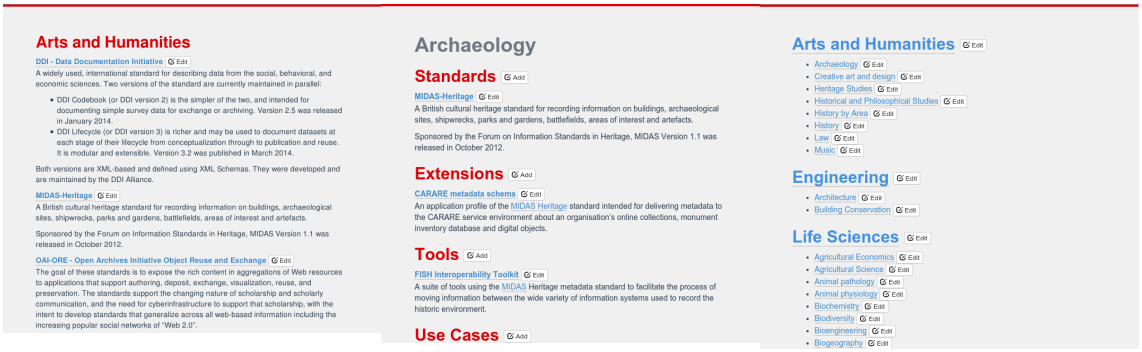


Figure 5: Ways of browsing the directory: (*left*) standards by subject area, (*middle*) entries by discipline, and (*right*) disciplines by subject.

¶ If you do contribute via the GitHub editor or your own local copy, the edits won't go in directly; they'll come to us as a pull request. This just gives us a chance to do some quality control and weed out any spam. We do welcome contributions and your edits will go in pretty quickly.

§ Another way you can help is to tell us about a standard on the issue tracker. It's very quick and easy to do, so if you don't feel up to adding a full entry you can at least put it on our to-do list so it doesn't get forgotten.

To date there have been 25 updates or additions suggested through the issue tracker and 21 contributed through the pull request mechanism.

Figure 6: Usage statistics showing commits to GitHub since December 2013.

¶ That gives you some idea of how well it is being developed. You can also see that in the commit history for the source code (Figure 6), but how well is it being used?

I'm afraid I don't have usage figures for the site on GitHub, but I can tell you that in 2015, the DCC directory received 27 355 visits, accounting for 6% of the traffic to the DCC website. That doesn't sound like much, so to give you a sense of what that means, it's the third most popular section of the site, after the events section with 14% of the traffic, and the How-to Guides section with 9% (see Figure 7).

¶ I think that's a good start, and really encouraging. But we can do better than that… There are still lots of ways we could develop it further to make it even more useful to people.

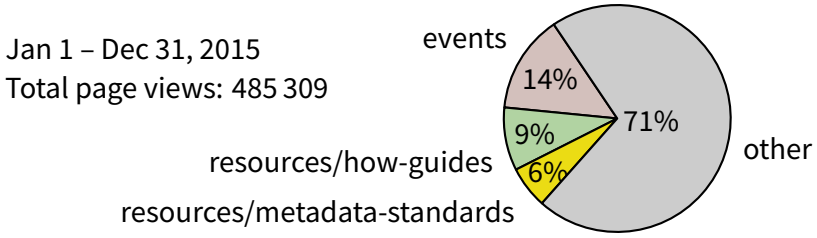


Figure 7: Usage statistics showing the proportion of page views attributable to sections of the DCC website since January 2015.

5 Next steps: The Metadata Standards Catalog

At the RDA Third Plenary in Dublin, we had a brainstorming session and thought up a large number of possible ways in which different stakeholders might use the directory. Many of them were too ambitious for what we were able to achieve in our 18-month time frame, but we kept them in mind as we developed the GitHub prototype, and now we're in a position to consider developing a much more capable catalogue. So on these next few slides I'll show you some of those use cases.

Data providers and *custodians* would like to use the Directory

- to search or browse for metadata standards by what they describe – *not only types of tabular data, but also* physical artifacts, video, etc.
- to compare standards side-by-side, especially to identify commonalities between the standards of different communities. *I can see that also being useful when constructing a crosswalk or common interface.*
- to obtain recommendations of standards to use based on criteria I provide. *Beyond the disciplinary aspect we already have, that might include something like number of use cases, or association with a standards body.*
- to look up the persistent ID (PID) for a standard, for robust linking. *Persistent IDs were on people's mind that day, as you'll see.*

Librarians would like to use the Directory

- to inspect existing profiles of a standard as a first step to constructing their own. *We support that now, so we mustn't lose that capability.*

¶ *Journal editors* would like to use the Directory

- to check the maturity and level of support of existing standards, so they know which to recommend to authors. *We can infer that now from our lists of tools and use cases, but we might present things differently to make it easier.*

Funders would like to use the Directory

- to find out of which standards they have funded the development, whether they are widely used, whether they have been kept up to date, and whether they might be merged into other standards. *There are overlaps here with previous use cases: search criteria, maturity, and comparison across standards.*

But our biggest list came from Tool developers, who would like

- to submit a whole or partial dataset and retrieve a list of metadata standards which could be used to document it. *For that we'd need a JHOVE or DROID-like format identifier and quite detailed knowledge of what data types the standards are intended to support.*
- to generate a 'first attempt' crosswalk between schemas automatically. *This is a more complex version of the comparison function from earlier.*
- ¶ to submit a set of field names to the Directory and retrieve the metadata standard from which they originate. *Again, this would require some sort of format identification routine.*
- to request from the Directory a sample of metadata records adhering to a specific standard. *For that, we'd need to assemble those samples.*
- to retrieve a list of appropriate metadata standards based on the partial content from a draft data management plan. *This one has actually rocketed up to the top of our list. We'll be looking at it in an RDA Europe Collaboration project, again with the DCC who also develop the DMPonline tool.*
- to submit a PID for a metadata standard to the Directory and retrieve the specification for the standard. *The group had it in mind that we would record PIDs where they already existed, but maybe mint our own for those without.*
- to submit a pair of PIDs for metadata standards to the Directory and retrieve a suggested migration pathway. *We do already collect information on crosswalks, so this would be a great use for that information.*

¶ That list may have sparked your imagination, so Can you think of any others? *(Audience participation.)*

¶ As I say, the current Directory is a great base to build on, but there is a lot more work to do. We Can only hope to satisfy such use cases

- with more detail about each standard/profile
  - *some of those use cases require* information at the element level? *Do we keep that information as*
  - specifications in 'native' form? *We already provide links, this is about storing and serving the specs ourselves.*
  - *Or do we keep the* specifications in a normalized form?

  - *Should we provide* converters in a normalized form? *Again, we provide links already, this is about storing and serving ready-to-use code.*
  - *Do we develop a classification of* data types for which the standard*s are* used?
  - *do we add flags that mark* whether *a standard is* tied to a given format, *or* whether *it is* RDF-friendly *and thus able to be used in Linked Open Data applications*?

- *We will certainly need* an API for automated access to the information
- *and* structured outputs so tools can act on the information

Also *we will* need to make it even easier for people to contribute, directly or via tools.

¶ To this end, we are starting a new Working Group, called the Metadata Standards Catalog Working Group. The move from Directory to Catalog is meant to signify that extra level of sophistication that will allow software to interrogate the resource and act on what it receives back.

- *The new group was* Recognized and endorsed on 18 January 2016 *and we've agreed a work plan as follows.*
- By 18 July 2016, *we will* collect and analyse *use cases* to produce requirements and technical specification for the Catalog
- By 18 January 2017, *we will* develop *a prototype Catalog* and identify adopters
- By 18 July 2017, *we will* evaluate and validate *the* Catalog; add *mappings* from *a* selection of standards to functional 'packages'; *and* refine *the* user interface and API.

Let me explain about the packages. The idea is similar to that of the CASRAI dictionary, or perhaps the inner workings of Pandoc if you're familiar with that conversion utility. The package would set out the metadata that is needed for a particular function, in a standards-neutral way. By mapping from each standard to the package, we can use that to match up elements doing the same job in different standards and generate a rough and ready converter to go between them. The package itself wouldn't be a rival standard, it would simply act as a way of simplifying translations between standards, reducing the complexity from n^2 *to* n*.*

Please join us! *You can find the page for the new group at the following URL.*

`https://rd-alliance.org/groups/metadata-standards-catalog-working-group.html`

6 Acknowledgements

All that remains is for me to acknowledge the many contributors to this work:[8]

Fellow MSDWG co-chairs

- Jane Greenberg, ⟨MRC⟩
- Keith Jeffery
- Rebecca Koskela, DataONE

DCC Disciplinary Metadata Catalogue

- Liz Bedford, DCC

Survey and GitHub work

- Sean Chen, ⟨MRC⟩
- Cristina Perez, ⟨MRC⟩
- Kate Anne Alderete, DataONE
- Adrian Ogletree, ⟨MRC⟩

I would like to thank the Working Group members who suggested directory entries, provided use cases, and helped to steer the work.

Finally, thank you for listening.

References

Ball, A. (2009), *Scientific Data Application Profile Scoping Study Report*, version 1.1, Scoping study (Bath, UK: UKOLN, University of Bath, 3 June), http://www.ukoln.ac.uk/projects/sdapss/papers/ball2009sda-v11.pdf.

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, http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/203.

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

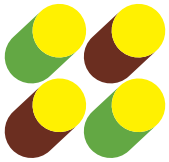
Qin, J., Small, R., and D'Ignazio, J. (2008), 'Metadata standards', Syracuse University, Science Data Literacy Project, http://sdl.syr.edu/?page_id=32.

Riley, J. and Becker, D. (2010), 'Seeing Standards: A Visualization of the Metadata Universe', Indiana University Libraries, http://www.dlib.indiana.edu/~jenlrile/metadatamap/.

Tanenbaum, A. S. (1988), *Computer Networks* (2nd edn., Upper Saddle River, NJ: Prentice-Hall), ISBN: 0-13-162959-X.

Tenopir, C. et al. (2011), 'Data Sharing by Scientists: Practices and Perceptions', *PLoS ONE*, 6/6: e21101, doi: 10.1371/journal.pone.0021101.

[8] ⟨MRC⟩ = Drexel University Metadata Research Center; DCC = UK Digital Curation Centre.



The Metadata Standards Directory Working Group is part of the Research Data Alliance, which is supported by the European Commission, the US Government and the Australian Government.

For more information, please visit https://rd-alliance.org/groups/metadata-standards-directory-working-group.html