Flying the flag: in support of metadata standards

Alex Ball

2017-05-25

It is well understood how open data can lead to more rigorous research, and unlock the possibilities for answering increasingly complex questions and challenges. Often overlooked is the vital role that metadata plays in allowing that to happen. Metadata is vital for discovering data, assessing its quality and relevance, and for reusing and repurposing it. Far too much metadata is either missing or in a form only the creator’s systems can use. We need to take seriously the challenge of providing standardized metadata as standard, and that will take commitment from systems designers, research data managers, and researchers alike.

Introduction

Title slide

My short biography.

Metadata specialists are like hamsters… DCES took four years to agree on 15 elements and another three to be published as a standard.

Two reasons:

• metadata is really important, and I hope to show you why;
• for any problem, there’s usually more than one satisfactory answer. The precise form doesn’t matter so much as it being there in a consistent manner.

Bloodbath at St Justin’s

Before that...

I read Physics and Philosophy as an undergraduate. In my first term I had logic lectures... Our logic lecturer was Bob Hargrave, and to make the subject more palatable, he organised a Fantasy Logic League: competitive puzzle solving. Many
fine and funny puzzles, but there was one of that really stayed with me, and it was this one...

What struck me about it was that it does what every book description is supposed to do: give a flavour without spoiling the full experience...

Look at how careful it is being not to reveal who the murderer is. And yet. Even with it trying so hard to conceal that one vital piece of data, it fails...

**Who killed the Chaplain?**

The relevance of this is, from an Amazon store perspective, that book description is metadata. It is surprising how powerful metadata can be, for revealing new insights into data, for organizing and manipulating data, and as a resource in its own right. Need to explain what ‘metadata’ means...

**Metadata**

**What is metadata?**

Facts about facts about stuff…

Not helpful, and probably too narrow.

People recognise metadata not by what they *are*, but what they *do*, and by the role they play…

High level view of the research process…

When you do this the first time, rely on *tacit information*: you know it because you’re the one doing it. But if you’re interrupted, you might need reminding of certain things; if someone else were to reproduce or validate your findings, chances are they’d be stuck…

Metadata are all the other things they would need to know…

None of these are the data that proves the point, but they make it possible to draw meaning from the data.

So if something is metadata because of its role, it follows that the same information can be data in one process and metadata in another.

**Example: Internet traffic**

When we use the Internet, we’re receiving and sending data. The idea of encryption protocols is that the data will be gobbledegook to anyone in the middle… But ISPs and servers take logs of of all the transmissions… Helps them to provide the service.
In popular media, this is the context where ‘metadata’ has been used… § Snowdon revelations, Investigatory Powers Act… People have been profiled using just the metadata, but to GCHQ, the information has now become data proper.

To give you an idea of how powerful this is, in 2012 in the US, the FBI used these techniques to uncover a sex scandal between David Petraeus, Director of the CIA, and his mistress Paula Broadwell.¹ They used IP addresses to demonstrate that they were both accessing the same anonymous email account.

**Types of metadata**

As a result, people now sometimes look at me funny… But there are plenty of less sinister things you can do with metadata. We have names for different sets of metadata depending on what they’re useful for…

**Why should I use a metadata standard?**

Answer is fundamentally about interoperability, but that’s something only those of us on the infrastructure side really care about. What researchers want to know is what the interoperation of metadata-powered systems can do for them.

**Better discovery**

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.

**Better context**

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.

**Better reuse**

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.

¹ See https://www.aclu.org/blog/surveillance-and-security-lessons-petraeus-scandal
**Better ecosystem**

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.

**Research Data Discovery**

**UK Research Data Discovery Service (phase 1)**

I was involved in the early work on setting up RDDS for the UK. (Another team is working on it now.) Idea was to set up service where every dataset in the country had a page like this... More information off the bottom of the screen

**Metadata requirements**

We wanted to enable to researchers to search and browse for interesting data, and so we needed to collect metadata...

**Collaborators**

Our collaborators included 9 national data centres and 9 universities, so we needed to find a way to harvest metadata from them all.

**Metadata flows**

Not a standing start; the national data centres and Oxford were already contributing to metadata aggregators. We planned to benefit from that as much as possible, but bring in records from more places. Thus we could contribute to this network of metadata flows by

- collating records from both data centres and institutional repositories;
- normalising and deduplicating, to provide a unified search interface;
- ultimately make the records visible in other places researchers might look.
RIF-CS data model

The platform we chose was ORCA, a catalogue system developed by ANDS for Research Data Australia. It was designed around a metadata profile called RIF-CS (Registry Interchange Format: Collections and Services) based on ISO 2146. If you are familiar with CERIF, it is similar in philosophy but about 20 times simpler. RIF-CS is not very well known or supported: couldn’t expect our partners to implement it and send us ready made records...

Metadata crosswalks

So we had to write crosswalks to harvest records in formats that were supported. Had OAI DC as a fallback that most repositories should support, but we also wanted to benefit from more detailed metadata that many repositories might be able to provide.

Ended up supporting 6 standards: DDI & UK Gemini were discipline-specific for national data centres. DataCite for DataCite. EPrints for EPrints. MODS for DSpace. Dublin Core for everything else. The hard bit was doing the matching and the mapping, but programming the crosswalk was fairly straightforward and we got good results.

Lessons learned

I won’t bore you with the nuts and bolts of how we did this. There is a report online with the details if you’re interested. But I do want to reflect a bit on what we learned from the experience...

...We could only work with what we were given. If there was a systematic error, we could correct for it, but if missing there was little we could do.

DataCite Metadata Schema v4.0

So to end this case study with a plea. Here is the DataCite scheme. Note the very small number of mandatory elements. We can do very little with such a little. We need the others as well for meaningful search. So if any of you are managing a repo and minting DOIs, please, please provide as much metadata as you can.

So why doesn’t everyone use a metadata standard?

No suitable standard?

I might not use a standard because I don’t know of any that are suitable for my field of research. Here 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.

**Too many standards?**

Another possibility is that there are just too many standards to choose from. 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.

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.

**Isn’t that really hard?**

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.

**Metadata Standards Catalog**

For this reason, I can’t tell the researchers I support, ‘Just use standards.’ It is a struggle sometimes just finding the right standard to use, let alone working out how to use it. For this reason, a group of metadata enthusiasts went to the the newly formed Research Data Alliance in 2013 and asked to set up a Metadata Standards Directory Working Group.

**RDA Metadata Standards Directory WG**

After some negotiation to get the terms of reference right, the group was set up and started work in August of that year. Their goals were...

- comprehensive = across disciplines, countries and use cases;
• easy = to spread the load of keeping it current;
• use cases = to help organize the records in the most natural way;

Existing work

This idea was not new, and there were and are many other resources out there enumerating metadata standards. Most of them were either one-off reports or discipline-specific maintained lists. There was only one other resource that was both generic and dynamic … and that was the one I was responsible for maintaining.

The Metadata Standards Directory

So we joined forces. I joined the RDA group and together we formulated a plan. The issue with the DCC Directory was that it was just another section of the DCC website, so for security reasons only DCC staff could edit it. So if anyone spotted an error or wanted to add anything, they had to go through the helpdesk, things would not necessarily happen immediately and the process was opaque. Plus the information was sort of trapped on the web pages and not very reusable.

In stage 1, we conducted a worldwide survey to update and extend the content of the DCC directory. In stage 2, we set up a new version of the directory on GitHub. It could do nearly everything the DCC directory did, though I will admit it wasn’t quite as pretty, and the underlying data was there in YAML files that anyone could inspect and use. The process for submitting information was opened up to public view, and if there were any queries we could hold the debate about them out in the open.

But there is more to be done

But while all this was going on, we were also gathering ideas of what people might want to use the directory for. Blue sky thinking. They came up with lots of great ideas that the Directory couldn’t handle. Only a statically generated site after all. Ideas included…

The Metadata Standards Catalog

So over the last year and a bit we have been working on a new Metadata Standards Catalog that can do everything the old Directory did plus a whole lot more. We’ll be launching it over the summer, but I can give you an idea of what it looks like. Here is what the page for a standard looks like. Everything on the page is meant to help answer an important question…

We also provide a search facility, where you can search by title, subject, identifier, data type or funder, and an API for getting the information in JSON format.

One of the biggest changes is a web form interface for editing the records, which you can access by signing in through Google, Twitter or various other providers.
**Future developments**

Even so, there are still more features that we haven’t managed to build into the Catalog yet, but set things up in such a way that we can add these in at a later date...

I think this, and the changes being made to some of the discipline-specific catalogues like BioSharing, have the potential to make a serious improvement in the standard of data documentation, with knock-on benefits for discovery, reuse and sharing, as I outlined at the beginning.

If you want to follow developments, you can get the latest news from the URL on the screen.

**Canonical metadata packages**

For the side-by-side comparisons, we are hoping to take advantage of an activity being conducted by the Metadata Interest Group, and this is, controversially, to define canonical metadata packages for different use cases.

The idea is to think about the most useful metadata elements for particular purposes, and then identify where these occur in various metadata standards. This way, when you need to, you’ll be able to grab the metadata you need from any given standard, and encode it again in any given standard to support others with the same use case. This will also help to make the task of maintaining mappings between standards, since you will only need $n$ instead of $n^2$.

These are the high level elements that we identified from use cases collected by all the metadata groups. But as you can see, they are at a high level of abstraction, so between now and September, we’re planning to unpack these elements into more precise concepts.

**Unpacking the elements**

Here is an example with spatial coordinates...

If you would like to join in with this activity, keep an eye on the MIG web page for announcements about telcons we’ll be holding to discuss each of the high-level elements in turn. Once we’re reasonably confident about them, we plan to publish them as an official RDA output.

**Call to action**

**Metadata → better data**

So I’d like to leave you with the main points again from my talk, and these will be relevant to you if you are a researcher, a data manager or are setting up a data archive...
...and above all, be consistent, because if there is a consistent error, we can correct for it and make allowances. If it’s a muddle then each one has to be done by hand, and that defeats the object. Metadata should make everything flow seamlessly so we can all make the most out of our data.

Thank you for listening.