Metadata: the Spirit of Research Data Management

Alex Ball

2017-05-25

To emphasise the importance of data, it has been described as the new oil, the new water, even the new light. But what is not so well understood is the vital role that metadata plays in unlocking the value of data. It is metadata that allows datasets to be discovered, accessed, understood, assessed, and ultimately reused in new studies. It is key to demonstrating the integrity of research, instrumental in increasing its impact, and a cornerstone of interoperability. If the full potential of data is to be realised, giving it global and long-standing significance, researchers must collect appropriate metadata and data archives must curate and syndicate it according to established standards. This means research institutions need to participate in wider research data infrastructures; encourage the use of standards and profiles wherever possible; and above all support their researchers with training, tools and guidance.

Introduction

Title slide

(My short biography.)

You've already heard today about the importance that FAIR data has for world class research, and so now it falls to me to explain why high quality research data, and high quality research data management, depend upon high quality metadata.

When preparing this talk, I was looking for an appropriate metaphor for describing the role that metadata plays in the research process. I looked to see what others had said, but what I found were metaphors about the role of data.

Data is the new oil

One that I found in many places is that data is the new oil. The earliest reference I could find to this was a talk by Clive Humby, an UK mathematician who set up a

loyalty card scheme for a major supermarket chain. These loyalty cards, of course, enabled that supermarket to track the spending habits of individual customers, and use that data to optimise their marketing, pricing, and supply chain strategies more precisely than ever before.

He describes data as the new oil because it is a highly valuable commodity but, crucially, not very useful in its raw form. In order to make it useful, it has to be broken down, refined, understood and analysed. It's telling that Clive goes on to say that for data to mean something, you need to place it in context. The way in which data is collected, the way it is cleaned, aggregated and analysed, makes an enormous difference to the way the results should be interpreted. The data are useless without metadata.

Data is the new water

I also saw a talk recently by Juan Bicarregui, who is a senior officer at STFC, the agency in the UK responsible for large research facilities, and until recently the Head of the Organizational Advisory Board of the Research Data Alliance (RDA). (Describe the RDA if not done already)

In describing how the RDA works, Juan used a different metaphor, that data is the new water. Partly this a comment on how data is as vital to research as water is to life. But the main point was actually about how the RDA groups relate to each other. Up in the branches of the tree we have the domain groups: social scientists, nuclear physicists, agriculturalists and so on, all coming together to improve data sharing within their own disciplines. Down in the ground we have the infrastructure groups, solving the technical challenges. In the middle we have the trunk of the tree, the machinery of RDA that allows the domain groups to extract and use the data that the infrastructure groups make available.

And if you look at the infrastructure groups, most are dealing with metadata in some form. Some, like the Metadata Interest Group, concentrate on the elements needed for certain tasks. Some, like the Data Type Registries Working Group, concentrate on supplying values for those elements. And others, like the Research Data Provenance Interest Group, do both. But almost all deal with metadata in one way or another.

Data is the new light

Lastly, here is an intriguing metaphor from Axelle Lemaire, Minister of State for Digital Technology in France. She sees data as the new light: ubiquitous and inexhaustible, a resource that does more good the wider it is spread, and that helps us to understand the world around us.

Mark Parsons, outgoing Secretary General of the RDA, took this metaphor a step further, noting that light is also an effective disinfectant. His point was that the only way to erase the lies and deception that seem to be infecting global politics is to shine light on them, or in other words, counter them with clear and unambiguous

facts. But those facts, that data, can only be bright and clear if the metadata are there to clarify them.

Metadata

So if data is the new oil, the new water, or the new light, what is a good metaphor for metadata? It's a tricky question to answer because metadata can be many different things and serve many different functions.

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, and its surprising how powerful it can be.

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 the security services, 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.

[1] See https://www.aclu.org/blog/surveillance-and-security-lessons-petraeus-scandal

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?

Now, simply documenting that information is a really good start, but it is even better to provide it in a structured form according to an agreed standard. Why is that?

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.

Metadata flows for data discovery

I had personal experience of this, actually. I was involved in Phase 1 of the development of a UK Research Data Discovery Service, led by Jisc. The idea was to set up an aggregator as I've just described, that would enable you to search across all the data centres and institutional data archives in the country at once. If we had to write a separate crosswalk for each one, it simply would not have been possible. But because of standards such as DDI in the social sciences, INSPIRE for geospatial data, and DataCite for just about anything, we were quite confident we could aggregate from all these data centres and archives with just six crosswalks.

We were helped that the concept had already been proven by other groups. The UK Data Archive was heavily involved in our project, and they already had experience of contributing their records to the CESSDA Catalogue. We were able to use the same feeds to make their records visible in even more places, or at least they will when the service launches.

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.

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.

So why doesn't everyone use a metadata standard?

So if using standards is so beneficial, why isn't everyone using them?

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 divided 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. Researchers need support to be able to do the right thing.

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;

The Metadata Standards Directory

As it turned out, the UK Digital Curation Centre, where I was working at the time, had already launched a directory that hit most of the criteria. It was aiming to be comprehensive across domains and was actively maintained. This is it, on the left.

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.

So we joined forces. I joined the RDA group and together we formulated a plan. 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.

RDA Metadata Standards Catalog WG

But while all this was going on, we were also gathering ideas of what people might want to use the directory for. We asked them to be creative, and they most certainly were. They came up with lots of great ideas that the Directory couldn't handle. Only a statically generated site after all.

So over the last year and a half we have been working on a new Metadata Standards Catalog that can do everything the old Directory did plus a whole lot more. This has involved migrating the records we hold to a richer data model, overhauling the user interface to allow searching and editing, and implementing a new machine-to-machine interface to allow information from the Catalog to be embedded in other applications. We'll be launching it in the coming weeks, but I can give you an idea of what it looks like.

The Metadata Standards Catalog

Here is what the page for a standard looks like. Everything on the page is meant to help answer an important question:

- To help with the choosing an appropriate standard, we have a title, description, tags for domains of study and data types. We also have sections for endorsements and version history, to give an idea of how actively it is developed.
- To help people use the standard, we have links to human-readable documentation, metadata scheme specifications, tools that read or write to that standard, sample records and examples of archives or other services that use the standard.
- To help people refer to it, or find it again, we record identifiers associated with the standard, plus other citation information like the maintaining organization.
- Lastly, to help people move data into and out of the standard, we list relationships to other standards: is it a profile of another standard or vice versa; are there mappings between this and other standards?

§ We also provide a search facility, where you can search by title, subject, identifier, data type or funder. The API allows you to run the same queries and get the information back 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.

In this way, we hope to break down many of the barriers that get in the way of people and systems developers using metadata standards, in the hope that we can all benefit from better discovery, context and reuse.

Call to action

So this brings me back to the question I raised at the beginning. What is the best metaphor for describing the role and importance of metadata to the research enterprise? I'd like to propose metadata as the spirit of research data management.

Metadata: the Spirit of Research Data Management?

Metadata brings data to life by imparting dull strings of codes and numbers with real significance, explaining their provenance, significance and value.

It is what powers your current research information systems and data archives, allowing you to properly manage and curate your data.

By flowing out from your local systems into the wider ecosystem, it allows your data to appear in many different locations, and be visible wherever people care to look.

And it breaks down the barriers to data sharing and reuse, by allowing data from different sources to be aligned, compared, combined and examined in all sorts of interesting new ways.

Put like that, it almost seems magical. But of course it isn't. It requires effort. And so I have some things to ask of you.

What you can do

Please use metadata standards wherever possible in your own systems, so that you can benefit from the experience and the efforts already put in by others, both here and abroad.

Participate in national and global initiatives, like CESSDA and DataCite. If you don't have national initiatives, get together and see if there are ways of working together for everyone's benefit.

Encourage all the researchers you come in contact with to engage with metadata standards. Okay, as we've discussed, you may not want to force them to hand code XML, but maybe you can make the standards work for them. If they need to package

up their data in DDI Lifecycle, give them a DDI editing programme they could use, and show them how much better their archived data looks as a result.

In other words, give them the training, tools and guidance they need to document their data effectively.

DataCite Metadata Schema v4.0

If that seems too daunting, start with what you know best. Use DDI for your social science data, SDMX for your statistics. It's also a good idea to target DataCite as well; you'll need to if you want to give your datasets a DOI.

If you do, I urge you to go beyond the bare minimum and use the recommended and optional elements. Believe me, the more you can provide the easier your datasets will be to discover. If you are using DDI, you will already know most of this information anyway, so it would be a shame not to use it.

Metadata → better data

These points are probably a good starting point. Whether you are a researcher, a data manager or are setting up a data archive, this is what I want you to take away from this talk…

…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.