Citation for published version:
Ball, A 2012, 'Metadata for Data Citation and Discovery' Paper presented at Describe, Disseminate, Discover: Metadata for Effective Data Citation, London, UK United Kingdom, 5/07/12 - 5/07/12, .

Publication date:
2012

Document Version
Early version, also known as pre-print

Link to publication

© Digital Curation Centre. This work is licensed under Creative Commons Attribution 2.5 Scotland: http://creativecommons.org/licenses/by/2.5/scotland/

University of Bath

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Metadata for Data Citation and Discovery

Alex Ball
6 July 2012

Abstract

The metadata needed to support users working with data is very different according to whether they are citing, discovering or reusing it. Only about five elements are strictly necessary to form a complete citation, whereas a useful generic data discovery service can be built using about sixteen to twenty elements. The metadata needed for reuse varies enormously between disciplines, data types and user communities.

Hello, my name is Alex Ball and I work for the Digital Curation Centre. I expect most people here will have heard of us, but in case not, we’re a centre of expertise in digital curation and research data management funded by JISC. We’re distributed across the three sites of Edinburgh, Glasgow, and Bath where I am based.

It’s a privilege for me to open this workshop with an overview of the issues surrounding metadata for data citation and discovery. It’s an area I’ve long been interested in and something I keep coming back to again and again from different angles.

I’ll start off with some definitions, then explain why we should care about data citation and the metadata used to support it, and then go over some of the metadata elements commonly used to support data citation. Then I’ll do the same thing again but this time for data discovery. Finally I’ll touch on some of the issues and challenges in this area and make some recommendations.

1 Citation, discovery and reuse

Many of you, I’m sure, will be familiar with this diagram (Figure 1) or one very like it. It’s the OAIS model for Information Packages, for SIPs, AIPs and DIPs. Briefly, you have Packaging Information that ties together Content Information – a Data Object and some information to help understand it – and Preservation Description Information. Outside the Package you have Descriptive Information that points to the package; this is the information used in catalogues and other discovery aids.

The reason I’m showing you this is because it helps to place into context the activities we’re discussing today and the metadata we’ll need for the job. What a citation does, fundamentally, is identify an item uniquely so that a reader of a scholarly work can assemble the exact body of literature that the author used. In library terms we’re talking about (transition) known item searches. In OAIS, this is the job of Reference Information. I’m fudging a little here because Reference Information is strictly speaking about identifiers, and we’ll be looking a little wider than that, but it captures the spirit of what we’re after.

If an author or reader does not have a particular item in mind, but is looking for something – anything – to satisfy a particular information need, they perform (transition)
a speculative search, and this is where Descriptive Information comes in. This is the information that tells you what the data is about, what it might be used for, and whether it might be useful in a particular research context.

In order to actually use the data, though (transition), the researcher is going to need the Data Object itself, the Representation Information that explains what it means, and perhaps some information that explains how it relates to other resources.

What we see from this is that the three tasks of citation, discovery and reuse require quite different sorts of information, and they get more detailed and specific as we go from one to three. In OAIS, Representation Information is always relative to a particular community, their conventions and shared knowledge, so discipline-specific metadata is going to be massively important here. Reference Information, on the other hand, is utterly generic: all you’re worried about is uniquely identifying the resource. Descriptive Information is somewhere inbetween: some of it will be generic, some will be more appropriate to some disciplines than others.

We’re not so worried about the third layer here, vital though it is, but just the first two. So, taking things in order, why do we need to support data citation?

## 2 Data citation

### 2.1 Motivation

To understand that, we have to think about scholarly communications more generally. Journals are the big success story in this area, but what made them so great (Figure 2)?

They provided a way of communicating research results such that others could verify the results and build on them, while also ensuring authors received due credit, and in time rewards, for their work. Formal publication also meant formal archiving could take place. The fact is, though, that as the process of conducting research has become more specialist and complicated, your average scientific journal paper can no longer contain all the information it needs to, to make the research reproducible (transition); we also need the underlying data. But we won’t get data routinely shared until all these things apply to
data as well as to journal papers. I would argue (Figure 3) that, given time, data citations are what will make it happen, because the citation model is well understood and trusted. But it won't happen unless the right metadata is in place.

Figure 2: What’s great about journal papers?

- Awareness raising
- Protection from plagiarism
- Verification of results
- Basis for future research
- Reward models
- Permanent access

Figure 3: What data citations provide.

2.2 Metadata

What is the right metadata? Well, here are four standard data citation styles I found in the literature.


There are others. I’m not so much interested in the format, as citation formats as are varied as snowflakes. What I’m interested in are the elements they have in common. Which elements do they use?

Author, Publication date, Title, Version, Feature (ISO 19101), Resource type, Publisher, Identifier, Location, Unique Numeric Fingerprint.
Altman and King (2007): Dataverse


Lawrence et al. (2008): BADC


Green (2010): OECD


Starr and Gastl (2011): DataCite


There are five elements that occur in all four styles, four of which have a long pedigree in scholarly citation:

- Author
- Publication date
- Title
- Location (= identifier)

Despite the fact that we’ve had ISBNs since 1970 and online catalogues in widespread use since the mid-1980s, identifiers didn’t really start to catch on in citations until the introduction of DOIs in the last five to ten years. I’d guess this is because with things like ISBNs there is no central register that allows you to look up the item; booksellers and libraries have had to build them up for themselves. So identifiers have tended to be used, if at all, more like checksums for making sure you had the right item, rather than as a way of accessing resources.

But the Web is changing all that. We now have ways of making locations persistent enough to be used as an identifier (transition). While it’s possible to do this by carefully managing URLs, it’s more usual to achieve it by using a fake location, made up of a resolver service and the identifier, that redirects to the real location. DOIs are getting the most traction for datasets that are considered ‘published’, with Handles and ARKs being used more for ephemeral datasets.

3 Data discovery

3.1 Motivation

Let’s turn now to the matter of data discovery. The motivation for data discovery services is rather more straightforward. I’ve listed on the slide here (Figure 4) some additional benefits of data sharing.
Needed to realise some benefits of data reuse:
- Maximum return on funder investment
- Reduced duplication of effort
- Reduced costs of data collection
- Broader scope of possible research

Figure 4: Motivation for a data discovery service.

Funders get more value for money if data collected by one project can be used by another to yield additional results. If data is shared, no-one will need to go and collect it again unless there’s a problem with it. The costs of data collection go down if researchers can learn from previous studies which techniques yield the best results. Lastly, having the data out there means it is available for use in comparative studies and data mining within a discipline, and for combination with other data in interdisciplinary studies. None of these benefits can be realised unless researchers can find the data they are interested in, and that means there needs to be a robust data discovery service in place, and the right metadata to power it.

3.2 Metadata

So what are the right metadata? This is an area where disciplinary differences start to become important, so let’s narrow down to, say, geospatial and environmental data (Figure 5).

Figure 5: Map of metadata standards for geospatial/environmental data.

Here’s a concept map showing a selection of metadata standards and profiles used in geospatial data and the mappings between them. These metadata standards provide both descriptive information for data discovery and representation information to help understand the data. One way of sorting out which is which would be to look at the search interfaces that are actually used for this kind of data.
Here’s (Figure 6) the search interface for the Earth System Grid. As you can see, it uses twelve facets, plus some additional fields such as funder, date, and related datasets and publications.

![Figure 6: Earth System Grid (http://www.earthsystemgrid.org/search.html).](image)

The NERC Data Catalogue Service (Figure 7) lets you search by Date, Geographic area or through a text search of fields such as Author, Data Originator, Data Format, Parameter, or Topic Category plus the ones you see on the screen (abstract, title, conditions of use/access).

![Figure 7: NERC Data Catalogue Service (http://data-search.nerc.ac.uk/).](image)

Now for an example of a generalist data catalogue. Research Data Australia (Figure 8) is the catalogue maintained by the Australian National Data Service. On the screen you’ll see we have title, abstract, geospatial and temporal coverage, subject, identifier, access and usage conditions, and ‘connections’ to researchers, activities, related datasets and
so on. You can’t see it on this screen but there’s also faceted search by research group, subject, and content type.

So that gives you a flavour of the kind of metadata currently being used in data discovery services. A few years ago now I compared fifteen metadata schemes used by data archives in the UK and extracted the elements they had in common. And the list I came up with was this (the numbers brackets show how many schemes shared the element).

**Identification**
- Dataset Name (15)
- Dataset Version (4)
- Dataset Date (13)
- Dataset Identifier (12)
- Metadata Scheme Name (7)
- Metadata Scheme Version (5)
- Metadata Record Date (10)
- Metadata Record Identifier (3)

**Responsibility**
- Project/Study/Series Name (9)
- Project/Study/Series Status (4)
- Rights/Restrictions (14)
- Agent (15)
- Agent Contact Details (11)

All the fields you’d expect are there for identifying the data set. Under ‘Responsibility’, Project Status refers to whether it is ongoing or completed. Under ‘Agent’ I grouped all the people and organisations involved: funder, principal investigator, co-investigator, project manager, data collector/creator/contributor, researcher, technical contact, distributor, etc.

---

Archiving

- Location (15)
- File Format(s) (10)
- Storage Medium (6)
- Size (7)

- Data Quality Information (5)
- Data Preview (4)
- Dataset Language (7)
- Dataset Status (8)

Again, Dataset Status refers to whether the database is planned, in progress or complete. Data Preview is a thumbnail visualisation of the data, most often used for geospatial data, as are the extent and resolution of the spatiotemporal coverage.

Spatiotemporal Coverage

- Spatial Extent (12)
- Spatial Resolution (7)

- Temporal Extent (15)
- Temporal Resolution (5)

Topical Coverage and Derivation

- Dataset Type (12)
- Subject/Keywords (13)
- Abstract/Summary/Description (14)
- Parameters Used (6)

- Methodology/Instrumentation (8)
- Processing Steps (6)
- Related Datasets (11)
- Derived Publications (11)

Dataset Type could be used either to distinguish datasets from other types of resource like text documents or sound recordings, or to distinguish simulation data from experimental data, or -omic data from non-omic data, say. Parameters refers to measured or controlled variables. Processing Steps includes software tools used in processing.

Restricting this to the elements shared by at least two-thirds of the scheme, we’re left with these.

- Dataset Name (15)
- Dataset Date (13)
- Dataset Identifier (12)
- Metadata Record Date (10)
- Rights/Restrictions (14)
- Agent\(^1\) (15)
- Agent Contact Details (11)
- Location\(^2\) (15)

- File Format(s) (10)
- Temporal Extent (15)
- Dataset Type (12)
- Abstract/Summary/Description (14)
- Related Datasets (11)
- Derived Publications (11)

1. DataCite uses Creator, Publisher, Contributor
2. DataCite holds this separately
3. DataCite also has Dataset Version (4), Dataset Language (7), Size (7)

Notice we have the five citation elements included in this list (transition). One thing I
didn’t notice until preparing these slides is that there’s a close correspondence between these and the four elements that were in all fifteen metadata schemes (transition). Comparing this list with the search interfaces we had before, you can see that despite disciplinary differences, it provides enough information to power a useful generic data discovery service. In fact, if we compare it to the DataCite metadata schema version 2.2 (transition), you can see it has a large overlap with our top sixteen.

4 Issues and challenges

So, at a broad level I think the guidance we can glean from all this is pretty clear. The devil is, of course, in the detail.

Take the author, for example. Authorship of a dataset is a strange concept. More natural roles might be a compiler, or a principal investigator, or a corporate owner. Furthermore, it is far easier to rack up a silly number of contributors with datasets than with textual publications, though in some disciplines they have a good go. If we’re going to use citation data in metrics of researcher impact, most likely we’ll need some sort of microattribution approach (Figure 9).

Figure 9: Microattribution table (http://dx.doi.org/10.1038/ng.785).

This spreadsheet was submitted as part of the supplementary data for an article published in Nature Genetics last year. You’ll see it attributes each genetic variation in the dataset to its contributor, as identified by a Thompson Reuter ResearcherID (other contributor ID schemes are available). This was very much a proof of concept. In future we might hope for this sort of information to be made available as linked data, preferably somewhere more accessible than supplementary data, like DataCite’s metadata store.

The other issue I want to talk about is dataset identifiers, and how they should be applied to dynamic datasets. There are two ways a dataset can be dynamic (Figure 10). The first (animate) is where the dataset is fairly stable in its extent, but points are revised every so often. A table of the masses of subatomic particles would fall under that category.

The other, more common case (animate) is where a dataset is continually expanded with new data, such as with sensor data.
There are three ways of making such datasets citable.

1. Differentiate versions by access date rather than ID

2. Take time slices

3. Take snapshots

The first option I know is adopted by the National Snow and Ice Data Center in the US, because first, in the disciplines they serve the dataset itself is more important than the version, and second, the Federation of Earth Science Information Partners of which they are a part believe that the identifiers they assign aren’t identifiers at all but locations, because you can resolve them to addresses.\(^2\) It’s not a view I share, and so I’m not keen on this option. The second approach really only makes sense with expanding datasets, and even then works best if the researchers tend to use one slice of the set at a time. Even so, it is possible to combine it with the first approach, or the third one which is the one I reckon is most generally suitable; if the rate of change is particularly frequent, it would probably be best to take these snapshots on demand rather than at predefined intervals.

The apparent downside of the third option is that it seems to involve massive duplication of data, but there’s nothing to stop the data backend generating these snapshots on the fly from a single master sequence.

## 5 Recommendations

There’s plenty more I could go on to talk about, but it’s about time for me to wrap up with some recommendations about metadata for citation and discovery.

For citation purposes, I recommend collecting the following metadata.

\(^2\)http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines#Note_on_Versioning_and_Locators
• Author
  – Record roles, identifiers and contact details as well as names.

• Publication date

• Title
  – It helps to avoid confusion if this is different from the article title.

• Location/Identifier
  – Express identifiers in location (http) form if possible.
  – Use DOIs for ‘published’ data and Handles, ARKs, PURLs, etc. for changeable, inaccessible, or draft data.
  – Use different DOIs for different versions.
  – Location given should be a catalogue record/landing page for the data.
  – Giving the Publisher or Host Archive as well provides some recourse if the link breaks.

For discovery purposes, in addition to the citation metadata, these should also be recorded where they apply:

• Contributors
• Abstract/Summary/Description
• Subject/Keywords
• Rights/Restrictions
• Spatial Coverage
• Temporal Coverage
• Derived Publications
• Related Datasets

• Resource Type
• File Format(s)
• Important Dates
  – Creation, Submission, Acceptance, Use By…
• Language*
• Version*
• Size*
• Metadata Record Date

I’ve re-ordered these into approximate descending order of usefulness (down then right). Again, for Contributors, record the roles, identifiers and contact details as well as names. Under rights and restrictions, a standard licence is best, but that’s a whole other talk. The ones with asterisks are in the DataCite Metadata Schema but not my top sixteen. Ideally language shouldn’t matter with datasets, but it can be an issue if the Representation Information is in an unfamiliar language, or if the data is largely textual. I don’t think anyone will search by version number, but it’s nice to know. And you need size next to your download link, but probably not in your search interface.

Alex Ball. DCC/UKOLN, University of Bath. http://www.ukoln.ac.uk/ukoln/staff/a.ball/

Except where otherwise stated, this work is licensed under Creative Commons Attribution 2.5 Scotland: http://creativecommons.org/licenses/by/2.5/scotland/

The DCC is funded by JISC.
For more information, please visit http://www.dcc.ac.uk/