Metadata for Impact

Lessons from the UK

Alex Ball

25 September 2014

Abstract

In the UK, Jisc is developing a national research data registry and discovery service. The service will be powered by metadata harvested from long-established data centres and the newer data repositories set up by universities. A pilot version helped to identify the challenges arising from combining metadata of various levels of detail and quality, leading to a preliminary set of recommendations for discovery metadata.

Greetings…talk about metadata quality from the perspective of a national-level aggregator. The UK doesn't currently have a research data discovery service, but I'm part of a team trying to build one. The name we have for it is the Jisc Research Data Registry and Discovery Service. ¶

While there's not much to show at the moment, I can at least tell you a little of why we're building it, what we're hoping to achieve, and what we've learned so far from trying to harvest and normalise metadata from a diverse range of organisations.

Contents

Acknowledgements

Project team Kevin Ashley, Alex Ball, Patrick McCann, Laura Molloy, Veerle Van den Eynden

Funder Jisc

1 Vision

The first thing to stress is that we are focusing on *UK research data*. There is some work we have to do to define what we mean by that, but so far we've found it convenient to interpret it as data that can be found in UK data repositories. That means we're harvesting from discipline-specific data centres, institutional data repositories, and in due course CRISes.

§ The second thing is that this is a *discovery service*, not a super-repository. We will not host any data ourselves, but will make data as visible as possible wherever it happens to be. . . . important for long-tail research, and for cross-disciplinary work; relevant data might be scattered across 170 UK HEIs and a dozen data centres, and no-one wants to search through them all individually.

§ Ultimate goal is to encourage *data sharing and reuse*, so that the data has maximum *impact*, researchers get due *credit* for it, funders get better *value for money* and we have a more complete and higher quality scientific *record*.

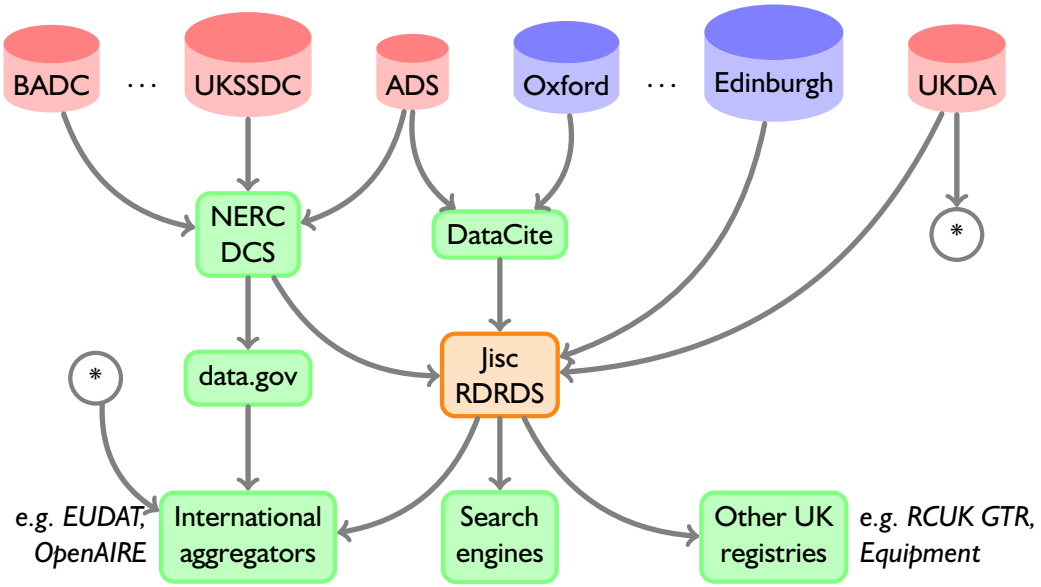

Figure 1: Place of the RDRDS in the repository landscape. Note that many repositories will also contribute directly to international aggregators.

¶ Of course, there are already services doing similar things, not least B2Find, but we have the potential to complement rather than compete with them (see Figure 1), 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.

2 Progress so far

How far have we got? We've completed Phase 1, in which we built a pilot registry to give us a better idea of the technical challenges and stakeholder requirements.

- Registry based on ORCA (Research Data Australia)
  - *we knew it worked well in Australia;*
  - *we at the DCC were already working with the developers to make the code more portable;*
  - *we liked the way it was search engine friendly, and provided citation and rights information up front.*
- Participating data repositories:
  - 9 universities: Edinburgh, Glasgow, Hull, Lincoln, Leeds, Oxford, Oxford Brookes, St Andrews, Southampton
    All of these had or were developing data repositories, and had data records we could harvest.
  - UKDA
  - Archaeology Data Service
  - 7 NERC Data Centres: BADC, BODC, EIDC, NEORC, NGDC, PDC, UKSSDC
    These gave us a wide disciplinary range, plus we could cheat a bit because all except the UKDA contribute to the NERC Data Catalogue Service.
- Crosswalks written *to* RIF-CS *from*
  - DataCite
  - DDI Codebook
  - EPrints (native + ReCollect)
  - MODS
  - OAI-PMH Dublin Core
  - UK Gemini 2
    These six were enough to harvest from all 18 repositories.

Finally, we set up accounts for all the participating repositories and asked them to try importing records into the registry. Which they did.

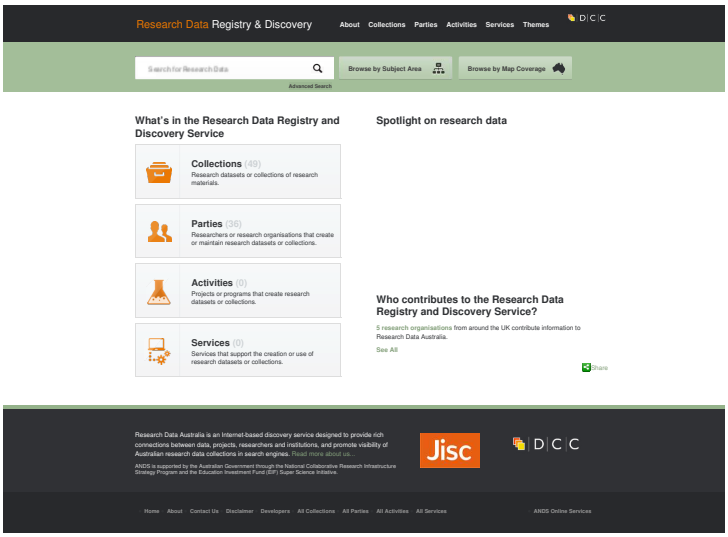

Figure 2: The pilot UK Research Data Registry and Repository Service, URL: `http://rdrds.cloudapp.net/`

¶ So here is what the pilot registry looks like (see Figure 2). There are still a few remnants of Research Data Australia . . . more work to do on making the code portable. . . . If you do visit it there isn't much to see as the participants haven't their records public. This is a feature of the system. Contributors get a chance to review the quality of the imported dataset records before making them public.

¶ They can check them manually, but to make things easier to see at a glance, the software has some automated quality checks that ensure that records have enough useful information. Quality Level 1 is basic: it checks that the incoming XML is valid.

Quality Level 2

- title
- description
- location (e.g. URL)
- IPR statement
- related party, e.g.
  - P.I./researcher
  - manager

Quality Level 3

- identifier
- citation information*
- subject
- date (e.g. of publication)
- spatial coverage
- temporal coverage
- related activity

* Such as 'publisher'; other relevant fields are already mentioned.

¶ This (Figure 3) is a record from the BODC imported using the UK Gemini crosswalk . . . Lots of opportunities for related item searches . . . *author* . . . *subject term* . . . *data centre* . . .

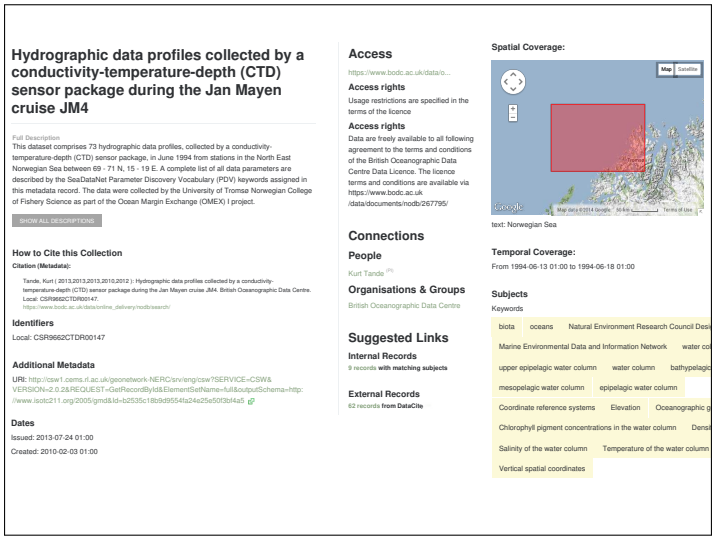

Figure 3: Sample record in the pilot registry

A note about the link to additional metadata. The metadata we present here is aimed at discovery. It is not suited to data reuse. The idea of this link to allow quick access to the detailed metadata that would be needed in order to use the data. But we can only provide that link if that metadata exists, and Jane Greenberg and myself are co-chairs of an RDA working group which has as one of its aims to help researchers collect and express that metadata in a standard appropriate to their discipline. You can look us up on the RDA website, under the Metadata Standards Directory Working Group.

We're now at the beginning of Phase 2, in which we'll develop this into a full service. ¶…Note the penultimate item in this list:

- Define a set of clear use cases and workflows. *These will be grounded in requirements gathered in Phase 1 and the early part of Phase 2, and will determine which functionality the service will and will not support.*
- Compare different possible platforms for the service and assess their suitability*, both in terms of the initial development of the service and its longer term maintenance*.
- Establish a working instance of the system, involving all UK data centres and university data repositories. *This will require a programme of engagement with the data providers.*
- Establish a simple workflow for adding more data sources to the service, adapting to changes in existing data sources, and avoiding duplication. *This will require a plan for continued engagement with data providers.*
- Test the system for usability. *This will require engagement with both data providers and potential users of the system.*
- *Produce recommendations for quality and standardisation of metadata records*.
- Evaluate the costs and benefits of the system. *This will inform planning for the long term sustainability of the service.*

Standardisation might be too strong a term, but we will need a certain level of consistency in the incoming metadata in order to provide a good service.

3 Metadata elements

So, getting down to the meat of the thing now, one of the key work packages of this project involves metadata. While we can't guarantee the metadata scheme we'll end up using in the finished service, or what forms we'll be able to accept, the pilot work has at least flagged up metadata that we're likely to need.

Title – *We want this for heading up dataset records, for listing in search results, and for sample citations.* We found few problems here, except

- some records did not provide one;
- some titles were duplicated because they did not mention which subset was included;
- some might have to be redacted if they contain sensitive information.

Description/Abstract – *Potential reusers need this in order to decide whether the dataset will be useful for them.*

- These were generally pretty good.
- There was variety in level of detail *but that's OK*.

Dataset identifier – *This is useful in searches, and is prominent in the dataset record and sample citations. It is also very important for deduplication, if we receive records about the same dataset from different sources.*

- Most places could provide a local ID.
- A handful supplied DOIs.
- One also supplied `<identifier type="citation">` *which suggests to me someone didn't understand the difference between identifying something and making reference to it.*

Subject – *Again, this is useful for deciding if the dataset is relevant, and also for finding similarly useful datasets whether by searching, browsing or clicking through. For this reason, use of a controlled vocabulary is preferred.*

- Most places could provide subject or topic terms.
- Some specified the scheme used.
- It was hard to transform these into our list of scheme identifiers.

URL of landing page – *This is used in the sample citation, and is used to facilitate access both to further metadata and the data itself.*

- Best case: derived from ID.
- Some provided multiple URLs: which to choose? *This was notably the case with the records from the NERC DCS; the typology of URLs used there was not applied wholly consistently, and in the system we were using we couldn't readily apply our own distinctions.*

Creator (+ID) – *This refers to the person or group that would go in the author slot of a citation. Information about other contributors will be welcome too. The pilot showed up some issues here…structured (forename, surname)…ID to avoid false (de)duplication.*

- In one case, only provided within a citation.
- In another, provided as '`FirstName,LastName`' contrary to the specs. *It should have been '`LastName, FirstName`'.*
- No IDs supplied: big trouble with duplication. *Depending on the approach we took, we either duplicated an author every time they appeared, missing the point of browsing connections between databases, or we combined records from everyone with the same name within the same harvest, risking attributing datasets to the wrong people. We're hoping that ORCID will solve these problems for us.*

Release date – *This is used in the sample citation and can be used for search too.* Various dates were supplied that could be used:

- published

- issued
- available

DataCite has all three. Should we choose? Take them all?

Rights information – *We expect to display this prominently as it is a vital factor in reusability.* Various types:

- access instructions
- access or usage restrictions
- licence statement
- licence URL

Not always easy to categorize them.

Spatial coverage – *This is a vital property for geospatial data, so we expect to display it and provide it as a search parameter.* We only had a problem with `oai_dc` records – `dc:coverage` can contain

- coordinates of different sorts
- place or region names
- date or date range
- period name

Temporal coverage – *This is another property that's useful for tying together diverse datasets, so we expect to provide it as a search parameter.*

- Not common, but usually consistent where provided.
- Only a problem in `dc:coverage`.

Publisher – *This is used in the sample citation, and might be used by researchers as an indicator of quality and as a backup access route.*

- Tend to generate this from holding repository.
- Rarely provided explicitly.

4 Conclusions

- Ideal source format: one where there is only one right way of doing things!
- Need for identifiers all round:
  - datasets
  - people

  - organisations
  - subject vocabularies
  - subject terms
- People provide higher quality metadata if they see the effect it has.

On that last point, I have seen it happen with institutional repositories: a researcher sees his colleague is getting more downloads because she has provided better metadata, and resolves to put in more effort next time. We've seen it with schema.org: web managers see the positive effect that providing that metadata has on the display of their site in Google search results, and they make the effort to provide it. I think if nothing else, initiatives like EUDAT and the Jisc RDRDS will have a really positive influence on metadata quality because the metadata will affect the image of researchers, institutions and data centres in these services and make a difference to the impact of datasets.

Indeed, we've seen this happening already with the RDRDS. The UKDA has been providing access to its holdings via OAI-PMH for some time now, and as part of that provided DDI records in XML. But while they'd moved on internally in terms of metadata versions and identifiers since that was set up, the XML feed hadn't. So they were still providing records in DDI Codebook 2.1, and missing out some information. As a result of their involvement with this project, which if nothing else proved that people might actually use be using that XML, the UKDA has updated and improved its feed so that it now uses DDI Codebook 2.5, includes DOIs for the datasets, and provides Semantic Web-friendly identifiers for the terms in its HASSET thesaurus.

It just goes to show that documenting datasets, even at the discovery level, can be a hard and thankless task if you're working in the dark. But if you can see the metadata being used in a system like the registry, and you can see what that means for the potential impact of the dataset, it suddenly becomes more rewarding to do a good job, and everyone wins.

Alex Ball. DCC/UKOLN, University of Bath. `http://alexball.me.uk/`

The DCC is supported by Jisc.

For more information, please visit `http://www.dcc.ac.uk/`

Jisc RDRDS Project: `http://www.dcc.ac.uk/projects/research-data-registry-pilot`