Building a Research Data Registry for the UK

Alex Ball[1] Kevin Ashley[2] Patrick McCann[3] Laura Molloy[3]
Veerle Van den Eynden[4]

12 June 2014

Abstract

The UK Research Data (Metadata) Registry (UKRDR) Pilot Project, funded by Jisc, is implementing a prototype registry for the research data assets created by UK researchers. The vision for the registry is to allow a single search to query the holdings of all UK data centres and institutional data repositories at once. We anticipate this will be useful not only for researchers working in disciplines without a dedicated data centre, or across disciplines, but also for funders and university administrators to track the outputs of research projects. In the first phase of the project, we set up a prototype registry using software written for ANDS' Research Data Australia, and used it to harvest metadata from nine data centres and nine institutional data repositories. We reflect on this experience, and what it tells us about the feasibility and value of developing the prototype into a full service.

Greetings… talking about project to build a Research Data Registry and Discovery Service for the UK. The service will allow people to find the data they are looking for regardless of whether it is held in a data centre like the UK Data Archive, or a data repository run by an institution such as the University of Glasgow.

Contents

[1]DCC/UKOLN Informatics, University of Bath, [2]DCC, University of Edinburgh, [3]DCC/HATII, University of Glasgow, [4]UK Data Archive.

1 Motivation

Figure 1: UK data landscape

In the UK we have a long history of working with data in a subject-specific fashion. The slide (Figure 1) shows a selection of our national data centres. They are convenient for researchers because it means there is a central place they can go to find data related to their subject.

But there are some emerging trends that this system is not well placed to cope with.¶

- Many research funders now require that data they have paid for be shared if possible. But not all datasets have a natural home among the data centres we have. So to fill that gap, UK universities are setting up their own data repositories.
- In parallel with this, many journals are asking authors to make their data available from a repository, and again, where subject-specific ones are lacking, there is a place for more generalist ones like Dryad or figshare.

  So searching efficiently in some disciplines requires a cross-search service.
- Also, in the past decade or so we have seen a rise in multidisciplinary and interdisciplinary research, and that often requires data from multiple disciplines to be combined and cross-referenced in new and interesting ways. So that's why we're seeing things like the INSPIRE initiative, to enable all of Europe's geospatial data to be searched at once.§
- Since datasets are gaining recognition as scholarly outputs, funders, assessors of research quality and universities are becoming more interested in tracking their impact.§
- Lastly, there are disciplines that would benefit from a greater pool of shared data, but researchers are reluctant to share because they don't believe the data will be reused, or they fear that if it is, they won't get due credit. To break this cycle, we need to make it easier to find data, cite it and measure its impact.

All of which points to the need for a portal that can search across all the data centres and institutional data repositories at once. And that is exactly what our proposed national data registry is supposed to achieve.

2 Previous work

We weren't the first to come to this conclusion by any means.

- ¶ INSPIRE is the Infrastructure for Spatial Information in the European Community, and provides a GeoPortal for discovering and accessing geospatial data.
- ¶ The NERC Data Catalogue Service is a step further back in the chain, providing a single access point for data held in Natural Environment Research Council's data centres.
- ¶ The CESSDA Catalogue allows one to cross-search all the social science data archives across Europe.
- ¶ Coming on to more general initiatives, OpenAIRE provides a portal for cross-searching participating European repositories. It uses the DataCite metadata schema for harvesting dataset records.
- ¶ This is something it has in common with the DataCite Metadata Store, which holds records for all datasets with a DOI.
- ¶ Lastly, Research Data Australia provides a portal for discovering research data from Australian institutions.

For our version, Jisc wanted to test the feasibility of the service before going any further, so they commissioned a six month project to set up a pilot registry. There wasn't time to perform a meaningful assessment of the various software platforms on offer, or to develop our own. So we needed to find a system we were confident could do the job, and run with it.

As it happened, the folks at the Australian National Data Service were in the process of opening up ORCA, the software behind Research Data Australia, and were looking for testers.¶

Now, we knew it was working well for Australia, and we at the DCC were familiar with the code base because we'd already done some early testing, but there were some other things we liked about the software. One thing is that it the records in the system are designed to show up in search engine results, making the datasets highly visible where researchers are already looking. Also, the system supports the idea of data as a first class research output by providing sample citations, and encourages re-use by making explicit any licence conditions or access restrictions so people know where they stand, up front.

The only thing that did make us pause was that it stores metadata in RIF-CS. RIF-CS is a metadata standard, it's a profile of an ISO standard in fact, and it is supported by all Australian universities. You could think of it as a very simple CERIF, and it's actually well suited to this context, apart from the fact that nowhere in the UK supports it.¶ So the plan we agreed for the pilot phase was this:

1. Implement a working instance of ORCA, *noting any difficulties encountered and refinements that had to be made*.
2. Assemble a group of contributors *(data centres and institutional data repositories)* and establish how their metadata *would* be harvested, *both on a technical level and in terms of policies regarding updates, deletions, scope and so on*.
3. Write crosswalks for transforming contributed metadata into RIF-CS *in collaboration with contributors*.
4. Populate the registry with metadata records provided by our contributors.
5. *Then finally, as part of the evaluation of the pilot phase, write* Reports on
   - *our experiences of* using the Research Data Australia software;
   - how harvesting from data centres went;
   - how harvesting from university repositories went;
   - *reflections on* the value of continuing to develop the registry *into a service*.

   (*These reports are being* Made available from *the project website:* `http://www.dcc.ac.uk/projects/research-data-registry-pilot`)

Now I'll take you through these stages a tell you a little of what we've done.

3 Architecture

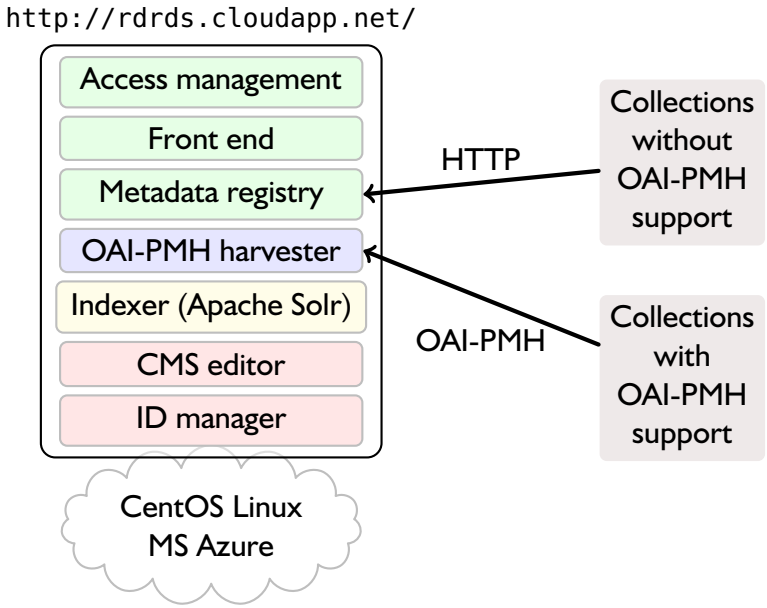

Figure 2: UKRDR architecture

We have indeed got an instance of the ANDS software installed and functioning, in a virtual machine running CentOS Linux in the cloud, specifically Microsoft's Azure service.

Here (Figure 2) is a very simple representation of the components of the registry. The core software is at the top, and the optional extras at the bottom. The code, by the way,

is all open source and available on GitHub, both the original and the adaptations we have made.

The core registry is able to fetch metadata in XML directly over HTTP, but it delegates the handling of OAI-PMH (Protocol for Metadata Harvesting) to a separate module. As it happens, some of our contributors support OAI-PMH, some don't, so we're using both methods.

4 Collaborators

Table 1: Collaborators in the UKRDR Pilot Project

Data centres:

- UK Data Archive
- NERC Data Catalogue Service
  - BADC
  - BODC
  - EIDC
  - NEODC
  - NGDC
  - PDC
  - UKSSDC
- Archaeology Data Service

Universities:

- Edinburgh
- Glasgow
- Hull
- Lincoln
- Leeds
- Oxford
- Oxford Brookes
- St Andrews
- Southampton

We are working with eighteen organisations in total: nine subject-based data centres and nine universities (Table 1).

The data centres were chosen for the disciplinary range they represented, giving us a diverse collection of datasets to work with. It also helped that eight of them already contribute to the NERC Data Catalogue Service, so we're actually harvesting through that rather than going to them all individually.

On the university side, we approached institutions we knew had or were developing data repositories, and the ones you see on the slide (Table 1) are those which had a functioning and populated data repository, and were interested in working with us. There are quite a few other institutions interested in working with us, but they don't have any records to share as yet.

As I mentioned before, the big challenge we faced was that none of these collaborators could provide us with metadata in RIF-CS format. Fortunately, when making the registry code portable, the ANDS developers a had introduced crosswalk functionality. The XML that is harvested (through either method) can be transformed before it is parsed as RIF-CS. So my [Alex's] main role in this project has been to agree and implement crosswalks we can use with all these repositories.

5 Metadata

We've worked on six mappings so far:

DataCite 3

- Archaeology Data Service
- Oxford

DDI Codebook 2.5

- UK Data Archive

EPrints 3

- Glasgow
- Leeds
- Southampton

MODS 3.5

- Edinburgh
- St Andrews
- Hull

OAI-PMH Dublin Core

- Oxford Brookes
- Lincoln

UK Gemini 2.2

- NERC Data Catalogue Service (incl. ADS)

The process we used for writing these crosswalks is this:

1. Match elements in RIF-CS to semantically equivalent elements in the source metadata standard, using both the specifications and sample records.
2. Consult with the collaborators to ensure that the semantic matching is correct, and confirm any syntactic conventions.
3. There are instances where RIF-CS records information in a different way to that that used by the source standards (e.g. repositories, subject terms), so we identify suitable transformations for getting from one representation to the other.
4. Code the crosswalk in PHP and run it on some sample records to ensure it works.
5. Consult with the collaborators to ensure that the RIF-CS version accurately reflects the original metadata record. A useful feature of the registry is that it allows contributors to review the records they have imported into the system before making them live.

¶ And this (Figure 3) shows the crosswalk in action. On the left is a record from the UK Data Archive. On the right is how it looks when imported into the registry. There has been some loss of information: the registry doesn't tell you anything about the methodology, for example. But if you were searching the registry for data on, say, alcohol consumption in 1980, then this record would be flagged up of possible interest and you could click through to the original record for more information. That's because our purpose here is not to replace all those individual catalogues but to make them more visible to those who wouldn't normally go there.

(Note the prominent sample citation and information on access rights.)

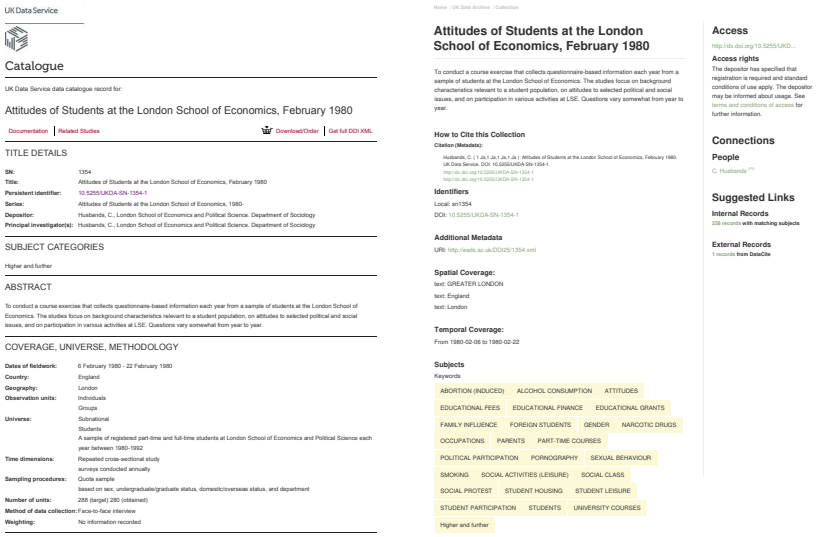

Figure 3: A comparison showing how a dataset record from the UK Data Archive (*left*) appears when imported into the registry (*right*).

6 Evaluation

We are now evaluating the project, and though we haven't finished, here are some early indications:

- Does the software work as intended? *Is the system stable and functional? Are the required basic functions available? Can we import metadata into the system?*

  Can import metadata, manage records, perform searches, etc. OAI-PMH harvest needed work. *ANDS only recently published how to get it to work for non-RIF-CS metadata formats.*

- Do the harvested records look useful and accurate?

  Most records imported at Quality Level 2. *ORCA defines 3 quality levels: 1 means the record is valid; 2 means it contains a title, description, location and IPR statement, and is related to a person or group; 3 means it has an identifier, citation, subject, date and spatiotemporal coverage, and is related to a project. Many records only missed level 3 because they didn't give project information.* No systematic problems so far. . . Contributors motivated to optimise OAI-PMH feeds/DataCite submissions

- Is the system straightforward to use?

  Interface could be more streamlined. *There are a lot of power options exposed, and it can be confusing for new users. Conversely some configuration options are well hidden. Overall more documentation is needed.*

- What might be improved?

  Better to ingest in original format and normalise later? Handling of citation data. *It would be nice if users could choose their preferred citation format.*

- *Are there any* functions *that the registry does not currently provide that* would be desirable *if we were to develop it into a service*?

  Richer set of relations supported? *For example, to data management plans, equipment registries and so on.*

7 Phase 2

- Define a set of clear use cases. *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.*
- Define a set of workflows for using the service. *These will relate to the use cases and form the basis for usability testing and demonstrations.*
- 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.*

8 Conclusions

- Institutional data repositories need to work harder to make their holdings visible than domain-specific repositories. Cooperating with registries and search engines can help!
- When dealing with metadata it is important to focus on the application you are trying to provide. *You can't satisfy everyone's use cases, but if you focus you can satisfy a few.*
- Working with the community is vital to ensure the service is relevant and useful*, otherwise there's no point*.
- A strong business case and value proposition is important for sustainability *so we're taking it very seriously*.

But here is an example of some value we've already managed to add to the community.

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.

And I think that's indicative of the kind of influence we can expect this project to have. 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[1], *Kevin Ashley*[2], *Patrick McCann*[3], *Laura Molloy*[3], *Veerle Van den Eynden*[4]. [1]*DCC/UKOLN Informatics, University of Bath,* [2]*DCC, University of Edinburgh,* [3]*DCC/HATII, University of Glasgow,* [4]*UK Data Archive.*

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`