Show me the data

The pilot UK Research Data Registry

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

2014-02-26

Abstract

The UK Research Data (Metadata) Registry (UKRDR) Pilot Project is implementing a prototype registry for the UKs research data assets, enabling the holdings of subject-based data centres and institutional data repositories alike to be searched from a single location. The purpose of the prototype is to prove the concept of the registry, and uncover challenges that will need to be addressed if and when the registry is developed into a sustainable service. The prototype is being tested using metadata records harvested from nine UK data centres and the data repositories of nine UK universities.

Greetings… talking about the UK Research Data Registry, which is completing its pilot phase as I speak. It is a project to set up a discovery portal for UK research data assets, so people can find the data they are looking for regardless of whether it is held by the UK Data Archive, say, or 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.

Figure 1: Repository landscape

1 Motivation

The UK has a long history of working with data in a subject-specific fashion. We have a network of national data centres catering for environmental science of various sorts, the social sciences, archaeology, the visual arts, and so on (Figure 1). Up until a few years ago, there was no call for searching across all these different repositories. If you wanted oceanography data you'd go to the BODC; if you wanted atmospheric data, you'd go to the BADC.

But that is changing.¶

- More funders require research data sharing, *even those that do not run data centres*.
- . . . EPSRC requires that universities ensure research data is preserved and disseminated, *but does not provide central facilities to support this*.
- . . . Universities need to run their own data repositories *for data without a natural home among the data centres; so if you wanted Engineering data, say, you'd have to search through each repository in turn*.
- Interdisciplinary and multidisciplinary research requires data drawn from diverse sources, *so as such research becomes more popular there is a greater need for discipline-agnostic data searches*.
- Data outputs will contribute to research assessments *(in the UK we have the Research Excellence Framework)*, so need to be tracked *by both funders and university administrators*.

All of which points to the need for a portal that can search across all the data centres and institutional data repositories at once. So Jisc (the body that provides a lot of information systems for UK higher education) decided to invest some money in piloting a research data registry for the UK, and the DCC was given the task of realising it in collaboration with the UKDA.

2 Project overview

There were several different routes we could have gone down, but we only had six months and what we couldn't fail to notice was that Australia already had a working registry (Figure 2).

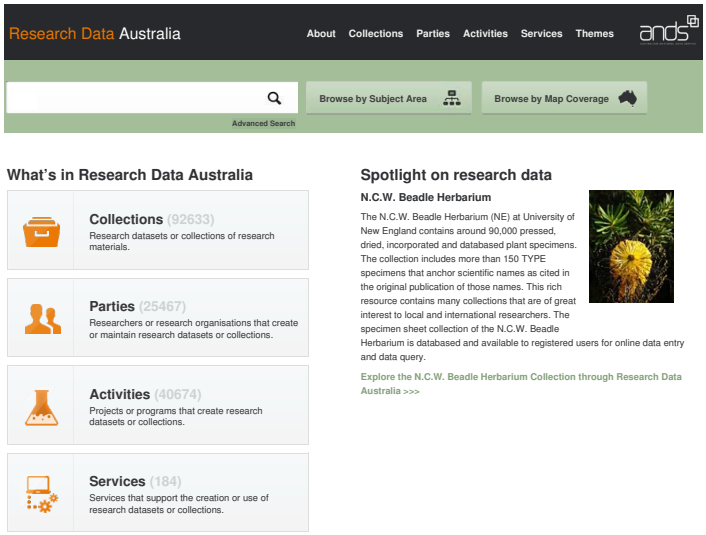

Figure 2: Research Data Australia

In fact, prior to starting the project we had already been working with the Australian National Data Service on making their code more portable, and exploring how to import metadata from Current Research Information Systems. So it was a natural choice for testing the feasibility of a registry in the UK.¶

Attractions of the Research Data Australia software:

- Familiar to project team
- Proven technology
- Plays nicely with search engines
- Displays sample citations and access/rights information up front

Challenges of using the software in the UK:

- Not used before outside Australia
- Uses uncommon metadata standard (RIF-CS) internally
- Original implementation only harvests in RIF-CS
- No UK data centre can output RIF-CS metadata

¶ So our task list for the six month project was this:

1. Implement a working instance of the ANDS software, *noting any difficulties encountered and refinements that had to be made*.
2. Assemble a group of contributors *(data centres and repositories)* and establish how their metadata will be harvested, *from both a technical and policy perspective*.
3. Write crosswalks for transforming contributed metadata into RIF-CS *in collaboration with contributors*.
4. Harvest metadata from contributors.

5. Reports on
    - using the Research Data Australia software;
    - how harvesting from data centres went;
    - how harvesting from university repositories went;
    - the value of continuing to develop the registry.

3 Architecture

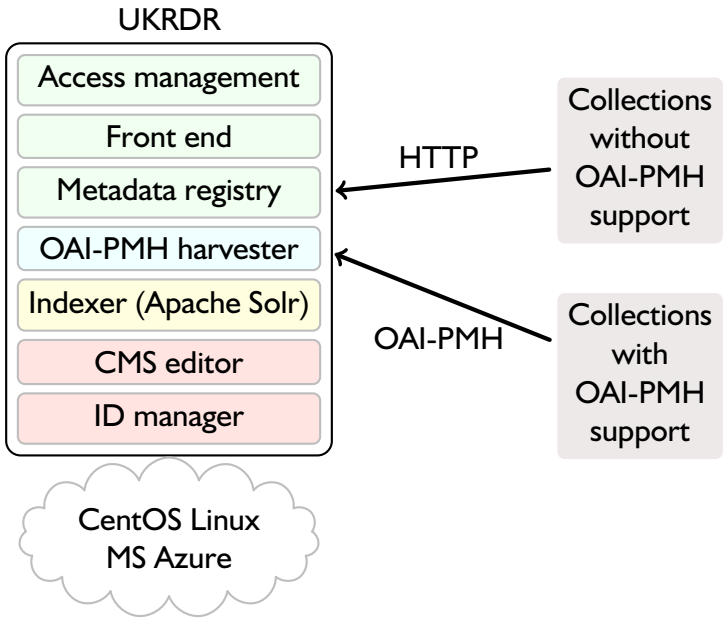

Figure 3: UKRDR architecture

Here (Figure 3) is a very simple representation of the components of the registry. I'll skip lightly over the details here as we haven't changed them that much. I'll just point out that we're hosting it on a CentOS system running in the Microsoft Azure cloud, and we are harvesting data both through directly through HTTP and via the dedicated OAI-PMH harvester module.

The collections that I refer to on the right hand side are a selection of data centres and university-based data repositories.

4 Collaborators

We are working with eighteen 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. On the university side, we approached institutions we knew had or were developing data repositories, and the ones you see

Table 1: Collaborators in the UKRDR Pilot Project

Data centres:

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

Universities:

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

on the slide (Table 1) are those which had a functioning and populated data repository, and were interested in working with us.

We held a conference call with our collaborators early in January to introduce and discuss the project, and we'll hold another one towards the end of March as part of our evaluations. We have also asked them individually some things we will need to know if the registry goes forward into service, and discussed matters of policy and metadata.

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. Put simply, the XML that is harvested (through either method) can be transformed before it is parsed as RIF-CS. So a major part of my time over the past few months has been to agree and implement crosswalks we can use with all these repositories.

5 Metadata

We've worked on five mappings so far:

DDI Codebook 2.5

- UK Data Archive

DataCite 3

- Edinburgh (TBC)
- Oxford (TBC)
- Hull (TBC)

OAI-PMH Dublin Core

- Oxford Brookes (TBC)

UK Gemini 2.2

- NERC Data Catalogue Service

EPrints 3

- Glasgow
- Leeds
- Lincoln (TBC)
- Southampton

The process we are using for writing these crosswalks is…

1. Match elements in RIF-CS to semantically equivalent elements in the target metadata standard. It's helpful if there are some sample records to work from as well as the spec.
2. Consult with the collaborators to ensure that the semantic matching is correct, and confirm any syntactic conventions.
3. Where needed, specify how element values and properties would need to be transformed.
4. Code the crosswalk in PHP using the class provided.
5. Run the crosswalk on some sample records to remove the obvious flaws.
6. 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.

6 Evaluation

Next month we start evaluating the project in earnest. These are the kind of questions we'll be asking:

- 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?*
- Do the harvested records look useful and accurate *compared to the originals, both to us and to the stakeholders*?
- Is the system straightforward to use*, both for harvesting metadata and searching for datasets*?
- What might be improved?
- What additional functions would be desirable?

We'll be answering some of these ourselves and some in conjunction with stakeholders. As befits a short project, the evaluation will be fairly lightweight, but if the registry gets approval for further development we will need something more formal. One thing we're looking at is the ROAMEF model, which is used by the UK Government for policy development.

- ROAMEF = *R*ationale, *O*bjectives, *A*ppraisal, *M*onitoring, *E*valuation, *F*eedback

In order for Jisc to consider funding the registry as a service, we will need to provide them with realistic estimates of the costs and benefits, and some evidence that the approach we've taken is appropriate.

7 Future

Another thing we have in mind for Phase 2 should it go ahead are to answer questions like:

- **Would another platform suit us better?** *CKAN, for example, is well used by government data portals and was recently installed at the University of Lincoln as their data repository.*
- **Would another internal metadata scheme suit us better than RIF-CS?**
- **What use cases should the registry target?** *We do have some in mind already, but it's about committing to some.*
- **How can we add value to the registry's records?** *Pulling in reviews from other sources to give an idea of quality?*
- **Could the registry add value to other systems?** *Perhaps by providing metadata quality checks? As a result of this project the UKDA has updated and improved the DDI metadata it surfaces externally.*

I would like to finish by thanking Jisc for funding this work and you for listening.

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/`

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

This project is funded by Jisc.