Increasing research impact

The national data registry

Alex Ball

11 March 2014

Abstract

Jisc is funding work to develop a national data registry for the UK, so that the holdings of subject-based data centres and institutional data repositories alike to be searched from a single location. The work begins with a pilot phase, testing the feasibility of the registry and uncovering challenges that will need to be addressed if and when the registry is developed into a sustainable service. The pilot project team is working with nine UK data centres and the data repositories of nine UK universities to harvest their metadata and provide a consistent search and browse interface for locating research data.

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 in a data centre like the UK Data Archive, or a data repository run by an institution such as the University of Glasgow.

Project Team

- **Kevin Ashley**, DCC (Edinburgh)
- **Alex Ball**, DCC (Bath)
- **Patrick McCann**, DCC (Glasgow)
- **Laura Molloy**, DCC (Glasgow)
- **Veerle Van den Eynden**, UKDA

Funded by **Jisc**

Contents

1 Motivation

Figure 1: UK data landscape

Here in the UK we have 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). This is a good system in at least a couple of ways:

- It means that the datasets in these repositories are given specialist curation: for example, they are packaged and documented in a way that makes them highly useful for researchers in that field.
- And it means that there is a central place where researchers can go to find data related to their subject. Indeed many national centres contribute to specialist international and global centres to make searching for data even more efficient.

So far, so good. But in the area of research data, there are some emerging trends that this system is not well placed to cope with.¶

- Many research funders, and certainly the major UK ones, 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. I have worked closely with Engineering researchers in the past, as an example, and there isn't an engineering data centre where they could deposit data.

  Engineers, of course, get much of their funding from the EPSRC, and as of May 2015 it will expect universities to take responsibility for making sure that the data outputs of funded research are preserved and visible to the outside world. In effect, they require that universities set up their own data repositories.

- In parallel with this, there are initiatives such as Dryad, which is an increasingly generalist repository that stores data underlying journal papers. While journals might not want to look after data themselves, more and more of them are asking authors to make their data available from some repository or other, and where subject-specific ones are not available that means using a generalist one.

  So we are seeing a proliferation of generalist data repositories of various sizes, and without some sort of cross-search service there will be no way of searching them efficiently, meaning the potential impact of their holdings is reduced.

- So does that mean that we need a two track system, where some disciplines go to their specialist data centre and the rest use the general search portal? Well, no. 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 combined and cross-referenced in new and interesting ways. So there are efficiency gains to be had by making all available research data searchable from a single location. And we're already seeing moves towards that with geospatial data, through services like NERC's Data Catalogue Service.

- Furthermore, we are seeing a shift in attitude away from data being the boring bit of science to data being regarded as an asset that may be mined repeatedly for new insights. This means that funders are interested in tracking the datasets they have paid for and measuring their impact in terms of both primary and secondary use.

  This year's Research Excellence Framework exercise allowed datasets, software and other digital assets to be submitted as evidence to the review panels. So in future we might expect university administrators to take more interest in the tracking the data outputs of their researchers.

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 the proposed national data registry is supposed to achieve.

2 Project overview

I think it is fair to say we have high hopes for what the registry might achieve (Figure 2). There is talk of

- joining it up in some way to the RCUK Gateway to Research,

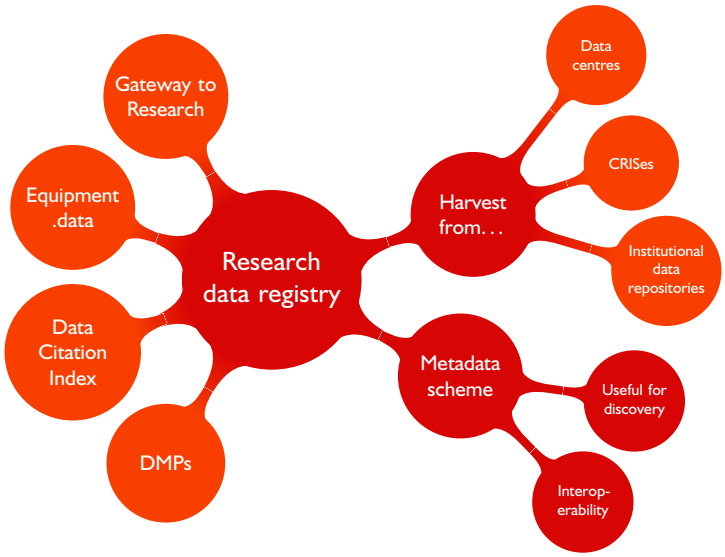

Figure 2: Brainstorming the national data registry

- using it as a short cut for getting metadata into Thomson Reuters' Data Citation Index, and
- connecting data with instruments through equipment.data.

It might also be nice to enhance the metadata records with links to data management plans (DMPs) where these are made available.

Finding the right technical solution to support all of this is quite a challenge, even more so when you consider the diversity of systems from which the metadata would have to be harvested. Plus there is the matter of the scheme used to record the metadata locally. It's a tricky business to find one that both serves the registry use cases and maps well to the various schemes already in use.

So it was decided to split the work into several stages. The first phase would be a pilot phase to test the feasibility of the registry and gain an understanding of the main challenges to overcome. If that worked out alright, then there would be a second phase which would attempt to develop the registry into a service.

As it turned out, we only had six months to complete the pilot phase, and that rather limited what we could do. There wasn't time to perform a meaningful assessment of the various software platforms on offer, or the many metadata schemes that could be used. There certainly wasn't time to develop our own. So we needed to find a system we were confident could do the job, and run with it.

And what we couldn't fail to notice was that Australia already had a working registry (¶ Figure 3).

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.¶

Aside with us being familiar with it, and the reassurance of knowing it was already powering a national data portal, there were some other things we liked about the

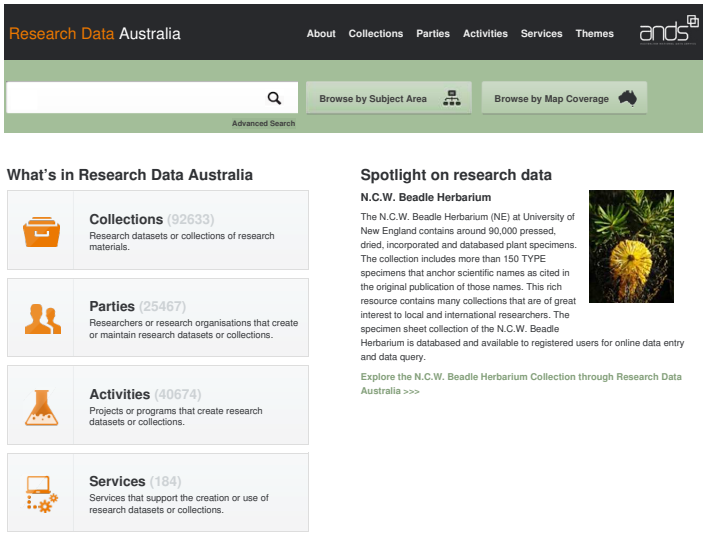

Figure 3: Research Data Australia

software. One thing is that it plays nicely with search engines: the records in the system show up in Google Search results, for example, so even if people are using generic search tools instead of coming to the site direct, they can still find the data they need through the system. Which is good for impact, of course. 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.

But it would be boring if it was all too easy. Aside from us being the first people to use the software outside its original Australian context, there was one major concern we had, and that was the metadata scheme it used. RIF-CS is a standard, it's a profile of an ISO standard in fact, and it is supported by all Australian universities. It is however, supported by precisely none of the UK data centres, and none of the universities who setting up their own repositories. So dealing with that became a major part of the work.¶ Here's the outline of what we agreed for this pilot phase of the project:

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 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*.

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

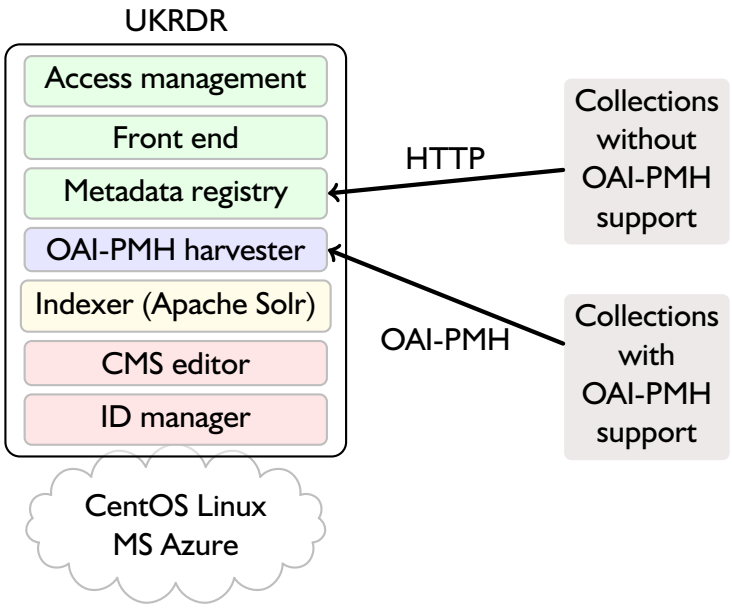

Figure 4: UKRDR architecture

3 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 4) 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. But to explain it a bit, the colour coding shows how the code is organised into modules, with the core software at the top and 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. But who are these mysterious contributors?

4 Collaborators

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

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

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.

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. 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 this:

1. Match elements in RIF-CS to semantically equivalent elements in the source metadata standard. It's helpful if there are some sample records to work from as well as the specification, as you often find places using undocumented conventions when implementing standards.
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, so we identify suitable transformations for getting from one representation to the other.
4. Code the crosswalk in PHP by implementing the interface provided in the registry code.
5. Run the crosswalk on some sample records to check it works and remove any 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.

RIF-CS 1.5 element	Source (using DDI 2.5 record)
collection[dateAccessioned]	codeBook > stdyDscr > citation > distStmt > depDate[date]
identifier[type=doi]	codeBook > stdyDscr > citation > titlStmt > IDNo[agency=datacite]
identifier[type=local]	codeBook > stdyDscr > citation > titlStmt > IDNo ([agency]:content)
name[type=primary] > namePart	codeBook > stdyDscr > citation > titlStmt > titl
name[type=alternative] > namePart	codeBook > stdyDscr > citation > titlStmt > altTitl
dates[type=dc.available, dc.issued] > date[type=dateFrom]	codeBook > stdyDscr > citation > distStmt > distDate
dates[type=dc.dateSubmitted] > date[type=dateFrom]	codeBook > stdyDscr > citation > distStmt > depDate[date]
location > address > electronic[type=url] > value	codeBook > stdyDscr > citation > holdings[URI]
subject[type=hasset]	codeBook > stdyDscr > stdyInfo > subject > keyword[vocab=S]
subject[termIdentifier]	codeBook > stdyDscr > stdyInfo > subject > keyword[vocab=S; vocabURI]
subject[type=ukdasc]	codeBook > stdyDscr > stdyInfo > subject > topClas
description[type=full]	codeBook > stdyDscr > stdyInfo > abstract
coverage > temporal > date[type=dateFrom]	codeBook > stdyDscr > stdyInfo > sumDscr > collDate[event=start, single; date], timePrd[event=start, single; date]
coverage > temporal > date[type=dateTo]	codeBook > stdyDscr > stdyInfo > sumDscr > collDate[event=end; date], timePrd[event=end; date]
coverage > spatial[type=text]	codeBook > stdyDscr > stdyInfo > sumDscr > geogCover, geogUnit, nation; codeBook > stdyDscr > stdyInfo > subject > keyword[vocab=G]
relatedInfo[type=metadata] > identifier[type=uri]	'http://esds.ac.uk/DDI25/' + UKDA ID + '.xml'

```php
<?php

class Ddi2pSToRifcs extends Crosswalk {

  private $oaipmh = null;
  private $rifcs = null;
  private $ddiProviders = array(
    "http://oai.ukdataservice.ac.uk/oai/provider" => "UK Data Archive",
  );

  function __construct(){
    require_once(REGISTRY_APP_PATH . "core/crosswalks/_crosswalk_helper.php");
    $this->rifcs = simplexml_load_string(CrosswalkHelper::RIFCS_WRAPPER);
  }

  public function identify(){
    return "DDI v2.5 to RIF-CS (Experimental)";
  }

  public function metadataFormat(){
    return "ddi_2.5";
  }

  public function payloadToRIFCS($payload){
    $this->load_payload($payload);
    foreach ($this->oaipmh->ListRecords->children() as $record){
      if ($record->getName() != "record") {
        continue;
      }
      $reg_obj = $this->rifcs->addChild("registryObject");
      if (array_key_exists((string) $this->oaipmh->request, $this->ddiProviders)) {
        $reg_obj->addAttribute("group", $this->ddiProviders[(string) $this->oaipmh->request]);
      }
      $key = $reg_obj->addChild("key", $record->header->identifier);
      $originatingSource = $reg_obj->addChild("originatingSource", $this->oaipmh->request);
      $coll = $reg_obj->addChild("collection");
      $coll->addAttribute("type", "dataset");
      $coll->addAttribute("dateModified", date(DATE_W3C));
      $citation = $coll->addChild("citationInfo");
      $citation_metadata = $citation->addChild("citationMetadata");
      $coverage = $coll->addChild("coverage");
      $rights = $coll->addChild("rights");
      foreach ($record->metadata->codeBook->stdyDscr->children() as $node){
        foreach ($node->children() as $subnode) {
          $func = "process_".$subnode->getName();
          if (is_callable(array($this, $func))){
            call_user_func(
              array($this, $func),
              $subnode,
              array(
                "registry_object" => $reg_obj,
                "key" => $key,
                "collection" => $coll,
                "citation_metadata" => $citation_metadata,
                "coverage" => $coverage,
                "rights" => $rights
              )
            );
          }
        }
      }
    }
    return $this->rifcs->asXML();
  }
```

Figure 5: DDI 2.5 crosswalk: (*left*) extract from the element matching table; (*right*) extract from the PHP crosswalk class.

¶ By way of example, here (Figure 5) is the crosswalk to RIF-CS from DDI Codebook 2.5, which is used by the UK Data Archive. On the left is part of the documentation of the crosswalks, showing how the elements match across the two schemes (the full version is available from the DCC website). On the right is a taster of how that looks when coded up in PHP (and that is available from the DCC's account on GitHub).

¶ And this (Figure 6) 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

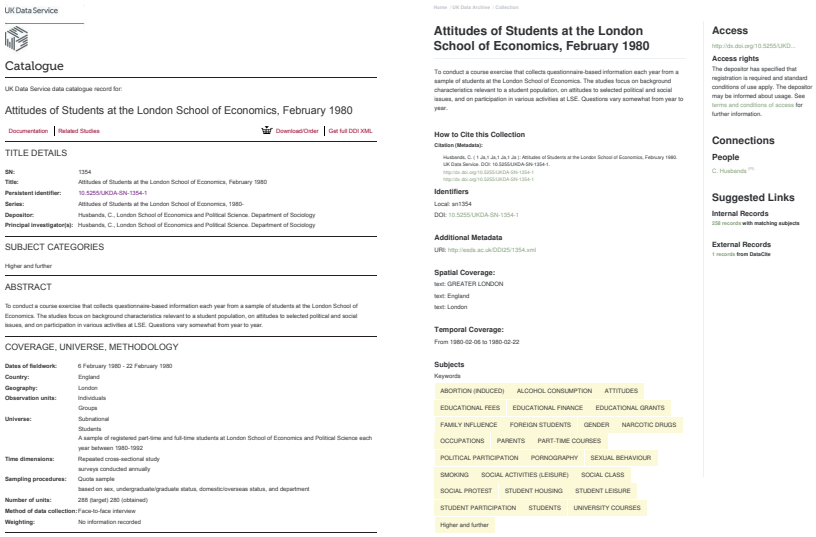

Figure 6: A comparison showing how a dataset record from the UK Data Archive (*left*) appears when imported into the registry (*right*).

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

6 Evaluation

We are now in the midst of evaluating the project, and these are the kind of questions we're 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 *both to us and to the stakeholders*, and accurate *compared to the originals*?
- Is the system straightforward to use, *both for harvesting metadata and searching for datasets*?
- *And finally,* What might be improved? *and*
- *Are there any* functions *that the registry does not currently provide that* would be desirable *if we were to develop it into a service*?

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 for the next phase (which promises to be somewhat longer) we will need something more formal.

7 Future

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. In order to provide that evidence, we'll have to ask questions like these:

- 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? *I have found it quite easy to work with, but it does miss some interesting information and it has a few quirks.*
- 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?*

And of course, those last two questions lead in to thoughts about what other systems we could synchronise with, and I showed you some of the possibilities at the start of this presentation.

But I'd like to leave you with an example of how the registry is already having a positive impact on the data landscape, even though it's not ready yet.

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.

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

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`