# Preparing for the UK Research Data Registry and Discovery Service

Alex Ball

31 July 2014

**Abstract**

In 2013, Jisc funded a project piloting a research data registry for the UK. Following the successful conclusion of that project, Jisc granted continuation funding to develop a fully fledged data discovery service for the UK. This work has only just started, with the service expected to launch in 2016. There are, however, things repositories can do now to ensure they are ready for the service when it launches, such as consider the metadata they collect for datasets, look at how they syndicate that metadata, and above all participate in the development process.

*Greetings*… talk about the Jisc Research Data Registry and Discovery Service… in its embryonic stages… can't give you firm details about what it will look like, how it will operate, or what technologies it will use, but ¶ in the course of this talk I hope to give you a good idea of what the service is supposed to achieve, the work we've done so far, and what repositories can do now to prepare for it when it finally comes online.

## 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 *research data*. Now, that doesn't tell you all that much…type of role, not a type of thing…Typically *not* interested in pre- or post-prints, learning objects or administrative data like the university's student database or finance records. Typically we're looking for collections of evidence underlying a written scholarly output. Many institutions are setting up dedicated data catalogues…we would work with them rather than the regular repository.

§ 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. There are about 170 UK HEIs…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*.

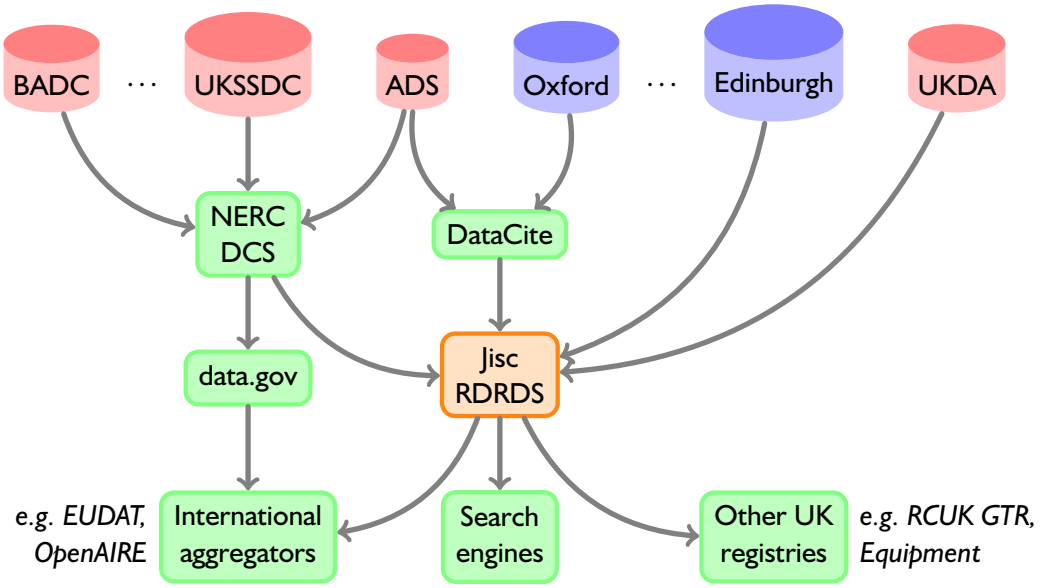


**Figure 1:** Place of the RDRDS in the repository landscape. Note that many repositories will also contribute directly to international aggregators.

¶ There are already services doing similar things, but none have quite the same scope. In fact we have the potential to complement existing services (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

I said at the start, this isn't built yet. While that is true, it's not vapourware either….just starting Phase 2 of the development work. In Phase 1, 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.

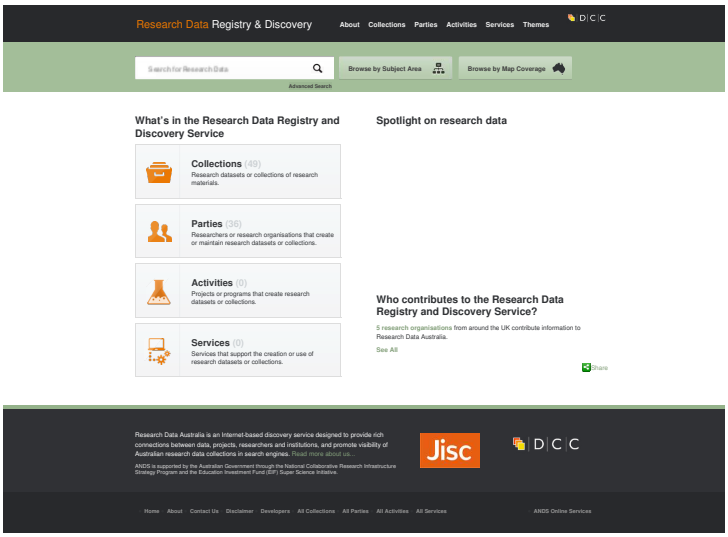


**Figure 2:** The pilot UK Research Data Registry and Repository Service, URL: `http://rdrds.cloudapp.net/`

¶ 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, but with an administrator's privilege I can show you what one looks like.

¶ 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* . . .

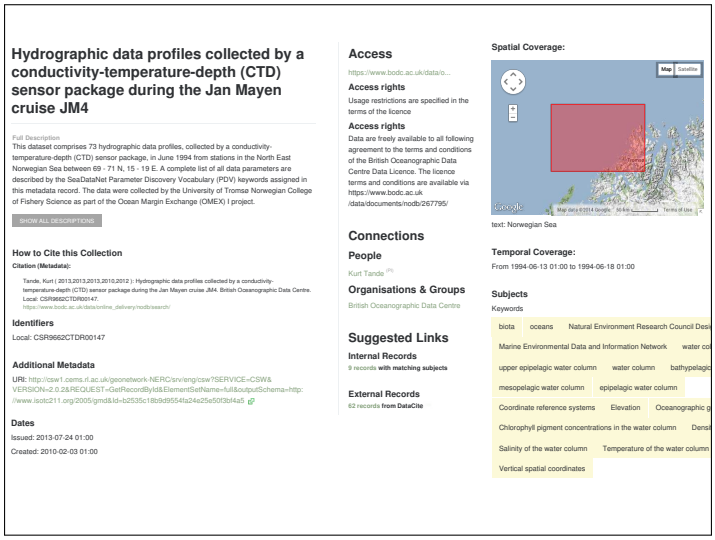


**Figure 3:** Sample record in the pilot registry

This is a good start, but there's much more to do before we can roll it out as a service. ¶

- 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.*

# 3 What this means for repositories

We have a long way to go, and two years to do it in, but in the meantime we're already being asked what repositories can do know to prepare for the service, and I think it boils down to three things: §

- collecting and storing structured *metadata* about your data holdings;
- consider how you will *syndicate* that metadata to the registry;
- *join in* with the development effort.

## 3.1 Metadata

While we can't guarantee the metadata scheme we'll be using, or what forms we'll be able to accept, the pilot work has flagged up metadata that we're likely to need. If you collect it and find a way of providing it in a structured way it should be trivial for you to contribute to the registry.

- Title – *We want this for heading up dataset records, for listing in search results, and for sample citations.*
- Description/Abstract – *Potential reusers need this in order to decide whether the dataset will be useful for them.*
- 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.*
- 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. Use of a controlled vocabulary is preferred.*
- 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.*
- 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.*
- Release date – *This is used in the sample citation and can be used for search too.*
- Rights information – *We expect to display this prominently as it is a vital factor in reusability.*
- Spatial coverage – *This is a vital property for geospatial data, so we expect to display it and provide it as a search parameter.*
- Temporal coverage – *This is another property that's useful for tying together diverse datasets, so we expect to provide it as a search parameter.*
- 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.*

§ Note that if you're thinking of registering DOIs for datasets, the DataCite minimum metadata gets you over halfway through this list, and the remaining elements would also make great additions to a DataCite metadata record.

## 3.2 Syndication

### Method

*We anticipate supporting several methods for harvesting metadata from repositories…*

- OAI-PMH
- CSW
- Atom/RSS?
- Other XML export over HTTP

**Scope**

*But whichever method you choose, it is key that you are able to separate the datasets you have from the other content…*

- Whole data repository
- Items of type 'dataset'
- Specific sets or collections – *OAI-PMH for example supports putting items into sets …how we harvest a single repository's data from DataCite.*

**Metadata format**

*We're not at the stage of recommending one, but* The more detail, the better!

## 3.3 Participation

Get involved by

- letting us know about issues that concern you – *one of our contributing data centres told us it was essential we preserved the citation order of data creators*;
- letting us know when you have data records we could look at *so we can perform a test harvest ourselves*;
- helping us test the evolving system:
  - set up and update an account on the system;
  - harvest your metadata into the system and check it;
  - see if we handle duplicates (and non-duplicates) correctly;
  - see how your records look on the system;
  - see how easy they are to find;
  - measure the visibility of your datasets before and after inclusion*: this will help us demonstrate the value of the service*.

Some of this will be some way down the line, but if you haven't already do get in touch if you are interested in helping us develop this exciting new service.

*Alex Ball. DCC/UKOLN, University of Bath.*

---