IJDC | *General Article*

Show Me The Data: The Pilot UK Research Data Registry

Alexander Ball
DCC/UKOLN Informatics
University of Bath

Kevin Ashley
DCC
University of Edinburgh

Patrick McCann, Laura Molloy
DCC/HATII
University of Glasgow

Veerle Van den Eynden
UK Data Archive

Abstract

The UK Research Data (Metadata) Registry (UKRDR) pilot project is implementing a prototype registry for the UK's 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.

Received 23 October 2013 | *Accepted* 26 February 2014

Correspondence should be addressed to Alexander Ball, UKOLN Informatics, University of Bath, Claverton Down, Bath, BA2 7AY. Email: a.ball@ukoln.ac.uk

An earlier version of this paper was presented at the 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The UK Research Data (Metadata) Registry (UKRDR) pilot project is a six-month project that aims to set up and evaluate a prototype registry of research data held by UK universities and key national and subject-based data centres. The purpose of the registry is to make the UK's research data assets as discoverable as possible, in order to encourage their reuse and to ease certain administrative burdens. The registry will collect and normalize metadata describing the research data but will not hold the data assets themselves. The pilot project will use the prototype registry to assess the feasibility of a full registry service and explore options for its sustainable provision.

The project is funded and steered by Jisc; it is led by the UK Digital Curation Centre (DCC) in partnership with the UK Data Archive (UKDA). The work is being conducted in collaboration with a small group of UK higher education institutions and subject-based data centres.

In this paper we will begin by outlining the motivation for the project. We will then provide a summary of the project and the major decisions that have been taken. We will continue with overviews of the architecture of the prototype registry and how it is being tested. Finally, we will discuss some of the issues we foresee in running the registry.

Motivation

One of the major arguments in favour of researchers sharing their data is that it promotes reuse. Where data is reused, this increases the return on the investment made in collecting the data, and when combined with other data sets provides opportunities for new research questions to be answered. But making the data available is just one step towards this: other researchers must be able to find and trust the data before they can make use of it.

There are several union data catalogues already in operation that allow multiple repositories to be cross-searched. One example is Research Data Australia,[1] set up by the Australian National Data Service, which provides a unified interface for searching the data holdings of Australian universities. It also provides information on researchers and projects producing the data.

Another is the DataCite Metadata Store. When a dataset is given a DOI through DataCite, a set of metadata describing it is added to the store. It is possible to query this metadata through the DataCite Metadata Search interface.[2]

Until recently, the need for a generalist union catalogue for data was not keenly felt in the UK. The UK has a long tradition of discipline-specific data centres funded by individual research councils to preserve and disseminate research data, making those available for further research:

- The Economic and Social Research Council (ESRC) funds the UK Data Archive.
- The Arts and Humanities Research Council (AHRC) supports the Archaeology Data Service, and formally funded data centres such as the Oxford Text Archive, the Visual Arts Data Service (VADS) and the History Data Service.

[1] Research Data Australia: http://researchdata.ands.org.au/
[2] DataCite Metadata Search: http://search.datacite.org/

 doi:10.2218/ijdc.v9i1.307

- The Natural Environment Research Council (NERC) funds the British Atmospheric Data Centre (BADC), British Oceanographic Data Centre (BODC), Environmental Information Data Centre (EIDC), National Geoscience Data Centre (NGDC), NERC Earth Observation Data Centre (NEODC), Polar Data Centre (PDC), and a range of other centres that hold data, such as the Marine Life Information Network for Britain and Ireland (MarLIN), and the UK Solar System Data Centre (UKSSDC).
- The Science and Technology Facilities Council (STFC) holds data generated by its large facilities.

These data centres acquire and curate data that result from research council grants. The UK also hosts several data centres which are not primarily funded by a single funding council, such as the EMBL European Bioinformatics Institute and the Cambridge Crystallographic Data Centre.

Each data centre uses a metadata standard and profile suited to its purpose and discipline, and has its own discovery catalogue. As researchers tend only to need data from within their own discipline, this has worked well. The one exception to this highly specialized picture was the development of a cross-search service for the NERC-funded data centres, the latest incarnation of which is the NERC Data Catalogue Service.[3]

Nevertheless, there are several factors which are starting to make a national, generalist search portal for research data more attractive. First, funders now require a greater variety of data to be shared, with institutions taking responsibility for archiving data without a natural home among the data centres. Indeed, EPSRC now expects institutions themselves to ensure research data resulting from its grants is preserved and disseminated. There is a danger that such data would remain largely invisible, as no researcher could be expected to perform speculative searches at each institution's data repository.

Second, funders now see data sets as valid research outputs, the impact of which is worth tracking, and this attitude is beginning to permeate into the UK's Research Excellence Framework. Funders and university administrators alike would therefore find it useful to search for data across both institutional data repositories and subject-based data centres, the former to track the impact of the projects they fund, the latter to keep abreast of the outputs from the university's researchers.

Third, research is tending to become more interdisciplinary and multidisciplinary. A generalist search portal would afford some convenience in being able to search for data in a discipline-agnostic way. It might also render some concerns about the data outputs of such research – whether data would be better held in one subject-based data centre rather than another, in an institutional repository, or scattered across several archives – a moot point from the point of view of visibility and impact.

Project Overview

The vision for the UKRDR is somewhat akin to that of Research Data Australia, in that it should focus on UK research data but should otherwise be inclusive of data regardless of discipline, the scheme used to identify it, or the type of repository in which it is held. It should provide an interface for searching and browsing the holdings of many different

[3] NERC Data Catalogue Service: http://data-search.nerc.ac.uk/

archives, for reviewing information about a particular data set, and for discovering how that dataset might be accessed (the registry would not itself hold data). Finally, it should enable the data outputs of particular researchers, projects and funding programmes to be tracked.

The original intention was for the pilot project to extend over 18 months. During this time, several possible software platforms for the registry were to have been evaluated, including CKAN[4] and the software underlying Research Data Australia. The proposal also left open the possibility of developing an original system, should none of the existing ones prove suitable. In addition, a new metadata scheme was to have been devised, tailored to the use cases of the registry while mapping easily to metadata schemes already in use in UK data centres and repositories, and informed by a large panel of experts.

Prior to the confirmation of funding for the project, the DCC conducted some exploratory work. As the software ANDS had developed for Research Data Australia was working well towards goals similar to those for the UKRDR, we concentrated on that in the first instance. We worked with the ANDS developers on ensuring the software could be installed and run outside of its native Australian context, and explored possibilities for harvesting metadata from Current Research Information Systems (CRISes) via the CERIF metadata standard (Tonkin & Russell, 2012; EuroCRIS, CERIF Task Group, 2013).

Various factors contributed to a delayed start to the project, thus it was reformulated as a six-month effort. As a result, the ambition for the pilot was scaled back. Given there was insufficient time to make a robust comparison of different systems within the new timescale, a single system had to be chosen at the outset. As we had already made some progress with the ANDS software, it made sense to build on that. Our initial concern about its portability had been addressed, and the fact that it was tied to a particular metadata scheme – RIF-CS[5] – was less of an issue now that we did not have the time to devise our own.

The agreed scope of the pilot project became as follows:

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 collaborators (data centres and repositories) willing to submit metadata to the prototype registry, and establish how the metadata will be harvested from both a technical and policy perspective.
3. Work with the collaborators to establish suitable mappings from their own metadata schemes to RIF-CS, as implemented by the ANDS software, and implement these as crosswalks.
4. Harvest metadata from collaborators, thereby populating the prototype registry with records.
5. Report on our experiences of using the ANDS software, the feasibility and implementation challenges of harvesting metadata from data centres and UK university data repositories, and the value of continuing to develop the registry.

It should be noted that the end result of this project will be a proof of concept rather than a fully fledged service. If the prototype registry shows promise, a continuation

[4] Comprehensive Knowledge Archive Network: http://ckan.org/
[5] RIF-CS: http://www.ands.org.au/resource/rif-cs.html

 doi:10.2218/ijdc.v9i1.307

project may be funded to develop the registry further. In which case, we expect that the questions of the most suitable technologies and metadata will be examined in more depth, alongside serious consideration of use cases, user experience, and integration with other services, such as the RCUK Gateway to Research portal.[6]

Architecture of the Prototype Registry

In the middle of 2013, ANDS launched an updated version of Research Data Australia, based on version 10 of their core software. The software was completely rewritten for that release, and the ANDS team subsequently made the code available on GitHub.[7] Prior to version 10, the code had been available for download from the ANDS website. Development has continued apace – version 11.1 was current at the end of 2013 and is the version being used for the pilot.

The core registry software incorporates a metadata registry, a front end portal and an access management system. It is an object-oriented PHP application following the Model-View-Controller design pattern and making use of a MySQL database. Apache Solr is used to index the metadata and enable searching. OAI-PMH harvesting is handled by a separate Java application and database. A suite of non-core components provides additional functionality, such as a CMS editor and an identifier-management front end.

We determined that a cloud platform would be the best place to host the software, in preference to one of the institutions involved in running the pilot. An approach was made to Eduserv with a view to using their platform. Unfortunately, their scale and remit meant that they were unable to accommodate a pilot project, such as ours. They suggested Microsoft's Azure platform, and made a referral to the team at Microsoft Research working on the Azure for Research Programme. The software is installed on an instance of the CentOS Linux distribution (the same operating system used by ANDS) on the Azure platform.

While registered users can manually enter the details of datasets or collections, it is expected that the registry will be populated chiefly through the harvesting of the metadata made available by repositories. The ANDS Registry Core software can import metadata made available at a URL over HTTP, while OAI-PMH imports are handled via the harvester component.

The system is tied to the RIF-CS metadata scheme, but includes a facility to perform crosswalks that convert harvested metadata described using other schemes to RIF-CS. This facility is provided as a PHP interface: crosswalks can be added by creating PHP classes implementing that interface.

The ANDS registry software allows the user to specify one of three harvesting modes:

- The default mode is 'Standard.' This mode simply ingests all records from the data source and can be used for both direct HTTP and OAI-PMH harvesting.

- 'Full Refresh' mode removes all records previously harvested from the data source and updates the dataset with the latest ingested records. Again, this mode is available for both direct HTTP and OAI-PMH harvesting.

[6] Gateway to Research: http://gtr.rcuk.ac.uk/
[7] Australian National Data Service on GitHub: https://github.com/au-research/

- 'Incremental' mode applies only to OAI-PMH harvesting. Using this mode means that only records that have been created or modified since the last harvest date will be ingested, with modified records replacing the previous version.

Manually added records attributed to a data source are not updated by any subsequent harvests from that data source. The frequency of recurring harvests can be specified, and a date can be set in advance for single harvests, or for recurring harvests to start. It is also possible to restrict harvests to a specified set in an OAI-PMH field.

Collaborators

In order to test the registry's ability to harvest data, we recruited a panel of collaborators representing both subject-based data centres and UK higher education institutions.

The data centres involved in the project are the UK Data Archive (one of the partners in the pilot project) and the primary NERC data centres, via the Data Catalogue Service (DCS). These two were selected as case studies for the social and natural sciences respectively, and since they each represent a diverse range of data collections.

The DCS is a union catalogue of metadata records from the six primary NERC data centres (BADC, BODC, EIDC, NEODC, NGDC, and PDC), UKSSDC and the Archaeology Data Service. Harvesting from the DCS has the advantage that the metadata records are already harmonized and standardized across the eight contributing data centres, reducing the number of mappings that need to be written. The DCS uses the NERC Discovery Metadata Standard, which is a profile of ISO 19115 (2003) and ISO 19119 (2005) and compatible with UK GEMINI 2,[8] INSPIRE,[9] and MEDIN (Seeley, Rapaport, Merritt & Charlesworth, 2013). DCS records already appear on the data.gov.uk portal.

The UK Data Archive holds data collections from all disciplines of the social sciences and humanities (the latter acquired via the now defunct History Data Service) and its metadata profile is based on the Data Documentation Initiative,[10] a metadata standard commonly used in the social sciences. The UK Data Archive's Discover portal is DDI 2.5 compliant.[11]

There are nine universities collaborating with the project: Edinburgh, Glasgow, Oxford, Oxford Brookes, Lincoln, Leeds, St Andrews, Southampton and Hull. Each of them have or are developing institutional data repositories, and are able to provide dataset records for harvesting by the registry. They are using a variety of different systems between them, including regular institutional repository software, such as DSpace and Eprints. The formats in which they are capable of exposing metadata therefore vary as well, but among the commonly supported ones are the EPrints ReCollect metadata profile, the DataCite metadata schema, and the base OAI-PMH metadata profile.

ReCollect is a plugin for EPrints software that turns EPrints into a research data repository with extra functionality and an expanded metadata profile for describing research data across academic disciplines. It was developed by the UK Data Archive and the University of Essex as part of the second JISC Managing Research Data (MRD)

[8] UK Gemini: http://www.agi.org.uk/uk-gemini/
[9] INSPIRE: http://inspire.ec.europa.eu/
[10] Data Documentation Initiative Alliance: http://www.ddialliance.org/
[11] UK Data Service Discover portal: http://discover.ukdataservice.ac.uk/

 doi:10.2218/ijdc.v9i1.307

programme (Van den Eynden, Ensom & Corti, 2013). Its metadata profile is based on common research data metadata schemas, such as DataCite, INSPIRE and DDI. The plugin is being used by the University of Leeds, University of Glasgow and University of East London to develop a data repository.

The DataCite Consortium is a registration agency for Digital Object Identifiers (DOIs), specialising in datasets. Any repository seeking to mint DOIs for its data holdings through DataCite must be able to provide metadata according to the DataCite Metadata Schema (DataCite, Metadata Working Group, 2013), a standard developed to support multidisciplinary and interdisciplinary data discovery.

The specification for OAI-PMH requires that, at a minimum, all compliant repositories provide metadata in a format based on the unqualified Dublin Core Metadata Element Set (Open Archives Initiative, 2008). The elements of this format enjoy near-ubiquitous support at the expense of detailed expressive power.

Metadata Mapping

We recognized early on that it would not be practical to require data archives and repositories to provide metadata in RIF-CS: to modify each archival system to allow this would be too much work to expect for a pilot project. Recent builds of the ANDS software allow metadata to be converted to RIF-CS after harvesting, so instead we opted to write converters for the metadata standards the collaborators could already work with.

The process we are following for writing these converters is as follows:

1. Match elements in RIF-CS to semantically equivalent elements in the target metadata standard, checking against sample records from collaborators using that standard.
2. Consult with the collaborators to ensure that the semantic matching is correct, and confirm any syntactic conventions.
3. Create a mapping from the target metadata standard to RIF-CS by specifying how element values and properties would need to be transformed, if applicable.
4. Implement the mapping as a crosswalk within the ANDS software.
5. Run the crosswalk on some sample records and perform an initial quality check.
6. Consult with the collaborators to ensure that the RIF-CS version accurately reflects the original metadata record.

Crosswalks are being written for DDI version 2.5 (as used by UKDA), the NERC Discovery Metadata Standard version 1.0, the DataCite Metadata Schema version 3.0, and the EPrints ReCollect metadata profile, alongside a minimal Dublin Core-based crosswalk that will be used as a fallback.

Evaluation Process

Due to the significant contraction of the initial pilot period and the necessary concentration of effort on the initiation and development of relationships with the collaborator

group, and the delivery of a pilot registry output, evaluation activities will be tailored accordingly to the scale of this phase of the project. In this highly participatory initiative, we are seeking iterative feedback from the initial set of collaborating institutions throughout our activity on the appropriateness of the approach taken and the solutions identified. In addition, the success factors to be considered at the end of this phase of development will be as follows:

1. Is the system stable and functional? Are the required basic functions available? Can we import metadata into the system?
2. Does the information in the registry look useful? Does the system appear straightforward to use, both for deposit and retrieval? Are the crosswalks effective? Are the results rendered accurately in comparison with the source repositories?
3. Is the information in the registry useful? Is the system straightforward to use? Is the deposit process clear? When searching the registry, can users find satisfactory results? What improvements are required? Are there additional functions which would be desirable?

The first two sets of factors will be evaluated by the project team themselves, while the third will be evaluated in consultation with the collaborator group.

Using both the ongoing user feedback and the results of this evaluation, we will identify improvements that might be made to the current system, and assess the resources required to do so. We will also consider what we have learned about the technical and user requirements for the UKRDR, and identify a set of criteria to be used to compare the ANDS software against possible alternative platforms.

In further phases of activity, the structured ROAMEF logic model approach will be considered as a more formal approach to evaluating impact and evidence of benefits of the UKRDR development. ROAMEF is a cyclical model of policy development and evaluation used by UK Government (HM Treasury, 2011a, 2011b), presented in a series of phases reflected in the acronym: Rationale, Objectives, Appraisal, Monitoring, Evaluation and Feedback. Whilst the model is not without its critics,[12] it appears to offer a useful method of structuring an evaluation of the impact of the intervention provided by the registry. Evaluation at this stage will include consideration of the potential of the registry to deliver the intended benefits, as well as the appropriateness of the technical and organizational approach, information on the costs of any further development and those of operating the registry as a sustainable service. Jisc would then consider how a UK Research Data Registry would fit into the Jisc research data proposition and the broader, national infrastructure as a sustainable service.

Discussion

The focus of the pilot project is to get a working implementation of the registry running, both to demonstrate what can be achieved and to uncover any potential pitfalls. Already our discussions with collaborators have brought up interesting practical challenges, such as how to check for modifications to and deletions of records; how to 'hide' metadata prior to quality control checks; and how to check for and merge duplicate records. But

[12] See, for example, Hallsworth, Parker and Rutter (2011).

 doi:10.2218/ijdc.v9i1.307

we are conducting this project with one eye on its future development, and thus we are considering some of the longer term implications of the work.

A major aim for the UKRDR is to increase the discoverability and visibility of data. It could do this on its own terms, in providing a portal that cross-searches many repositories, but that requires researchers to know about the registry and make a special effort to visit it. One of the attractions of the Research Data Australia software is that it has been optimized for search engine visibility. In other words, its records are intended to show up in search engine results, which means the information is pushed out to where researchers are already looking.

Another place researchers already look is in the reference lists of published literature. We would like to encourage researchers to cite the data they use, and one way we could do that is by providing sample citations. Within the RIF-CS standard are a set of optional elements for recording a fully formatted citation, or the bibliographic information that would be used to construct one. When writing the metadata converters, we have tried to populate the bibliographic information elements as far as possible, so that in due course we have the option of providing citations in multiple different formats.

When it comes to reusing data, it is important for researchers to know the intellectual property rights status of the dataset (e.g. the licence under which it is released), its provenance, and how other researchers have used or critiqued it. For this reason we also aim to collect, wherever possible, rights information concerning the dataset, and references to papers that have used or reviewed the data.

Conclusions

The UKRDR is still in its early phases. At the time of writing, we have set up a working instance of the registry software, assembled a group of collaborators who will contribute test metadata to the registry, and initiated the creation of crosswalks. Once the crosswalks are in place and sample records have been harvested, the registry will be evaluated.

If development of the registry continues, we plan to look again at alternative software platforms and the metadata stored by the registry. We will also consider how the registry will be used, how it should be made fit for those purposes, and how it might add value to its own records or those of others by establishing links with other national and international systems.

We hope in this way to ensure that the data assets produced by UK researchers receive the visibility they deserve, regardless of whether they are archived in a subject-based data centre or an institutional data repository.

Acknowledgements

The UK Research Data (Metadata) Registry (UKRDR) pilot project is funded by Jisc.

References

DataCite, Metadata Working Group. (2013). *DataCite Metadata Schema for the Publication and Citation of Research Data*. Retrieved from http://dx.doi.org/10.5438/0008

EuroCRIS, CERIF Task Group. (2013). CERIF 1.6 full data model. Retrieved from http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.6/documentation/MInfo.html

Hallsworth, M., Parker, S. & Rutter, J. (2011). *Policy making in the real world: Evidence and analysis*. London: Institute for Government. Retrieved from http://www.instituteforgovernment.org.uk/publications/policy-making-real-world

HM Treasury. (2011a). *The Green Book: Appraisal and evaluation in Central Government*. London: The Stationery Office. Retrieved from https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-governent

HM Treasury. (2011b). *The Magenta Book: Guidance for evaluation*. London: Author. Retrieved from https://www.gov.uk/government/publications/the-magenta-book

ISO 19115. (2003). Geographic information – Metadata. International Organization for Standardization.

ISO 19119. (2005). Geographic information – Services. International Organization for Standardization.

Open Archives Initiative. (2008). *The Open Archives Initiative Protocol for Metadata Harvesting: Protocol version 2.0 of 2002-06-14*. Retrieved from http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Seeley, B., Rapaport, J., Merritt, O. & Charlesworth, M. (2013). *Guidance notes for the production of discovery metadata for the Marine Environmental Data and Information Network (MEDIN)*. Retrieved from Marine Environmental Data and Information Network website: http://bit.ly/1gHYldk

Tonkin, E. & Russell, R. (2012). RIF-CS and CERIF alignment study. Retrieved from University of Bath website: http://opus.bath.ac.uk/30220/

Van den Eynden, V., Ensom, T. & Corti, L. (2013). *Research Data @ Essex final report*. Colchester: University of Essex. Retrieved from http://data-archive.ac.uk/media/402404/researchdataessex_finalreport_01.pdf