UK Research Data Registry and Discovery Service: pilot project

WP2: Infrastructure Implementation: Appraisal of ANDS software and adaptation to UK circumstances

Project Information			
Project Identifier	To be completed by Jisc		
Project Title	UK Research Data (Metadata) Registry: Initial Pilot		
Project Hashtag	#jiscRDS		
Start Date	1 October 2013	End Date	31 Mar 2014
Lead Institution	Jisc		
Project Director	Rachel Bruce		
Project Manager	Laura Molloy		
Contact email			
Partner Institutions	Digital Curation Centre (Universities of Edinburgh, Glasgow, Bath); UK Data Archive (University of Essex)		
Project Webpage URL	http://www.dcc.ac.uk/projects/research-data-registry-pilot		
Programme Name	Jisc Capital Programme		
Programme Manager	NA		

Document Information			
Author(s)	Patrick McCann		
Project Role(s)	Technical systems developer		
Date	27 May 2014	Filename	UKResearchDataRegistryPilot_reportWP2_03-4
URL	http://www.dcc.ac.uk/projects/research-data-registry-pilot		
Access	This report is for general dissemination		

Document History		
Version	Date	Comments
01	2014-05-13	First version of WP2 report; for review by project team
02 - 03	2014-05-27	Versions revised by project team
04	2014-06-13	Final for public dissemination

Table of Contents

Table of Contents 2
1. Background 2
2. Objectives 2
3. Activity 3
3.1. Implementation 3
3.2. Customisation 3
3.2.1. Crosswalks 3
3.2.2. Visual Customisations 4
3.3. Importing/Harvesting 4
4. Conclusions on Suitability 5

1. Background

This document reports on activities and findings of work package 2 (Infrastructure Implementation) of the UK Research Data Registry pilot project. This project aims to develop a pilot UK-wide registry for research data collections held in UK research institutions and subject data centres. The pilot tested approaches for a service that aggregates metadata relating to data collections or datasets held in institutional data repositories and established data centres, by harvesting metadata from their respective catalogues.

Such a registry and discovery service aims to provide a coherent point of access to discoverable, searchable, browsable and actionable descriptions of given datasets and how to access them, and thereby showcase the wealth of UK research data.

In this pilot phase, the project team adapted software developed by the Australian National Data Service (ANDS) for Research Data Australia[1] to develop a proof-of-concept pilot registry. Metadata profiles used by institutional repositories and data centres were mapped to Registry Interchange Format – Collections and Services (RIF-CS), the metadata schema used for the software, and metadata imported into the registry via OAI-PMH or other modes.

2. Objectives of the work package

Work Package 2: Infrastructure Implementation is concerned with the implementation of the Research Data Australia software[2] as the registry platform and the development of crosswalks to facilitate the harvesting and/or importing of metadata from repositories in the UK.

1 Research Data Australia: http://researchdata.ands.org.au/

2 ANDS Registry Core on GitHub: https://github.com/au-research/ANDS-Registry-Core

3. Activity

3.1. Implementation

Getting an instance of the ANDS Registry software up and running (i.e. without any specific customisation for use in the UK context) did not present any unexpected challenges. Deployment of the software was not difficult. The core Registry software is a PHP application with a MySQL database, readily deployed on an Apache web server within a Linux operating system. Apache Solr[3] is used for indexing and searching records and is straightforward to deploy using the Tomcat servlet container, as is the Harvester component. The combination of Microsoft's Azure cloud platform, the use of which was generously provided by Microsoft Research, and gracious assistance from the RDA team at ANDS allowed for the software to be deployed using a CentOS Linux system very similar to the one in use for RDA.

3.2. Customisation

3.2.1. Crosswalks

While the core metadata components of the application may be schema-agnostic, as contacts at ANDS have asserted to the project team, the application as a whole is tightly coupled to RIF-CS. As such, crosswalks must be provided to convert metadata encoded in other formats to RIF-CS. The software includes provision for this, providing an interface which crosswalk classes must implement and a location within the file structure where such classes can be placed so as to automatically available within the registry. Small modifications were made to the crosswalk interface by the project team which were passed back to the ANDS team and included in subsequent versions of the application.

Crosswalks were developed based on the work carried out in WP3 for converting the Eprints, DDI, MODS, DataCite, Gemini and Dublin Core metadata formats to RIF-CS. A number of them have been used with some success in the registry, notably those for Eprints and DDI; however, they remain experimental, having been tested against a relatively small selection of records and while they may not have reported errors when converting those records that does not guarantee the correctness of the records after conversion. There are particular issues around controlled vocabularies and the mechanisms available within the application to manage the translation from one to another in a sustainable way. Also, it must be remembered that the only information available for use in a crosswalk is that included within the source XML itself – other contextual information, e.g. the URL from which the metadata is harvested, is not available for inclusion in the target XML. While there is scope for improvement, it has been demonstrated that it is possible to create crosswalks to convert other metadata formats to RIF-CS.

Recently, ANDS have added crosswalks which make use of XSL to transform XML from one format into another. This is not an approach used in this pilot, but may be useful in future phases of activity. It is worth noting that the XSL component of such a crosswalk would also be of use to anyone else desiring to perform this XML transformation in a way in which pure PHP crosswalks would not be.

It should be noted that the lack of a standard metadata schema across all pilot participants means that crosswalks would be necessary for most participants regardless of the metadata format being used by the software. When converting metadata from one format to another there is a risk, depending on the specific crosswalk, that information may be lost or changed.

[3] Apache Solr: http://lucene.apache.org/solr/

- If an element of the source format has no corresponding element in the target format then information may be lost completely.
- Imperfect mappings may mean that information may seem to have a slightly different meaning in one format than in the other.
- Required fields in the target format with no equivalent in the source present a serious problem.

It would seem that a solution would be to store data in whatever format it is provided and that mappings could be used for the purpose of indexing and searching data. This is not currently possible with the ANDS software.

3.2.2. Visual Customisations

Some simple modifications were made to the user interface, replacing some of the Australia-specific references on the homepage with ones suitable for the pilot. A more complete overhaul of the interface would take significantly greater effort – the software has clearly been developed by ANDS for their own use, with references to the Australian context embedded within the code, rather than those elements being readily configurable for use in other settings.

3.3. Importing/Harvesting

The process of importing/harvesting records has proved challenging. Most significantly, the choice of crosswalk is ignored when choosing to harvest via OAI-PMH using the ANDS Harvester component, with an error being returned if the source URI does not return metadata in the RIF-CS format. This is not immediately apparent from the user interface or the available documentation. As of May 2014, the team at ANDS are taking steps to address this in both the Harvester and core Registry software, but this work comes too late for the purposes of this pilot.

The remaining options are direct HTTP harvesting (either one-off or scheduled), import from URL, import from pasted XML and manual entry. Whilst direct HTTP harvesting has also proved error-prone, import from URL and pasted XML work reasonably well, depending on the crosswalk in use, with only some minor interface issues. However, the details of the crosswalk selected for the data source are not displayed on the page on which importing is conducted, and some users (with the ability to import records) may not have access to this information at all.

Once records have been imported, the interface for reviewing and publishing them is quite complex, with common actions (view, edit) hidden away in sub-menus. Imported records are given a metadata quality level between 1 and 3:

1. [Includes all] Required RIF-CS Schema Elements
2. [Meets] Required Metadata Content Requirements
3. [Meets] Recommended Metadata Content Requirements

4 kinds of records are recognised by the system:

- Collections: Research Datasets or collections of research materials
- Parties: Researchers or research organisations that create or maintain research datasets or collections
- Activities: Projects or programs that create research datasets or collections
- Services: Services that support the creation or use of research datasets or collections

Typically, the crosswalks used in the pilot converted a record in the source format into a Collection and one or more Parties.

1035 records were imported during the pilot period from the UK Data Archive via the “Import Records from a URL” functionality and a DDI to RIF-CS crosswalk. 1030 of those records had

a metadata quality level of 2 and 5 had a quality level of 1. 601 of those records were Collections, 434 were Parties. 12 of those records (7 Collections and 5 Parties) were published.

26 Collections and 7 Parties were imported from the University of Glasgow using the "Import Records from pasted XML" function, all with a quality level of 1; all were published. Also, 6 Collections and 4 Parties were imported from the University of Lincoln. Both Glasgow and Lincoln used an Eprints to RIF-CS crosswalk.

The Dublin Core to RIF-CS crosswalk was used to import 35 Collections and 16 Parties from the Archaeology Data Service via URL. The 26 records with a quality rating of 1 or 2 were published (the others did not meet the minimum requirements). That crosswalk was also used to import 35 Collections and 10 Parties from the University of Oxford. The 17 published Oxford records all had a quality rating of 2.

The MODS to RIF-CS crosswalk was used to import 100 Collections and 4 Parties from the University of Edinburgh, 8 of which had a quality rating of 1 or 2 and were published. The same crosswalk was used to import 250 Collections and 533 Parties from the University of Hull, all with a quality level of 1 or 2, but they were all papers rather than datasets so none were published.

4. Conclusions

While the Research Data Australia software is available under an open-source licence (version 2.0 of the Apache Licence), it is an application which has been developed by a team at ANDS for their own use. They are continuing to develop it, with version 12 released in March 2014; this pilot used version 11.1, released in December 2013. It is not a collaborative open-source project, nor is it designed to be easily configurable for use by others e.g. there are numerous interface elements mentioning ANDS or Research Data Australia which are hard-coded. The separate (OAI-OMH) Harvester component[4], also developed by ANDS, is an older piece of software with little documentation available. In particular, the way in which these two pieces of software interact is not entirely clear from the available documentation. None of this is meant as a criticism of the ANDS team, who have put together a software suite which works well for their purposes, have chosen to do so transparently on GitHub, have made it available for reuse and modification by others and have been extremely helpful with regard to its use in this project. That said, the time difference greatly slows the response times for support requests, with emails sent to the ANDS team being replied to overnight, UK time.

Embarking on a process of implementing a UK Research Data Registry service based on the ANDS software would require considerable development effort. The most obvious way to do this, which is what has been done to some extent in this pilot, is to modify the ANDS software for use in a UK context. However, ANDS are continuing to develop their software and modifying the software in this way would make taking advantage of any future developments more difficult, requiring similar modifications to be made to those future versions before they could be used.

Alternatively, it might make sense to engage in a collaborative relationship with ANDS to convert the application into one which is readily configurable for deployment in disparate contexts. This would be a significant undertaking, but one with greater benefits for the wider community. Regardless of approach, configuring the software for the UK context could include using a metadata schema other than RIF-CS; however, the development and maintenance of crosswalks would still be necessary given the range of metadata schemas in use across the UK.

[4] ANDS Harvester on GitHub: https://github.com/au-research/ANDS-Harvester

The development of a custom registry platform from scratch would require considerable development effort over an extended period and risks 'reinventing the wheel'. All efforts should be made to make use of available solutions before this option is considered.

An alternative would be to use a suitable open-source application which has been developed by a community so as to be readily configurable and deployable in any context. CKAN[5] from the Open Knowledge Foundation[6] would seem to be the outstanding candidate of this kind. Whilst primarily used for the dissemination of government data (e.g. data.gov, data.gov.uk), it has been used to some extent within a research data context in the UK, most notably at Lincoln[7] and Bristol[8], and there is a fledgling CKAN for RDM community. The core software is simple to download and deploy, is easily configured in terms of branding etc. and the extensible architecture should in principle allow for the relatively straightforward addition of any required functionality that is missing. Any required development of CKAN extensions should constitute relatively small, self-contained pieces of work which would be of value to the wider research data management and CKAN communities.

A full evaluation of the suitability of CKAN for a UK Research Data Registry has not been carried out in this pilot, and there may be other applications which are suitable candidates. However, it is commended for serious consideration when deciding how to proceed in further phases of activity.

5. Recommendations

The first phase of pilot activity has provided us with experience of deploying the ANDS software in the new context of the UK HE and research community. Accordingly, the following are recommendations to considering in further phases of project activity.

- An evaluation of the suitability of CKAN - and any other applications which are suitable candidates for a UK Research Data Registry - should be carried out to compare feasibility of use and to gather user feedback on each application. The evaluation should cover the work required to produce extensions or similar necessary to make the application suitable for use as a UK Research Data Registry and Discovery.
- Efforts across the HE and research datacentre communities to agree more broadly on metadata schemas suitable for use when harvesting metadata into a cross-disciplinary registry, in order to reduce the amount of variety, would considerably assist in future development of the service.
- If the ANDS software is preferred, an exploration should be conducted of the possibility of collaborating with ANDS to:
    - Turn the ANDS software into an application readily deployed in a range of contexts, allowing future changes to be easily implemented. At the most basic level this covers the visual presentation of the software, but may extend to other features.
    - Improve the documentation accordingly.

5 CKAN: http://ckan.org

6 Open Knowledge Foundation: https://okfn.org

7 Orbital: http://orbital.blogs.lincoln.ac.uk

8 data.bris: http://data.bris.ac.uk/project-blog/

    - Shape the future direction of the software, preferably contributing development effort.
    - Develop associated software components, particularly the Harvester, such that they are suitable for the UK context (and preferably others).
- If ANDS software is preferred, significant effort will be required to further develop and thoroughly test crosswalks. They should be implemented in such a way as to maximise utility to the wider community.