JISC

UK Research Data Registry pilot project Feasibility and implementation challenges of harvesting metadata from data centres (WP4)

Project Information			
Project Identifier	*To be completed by Jisc*		
Project Title	UK Research Data (Metadata) Registry Pilot		
Project Hashtag	#jiscRDR		
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/news/research-data-registry		
Programme Name	Jisc Capital Programme		
Programme Manager	NA		

Document Information			
Author(s)	Veerle Van den Eynden		
Project Role(s)	Data Centre liaison		
Date	04 June 2014	**Filename**	UKResearchDataRegistryPilot_reportWP4_03
URL	*http://www.dcc.ac.uk/projects/research-data-registry-pilot*		
Access	This report is for general dissemination		

Document History		
Version	**Date**	**Comments**
01	2014-05-12	First version of WP4 report; for review by project team
02	2014-05-28	Version reviewed by project manager
03	2014-06-04	Published version

Table of Contents

This documents reports on activities and findings of WP4 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 an approach 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 would aim 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.

1. Background

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:

- ESRC funds the UK Data Archive;
- AHRC supports the Archaeology Data Service, and in the past funded data centres such as the Oxford Text Archive, the Visual Arts Data Service (VADS) and the History Data Service;
- 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), NERC Environmental Bioinformatics Centre)NEBC); 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);
- STFC funds the ISIS Data Catalogue (ICAT) and holds data generated by its large facilities.

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

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

Each data centre uses a metadata standard and profile suited to its purpose and discipline, and has its own discovery catalogue.

NERC recently developed its NERC Data Catalogue Service (DCS)[2] as a central data discovery portal that brings together all metadata records from data holdings of eight different data centres: BADC, BODC, EIDC, NEODC, NGDC, PDC, UKSSDC and the Archaeology Data Service.

2. Objectives

The aim of this work package (WP4) was to liaise with selected data centres and determine and test how existing metadata records for data collections held at established data centres can be harvested from their existing metadata catalogues - with the minimum of adjustment – into a central registry, to enhance findability and inter-repository searching. For selected data centres, their metadata profile was mapped to the pilot registry metadata schema Registry Interchange Format – Collections and Services (RIF-CS) (WP3 activities)[3] and a mapping service developed (WP2 activities) to import metadata records into the pilot registry.

3. Activities

To test the pilot research data registry's ability to harvest metadata records for data collections, we recruited collaborators representing both subject-based data centres (WP4) and UK higher education institutions (WP5).

The data centres involved in the pilot project were the UK Data Archive (one of the partners in the pilot project), the Archaeology Data Centre, and various NERC data centres, via the NERC Data Catalogue Service (DCS). These were selected as case studies since they each represent a diverse range of data collections.

3.1. UK Data Archive

The UK Data Archive[4] 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,[5] a metadata standard commonly used in the social sciences. The UK Data Archive's Discover portal[6] is DDI 2.5 compliant.

Discussions at the UK Data Archive in November 2013, about the pilot registry harvesting the Archive's metadata via OAI-PMH, prompted activities to upgrade its OAI stream, in line with recent developments for the new and improved Discover portal that uses more standardised controlled vocabularies, is DDI2.5 compliant, and contains DataCite Digital Object Identifiers (DOIs) for each collection. The existing live OAI stream in November 2013 was still DDI2.1 compliant and did not yet contain DOIs. The Archive released its new DDI2.5 compliant OAI stream in February 2014[7].

2 NERC Data Catalogue Service: http://data-search.nerc.ac.uk/
3 UK Research Data Registry Mapping Schemes: http://www.dcc.ac.uk/sites/default/files/documents/registry/uk-rdr-mapping-v09.pdf
4 UK Data Archive: http://www.data-archive.ac.uk
5 Data Documentation Initiative Alliance: http://www.ddialliance.org/
6 UK Data Service Discover portal: http://discover.ukdataservice.ac.uk/
7 UK Data Service OAI-PMH repository: http://oai.ukdataservice.ac.uk/oai/

The Archive's metadata profile was mapped to the pilot registry's metadata schema RIF-CS as part of WP3, and the mapping verified by Archive staff. After the mapping service was implemented in the pilot registry, import of metadata records was tested, with several hundred metadata records for data collections and responsible parties (data owners) successfully imported and published in the pilot registry. The mapping and import will be validated and fine-tuned in the next phase of the registry project.

3.2. NERC data centres

Discussions with various NERC data centres indicated that the preferred route for harvesting metadata records for their data centres would be via the newly developed DCS. The DCS uses the NERC Discovery Metadata Standard, which is a profile of ISO 19115 and ISO 19119, and compatible with UK GEMINI 2[8], INSPIRE[9], and MEDIN[10]. DCS records already appear on the data.gov.uk portal. Alongside the DCS, NERC developed a GeoNetwork Catalogue Services for the Web (CSW) node[11] to support the DCS portal, and from where metadata can be harvested. The CSW became available in early 2014.

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 metadata mappings that need to be written. The NERC DCS metadata profile was mapped to the pilot registry's metadata schema RIF-CS as part of WP3[3], and the mapping verified by NERC data experts. A mapping service to import from the NERC CSW was developed but remains to be tested.

3.3. Archaeology Data Service

The Archaeology Data Service[12] metadata profile follows a bespoke ADS schema called *ads_archive* and exposes its metadata via OAI-PMH. This service also publishes NERC-funded data collections into the NERC Data Catalogue Service (which is a subset of the total data collections); and publishes metadata records to the DataCite Metadata Store[13] when minting DOIs. Harvesting metadata records from the DataCite Metadata Store (subset BL.ADS) was deemed to be the preferred option for the registry. To date, this import has not yet been tested.

3.4. Further UK data centres

Besides those three pilot cases, initial contacts have also been made with the Visual Arts Data Service (VADS)[14], the Cambridge Crystallographic Data Centre[15] (CCDC) and the ISIS ICAT data catalogue[16] to explore the technicalities of metadata harvest and their metadata profiles to inform future research data registry work. All showed interest in being involved in future registry activities. VADS uses a Dublin Core metadata schema and publish OAI-PMH metadata. CDCC publishes its metadata to the DataCite Metadata Store. ICAT uses the Core Scientific Metadata (CSMD) standard.

8 UK Gemini: http://www.agi.org.uk/uk-gemini/
9 INSPIRE: http://inspire.ec.europa.eu/
10 Marine Environmental Data and Information Network metadata: http://www.oceannet.org/marine_data_standards/documents/medin_schema_doc_2_3_8.pdf
11 NERC Catalogue Services for the Web: http://csw1.cems.rl.ac.uk/geonetwork-NERC/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities
12 Archaeology Data Service: http://archaeologydataservice.ac.uk/
13 DataCite Metadata Store: https://mds.datacite.org/
14 Visual Arts data Service: http://www.vads.ac.uk/
15 Cambridge Crystallographic Data Centre: http://www.ccdc.cam.ac.uk
16 ISIS ICAT data catalogue: http://www.isis.stfc.ac.uk/groups/computing/data/icat11680.html

Table 1. UK data centres and their metadata characteristics

Data Centre	Metadata profile	Metadata harvest method for registry	Metadata harvest end point
UK Data Archive	DDI2.5	OAI-PMH	http://oai.ukdataservice.ac.uk/oai/
BADC	UK Gemini2	CSW	http://csw1.cems.rl.ac.uk/geonetwork-NERC/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities
BODC	UK Gemini2	CSW	http://csw1.cems.rl.ac.uk/geonetwork-NERC/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities
EIDC	UK Gemini2	CSW	http://csw1.cems.rl.ac.uk/geonetwork-NERC/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities
NEODC	UK Gemini2	CSW	http://csw1.cems.rl.ac.uk/geonetwork-NERC/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities
NGDC	UK Gemini2	CSW	http://csw1.cems.rl.ac.uk/geonetwork-NERC/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities
PDC	UK Gemini2	CSW	http://csw1.cems.rl.ac.uk/geonetwork-NERC/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities
UKSSDC	UK Gemini2	CSW	http://csw1.cems.rl.ac.uk/geonetwork-NERC/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities
ADS	ads_archive	OAI-PMH; DataCite Metadata Store	http://archaeologydataservice.ac.uk/oai/ ; https://mds.datacite.org/
VADS	DC	OAI-PMH	http://www.vads.ac.uk/oai
ISIS / ICAT	CSMD		
CDCC	DataCite	DataCite Metadata Store	https://mds.datacite.org/

4. Feasibility and implementation challenges for metadata harvest

All data centres that have been contacted in the course of this pilot project, either as pilot case study, or to discuss future engagement with a registry project, have shown immediate

interest in having their metadata records included in a UK-wide research data registry. All see the immediate importance in having a central discovery service for research data. In a next phase, a registry project can therefore aim to incorporate metadata records of all UK data centres into a central registry.

Significant progress has been made during the current pilot project to map specific metadata schemas to RIF-CS. Where mapping services for import into the pilot registry were developed and tested, this was successful. Where mapping services for data centres require further testing, we do not foresee problems in being able to ingest metadata records into the registry. This proof of concept pilot shows that developing a UK-wide registry that aggregates metadata relating to data collections or datasets held in institutional data repositories and data centres, is feasible. Although in a next phase, the suitability of the ANDS software and RIF-CS as a standard, will remain to be further evaluated and discussed across the wider stakeholder groups representing repositories and data centres.

The research data registry has the advantage of being able to build upon existing initiatives that already aggregate and standardise metadata records across disparate repositories holding collections of data, such as the DataCite Metadata Store (where repositories publish metadata when minting DOIs) and the NERC DCS. Both form aggregating points from where standardised metadata records from disparate sources can be harvested into the registry.

The fact that many data centres have been long established in the UK means that they have gained much expertise and experience on discoverability of data, such as how users want to search for data they need for their research. The registry project can build upon this expertise to implement efficient visibility and searchability of data.

Challenges that have been highlighted by data centres (and repositories) during this pilot phase are:

- promoting visibility of data to generic search engines such as Google;
- avoiding duplication when the same metadata records may exist in different places, e.g. in an institutional repository and a data centre, or in ADS and NERC DCS;
- the quality of metadata exposed by different data centres (and repositories);
- the diversity in mandatory and optional metadata fields used by data centres (and repositories); a minimum set of mandatory or recommended metadata fields may be required for a registry;
- the copyright of metadata where those are provided by data depositors; and whether or not all metadata can be licensed via a CC0 licence;
- being able to handle changes and deletions of metadata records at source, to ensure that a registry always accurately reflects the correct information;
- optimal workflows for data centres on when to push or pull metadata to a registry, and who should do this.

In addition, further testing and validation of metadata records for the pilot data centres is needed.

Whilst these challenges have been identified as requiring consideration during a next development phase of a research data registry, we foresee that all will be able to be overcome in the development of a UK-wide research data registry.