JISC Final Report

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	DCC/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 Web URL	http://www.dcc.ac.uk/news/research-data-registry		
Programme Name	Jisc Capital Programme		
Programme Manager	NA		

Document Information			
Author(s)	Compiled by Laura Molloy with contributions from RDRDS pilot project team		
Project Role(s)	Project Manager		
Date	June-Aug 2014	**Filename**	
URL	*If this report is on your project web site*		
Access	This report is for general dissemination		

Document History		
Version	**Date**	**Comments**
1.1	2014-06-06	First draft for review by DCC director
1.2	2014-08-03	Revisions and comments by KGA
1.3	2014-08-10	Third draft for review by project team
1.4	2014-08-27	Final version for delivery to Jisc

Table of Contents

Acknowledgements

This work is part of the Jisc capital programme and was funded by Jisc. We are grateful to the cooperation of all the institutions that freely participated in this experimental work and were so generous with their time and input.

1 Project Summary

In this initial six-month phase (October 2013 - March 2014), the Digital Curation Centre (DCC) aimed to test approaches for a service which would help to make research data more discoverable by aggregating metadata relating to data collections or datasets held in UK research institutions and subject data centres. Neither the pilot nor any further service will act as a repository for the datasets themselves. Rather, the project aimed to increase the likelihood of reuse of research data by providing a coherent point of access to discoverable, searchable, browsable and actionable descriptions of given datasets and how to access them, and so showcase the wealth of UK research data.

The pilot was delivered by the DCC and the UK Data Archive, working with RCUK data centres and a small group of universities with working data repositories. Liaison with the participant community was undertaken continuously throughout the work and the participant network provided monitoring and feedback for the direction of project activity.

2 Main Body of Report

2.1 Background and context

The changing practice of research increasingly requires the data and other sources that constitute the evidence underpinning findings to be made available for verification and reuse. For this shift to take place across the UK research sector, appropriate technological and skills infrastructure must be developed. This has been recognised by a number of initiatives over the last decade. In 2004, the UK Government Treasury, Department of Trade and Industry (DTI) and Department for Education and Skills (DfES) published the Science and Innovation Investment Framework 2004-14[1], in response to which the Office of Science and Innovation e-Infrastructure Working Group was formed, publishing findings in 2007[2]. These included a vision for a national e-infrastructure to support data-intensive forms of research. The UK Research Data Service (UKRDS) feasibility study then built on this work and in 2010, proposed and described a UK-wide Research Data Service[3]. This proposal called for, *inter alia*, 'a collaborative strategy for more efficient use of resources across the sector, including national data centres, libraries and archives and shared data centres' to provide 'better access to datasets of good provenance for researchers'.[4]

While the facility to verify findings has always been a central principle of research integrity, the ability easily to communicate, reanalyse, combine and re-use digital data has led to calls for increasing openness in research practice, most recently in the Royal Society's *Science as an Open Enterprise* report.[5] In step with this development, and keen to see the greatest return on investment in research through the maximum use and reuse of data assets, research funders have issued policies requiring grant holders to make research data accessible to the greatest extent appropriate. Increasingly, too, scholarly journals have published policies recommending, and in many cases requiring as a condition

[1] Now available from the National Archives at http://webarchive.nationalarchives.gov.uk/+/http:/www.hm-treasury.gov.uk/spending_sr04_science.htm.

[2] OSI e-Infrastructure Working Group (2007). *Report of the OSI e-Infrastructure Working Group:* http://www.nesc.ac.uk/documents/OSI/report.pdf

[3] UKRDS (2010), *UK Research Data Service: Proposal and Business Plan for the Initial Pathfinder Development Phase* https://web.archive.org/web/20110604073659/http://www.ukrds.ac.uk/resources/download/id/47

[4] UKRDS (2010), p4.

[5] Royal Society 2012, *Science as an Open Enterprise* http://royalsociety.org/policy/projects/science-public-enterprise/report/

of publication, that data substantiating published findings should be deposited in an appropriate database or archive, or otherwise made available.

Universities have an essential role in realising these aspirations and supporting researchers to adjust practice in response to such policies. There are many international, discipline-oriented databases for specific types of data that have been established in response to the needs of specific research communities. The UK has a particularly strong infrastructure of national, discipline-specific data archives (including the UKDA / UK Data Service and the various NERC data centres). However, significant gaps remain. The recent EPSRC Policy Framework on Research Data lays out expectations that place responsibility for ensuring the preservation and availability of research data with universities and other research organisations in receipt of EPSRC funding.[6] Even before this mandate, it was clear that universities face a significant challenge in supporting good practice management of digital research data through the lifecycle. There are concerns about the availability of robust storage and backup facilities and the capacity (skills, resources, infrastructure) to manage and preserve digital research assets.

Jisc has responded to these developments in a number of ways. At the core of the second Jisc Managing Research Data (JISCMRD) programme were seventeen large projects developing research data services (including institutional policies, support services, repository infrastructure and data catalogues).[7] In this way the programme develops practice and solutions that can be shared beyond these participating institutions. The Digital Curation Centre (DCC) provides guidance and support to UK universities, and more tailored advice in a series of Institutional Engagements.[8] The UMF investment allowed the development of tools and the offer of services in the cloud, brokered by Janet. Part of the UMF vision for the DCC was to test technical solutions for a metadata registry for research data. The current project undertook that with added stakeholder engagement.

In order to be re-used, research data must be discoverable. Universities are making research data assets available through repositories or other data portals. EPSRC requires research organisations to maintain a data catalogue. It is likely that some mechanism for aggregation will be necessary to increase visibility and to promote discovery and linking between datasets in related subject areas held in different institutions. Whereas document repositories can, in principle, make articles open to full-text searching by Google, this recourse is not available to data archives relying on metadata.

A registry solution that aggregates simple, but textually rich, metadata records for research data assets held in Australian universities and data centres has been developed by the Australian National Data Service (ANDS).[9] Research Data Australia provides a discovery service for Australian research data collections.[10] Research Data Australia presents records as web pages and thus promotes the visibility of data resources to search engines. The information architecture establishes connections between data collections, thus promoting discovery. Two important related use-cases of Research Data Australia are:

1. to break down data silos, encouraging linking and reuse of related data collections, particularly in interdisciplinary research;
2. to facilitate linking data to other research outputs, making data citation and referencing easier, thereby incorporating data in research achievements and impact.

The ANDS approach has demonstrated success and the software appeared potentially re-purposable. As UK universities become more involved in the management of research data and capacity develops, the use case for a UK Research Data Discovery Service grew. However, for such a service to be genuinely useful to the UK higher education and research community, it is crucial that the user community has a central role in setting out requirements and providing feedback on the various forms that such a service might take throughout the development and piloting process.

[6] EPSRC Policy Framework on Research Data http://www.epsrc.ac.uk/about/standards/researchdata/Pages/default.aspx
[7] JISCMRD Programme, Research Data Management Infrastructure Projects http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata/infrastructure.aspx
[8] DCC http://www.dcc.ac.uk/
[9] ANDS http://www.ands.org.au/
[10] Research Data Australia http://researchdata.ands.org.au/

The work in this pilot project was originally planned to take place over a 12-month period. Constraints arising from the Jisc transition required the compression of the timetable and sacrificing one or two desirable areas of work.

2.2 Project Outputs and Outcomes

Output / Outcome Type *(e.g. report, publication, software, knowledge built)*	**Brief Description and URLs (where applicable)**
Instance of the UK research data registry (pilot)	Test instance of the ORCA software initially deployed by the Australian National Data Service (ANDS), adapted for UK as proof-of-concept registry. Available at the time of writing at http://rdrds.cloudapp.net/registry/
Changes / fixes to ANDS registry software	A number of fixes and suggestions for improvement were supplied to the ANDS team during the development process and as a result of the requirements of this project, i.e. for repurposing in another national context.
Stakeholder group	Group of representatives from interested HEIs and subject-specific data centres who have contributed knowledge and effort towards trialling approaches, and have provided feedback and requirements.
Metadata mapping	This report documents the mappings created for importing metadata into the pilot UK Research Data Registry implementation. Available at http://www.dcc.ac.uk/projects/research-data-registry-pilot and in the University of Bath repository at http://opus.bath.ac.uk/40016/
Overview of current partner repository landscape	Analysis of questionnaires distributed to Phase 1 partner repositories (subject datacentre and HEI-based) in order to understand current scale and approaches of this sample of research data repositories. Findings are incorporated into the current report at Appendix A.
Brief report WP2	Brief report: appraisal of ANDS software and adaptation to UK circumstances (WP2): http://www.dcc.ac.uk/projects/research-data-registry-pilot
Brief report WP3	Brief report: metadata development: http://www.dcc.ac.uk/projects/research-data-registry-pilot
Brief report WP4	Brief report: feasibility and implementation challenges for data centres: http://www.dcc.ac.uk/projects/research-data-registry-pilot
Brief report WP5	Brief report: feasibility and implementation challenges for universities: http://www.dcc.ac.uk/projects/research-data-registry-pilot
Final report	The current document. Incorporates the individual workpackage reports: http://www.dcc.ac.uk/projects/research-data-registry-pilot

2.3 How did you go about achieving your outputs / outcomes?

2.3.1 Aims and objectives

Component objectives were:

- To test and appraise the ANDS software and its applicability to the UK context and requirements.
- To develop initial stakeholder engagement around the pilot service in order to demonstrate and evaluate its benefits.
- To move towards agreement, among an initial group of key experts and stakeholders, around the metadata profile to be used and the harvesting mechanism.
- To demonstrate how the platform may beneficially be used to aggregate metadata for discovery from existing national data centres and emerging university-based data repositories.
- To tackle feasibility and implementation challenges for metadata stores in a set of university-based pilots.
- To develop recommendations for useful further pilot work, which will in turn allow the development of a business model presenting options for a sustainable funding as a service.

The scope of this project was to pilot an initial approach for a UK Research Data Discovery Service. It was felt essential to trial the approach with a range of national discipline-specific data centres, as well as working with a number of university-based data repositories. The project sought to tackle technical and organisational challenges and to implement a pilot that could form the basis of a broader service. Along the way, challenges and benefits were appraised through liaison with stakeholders.

Ideally we would have preferred to compare more than one technical solution but this was not possible due to resource limits. Specifically, in a compressed timetable it was apparent that we did not have sufficient developer resource to run dependable parallel trials and create a full and accurate assessment of their respective performance and the effort available from partner institutions was necessarily limited as it was provided by them on a voluntary basis. However, the project had an evaluative component, both of the pre-intervention landscape and feedback on the outcomes of the current phase's particular approach to this service and its ability to deliver the intended benefits as well as the appropriateness of the technical and organisational approach. It is hoped similar activities can be run with alternative technical solutions in further phases of activity, to achieve a reliable and useful comparison for the relevant user community.

2.3.2 Technical development

The project piloted an approach for a UK Research Data Discovery Service to provide a better understanding of the technical and organizational requirements for such a service, in addition to a demonstration of its feasibility.

Work Package 2: 'Infrastructure Implementation' was concerned with the implementation of the Research Data Australia software[11] as the registry platform and the development of crosswalks to facilitate the harvesting and/or importing of metadata from repositories in the UK.

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[12] is used for indexing and searching records and is straightforward to deploy using the Tomcat servlet container, as was the Harvester component at the time of the pilot work. (However, since the end of the pilot phase, the Harvester has since been reworked as a Python application which can be deployed as a Linux service.) 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.

[11] ANDS Registry Core on GitHub: https://github.com/au-research/ANDS-Registry-Core
[12] Apache Solr: http://lucene.apache.org/solr/

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 a schema named Registry Interchange Format – Collections and Services (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 that crosswalk classes must implement and a location within the file structure where such classes can be placed so as to be automatically available within the registry. This crosswalk capability was added in anticipation of the requirements of the RDRDS pilot following discussions between ANDS and DCC staff at the time of development of the original pilot project plan. Small modifications were made to the crosswalk interface by the project team. These 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.

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.

The process of importing/harvesting records has proved challenging. Most significantly, the choice of crosswalk is ignored when choosing to harvest via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[13] 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.

(It is worth noting that this ANDS work has moved on considerably since the end of our pilot activity[14]. Updated documentation available from ANDS indicates that XSLT must be used to convert other metadata formats to RIF-CS within the Harvester. With the crosswalks needing to be present within the core registry software to enable the other import methods, those crosswalks will also need to be implemented in XSL if duplication is to be avoided and consistency maintained. The overall result is that the Harvester software may as a result be more fit for the purpose of the current project.)

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:

[13] Open Archives Initiative Protocol for Metadata Harvesting: http://www.openarchives.org/pmh/
[14] see https://github.com/au-research/ANDS-Harvester

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

Four 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. Of those records, 1030 had a metadata quality level of 2. Five records had a quality level of 1. Of the records imported during the pilot period, 601 of those records were Collections, 434 were Parties. Twelve of those records (7 Collections and 5 Parties) were published.

From the University of Glasgow, 26 Collections and 7 Parties were imported 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.

The initial pilot phase allowed us to start to scope many of the challenges of this particular technical platform and, crucially, the scale of the work that would be required to enable its successful full redeployment as a working service in the UK HE context.

2.3.3 Metadata development

One of the purposes of work package 3, 'Metadata development and standardisation', was to move towards an agreed metadata standard for use with the registry pilot phase 1. The decision to use the ANDS ORCA software as the platform for the registry considerably simplified this task: since ORCA is set up by default to use RIF-CS as its underlying metadata scheme, the effort required to switch it to another would have been greater than any potential benefit, in the time available. Indeed, working with RIF-CS has shown it to have some admirable qualities:

- With its four-entity data model, it has some of the power of a highly structured and relational scheme like CERIF while having a much gentler learning curve;
- It has a minimal number of required fields, meaning that records can be generated even if little information is available to populate them;
- It allows information to be recorded in a structured way, with semantic mark-up and controlled vocabularies, making it easier for automated tools to interact with it downstream, but provides distinct ways of providing unstructured information if that is all that can be harvested.

Even so, the process of using RIF-CS in practice raised some questions that need to be resolved.

- *Controlled vocabularies*: RIF-CS uses a large number of controlled vocabularies, an approach which has its advantages for downstream re-use of the information, but can be problematic when the vocabulary is not expressive enough for a given context. For example, we found that the (meta) vocabulary used for subject vocabularies, while extensive, did not contain terms for the subject schemes used by our data contributors. This left us with the dilemma of whether we should follow the standard exactly, and thereby lose knowledge of which schemes were used under the catch-all 'local' term, or extend the vocabulary with additional terms. We decided to take the latter approach on the understanding that if our extensions were used in the eventual Jisc service, we should attempt to feed our terms back into the RIF-CS standard.
- *Dates*: RIF-CS allows many different types of date to be recorded. Some of them have overlapping semantics, e.g. 'published', 'available' and 'issued'. In such cases it is not clear when harvesting a date whether to represent it using all three dates or just one, when the semantics appear to match all three. The former approach might assist downstream applications, but also appears to clutter the interface of the registry application: indeed, it seems the generated sample citation makes use of all available dates in a rather unhelpful matter.
- *Links*: In contrast, RIF-CS does not provide a way to distinguishing different types of links, beyond treating an alternative metadata record for the object as a separate, related entity. Thus there is no way of distinguishing a directed download link from a landing page link if both are provided. The registry software does, however, know how to generate a link from a plain DOI, so if the DOI in link form is separately provided, this leads to duplication of the link in the web page for the record.

The other objective for this work package was to develop crosswalks from the metadata schemes used by our data contributors to the one used by the registry. This approach was taken so as to ensure that the additional work required to deliver the pilot was carried out by those funded to do it. (A much earlier version of the project plan allocated funding to contributing institutions; this would have allowed alternative approaches to be explored.)

This was a three-stage process:

1. We had to decide which metadata schemes to support, as several data contributors could supply metadata according to multiple schemes.
2. We determined how, in theory, a RIF-CS record should be populated from information encoded in these supported metadata schemes. First, we matched the semantics of elements in RIF-CS with those of a particular source scheme. Then we determined how to transform information encoded according to the conventions of the source scheme into the formats and vocabulary terms expected by RIF-CS.
3. We wrote PHP crosswalks that instantiated these transformations and tested them against sample records provided by our data contributors. In many cases we had to write more complex transformations than originally planned in order to support real-world usage of the source metadata schemes.

The schemes we chose were as follows:

- DDI Codebook 2.5, the most detailed scheme available at the time of project activity from the UK Data Archive;
- UKGEMINI 2 (version 2.2 was used, for reference), the scheme used by the NERC Data Catalogue Service (DCS);
- OAI Dublin Core (oai_dc), the fallback scheme supported by all OAI-PMH endpoints;
- The EPrints metadata export scheme with support for extensions provided by the ReCollect plugin and other local variations used by our data contributors;
- DataCite (version 3 was used, for reference), used by the Archaeology Data Service and other repositories registering DOIs for their datasets, but in practice only available from DataCite's own OAI-PMH endpoint;
- MODS (version 3.5 was used for reference), a scheme commonly available from DSpace-powered OAI-PMH endpoints.

CERIF was considered but as support for datasets was still immature in CERIF at the time of project activity, and there were few sample records with which we could work, we felt it would be better to postpone writing a CERIF crosswalk until a later phase. The addition of CERIF support is considered highly desirable not just in a UK context but also as a vehicle for wider international uptake.

2.3.4 Data centre and HEI liaison

This first phase of pilot work established a group of initial stakeholders and a further group of interested data-holding institutions who may become active stakeholders for subsequent phases of work. Liaison activity was undertaken in WP4 (targeting discipline-specific data centres) and WP5 (targeting Higher Education Institutions) to build this network and to support the communication between the project team and the active participating institutions in order to achieve the technical and metadata work described above.

The aim of both of these work packages was to liaise with active participants from data centres and university data repositories to determine and test how existing metadata records for data collections held at these institutions can be harvested from their existing metadata catalogues into a central registry, in order to enhance 'findability' and inter-repository searching. The metadata profile used by each of the participating institutions was mapped to the pilot registry metadata schema RIF-CS (see WP3 activities, described above) and a mapping service developed (WP2 and WP3 activities) to import metadata records into the pilot registry.

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.

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.

These data centres acquire and curate research data that result from research council grants which was felt to be an appropriate starting point for forming the group of active participants for this first phase of activity.

The UK also hosts several discipline-specific 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. It is hoped that a further phase of activity will allow expansion of the stakeholder group to include these further bodies.

More recently, as recommended by the UKRDS final report and encouraged by efforts such as the Jisc Managing Research Data programmes of 2009-13[15] and the Digital Curation Centre, higher

[15] Jisc Managing Research Data programme 2009-11: http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx and http://www.jisc.ac.uk/whatwedo/programmes/mrd/outputs.aspx; and 2011-13: http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx

education institutions have also been considering and implementing ways to better manage their research data through advocacy and training in research data management and digital curation, and the development of technical solutions to ingest, manage and –where appropriate – expose data for sharing.

In this way it is hoped that discipline-specific data centres and HEI-based data repositories will move towards a more cohesive research data management landscape for the benefit of researchers and the funders and institutions that support research practice.

2.3.4.1 Participating data centres

UK Data Archive
The UK Data Archive[16] 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 (DDI)[17], a metadata standard commonly used in the social sciences. The UK Data Archive's Discover portal[18] 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. Discover 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 DDI 2.1 compliant and did not yet contain DOIs. The Archive released its new DDI 2.5-compliant OAI stream in February 2014[19].

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.

Natural Environment Research Council (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 Data Catalogue Service (DCS). The DCS uses the NERC Discovery Metadata Standard, which is a profile of ISO 19115 and ISO 19119, and compatible with UK GEMINI[20], INSPIRE[21], and MEDIN[22]. DCS records already appear on the *data.gov.uk* portal. Alongside the DCS, NERC developed a GeoNetwork Catalogue Services for the Web (CSW) node[23] 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, and the mapping verified by NERC data experts. A mapping service to import from the NERC CSW was developed but remains to be tested.

Archaeology Data Service

[16] UK Data Archive: http://www.data archive.ac.uk
[17] Data Documentation Initiative Alliance: http://www.ddialliance.org/
[18] UKDA Discover portal: http://discover.ukdataservice.ac.uk/
[19] UK Data Service OAI-PMH repository: http://oai.ukdataservice.ac.uk/oai/
[20] UK Gemini: http://www.agi.org.uk/uk-gemini
[21] EU INSPIRE: http://inspire.ec.europa.eu/
[22] Marine Environmental Data and Information Network (MEDIN) metadata: http://www.oceannet.org/marine_data_standards/documents/medin_schema_doc_2_3_8.pdf
[23] NERC Catalogue Services for the Web: http://csw1.cems.rl.ac.uk/geonetworkNERC/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities

The Archaeology Data Service[24] 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[25] 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.

Further UK data centres
Besides those three pilot cases, initial contacts have also been made with the Visual Arts Data Service (VADS)[26], the Cambridge Crystallographic Data Centre[27] (CCDC) and the ISIS ICAT data catalogue[28] 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 publishes OAI-PMH-harvestable metadata. CDCC publishes its metadata to the DataCite Metadata Store. ICAT uses the Core Scientific Metadata (CSMD) standard.

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	http://archaeologydataservice.ac.uk/oai/ ; https://mds.datacite.org/

[24] Archaeology Data Service: http://archaeologydataservice.ac.uk/
[25] DataCite Metadata Store: https://mds.datacite.org/
[26] Visual Arts Data Service: http://www.vads.ac.uk
[27] Cambridge Crystallographic Data Centre: http://www.ccdc.cam.ac.uk
[28] ISIS ICAT data catalogue: http://www.isis.stfc.ac.uk/groups/computing/data/icat11680.html

		Metadata Store	
VADS	DC	OAI-PMH	http://www.vads.ac.uk/oai
ISIS / ICAT	CSMD		
CDCC	DataCite	DataCite Metadata Store	https://mds.datacite.org/

Table 1: UK datacentres and their metadata characteristics

2.3.4.2 Participating HEI data repositories

Participating HEIs were initially identified for WP5 activity via existing relevant networks such as, *inter alia*, those institutions that had participated in the Jisc Managing Research Data programmes and/or DCC Institutional Engagements, and whom we were therefore aware were interested in the pilot registry initiative and likely to be able to participate. An open invitation was also published on the DCC website[29] and circulated through DCC publicity channels. In order to recruit appropriate participants for the initial active stage of pilot work, the following requirements were used:

The institution needs to:

- have some research data,
- be able to expose metadata about it; and
- have that metadata provide a route to the data itself.

Nine universities were ultimately involved in this initial pilot phase as active participants: Edinburgh, Glasgow, Hull, Leeds, Lincoln, Oxford, Oxford Brookes, Southampton and St Andrews.

Any other institutions that were interested but not yet able to actively participate were kept informed on project activity and invited to contribute feedback throughout the pilot phase. They will be invited to join as active participants in phase two if their circumstances have made such participation possible by that time. It is anticipated that active participants from the current phase of activity will form the core of the phase 2 advisory group.

Liaison between the project team and the participating HEIs was actively undertaken throughout the project in parallel with datacentre liaison work (WP4) and included the selection of partners, promotion of the aims of the work in HEI-appropriate language and through channels already familiar to the HE sector (including blog posts on the DCC website, Twitter promotion and communication through the 'Pipeline' newsletter), and provision of regular opportunities for HEI-based repository-related staff to interact with the pilot service and feed back on their requirements and experience using the pilot service including through group calls and face-to-face meetings. In this way, the stakeholder group was a mechanism by which we tested ideas and gathered intelligence from a sample of our likely ultimate user group.

It is important to an experimental project such as this that the needs of the likely user communities are understood as far as possible so as to ultimately deliver a service with the best chance of meeting those needs. This is perhaps a slightly clearer task when considering the discipline-specific datacentre community which has a longer heritage of established approaches for managing research datasets and more experience in collaborative approaches to their curation and exposure. We were aware that in contrast, the development and use of research data repositories and metadata catalogues are currently much less mature in the HEI community; in many institutions, these functions are still to be developed. These considerations lend complexity to the current project, including WP5,

[29] http://www.dcc.ac.uk/projects/research-data-registry-pilot

but also suggest that the current development of a UK research data registry and discovery service is a timely initiative.

In order to clarify and quantify the state of development of research data infrastructure in terms of both provision and use, and so to better understand current user requirements, a questionnaire was circulated to participating institutions to gather key facts about their repositories. Findings are discussed in section 2.4.4 below.

Throughout the reporting and communication activities of this current phase[30], further HEIs became aware of our initiative and indicated their interest in future participation. These institutions have been added to our contact list and will be invited to participate should further funding be confirmed.

2.4 What did you learn?

2.4.1 Technical development

While there is scope for improvement, it has been demonstrated that it is possible to redeploy the ANDS software to harvest UK-based research data repositories and data centres, and in order to do so, to create crosswalks to convert other metadata formats to RIF-CS.

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.

- 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.

This risk of loss may have negligible consequences if the information lost has minimal function in the target environment. For instance, some information held in a CERIF source may be critical for the purposes of research administration but may add little to a dataset's potential for discovery, and hence its loss in a crosswalk is acceptable. If the registry were required to act as a metadata gateway (that is, if information was to be passed through the registry between two otherwise unconnected sources) then the consequences of loss are potentially more significant, even if the information lost is not of use in the discovery service.

A number of valuable points were learned during the technical development in this first phase of activity:

While the Research Data Australia software is available under an open-source licence (version 2.0 of the Apache Licence), it is an application that 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 yet 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 that are hard-coded. At the time of pilot activity, the separate (OAI-OMH) Harvester component[31], also developed by ANDS, was an older piece of software with little documentation available. In particular, the way in which these two pieces of software interact was 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

[30] See 'Presentations and other publications' at http://www.dcc.ac.uk/projects/research-data-registry-pilot.
[31] ANDS Harvester on GitHub: https://github.com/au-research/ANDS-Harvester

so transparently on GitHub, have made it available for re-use and modification by others and have been extremely helpful with regard to its use in this project. That said, the time difference inevitably slows the response times for support requests significantly, with emails sent to the ANDS team being replied to overnight, UK time. We note, however, that since the end of the current pilot phase, the Harvester software and its documentation appear to have been improved and updated by ANDS,

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 is continuing to develop its 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 that is readily configurable for deployment in disparate contexts. This would be a significant undertaking, but one with greater benefits for the wider community. However, substantial and sustained developer resource would be required to be committed to this approach.

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 currently in use across the UK.

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 that has been developed by a community so as to be readily configurable and deployable in any context. CKAN[32] from the Open Knowledge Foundation[33] 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[34] and Bristol[35], and there is a fledgling CKAN for RDM community[36].

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 that would be of value to the wider research data management and CKAN communities. However, again, substantial and sustained development effort would need to be committed to fulfil this ambition.

2.4.2 Metadata development

In the majority of cases we found it possible to create high quality RIF-CS records from harvested metadata, even with OAI Dublin Core, our simplest source scheme. Whilst much depended on how much information was actually provided, of course, in many cases records only missed ORCA's Quality Level 3 – the highest – due to missing association with an activity (project) record. The more permissive and flexible the source scheme, though, the more work had to be done in the crosswalk to achieve a high quality record. An extreme example of this is the oai_dc 'coverage' element, which can contain geospatial or temporal information in any number of formats. In such cases we chose to write an algorithm for determining the type of information and represent it appropriately, rather than ignore the field, as we felt it was better to risk the occasional mistake than to lose the information entirely.

[32] CKAN: http://ckan.org
[33] Open Knowledge Foundation: https://okfn.org
[34] Jisc MRD project Orbital: http://orbital.blogs.lincoln.ac.uk/
[35] Jisc MRD project data.bris: http://data.bris.ac.uk/project-blog /
[36] E.g. as described in late 2013 by Birgit Plietzsch: https://research-computing.wp.st-andrews.ac.uk/category/rdm/ckan/

There were some instances where we could not in practice complete the mappings as originally intended. Both DDI and UK GEMINI can record information about related publications, but they do so in a free-text manner; this makes it hard to extract the title and identifier information needed to record the information in RIF-CS. We did not have the resources to write a reliable routine to achieve this in the current phase of activity but if the ANDS software was to be sustained as the preferred option for our registry and discovery service, this would be a useful task to complete. Furthermore, several schemes can record funder information, but without information about the activity (project) through which the dataset was funded it is not possible to record this in RIF-CS in a meaningful way. It may be hoped that in a future development of this service more comprehensive information might be harvested from Current Research information Systems.

There were some technical challenges to implementing the crosswalks:

- The relevant functionality was newly implemented in ORCA and there were some coding issues that needed to be addressed before the crosswalks could be written.
- Also, since the crosswalks work on a whole XML file level, rather than on extracted records, they had to take account of how records were harvested as this affects if and how they are 'wrapped' in XML.
- Lastly, the crosswalks could not be tested on a live OAI-PMH harvest, as the OAI harvest method provided by ORCA was hard coded to request RIF-CS metadata.

Concentrating on record quality meant we did not have the resources to address some other points of concern we had, such as how to handle record deletions, and how to deal with records that might be harvested from several sources. We believe the ORCA APIs would be useful in finding appropriate solutions, but have not been able to verify this in practice.

Another point we would like to consider in a future phase, if the ANDS software is persisted with, is whether the workflow used by ORCA – harvesting metadata and transforming it to RIF-CS before storing it – is optimal. A different workflow to consider would be to harvest and store the 'original' metadata records and normalise them either on the fly or as a regular internal registry operation. That way, if improvements are made to the crosswalk, the benefits can be realised immediately instead of waiting for the next harvest.

In summary, then, we found it possible to normalise metadata from a wide variety of data sources into high quality records that could be used to operate a data discovery service. There are certainly some quirks to be addressed, and the solution we have found may not be robust enough in their current state to cope with all possible implementations of the chose metadata schemes. Nevertheless, the progress we have made in this short phase bodes well for future development.

2.4.3 Datacentre liaison

All data centres that have been contacted in the course of this pilot phase, either as a pilot case study or to discuss future engagement with registry and discovery service pilot work, have shown immediate interest in having their metadata records included in a UK-wide research data registry and discovery service. All see the immediate importance in having a central discovery service for research data. In a further phase, a registry and discovery service pilot project can therefore aim to move significantly further towards the incorporation of 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. 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 upo 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.

Challenges that have been highlighted by data centres (and HEI-based repositories) during the work of this pilot phase (as distinct from those identified at the closing workshop in June 2014 and included here as Appendix B) 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 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.

It was clear from the June 2014 workshop that there is a large network of discipline-specific data centres and databases with which we have not yet connected, and with whom it would be useful to engage. In future activity, we would plan to expand the datacentre representation to the Visual Arts Data Service (VADS)[37], the Cambridge Crystallographic Data Centre[38] (CCDC) and the ISIS ICAT data catalogue[39], all of whom have been contacted and have shown interest in participation, alongside any other data centres who are interested in and capable of participation. We aim to explore the technicalities of metadata harvest and their metadata profiles to inform future research data registry work.

2.4.4 HEI liaison

We found from WP5 activity that there is appetite amongst UK HEIs for a research data registry and discovery service but that not all are currently in a position to actively participate in such an initiative. Nine HEIs were able to participate, meeting the criteria outlined in section 2.3.4.2 above, and were willing to do so on a voluntary basis for a set period of time. Any other institutions that were interested but not yet able to actively participate were kept informed on project activity and invited to contribute feedback throughout the pilot phase. They will be invited to join as active participants in phase two if their circumstances have made such participation possible by that time. It is anticipated that active participants from the current phase of activity will form the core of the phase 2 advisory group.

One of the main findings from the evaluative and landscape description activity of this work package was confirmation that datacentres have, as suspected, a significantly longer history than HEIs in managing research data collections and making them available for re-use: 80% of participating university repositories have become open to users (both depositors and seekers) within the last 5 years, whereas all participating datacentres have been available to depositors longer than 5 years, with the oldest respondent launched in 1969. We know there is a significant amount of relevant expertise in the discipline-specific research datacentre community and we are hopeful that the current pilot work, with any future phases, can be a mechanism - alongside other existing mechanisms - to

[37] Visual Arts Data Service: http://www.vads.ac.uk
[38] Cambridge Crystallographic Data Centre: http://www.ccdc.cam.ac.uk
[39] ISIS ICAT data catalogue: http://www.isis.stfc.ac.uk/groups/computing/data/icat11680.html

help share that expertise across datacentres and also with the universities who are developing research data repositories.

Several of the universities commented either in response to the survey or informally in person that whilst they had struggled to pull together the information requested by the survey, they recognised the activity as worthwhile and confirmed it was useful to know how to get such statistics together for future use. It may be worth noting that although the deadline given for responses to the survey was 5 February 2014, only one response was received by then. The latest response received was at the end of April 2014. This appeared to demonstrate that whilst institutions and data centres were in principle prepared to agree to participate voluntarily in initiatives such as the pilot phase of this project, they can find it difficult to prioritise such activities.

The survey found that many of the HEI-based data repositories had much lower staffing levels but also much smaller collections, compared to the participating data centres. They appeared much less able to report on their collections than data centres. However, clearly the burden of reporting is not entirely related to the size of the collections, so this would have been a heavier burden on the much smaller level of FTE per HEI repository. The low level of resource able to be committed to each HEI repository was a key economic argument for the central discovery service proposed in the UKRDS report.

As the exercise gathered such details from HEIs as the software used, OAI-PMH endpoint, use and type of PIDs, current FTE, and the scale of the data collections they held, the survey work proved immediately useful for supporting the work of the registry project team; but the results also serve as a yardstick from which we aim to measure development of the UK HEI data repository landscape as the field develops.

The full analysis of the research undertaken into the current UK data centre and repository landscape, as represented by our active participants, can be found as Appendix A of this report.

2.5 Immediate Impact

The current phase of activity has done much to scrutinise, improve and question the current research data repository landscape, at least amongst the participation institutions of this first phase.

Firstly, liaison efforts and project management activities have brought together a group of 9 HEIs and 9 discipline-specific data centres specifically to examine, discuss and suggest strategies for a common mechanism by which research data metadata can be harvested and promoted for greater findability. These discussions have focused minds by bringing the issue – and the practical ramifications of taking necessary action - to the attention of these 18 partner institutions.

We have also progressed a substantial distance towards creating a positive attitude amongst stakeholders about the possibility of a research data registry and discovery service that can be genuinely helpful in promoting the research datasets they steward. Interest in such a service can be partially illustrated by a couple of measures: the initial webpage describing the initiative has to date received 260 ‘shares’; the end of phase 1 workshop attracted 30 participants in a 2-week registration period to attend a day-long session of reporting and structured discussion about requirements. A blogpost describing the end of phase 1 workshop has to date received 81 ‘shares’ and 64 tweets. Tony Hey of Microsoft Research mentioned our efforts in his keynote address to several hundred delegates at the Research Data Alliance 3rd Plenary in March 2014.

Discussions initiated by the current phase of activity have also promoted positive developments from specific institutions: for example, as noted in section 2.3.4.1 above, 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. This resulted in a newly DDI 2.5-compliant OAI-PMH stream available in February 2014.

Our activities have also had a positive impact upon collaborative efforts across institutions – for example, RDRDS pilot activity prompted a discussion amongst NERC-funded data centres, which resulted in confirmation of the suitability for the CSW for harvesting.

Technical development in the current phase has also had a positive immediate impact: primarily the achievement of ingest of 1035 metadata records from across our participant group. This will have had the further positive impact of involving participating repository staff in our learning process as we work together to understand the tools employed, resulting in increased skills and knowledge in OAI-PMH deployment and the characteristics of the various repositories involved.

Other positive impact has also been made by our technical work: for example, a number of suggestions and 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. We are hopeful that our work has also, at least in part, helped inform and/or encourage the recent further development of the Harvester.

Work in and preceding this project has also been one of the spurs to the creation of an RDA Working Group[40] on the interoperability issues that will be faced by such services around the world.

The initial stakeholder surveying work has highlighted the importance to a number of participating institutions of routine gathering and recording of statistics relating to their data repository: this exercise has the immediate positive impact of gathering key statistics which may form the basis of subsequent comparative numbers in order to assist individual repositories with their reporting and user requirements research.

2.6 Future Impact

The key impact of any subsequent work was set forth in the business case for the service set out in the final UKRDS report – to increase the discoverability, and hence reuse, of UK research data in the most cost-effective way. National provision reduces the need for expert effort at local level and has particular value for those institutions with relatively small research data collections. These are also the universities least able to deploy and resource effective local services.

Our experience and the lessons learned from this initial pilot phase will primarily have an immense impact on further phases of activity. Our experimental work to redeploy the ANDS software in the UK context has allowed us to better understand the opportunities and limits of the platform, and allow a comparison with future pilots of alternative technical solutions. The metadata development work to date also provides a guide to possible future directions towards agreement of an approach to metadata for registry purposes, and may provide useful and reusable expansion of the controlled vocabulary used by the RIF-CS schema, should we have the opportunity to feed in our recommended terms.. The liaison work with HEIs and discipline-specific data centres has created a stakeholder group and the beginnings of a user community which will impact the scope and direction of future development and application across the UK research sector, by provision of sector-specific intelligence and user requirements and priorities.

Any further funded phase of the current initiative will involve some comparative work at an appropriate stage to redeploy the stakeholder survey described in Appendix A of the current document. This may serve as a useful way to track the development of the repository sector in the UK as work on the RDRDS continues. Whilst it would possibly be rash to attribute its expansion specifically to the impact of the RDRDS activity, we hope to see at least a concomitant growth and maturity of the HEI repository landscape in particular.

Future funded phases will include evaluation work to understand the progress made in each stage of the project and attempt to track impact attained.

[40] Described at https://www.rd-alliance.org/group/data-description-registry-interoperability.html

3 Conclusions

There is appetite in the UK research data repository community for a harvesting mechanism that will contribute to better exposure of research datasets. There is also a considerable, and so far un-met, need for evidence-backed guidance on appropriate metadata standards for the description of research data together with evidence on which metadata elements make the greatest contribution to discovery and reuse potential. However, the development of this concept, its technical platform and the associated user community can only be done in an embryonic way in a short-term intervention such as the current pilot effort. We adjusted our ambitions accordingly in order to concentrate on proving the concept of such a registry and discovery service and to begin building a stakeholder group. From these activities we conclude that whilst there are a number of intricate technical, metadata and advocacy tasks required for further progress, there is a need and desire to bring about this service and to collect and interpret the evidence that it will make possible.

The initial pilot phase allowed us to start to scope many of the challenges of a particular technical platform and, crucially, the scale of the work that would be required to enable its successful full redeployment as a working service in the UK HE context.

The core Research Data Australia software is simple to download and deploy and the extensible architecture should in principle allow for the relatively straightforward addition of any required functionality that is missing. However, we are clear that it is not the only possible solution for a research data discovery service and that all viable options should be investigated and assessed by the target user community. Nonetheless, the willingness of the ANDS team to move to a more collaborative approach for future software development is a very encouraging development.

There is a compelling counter-argument to further evaluation of alternatives, which rests on two key points. One is that continued delay to enhancing discoverability of research data at a national level loses the value that could be realised from existing data whilst evaluation continues. The second is that delay at this stage is likely to result in greater effort being expended within individual HEIs to do what a national service would have done for them; it is also likely to result in even greater diversity of metadata standards in the absence of clear guidance and infrastructure at national level. There is thus a considerable cost to the delay that can only be justified if there is potential for even greater gains if one of the alternatives is adopted. There is little evidence at this stage that this is true. However, we also recognise that key to the success of this initiative is an open and public process by which the user community is actively engaged in the consideration of approaches to the service, and so the provision of two working instances for comparison seems a reasonable strategy to employ.

It will be an area of priority in future activity to work towards a service which provides a reliable way for researchers to search for and find research data across UK repositories and data centres. It is hoped this can be done in a way which allows the easy access to key statistics for participating institutions, and which can forge and build on productive cooperation with related initiatives such as data citation indexes, registries of repositories, research information system development initiatives, research data metadata development initiatives and other relevant activities. However, our current priority remains the provision of at least two viable alternatives of the working platform to as large a user community as possible in order to deliver the most appropriate solution for the UK research data repository and data centre context.

Further work needs to be done in as sustained an effort as possible. Our experience from this first phase of activity indicates a two year period as the practical minimum to test a comparative technical system both technically and with the user community, to develop and refine required functionality and to perform the necessary analysis of user reactions to the system from across an expanded stakeholder group.

4 Recommendations

A number of recommendations emerged from this first phase of activity.

Technical development:

The first phase of the 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 consider in further phases of project activity.

- An evaluation of the suitability of CKAN – and any other applications that are suitable candidates for a UK Research Data Registry and Discovery Service – 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 research data registry and discovery service in the UK context. However, this should not be allowed to delay deployment of a service based on the ANDS system and the gathering of evidence about data discoverability using this system.
- 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 change 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.
    - 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 other international contexts too). As indicated above, this has to some extend been undertaken after the end of the current pilot phase but as such we have not tested the results.
- 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. However, this effort is likely to be required whatever solution is adopted.
- Sufficient developer effort must be funded for the necessary period of time, and on as continuous a basis as possible, to minimise the loss of knowledge and the fracturing of relationships with useful contacts which occurs with staff changeover.

Metadata development:

- Agreement by HE and research data centre communities on metadata schemas suitable for use when harvesting metadata to a cross-disciplinary registry in order to reduce the amount of variety, would considerably assist in future development of the service.
- Clear evidence-backed guidance at a national level has the potential to speed adoption at individual HEIs, reduce costs, increase researcher engagement and assist commercial suppliers to provide compliant solutions.

Liaison activities (data centre and HEI):

- Future work should broaden data centre liaison from RCUK data centres to include such trans-national UK-hosted data centres as CCDC.
- Returns from HEIs to the questionnaires indicate that many of the Phase 1 participants are in the early stages of setting up and implementing their research data repository. Any Phase 2 activity must be pragmatic about the rate at which progress can be made given the state of development at many UK HEIs.
- If funding is made available to allow participant institutions to commit resource to work on implementation, care should be taken that those institutions who voluntarily engaged with Phase 1 do not find themselves unable to access such funds. It would be better to offer no funding than for some Phase 1 institutions (e.g. those in Scotland) to be unable to access funds. Such a situation would potentially damage our relationships with such institutions and the overall credibility of, and goodwill towards, the initiative. It is recommended that if funds are made available for participating institutions to help resource the necessary effort of participation, these are offered first to those active in Phase 1, who have already volunteered resource and expertise to the efforts of the pilot, and the remainder via a lightweight competitive process.

5 Implications for the future

It is hoped that future phases of activity will build towards a research data registry and discovery service with more specific, measurable impact on the UK research community and, ultimately the international research community. Researchers will be able to access metadata describing datasets in their disciplines of interest, which have enormous potential to expand and accelerate the generation of knowledge on a national and international basis. Participating HEIs and research data centres would be likely to see a marked increase in traffic to their holdings as researcher use increases. Researchers could – through effective deployment of data citation practices – see an increase in the re-use of their deposited datasets: ideally alternative metrics will be developing along with such initiatives as ours to more effectively, completely and easily allow researchers to access and understand their citation metrics and this, hopefully can help drive change in researcher reward structures as data re-use becomes more readily understood as a marker of research impact. In addition to the benefits for researchers, research administrators at HEIs may be able to benefit from access to statistics about the re-use and impact of research data generated at their institution, providing another driver for recognition of research datasets as primary outputs of research and an institutional asset in their own right.

We also note that our work may have an impact on other relevant stakeholders in the research data repository landscape. For example, we note that RCUK research councils have agreed to use Researchfish[41] as the means of collecting data on the outputs and impact of RCUK-funded research from September 2014. Researchfish provides a way to collect output information from projects of all research disciplines, and enables funders to obtain a common qualitative and quantitative view of the details, productivity and impact of the research they individually and collectively support. This complementary initiative focusses on funder needs, much as CRIS systems in HEIs focus on the needs of research administration. However, much information is shared between these use cases and that of data discovery. It is important that requirements for such systems in future focus on holistic requirements for the research and education sector, allowing (for instance) free flow of metadata into and out of such systems – something which is not possible at present with many widely-used commercial CRIS systems.

Additionally, a data registry has the potential of forming a key part in the infrastructure for measuring and understanding data re-use. Thomson Reuters run a Data Citation Index[42]. They are keen to be able to harvest quality metadata from a single source and are already collaborating with ANDS to use Research Data Australia. Thomson Reuters have shown interest in collaboration with this initiative in due course.

Technical development
New development work that could take our progress further is detailed in section 4 above, 'Recommendations'. Work on both the ANDS software and an implementation of CKAN would provide a working instance of each solution for the consideration of our user community.

In addition, we would like to note that ANDS have recently 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, which can be used in the newly-developed version of the Harvester, 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. This may have impact beyond the scope of the registry.

Metadata development
Achievement of a more confluent approach across a significant number of UK research data centres and HEI-based repositories towards metadata for research datasets would have far reaching impact across the UK research sector.

Data centre and HEI liaison
The stakeholder group which has been initiated and developed in the first phase of activity is comprised of individuals who separately and collectively hold the knowledge and experience gained

[41] https://www.researchfish.com/
[42] http://thomsonreuters.com/data-citation-index/

by their interaction with the RDRDS pilot initiative. This group includes representatives of HEIs, research data centres and the DCC. We hope to build on this group with representatives of further research data centres, HEIs and other relevant parties to create a larger core user community who represent broad and deep knowledge, across discipline and professional boundaries, of research data and its curation and re-use. It is hoped that their expertise not only informed this current phase and will inform and influence future funded phases of this initiative, but that – particularly as the RDRDS pilot work matures - they will also provide a flow of information into their respective discipline and professional areas about the aims and benefits of the RDRDS and encourage engagement with the RDRDS from their own professional networks.

Appendices

Appendix A: RDRDS pilot phase 1: Key facts questionnaires: April-May 2014: analysis and question schema

Jisc Research Data Registry and Discovery Service (RDRDS): pilot project, phase 1. Key facts questionnaires: analysis, April-May 2014

Rationale

In the first phase of the Jisc Research Data Registry and Discovery Service (RDRDS) pilot work43, UK higher education institutions (HEIs) and discipline-specific datacentres were openly invited to become active participants. Nine universities44 and 4 datacentres45 became active participants in this initial pilot phase, with further institutions becoming involved as the work progressed.
In order to supply useful information for work in this and subsequent phases, and to inform an overall impression of the current landscape of institutional and datacentre research data repositories at an early stage of project activity, a set of questions was formulated and circulated to stakeholders and to some of the datacentres who were not actively participating in other ways, but were still willing to complete the survey. The question schema is available as an appendix to this document. Numbering of the paragraphs below corresponds to the numbering of questions in the schema. The question schema does not pretend to constitute a sophisticated research enquiry; rather, it was intended to quickly gather a set of basic facts and statistics about participating research data resources which, it was expected, would be readily available.

Response rate

The questionnaires were sent to 18 organisations (9 universities, 9 datacentres). Seventeen responses were received, providing key facts about repositories hosted by 15 organisations (9 universities and 6 datacentres). This represents an institutional response rate of 83%. Three datacentres failed to respond. Two universities returned 2 responses each, representing discrete research data repository systems. For clarity, in the overview below the term 'repositories' will be used to refer to the 17 overall collections or holdings of research data for which a given organisation has a responsibility and on which information was returned, as opposed to the 15 organisations which returned questionnaires. It is recognised, however, that not all of these overall data holdings are necessarily understood as repositories per se; the term is used here as a shorthand. Of the 17 questionnaires returned, 2 contained responses to every question. Many respondents commented that they had not gathered these key facts about their service before; 2 organisations specifically noted that it had been a useful exercise.

1. Access URL

Sixteen (of 17, 94%[46]) responses provided at least one access URL. Of these, all but one (a university repository) used the '.ac.uk' suffix for this location.

[43] Described at http://www.dcc.ac.uk/projects/research-data-registry-pilot

[44] Namely: Edinburgh, Glasgow, Hull, Leeds, Lincoln, Oxford, Oxford Brookes, Southampton, St Andrews.

[45] Namely: ADS, BODC, EIDC, UKDA.

[46] All percentages have been rounded to the nearest whole number.

2. OAI-PMH endpoint

Ten (of 17) responses (59%) specified an OAI-PMH endpoint. Eight of these were university-based, the other 2 from datacentres. The remaining 4 datacentres indicated that specification of an OAI-PMH endpoint was not applicable, as their data is harvested and exposed by the NERC-wide Catalogue Service for Web (CSW) infrastructure.

3. Date from which repository / datacentre available to users (i.e. both data depositors and data seekers):

Fifteen (of 17) repositories (88%) both accept dataset deposits and provide a search interface to data seekers. Of the remaining 2 repositories, one is still in development, and the other accepts deposits but is still working to secure funding to enable access for data seekers.

a) **data depositors**

University responses were as follows:

Oct 2004; Feb 2009; 2010; Jul 2010; Mar 2012; Sep 2012; Jan 2013; Feb 2013; May 2013.

- The majority (7/9; 78%) became available to depositors within last 5 years.

Datacentre responses were as follows:

Apr 1969; 1984; 1994; 1997; 1998; 2009.

- All (6/6, 100%) available to depositors for 5 years or longer at time of study.

b) **data seekers**

University responses were as follows:

Oct 2004; Feb 2009; Jul 2010; 2011; Sep 2011; Sep 2012; Jan 2013; Feb 2013; May 2013.

- Majority (6/9; 67%) became available to depositors within last 5 years.

Datacentre responses were as follows:

Apr 1969; 1984; 1994; 1997; 1998; 2010.

- Majority (5/6; 83%) available for 5 years or longer at time of study.

From these responses we can see that discipline-specific research datacentres in the UK have a significantly longer history of dedicated dataset curation than research institutions such as universities: More than two-thirds of participating university repositories have become open to users (depositors + seekers) within the last 5 years, whereas all participating datacentres have been available to depositors longer than 5 years, with the oldest respondent launched in 1969 and half of datacentre responses providing availability dates for both depositors and seekers in the 1990s. As organisations with substantial research data management and digital preservation experience through varying economic and political climates, the importance of the older datacentres as stakeholders and advisers to the current project is highlighted.

Where both dates are supplied (15 responses), we can see that 12/15 (80% of this subset, 71% of all 17 responses) of repositories made their services available to data depositors and seekers at the same time, with one organisation (an HEI) intriguingly opening first to data seekers, and 2 repositories reporting earlier access to depositors than seekers.

4. Access conditions for depositors

Of 17 responses, 10 (59%) required depositors to be registered using credentials which proved an affiliation specifically to the hosting organisation. Three organisations required depositors to register using some sort of institutional credentials (these could be from the hosting organisation or another recognised organisation). Two more required user registration without necessarily requiring any kind of institutional affiliation. One organisation did not require depositors to register. In the remaining response, user registration is not required but the repository operates a whitelist of research activities from which they will accept data.

A charge for deposit and archival storage of data was reported by 2/17 repositories (12%) although the author is aware of at least one further participating repository where this charge is levied but has not been reported here.

5. Access conditions for data seekers

Of 17 responses, 9 repositories (53%) report no existing access conditions for data seekers. Eight repositories (47%) require seekers to register for access to some or all datasets; one of those institutions clarifies that browsing is free but registration is required in order to download data. Three repositories require user registration for all data seekers. One organisation charges a fee for some of their datasets but 'only for a small number of added-value products'.

Comparing the answers to question 4 and 5 suggests there is currently a more open approach to access to repositories for users who wish to seek, rather than deposit, datasets. There is a need to ascertain the provenance of deposited datasets, including the context from which they emerge, their licensing arrangements and other relevant descriptive information; the higher level of user identification required for depositors seems appropriate to achieve these aims. However, this overview of the access conditions of participating repositories shows that many do not participate in the 'open access' agenda, at least as defined by, for example, the Budapest Open Access Initiative[47], the Panton Principles for Open Data[48] and similar initiatives. This in itself is not necessarily a problem as long as such practices are acceptable to their funders and user communities, and as long as the repositories who apply access conditions – particularly to data seekers – are careful to refrain from presenting their resources as open access-compliant.

6. Total number of registered data users (since launch of service)

Ten organisations (10/17, 59%) were able to supply a figure of their registered users. The lowest specified figure was 0, the highest 40,516. However at least two of these were projected, rather than actual, figures.

[47] http://www.budapestopenaccessinitiative.org/
[48] http://pantonprinciples.org/

Table 1: Total number of registered data users (since launch of service)

Number of registered users	Number of organisations
<100	2
100-999	2
1,000 – 4,999	4 (including one estimate of potential users)
5,000 – 9,999	0
10,000+	2

It is possibly not surprising that the datacentre group was more likely to return a specific figure of registered users rather than noting an estimate of potential users or returning no figure at all: 5/6 datacentres (83%) reported a specific figure of registered users, compared with 4/11 (36%) of university-based repositories.

a) **Estimate of total users if no registration used:**

Few repositories (3/17, 18%) provided an estimate of their total users to date, which again were widely dispersed (lowest reported figure: 22, highest: 3949). However, two of these responses were inconsistent with the rest of the responses from that institution, in one case at least being offered as potential rather than actual user numbers (for example, the size of the institution's research population).

7. Repository software used

Seventeen responses were received. Of those, 7/17 responses (41%) indicate substantial in-house development work to create a bespoke solution, although it is acknowledged that some of the other solutions reported will also have required significant technical effort to install and configure and it is acknowledge that it is difficult to draw a distinct line between configuring an off-the-shelf system and building one from existing components.

In order of popularity, named 'ready-made solution' tools include: ePrints (4/17), plus Hydra, DSpace, CKAN, Pure, Fedora and Equella all used by one repository each.

8. Number of datasets ingested to date (total)

Of 17 responses, 13 (76%) offered a specific figure for number of datasets ingested, including 2 estimated figures. Two more offered a figure for metadata records but not datasets themselves. A further one organisation offered various estimated figures based upon the 'definition of a dataset' (which may be considered slightly surprising ambiguity from a datacentre). There was one non-response.

From the 13 specific (including 2 estimated) figures for datasets held, e.g. ranging from 0 to 3000, we find that most repositories hold relatively small numbers of datasets.

Table 2: number of datasets ingested to date (total)

Number of ingested datasets	Number of responses
<100	7
100-999	5
1,000 – 4,999	1

Two respondents specifically questioned the definition of a dataset, or found it difficult to count datasets in the context of this study. It should be noted that one of the estimated figures supplied by one of these two respondents allowed for the possibility of their repository holding between 980 and 3.6million data items, depending on the definition used. Due to the range of possible responses offered by this particular repository, it is not included in Table 2.

9. Page visits to date (total OR during 2013, as specified by respondent)

Four responses supplied specific figures for this:

- 2013: 2,480,808
- Total (*i.e. since launch between Jan 2013 and Feb 2014*): c. 600
- 27,675 (*date range unspecified*)
- 67,169 since July 2010

Four further organisations supplied figures but were clear these were for the top-level URL of their online service as opposed to specifically the pages providing access to research datasets and / or their metadata:

- Total: 156,346 (to the repository as a whole. It is a mixed repository and we do not have access to data only figures at this time). This led to 643,541 page views.
- 506,402 page views from Apr 12 – Mar 13 (*top level data centre URL.ac.uk*)
- (*Data centre*) webpages: 321,400
- 11,779,617 (this applies to both data and publications) from 31st May 2013 to 24th April 2014

It is unclear across all responses at which level of each website the page visit total has been drawn - specifically, if all page visits reported are to pages which are directly connected to dataset discoverability as opposed to, for example, general introductory or guidance pages.

The key finding from this question is that the majority of responses (9/17, 53%) did not supply any figure for this question, and half of those who did were unable to specify that these view counts were specifically for visits to pages directly connected to dataset discovery. Seven repositories (41%) indicated this statistic was not available to them and 2 further organisations left the question blank. University and datacentre responses contained 'unknown' in the same proportion, i.e. about half in each category. This would seem to suggest that the gathering of these fairly core facts about data repositories is not yet a widespread practice.

10. Number of datasets currently available to data seekers

All repositories but one supplied a figure in response to this question (see Table 3). However, one response specified it related to metadata records, not datasets. Extrapolating from Q8, one more repository can also be identified as providing the metadata only. These are not included in Table 3.

In most cases, the number of datasets ingested and the number available to data seekers are broadly or exactly similar, relative to the size of the collection as can readily be observed by the 'percentage of availability' column in Table 3, and in analysis of the distribution of results (Table 4), which shows the most frequent percentage of availability to be 100%, whether the responses of institution 15 are included or excluded from the analysis.

Only 2 organisations report noticeably low percentages of ingested datasets available to data seekers (see Table 4). In one case (HEI), 3,000 datasets are reported to be ingested and 268 available (less than 9%) to data seekers. In another (a datacentre), the respective figures are 214 and 106, i.e. less than half of datasets ingested are available to seekers.

Table 3: comparison of datasets ingested and datasets currently available to data seekers

Institution ID	A: Number of ingested datasets	B: Number of datasets currently available to data seekers	A-B	% ingested datasets that are available[49]
2	166	164	-2	98.8%
3	12	10	-2	83.3%
4	28	28	0	100%
5	0	0	0	100%
7	3,000	268	-2,732	8.9%
9	6	3	-3	50%
10	20	16	-4	80%
11	116	115	-1	99.1%
12	12	6 (+2 only available within institution)	-6 (public) -4 (within institution)	Public: 50% In institution: 66.7%
13	615	615	0	100%
14	214	106	-108	49.5%
16	~253	~253	0	100%
17	~30	~30	0	100%

[49] Calculated to one decimal place, for clarity: this also applies to all further tables.

Additionally, the organisation which gave four separate figures for its one repository, depending on how 'dataset' is defined, including only these options where data is available using online repository:				
15	~980	~960	-20	98%
15	~96,500	~88,129	8,371	91.3%
15	~44,000	~44,000	0	100%

Table 4: Distribution of percentage availability of ingested datasets across sample (when inst. 15 included, n=16 responses, with inst. 12 being treated as one response and inst. 15 as three responses; when inst. 15 excluded, n=13)

Percentage of ingested datasets currently available	**Number of responses including inst. 15. (n=16)**		**Number of responses excluding inst. 15 (n=13)**	
	Absolute	**%**	**Absolute**	**%**
<50%	2	12.5	2	15.4
50-79%	2	12.5	2	15.4
80-89%	2	12.5	2	15.4
90-99%	4	25.0	2	15.4
100%	6	37.5	5	38.5

11. How many of your datasets have been downloaded at least once?

Reports were sparser for this point, with a specific number of datasets downloaded provided by 4/17 repositories (23.5%) and an estimated figure offered by a further 5/17 repositories (29.4%).

Table 5: Number of datasets available to data seekers downloaded at least once and percentage of those publicly available

Number of datasets available to seekers	**Number of datasets downloaded at least once**	**Percentage of publicly available datasets, downloaded at least once**
164	"Unknown … approx 50%"	~50%
268	268	100%
115	"Almost certainly all."	Interpreted as ~100%
6 public + 2 available within institution	7	87.5% of those available from within institution (7/8)

615	615	100%
106	90	84.9%
~96,500	~61,500	~63.7%
~253	~253	100%
~30	~30	~100%

Of these 9 responses, 5 repositories (56% of those who responded to this question) reported that 100% of the datasets available to data seekers have been downloaded at least once, either as a definite or estimated figure. This equates to 29% of all 17 repositories.

Again, the most significant finding of this question is the difficulty in ascertaining a definite figure, which echoes the observation made earlier that tracking of key statistics such as downloads is not (yet) a routine activity for many repositories. Indeed, some HEI-based repositories are actively hampered in doing so: one HEI reported that "our system doesn't allow us to determine this specifically …".

Data centres again showed a stronger overall ability to report on this point than HEIs, with 5/6 (83%) data centres (80% of data centres) returning a specific figure (albeit for one data centre, the figure is only for a subset of their collection), compared with 2/11 HEI-based responses (18% of HEI).

12. Current staff resource (FTE)

There was a clear divide between the levels of staffing for HEI-based repositories and datacentres.

Of the 11 HEI-based repositories, all but one provided a response and 8 of these specified a figure. The lowest figure provided was 0.25FTE/repository. The highest FTE reported by the 9 institutions was 2.4FTE: however, this was staffing for 2 repositories hosted by the same institution. For the purposes of analysis, I have attributed 1.2FTE to each of the 2 repositories hosted by that organisation, leaving the highest value per repository at 2FTE, and the mean value across the range at 1.3FTE per repository.

Data centres, possibly predictably, reported much higher staffing levels. Four of 6 (67%) returned a figure; the lowest was 4 and the highest 17. One figure was expressed as 14-15: this has been determined as 14.5 for the current analysis. The mean value across the range is 12.4FTE per datacentre.

13. Total number of searches on your discovery system

Specific or estimated figures were received from 3/11 HEI-based repositories (27%) and 3/6 (50%) of datacentres. Datacentre responses indicated complications in providing a straightforward answer due to multiple ways for seekers to search for datasets.

Across the 6 numerical responses received, the lowest was 0 and the highest 430,068. Again, however, a significant point is the lack of information available here: there was no figure reported for 11/17 (65%) of repositories.

14. Total number of downloads of datasets to date (all datasets)

Over both groups, 8/17 (47%) provided a definite number of downloads. Two further responses provided the amount of data downloaded in Tb and one further response provided a broad estimate ("In excess of one million per year"). Table 6 provides comparison of those figures with the number of datasets available to data seekers. Repositories returning zero values for both measures have been excluded from Table 6 and only institutions which have returned figures for both measures are compared here.

Table 6: Number of overall downloads of datasets compared to number of datasets available to data seekers

Institution number	Number of datasets available to seekers	Number of downloads to date	Avg number of downloads per dataset
7	268	11,421	42.6
11	115	61,634	535.9
12	6 public + 2 available within institution = 8 total	437	54.6
13	615	179,900	292.5
14	106	1864	17.6

15. Number of downloads of most popular dataset to date (where popular = highest number of downloads in total, as opposed to e.g. views)

Nine repositories (9/17, 52.9%) provided a specific figure in their response. The lowest of these was 81 and the highest 58,149.

One further response suggested the most popular dataset download may be one of two resources: image downloads of a particular map numbered more than 6 million in 2013, and the most popular file download from the same repository was downloaded around 2000 times per month during an unspecified timescale. Due to the ambiguity of these responses, they have not been included in the analysis here but remain worth reporting.

16. Permanent identifiers assigned to datasets?

Thirteen of 17 (76%) of repositories report they assign a permanent identifier to some or all of their datasets. Nine repositories specifically mention use of DataCite DOIs as part of their current or planned activity. A further 3 mention DOIs but do not specify the registration agency.

17. Any other relevant characteristics or issues about the current setup and use of the data repository?

No particular trends emerged from responses to this question.

18. Requirement to keep responses anonymous in public reporting

One respondent asked for this.

Question schema

Jisc research data registry project: pilot 2013-14: Collaborator key facts
Today's date:
Contact name:
Job title:
Institution:

Per repository:

1. Access URL:
2. OAI-PMH endpoint, if applicable:
3. Date from which repository available to users (or please indicate if not yet available):
    a. Depositors:
    b. Data seekers:
4. Access conditions for DEPOSITORS (check any/all that apply):
    a. none ☐
    b. user registration required ☐
    c. user registration required, institutional affiliation necessary ☐
    d. user registration required, affiliation to my institution necessary ☐
    e. user fee required ☐
    f. other:
5. Access conditions for DATA SEEKERS (check any/all that apply):
    a. none ☐
    b. user registration required ☐
    c. user registration required, institutional affiliation necessary ☐
    d. user registration required, affiliation to my institution necessary ☐
    e. user fee required ☐
    f. other:
6. Total number of registered data users:

   a. Estimate of total users if no registration method used:
7. Repository software used:
8. Number of data sets ingested to date (total):
9. Page visits to date (total):
10. Number of data sets currently available to DATA SEEKERS:
11. How many of your available data sets have been downloaded at least once, to date?

****For 2013****

12. Current staff resource (FTE):
13. Total number of searches on your discovery system:
14. Total number of downloads of data sets to date (all data sets):
15. Number of downloads of most popular data set to date (popular = highest number of downloads in total):
16. Permanent identifiers assigned to data sets? Yes / No
    a. If yes: convention used (e.g. DOI, DataCite DOI, etc):
17. Any other relevant characteristics or issues about the current set-up and use of the data repository?

****

Use of your responses:
Thank you for completing this questionnaire. This information is gathered for use solely by the Jisc Research Data Registry project as part of our evaluation activities and as such is very valuable to us. Your answers will be used and stored securely and only for the purposes of this project. Your answers will be aggregated with answers from other project collaborators to give us an impression of current activity in the sector. However, responses of individual institutions may be publicly identified only in the reporting activities of this project, where we may for example name an institution that has experienced particularly growth in use of their data repository.

- Please indicate against any response if you would like that particular response to be anonymous in our reporting.
- Alternatively, indicate if you would like all your responses to be anonymous in such public reporting ☐

Please return to laura.molloy@glasgow.ac.uk by Wed 5 Feb 2014.

END

Appendix B: RDRDS pilot phase 1: requirements gathered at closing workshop, 16 June 2014

Requirements, desiderata and other points to consider in future development of the RDRDS: points emerging from user community discussion groups

Discussion group: Technical aspects

Required:

- Harvesting metadata
- Indexing for searching
- Transforming between formats
- Inbuilt QA
- Platform: open source and metadata schemas, OS too
- Should be modular
- Resilient regarding metadata changes in terms of source
- The registry should be harvestable
- Handle DOIs etc
- Differentiate between research outputs and what is part of ongoing research
- Report on errors to the data owners for correction
- Formal system providers can request their data to be included
- Responsive design
- Resilient
- Searchable on Google and have analytics
- It must be sustainable

Desired:

- Version control
- Crosswalks intelligently with additional fields
- Automated push
- Links to publications/grants
- Statistics tracking,
- Faceted
- Appropriate controlled vocabularies
- Choice of harvesting formats when harvesting FROM the registry
- In the absence of identifiers, have ones based on context
- De-duplication and merging of records possible

Optional:

- Ability to add comments, corrections, improvements
- Show related records
- Ways of sharing
- New data alerts, based on search criteria
- Favourites

Discussion group: Metadata development

Essential:

- Problems solved by having clear rules – have policy on preferred record sources, which to take/are preferred
- A policy on deletion – this should try to reflect what the original repository is doing
- Need Ts and Cs on usage to cope with copyright, mainly for downstream users
- Set of recommendations on what metadata should be provided – need clarity on the degree of flexibility on this
- Clear semantics should be included in DMPs
- Metadata must be presented in a search engine-friendly way
- Want local copy of original metadata record so can correct any errors in crosswalk without having to then re-harvest

Desirable:

- Subject terms – very important access point – nominate preferred scheme?
- Fuzzy matching
- Filter out metadata field which are dodgy or not useful
- Unique ID for people, organisations too when possible. Could even look up pre-existing ones to use when possible
- Indication of type of data, eg. image, multimedia
- Typology for links provided by registry eg. further info, data download

Discussion group: workflows at partner institutions

There was no classification of requirements and desiderata from this user group. The following points were provided:

- How should things be updated if something is withdrawn from a repository? Have an urgent take-down policy?
- Should the registry exclude anything? If published (by participating repositories) should be available for harvesting.
- 2-way information would be ideal: Information should be available to institutions so they know who is using their data and institutions should report benefits back to registry to help justify its existence.
- Feedback on quality of data from registry to institutions
- Metadata quality is key
- Sometimes quick updates are needed – registry should pull frequently when needed
- If there are links between repositories it could be useful to show these, but how to do this automatically? May need to be provided in the metadata by depositing institutions.
- We would like to participate in international network of services, but we want practical experience at a national level before considering this. Before committing to include some of the functionality above we want to know who some of the technical requirements would benefit.
- Planned to have an advisory group and a technical group but are more needed to get specific input as required.

- A sustainability working group could be helpful.

Discussion group: use cases

There was no classification of requirements and desiderata from this user group. The following points were provided:

There were considered from different stakeholder perspectives:

- Interest in increased visibility of datasets and how this would be provided for was a key thought had by this group.
- Click through stats for institutions and data centres
- Datasets with DOI would be good (but bearing in mind not all datasets have a DOI attached – approach shouldn't exclude those)
- Subject classifications – which to use? How can they add to the metadata harvested?
- Everyone is interested in finding related datasets and how the registry will be able to point to related datasets.