Jisc Research Data Registry and Discovery Service (RDRDS): pilot project, phase 1

Key facts questionnaires: analysis, April-May 2014

Laura Molloy: laura.molloy@glasgow.ac.uk

Rationale

In the first phase of the Jisc Research Data Registry and Discovery Service (RDRDS) pilot work[1], UK higher education institutions (HEIs) and discipline-specific datacentres were openly invited to become active participants. Nine universities[2] and four datacentres[3] became active participants in this initial pilot phase, with further institutions involved as 'in the loop' participants, and others 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 eighteen organisations (nine universities, nine datacentres). Seventeen responses were received, providing key facts about repositories hosted by fifteen organisations (nine universities and six datacentres). This represents an institutional response rate of 83%. Three datacentres failed to respond. Two universities returned two responses each, representing discrete research data repository systems.

For clarity, in the overview below the term 'repositories' will be used to refer to the seventeen 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 fifteen 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 seventeen questionnaires returned, two contained responses to every question. Many respondents commented that they had not gathered these key facts about their service before; two organisations specifically noted that it had been a useful exercise.

1. Access URL

Sixteen (of 17, 94%[4]) responses provided at least one access URL. Of these, all but one (a university repository) used the '.ac.uk' suffix for this location.

---

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

2 Namely: Edinburgh, Glasgow, Hull, Leeds, Lincoln, Oxford, Oxford Brookes, Southampton, St Andrews.

3 Namely: ADS, BODC, EIDC, UKDA.

4 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 two from datacentres. The remaining four 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 two 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 five years.

Datacentre responses were as follows:

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

- All (6/6, 100%) available to depositors for five 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 five years.

Datacentre responses were as follows:

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

- Majority (5/6; 83%) available for five 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 five years, whereas all participating datacentres have been available to depositors longer than five 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 (fifteen responses), we can see that 12/15 (80% of this subset, 71% of all seventeen 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 two repositories reporting earlier access to depositors than seekers.

4. Access conditions for depositors

Of seventeen 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 seventeen responses, nine 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[5], the Panton Principles for Open Data[6] 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 repositories (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.

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

[5] http://www.budapestopenaccessinitiative.org/

[6] http://pantonprinciples.org/

Number of registered users	Number of repositories
<100	2
100-999	2
1,000 – 4,999	4 (including one estimate of potential users)
5,000 – 9,999	0
10,000+	2
No response	7

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 seventeen responses, 13 (76%) offered a specific figure for number of datasets ingested, including two 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 thirteen specific (including two 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 datacentre URL.ac.uk*)
- (*Datacentre*) 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 two 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[7]**
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%
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%

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

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%).

Of these nine responses, five 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 seventeen 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 …".

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

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%

12. Current staff resource (FTE)

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

Of the eleven HEI-based repositories, all but one provided a response and eight of these specified a figure. The lowest figure provided was 0.25FTE/repository. The highest FTE reported by the nine institutions was 2.4FTE: however, this was staffing for two repositories hosted by the same institution. For the purposes of analysis, I have attributed 1.2FTE to each of the two 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.

Datacentres, possibly predictably, reported much higher staffing levels. Four of six (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 six 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 three 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.

Appendix A: 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.