Research Data Assessment Support

Findings of the 2016 data assessment framework (DAF) surveys

Executive Summary

**Context**

Jisc is developing a pilot research data shared service (RDSS), which is intended to offer easy deposit of data along with features such as discovery, safe storage, long-term archiving, and preservation. Thus, Jisc requires information on the current perception of Research Data Management (RDM) in Higher Education Institutions (HEIs), which are meant to be the primary users of the service.

In order to achieve this, Research Consulting developed a refined version of the existing data assessment framework (DAF) tool and run a survey in six pilot institutions. Such a refinement was achieved by analysing the existing DAF toolkit and RDM surveys previously used in other institutions. The survey was further improved by seeking the feedback of the pilot institutions involved in the RDSS project. The topics analysed in the survey are active data, data preservation, data sharing, and institutional support services.

The results of this survey were analysed by Research Consulting and are gathered in this report. The anonymised survey results are available online[[1]](#footnote-2). In this report, the survey results are analysed in aggregated form, by institution, by role, and by REF panel[[2]](#footnote-3). The analysis by role shows differences between junior and senior members of research staff, while the analysis by REF panels compares the habits of researchers from different fields.

**Active data**

Documents, reports and spreadsheets are the most commonly used types of digital data. Several survey respondents also store non-digital data (58%), with the most common types being notebooks/lab books, paper records/portfolios, and samples.

Use on data management plans remains generally low. The most common reasons for having a data management plan are that it constitutes good research practice (72%) and that it is required by the project funder (53%). On the other hand, reasons for not having data management plans include that it is not required or appropriate for the field of research (47%) or funder (45%), or that researchers lack the skills or knowledge to create one (32%).

Survey respondents expect that their data will be mostly accessed by themselves and other researchers (in the same or in other institutions). Researchers apply security measures to protect their data in 59% of cases, with 41% holding some form of personal or sensitive data. Respondents from REF panels A, C, and D mostly hold personal data about identifiable individuals, while those from panel B mostly hold commercially sensitive information. Where security measures are applied, the most common are password protection of files (59%), physical security (47%), and access logging (45%).

A large share of survey respondents hold less than 50GB of data by volume (40%), however, a minority hold up to 10TB, with one as much as 2PB of data. These data were mostly collected in the past 1-3 years (39%). Only 10% of respondents continue to store data that was gathered more than 10 years ago. Respondents holding at least 501GB of data expect their storage requirements to increase in the next 5 years, either slightly or substantially. Postgraduate students and professors have slightly different habits when storing data. The former prefer external flash drives and hard drives of privately-owned computers, while the latter mostly use hard drives of university-owned computes and university network storage. When using cloud storage, personal accounts are much more common than institutional ones.

**Data preservation**

92% of survey respondents backup at least some of their data, usually personally with a weekly or daily frequency. The most common backup solutions are external drives or memory sticks, university-managed backup storage, and cloud drives.

17% of the respondents reported that they had lost data during their career. In these cases, the most common causes for the data loss were hardware failure, human error, and stolen property. Among the impacts of the data losses the most common is wasted research effort due to the need to replicate research.

Most respondents from REF panels A and B move data with long-term value to a different location for preservation and storage, while this is not as common in panels C and D. The amount of data that the survey respondents believe is worth storing for long-term preservation is not materially different from the total volume held. Survey respondents would generally expect to store their data in an institutional repository, though some respondents indicated they would simply use personal external hard drive for this purpose.

The results suggest that few researchers are aware of digital preservation methods outside of bit-level preservation and that training and awareness raising around preservation is required.

Survey respondents generally do not track their archived data and very few follow guidelines for the preparation of metadata.

**Data sharing**

Sharing of research data is usually handled through the use of cloud storage services (67%), emailing data files (60%), and using portable storage devices (32%). Survey respondents are generally happy to share their data, mostly because research is a public good, there is potential for re-use, and their research findings can be verified independently. For public sharing of data (i.e., data publication), respondents mentioned the use of ArrayExpress, the British Atmospheric Data Centre (BADC), the EMBL Nucleotide Sequence Database (ENA), GenBank, the Gene Expression Omnibus (GEO), Github, the NCBI Sequence Read Archive (SRA), and Zenodo. In these databases, documentation in the form of metadata is usually present.

Among the reasons not to share data, the respondents mentioned confidentiality, issues with permissions, and the desire to hold the data to work more on it. Only a minority of respondents stated that they have already re-used someone else’s data.

**Institutional support services**

Survey respondents showed low levels of awareness of the institutional support services available to them on data management and sharing. Even among those who are aware of these services, most are not currently using them and 10% do not expect to use them in the future.

In terms of training, the largest gaps relate to long-term storage, sharing (including publication), and security of data.

Introduction

Background

Jisc is developing a pilot research data shared service (RDSS), to help Higher Education Institutions (HEIs) and researchers meet the requirements for the deposit and sharing of research data[[3]](#footnote-4). The service is intended to offer easy deposit of data along with features such as discovery, safe storage, long-term archiving, and preservation. The aim of the project is to offer HEIs a managed service, which is expected to simplify the workflow for the preservation of research data.

In order to develop an effective service, Jisc wished to evaluate the current perception and understanding of Research Data Management (RDM) in HEIs. In this respect, Research Consulting was tasked with:

* developing a refined version of the data assessment framework (DAF) tool
* disseminating an online survey based on the refined DAF tool to gather data on existing RDM practices in six pilot institutions.

The survey findings summarised in this report provide an evidence base to benchmark the types of approaches to RDM chosen by the pilot institutions, highlight good practice, predict future trends, and identify relevant requirements for the new service proposed by Jisc.

Terms of Reference

Research Consulting was tasked with refining the DAF tool and using it to benchmark RDM practices in pilot institutions. Our findings aim to enhance RDM provision, establish common ground, and feed requirements to the Jisc RDM Shared Service system suppliers.

More specifically, Research Consulting was tasked with the following:

* Assist and advise pilot institutions in assessing the current status of their research data (numbers, volume, types, complexity, status, etc.) and in using a light weight Data Asset Framework[[4]](#footnote-5)
* Assess the existing arrangements these institutions have in place to support researchers with research data management and planning and make suggestions about how these can be improved and what additional help may be required
* Work with the Jisc team to feed in these findings and understand their implications in the context of the overall RDM Shared Service Project, including implications for the technical development and capacity needs in the RDM Shared Service
* Produce guidance (drawing from and building on good practice) on how to undertake an assessment of data needs so this can be disseminated to other HEI’s and other relevant stakeholders.

This report includes the findings of the DAF survey run in six pilot institutions, while the guidance on how to implement and perform a DAF survey will be produced as a separate document called the “Data Asset Framework Toolkit 2016”. The implications of our work for the implementation of the Jisc RDM shared service will be captured in a separate document, to be delivered to Jisc in late 2016.

Methodology

Design and dissemination of the refined DAF survey

Research Consulting analysed the existing DAF toolkit[[5]](#footnote-6) and RDM surveys that were previously run in some of the pilot institutions. Then, based on these, a master survey template was prepared and shared with all the pilot institutions listed in Appendix A. After receiving feedback from all the pilot institutions, the revised survey template was shared with the following institutions to allow some customisation before its final implementation:

* CREST (consortium of HEIs)[[6]](#footnote-7)
* Lancaster University
* Plymouth University
* The Royal College of Music (RCM)
* The University of Cambridge
* The University of St Andrews

Each of these institutions gave specific feedback on the survey questions, with some elements being customised to individual institutions’ needs. Six versions of the survey were then implemented on the SurveyGizmo[[7]](#footnote-8) platform, with appropriate institutional branding. A full set of the survey questions is available on Figshare[[8]](#footnote-9).

The survey was disseminated in several ways within the institutions:

* Email
* Online bulletins
* Newsletters
* Twitter
* University intranet
* Links on RDM guidance pages
* Library/Research Office blogs
* Staff/PGR portals

In addition, email reminders were sent two weeks and a few days before the survey closure date. For details on dates of the survey in each institution, see Appendix B.

Study and interpretation of the survey results

A total of 1,185 responses were obtained from the institutions and analysed using Microsoft Excel. All views have been treated in confidence in accordance with the Market Research Society's Code of Conduct[[9]](#footnote-10).

The quantitative analysis and narrative observations presented in this report are based on the results of the online survey. The results are sometimes represented based on selected criteria, i.e., by institution, by role, or by REF panel[[10]](#footnote-11). Limited qualitative evidence is presented in the form of quotes obtained from the survey.

Throughout this report, the number of unique respondents for each survey question is described with the symbol “N”. The number of responses is equal to the number of survey respondents when the question type is a radio button or a Yes/No question. On the other hand, when the question type is “tick all that apply”, there are usually more responses than there are respondents.

# Limitations of the Study

This study is based on responses from researchers at six UK higher education institutions of different sizes and levels of research intensity. Given the small sample size, the results in this report cannot be considered be representative for the UK higher education sector as a whole.

In addition, the survey results may not be representative of the researcher population within each of the six institutions due to ‘selection bias’. This arises from the fact that specific groups or categories of researchers, e.g., professors or researchers in the humanities field, are likely to be under-represented in the survey sample. As a result, the survey results may not be representative of the whole population under analysis.

Finally, the survey was run during summer months. This may have caused a lower response rate due to some staff members being on leave.

# Acknowledgements

The outcomes of the 2016 DAF surveys are the result of a consultative process that was possible only thanks to a collaboration between several stakeholders. Therefore, Research Consulting wishes to thank all the RDMSS pilot institutions (see Appendix A) and the Jisc staff who contributed to the success to this work.

Finally, we gratefully acknowledge the original work undertaken by the University of Glasgow and the DCC to develop the DAF survey.

Results of the survey

Data availability statement

The survey results underlying this report have been anonymised and aggregated to form a single dataset. This is openly available online on Figshare[[11]](#footnote-12) along with a guidance document including all the questions asked.

Profile of respondents

Survey respondents from the six institutions held a range of different positions. Postgraduate students accounted for 38% of the respondents and were by far the most represented category. When considered together, senior researchers (Lecturers/research fellows and senior lecturers/research fellows) accounted overall for 43% of the responses, while professors and assistant/associate professors provided 13% of the responses. The remaining respondents primarily held administrative and technical roles.

|  |
| --- |
| 1. Which of the following best describes your role? |

The participating institutions had different numbers of respondents, distributed as follows (see Appendix A):

* The University of Cambridge: 37% (N=440)
* The University of St Andrews: 25% (N=300)
* Plymouth University: 21% (N=252)
* Lancaster University: 10% (N=124)
* CREST: 5% (N=62)

Royal College of Music[[12]](#footnote-13): 1% (N=7)

|  |  |
| --- | --- |
| 1. **Is any of your research activity supported by an external funding body?** | |
| **REF panel A (N=384)** | **REF panel B (N=403)** |
| **REF panel C (N=212)** | **REF panel D (N=186)** |

The field of study of the respondents was assessed by the means of units of assessments within REF panels. The dominant REF panel between the respondents was panel B, covering 34% of the responses. Panels A, C, and D gathered 32%, 18%, and 16% of the responses, respectively. The most represented units of assessment within the panels are shown in Table 1. For a full breakdown, see Appendix C.

The survey respondents were also asked whether they had access to external funding. Researchers at the Universities of Cambridge, St Andrews, and Lancaster and the Royal College of Music are mostly supported by external funding bodies (at least 60% of respondents). On the other hand, Plymouth University and CREST have much lower shares of respondents who have access to external funding (32% and 23%, respectively).

Table 1 – Top 3 fields of study of the survey respondents (by REF panel).

|  |  |
| --- | --- |
| REF panel A (N=384) | |
| Percentage of respondents in the panel | REF unit of assessment |
| 50% | 5. Biological Sciences |
| 19% | 4. Psychology, Psychiatry and Neuroscience |
| 9% | 3. Allied Health Professions, Dentistry, Nursing and Pharmacy |
| REF panel B (N=403) | |
| 21% | 9. Physics |
| 19% | 8. Chemistry |
| 14% | 10. Mathematical Sciences |
| REF panel C (N=212) | |
| 21% | 25. Education |
| 20% | 19. Business and Management Studies |
| 15% | 17. Geography, Environmental Studies and Archaeology |
| REF panel D (N=186) | |
| 23% | 30. History |
| 15% | 33. Theology and Religious Studies |
| 15% | 35. Music, Drama, Dance and Performing Arts |

The answer to this question were further analysed by REF panel (Box 2). Respondents from REF panels A and B are heavily funded by external bodies (71% and 75% of respondents, respectively), while panels C and D report much lower percentages. This reflects the nature of the research conducted in the different panels, with more experimental studies (typical of panels A and B) more likely to require external funding.

The top 10 funders mentioned by the survey respondents were:

1. Engineering and Physical Sciences Research Council (EPSRC)
2. Wellcome Trust
3. European Commission
4. Biotechnology and Biological Sciences Research Council (BBSRC)
5. Economic and Social Research Council (ESRC)
6. Leverhulme Trust
7. Medical Research Council (MRC)
8. Natural Environment Research Council (NERC)
9. Arts and Humanities Research Council (AHRC)
10. Science and Technology Facilities Council (STFC)

Among these funders, EPSRC particularly stands out, being named by more than 25% of respondents.

Active data

Types of research data

For the purposes of the DAF survey, the research data life cycle spans three phases, i.e., generation and use (active data), sharing and publication, and preservation. These phases are reflected in the present section and the next two.

Active data is all the data being generated and used throughout a project or research activity. Digital data is the most common form of data (as opposed to non-digital data) and can be split in a number of different types (see Box 3). The popularity of the different types of digital data generally depends on the field of study and can be summarised as follows:

* Documents/reports are the only type of digital data to be widely used by respondents in all REF panels (at least 70% of responses)
* Spreadsheets are very common across all panels except for panel D
* Computationally generated and sensor data are used mostly in panels A and B.
* Audio files, survey results, and interview transcripts are mostly used in panel C
* Blog posts are common only in panel D

These differences arise from the fact that panels A and B deal mostly with quantitative research (e.g., computational, experimental), while panels C and D usually work on more qualitative types of analyses.

If the same data in Box 3 is studied by institution, the top 2 types of digital data stay the same:

* documents or reports were ticked by at least 70% of respondents in all institutions
* spreadsheets were selected by at least 39% of respondents in all institutions.

For the data automatically generated from or by computer programs and the data collected by sensors/instruments, the situation changes quite dramatically. For the University of Cambridge and the University of St Andrews these types of data are in the same order as in the aggregated dataset. On the other hand, for Plymouth University, Lancaster University, CREST, and the RCM, these types are way further down the list.

Even though digital data is more widespread, a large number of survey respondents reported that they store non-digital data (e.g., notebooks, physical samples field notes, etc.):

* 73% of respondents in REF panel A (N=384)
* 53% of respondents in REF panel B (N=403)
* 49% of respondents in REF panel C (N=212)
* 58% of respondents in REF panel D (N=186)

The top 5 types of non-digital data stored in the participating institutions are (N=552):

* Notebooks/Lab books (74%)
* Paper records/portfolios (56%)
* Samples (43%)
* Consent forms (23%)

Specimens (20%).

|  |
| --- |
| 1. Which types of digital research data are generated from your research? |

Research data management plans

“I am concerned about the balance between ensuring good RDM, and the inevitable time lost/needed to administer that RDM.”

Awareness and use of data management plans (DMPs) remains low. In the surveyed institutions, at least 29% of respondents stated that they do not have a DMP and a further 16% to 20% were unsure (the only exception being the RCM, likely due to the low number of responses).

When the same question is analysed by REF panel (Box 4), it seems that only respondents in REF panel A make frequent use of data management plans. Most respondents from the remaining panels do not currently use data management plans.

|  |  |
| --- | --- |
| 1. **Do you have a data management plan for any of your research projects?** | |
| **REF panel A (N=384)** | **REF panel B (N=403)** |
| **REF panel C (N=212)** | **REF panel D (N=186)** |

The most common reason for researchers to have data management plans is not the mandate from the funders, but the fact that it simply is good research practice to have one (Box 5). About 20% of respondents also report that such a plan is required by either their institution or supervisor/research group leader.

|  |
| --- |
| 1. What are your reasons for having a data management plan? (N=100) |

On the other hand, the most common reasons for not having data management plans are that (N=171):

* it is not required/appropriate to the field of research (47%)
* it is not required by the project funder (45%)

there is a lack of skills/tools to create one (32%)

The majority of respondents have arrangements in place to track and manage their research data (Box 6). The most common solutions are spreadsheets (28%), local databases (26%), and paper log books (22%). It is also important to point out that the adoption of dedicated data management software in the pilot institutions is low (6%).

|  |
| --- |
| 1. How do you currently track and manage your active research data? (N=737) |

Access to and use of research data

Most researchers expect their research data to be accessed or used by other researchers, either in the same or in other institutions (Box 7).

|  |
| --- |
| 1. Who would you expect to access and use your research data, apart from yourself? \_\_ \_(N=547) |

A very large portion of the respondents (48%) reported that they expect that their data will only be used by themselves. Other stakeholders expected to access and use research data include non-academic partners, publishers, funders, or the general public.

Sensitive data

“I would like clear guidance on what kind of data (personal, sensitive, etc.) my research data should be categorised as, along with how it should be securely treated and stored.”

Appropriate management and control of access to research data is crucial when dealing with sensitive data. 59% of the respondents of the DAF survey (N=1185) stated that they apply security measures to protect their data, while only 35% stated that they do not, with 6% unsure whether any measures are applied.

This is possibly a consequence of the fact that a reasonably large portion of the survey respondents (41% of the total) hold some form of personal or sensitive data. The nature of this data varies by REF panel: respondents in panels A, C, and D tend to hold personal data about identifiable living individuals, while those in panel B hold mostly commercially sensitive information (Box 8). However, panels A and C also hold a wider mix of confidential or restricted data, probably due to the nature of the research undertaken in these disciplines.

The most common ways the survey respondents (N=688) protect their data are: password protection of files (59%), physical security (47%), access logging (45%), anonymisation (35%), and encryption (29%).

|  |
| --- |
| 1. Does your data contain any of the following? |

Volume of data

Survey respondents were asked how much data they held (Box 9, Box 10). A large share of respondents from all REF panels stated that they hold less than 50 GB of data, while very few selected the option “more than 1 TB”. Most users having a very large volume of data belong to panels A and B and the most common volume of data above 1TB is 2TB, with 30 occurrences. However, at least 10 respondents stated that they hold 3TB, 5TB, and 10TB of data respectively. Finally, a respondent from the field of physics mentioned holding 2PB (2000TB) of data. This, however, is related to his/her field, as the data comes from the Large Hadron Collider (LHC) and is certainly an exception to the norm. For the purposes of the RDSS, such a high volume of data is out of scope.

The intermediate categories in Box 9 have approximately similar shares of respondents for panels A and B, with those for panels C and D decreasing as the volume of data increases.

|  |
| --- |
| 1. Approximately how much research data do you have at present? (by REF panel) |

Location of data and future storage requirements

“I desperately need access to a substantial storage cloud.”

If the survey responses on the volume of data are filtered by role, the storage needs of professors and postgraduate students can be compared (Box 10). It seems clear that the storage requirements increase with seniority, as there are more than twice as many professors holding more than 501GB of data than there are students. Almost 50% of postgraduates hold less than 50GB of data, which is in line with their junior position within universities. In addition, 20% of professors are not sure about how much data they hold. This value is higher than that found for the aggregated dataset (17%).

The timeframe over which respondents’ data was collected was:

* Within the last 12 months: 26% of respondents
* 1-3 years: 32% of respondents
* 3-5 years: 20% of respondents
* 5-10 years: 11% of respondents
* More than 10 years: 10% of respondents

Therefore, most data held by the survey respondents is rather recent, with 78% of data having been generated in the last 5 years. The shares seen above are roughly similar to those found for the REF panels separately and the only one different from the norm is panel D. In its case, a larger amount of data (39%) has been generated in the last 1-3 years, while lower percentages of 15-16% are found for the other categories.

Respondents from REF panels A and B mostly expect the volume of their data to increase substantially or slightly (more than 70% of respondents) in the next 5 years, with very few forecasting a decrease in their storage requirements (Box 11).

|  |
| --- |
| 1. Approximately how much research data do you have at present? (by role) |

|  |  |
| --- | --- |
| 1. **How do you expect your data storage requirements to change over the next 5 years?**  \_ \_\_\_**(researchers with >500GB of data only)** | |
| **REF panel A (N=100)** | **REF panel B (N=109)** |
| **REF panel C (N=12)** | **REF panel D (N=17)** |

Note that only users with at least 501GB of data were asked this question. Respondents from panels C and D foresee a slightly lower increase in the volume of their data, however, they still expect a reasonably strong growth. In panels C and D, significant shares of respondents (above 25%) expect the volume of their data to stay the same.

Finally, it is important to know where survey respondents hold their data (Box 12). This kind of information is particularly interesting when split by career stage, so that differences between generations of scholars can be seen:

* Professors prefer hard drives of university-owned computes and university network storage
* Postgraduate students prefer external flash drives or hard drives of privately-owned computers
* Usage of cloud services and personal email accounts are more popular among postgraduate students
* A reasonably large share of professors (27%) still hold data on physical media like CDs/DVDs

When considering cloud accounts, further analysis of the data by REF panel and institution indicates the following:

* The use of personal cloud accounts (as opposed to institutional ones) dominates across different REF panels, institutions, and roles
* The use of institutional accounts is particularly high at Lancaster University (31%) and Plymouth University (20%), reflecting the fact these institutions have implemented services from Box and OneDrive, respectively
* At Lancaster University and Plymouth University, the use of combined solutions (personal and institutional accounts) is also high (29% and 36%, respectively)

The use of institutional accounts or combined solutions is highest among professors (22% and 30%, respectively), while PGR students have a clear preference for personal accounts (70%)

In addition, it is relevant to point out that 49% of respondents have at least some data on Dropbox, while 20% have data on Google Drive. However, these services are not appropriate for long-term data-sharing and preservation, and may not meet regulatory requirements for the storage of sensitive and commercial data.

|  |
| --- |
| 1. Please indicate roughly how much of your digital research data is held in each of \_\_ \_ the following locations (N=1185). |

|  |
| --- |
| 1. What is your primary backup solution for your digital research data? |

Data backup and preservation

Backup of data

“Currently, I make semi-regular copies of my files between my work computer, home computer and external hard discs, but this is time consuming and error prone.”

92% of survey respondents (N=1153) stated that they backup at least some of their data, with 56% of the total backing up all of it. The vast majority of respondents take personal responsibility for backing up their data (86% of respondents, N=636). Backups usually occur with a weekly (25%, N=330) or daily (22%) frequency, however, 28% of respondents stated that backups are performed only when they remember. Some respondents stated that their data is backed up by their colleagues or institution (33%). In this case, the frequency of the backup is most often daily (41%, N=146).

The most popular backup solution is the use of external hard drives or memory sticks (Box 13). This solution is followed by university-managed backup storage and by the use of cloud drives. At the University of Cambridge, at Lancaster University, and at the Royal College of Music, the use of university-managed backup storage is particularly high. Little or no use is made of repositories for backup purposes. This reflects the fact their primary role is to host and make public final datasets, and provide a route into preservation services, rather than serving as a backup solution for active research data.

Loss of data

“A lot of knowledge and research data were lost due to bad documentation when PhD students left the group. “

17% of the survey respondents (N=1176) had suffered a data loss at some point in their career. The top 5 causes for data loss reported by the survey respondents are hardware failure, human error, stolen property (e.g., laptop), obsolescence of storage medium, and lost equipment. Hardware failures and human errors were more than 20 times more frequent than the loss of equipment.

|  |
| --- |
| 1. What was the impact of the data loss? (N=191) |

The most frequent impact of data losses (Box 14) was wasted research effort (77%), followed by delays to publication (20%) and reduction in the quality of research outputs (10%). The delays caused by data losses ranged between 1 day and up to more than 3 months, depending on the significance of the loss. A small minority of respondents had suffered a failure to meet funder or regulatory requirements as a result of data loss (<2%).

Long-term preservation

“An open problem seems to be deciding when data no longer has any potential future use.”

When projects are completed, the data generated throughout their life should be considered for long-term preservation. Around half of survey respondents from REF panels A and B stated that they move data with long-term value to a different location for preservation and storage (50% and 47%, respectively). On the other hand, much lower shares of respondents from panels C and D do the same (33% and 37%, respectively).

|  |
| --- |
| 1. Approximately how much of your research data would you consider to have long- …….term value? (by REF panel) |

The volume of data that researchers believe to have long-term value is roughly the same as the total volume of data held (Box 15). This suggests researchers are reluctant to discard any data, and would find it difficult to identify only a subset of their data which is appropriate for preservation. Therefore, it is likely that researchers will simply deposit all their data for preservation into an RDM system, unless awareness and training can encourage them to be more selective.

When the aggregated results in Box 15 are split by career stage, the following conclusions can be drawn:

* Professors are 4 times more likely than postgraduate students to hold more than 1TB of data with long-term value
* Professors are 3 times more likely than postgraduate students to hold 501GB-1TB of data with long-term value
* Postgraduate students are more than 3 times more likely than professors to hold less than 50 GB of data

These figures are in accordance with the career stages of the above-mentioned stakeholders. In addition, Professors tend to be involved in multiple projects at once, and therefore, are expected to be in charge of a larger amount of data.

The survey found 59% of researchers generate non-digital data, with the long-term preservation of this a concern to many. The focus groups found researchers with large volumes of paper who requested preservation of these records beyond simple office storage. This raises the issue that researchers are holding significant paper records that could be converted to digital format by scanning to images of PDF. If encouraged, this would raise the volume of digital data greatly and increase the storage costs of any long-term preservation system. Steps should be taken to mitigate this by using lossless compression of files and identification of files that do not contribute to the overall quality of the data set.

|  |
| --- |
| 1. For how long would you expect this data to be preserved? |

Survey respondents from different REF panels have different expectations on the timeframe for the preservation of their data (Box 16). Respondents from panels A, B, and C generally expect their data to be preserved for 5-10 years (39%-45%), while respondents from panel D expect a period longer than 10 years (38%).

38% of respondents (N=105) stated that they do not remove their data from its original storage location once it has been moved to long-term storage. In addition, only 18% of respondents always delete their original data once it is deposited elsewhere for storage, while 44% do so occasionally.

Most respondents would expect to use an institutional repository for long-term storage (Box 17), which is in line with the purposes of the Jisc shared service. Small differences exist between the REF panels, with more than 20% of respondents from panels C and D responding “Other”. These respondents mostly indicated that they would rely on personal external hard drives for long-term preservation.

|  |
| --- |
| 1. Which of the following would you expect to use for long-term preservation and …….storage? |

The results suggest that few researchers are aware of digital preservation methods outside of bit-level preservation. Although the survey did not explore this area in great depth, follow-up focus groups across the pilot institutions found researchers equate the term ‘preservation’ to mean digital backup to another storage medium, such as a subject repository or university storage. At a minimum, an RDM system should replicate files to a secure long-term medium to deter researchers from relying upon temporary storage such as USB hard drives or flash storage.

Digital preservation methods such as replication of data in separated geographic locations, audits for file format obsolescence and normalisation of files to open formats were not mentioned. The underlying expectation from researchers is that data they submit for preservation, will be available in the future for reuse and their role is simply to provide the files for ingest. Therefore underlying preservation techniques such as the persistence of the media, how the system keeps their data safe or additional preservation details found in models such as PREMIS[[13]](#footnote-14) may be kept ‘hidden’ from the researcher.

Although there was no explicit mention of digital preservation methods by name, the focus groups revealed researchers are aware that the digital data they generate today, may not be accessible in the future due to version changes in software and licence issues for proprietary software. They were largely concerned with their own reuse of data rather than the reuse by others, but it does highlight the need for digital preservation methods beyond bit-level preservation and the need to raise awareness that preservation is not simply backup.

File transformation to open standards, auditing of the software required to reuse data and the potential reuse of data should all play a part in the workflow of researchers if awareness of preservation is to improve. Although, the survey revealed low awareness at present, it is expected this situation will improve as funding and journal data sharing policies mandate sharing of data outputs and awareness of this increases amongst researchers.

Tracking of archived data

Most survey respondents do not formally track or document their archived data (Box 18). Among the respondents who do track the location of their data, the most popular approaches are the use of a local database (24%) or a spreadsheet (20%). At present only a small minority use online profiles such as ORCID (7%) or institutional systems (6%) for this purpose.

In addition, it appears that the documentation of archived data is not a uniform practice, and this is often done in an unstructured way. The proportion of respondents who prepare metadata for their work following pre-defined guidelines are:

* 24% in REF panel A (N=380)
* 17% in REF panel B (N=403)
* 14% in REF panel C (N=210)
* 14% in REF panel D (N=183).

Among the respondents preparing metadata for their work, the guidelines mentioned at least twice were:

* University/research group guidelines
* ESRC guidelines[[14]](#footnote-15)
* GIS metadata format[[15]](#footnote-16)
* MEDIN guidelines[[16]](#footnote-17)
* UKDA guidelines[[17]](#footnote-18)

|  |
| --- |
| 1. How do you track the location of your archived data? (N=244) |

Data sharing and publication

“We run our own server and backup to manage our 5TB of data. We are interested in some of our data being open, and some closed. “

The most common approaches to data sharing among the survey respondents (Box 19) are the use of email (60%), cloud storage services (35%), and portable storage devices (20%). Comparing the data sharing habits of postgraduate research students and professors shows that:

* Students are 9% more likely to email data files
* Professors are 10% more likely to use cloud storage services

Professors are more than 3 times more likely to share files on academic social networks

|  |
| --- |
| 1. How do you currently share data with others? |

For public sharing (publication of research data), survey respondents mentioned the use of:

* ArrayExpress
* BADC
* EMBL Nucleotide Sequence Database (ENA)
* GenBank
* Gene Expression Omnibus (GEO)
* Github
* NCBI Sequence Read Archive (SRA)
* Zenodo

These databases were present at least twice in the survey data and were collected through a free text question with a low number of responses (N=78). Therefore, it is likely that other pathways to data publication exist but were not mentioned by survey respondents. Note that the above-mentioned databases include complete documentation of the hosted data (metadata). This suggests that, at least in certain fields of research, data documentation is a widespread practice.

Survey respondents are generally inclined towards sharing data (N=929):

* 33% stated that they already share data
* 35% stated that they expect to share data in the future but don’t currently

Only 10% stated that they would not want to share their work

|  |
| --- |
| 1. What motivates you to share your data publicly? (N=630) |

When asked why they share data (Box 20), survey respondents mostly agree that there are three main reasons for it:

* Research is a public good and should be open to all (74%)
* There is some potential for data re-use (72%)
* It is possible to verify research findings (59%)

However, many respondents also mentioned increased citations and impact and safeguarding research integrity among their motivations for sharing data. On the other hand, time or efficiency savings, which are usually advertised as one of the main motives for data sharing, were not among the most frequent choices.

In addition, most survey respondents stated that they still have not seen any practical advantages from sharing their data. In a few cases, however, data sharing had resulted in independent verification of findings, increased citations, or new collaborations.

In fact, while most respondents recognise the benefits of sharing data, only a minority have actually re-used someone else’s data:

* 28% in REF panel A (N=383)
* 34% in REF panel B (N=402)
* 21% in REF panel C (N=208)

15% in REF panel D (N=182).

Finally, the top 5 reasons against data sharing (Box 21) are the presence of sensitive or confidential data, the lack of permissions, and the potential for the re-use of data within the research group that has generated it.

|  |
| --- |
| 1. Why wouldn't you be willing to share your data publicly? |

Respondents from REF panel C are particularly concerned about sensitive/confidential data (64%), which is in line with the type of research they conduct. On the other hand, respondents from panel A and B are less likely to want to keep their data for further research. A small but significant minority of researchers across all fields have reservations about their data being used for commercial purposes.

Institutional support services

“I liked the fact that someone emailed me at the end of my project to check that I was going to archive the data as per EPSRC guidelines.”

Survey respondents seem interested in a number of topics for training (Box 22). It appears that the largest knowledge gaps for survey respondents relate to the long-term storage of data, sharing (including publication), and security. In addition, training on the development of data management plans seems desirable, reflecting the growing tendency for research funders to require these.

|  |
| --- |
| 1. Would you value training on any of the following? (N=875) |

|  |  |  |
| --- | --- | --- |
| 1. Do you expect to make use of University services designed to support data …….management and sharing?   **Aggregated data (N=928)** | | |
| **University of Cambridge (N=438)** | **University of St Andrews (N=298)** | **Plymouth University**  Question not asked |
| **Lancaster University (N=123)** | **CREST (N=62)** | **Royal College of Music (N=7)** |

Awareness of the institutional services available to support data management and sharing remains low (Box 23). Respondents from the universities of Cambridge and St Andrews have the lowest levels of awareness of available services (39%), while respondents from Lancaster University, CREST, and the Royal College of Music show higher levels of awareness. Most survey respondents, however, are not currently using university support services (less than 23% in all institutions except for the Royal College of Music) and a small subset of respondents (10% in the aggregated survey results) do not expect to use them in the future.

1. Pilot Institutions

The following table includes a list of all the pilot institutions involved in the Jisc RDM shared service project.

Table A1 – Institutions taking part in the Jisc RDM shared service project.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Institution | Typology | Research Income | Staff | RDM Systems | Publications | CRIS |
| University of Lincoln | Business engaged | 4% | 855 | - | EPrints | - |
| University of Plymouth | Business engaged | 5% | 1285 | - | Dspace | Symplectic Elements |
| CREST | Consortium | - | - | - | EPrints | - |
| Middlesex University | Modern and diverse offerings | 2% | 950 | - | EPrints | - |
| Cardiff University | Research intensive | 21% | 3295 | GlusterFS, Converis | Eprints, Converis | Converis |
| Imperial College | Research intensive | 41% | 4055 | Symplectic, DSpace | DSpace | Symplectic Elements |
| University of Cambridge | Research intensive | 25% | 5430 | DSpace | DSpace | Symplectic Elements |
| University of St Andrews | Research intensive | 20% | 1185 | Pure | Dspace, PURE | Pure |
| University of York | Research intensive | 18% | 1605 | EPrints, Archivematica | Eprints | PURE |
| Royal College of Music | Small and specialist | 1% | 285 | - | - | - |
| St George's Hospital Medical School | Small and specialist | 14% | 510 | - | EPrints | Symplectic Elements |
| University of Lancaster | Vocational and applied research | 14% | 1650 | Pure | EPrints, PURE, Hydra | Pure |
| University of Surrey | Vocational and applied research | 14% | 1390 | Symplectic, EPrints | EPrints | Symplectic |

1. Institutions that took part in the survey

The following table includes a list of the institutions that took part in the pilot of the revised DAF survey.

Table B1 – Institutions taking part in the pilot survey, number of responses, and main contact.

|  |  |  |  |
| --- | --- | --- | --- |
| Institution | Number of responses | Main contact(s) | Role |
| The University of Cambridge | 440 | Marta Teperek | Research Data Facility Manager |
| St Andrews University | 300 | Anna Clements | Assistant Director (Digital Research) |
| Plymouth University | 252 | Elena Menéndez-Alonso | Digital Curator |
| Lancaster University | 124 | Hardy Schwamm | Research Data and Repository Manager |
| CREST | 62 | Matthew Guest  Tijana Close | Senior Policy Advisor (Research & Innovation)  CREST Research Network Coordinator |
| The Royal College of Music (RCM) | 7 | Emma Hewett | Research & Knowledge Exchange Manager |

Table B2 – Survey opening and closure dates in each participating institution.

|  |  |  |
| --- | --- | --- |
| Institution | Survey opening date | Survey closure date |
| The University of Cambridge | 07/07/2016 | 01/08/2016 |
| St Andrews University | 04/07/2016 | 01/08/2016 |
| Plymouth University | 27/06/2016 | 15/08/2016 |
| Lancaster University | 13/07/2016 | 01/08/2016 |
| CREST | 07/07/2016 | 05/08/2016 |
| The Royal College of Music (RCM) | 04/07/2016 | 01/08/2016 |

1. Field of study of survey respondents

Table C1 – REF unit of assessments of the survey respondents.

|  |  |
| --- | --- |
| REF panel A (N=384) | |
| Percentage of respondents in the panel | REF unit of assessment |
| 50% | 5. Biological Sciences |
| 19% | 4. Psychology, Psychiatry and Neuroscience |
| 9% | 3. Allied Health Professions, Dentistry, Nursing and Pharmacy |
| 9% | 1. Clinical Medicine |
| 8% | 2. Public Health, Health Services and Primary Care |
| 5% | 6. Agriculture, Veterinary and Food Science |
| REF panel B (N=403) | |
| 21% | 9. Physics |
| 19% | 8. Chemistry |
| 14% | 10. Mathematical Sciences |
| 13% | 11. Computer Science and Informatics |
| 7% | 7. Earth Systems and Environmental Sciences |
| 7% | 12. Aeronautical, Mechanical, Chemical and Manufacturing Engineering |
| 7% | 13. Electrical and Electronic Engineering, Metallurgy and Materials |
| 6% | 14. Civil and Construction Engineering |
| 5% | 15. General Engineering |
| REF panel C (N=212) | |
| 21% | 25. Education |
| 20% | 19. Business and Management Studies |
| 15% | 17. Geography, Environmental Studies and Archaeology |
| 9% | 23. Sociology |
| 9% | 21. Politics and International Studies |
| 6% | 18. Economics and Econometrics |
| 4% | 20. Law |
| 4% | 24. Anthropology and Development Studies |
| 4% | 22. Social Work and Social Policy |
| 3% | 16. Architecture, Built Environment and Planning |
| REF panel D (N=186) | |
| 23% | 30. History |
| 15% | 33. Theology and Religious Studies |
| 15% | 35. Music, Drama, Dance and Performing Arts |
| 15% | 34. Art and Design: History, Practice and Theory |
| 12% | 29. English Language and Literature |
| 7% | 28. Modern Languages and Linguistics |
| 4% | 31. Classics |
| 4% | 27. Area Studies |
| 3% | 32. Philosophy |
| 3% | 36. Communication, Cultural and Media Studies, Library and Information Management |

1. Please see <https://dx.doi.org/10.6084/m9.figshare.3796305> for more information. [↑](#footnote-ref-2)
2. The Research Excellence Framework (REF) is the system for assessing the quality of research in UK higher education institutions. In the most recent REF, which took place in 2014, Institutions were invited to make submissions in 36 [units of assessment](http://www.ref.ac.uk/panels/unitsofassessment/) (UOAs), grouped into four main panels. Please see [http://www.ref.ac.uk/panels/unitsofassessment](http://www.ref.ac.uk/panels/unitsofassessment/) for more information. [↑](#footnote-ref-3)
3. Please see <https://www.jisc.ac.uk/rd/projects/research-data-shared-service> for more information. [↑](#footnote-ref-4)
4. Please see <http://www.data-audit.eu/> for more information. [↑](#footnote-ref-5)
5. Please see <http://www.data-audit.eu/> for more information. [↑](#footnote-ref-6)
6. Please see <http://www.crest.ac.uk/member/> to find out more about CREST member institutions. [↑](#footnote-ref-7)
7. Please see [www.surveygizmo.eu](http://www.surveygizmo.eu) for more information. All data is held within the EU. [↑](#footnote-ref-8)
8. Please see <https://dx.doi.org/10.6084/m9.figshare.3796305> for more information. [↑](#footnote-ref-9)
9. Please see <https://www.mrs.org.uk/standards/code_of_conduct> for more information. [↑](#footnote-ref-10)
10. Please see [http://www.ref.ac.uk/panels/unitsofassessment](http://www.ref.ac.uk/panels/unitsofassessment/) for more information. [↑](#footnote-ref-11)
11. Please see <https://dx.doi.org/10.6084/m9.figshare.3796305> for more information. [↑](#footnote-ref-12)
12. Note that, from this point onwards, results about the Royal College of Music may not be representative due to the very low number of responses that were gathered. [↑](#footnote-ref-13)
13. Please see <http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf> for more information [↑](#footnote-ref-14)
14. Please see <http://www.esrc.ac.uk/files/about-us/policies-and-standards/esrc-research-data-policy/> for more information. [↑](#footnote-ref-15)
15. Please see <https://www.fgdc.gov/metadata/geospatial-metadata-standards> for more information. [↑](#footnote-ref-16)
16. Please see <http://www.oceannet.org/marine_data_standards/medin_data_guidelines.html> for more information. [↑](#footnote-ref-17)
17. Please see <http://www.data-archive.ac.uk/create-manage/document/overview> for more information. [↑](#footnote-ref-18)