Research Data Management at SGUL

Web survey analysis: Sept – Nov 2015

Jenny Basford
researchdata@sgul.ac.uk

January 2016

Contents

Summary

This report describes the findings of the SGUL Research Data Management (RDM) survey, conducted over three months in 2015, which investigated existing institutional data management practice and knowledge. The questionnaire drew upon the Jisc-funded Data Asset Framework (DAF) methodology and advice from the Digital Curation Centre.

In accordance with best practice, it utilised a similar format to surveys of institutional research data carried out by the London School of Hygiene and Tropical Medicine, the University of Bath, the University of Leeds and the University of Nottingham.

The online survey received 86 responses (81 in the main survey plus five from the RDM Working Group during a pilot phase). Respondents included both staff and PhD students from all SGUL Research Institutes.

Key points:

1. Although most researchers currently produce relatively small amounts of data at present (under 10GB per project), many reported problems with the present allocation of active data storage and would like this to be increased
2. Respondents revealed that the majority use the SGUL shared network drives for storage and to collaborate internally but a significant number also used cloud-based sharing systems such as Dropbox, Box or Amazon Web Services, to share data. Although researchers may not transfer confidential data via these means, it demonstrates a need for quick and easy to use cloud-based storage and collaboration. SGUL OneDrive does not appear to have had a significant uptake in use thus far
3. Most respondents expected to share their research data, with over half predicting this would be via restricted access granted by the project's PI, although a quarter of respondents are preparing to deposit their data in a third-party service or subject data repository
4. Training needs with regards data management were evenly split between guidance on data management plans, funder requirements, ethics and consent issues and data storage

Overview

Rationale

The survey was intended to determine current research data management practices throughout SGUL for the RDM Working Group (RDMWG). These findings will be used to inform decisions and recommendations on institution-provided training, policy and service provision for RDM.

Aims:

1. To investigate current data management knowledge and practice at SGUL
2. To identify particular issues and challenges faced by researchers, both managing active data and sharing archived data

The online survey will be followed up with a number of more in-depth interviews with researchers from all RIs to inform the design of the new SGUL Research Data Management Service.

Survey design and implementation

The online survey was designed in accordance with advice from several sources including the Data Asset Framework and guidance from the DCC. It is similar to those from the London School of Hygiene and Tropical Medicine and draws upon those conducted by Bath, Leeds and Nottingham. The survey was piloted by the RDMWG before going live to the rest of the institution between September and November 2015. It was promoted extensively by a mailshot to all research staff, via the Research Institute Managers, at RI seminars, the JREO Research Grants Day and in the George's Weekly email as well as personal emails to around sixty researchers as identified by the RDMWG. The survey was structured into five sections.

Survey sections:

1. About You
2. About Your Research
3. Research Data Security
4. Sharing and Collaborating with Research Data
5. The SGUL Research Data Management Service

Findings

Section 1: 'About You'

The online survey garnered 86 responses, although not all were completed. The three largest Research Institutes were fairly equally represented, with 11% of respondents from IMBE also contributing. In total, this is around a third of all academic (including research-only) staff at SGUL and is one of the highest responses to RDM audits conducted by other institutions.

At the request of the RDMWG, responses were not anonymous and providing your name and Research Institute were mandatory questions to help analysis of the survey results and to follow up any particular RDM-related queries that respondents may have flagged in the open-ended questions.

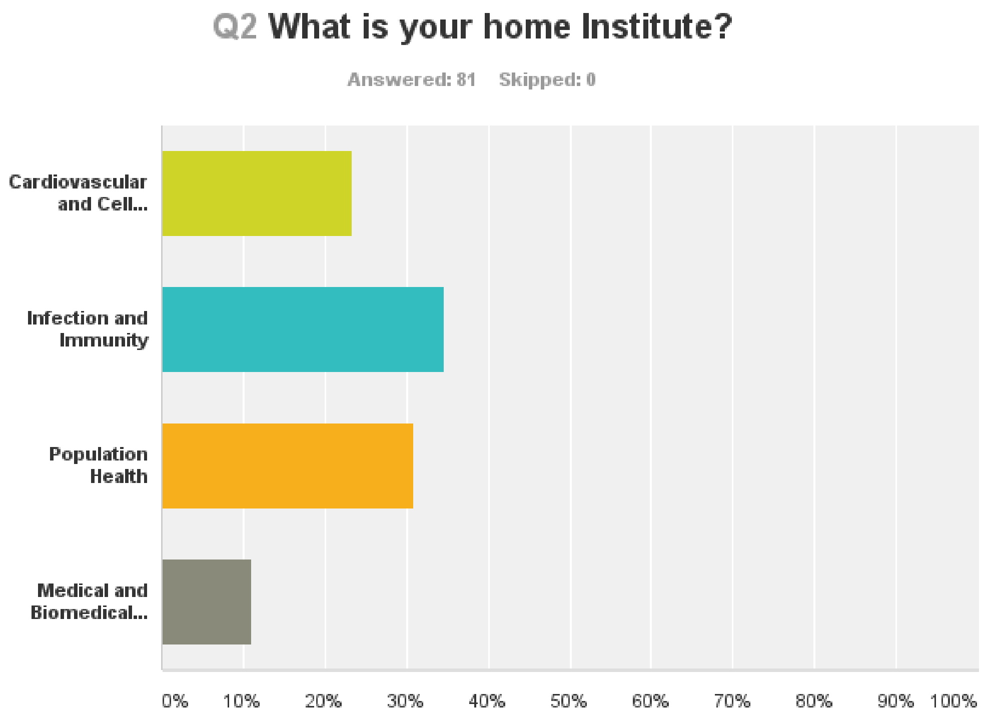


Fig. 1 *Q2, Survey respondents by Institute*[1]

Respondents were asked about their sources of funding, as some organisations have specific data policies and as such it was important to identify any that were at risk (e.g. EPSRC). The survey revealed that SGUL researchers are funded from a variety of different external sources, with the National Institute for Health Research (NIHR) and Medical Research Council (MRC) being the most popular sources of grant support (27% and 26% respectively). Otherwise, researchers had chiefly won funding from charitable organisations (over 50% of respondents) such as the Wellcome Trust, Cancer Research UK, the Cancer Vaccine Institute, the British Heart Foundation and Arthritis Research UK. Cancer Research UK and other charities (particularly those aligned to the COAF group) have data sharing policies and expect data to be shared in a timely fashion, as with RCUK.[2]

[1] Results shown in graphs in this report show only the results from the main survey, not those of the RDMWG, as this conducted in a separate survey.

[2] DCC, 'Overview of funders' data policies', http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies (accessed 17/12/15)

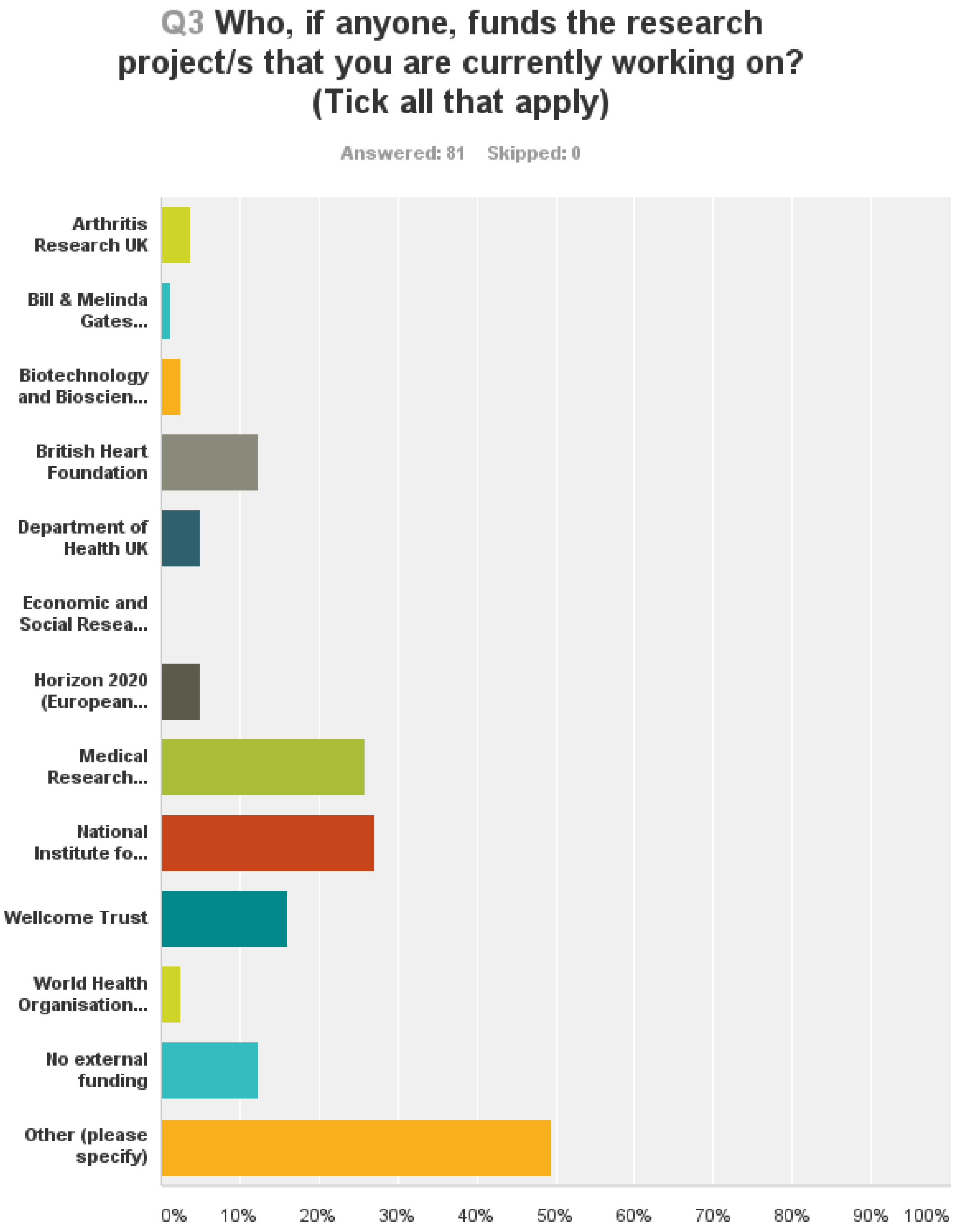


Fig. 2, *Q3, Sources of funding*

Section 2: ‘About your data’

‘Research data’ can mean different things for different disciplines. The survey asked respondents to list the variety of data that they produce, collect and work with generally. Respondents were encouraged to tick all that were applicable and to provide a description of any other types. This question revealed a very rich and complex range of data types, both digital and print.

Laboratory data was the most common type, with 61% of respondents noting that they either created or accessed this during their work. 50% of respondents used observational data, and a close 49% used paper lab notebooks (27% used electronic lab notebooks). Questionnaires featured in 47% of respondents' research. HSCIC data was used by 28% of researchers, although it was noted in the 'other' section by a further six respondents that they used patient or clinical data from their own practice, e.g. case report forms. At the lower end of the scale, audio and photograph files were used by fewer respondents (12% and 16% respectively).

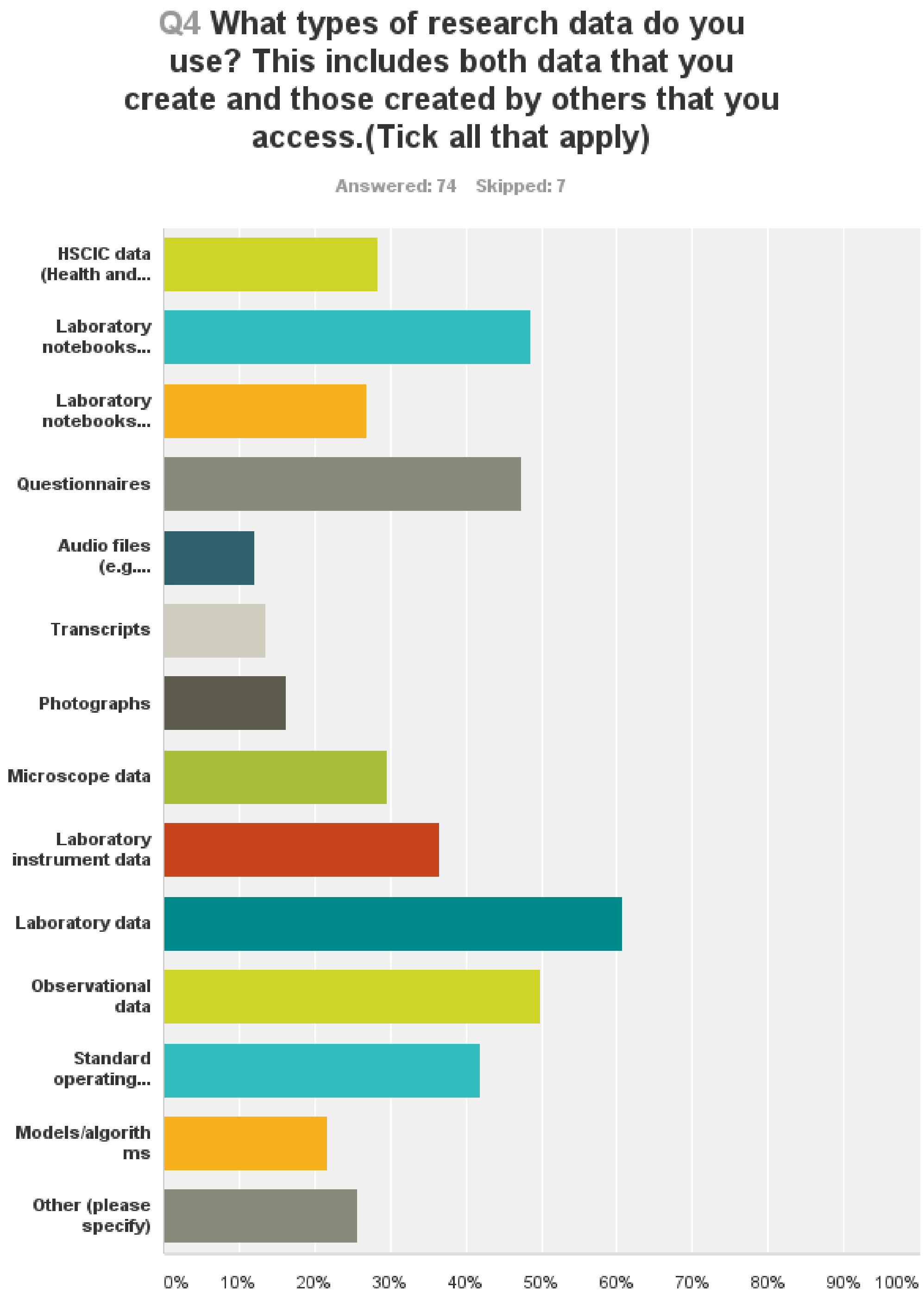


Fig. 3, *Q4, Research data types*

Virtually all respondents (82%) created data during the course of their work (as opposed to interrogated data from other sources, such as HES data). Over half of respondents noted that they used personally-identifiable information at some stage of their projects, with a further 14% noting that they worked with such data on occasion for specific projects.

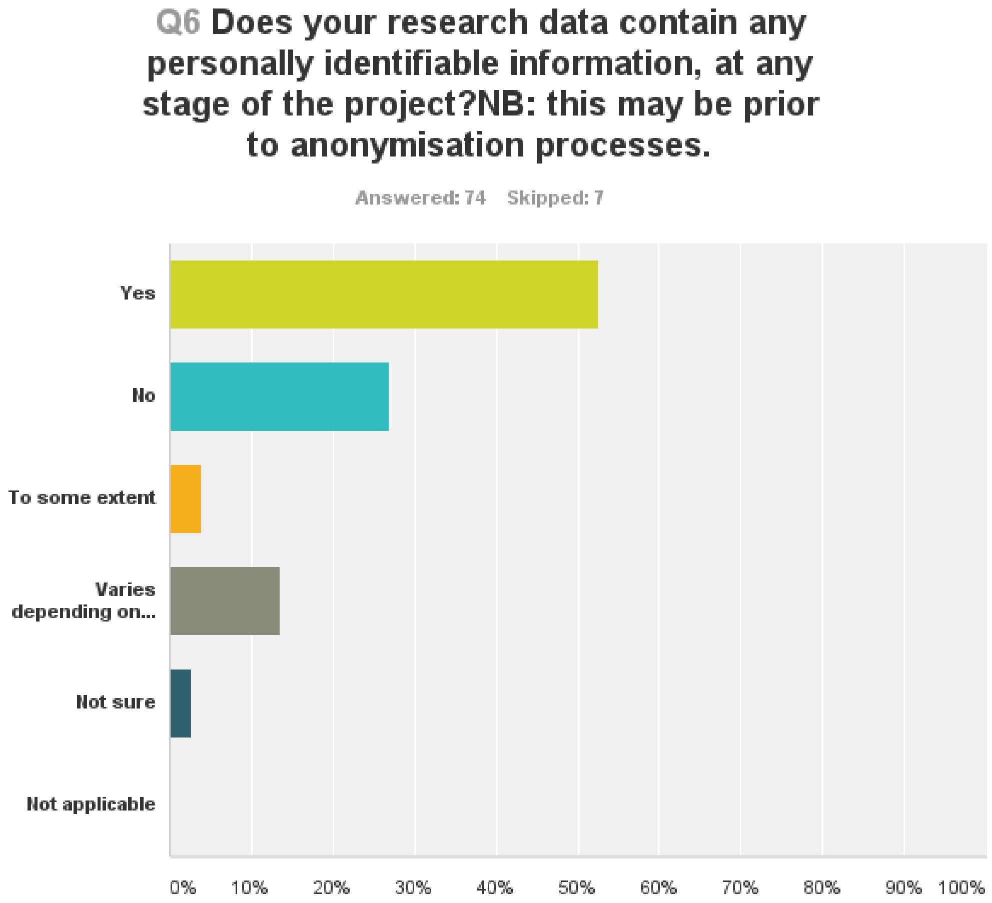
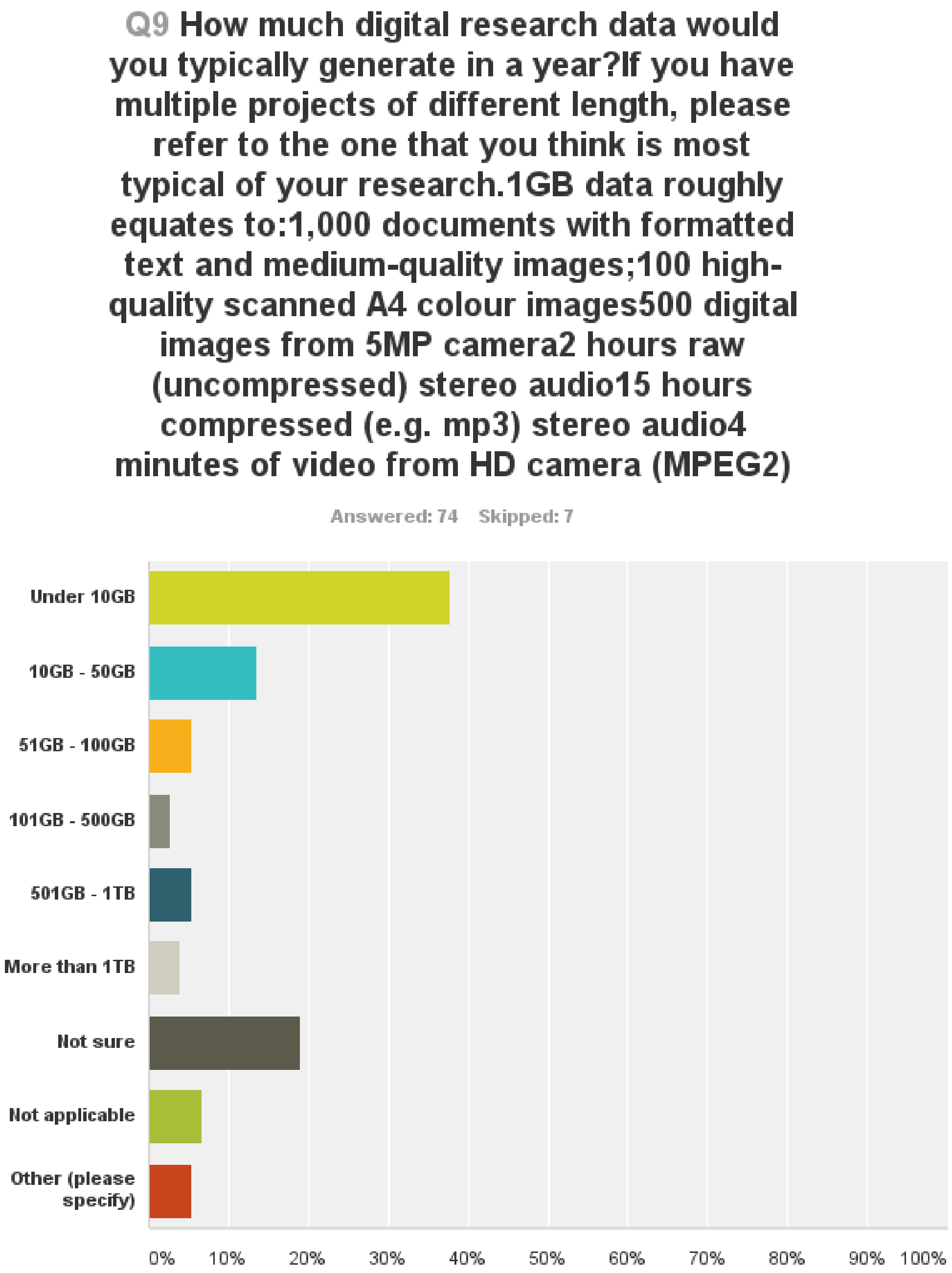


Fig. 4, *Q6, Use of personally identifiable data*

The survey asked respondents to list the three pieces of software that they used most frequently in the creation and analysis of their research data. This information helps us to predict the types of data that our researchers would deposit in an SGUL data registry and repository. As has been found by other institutions' data audits, Microsoft Office was by far the most frequently-used programme, with Excel receiving 50 mentions, Word 24 mentions and Access 12 (respondents also referred to the entire Office suite as a collective). STATA (22 answers), GraphPad Prism (18 responses) and SPSS (16) were the next most popular programmes. There was also a variety of visual data analysis-related software, including ImageJ, Image Studio, Illustrator, GIMP and CorelDraw and other analytical packages such as MATLAB, Winfluor and REDCap, and single mentions for approximately eighteen other analytical and processing programmes.

To uncover more detail about expected deposit sizes for a data repository, the survey asked, 'How much digital research data would you typically generate in a year?' and provided examples to help researchers quantify their data. 19% of respondents were unsure about how much data they produced.

In line with similar surveys at other institutions, most researchers generated a reasonably small amount of data – less than 10GB – per year, although as noted in the summary, many expressed a desire elsewhere in the survey for a greater amount of active storage than provided. There was also concern from some respondents that they expected this to increase in size significantly, with 14% of replies showing that they produced between 10GB and 50GB of data yearly and 11% producing between 101GB to over 1TB of data on an annual basis.



Q9 How much digital research data would you typically generate in a year?If you have multiple projects of different length, please refer to the one that you think is most typical of your research.1GB data roughly equates to:1,000 documents with formatted text and medium-quality images;100 high-quality scanned A4 colour images500 digital images from 5MP camera2 hours raw (uncompressed) stereo audio15 hours compressed (e.g. mp3) stereo audio4 minutes of video from HD camera (MPEG2)

Answered: 74 Skipped: 7

Fig. 5, *Q9, Amount of research data generated yearly*

The survey also asked researchers to outline all the places in which they stored their active data, i.e. data that are still being generated or analysed, rather than being prepared for long-term preservation. While the majority (72%) used the shared SGUL network drives developed for this express purpose, 55% of respondents stored their data on the hard drive of a networked computer, and 22% on the hard drive of a laptop. Respondents show a tendency to use USB sticks to transfer data (38%), and 20% using cloud-based storage services such as Dropbox. SGUL's cloud solution, OneDrive, had a relatively low uptake thus far, with 9% of respondents using it. Around 22% of researchers said they used NHS storage services. Paper, for lab notebooks or questionnaires, is still commonly used (37%) although a separate question (Q8) revealed that the majority of researchers do not usually digitise their paper data.

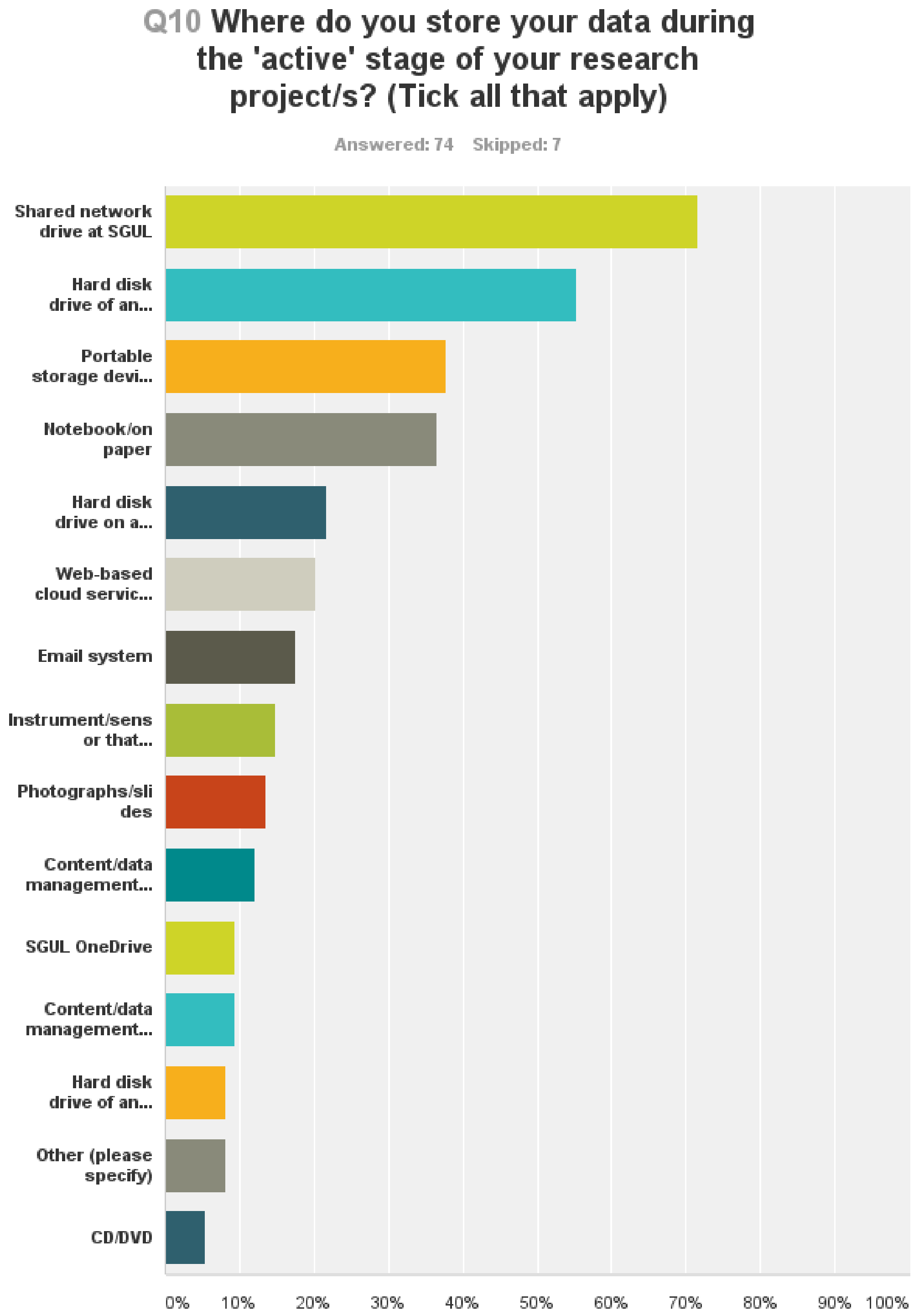


Fig. 6, *Q10, Locations for active data storage*

The survey asked researchers about the frequency with which they backed up their data: ‘How often do you manually back up your digital data, in addition to any automatic back up procedures?' This revealed a split with 26% of respondents answering daily, and 42% answering that they did this on a more ad-hoc process. 9% of respondents revealed that they never manually backed up their data. Comments provided in the 'other' field of this question suggested that those using SGUL or other collaborator's institutional storage (depending on the location of their data), relied heavily on the back up procedures performed by Computing Services each night.

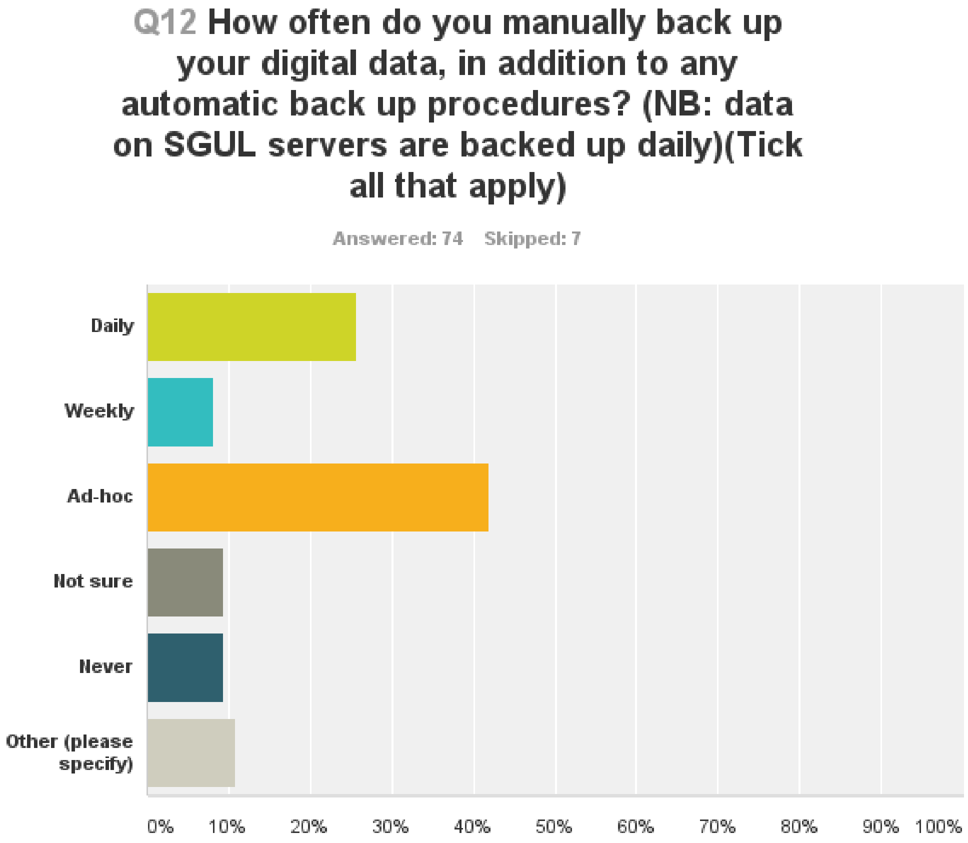


Fig. 7, *Q12, Frequency of manual backups for digital data*

Section 3: 'Research Data Security'

Respondents seemed well acquainted with the legislative and institutional regulations concerning their research data, perhaps unsurprisingly for an institution that uses a high amount of personally identifiable data. 61% of respondents stated that the Data Protection Act influenced the storage, management and sharing of their research data. This was closely followed by 55% of respondents acknowledging the influence of SGUL-specific requirements such as research governance, and 54% of respondents noting the importance of NHS ethics/National Research Ethics Service (NRES) in their work.

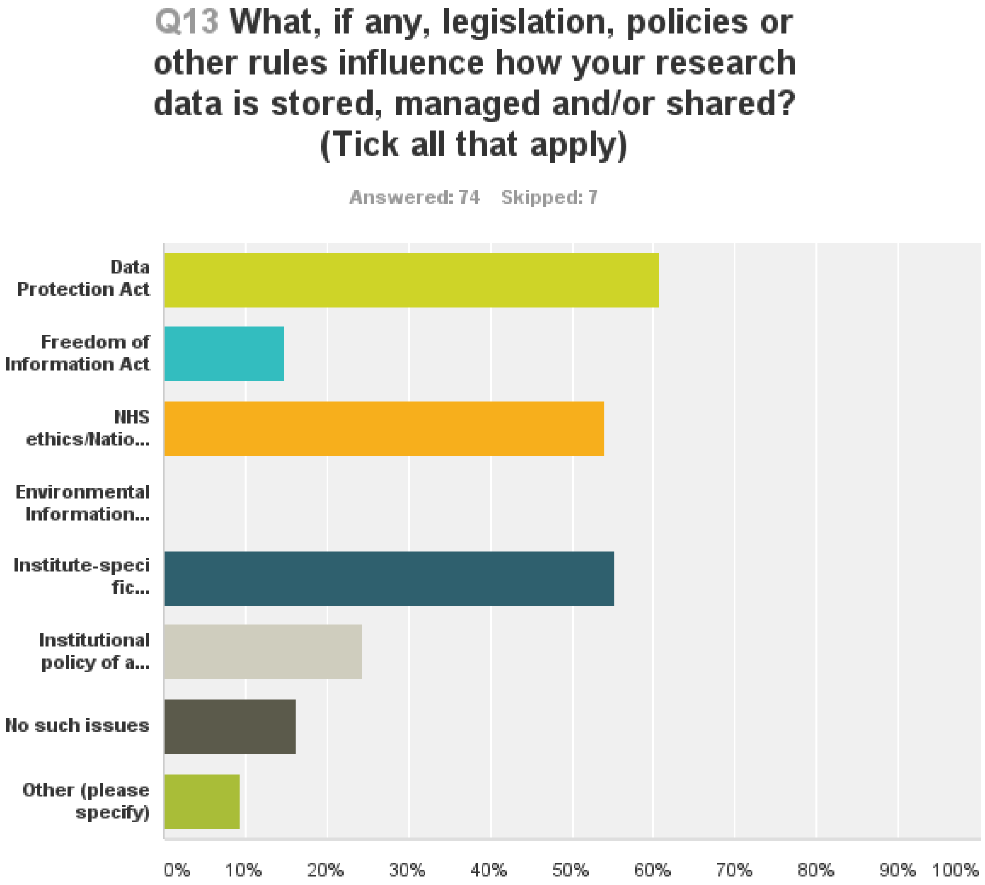


Fig. 8, *Q13, Influence of legislation or policies on data management*

Respondents showed a robust set of various security measures in place to protect their data, and many noted that they used multiple methods to safeguard their data. The survey asked, 'What, if any, security measure/s do you currently employ to all or some of the research data that you create and/or use?'

The majority (64%) used anonymisation, closely followed by 62% that used password protection of files (62%). Over half of respondents (52%) noted that their data was kept in a controlled access area, with 34% using access logging and 30% using encryption. A further two respondents noted that access to relevant drives was restricted. Just 4% of respondents were not sure what security measures they used and 7% said they did not impose any at all.

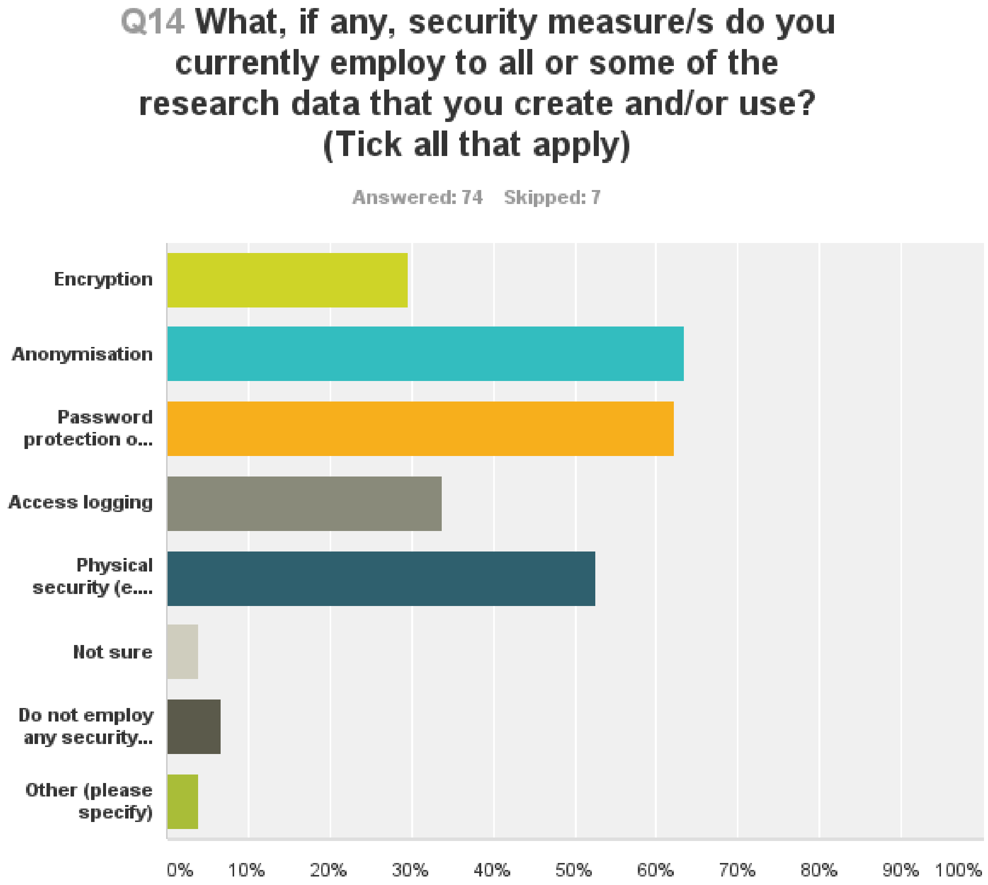


Fig. 9, *Q14, Data security measures taken*

Section 4: 'Sharing and Collaborating with Research Data'

Most researchers had an idea of how long they were required to keep their data for in accordance with their funder's mandate with the majority (30%) asserting that they were obliged to retain data for 6-10 years (most RCUK data policies require a ten-year minimum). 19% were unsure and 23% who had support from multiple sources, said it varied according to their funder. In accordance with the MRC's RDM policy, the SGUL RDM policy requires data to be kept for ten years, after which point it is reviewed. Certain types of data collected during MRC studies require longer preservation, e.g. 25 years for population studies.

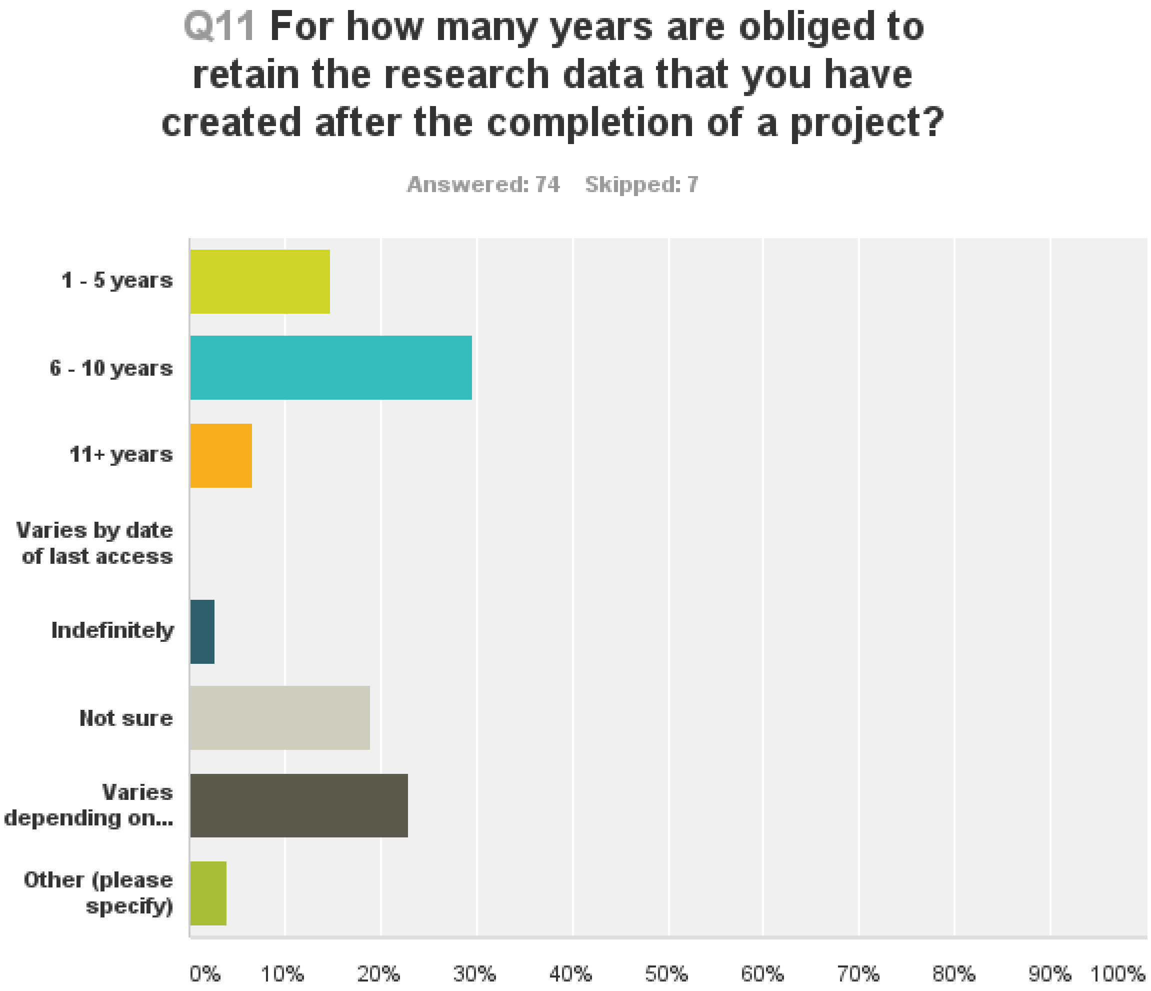


Fig. 10, *Q11, Retaining data*

The survey asked researchers about the possibly multiple entities with whom they share their active data. Unsurprisingly, a significant majority shared it with their project team (96%) and 64% shared with collaborators at other institutions or organisations. 34% shared with other members of their Research Institute (someone not necessarily in their project team). 23% gave data to their research funder.

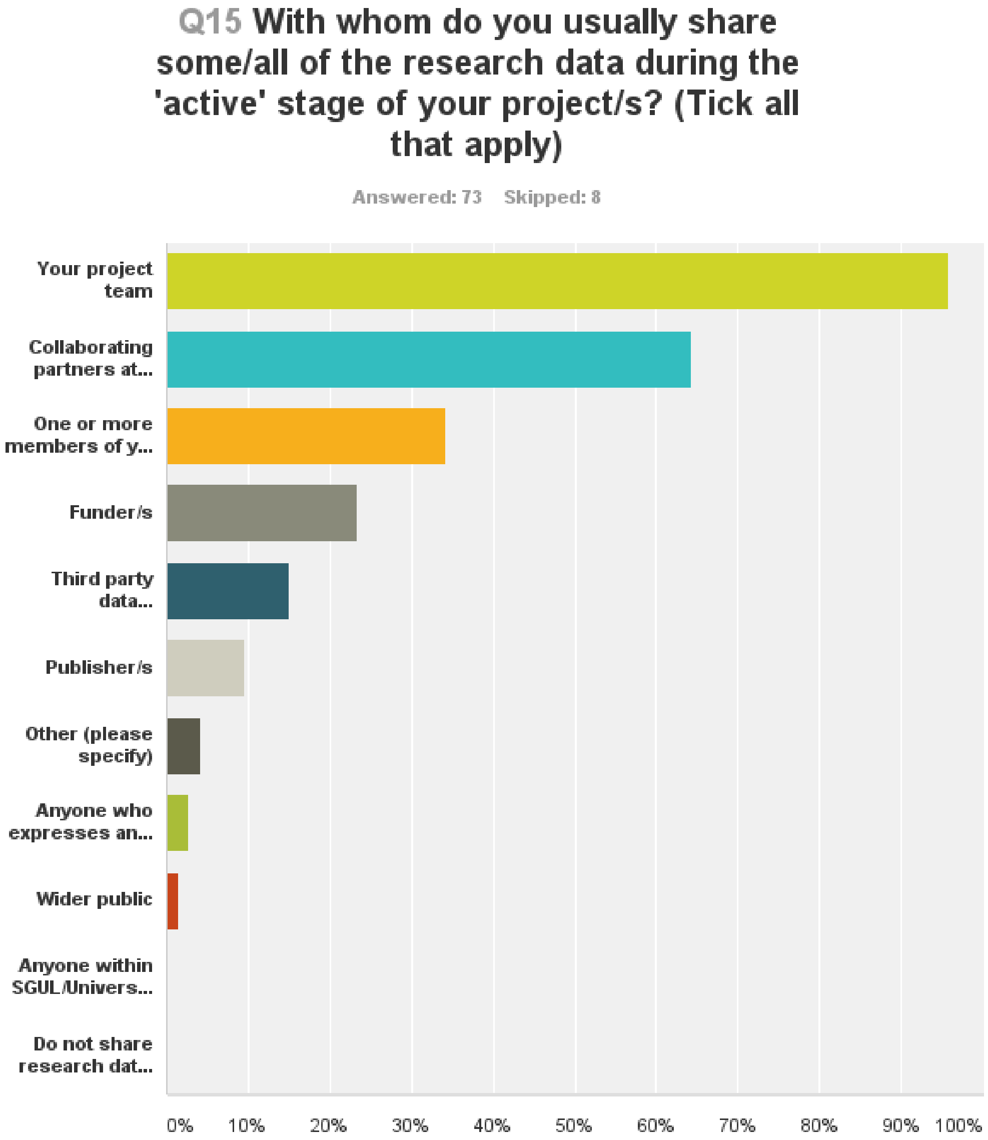
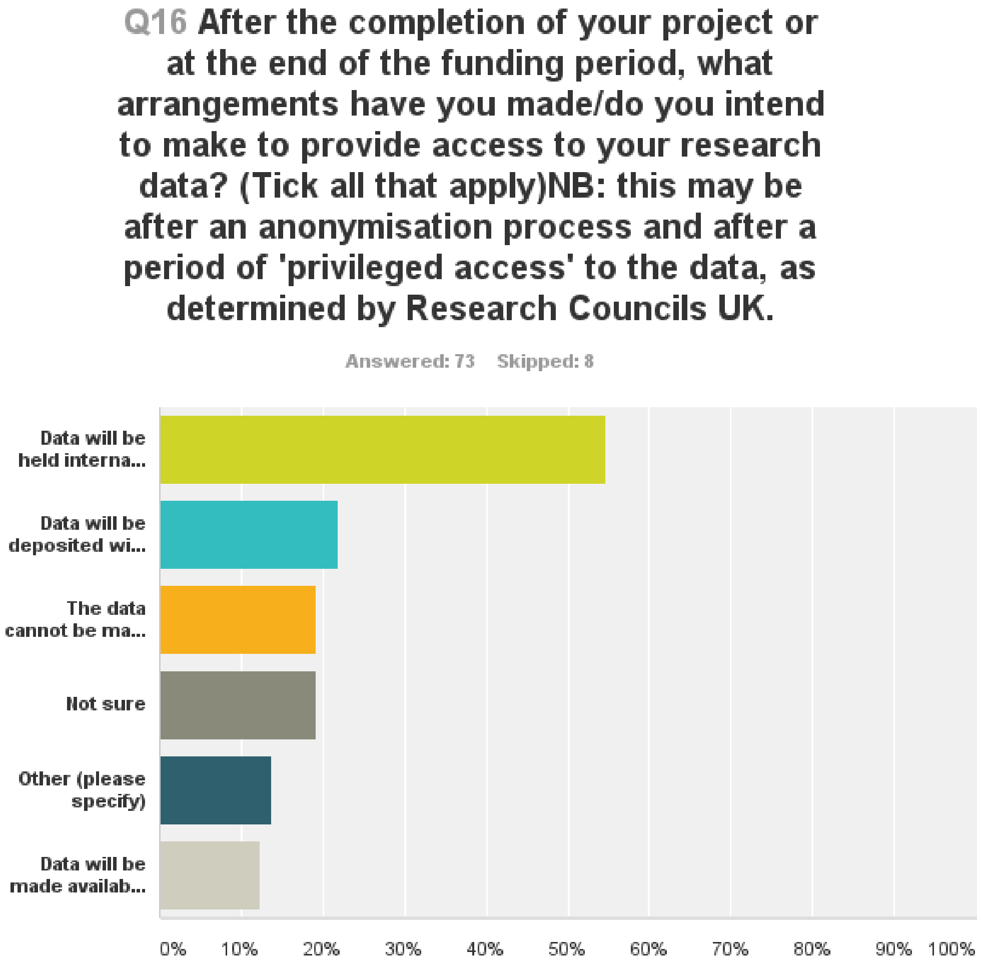


Fig. 11, *Q15, Data sharing*

Respondents were also prompted to share what arrangements they intended to make to share their data after the completion of their project or at the end of the funding period. The survey acknowledged that this may be after a period of ‘privileged access’ to the data to write up publications or to produce follow-on grant applications.

Over half of researchers (55%) stated that their data would be made available subject to access requests and would require approval from the project PI and/or Institute Research Committee. 22% noted that data would be deposited with a third party data service/archive, such as the UK Data Service, figshare, GenBank, or other subject data repository. Just 19% of respondents felt that their data could not be made available under any circumstances and would be restricted to the PI, project team and designated individuals. This suggests that it is important for SGUL to provide a means of discovery for the majority of researchers, such as a data catalogue and/or repository with access request and logging functionality.



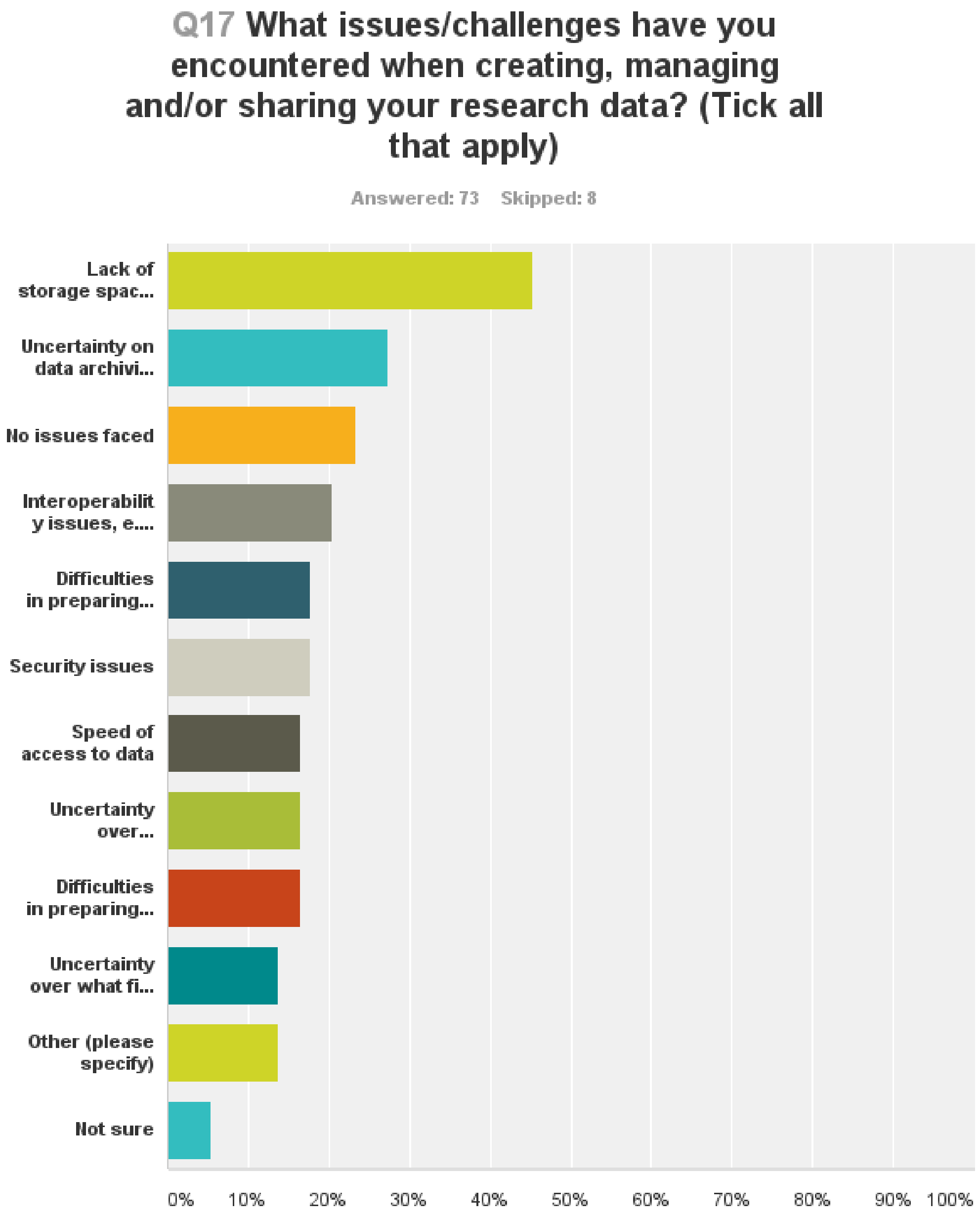
Q16 After the completion of your project or at the end of the funding period, what arrangements have you made/do you intend to make to provide access to your research data? (Tick all that apply)NB: this may be after an anonymisation process and after a period of 'privileged access' to the data, as determined by Research Councils UK.

Answered: 73 Skipped: 8

Fig. 12, *Q16, Long term data sharing (non-active data)*

Respondents were asked about any issues or challenges they had faced in the data management process, including creating and sharing data. Researchers were allowed to select multiple answers. Lack of storage space was by far the most prominent area for concern in data management with 45% of respondents. Uncertainty on data archiving practices was noted by 27% of respondents but by contrast 23% said they had no such issues. Other concerns, such as interoperability problems, difficulties in preparing data management plans, security issues, speed of access to data, uncertainty over documentation standards, difficulties in preparing data sharing agreements and uncertainty over file formats drew roughly the same amount of respondents each (14-18%).

Storage space for active data is a challenging issue across UK HEIs.[3] The RDM Service can provide a conduit to Computing Services via the web pages, giving clear guidance about where to ask for more storage space and any related costs that may be incurred as a result. Training or one-to-one guidance on how and where to archive data is also offered by the RDM Service on request. As SGUL does not yet have an institutional data repository, data may be deposited in subject repositories or other free-to-use locations such as figshare or Zenodo.[4]



[3] See questions on data storage and allocation: Martin Hamilton, ‘Metadata is a love note to the future’ – UK Higher Education Research Data Management (RDM) Survey, http://blog.martinh.net/2013/10/metadata-is-love-note-to-future-uk.html (accessed January 2016).
[4] Figshare, http://figshare.com; Zenodo: http://zenodo.org

Fig. 13, *Q17, Issues/challenges currently encountered in data management*

Respondents were invited to provide more detail about any concerns or difficulties that they had faced in managing their data. These varied between training needs on specific database software or in data management planning, but a recurrent theme was the lack of storage space and difficulties in sharing data, with one respondent noting 'the SGUL VPN is unreliable and difficult to use. OneDrive could be a step forward but it's not as good as DropBox. But for using reasons that are inexplicable to me, using DropBox is frowned upon.'

Section 5: 'The SGUL Research Data Management Service'

The final part of the survey focused on what provision respondents felt the University should provide in order to help them with their data management concerns and to meet expectations of their funding bodies. The vast majority listed increased storage capacity and speeds to improve productivity. Clarification on information security procedures were seen as critical by a considerable number of respondents. A significant number also requested a data repository or archive. Funder-specific templates for data management plans were also popular. Some expressed a desire for training on database software or for more software to be provided institutionally.

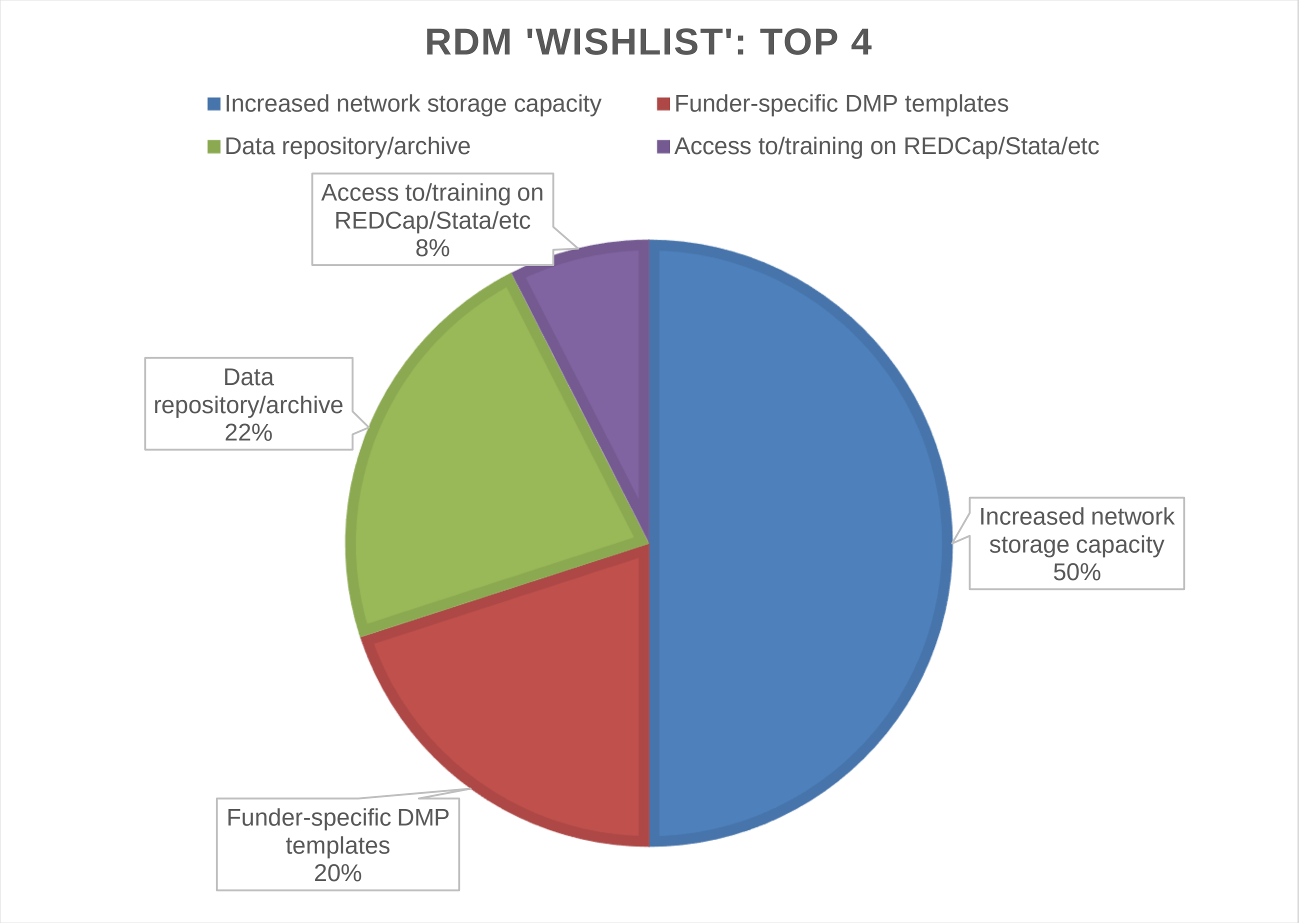


Fig. 14, *RDM 'Wishlist': top 4 most requested facets of an RDM service at SGUL*

The survey closed by asking respondents for their level of interest in training specific aspects of RDM, including writing data management plans, learning about funder requirements for RDM and ethics/consent issues. These were the most popular, particularly the need for clarification on funder requirements for RDM, closely followed by requests for help with writing data management plans. Researchers were less interested in training in documenting data and writing metadata about the data, both of which are necessary for the effective discovery and re-use of data. This suggests that any required metadata fields for data deposit and/or minting of DOIs should be kept to a minimum to reduce the effort required of researchers as much as possible when cataloguing datasets.

Survey conclusions

The survey covered the broad areas of the makeup of the data types created and analysed at SGUL, live and archival data storage, data security, data sharing and training needs.

1. **Data types:** single datasets are rarely excessively large in size, although researchers expect this to increase in the future. Researchers use a wide range of analytical software to interpret data.
2. **Active storage:** this was mentioned more than any other RDM-related issue in the survey. Uncertainty about obtaining additional active storage space and the charging model for this was raised.
3. **Sharing active data:** OneDrive is not yet used by many academics as it is a relatively new service but cloud storage and sharing functionalities are seen as very important by many researchers. OneDrive may not be fulfilling the needs of researchers that are found in commercially-provided alternatives such as Dropbox.
4. **Sharing archived data and digital preservation:** respondents have a positive attitude towards sharing data alongside publications or with other researchers, but many need assistance in doing so. A significant number of respondents indicated an SGUL data repository is essential.
5. **Data security:** generally seen as adequate although researchers engaged in obtaining data from HSCIC have faced considerable problems with SGUL’s IG Toolkit status.
6. **Training needs**: are equally split between writing DMPs, funder requirements, ethics and issues of consent around sharing data.

Recommendations

There are several facets to an effective RDM Service, which can be split into the three main stages of research: before, during and after. The SGUL RDM Service should aim to provide solutions that addresses all these stages, reduces the amount of new workflows for researchers as much as possible and consolidates any existing systems to achieve this goal.

SGUL RDM: Before, During and After phases

Before research begins

1. **Sample/pre-filled DMPs with answers to generic questions** and one-to-one DMP clinics on application to funders (‘Before’ stage)
2. **Training on key areas, particularly DMPs, funder requirements and ethics should be provided by the RDM Service and the JREO.** Training on ethics should continue to be provided by the JREO, whereas the Library will focus on deliver training on DMPs, funder requirements and where/how to deposit and archive data (‘Before’ and ‘During’ stages)

During a research project

3. **Easily accessible guidance on storage of active data and advocacy for existing systems** e.g. OneDrive (if appropriate) (‘During’)
4. The University should **continue to provide backup of data as a key service**, whether in OneDrive or elsewhere on university servers (‘During’)

After a project has finished

5. **A data repository and DOI minting service that is easy for academics to deposit in** – no lengthy metadata or documentation forms to fill out in order to deposit. This must support controlled access as the majority of researchers anticipate their data cannot be wholly ‘open’ (‘After’)
6. Along with the rest of the University, the RDM Service should be **compliant with the NHS Information Governance Toolkit**. Securing this status would significantly facilitate the work of the RDM Service by bringing together (and updating) the relevant institution-wide policies from the Information Directorate and JREO. (All stages)

Appendix 1: RDM Service Implementation case studies

Essential to the latter two stages of the RDM Service is the need for a data management platform. SGUL has been successful in its application to become a pilot institution for the Jisc RDM Shared Service. This will commence in January 2016 and end in September 2017, so interim solutions will need to be provided for researchers. Below are three case studies that demonstrate the range of approaches to the design and delivery of an RDM service currently used by other HEIs.

The modular approach (Imperial)

The 'light touch', modular RDM strategy offered at Imperial College London utilises a number of externally-provided products (some of which their researchers were already using), rather than investing in an 'end-to-end' RDM solution that would effectively be imposed upon their researchers.[5] Imperial uses DMPOnline to guide researchers through data management plans[6]; they advise researchers to use Box for active (cloud) storage (which will be compliant with UK/EU legislation on confidential data from 2016);[7] and suggest that researchers deposit data in, and obtain a DOI for their data from, a subject repository. Where no such repository exists, they recommend that researchers deposit data in the CERN-maintained and EU-funded Zenodo repository.[8] Zenodo mints a DOI for all data deposited within it for no cost, although it would be a Zenodo DOI.

Benefits of this model to SGUL:

This modular approach for provision of RDM tools is not only cost effective but it also decreases the risk of total service failure presented by the use of a totally integrated RDM system. It also enables developments to be implemented to one area of the service at a time, without impinging on the other areas. It is possible to create an institutional community within Zenodo to guide researchers towards.

Challenges of this model to SGUL:

Zenodo is an excellent free tool and makes use of established open standards that promote discoverability of its contents. There are two APIs currently available that would require support from SGUL Computing Services and Symplectic Elements to integrate with SGUL's current research information system (CRIS), which would help to monitor institutional compliance and present a consistent message to SGUL researchers about how to deposit data and publications.

The Elements/figshare/Arkivum approach (Loughborough)

[5] 'Policy guidance | Imperial College London', http://www.imperial.ac.uk/research-and-innovation/support-for-staff/scholarly-communication/research-data-management/imperial-policy/guidance/ (accessed 1/12/15)

[6] http://dmponline.dcc.ac.uk/

[7] 'Data storage | Imperial College London', http://www.imperial.ac.uk/research-and-innovation/support-for-staff/scholarly-communication/research-data-management/imperial-policy/guidance/data-storage/ (accessed 1/12/15)

[8] http://www.zenodo.org/

Like the Imperial model, Loughborough have made use of an already-familiar system to try to reduce the impact on researchers' time when depositing data. Loughborough's RDM Service does not directly address active data storage (as Imperial does) but focuses on delivering a one-stop data discovery and archiving solution for all disciplines that keeps interruptions to the researcher's current workflow to a minimum. Loughborough use figshare to make their data discoverable, sharable and citable and Arkivum to preserve that data and to make it available to users. The figshare/Arkivum platform has been integrated with Symplectic Elements, Loughborough's current research information system (CRIS), so that any data deposited in figshare will automatically be added to a staff member's research profile.

Loughborough researchers deposit data into their figshare data platform and provide a few details to describe them, such as creator, date of creation, keywords and/or a description. [9] Figshare mints a DOI for the data so that they have a persistent identifier and can always be located. The metadata describing the dataset and the DOI are then automatically harvested into the CRIS.

Loughborough researchers therefore have a list of all their outputs, including both publications and data, in the CRIS. Long-term data archiving and preservation needs are fulfilled by Arkivum, with figshare transferring large-scale data and data intended for 'deep storage' which is not expected to be accessed on a regular basis, to Arkivum on Loughborough's behalf. [10]

Benefits of this model to SGUL:

The Elements/figshare/Arkivum method ensures that Loughborough's research data is highly visible, is preserved for the long term and ensures that records are being kept within the University that will facilitate compliance monitoring with funding body data sharing requirements. Figshare has a clean and intuitive deposit process for researchers and have plans in 2016 to develop dashboards to help institutions ascertain compliance and track the impact of their research.[11] Loughborough is working with Symplectic to enable deposit of data into figshare via Elements in 2016 (in the same way in which publications at SGUL are deposited with the publications repository via the CRIS).[12]

Challenges of this model to SGUL:

Use of figshare is free to individual users but institutionally the price is subject to research intensity of the university.[13] The cost is split into an annual subscription to the figshare platform plus storage (from Arkivum). Storage costs vary depending on whether an institution selects the pre-pay or pay-as-you-go model. Although figshare will mint a DOI as part of this service, to obtain an institution-specific DOI for data deposited in figshare as

[9] 'figshare | Loughborough University', https://lboro.figshare.com/ (accessed 14/12/15)

[10] Brewerton, Gary, 'RDM – the Loughborough Solution', https://dx.doi.org/10.6084/m9.figshare.1604781 (figshare, 2015)

[11] 'figshare fest', London 12/11/15

[12] Brewer, Gary and Cole, Gareth, 'Research Data Management Case Study: Loughborough University' (figshare, 2015), https://dx.doi.org/10.6084/m9.figshare.1492975

[13] 'figshare – Institutions', https://figshare.com/services/institutions (accessed 15/12/15)

Loughborough have opted for, a subscription to DataCite is required.[14] SGUL already has an institutional arrangement with Symplectic for the use of the Elements CRIS, although integration of the SGUL CRIS with a data repository may incur a cost.

The ULCC EPrints/Arkivum approach (Reading and Sheffield Hallam)

The University of Reading and Sheffield Hallam University (SHU) shared similar drivers in developing its RDM Service with Loughborough, most pressingly the EPSRC's deadline of 1 May 2015, and as such their focus has been on providing a data archive, although SHU have introduced central – not cloud – data storage for active projects. Both chose to capitalise on their institutional familiarity with EPrints publications repositories to develop a data archive.[15] They worked with the University of London Computing Centre (ULCC) to implement and host an EPrints data repository that could be integrated with Arkivum to hold large datasets and meet their digital preservation requirements, based on ULCC's successful track record with creating data repositories at LSHTM and the University of East London (UEL).[16] Like SHU and Reading, SGUL's publications repository, SORA, is powered by EPrints, although unlike SHU and Reading, SORA is already hosted by ULCC.

ULCC use a standard data repository plugin, EPrints-ReCollect, to create data repositories that are then tailored to each institution's requirements. Reading does not have a CRIS and has no plans to procure one; at present SHU uses only the pre-award module of Thomson-Reuter's Converis CRIS. As a result, rather than linking a CRIS to the repository, data are deposited by researchers or the RDM Service by logging directly into EPrints, using the integrated institutional log-in. Access controls for items in EPrints can be divided into: restricted, registered or open.[17] Restricted items are available only to the depositor and repository staff, whereas registered items can be made available to all institutional users via the log-in and external users need to complete a request form for permission to view it. Open items are available to anybody. If all users are required to register to download data, then it is also possible to track when an item was last accessed and by whom. EPSRC-funded data must be kept ten year from date of last access.[18]

Benefits of this approach to SGUL:

EPrints, like figshare, is a highly visible platform on which to store data, as SGUL is already aware through its use of EPrints for SORA. ULCC are already familiar with SGUL's infrastructure as it hosts SORA, so this may be an efficient solution in terms of integration and set up costs. SORA is already integrated with Symplectic Elements to upload publication records into the repository, so it is likely that this could be developed as a mechanism for researchers to upload their datasets. Like figshare, digital preservation

[14] Brewerton, Gary, 'RDM – the Loughborough Solution', https://dx.doi.org/10.6084/m9.figshare.1604781 (figshare, 2015); Presentation by Gary Brewerton at 'figshare fest', London 12/11/15

[15] http://centaur.reading.ac.uk/

[16] Darby, Robert; Verbaan, Eddy and McNicholl, Rory, 'Institutional case studies – Reading, Sheffield Hallam and ULCC' (RDMF'14, York, November 2015), http://www.dcc.ac.uk/sites/default/files/documents/RDMF/RDMF14/04%20McNicholl%2C%20Darby%2C%20Verbaan%20-%20RDMF14.pdf

[17] 'Decide how access should be provided | LSHTM', http://www.lshtm.ac.uk/research/researchdataman/depositdata/access_permissions.html (accessed 15/12/15)

[18] EPSRC, 'Expectations', https://www.epsrc.ac.uk/about/standards/researchdata/expectations/ (accessed 16/12/15)

issues and storage of large datasets are addressed through the integration with Arkivum. Arkivum storage via ULCC is competitively priced in 1TB instalments.[19]

An existing EPrints plugin, RIOXX, has been created to monitor compliance for open access publications and can be used for data as well. ULCC are heavily involved in the development of EPrints as a data repository: they are working with Lancaster University to integrate the Data Management Administration Online tool (a Jisc-funded project) with EPrints to facilitate reporting on RDM within institutions, and are working with Arkivum on a prototype to remedy the issues of upload/download of large datasets. Both are involved in other Jisc-funded RDM projects via the 'Data Spring' challenge, which demonstrates a willingness to innovate and an understanding of this changing policy landscape.[20]

Challenges of this approach to SGUL:

There is a DataCite plugin to enable minting of DOIs for data deposited in EPrints, which would involve an additional cost to the institution. EPrints is less intuitive to use than other platforms for non-library users, as direct deposit is a multi-stage process of between five to seven screens at Reading and SHU.[21] Integration with Symplectic Elements would be a key part of an EPrints service for SGUL so that there is a consistent message about how to deposit publications and data. There may be a cost to implement this. There may also be additional development and implementation costs if SGUL is the first to commission these extensions to EPrints' current functionality.

[19] ULCC, 'ULCC RDM Solution – Compliance doesn't have to be complicated', (ULCC webinar, October 2015), http://www.slideshare.net/ULCCEvents/ulcc-rdm-solution-compliance-doesnt-have-to-be-complicated (accessed 14/12/15)

[20] Jisc, 'Research Data Spring', https://www.jisc.ac.uk/rd/projects/research-data-spring (accessed 16/12/15)

[21] Darby, Robert; Verbaan, Eddy and McNicholl, Rory, 'Institutional case studies – Reading, Sheffield Hallam and ULCC' (RDMF'14, York, November 2015), http://www.dcc.ac.uk/sites/default/files/documents/RDMF/RDMF14/04%20McNicholl%2C%20Darby%2C%20Verbaan%20-%20RDMF14.pdf

Appendix 2: active data solutions

Researchers reported issues with running out of storage space and in sharing their active data, either with themselves when working off-campus or with collaborators. This matrix summarises the options available to SGUL researchers and the benefits and potential problems that may arise in using these tools.

Solution	Benefits to SGUL researchers	Challenges to integration with researcher practices
SGUL shared drives, e.g. H:\ etc	• Secure enough for medical data – additional security levels can be added for HSCIC/ONS data • Backed up every night and for the long term in a different location (ten years)	• Initial quota is c.300MB per user • Lack of awareness about how researchers can request additional space • Uncertainty about actual cost: no price structure currently in place, so difficult to cost for this in funding application Data Management Plans • Off-site access set-up less simple than commercial storage solutions
SGUL OneDrive (OneDrive for Business)	• 1TB active storage for every user – presently this is enough for the majority of researchers • File sync across devices • Built into each SGUL user account when saving files on an SGUL PC • Real-time online collaboration	• Not appropriate for medical data – at present there is no guarantee that data is kept in the UK/EU • Not backed up every night by SGUL • Deleted data will be not be recoverable by SGUL • Can share documents only with other users with either a Microsoft or SGUL account • Not perceived to be as intuitive as other cloud solutions • Researchers have reported that SGUL OneDrive is much slower than the free personal OneDrive
OneDrive (Personal)		• Files cannot be larger than 10GB • Security levels not yet appropriate for medical data – at present there is no guarantee that data is kept in the UK/EU • Not backed up every night by SGUL • Deleted data will be not be recoverable by SGUL

DropBox (Personal)	• Initial Basic accounts begin with 2GB	• Security not yet appropriate for medical data – at present there is no guarantee that data is kept in the UK/EU and is • Not backed up every night by SGUL • Deleted data will be not be recoverable by SGUL
Box (Personal)	• Not yet appropriate for UK/EU medical data, but will be compliant during 2016 (RDMF'14)	• Initial free quota is 250MB • Not appropriate for medical data – at present there is no guarantee that data is kept in the UK/EU • Not backed up every night by SGUL • Deleted data will be not be recoverable by SGUL

NB: Storage capacity sizes taken from: http://www.cnet.com/uk/how-to/onedrive-dropbox-google-drive-and-box-which-cloud-storage-service-is-right-for-you/ (updated November 2015)

Other options for the institution include ownCloud and SpiderOak, which would encrypt data on University servers before it is transmitted to their storage, making it more secure than DropBox.