IJDC | *General Article*

Introducing Safe Access for Sensitive Data at the University of Bristol

Debra Hiom
University of Bristol

Stephen Gray
University of Bristol

Damian Steer
University of Bristol

Kirsty Merrett
University of Bristol

Kellie Snow
University of Bristol

Zosia Beckles
University of Bristol

Abstract

The economic and societal benefits of making research data available for reuse and verification are now widely understood and accepted. However, there are some research studies, particularly those involving human participants, which face particular challenges in making their data openly available due to the sensitivities of the data. Despite its potential value to society this material is invariably kept locked away due to concerns over its inappropriate disclosure. The University of Bristol's Research Data Service has developed the institutional infrastructure, including policies and procedures, required to safely grant access to sensitive research data in a way that is transparent, secure, sustainable and crucially, replicable by other institutions.

This paper looks at the background and challenges faced by the institution in dealing with sensitive data, outlines the approach taken and some of the outstanding issues to be tackled.

Received 23 February 2017 ~ *Accepted* 4 December 2017

Correspondence should be addressed to Debra Hiom, Augustine's Courtyard, Orchard Lane, Bristol, BS1 5DS. d.hiom@bristol.ac.uk

An earlier version of this paper was presented at the 12th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v12i2.506
DOI: 10.2218/ijdc.v12i2.506

Introduction

The Research Data Service was introduced into the University of Bristol as a core part of the Library Service in 2014 after a three year project/pilot period, as described in Hiom et al. (2015). The service works in close collaboration with a range of groups across the University, including IT Services, Research Enterprise and Development and the Secretary's Office. The Research Data Service provides specialist support for data management across the whole research lifecycle and 'touches' all levels of research, from PGRs to Principle Investigators. The key service areas are data management planning support, advocacy and training, and support for publication of research data.

Background and Challenges

One of the key functions of the service is to provide an online repository to publish the University's research. This was initially conceived as an 'Open' repository i.e. the datasets are made available for download without registration. However, it quickly became apparent that much valuable data could not be shared as 'Open' data. The problem of sharing data in the absence of clear participant consent is well known (Navarro, 2008), and the service received a number of queries from research studies who wanted to find a way to share their data with other researchers but were not clear about the best way to do this in a way that ensured the terms of participant consent were adhered to and compliance with legal and regulatory requirements were met. The types of data included interview transcripts, medical imaging, clinical records and politically sensitive material. Much of this data was generated some time ago, before re-use potential was fully considered and therefore without asking participants for the appropriate permissions to share more widely. Other datasets were impossible to anonymise and some conflated commercial restrictions with ethical challenges.

It was important to introduce the means to securely share this material as access requests were being received regularly which could not realistically be met. The specific challenges were:

- **Data ownership and contractual obligations**: This required a policy which detailed the 'default' ownership of research data and a workflow for identifying contractual obligations, such as agreements with external collaborators, industrial partners and research funding agreements. These must be taken into account before deciding on whether or not to grant access to any particular sensitive dataset.

- **Governance:** Best practice required a nominated group, rather than any one individual, to assess requests and make justifiable and transparent decisions on the sharing of sensitive data.

- **Terms of data release:** The University needed fair and equitable terms of data release which protected participants but did not place unnecessary barriers in the path of future breakthroughs.

- **Storage and technical infrastructure:** Data was scattered across the University but should ideally be collated and stored centrally to facilitate sharing. A secure means to deliver data was also required.
- Going forward, ensuring that data access was considered during the set-up of projects to avoid future difficulties, e.g. through ethics and/or contractual negotiations and providing researcher training and guidance on sharing sensitive materials.

Approach Taken

Data Access Task and Finish Group

A cross-University Task and Finish Group was established in June 2015 to address some of the issues that were being raised within the institution. The group consisted of senior Professional Services staff from the Research Data Service, Research Governance, IT Services, Research Contracts and the Secretary's Office. There was also representation from senior academics from Social Sciences and Law and Social and Community Medicine who had specific interests in the ethics or legal aspects of this area. The remit of the group was to:

- Gain a better understanding of the issues of data access and sharing in relation to sensitive data,
- Assess and identify existing University processes and quality assurance mechanisms that impact on research data access/sharing,
- Develop an action plan highlighting where processes need adapting or developing to strengthen the University's compliance with the requirements for data sharing and access.

When looking at the issues the group also drew on expertise outside of the institution, including advice from national data centres and in particular by the report on Governance of Data Access published by the Expert Advisory Group on Data Access (EADGA, 2015).

The group met three times between June and October 2015 to consider the issues and the final output was a set of recommendations that was reported to, and accepted by, the University Ethics of Research Committee in November 2015. These recommendations were also approved by the University Research Committee in 2016.

Recommendations

This section outlines some of the actions and recommendations developed as a result of the work of the Task and Finish Group.

Levels of Data Access

Our initial idea was to provide a single alternative to ‘Open’ data. After several discussions with academics and the Research Governance team we realised this approach would be inefficient. While a route which involved a Data Access Committee (DAC) would undoubtedly be needed in order to provide transparent and impartial decisions on data sharing (EAGDA, 2015), this ‘Controlled’ process would also be necessarily very labour intensive. A ‘lighter touch’ and less time consuming alternative would be needed for less sensitive data; this became our ‘Restricted’ process. Currently, datasets can be published in the data.bris research data repository under three different access arrangements:

- **Open data:** The most permissive data access level, suitable for data where there are no particular sensitivities. Where research participants are involved, they have given consent to share anonymised data as ‘Open data’; the risk of re-identification is considered as extremely low.
- **Restricted data:** This covers anonymised human participant data that has clearance for sharing data for research purposes but not for making available as ‘Open data’ or where there are contractual restrictions allowing access to defined groups. Data is made available to approved bona fide researchers, after they have signed a data access agreement which governs access and use.
- **Controlled data:** This level relates to human participant data which can’t be anonymised or which has no ethical clearance for sharing and/or is subject to contractual constraints. Requests are referred to an appropriate Data Access Committee (DAC) for approval before data can be shared under a data access agreement.

Additionally a ‘Closed’ level exists which involves providing access to data for research funders and legislative bodies only. Although this category level is yet to be used, it’s envisaged that some highly sensitive ethical or contractual factors may result in data which must be retained yet cannot be shared.

Data Access Committee

A cross-University Data Access Committee (DAC) was formally set up in the spring of 2016. The membership was broadly similar to the make-up of the Task and Finish Group but invitations to the Committee are also extended to the Principal Investigator or Steward of the dataset in question and additional staff depending on the nature of the request. The Data Access Committee includes the Information Rights and Information Governance Manager, the Academic Research Facilitator for the Research Data Storage Facility, and a number of senior research staff across different faculties. This provides the committee with a good balance of regulatory, technical and ethical experience, which improves the decision-making process.

The role of the Committee is to monitor and oversee data access requests, in particular:

- Establishing that the research request is reasonable,
- Ensuring terms of participant consent are adhered to,
- Overseeing the compliance with legal and regulatory requirements,

- Assessing any risks related to granting access (e.g. re-identification),
- Reviewing policy relating to data sharing within the University of Bristol.

The Committee is administered by Research Data Service staff but chaired by a senior academic in the institution. The group meet face to face roughly every two months depending on whether there are any requests. Requests are generally handled within 20 working days so the group can also consider requests via email.

Data Access Agreements

One of the major pieces of work was the creation of standard data access agreement that could be used to administer the sharing of restricted and controlled datasets. This was drawn up with the help of an external lawyer who had considerable expertise in digital technology. The agreement effectively grants a non-exclusive, non-transferable, revocable, royalty-free licence to allow the access and use of the data for an authorised piece of research. It requires the recipient to become the data controller (as defined by the Data Protection Act, 1998) in relation to any protected data that they may receive under the agreement and to not permit any third parties (other than the Investigator and bona-fide members of the Investigator's research team) to have access to the data. For this reason a data access agreement has to be signed by an appropriate institutional signatory; this might be from recipient's contracts department, legal department or research office, depending on the nature of the institution.

Administrative Workflows

Workflows for the two sensitive categories were proposed by the Research Data Service and agreed by the Data Access Committee. The workflows for restricted and controlled data are broadly similar and so are shown as one workflow below, however, as controlled data release is agreed by the Data Access Committee, there are two routes at step two.

When an enquiry is made to the Research Data Service the requestor is sent a 'data access request form'. This form is required for both levels of access and provides the team with the basic information needed to proceed with the request. When the completed form is returned, the team create a folder for the request and follow the workflow below to monitor progress and to provide an audit trail.

Step 1: Receipt of form is acknowledged

- We use a template response to acknowledge receipt and include the date by which the applicant will be notified. We endeavour for 20 working days, in line with Freedom of Information requests (Freedom of Information Act, 2000 s.10(1)).
- The required response date is added to the team calendar to prompt response.

Step 2: Access request form checks

- We check the institutional affiliation and contact details, and verify the institutional signatory is from the contracts department, legal department or research office, depending on the nature of the institution.

- Screenshots of the institutional affiliation and email contact pages are saved to the request folder, including an onscreen date.
- We check that the applicant has ethical approval is in place (if required).
- We check that the planned research does not contravene the Data Access Agreement.
- We check that we have (or have access to) the requested data.

If further clarification is required the applicant is sent a template email. The 20 working day period resets (to day one) when any new information is received. If clarification is not received within the 20 working day period, the application is rejected on the grounds of having incomplete information. If clarification is unsatisfactory a rejection template email is sent which includes appeals information.

Step 2a: Restricted access

Restricted data requests are reviewed at the next Research Data Service team meeting. If approved, a filled in Data Access Agreement is sent to the institutional signatory, with a covering letter requesting two signed paper copies are returned by post.

Step 2a: Controlled access

Controlled data requests are reviewed by the Data Access Committee. The Research Data Service performs the same initial checks as outlined under ‘restricted’ data. If the application meets the criteria, the Research Data Service team forward the request to the data steward to check that the planned research can be supported by the dataset. If it cannot, we offer the requestor an opportunity to amend the planned research. This is a conversation between the data steward and the requestor of the data, and it may take time to come to a resolution.

When the application proceeds, the request is reviewed by the University of Bristol Data Access Committee. This committee includes senior representatives from IT Services, the University Secretary’s Office, Research Enterprise and Development and Library, and may include the data steward. A committee meeting is convened by the Research Data Service, and an information pack containing the supporting documents is provided to all members.

If the request is approved, a filled out Data Access Agreement is sent to the institutional signatory, with a covering letter requesting two signed paper copies are returned by post.

Step 3: Signed Data Access Agreement is received

The two signed copies are sent to the University of Bristol’s signatory to sign on behalf of the University.

Once signed on behalf of both parties, one copy is stored in a locked pedestal at the Research Data Service, and the other copy is returned to the requestor’s institutional signatory.

The dataset size is assessed by our Senior Technical Researcher.

Step 4: Data delivery is arranged

Detail of the delivery mechanisms are described in the Technical Workflows section below.

All controlled and restricted data requests are monitored on a spreadsheet which contains the workflow steps, annotated with dates and notes. The team are looking at procuring a case management system, so all documents and emails are stored more efficiently.

Technical Workflows

Research Data Storage at Bristol

The University's Advanced Computing Research Centre operates a secure file store specifically for holding research data, known as the 'Research Data Storage Facility' (*RDSF*). Research staff, typically Principle Investigators (PIs) may become *Data Stewards*, a role which allows them to request *Data Projects*: a networked file share with its own set of users, stored on this robust infrastructure.

Data Publishing

Data publishing at Bristol at the technical level is at heart an extension of the existing RDSF. The publishing system has a special space on the research file store to hold publications, which mirrors the project space (a convenience to make it obvious who the publication belongs to). The space is only accessible by Research Data Service staff.

The open publishing process (much simplified) is as follows:

1. Depositor prepares a folder containing their data in their project space /projects/my-project/data-bris/for-publication.
2. Depositor complete a metadata form for the publication and indicates the location of the data (for-publication).
3. The publication system copies /projects/a-project/data-bris/for-publication to /secure/a-project/deposits/abcd. abcd is the id for the publication, and will form the unique part of the final DOI.
4. If the deposit is accepted by the data librarians then:
5. The deposit is indexed, checksummed etc (this information is added to the metadata).
6. The metadata is added to the deposit folder.
7. A DOI is assigned and metadata sent to DataCite.
8. The deposit is moved to its final location /secure/a-project/public/abcd.
9. The folder is made available on the web.

Non-open data follows much the same process, but the final stored location is either restricted/adcd or controlled/abcd, which are inaccessible outside the RDS staff.

In addition we create a public/abcd folder containing a limited metadata record: it contains no details of the data content we index. We include this to fit with our existing publication infrastructure.

Transferring the Data

As has been said we typically make non-open data transfers over the internet using encrypted zips, or (for large deposits – we have some approaching 100GB) send encrypted disks. In either case this adds an additional step to the transfer since a password needs to be sent securely – the well-known key distribution problem.

A more successful approach we have found is using the common ssh ('Secure Shell') tool as part of scp or rsync. This relies on the recipient having a server with ssh, common enough at research institutions (though technical help may be needed). The transfer then involves:

1. We send recipient our ssh public key (this transfer does not need to be secure).
2. Recipient institution installs this on a server.
3. (Optional, for additional security) Recipient sends us their ssh public key.
4. File transfer proceeds over ssh, initiated by us.

This has many virtues: the public keys ensure that the transfer is secure and (in the case of Step 3) they are going to the right place, without needing any special treatment: they can be attached to emails, for example. This method doesn't expose our systems to any greater threat, such as firewall changes. It can also handle transferring large datasets.

Controlled Data and Individual Releases

Although controlled data is distinguished by the approval process, in practice it also often differs from restricted data in the nature of the release. Controlled data requests may concern a particular subset of the deposit: specific variables in a database, for example, or particular topics in a video interview archive.

As a consequence the release is not identical with the published data, and so needs to be recorded separately. Although there may be more compact ways to describe this release, such as a database query involving the requested variables, in general our only option is to hold the release in its entirety. So our example /secure/a-project/controlled/abcd may over time gain siblings /secure/a-project/controlled/abcd-release-1, /secure/a-project/controlled/abcd-release-2, and so on, plus the associated metadata concerning the release. This is potentially quiet expensive, and in future we may consider moving these to much cheaper tape storage.

Data Access in the Research Lifecycle

We encourage data producers to consider the most appropriate level of data access as early as possible in the research lifecycle. Several Principal Investigators have elected to include this information in a Data Management Plan, before research funding has been awarded, in an attempt to clarify data access plans for a potential research funder. More typically though, we see a level of data access nominated in either the ethical planning or research contracts stage.

The University of Bristol has several different ethical planning processes for researchers depending on the nature of the proposed research. Ethical planning usually occurs after a research grant has been awarded and always before research commences. We've introduced prompts into each of these routes which encourage researchers to state whether they expect any new data generated to be Open, Restricted, Controlled or Closed. Where a Faculty Ethics Committee assesses an ethical plan, the committee is also expected to comment on the appropriateness of the proposed data access level. As research progresses the elected level of access may change but we feel it's important that data access is considered as soon as possible.

We're currently running a pilot with our Research Contracts department which essentially extends this process beyond ethical planning. Many research projects have no particular ethical concerns and so do not require an ethical plan, yet all projects benefit from clear data access arrangements.

The research contracts stage involves the university formally entering into an agreement with an external organisation in order to carry out a defined programme of research. This stage precedes the commencement of research activities. As with the ethical planning stage, the contracts stage has been identified as an opportunity to prompt an appropriate level of data access to be considered. The Research Data Service is working in partnership with the Research Contract team to determine the likely data access level of new projects before associated contracts are finalised.

If a data access level has not been chosen before the data publication stage, the researcher primarily responsible will nominate a level and provide justification when they publish the data. This will be assessed by the Research Data Service team before data is made available.

Training and Guidance

Guidance and training on the policies and processes put in place around sensitive data are essential to ensure researchers understand the options available to them and requestors are aware of the actions they should undertake in order to access datasets. The Research Data Service website is the main area where guidance is located, through a dedicated 'Sensitive research data' webpage[1]. Information around data access levels and the application process is provided to potential requestors through a Frequently Asked Questions document[2]. University researchers who wish to learn more about the process before they choose an access level can equally consult this guidance. Generic blank copies of the Data Access Agreement for both restricted[3] and controlled[4] levels are also viewable. In addition, a short video has been developed which outlines the reasons for sharing sensitive data and how the University of Bristol Research Data Repository can help[5]. Information on the data access levels available also appears in more general guidance around sharing data concerning human participants[6].

[1] University of Bristol – Sensitive Research Data: https://data.bris.ac.uk/sensitive-research-data

[2] Data Access Levels of the data.bris University Research Data Repository: https://drive.google.com/drive/folders/0B-sxe4ro-QTTN19yNnBwZXFGSlE

[3] Data Access Agreement – Restricted Level: https://drive.google.com/drive/folders/0B-6vGrKq6udaUHlBTUJPeDJ5NEE

[4] Data Access Agreement – Controlled Level: https://drive.google.com/drive/folders/0B-6vGrKq6udaemFPU29PSXo2WWc

[5] Sharing data from research participants: https://vimeo.com/134607933

[6] Sharing research data concerning human participants: https://drive.google.com/drive/folders/0B-

In terms of training around sensitive data, we have developed a ‘Sensitive Data Bootcamp’ online tutorial, which as well as explaining the general concepts of sharing sensitive data, outlines the data access levels provided by the University[7]. The ability to publish sensitive data under restricted access levels is communicated during general research data management workshops, and plans are in place to deliver a dedicated sharing sensitive data course which includes an outline of the policies and processes in place.

Lessons Learned

Unforeseen challenges to the process included:

- Procedure for checking if applicants are ‘bona fide’; we have tried to reduce risks here by requiring a researcher to be affiliated with an institution that has research governance processes in place and requiring the institution rather than an individual to sign the agreement governing access and use of the data.
- Procedure for handling and retaining Data Access Agreements; ideally we would have the information stored in a case management system accessible to all departments that need to feed into the decision process, but currently this is still quite disjointed.
- There may be challenges from researchers to the data access level assigned to a dataset, therefore need to put in place an appeals process/escalation route to the DAC.
- Storage location of datasets; we are not always able to insist that controlled datasets are stored within the official RDSF storage area, which potentially may cause data audit problems further down the line.
- Supplying large datasets; there is no ‘typical’ dataset size but several datasets have been much larger than we had originally expected, so we needed to use encrypted disks for some transfers.
- Some research projects wish to engage with only part of the process described above. For example, a project may have its own steering group that wishes to act as a Data Access Committee, but the same project may have no formal sharing mechanism, such as a Data Access Agreement. In such cases the Research Data Service endeavours to support the project’s data sharing plans as much as possible by sharing documents, technical infrastructure and expertise. However, by ‘facilitating’ data sharing at arm’s length, rather than standardising and centralising the process, several complexities arise. Local conventions and project-specific arrangements may be different to institutional policy or project steering groups may be relatively short lived, raising questions about longevity. The Research Data Service is only just beginning to understand how and under what circumstances this kind of de-centralised data sharing can be supported.

sxe4ro-QTTZGhEaVcxaFB2SnM

[7] Sensitive Data Bootcamp: https://data.bris.ac.uk/bootcampSD/

Conclusion

One of major benefits of the exercise has been the close relationships built with other University services which are not normally communicated with by the Library. The policies, procedures and techniques that have been introduced through this work have provided the University with a way to ensure that research data is properly utilised. By setting up a permanent Data Access Committee to act as gatekeeper of data stored within the Research Data Storage Facility we have built in longevity and have moved away from temporary, project-funded solutions to the challenge of continuing accessibility. As a replicable model, this contribution will hopefully prove to be durable and long-lasting.

Acknowledgements

The Research Data Service team would like to acknowledge the work of the members of the University of Bristol Data Access Task and Finish Group and the Data Access Committee for their contribution to the workflows and issues described in the paper.

References

Data Protection Act. (1998). London: TSO.

Expert Advisory Group on Data Access. (2015). *Governance of data access*. Retrieved from https://wellcome.ac.uk/sites/default/files/governance-of-data-access-eagda-jun15.pdf

Freedom of Information Act. (2000). London: TSO.

Hiom, D., Fripp, D., Gray, S., Snow, K., & Steer, D. (2015). Research data management at the University of Bristol: Charting a course from project to service. *Program: electronic library and information systems, 49(4):475-493*.

Navarro, R. (2008). An ethical framework for sharing patient data without consent. *Journal of Innovation in Health Informatics, 16(4):257-262.*