Policy for the use of the Research Data Storage Facility

Introduction:

The University's High Performance Computing (HPC) facility went live to users in May 2007. Access to this world-class HPC facility allowed Bristol academics to intensify their computational research to levels not previously possible. As a result some groups, particularly in the Particle Physics and Climatology domains, were generating many terabytes of data and it became apparent that more data storage capacity was required.

This demand led to the creation of the University Research Data Storage Facility (RDSF), which provides an integrated resilient facility, available to researchers from all disciplines.

The RDSF data storage is a mix of volatile storage on disk (i.e. one copy only is kept) and high availability mirrored storage, where data which has to be available at all times and which also cannot easily be replaced, held on disk in two geographically separated places (i.e. the HPC machine rooms in Merchant Venturers Building (MVB) and Physics.) All the data is also backed up to tape, with the tape backup sited in the IT Services (ITS) machine room. Options for a tape archive and for offsite storage are still being explored. The next major investment in the RDSF is planned for 2014, and is likely to implement a Hierarchical Storage Management system which will allow less used data to seamlessly migrate off expensive disk storage onto cheaper tape storage, whilst still being accessible to the end-user.

Early work undertaken in developing a University research data retention policy by Research and Enterprise Development (RED) is detailed in the *Research Governance and Integrity Policy* and *Research Governance Guidance on Record Retention and Archiving for studies requiring a research sponsor* - http://www.bristol.ac.uk/red/research-governance/practice-training/policies-guidance/index.html.

Research data storage policy and procedure was formulated by the Storage Project Board and Research and Enterprise Development (RED) with input from ITS/ IT Services R & D/ILRT and the Library.

The Pro Vice-Chancellor Research is the owner of the Research Data Storage Policy. The Research Data Storage and Management Board will modify this policy in the light of developments and guidance from funders and other bodies. It is intended that the policy should align with the requirements of the UK Research Councils (RCUK). The Research Data Storage and Management Board reports through the HPC Executive, to the HPC Board as set out in its terms of reference, https://www.acrc.bris.ac.uk/acrc/rdsmb_tor.pdf.

Funder policy

RCUK published a set of Common Principles on Data Policy http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx in 2011 which provide an overarching framework for individual Research Council policies on data management. A declaration on Sharing Research Data to Improve Public Health was signed in January 2011 by 17 major international public health research funders - http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm.

Some funders have updated their guidance recently in the light of the RCUK principles, whilst new guidance from some others is awaited. The policies of the research councils in this area are summarised on the Digital Curation Centre (DCC) website - http://www.dcc.ac.uk/resources/policy-

and-legal/overview-funders-data-policies. The guidance that has driven much of the debate in this area is EPSRC's Data Policy Framework http://www.epsrc.ac.uk/about/standards/researchdata/Pages/default.aspx, which requires research data to be held for at least 10 years from the date of last access.

JISC has produced guidance written specifically for researchers on Freedom of Information and managing research records - http://www.jisc.ac.uk/publications/programmerelated/2010/foiresearchdata.aspx.

The DCC (http://www.dcc.ac.uk) has produced several tools to assist researchers in creating a Data Management Plan (DMP) for their project including a DMP template. A Data Management Plan provides a summary of the data which a researcher expects to create during the life time of the project and detail issues around its ownership, use, reuse, sharing, metadata, storage and preservation. The DCC's DMP Online tool is now available - http://www.dcc.ac.uk/resources/data-management-plans/ **.**

RDSF pricing policy

The current costs of storing data in the RDSF are detailed in Costs of using the Research Data Storage Facility. The principle underlying the costs is modelled on the Princeton 'Pay Once, Store Forever' (POSF) model , which calculates a one-off charge for data storage which will cover future refresh of the media by applying Moore's Law to storage costs.

We have defined 'forever' in this context to be 20 years.

Research data management principles - http://data.bris.ac.uk/principles

A set of high level University research data management principles, outlining the responsibilities of the University and of researchers, have been published in draft form and are in course of approval by the relevant University committees.

data.bris **research data service -** http://data.bris.ac.uk

data.bris, the University's research data service is currently being run as a service pilot across the University between 2013 and 2015 and will then be developed into a full service beyond 2015.

A data repository is being developed to enable researchers to upload subsets of data from the RDSF and make them publicly available, linking them where appropriate to a publication through creation of a Digital Object Identifier (DOI).

Policy for the use of the Research Data Storage Facility

V 1.7
19/7/2013

- This document should be read in conjunction with the Research Data Storage Facility Terms of Use .
- For clarification of terminology, refer to the Glossary in Appendix 1 of this document.
- This policy will be reviewed by the Research Data Storage and Management Board on a regular basis.

1. Management of the Research Data Storage Facility (the Facility)

1.1. The scope of this policy concerns only the digital research data assets held in the Facility.

1.2. The Facility must not be used for the storage of non-research data. Other University facilities exist for such data.

1.3. The Facility not only provides storage for the research data assets, but allows them to be processed, manipulated and mined.

1.4. The Facility will store both live data and archived data. Live data will be active with changes being made to it, whilst archived data will only be accessible in its current form without the possibility of making changes.

1.5. The Advanced Computing Research Centre (ACRC) will assume responsibility for maintaining the Facility infrastructure, that is, the storage hardware, storage file system (the IBM General Parallel File System - GPFS) and storage media. Maintaining the integrity and validity of the data is the responsibility of the Data Steward (see 1.9-1.13).

1.6. Day to day management of the Research Data Storage Facility is undertaken by the ACRC under the direction of the Research Data Storage and Management Board.

1.7. Assuming appropriate capital expenditure is available, the data held will be migrated as appropriate on to new media as required and as technologies change. If appropriate capital expenditure is not available, the data will be held on low-cost media such as tape until the media decays, at which point the data will no longer be available.

1.8. The ACRC will support researchers to meet the requirements of the University's *Research Governance Guidance on Record Retention and Archiving for studies requiring a research sponsor* - http://www.bristol.ac.uk/red/support/governance/. This will be achieved by endeavouring to provide storage for the research data that meets both the legal and regulatory framework for particular types of research and the terms and conditions imposed by external research funders. The ACRC will also support researchers as they seek to comply with the University's high level research data principles - http://data.bris.ac.uk/principles.

1.9. All data stored within the Facility, either in a Standard Project or a Collaboration Project, will have an owner, who must be a member of the University. This will normally be the PI of the project in the first instance, but in every case, the data owner must be a University of Bristol staff member.

The person within the University of Bristol with responsibility for the data will be known as the Data Steward. If that person subsequently leaves the University or is absent for a prolonged period of time (e.g. on sabbatical), their line manager will, in the first instance, assume the responsibilities of Data Steward.

Where personal data (as defined by the Data Protection Act 1998) is processed for research purposes in the course of a researcher's employment at the University of Bristol, the University will be the Data Controller under the Act for that Data. This will include where personal data is stored in the Facility.

However, in all circumstances, the Data Steward will be personally responsible to the University for ensuring the proper administration and oversight of any data they have caused to be stored in the Facility. This will include the Data Steward providing such information as is reasonably required by the Research Data Storage and Management Board to make an adequate risk assessment for data storage, and to meet the University's legal and ethical obligations.

1.10. Ensuring that the research project data is readable and accessible in future is the responsibility of the Data Steward. Readability and accessibility could be affected by, for example, changing document formats. It may not be possible to exactly reproduce computational data on a long term basis if the computer on which the data was originally run no longer exists or is no longer functional. Data will be returned to the Data Steward in its original form, subject to any changes of format. Data will be treated as raw data, not interpreted data.

1.11. It is the responsibility of the Data Steward to ensure that the data is stored in as resilient a state as is appropriate to the data type and research. The University cannot be held responsible for lost data if data is held in a volatile class (i.e. only one copy on disk or tape, plus a backup copy on tape). The University will use its reasonable endeavours to recover any data which may be lost.

1.12. It is the responsibility of the Data Steward of the research project to choose appropriate data storage classes for their data.

1.13. It is the responsibility of the Data Steward to validate any data to be uploaded into the *data.bris* repository for publication and to authorise the upload. As part of the validation process, the Data Steward will be asked to reconfirm compliance with the Data Protection Act , Freedom of Information Act and the University's ethical requirements, as the data being deposited will become 'read only' and public.

1.14. If the capacity of the Facility nears saturation, the Research Data Storage and Management Board will review the data held and may contact a Data Steward to see whether part or all the data held on their behalf is still required. Access to the data will be monitored (refer to Section 2) by the Research Data Storage and Management Board which will in turn report to the HPC Board. The ACRC will aim to work with the Data Steward to move data to another infrastructure where appropriate. The Data Steward and his/her line manager can choose to hold a copy of the data - where feasible - after it has been removed from the Facility.

1.15. Each project must have an exit strategy in line with the requirements of the relevant funder. The RDSF project application form will ask how long the data needs to be held for after the end of the project. Data will be kept for 12 months if a Data Steward leaves without appointing a replacement, and then the Research Data Storage and Management Board will consider at the next review whether the data remains within the Facility or is transferred elsewhere.

2. Application and review process

2.1. Data Stewards will apply to use the storage facility using a web-based application form detailing the amount of storage required within the appropriate classes of data (i.e. volatile, mirrored, resilient, backup, archive), and giving a brief description (not more than 500 words) of the data to be stored and for how long it should be stored. Requirements for levels of security will also be requested (see sections 5.5, 7.3 and 7.6 of the Research Data Storage Facility Terms of Use and IT Services policies relating to information security and management - http://www.bris.ac.uk/infosec/policies/docs/isp-01.pdf.) Each application will be given the actual or notional cost of the storage that they are applying for (see Research Data Storage Facility: Costs of using the Facility).

If a Data Steward wishes to share data with a researcher at another institution he/she can apply for a Collaboration Project and arrange for the External Data User to register to join the Collaboration Project. The Data Steward then authorises the External Data User to become a member of his/her

A Data Management Plan or a similar set of guidelines should be submitted with the application. This plan or set of guidelines can be more or less detailed depending on the amount and type of storage requested and the complexity of the data to be stored.

If a Data Steward subsequently finds that more data storage is required than originally requested, an amended request will need to be submitted to the Research Data Storage and Management Board.

2.2. The Data Steward will be responsible for the production and storage of metadata to address the semantic issues of what the data means.

2.3. Where the project generating the data has required an ethical review by an Ethics Committee, such as a Department or Faculty Ethics Committee, the Data Steward will be responsible for providing the Research Data Storage and Management Board with a copy of the ethics application and documented evidence of the Committee's approval and any conditions.

Proposed storage of unanonymised sensitive and/or personal data (as defined by the Data Protection Act 1998) will require the Data Steward to provide written approval from the University Secretary's Office, together with a copy of the relevant ethical review application and documented evidence of the Ethics Committee approval. Please also refer to section 2 of the Research Data Storage Facility Terms of Use.

2.4. Under the Freedom of Information Act 2000, third parties may request access to information held by Public Authorities, subject to certain exemptions. Such exemptions are interpreted strictly. Universities are defined as Public Authorities under the Act and research data may thus be requested under FOI legislation.

The Data Steward must first consult with the University Secretary's Office, and then inform the Research Data Storage and Management Board, at the time of application, if they believe an exemption to third party access under the Freedom of Information Act 2000 should apply to the data they wish to store in the Facility. They must provide details to the Secretary's Office of why the exemption applies to the data, and how long the exemption should last. If a FOI request is received, whether the exemption applies will be assessed by the Secretary's Office on receipt of the request. Please refer to section 3 of the Research Data Storage Facility Terms of Use.

2.5. The Data Steward must consult members of the ACRC when completing the grant application and Data Management Plan if a significant and long-term storage need is identified.

2.6. The application form will be reviewed by the members of the Research Data Storage and Management Board. Once approved, the Data Steward's storage account will be created.

2.7. Data Stewards will be asked annually if the data is still required, and if so, whether the work that produced it has been cited. Details will be requested of further research projects using the data and any grants related to the research data held. This annual review and impact analysis will help in ensuring sustainability of the Facility.

2.8. Storage holdings and assets will be reviewed by the Research Data Storage and Management Board. If there has been no demonstrable activity with the data within that 12 month period**,** the Research Data Storage and Management Board will consider the need to migrate the data down the storage hierarchy, for example from disk to tape media. The Research Data Storage and Management Board reserves the right to review how data is held in the light of overall storage requirements.

2.9. The conflict resolution process will involve disputes being escalated by the Research Data Storage and Management Board to the HPC Executive and then to the HPC Board, who will act as the final arbiter.

2.10. Users are encouraged to mention use of the storage facility in publications, where the data produced by or underpinning the research is stored in the Facility. The suggested template wording is "*The work was made possible in part by using the Research Data Storage Facility of the University of Bristol - http://www.bris.ac.uk/acrc/storage*".

2.11. If an external party is part of a joint project with a member of staff from the University of Bristol and requires access to the Facility, they may apply to join a Collaboration Project as an External Data User, as set out in section 2.1 for a maximum period of 12 months. If the External Data User's account is still required after 12 months, the Data Steward can ask External Data User to re-register. The University reserves the right to terminate external accounts if circumstances require.

Appendix 1

Glossary

Storage definitions:

Volatile Storage – a single copy of digital data held in one place on one site.

Mirrored/Highly Available Storage – two copies of digital data held on disk in two geographically separate sites.

Resilient Storage – digital data held in either volatile or mirrored storage, plus one copy on a different media such as tape.

Traditional Backup – a copy of the digital data that is held for at most 3 months, providing a snapshot of data at the current point in time.

Archive – a separate, static copy of the digital data held, usually on a read-only basis.

Other definitions:

Collaboration Project – a project registered by a Data Steward, specifically for sharing data with researchers from outside the University of Bristol.

Data Management Plan (DMP) - a structured plan increasingly asked for by funding bodies, usually developed in two stages: an initial version at the grant application stage, and a fuller version which is developed at the early-project stage. It provides a summary of the data that will be created during the project and how it will be stored, curated and made accessible.

Data Steward - the person with responsibility for the data, usually the Principal Investigator (PI) of the project. The Data Steward must be an employee of the University of Bristol.

Data User – someone authorised by the Data Steward to have access to the data assets of the project. The Data User must be an employee of the University of Bristol.

Digital Object Identifier (DOI) – a unique and permanent character string which identifies an electronic document (for example a publication or a data set.)

External Data User – someone collaborating with a Data Steward, who is not an employee of the University of Bristol, but is authorised by the Data Steward to have access to the data assets of a Collaboration Project.

General Parallel File System (GPFS) – IBM's scalable, highly-available, high performance file system, which is optimized for multi-petabyte storage management.

Petabyte (PB) - a unit of information equal to one quadrillion bytes, i.e. 1,000,000,000,000,000 bytes or 10^{15} bytes. 1 petabyte = 1,000,000 gigabytes.

Petascale - in this instance, a storage system capable of storing one or more Petabytes of data.(see Petabyte above.)

Research Data Storage Facility (RDSF) – the University's integrated, resilient facility for the storage of research data, also known as 'BluePeta'.

Standard Project – a project registered by a Data Steward to enable use of the RDSF and which can be accessed only by members of the University of Bristol.

Terabyte (TB) - a unit of information equal to one trillion bytes, i.e. 1,000,000,000,000 bytes or 10^{12} bytes. 1 terabyte = 1,000 gigabytes.