Edinburgh University Data Library Research Data Management Handbook

v.1.0 (Aug. 2011)

Table of Contents

Section 1.

Why manage research data?

Data management is one of the essential areas of responsible conduct of research.

Before starting a new research project, the Principal Investigators (PIs) and the research teams must address issues related to data management.

By managing your data you will:

- Meet funding body grant requirements.
- Ensure research integrity and replication.
- Ensure research data and records are accurate, complete, authentic and reliable.
- Increase your research efficiency.
- Save time and resources in the long run.
- Enhance data security and minimise the risk of data loss.
- Prevent duplication of effort by enabling others to use your data.
- Comply with practices conducted in industry and commerce.

Defining research data

Research data, unlike other types of information, is collected, observed, or created, for purposes of analysis to produce original research results.

Classification of research data

Research data can be generated for different purposes and through different processes (Research Information Network classification):

- Observational: data captured in real-time, usually irreplaceable. For example, sensor data, survey data, sample data, neuroimages.
- Experimental: ldata from lab equipment, often reproducible, but can be expensive. For example, gene sequences, chromatograms, toroid magnetic field data.
- Simulation: data generated from test models where model and metadata are more important than output data. For example, climate models, economic models.
- Derived or compiled: data is reproducible but expensive. For example, text and data mining, compiled database, 3D models.
- Reference or canonical: a (static or organic) conglomeration or collection of smaller (peer-reviewed) datasets, most probably published and curated. For example, gene sequence databanks, chemical structures, or spatial data portals.

Research data formats

Research data comes in many varied formats:

- Text - flat text files, Word, Portable Document Format (PDF), Rich Text Format (RTF), Extensible Markup Languague (XML).
- Numerical - Statistical Package for the Social Sciences (SPSS), Stata, Excel.
- Multimedia - jpeg, tiff, dicom, mpeg, quicktime.
- Models - 3D, statistical.
- Software - Java, C.
- Discipline specific - Flexible Image Transport System (FITS) in astronomy, Crystallographic Information File (CIF) in chemistry.
- Instrument specific - Olympus Confocal Microscope Data Format, Carl Zeiss Digital Microscopic Image Format (ZVI).

Research data (traditional and electronic research) may include all of the following:

- Documents (text, Word), spreadsheets
- Laboratory notebooks, field notebooks, diaries
- Questionnaires, transcripts, codebooks
- Audiotapes, videotapes
- Photographs, films
- Test responses
- Slides, artefacts, specimens, samples
- Collection of digital objects acquired and generated during the process of research
- Data files
- Database contents (video, audio, text, images)
- Models, algorithms, scripts
- Contents of an application (input, output, logfiles for analysis software, simulation software, schemas)
- Methodologies and workflows
- Standard operating procedures and protocols

The following **research records** may also be important to manage during and beyond the life of a project:

- Correspondence (electronic mail and paper-based correspondence)
- Project files
- Grant applications
- Ethics applications
- Technical reports
- Research reports
- Master lists
- Signed consent forms

Funders' policies and guidelines

In the UK, the following research funders have some data policies in place:

The Arts and Humanities Research Council's (AHRC) policy came into effect from April 2008. AHRC funding for the Arts and Humanities Data Service ceased on 31 March 2008.

The Biotechnology and Biological Sciences Research Council (BBSRC) takes a devolved approach and its policy came into effect in April 2007.

The Economic and Social Research Council (ESRC) requires data to be offered to the UK Data Archive for potential dissemination through the Economic and Social Data Service. The council updated its Research Data Policy in September, 2010 to require a data management and sharing plan.

The Engineering and Physical Sciences Research Council (EPSRC) has no policy on data as yet.

The Medical Research Council (MRC) has no data centres but adopted a data sharing policy with effect from April 2006.

The Natural Environment Research Council (NERC), which has a detailed data policy handbook and guidelines for grant-holders, has seven designated data centres where grant-holders can deposit their data.

The Science and Technology Facilities Council (STFC), formed in 2007 by a merger of the Council for the Central Laboratory of the Research Councils (CCLRC) and Particle Physics and Astronomy Research Council (PPARC), has yet to develop its formal data sharing policy, though its facilities have well-developed individual policies.

The Wellcome Trust has both a statement on open access of research outputs and a brief policy regarding data management and sharing.

Research funders' policies - The Digital Curation Centre (DCC)

A detailed comparison of main UK research funders' policies and the support infrastructure is available from the Digital Curation Centre (DCC) website:

Research funders' policies - http://www.dcc.ac.uk/resources/policy-and-legal#policies

Data management planning

Data planning checklist and data management plan are the first two crucial steps of data management planning process.

Data planning checklist

The first step in data management involves producing a data planning checklist to evaluate your data needs.

A suggested data planning checklist

The following is a suggested data planning checklist that should be considered before you embark on your project:

- What type of data will be produced? Will it be reproducible? What would happen if it got lost or became unusable later?
- How much data will be generated and how often will it change?
- Who will be the audience for your data and how will they use it now, and in the long run?
- Who controls it (Principal Investigator, yourself, computing officer in your school)?
- How long should it be retained? For example, 5 years, up to 10 years, or permanently.
- Are there tools or software needed to create/process/visualise the data?
- Are there any special privacy or security requirements? For example, personal data, high-security data?
- Are there any sharing requirements? e.g. funding body's data sharing policy?
- Are there any other funding body requirements? e.g. data management plan in proposal?
- Is there good project and data documentation?
- What directory and file naming convention will be used?
- What project and data identifiers will be assigned?
- What file formats will be used? Are they long-lived?
- What will be the storage and backup strategy?
- When and where the data will be published?
- Is there a community standard for data sharing/integration?

Data management plan

Data management plans ensure that all aspects of data management are fully understood at the start of a project.

What is a data management plan?

A data management plan is a document which describes:

- What research data will be created.
- What policies (funding, institutional, and legal) apply to the data.
- What data management practices will be used (backups, storage, access control).
- What facilities and equipment will be required (hard-disk space, backup server, repository).
- Who will own and have access to the data.
- Who will be responsible for each aspect of the plan.
- How its reuse will be enabled and long-term preservation ensured after the original research is completed.

The data management plan must be continuously maintained and kept up-to-date throughout the course of research.

Components of a data management plan

A detailed data management plan should give answers to the following questions:

- What kind of data will be collected?
- How will the data be collected?
- Who holds the copyright and intellectual property rights of the data?
- What kind of possession issues are involved?
- Who will decide on access to the data?
- How will the research participants be informed?
- Which software will be used in storing and processing the data?
- How will the (technical) quality of the data be assured?
- Which data formats and storage media will be used?
- What kind of rights will be granted to different user groups for reading and managing data files?
- What kind of data and file backup procedures will be used?
- How will data processing be documented?
- How will the metadata on the data collection and dataset content be stored?
- How will confidentiality be ensured?
- How will data protection be carried out?
- What will happen to the data after the original research is completed?

Data documentation and metadata

To be usable, a dataset (or other research data object) needs to be well-documented.

Documentation

It is good practice in research to ensure that all data generated or collected through the course of your research is easy to understand and analyse.

Producing good documentation and metadata ('data about data') provides context for your data, tracks its provenance, and makes it easier to find and use your data in the long term, or for others to discover on the internet.

Documentation and metadata requirements should be identified from the start of your project and considered throughout the lifecycle of your data. This is the essence of good 'data curation'.

Where a research protocol is used, much of the documentation needed will already exist. If instrumentation is used, calibration and other details need to be captured for the data to remain useful. Lab notebooks are perhaps the most rigorous form of documentation. Is it possible to put it into digital format?

For qualitative data or small-scale surveys, the documentation might exist only in your head.Take the time to write it down while it is fresh on your mind.

This may include writing methodology reports, creating codebooks with full variable and value labels, documenting decisions about software, tracking changes to different versions of the dataset, recording assumptions made during analysis.

Have you created a "readme.txt" file to describe the contents of files in a folder? Such a simple act can be invaluable at a later date.

Ultimately the amount of effort put into documenting your data will depend on the intended lifespan and how broadly you intend to share it.

Ideally these decisions should be made at the outset to avoid having to carry out a rescue mission on the data, sometimes known as 'digital archaeology', when a key member of staff leaves, or renewed interest in a topic suddenly puts a dataset in demand.

Metadata

Usually, metadata are standards-based and serve a particular purpose in data processing and machine-to-machine interoperability. Three broad categories of metadata are:

- Descriptive - common fields such as title, author, abstract, keywords which help users to discover online sources through searching and browsing.
- Administrative - preservation, rights management, and technical metadata about formats.
- Structural - how different components of a set of associated data relate to one another, such as tables in a database.

Data storage, backup and security

Guidance on securely storing and backing up research data.

Data storage

Through the course of your research you must ensure that all research data, regardless of format, is stored securely and backed up or copied regularly.

You can store your research data on:

Networked drives

These are managed by IT staff centrally or within your School or College. It is highly recommended that you store your research data on regularly backed-up networked drives such as:

- Fileservers managed by your research group or school.
- Fileservers managed by Information Services.
- Storage Area Network (SAN) - either the Infrastructure SAN or the Edinburgh Compute and Data Facility (ECDF) SAN.

This way you will ensure that your data will be:

- Stored in a single place and backed up regularly.
- Available to you as and when required.
- Stored securely minimising the risk of loss, theft or unauthorised use.

Personal computers and laptops

These are convenient for storing your data temporarily but should not be used for storing master copies of your data. Local drives may fail or PCs and laptops may be lost or stolen leading to an inevitable loss of your data.

External storage devices

External storage devices such as hard drives, USB flash drives, CDs and DVDs, can be an attractive option for storing your data due to their low cost and portability. However, they are not recommended for the long term storage of your data, particularly, your master copies as:

- Their longevity is not guaranteed, especially if they are not stored correctly, for example, Compacts Discs (CDs) degrade, tapes shrink in the long term.

- They can be easily damaged, misplaced or lost.
- Errors with writing to compact discs and digital video discs (CDs and DVDs) are common.
- They may not be big enough for all the research data, so multiple disks or drives may be needed.
- They pose a security risk.

If you choose to use CDs, DVDs and USB flash drives (for example, for working data or extra backup copies), you should:

- Choose high quality products from reputable manufacturers.
- Follow the instructions provided by the manufacturer for care and handling, including environmental conditions and labelling.
- Regularly check the media to make sure that they are not failing, and periodically 'refresh' the data (that is, copy to a new disk or new USB flash drive).
- Ensure that any private or confidential data is password-protected and/or encrypted.

Remote or online back-up services

These provide users with an online system for storing and backing-up computer files e.g. Dropbox, Mozy, A-Drive.

Typically, they:

- Allow users to store and synchronise data files online and between computers.
- Employ cloud computing storage facilities (e.g. Amazon S3).
- Provide the first few gigabytes free and users pay for more facilities, including space.

Advantages

- No user intervention required (change tapes, label CDs, perform manual tasks).
- Remote back-up maintains data offsite.
- Most provide versioning and encryption.
- Multi-platform.

Disadvantages

- Restoration of data may be slow (dependent upon network bandwith).
- Stored data may not be entirely private (thus pre-encryption).
- Service provider may go out of business.
- Protracted intellectual property rights/copyright/data protection licences.

Data backup

Keeping backups is probably the most important data management task. There is a real risk of losing data through hard drive failure or accidental deletion.

It is therefore recommended to keep at least 3 copies of your data, for example, original, external/local, and external/remote, and have a policy for maintaining regular backups.

When considering your backup strategy you need to know:

- Whether all data, or only changed data, will be backed up. (A backup of changed data is known as an "incremental backup", while a backup of all data is known as a "full backup").
- How often full and incremental backups will be made.
- How long will backups be stored.
- How much hard-drive space or number of Digital Video Discs (DVDs) will be required to maintain this backup schedule.
- If the data is sensitive, how will it be secured and (possibly) destroyed.
- What backup services are available that meet these needs and, if none, what will be done about it.
- Who will be responsible for ensuring backups are available.

Data security

Data security is the means of ensuring that data are kept safe from corruption and that access is suitably controlled.

It is important to consider the security of your data to prevent:

- Accidental or malicious damage/modification to data.
- Theft of valuable data.
- Breach of confidentiality agreements and privacy laws.
- Premature release of data, which can void intellectual property claims.
- Release before data have been checked for accuracy and authenticity.

Securing digital research data is part of the issue of information technology security. You should always have up-to-date anti-virus software installed on your office and home computer.

If you have sensitive data that are covered by privacy laws or confidentiality agreements, it is best to store them on a computer that is not connected to any network. If this is not possible, then you can also consider encrypting your data.

The final issue to consider is physical security. A computer that is not connected to a network is still vulnerable to theft and malicious damage/modification to data.

For highly sensitive data you can use an external hard drive and store it in a locked safe overnight. Such data should not be stored on portable drives such as laptops and flash drives unless absolutely necessary and without encryption.

Encryption

Encryption, whereby data is transformed into code, is a good way of ensuring its confidentiality and security. You will find the University's guidance on this topic useful.

- Encryption

http://www.ed.ac.uk/schools-departments/information-services/services/computing/desktop-personal/security/encryption

Confidential disposal of research data

The University has a comprehensive guide to the disposal of confidential and/or sensitive waste held on paper, CDs, DVDs, tapes, discs and other holding devices.

- Guidance on disposal of confidential waste –

http://www.ed.ac.uk/schools-departments/estates-buildings/waste-recycling/a-z/confidential-waste

Section 2.

Data sharing and preservation

Guidance and support for sharing with others and preserving research data in the long term.

Why share research data?

Why should I consider sharing research data that I collect, create or collate in the course of my research? Are there legitimate reasons not to share?

Reasons to share

- Scientific integrity - publishing your data and citing its location in published research papers can allow others to replicate, validate, or correct your results, thereby improving the scientific record.
- Publicly funded research - there is a growing movement for making publicly funded research available to the public, as indicated for example, in the Organisation for Economic Co-operation and Development (OECD) Principles and Guidelines for Access to Research Data from Public Funding.
- Funding mandates - UK research councils are increasingly mandating data sharing so as to avoid duplication of effort and save costs.
- The University of Edinburgh's mission - "the creation, dissemination and curation of knowledge" implies transparency about the research that is conducted in its name.
- Increase the impact of your research - those who make use of your data and cite it in their own research will help to increase your impact within your field and beyond it. Users of your data may include those in other disciplines, sectors, and countries.
- Preserve your data for your own future use - by preparing your data for sharing with others, you will benefit by being able to identify, retrieve, and understand the data yourself after you have lost familiarity with it, perhaps several years hence.
- Teaching purposes - your data may be ideal for students to learn how to collect and analyse similar types of data themselves.

Reasons not to share

If your data has financial value or is the basis for potentially valuable patents that could be exploited by the University, it may be unwise to share it, even with a data license or terms and conditions attached.

Edinburgh Research and Innovation (ERI) can assist you in determining the value of your research data for these purposes.

If the data contains sensitive, personal information about human subjects, it may violate the Data Protection Act, ethics codes, or your own written consent forms to share it, even with other researchers.

Often there are ways to anonymise the data to remove the personally identifying information from it, thus making it sharable as a public use dataset.

If parts of the data are owned by others, such as commercial entities or authors, then even if you have derived wholly new data from the original sources you may not have the rights to share the data with others.

By writing a data management plan near the beginning of your research project, you can work through these issues and determine if you will be able to produce a version of your data that can be shared with others.

Methods for data sharing

There are currently a number of ways in which data may be shared, both informal and formal.

Informal methods

Many researchers are accustomed to sharing data within their research group or extended 'virtual organisation' or an international research partnership.

In these situations, risk is perceived as low and trust is high, particularly with co-investigators or co-authors.

Data sharing within a group can be accomplished in a number of ways, but most of them require centrally administered, authenticated access.

- shared network file server (for example, within a research group, school or department)
- intranets or content management systems
- research project wikis (such as through the Confluence service offered by Information Services)
- online collaboration tools or groupware (for example, Google Docs and Spreadsheets)
- proprietary enterprise software (for example, Microsoft Sharepoint)
- computer 'grid' network via custom data applications
- social networking sites for researchers (such as myExperiment)

Researchers may also be persuaded to share data with other researchers or students who make a personal request based on knowledge of the researcher's work.

Direct communication means that the researcher can ask questions about the intent of use and the user's experience, and that the user can follow up with questions if use of the data is not clear-cut.

In these situations, direct or 'peer to peer' sharing of data files may take place via:

- email attachments
- posting files on a website or ftp (file transfer protocol) server, with the link sent by email
- laptop to flashdrive (for example, at a conference)
- Compact Disc (CD) or Digital Video Disc (DVD) (in the post)
- social networking sites
- file sharing sites
- mobile devices; bluetooth networking

Formal methods

Researchers often have options or requirements (from publishers or funders) to share their data more broadly.

This normally involves additional effort such as preparing a public use version of the dataset (including anonymisation techniques), 'cleaning' the data of routine errors, documenting or annotating the data to improve its understandability, and possibly reformatting the data, so that it can be re-used in another research context.

Once this work is complete the data may be disseminated or 'published' through one of the following methods.

- post on a University website
- post on a public website or wiki
- post on a publisher's website
- upload to a distributed, dynamic database (discipline-specific, such as a protein or gene database, such as those hosted by the European Bioinformatics Institute)
- deposit in a central data archive (such as the UK Data Archive, funded by the Economic and Social Research Council)
- deposit in a data centre (such as one of the Natural Environment Research Council's data centres)
- deposit in a local data repository (such as Edinburgh DataShare)

Access and restrictions

How quickly and how broadly to allow others access to your research data are questions that should be answered as part of the data management planning process.

Forms of control

For many researchers, the question is not whether to share data, but how broadly to share: within a closed research group or department, with students, with other researchers in an institution, subject-discipline, country, or with the public at large.

As indicated on the previous page (methods for sharing data), sharing can be done by informal or formal means.

Clearly an advantage of informal sharing is the researcher retains control over who uses the data and how.

With a popular dataset, or over a longer period of time, this may become a disadvantage, and more formal methods involving publishing or depositing the data with a 'trusted repository' or service may be more desirable.

Options for restricting access include:

- Requiring the user to confirm their agreement to follow stated Terms of Use.
- Attaching a legal license to the data, free or fee-based.
- Requiring password protected registration and contact details such as an email address (enforced through automatic email confirmation).
- Requiring authenticated registration for users from certain institutions (for example EASE for University of Edinburgh, UK Access Federation for UK Higher Education institutions).
- Authorising a written statement about how the data will be used for research purposes.
- Publishing a metadata record only (title, author, abstract) with an email contact to request access to the dataset.
- Limiting the number of concurrent users at a given time.
- Limiting access to certain Internet Protocol (IP) addresses.
- Limiting physical access to a secure non-networked server (a data enclave).

Many of the options above will hinder users' ability to discover the existence of the data through search engines.

Open data

> *'Open Data is a philosophy and practice requiring that certain data are freely available to everyone, without restrictions from copyright, patents or other mechanisms of control.'*
>
> Peter Murray-Rust, Cambridge University (in Wikipedia)

Those who advocate for open data are often interested in new computing techniques for combining, analysing, and integrating online data to create new forms of data or knowledge.

Text mining and "mashups" - such as spatial visualisations enabled through the use of Google maps - are two such techniques which are hindered by restrictive copyright permission, licensing or registration requirements.

> *The coolest thing to do with your data will be thought of by someone else.*
>
> Rufus Pollock, Cambridge University and Open Knowledge Foundation

This thought is exciting to some and disturbing to others. In itself it is a good reason to think through your reasons for restricting your data.

Data protection and confidentiality

If your research involves human subjects, you will need to consider both legal and ethical obligations regarding sharing your data.

The 1998 Data Protection Act affects the processing of personal or sensitive data and the circumstances under which you can share it with others.

The University's Records Management Section has online guidance and offers direct support for decisions about information disclosure.

The role of the University's various research ethics committees is to develop policy and general guidance for Colleges and Schools on ethical issues arising from non-medical research involving human participants. Your School may have its own research ethics committee or guidance.

Avoidance of disclosure of personal or sensitive data can be accomplished in a number of ways, including anonymisation techniques or data aggregation for numeric data, editing of video or sound recordings, use of pseudonyms in qualitative data.

Different methods have different consequences for data quality, and should be considered in tandem with the consent process, for example, what sort of informed consent you seek from your subjects.

The UK Data Archive has an excellent guide on consent, confidentiality and ethics as part of their Managing and Sharing Data guide.

Long-term preservation

Data which are not archived or managed in a systematic way are in danger of becoming effectively lost. The passage of time increases this risk, and so active long-term preservation is needed.

As part of your data management planning, you hopefully will have identified the period of retention for which the data remain valuable: length of the project only, 5, 10, 20 years or indefinitely. This decision can be revisited after the data are in use. New uses for the data may come to light, or a historical value may be anticipated.

Fragility of data

Digital data - made up of bits and bytes - are in many ways more fragile than paper records for a number of reasons. Depending on the type of media on which the data are stored (magnetic, optical, and so forth), over time they are subject to different forms of 'bit rot' or decay, in which the electrical charge representing a bit disperses.

This gradually introduces either minor or major errors in the data, and their ability to be read by computer software.

Strategies

- Refreshment - move data files onto new storage media well within the projected lifespan of the media.
- Replication - by keeping more than one copy of a data file, the risk of losing a readable copy over time is reduced.

These strategies apply to both online and offline storage media. Where data are kept on a server, backup procedures and disaster recovery planning may take into account the necessary procedures. Ask your system administrator about their procedures and tests.

Offline storage media include optical discs such as compact discs (CDs) and digital video discs (DVDs). Depending on the quality, these may need to be refreshed every ten years or less. Portable flash drives can be useful for short-term backup and portability but are not reliable for preservation purposes.

Software obsolescence

Another threat to long-term accessibility of datasets is software obsolescence. When a new version of a software product is unable to render a file created in an older version, or when a software company retires a product, goes bankrupt, etc, there may be no available version of the software to be used on newer operating system platforms.

Strategies

- Migration - when a new software version has become established, the data file is converted or 'migrated' to the new software version or package.
- Emulation - a specialised strategy to recreate the functionality of the obsolete software package on a new operating system, or, for example, on a Java Virtual Machine system.
- Format conversion - the most pro-active method is to select a format that is most easily imported into a number of suitable software programs, or that is based on a universal standard.

Preservation planning

As most of the activity described above needs to be done at some future point(s) in time, planning for preservation is usually deemed to be essential.

The Open Archival Information Standard Reference Model (OAIS), originally developed by space scientists, is the pre-eminent model for preservation planning and is an International Standards Organisation (ISO) standard.

The Digital Curation Lifecycle Model is useful for mapping planned preservation and curation activities onto a lifecycle view of a digital object.

Two online books produced in the UK - Preservation Management of Digital Materials: The Handbook, by the Digital Preservation Coalition and the Digital Curation Manual by the Digital Curation Centre (DCC) are excellent sources of detailed information for those tasked with long-term preservation.

Data archives and repositories

Services may exist that could relieve you as a researcher of taking on long-term preservation of data yourself.

Digital preservation and data curation are represented by emerging professional fields that are increasingly specialised. Specialists are knowledgeable about preservation planning and procedures, as well as standards, informatics, and discipline-specific knowledge and norms.

A big advantage of depositing your data in an archive or repository is that it will be preserved - even for your own future use!

Deposit your data!

A big advantage of sharing research data by depositing them in an archive or repository is that they will be preserved - even for your own future use.

National services

Some of the national research councils fund data services to curate, disseminate, and preserve data created as part of their funded programmes e.g. the Economic and Social Research Council's UK Data Archive (also known as Economic and Social Data Service) and the designated data centres from the Natural Environment Research Council.

If you have created teaching materials around research data for use by students, these can be deposited in Jorum, a national repository of learning objects for UK Higher and Further Education.

- Depositing in the UK Data Archive (for ESRC grant-holders)

http://www.data-archive.ac.uk/deposit

- Natural Environment Research Council Data Centres

http://www.nerc.ac.uk/research/sites/data/

- Jorum learning object repository

http://www.jorum.ac.uk/

University data repository

Edinburgh DataShare is an institutional data repository set up by the Data Library to allow University researchers to deposit, share, and license their data resources for online discovery and use by others, either openly or on a restricted basis. Resources are managed according to a written preservation policy.

The Library offers guidance to PhD students on submitting their thesis and associated data to the University and whether to allow the thesis to be openly shared. Postgraduate students may choose to share their data openly through Edinburgh DataShare, with a link to their thesis in the Publications Repository.

- Edinburgh DataShare

http://datashare.is.ed.ac.uk/