Research Data Management
For Researchers

Dr Joanna Goodger
Information Hertfordshire

With Bill Worthington, Sara Hajnassiri, and Mohamed Hansraj
The majority of data is funded by the Government through Research Councils and Universities.

There are seven Research Councils, each covering a broad disciplinary area. Research Councils UK (RCUK) is a strategic partnership between the Councils, enabling them to work together and enhance the overall performance and impact of UK research.

Each council operates through grant programmes and some run their own research institutes and data centres.

The requirements laid down by the research councils cover various aspects of the research life cycle.
In response to these government policy changes, UH has reviewed its data policy UPR12. The University of Hertfordshire’s research data policy states:

The University owns data generated by its research; it expects data to be managed and shared in a robust and professional manner; and it places the responsibility for proper research data management with the Principal Investigator.

The University’s data policy also has the following clauses from the main policy:

4.2 The University recognises the value of data as an institutional resource and considers that value to be increased through the widespread and appropriate use of data and by virtue of data quality.

4.3 The University considers the value of data to be diminished through misuse, misinterpretation, or unnecessary access restrictions.
Why you should but maybe can’t share your data?

Spend a couple of minutes thinking about the benefits of sharing your data, having other data available for you, and what you might not be able to share.

Then we’ll combine your answers and see if we can resolve the issues.
Research Data Management
Benefits to Researchers

• More data is available for contrast and comparison
• More opportunities for collaboration as data is discovered
• Credit awarded on data as well as related publications
• Data preserved in managed archives
• Data is not lost when researchers leave research
• Student project data are managed assisting audit and reuse
Research Data Management

TOP TIPS
The DMP covers the whole project lifecycle.

The plan is divided into 6 sections, which cover the lifecycle but are divided into aspects of the project, not chronological steps.
What is data?

What does data mean to you? Spend a couple of minutes thinking about what data you will be working with, throughout your project.

Then we’ll combine your ideas and compare them between disciplines.
Research Data Management
Research Data Plan

Available from the RDM website:

http://rdm.herts.ac.uk/rdm/rdm-courses.html

Covers the lifecycle of your project on four sides of A4.
When planning your consent form, keep in mind that you and others may want to reuse the data.

Do not start with “We will not share or reuse your data”!

- Get permission for reuse in research
- Get permission to publish the anonymised data

Think about who might want to use it and how you can make that possible.

Also think about how to keep the impact of the data whilst making it anonymous.
Research Data Management
Legal and ethical issues

Legal Issues
- Do you have copyright issues?
- Is there a patent pending on your work?
- Is the data personal?

Ethics
- Who owns your data? UH? STFC? Third party company?
- How will the data be licenced?
- How will you deal with disputes?
Filing is more than saving files, it’s making sure you can find them later in your project.

- Naming
- Directory Structure
- File Types
- Versioning

All these help to keep your data safe and accessible.
<table>
<thead>
<tr>
<th>Formats</th>
<th>Uses</th>
<th>Considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Images</td>
<td>Reuse, paper, talk, poster, archive, web</td>
<td>Use, size, longevity</td>
</tr>
<tr>
<td>Raw, Processed, Plotted, Photos, Scans, CAD</td>
<td>FITS, JPG, PNG, BMP, PS</td>
<td></td>
</tr>
<tr>
<td>Tables</td>
<td>Code input, spectra, plot, paper, CDS</td>
<td>Use, metadata, accessibility</td>
</tr>
<tr>
<td>Catalogues, Query results, Calculations, Measurements</td>
<td>Text files, FITS, spread sheets</td>
<td></td>
</tr>
<tr>
<td>Source code</td>
<td>Third party edit, run. paper, web</td>
<td>User friendly; functions, size</td>
</tr>
<tr>
<td>Models, simulations, scripts, inputs, outputs, instructions</td>
<td>.c, .pl, .py, .idl, README, Make file, input, output</td>
<td></td>
</tr>
<tr>
<td>Interviews</td>
<td>Producing transcripts, further analysis</td>
<td>Format, longevity, security, metadata</td>
</tr>
<tr>
<td>Audio, Video, Written Transcript</td>
<td>.txt, .odt, .doc., mp3, .mp4, .avi</td>
<td></td>
</tr>
</tbody>
</table>
The format you choose will also affect the compression of your data and how much storage space you’re going to need to keep your data safe and accessible.

Consider a 5 Megapixel image.

The table below gives the size of that file in different standard formats. You can see what a difference your format makes to your storage requirements. You should think about which is best for your outputs: For the RDM website, resizing the image saves space and prevents the image becoming distorted by compression by the browser.

<table>
<thead>
<tr>
<th>Format</th>
<th>Size (MB)</th>
<th>Format</th>
<th>Size (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPG</td>
<td>1.5 MB</td>
<td>JPG resized (1024 x 776)</td>
<td>0.2 MB</td>
</tr>
<tr>
<td>PNG</td>
<td>9.0 MB</td>
<td>BMP</td>
<td>15.0 MB</td>
</tr>
<tr>
<td>BMP</td>
<td>15.0 MB</td>
<td>TIFF</td>
<td>3.0 MB</td>
</tr>
<tr>
<td>TIFF</td>
<td>3.0 MB</td>
<td>PDF</td>
<td>0.8 MB</td>
</tr>
</tbody>
</table>
Research Data Management
Data types and formats

Images / Photos

Plots

Code

Tables

Transcripts

Audio-Visual

Doctorial Training Conference
05/09/2013

Dr Joanna Goodger
j.l.goodger@herts.ac.uk
What is metadata?

Metadata is additional information that is required to make sense of your files – it’s data about data.

This is not a new idea; consider your music or film collection;

At least the title, authors, release date, producers, directors, etc.

Maybe the artwork, the studio, or the format it was released in such as LP (shown right), tape, CD, MD, Video, super 8, DVD, Blu-ray, 3D, etc.

All this information is metadata and allows you to make sense of the data and search the collection for the track that you're looking for.
What contextual details are needed?

Without additional information we do not know

- Who is in this picture?
- When was it taken?
- Where are they?
- Who took this photo?
- How was this picture taken?

All this information puts this image in context. Without it, it could be photo taken in the 1800s of Mr and Mrs Straus who died on the Titanic, or a Photoshop adjusted image of a young couple dressing up at Brighton pier in 2005.

Without additional information we just don’t know.
What is metadata?

Metadata is additional information that is required to make sense of your files – it’s data about data. You need to consider;

- **What contextual details are needed?**
  e.g. a description of the capture methods and data analysis.

- **How will you capture addition information?**
  e.g. in papers, in a database, in a ‘readme’ text file, in file properties/headers.

- **Which standards will you use and why?**
  Data centre recommendations for metadata, controlled vocabularies, and required documentation.

- **Whether there any encoding guidelines you should follow?**
Research Data Management
Activity

How is your data at risk?

What precautions do you have in place to safeguard your data? Spend a couple of minutes thinking about how your data could be lost, damaged or stolen?

Then we’ll combine your ideas and compare them to the DMP.
Backing up should be an automatic part of your everyday research activities.

In 2005, an electrical fault in the electronics and laser research building at the University of Southampton cost £50-100M including temporary building hire and transfer of work to Holland.

**Image if a fire or similar disaster happened at UH**

**How much would it cost you?**

Storing your data on the UH network means that it is stored at de Havilland and at college lane.
Research Data Management
Back Up

External HD
DVD, Tape

laptop

UH PC

local drive

Networked drives U: and X:

Cluster

UH server

5G

Doctorial Training Conference 05/09/2013

Dr Joanna Goodger
j.l.goodger@herts.ac.uk
Updates the changes to files between two directories and servers.

`/usr/bin/rsync -avu /data/jgoodger/ /local/data/`

Timed schedule to perform tasks – your rsync for example.

`SHELL=/bin/tcsh
PATH=/sbin:/bin:/usr/sbin:/usr/bin
MAILTO=a.user@herts.ac.uk`

`17 3 * * * /bin//usr/bin/rsync -avu /data/auser/ /local/data/`
Research Data Management
Short-term storage and data management

Windows Backup and restore

- Set an automated backup through control panel

Mac Time machine

- Back up your entire content to another disk or to the net.
Why keep a Lab Book?

Records are important for development and writing up of your research. You should keep a lab book of your research.

- a complete reconstruction of the experiment or measurement can be redone later
- the work can be repeated for re-evaluation of the reported results
- steps that led to the success or failure of a large project can be extracted
- patent lawyers need properly documented evidence of inventions
Paper lab books are at risk of loss or damage, and cannot be easily searched.

An electronic lab notebook (ELN) is a computer program designed to replace paper lab books;

- easier to search upon,
- simplify data copying and backups,
- and support collaboration
A good log should include:

- Steps and procedures and precautions which are not obvious
- References to other people's work, ideas, hints, and inputs
- Parameters which might affect the outcome of the experiment
- Equipment used, type numbers, serial numbers, any calibration steps taken
- Sketches of experimental layout and traces on recorders, oscilloscopes, etc.
- The date and time, names of other people observing
- Rough error analyses taken during the experiment, repeat observations of doubtful readings, calibration errors allowed for
## Research Data Management
### UH Shared Drives

<table>
<thead>
<tr>
<th>Drive</th>
<th>Capacity</th>
<th>Security</th>
<th>Back Up</th>
<th>Help</th>
</tr>
</thead>
<tbody>
<tr>
<td>My Docs on local machine</td>
<td>machine disc capacity</td>
<td>password protected</td>
<td>Up to you!</td>
<td>contact your local IT staff</td>
</tr>
<tr>
<td>U: personal network drive</td>
<td>Staff, 5GB students, 2GB</td>
<td>password protected</td>
<td></td>
<td>Helpdesk: ext. 4678 or email RDM Website</td>
</tr>
<tr>
<td>Research networked drive</td>
<td>&gt; 5GB</td>
<td>password protected and accessed by authorised members only</td>
<td>Daily UH backups held at College Lane and de Havilland data centres.</td>
<td></td>
</tr>
<tr>
<td>School networked drive</td>
<td>Unlimited (within reason)</td>
<td>password protected but accessed by all of the school members.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document Management System (DMS)</td>
<td>&gt; 5BG</td>
<td>password protected and accessed limited to group members only as selected by the project PI.</td>
<td>Replicated the Disaster Recovery System on the other campus. Nightly backups to tape.</td>
<td></td>
</tr>
</tbody>
</table>
Research Data Management
UH Shared Drives

<table>
<thead>
<tr>
<th>Document Management System</th>
<th>Research Drive</th>
</tr>
</thead>
<tbody>
<tr>
<td>External access</td>
<td>Free of file structure</td>
</tr>
<tr>
<td>Device independent</td>
<td>No automatic version control</td>
</tr>
<tr>
<td>Consistent file structure in place</td>
<td></td>
</tr>
<tr>
<td>Granular security</td>
<td></td>
</tr>
<tr>
<td>Automatic retention</td>
<td>Ideal for large files which are not documents, and would be costly to duplicate after every alteration.</td>
</tr>
<tr>
<td>Automatic file reporting and auditing</td>
<td></td>
</tr>
<tr>
<td>Full text search</td>
<td></td>
</tr>
<tr>
<td>Scanning straight to file</td>
<td></td>
</tr>
</tbody>
</table>

Dr Joanna Goodger
j.l.goodger@herts.ac.uk
The shared storage is accessible using Novell and the UH intranet.

**On Windows** machines, from off campus, or when connected to the student network, you need to activate Network Connect before logging into Novell. This allocates your connection a UH IP address.

**On Linux** machines, you need to use a Windows Virtual Machine to run Novell, but you can configure the network settings directly and do not need to run Network Connect.

**Mac** users should be able to connect direct to the drives without using Novell.
Sign in with your staff user name and password. In order to use the “drag and drop” facility you need to download additional add-ons.

Your project will be equipped with a project file structure, secure to your group with additional security for sensitive files in personnel and consent folders.
Research Data Management
Sharing

Many methods for sharing data and documents during the analysis and the writing up of projects have been developed by researchers, but most are unsecure and violate the UH data policy.

- Emails have a limit to the attachment size and can be intercepted or simply miss-sent
- Web space is generally unsecured and openly accessible
- Cloud storage is unreliable and gives third parties access to your data
- Storage media, such as DVDs and flash pens, can be lost, intercepted, or broken in transit
Research Data Management
Sharing

DMS
- Access only to UH members
- Versioning
- OS independent
- Set file structure

Research Drives
- Access only to UH members
- Undefined file structure
- No versioning

UH FTP
- Send large files using the UH server
- Web based only
- Open and Free
Research Data Management
Sharing

UH Sharing solutions are

• more secure and reliable
• accessible on multiple operating systems and on the web,
• accessible on and off campus.
• accessed by listed members only

Collaborators can be granted visiting member status; on a separate hierarchy so your collaborators will only be able to access folders that you assign to them.
Research Data Management
Document Management System (DMS)

- Research Design and Planning
  - Literature Review
  - Protocols and Research Methods
  - Regulation and Approval
  - Plans and Proposals
  - Technical Designs

- Research Funding Administration
  - Financial Reports
  - Budgets and Expenditure
  - Funding Award

- Personnel and Site Records
  - Contacts
  - Contracts and Agreements
  - Project Closure

- Project Administration
  - Progress Reports
  - Meetings

- Research Conduct
  - Site/Laboratory Management
  - Pharmacy and Medicinas Management
  - Equipment and Materials Management
  - Data Analysis
  - Risk and Adverse Event Monitoring
  - Logs
  - Data Collection
  - Participant Consent

- Research Reporting
  - Presentation
  - Publication

Key:
- Top Level Classification Folders
- Scientific/Clinical Trial Folders
- Highly Secure Folders
- Reference to Other Area
Most data needs some level of security:

- Sensitive Personal Information
- Proprietary data
- New discoveries
- Revolutionary code / software

All of it needs to be accessible, but secure during storing, sharing, and publishing.

If you lost it, who would be able to access your data?
Keep your data secure in an encrypted folder

**Bitlocker** is available on Windows 7 +

**Truecrypt** works on any operating system.

- Open Source Encryption that works with Windows, Mac and Linux
- Pack your files into an encrypted volume
- Send by email, shared drive, cloud storage, web space
- Password access
- Variable encryption algorithms available
Currently, selection, methods, algorithms, results, plots, and conclusions are in papers, published in journals and open archived in the ArXiv.

In the future you’ll need to select supporting data, including material with independent scientific merit for publication online in open access archives; either subject specific or in the UHRA.
Research Data Management

http://rdm.herts.ac.uk/

Planning

DMP\textit{online}

Subject Case Studies

Archiving

Document management system - DMS

UHRA D\textit{Space} Data Archive (Coming Soon)

Sharing

U:, R:, and S: drives

UHRA D\textit{Space} Data Archive (Coming Soon)

UH File Exchange:  https://fileexchange.herts.ac.uk/

Security

\textit{TrueCrypt} - Solution for encryption
Research Data Management (RDM)

A data legacy for future research

Since all research is fundamentally funded by the public, all published data needs to be open for reuse and verification. Being an effective researcher also means knowing how to look after your research data at all stages of your research project and beyond. The information in these pages will help you throughout your project, from planning and collecting your data to publication and archiving, helping you to manage your data with ease, efficiency and reliability for the future.

The information provided on these pages is the result of the Service Oriented Toolkit for Research Data Management project which ran from October 2011 to June 2013 at the University of Hertfordshire. The project was jointly funded by the University of Hertfordshire and the JISC Managing Research Data Programme.