SoMS Research Data Storage Guidelines

Never Stand Still Medicine Medical Sciences

The Australian Code for the Responsible Conduct of Research has been developed by the HRMC, ARC and Universities Australia. Compliance with the Code is a prerequisite for receipt of National Health and Medical Research Council funding.

§1.3 “It is important that institutions provide induction, formal training and continuing education for all research staff, including research trainees. Training should cover research methods, ethics, confidentiality, data storage and records retention, as well as regulation and governance.”

Data storage (§2) underpins top level aims of good stewardship of public resources and responsible communication of research results.

Introduction

These guidelines are designed to ensure that the contents of SoMS' servers are

- relevant,
- identifiable, and
- secure.

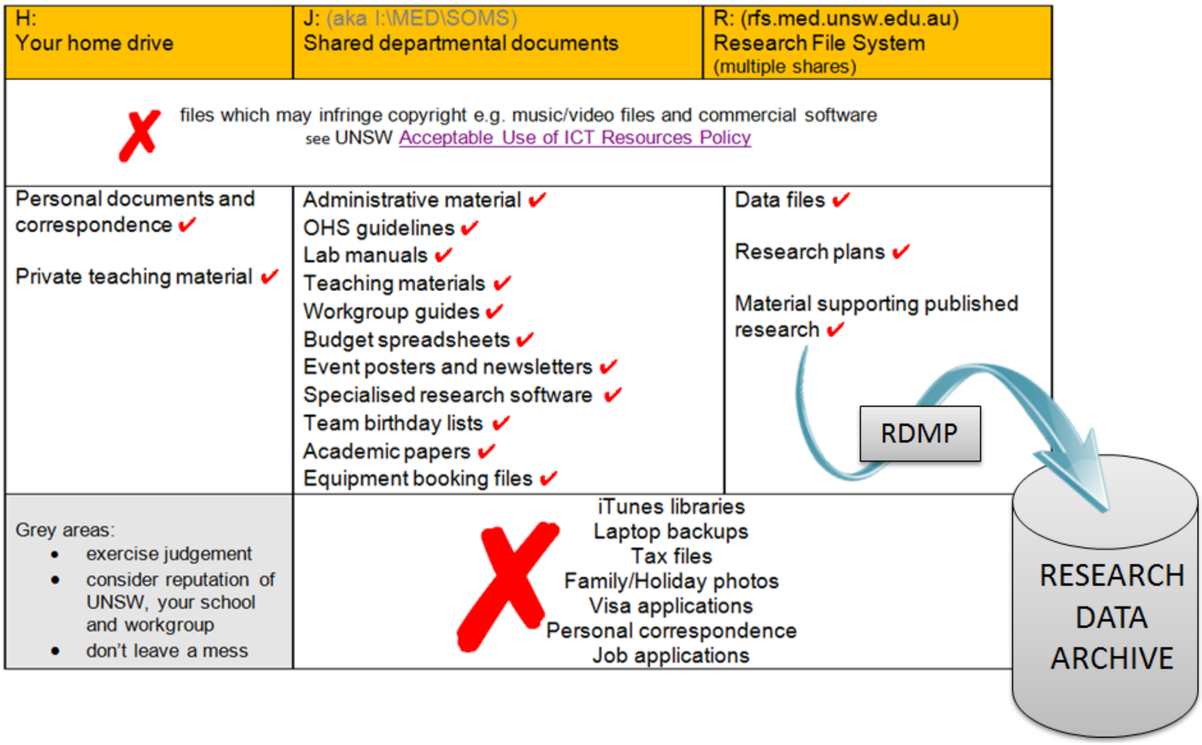
Storing the right data in the right place protects staff, teaching and research work, and the University's reputation. Failing to do so may mean re-creating documents, redoing experiments, paying to repeat microscopy, hours of tedious administrative clean-up and searching through multiple backup tapes, or even retracting a published paper.

Current information technology and data gathering devices (satellites, microscopes, scanners, …) allow us to keep vastly more information than ever before. Our *working* server space is limited, but we have virtually unlimited long term storage or archive space.

In order to preserve research data in archives and recall it when needed, the files must be well labelled with details of project, personnel, dates, associated funding etc. It's easier to "label as you go" than try to recall who owns what months or years down the track.

UNSW provides several options for storing data on its networks.

- H: (Home) drive – your personal documents
- J: files you share across your school, department or research group – generally DOC, XLS, PPT, PDF and other standard document types. May include training videos e.g. for laboratory procedures.
- R: research files – could be any file type, but with biomedical research, slide images and behavioural videos commonly translate to an abundance of large TIFF and video files.
- RDS (Research Data Store) – long term store for projects old and new. Most data on R: should be duplicated there, or sent there for permanent archive. It is *not* for storing files as you work with them. RDS will store multiple versions of files automatically. A Research Data Management Plan (RDMP) must be completed to gain access to RDS.



H: Your home drive	J: (aka I:\MED\SOMS) Shared departmental documents	R: (rfs.med.unsw.edu.au) Research File System (multiple shares)
✗ files which may infringe copyright e.g. music/video files and commercial software see UNSW Acceptable Use of ICT Resources Policy		
Personal documents and correspondence ✓ Private teaching material ✓	Administrative material ✓ OHS guidelines ✓ Lab manuals ✓ Teaching materials ✓ Workgroup guides ✓ Budget spreadsheets ✓ Event posters and newsletters ✓ Specialised research software ✓ Team birthday lists ✓ Academic papers ✓ Equipment booking files ✓	Data files ✓ Research plans ✓ Material supporting published research ✓
Grey areas: • exercise judgement • consider reputation of UNSW, your school and workgroup • don't leave a mess	✗ iTunes libraries Laptop backups Tax files Family/Holiday photos Visa applications Personal correspondence Job applications	

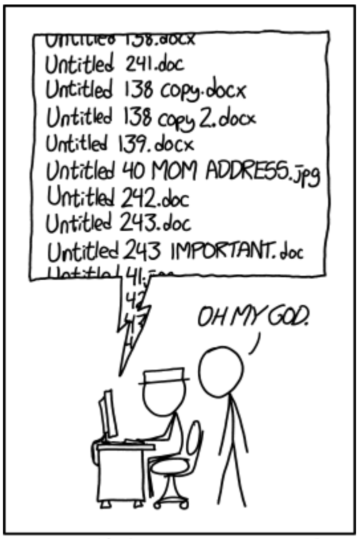
Principles

Relevant

- The files relate to the research, teaching and administrative work of SoMS
- Files belonging to departed staff & students, or pertaining to completed projects should be properly archived.
- Personal files (photos, correspondence, tax files etc) should go to personal **H:** drives or possibly private cloud storage (DropBox, OneDrive etc)
- Copyright-infringing music, video, software etc should not be stored on UNSW systems (see the Acceptable Use of ICT Resources policy)
- Backups of desktop or portable computers do not belong on our network drives
- When staff/students leave, their files should be sorted, then redistributed, archived or deleted as appropriate. Ideally this is done before they leave as part of exit process. Responsibility for this is owned by their supervisor/manager/sponsor (for visitors).

Identifiable

- Files are stored with their appropriate workgroup, teaching or research unit.
  - Don't store a second/third/… copy under another group's folders. Apart from the space usage – about 20% of the 5 million files on our servers appear to be duplicates – it is rarely apparent which set of files is the master version.
- Folders are identified by project names and/or **full names** of individuals. A text document in the root folder may supply additional details not covered by the folder's name or position in folder hierarchy such as dates, affiliated researchers, capture device…
  - "Lastname, Firstname" or "Firstname Lastname" are both fine, just have a system and stick to it.
- Check your spelling – folders and documents are hard to locate when you misspell them e.g. fl*uo*rescent is not fl*ou*rescent, and there is only *one* way to spell *Aperio*.



PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

Secure

- Your files are kept on backed up network servers that require UNSW zPass authorisation matching your work unit.
- Research data is matched to a RDMP (Research Data Management Plan) and archived to university's long term data store. Archiving should be done regularly and from project inception. Don't wait till the end of a project to archive. Research data sent to the archive may include budgets, timetables, and other relevant material.
- Avoid very long file and folder names (typically > 250 characters) due to the potential for file loss or misreading when opening or moving them.
- Data must not be kept exclusively on local or attached hard-drives, or USB sticks. The UNSW Research Data Guidelines are quite clear:
  - §5.2 : "Retention solely by the individual researcher is not permitted, as it may not protect the researcher or UNSW in the event that the veracity of the data is questioned."
  - §9.2 : "The Deputy Vice-Chancellor (Research) may determine that a breach of this Procedure may be dealt with as a Breach of the Research Code, or Research Misconduct."
- Confidential data is subject to loss, theft and malware when stored on portable drives.

FILING: DOs and DON'Ts

No single filing system is going to work for all research groups. The important thing is to have "a" system that others can easily follow, rather than shoehorn everyone into the same system.

The most common strategy is to organising around project titles and and/or full names of researchers. I suggest having an ARCHIVE folder to progressively file completed material ahead of transfer to long term storage.

- DON'T put filler words into file and folder names e.g. MAX'S FILES – obviously contains files, but it's not obvious who Max is. Ditto MAX'S FOLDER.
  - "NEW FOLDER", "NEW FOLDER(2)" etc – has no meaning
  - "FOR JENNY" – for Jenny who? For how long?
- Avoid redundancy – C:\BROWN HAIR\BROWN HAIR AND GREEN EYES\BROWN HAIR, THREE LEGS AND GREEN EYES\ETC which could be expressed as C:\BROWN HAIR\GREEN EYES\THREE LEGS
  - SUGGESTION: Put text files in root folder with extra descriptive text, especially if folder is full of raw data files and images.
- Avoid characters that may not travel well between Windows, Mac and web views: smart quotes **""**, ampersands **&**, trailing or leading spaces: C:\ RESEARCH \2014.
- DON'T rely on file-system dates to identify files and folders – these may change when they are moved to another drive.
- If naming folders or files based on dates, format them as YYYYMMDD – these sort alphabetically into the correct chronological order. Sorting by day (DDMMYY) or month-name (NOV 11, 2014) does not. YYYY-MM-DD also works, just be consistent
- When working with private personal data of human research subjects, don't put their names in file or folder names where they can be read by anyone casually browsing.
  - SUGGESTION: keep such data in a password-secured zip file, with the password noted in a safe place and/or keep the password in your secured long term data store.
- Don't *move* files between server folders with different access permissions (e.g. J:\COMMON and J:\CGMU). *Copy* the files and then delete the originals. This is because access permissions travel with the files. If you *move* files from folder A to B, then those viewing folder B will not be able to see or open the files as they retain folder A permissions. When files are *copied* the new versions inherit the permissions of the folder they were created in.
- Don't store files on your local PC/Mac desktop – these files are not automatically backed up!
- If you only store files on an external drive or local server, this is not a backup. It is data loss waiting to happen.
- As noted above, backups of other computers don't belong on the servers. Folders named 2010 BACKUP OF BROKEN PC or JILL'S OLD HARD-DRIVE (usually with dozens of subfolders containing random desktop files, unused toolbar links, sample videos and antivirus reports from 1998) are about as ordered as a desk drawer. Most of it is junk and very likely contains personal files overlooked by their owner. Clean it up while the content and structure has meaning!
- Specialised file formats may require specialised software to read them. Please consider backing up a copy of the software (appropriate version) as well.

Files don't just have a location; they also have a use-by date. The end of a project, the departure of a colleague or student signal occasions to carefully assess the state of associated files and folders.

FAQ

How do I archive content that cannot be attributed to a single person or project?
We have data collections that are the result of pooled research efforts, or the consequence of departed staff and students who have not fully labelled their data.
In either case, prepare an RDMP that contains as much information as you can about the collection, and identifies staff past and present who may require access to these files.

Our team pools research data and generates papers on subsets of that data. How do we archive?
Maintain an RDMP/archive for the pooled data, and create additional RDMPs for each paper. Archive an additional copy of the relevant raw and processed data for the paper. This will simplify the process of sharing data with other researchers.

The metadata for the two RDMPs should have enough details (personnel, FoR codes, organisational structure: Faculty of Medicine|School of Medical Sciences|Chief Investigator) to link the two in case the smaller archive needs to be supplemented with data from the team pool.

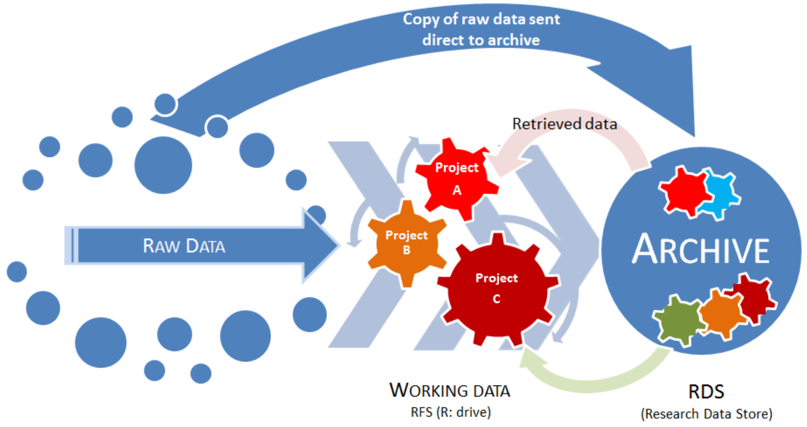
Why aren't we just putting everything in the cloud?
The university is exploring this option. However it is important to note

1. We can't just transfer a mess of data from local servers to the cloud. It has to be curated: trimmed and organised. If data is to be released to a third party, it should look as professionally presented as your resume, not like the forgotten shelf of a domestic refrigerator.
2. Hooking up cloud drives to work transparently with UNSW login, and apply workgroup security controls is not trivial.
3. Storage and network performance have significant $ costs attached. Do you want to pay for perpetual storage of someone else's messy data?

Are there limits on total file size a user can place on the RFS server (R: drive)?
Yes, the RFS drives are finite and will not be expanded. If you are planning to add a substantial amount of data (500GB or more) then you should advise SoMS administration ASAP.

If your project generates large amounts of data (> 1TB) it is in your interest and the School's interest to advise in advance so that server space provision may be made. New research partnerships, new students, new devices and firmware updates to existing devices will rapidly drive up file size and quantity. This should be communicated up through your organisation and also directly to SoMS administration.

Think about how much of your data needs to be shared and quickly accessible via the RFS server. Much of it may be better sent to the archive after processing and retrieved on demand.



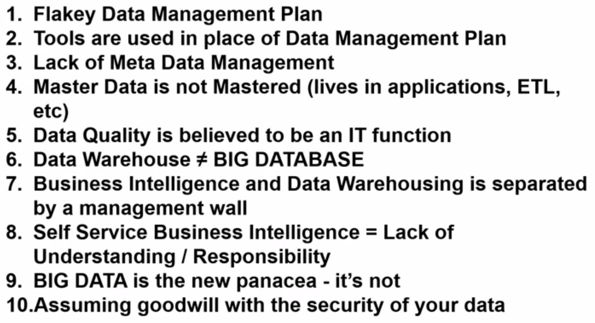
Further Reading:

The Australian Code for the Responsible Conduct of Research
(https://www.nhmrc.gov.au/guidelines/publications/r39)

Digital Curation Centre – What is digital curation? How-to guides.

University of Queensland – UQ Library – Research Data Management
Oxford University – Infrastructure for Research Data Management (video)

Top 10 Mistakes in Data Management (video)



1. Flakey Data Management Plan
2. Tools are used in place of Data Management Plan
3. Lack of Meta Data Management
4. Master Data is not Mastered (lives in applications, ETL, etc)
5. Data Quality is believed to be an IT function
6. Data Warehouse ≠ BIG DATABASE
7. Business Intelligence and Data Warehousing is separated by a management wall
8. Self Service Business Intelligence = Lack of Understanding / Responsibility
9. BIG DATA is the new panacea - it's not
10. Assuming goodwill with the security of your data

Figshare allows researchers to publish all of their data in a citable, searchable and sharable manner. All data is persistently stored online under the most liberal Creative Commons licence, waiving copyright where possible. This allows scientists to access and share the information from anywhere in the world with minimal friction.

Document History

VERSION	DATE	CHANGES	AUTHOR
1.0	2014-11-14	Document created	Mike Williams
1.1	2014-11-18	Reviewer feedback incorporated	Mike Williams
1.2	2014-11-25	Minor edits	Mike Williams
1.3	2014-11-25	Edits and FAQ additions	Mike Williams
1.4	2014-12-09	Minor edits	Mike Williams
1.5	2014-12-17	Edits, software backups, added Further Reading	Mike Williams
1.6	2014-01-05	Australian Code for the Responsible Conduct of Research	Mike Williams
1.7	2014-02-04	Edits throughout	Mike Williams