Information Literacy Program

ANU Data Management Manual

Managing Digital Research Data at the Australian National University

2017
# Table of Contents

1 Introduction .................................................................................................................. 1
   1.1 Objectives .................................................................................................................. 1
   1.1.1 Data Management at ANU .................................................................................. 1
   1.1.2 Outline ................................................................................................................. 1
   1.1.3 Data Management ................................................................................................. 2
   1.1.4 Benefits and Requirements ................................................................................. 2
   1.1.5 Methods of Data Management .......................................................................... 2
   1.1.6 ANU’s Data Management Services ................................................................. 2
   1.1.7 Writing a Data Management Plan .................................................................. 2

2 Data Management ...................................................................................................... 3
   2.1 Data ......................................................................................................................... 3
   2.2 Data Management ................................................................................................. 3
   2.3 Data Management Plan .......................................................................................... 4

3 Benefits and Requirements ....................................................................................... 5
   3.1 Benefits of Data Management .............................................................................. 5
   3.1.1 Efficiency .......................................................................................................... 5
   3.1.2 Protection ......................................................................................................... 5
   3.1.3 Quality .............................................................................................................. 5
   3.1.4 Exposure ......................................................................................................... 6
   3.1.5 Reproducibility ............................................................................................... 6
   3.2 Benefits of Data Archiving and Sharing .............................................................. 6
   3.3 Funding and Legislative Requirements ............................................................... 7
   3.3.1 ANU Responsible Practice of Research Policy (summary) .............................. 7
   3.3.2 Australian Code for the Responsible Conduct of Research ......................... 7
   3.3.3 ARC Funding Agreement for Discovery Projects ........................................... 7
   3.3.4 NHMRC Principles for Accessing and Using Publicly Funded Data for Health Research .............. 8

4 Methods of Data Management ............................................................................... 9
   4.1 Data Organization ................................................................................................. 9
   4.1.1 Bibliography Management ............................................................................... 9
   4.1.2 File Transfers and Remote Access .................................................................. 9
   4.1.3 File Synchronization ....................................................................................... 11
   4.1.4 Collaboration .................................................................................................. 11
   4.1.5 Revision Control ............................................................................................ 11
   4.2 Data Administration ............................................................................................. 12
   4.2.1 Backups ......................................................................................................... 12
   4.2.2 Data Validation and Authentication ................................................................ 13
   4.2.3 Documentation ................................................................................................ 14
   4.2.4 Access Controls ............................................................................................. 14
   4.2.5 IT Security ...................................................................................................... 14
   4.3 Data Archiving and Sharing ................................................................................ 15
   4.3.1 Data Sharing Methods .................................................................................... 15
   4.3.2 Copyright and Licensing .................................................................................. 16
   4.3.3 File Formats and Standards ............................................................................ 17
   4.3.4 Access Restrictions ......................................................................................... 17
   4.3.5 Archiving ......................................................................................................... 17
   4.3.6 Metadata ......................................................................................................... 18

Information Literacy Program
Managing Digital Research Data

5 ANU’s Data Management Services .............................................................. 19
  5.1 Local IT Support Staff (LITSS) ................................................................. 19
  5.2 Information Technology Services (ITS) .................................................... 20
    5.2.1 HomeDrive ....................................................................................... 20
    5.2.2 Alliance ........................................................................................... 20
    5.2.3 Webserver ....................................................................................... 20
  5.3 Information Literacy Program ................................................................. 21
  5.4 Open Research@ANU ............................................................................ 21
  5.5 ANU Data Commons ............................................................................. 21
  5.6 National Computational Infrastructure (NCI) .......................................... 22
  5.7 Australian Data Archive (ADA) .............................................................. 22
  5.8 Australian National Data Service (ANDS) ............................................. 23

6 Writing a Data Management Plan ................................................................ 24
  Project description ....................................................................................... 24
  6.1 Survey of existing data ........................................................................... 24
  6.2 Data to be created .................................................................................. 24
  6.3 Data organization methods ..................................................................... 24
  6.4 Data administration issues ...................................................................... 25
    6.4.1 Funding and legislative requirements ................................................ 25
    6.4.2 Data owners and stakeholders ........................................................... 25
    6.4.3 Access and security ......................................................................... 25
    6.4.4 Backups .......................................................................................... 25
  6.5 Data sharing and archiving ..................................................................... 25
    6.5.1 File formats, standards and conventions ........................................... 25
    6.5.2 Sharing ............................................................................................. 26
    6.5.3 Archiving and disposal ..................................................................... 26
  6.6 Responsibilities ....................................................................................... 26
  6.7 Budget .................................................................................................... 26

7 Glossary of Terms ..................................................................................... 27

8 References and Links ............................................................................... 30
  Other Resources ......................................................................................... 31
1 Introduction

Data Management is an integral part of modern research. Almost all researchers manage various forms of digital data — measurements from instruments, survey records, multimedia, or documentation.

Data management involves activities such as backups, collaborative work, data security, and archiving.

Managing your data allows you to work more efficiently, produce higher quality data, achieve greater exposure for your research, and protect your data from being lost or misused.

This document gives an overview of data management at the Australian National University.

1.1 Objectives

- Understand what research data is and why it needs to be managed.
- Appreciate legal, institutional and funding issues related to data.
- Learn how various data management methods can help you work more effectively with your data.
- Develop an awareness of existing data management services at ANU.
- Write a data management plan.

1.1.1 Data Management at ANU

The Information Literacy Program (ILP) offers an introduction to Data Management¹ at the Australian National University. The Data Management manual can be accessed at Research & Learn section of the ANU Library².

1.1.2 Outline

While this document is intended to be read in order, it can also be used as a handy reference. For instance, if your aim is to quickly put together a Data Management Plan (see Chapter 6), or begin using the ANU’s Data Management Services (see Chapter 5), you can skip ahead to those chapters and refer back to the earlier sections as needed.

¹ Workshop schedule and bookings at https://services.anu.edu.au/training/data-management-ildm01
Managing Digital Research Data

An outline of this document and a summary of the key points are as follows:

1.1.3 Data Management

- All researchers have digital data. While publications can be the most commonly encountered forms of digital data, researchers may also make use of data concerning measurements, survey responses, multimedia references, etc.
- Data Management can be loosely defined as "Anything outside of actually using the data." For example, organization, protection, and distribution of data.
- A Data Management Plan (DMP) is a document that describes what data will be created during a project, and how it will be managed.

1.1.4 Benefits and Requirements

- The key motivation for doing good data management is so you can spend more time using the data to comply with data management policies.
- There are a number of policies relating to data management: the ANU Policy for Responsible Practice of Research, the Australian Code for the Responsible Conduct of Research and the ARC Funding Agreement for Discovery Projects. Most relate to the ethics and long term storage (archiving) of data.

1.1.5 Methods of Data Management

- Data Organization: Description of various methods for working more efficiently with data.
- Data Administration: Discussion of methods to protect and improve the quality of data.
- Data Archiving and Sharing: Details of Data Archiving for preservation, and Data Sharing for exposure and open research.

1.1.6 ANU’s Data Management Services

- Local IT Support Staff (LITSS) – provide your computer and software. May also provide a fileserver for backups and a webserver.
- ITS Infrastructure as a Service (IaaS) – manage ANU’s central file storage (HomeDrive) and webserver. They also manage Alliance, which is an online collaborative environment.
- Information Literacy Program (ILP) – provide training in using software and general IT skills.
- ANU Data Commons – ANU’s repository for long-term storage and dissemination of data.
- ANU Supercomputing Facility – High performance computing, visualization, and large data storage.
- Discipline specific archives – ADA (Social Sciences)

1.1.7 Writing a Data Management Plan

- Recommended structure of a Data Management Plan
2 Data Management

This chapter defines key terms such as data, data management, and data management plans.

Other commonly used terms (such as fileserver, FTP, and Open Access) can be found in the Glossary [see section 7].

2.1 Data

Throughout this document, ‘data’ will refer to digital research data. Digital research data is any data that is created during research that can be stored on a computer. This includes field notes, analog recordings, and non-digital images as they can be converted to digital images. Physical data such as biological specimens, soil samples etc. are not considered.

Digital research data can additionally include:

- Numerical data: instrument measurements, survey responses.
- Documentation: Publications, experimental methods, field notes, analytical methods, technical reports, dataset descriptions.
- Digital images: photographs, diagrams, graphs.
- Digital audio: audio data, interviews, wildlife recordings, language recordings.
- Digital video: high-speed recordings, interviews.
- Configuration data: Configuration and optimization settings for simulation and in-silico experimentation.

Although not strictly data, you may also wish to consider the storage of any software developed to analyze the data, as an aid to being able to reproduce the results of your research.

2.2 Data Management

Data Management can generally be considered as any activity involving data outside of actually using the data.

Data management is best defined as any and all of the following examples:

- Organizing data into directories/folders and using meaningful filenames.
- Keeping backups of data in case you accidentally delete or lose data.
- Storing final state data in an archive.
- Making data available to others via an archive or website.
- Ensuring security of confidential data.
- Collaboratively creating and using data with other researchers.
- Synchronizing data between desktop, laptop, USB key, cloud storage, etc.
- Maintaining a bibliography and electronic copies of relevant literature.

Data management involves organizing, protecting and distributing the data. Data management does not produce results but is an unavoidable consequence of working with data. The aim is therefore to spend as little time doing data management as possible so that more time is spent using the data productively. Typically, people only do data management when it is needed and therefore tend to use the most obvious methods. The obvious methods are often the most inefficient, i.e., they are time-
Managing Digital Research Data

...-consuming and error prone. Using more advanced and automated methods will reduce the amount of time spent managing data.

2.3 Data Management Plan

A ‘data management plan’ is a document that describes what research data will be created, what policies (funding, institutional, and legal) apply to the data, who will own and have access to the data, what data management practices (backups, access control, archiving) will be used, what facilities and equipment will be required (hard-disk space, backup server, repository), and who will be responsible for each aspect of the plan.

The best time to develop your data management plan is at the beginning of your research. Any time spent on creating a robust and easy to use data management framework will be rewarded many times over during your research.
3 Benefits and Requirements

This chapter describes the benefits to researchers of data management as well as some of the institutional and funding requirements related to data management.

3.1 Benefits of Data Management

Research data is a valuable asset and data management should be seen as a necessary part of good research. The benefits of data management are:

- **Efficiency** of research through good organization, collaboration and documentation of data.
- **Protection** of data against becoming lost, unusable, forgotten, or improperly released.
- **Quality** of data through procedures to ensure data is accurate and authentic.
- **Exposure** of research outcomes through collaborations with others and dissemination of results and publications.
- **Reproducibility** of experimental and computational outcomes enables easy validation and verification of results.

3.1.1 Efficiency

Data management can improve the efficiency with which you work with your data. Typically organization and documentation of data are only done when they are absolutely necessary. Using software for version control and collaboration, and documenting data when it is created, will save time and allow you to work more efficiently with your data.

3.1.2 Protection

Data is a valuable asset so it is worthwhile protecting it from accidental loss or improper release.

Most people recognize the risk associated with losing data through accidental deletion and equipment failure, theft or destruction. Multiple and backup copies are therefore often kept for important data, but researchers should also consider using automated backup facilities to back up all their data.

Data management also protects the data from being improperly released. This is important where the data contains confidential or commercially valuable information. Improperly releasing data can violate privacy laws, confidentiality agreements, and possibly void intellectual property claims. It is therefore important to have well defined access rules for your data.

3.1.3 Quality

It is important to ensure the quality and authenticity of data that will be used for analysis and generating conclusions. Inaccurate data can invalidate results and conclusions resulting in lost time and damaging reputations.

Likewise, making any software or data analysis scripts available along with your data can help substantiate your results. This is particularly important when dealing with large datasets or complex analyses.
Managing Digital Research Data

It is also important to ensure the authenticity of data to avoid claims of plagiarism and ownership disputes.

### 3.1.4 Exposure

Creating a website for your research and placing your publications and research data in an archive greatly increases the exposure of your research. Research has shown that Open Access (OA) publications receive 2-3 times as many citations as articles that are only available via journal subscription\(^3\).

### 3.1.5 Reproducibility

Reproducibility of results and independent verification is an important criterion for research\(^4\). Ensuring a proper record of provenance and context trail facilitates recreation and analysis of critical research hypotheses and data. Again, making any software developed available can aid the verification process.

### 3.2 Benefits of Data Archiving and Sharing

Data sharing makes for good research as it allows for independent verification of results and conclusions and further analysis through the reuse of data.

An excellent list of the benefits of data sharing is given by the ICPSR’s *Guide to Social Science Data Preparation and Archiving*\(^5\):

- Reinforces open scientific inquiry. When data are widely available, the self-correcting features of science work most effectively.
- Encourages diversity of analysis and opinions. Researchers having access to the same data can challenge each others’ analyses and conclusions.
- Promotes new research and allows for the testing of new or alternative methods. Examples of data being used in ways that the original investigators had not envisioned are numerous.
- Improved methods of data collection and measurement through the scrutiny of others. Making data publicly available allows the scientific community to reach consensus on methods.
- Reduces costs by avoiding duplicate data collection efforts. Some standard datasets, such as the General Social Survey and the National Election Studies, have produced literally thousands of papers that could not have been produced if the authors had to collect their own data. Archiving makes known to the field what data have been collected so that additional resources are not spent to gather essentially the same information.
- Provides an important resource for training in research. Secondary data are extremely valuable to students, who then have access to high-quality data as a model for their own work.

---


3.3 Funding and Legislative Requirements

There are some key funding and legislative requirements relating to data management. Basic data management is required by the *Australian Code for the Responsible Conduct of Research*. Compliance with the Code is already a requirement for ARC and NHMRC funding and is likely to be mandated by other funding bodies, government and institutions in the near future.

The following sections summarize the policies relating to data management.

### 3.3.1 ANU Responsible Practice of Research Policy (summary)

- ANU policy document
- Data management should comply with the *Privacy Act (1988)*
- Research results should be open to scrutiny. Non-confidential data related to publications must be made available
- Data must be retained for at least 5 years. Retention must comply with the *Archives Act (1983)*
- Researchers should not unnecessarily enter into research agreements that limit access to information
- Researchers are responsible for data security
- Supervisors must ensure the validity of data gathered by their students
- Research misconduct includes: fabrication, falsification, and interference with data

### 3.3.2 Australian Code for the Responsible Conduct of Research

- The Code document
- Published research data should be retained for at least the minimum period specified by institutional policy and as long as scholarly interest and discussion persist
- Research data should be made available unless prevented by ethical, privacy, or confidentiality matters
- If research results are challenged, all data must be retained until the matter is resolved
- Researchers must keep records of research methods and data sources
- Researchers must retain research data in a durable, indexed, and retrievable form and maintain a *catalogue* of the data in an accessible form

### 3.3.3 ARC Funding Agreement for Discovery Projects

- Data from research in the social sciences should be archived with the Australian Data Archive (ADA, see Section 5.7) within two years
- The ARC strongly encourages the depositing of data arising from a Project in an appropriate publicly accessible subject and/or institutional repository.

---

3.3.4 NHMRC Principles for Accessing and Using Publicly Funded Data for Health Research

- The Principles provide information and guidance for researchers and organisations when researchers seek permission from organisations to access and use data for their research.\(^{10}\)
- The Principles have been developed by NHMRC on the advice of researchers, consumer representatives and organisations that hold data. They represent a common view about sharing the data, and the roles and responsibilities of all parties.
4 Methods of Data Management

This chapter outlines some of the methods of data management.

4.1 Data Organization

Data organization is about working more efficiently with data. Creating and using data requires some level of data organization. Often this organization becomes time consuming and error prone, in which case automated data organization methods should be considered.

Each section lists the standard methods of dealing with data organization and their drawbacks. Some automated and more efficient alternatives are suggested, but keep in mind that they often require some configuration and familiarization with the software. If the standard methods are adequate for your needs, then it is best to continue using them. If you think you are spending too much time organizing your data, then you should consider looking into the advanced methods.

4.1.1 Bibliography Management

Creating a bibliography manually is time consuming and error prone. Journals and conferences will usually specify a particular citation style, so it is best to generate citations automatically to save time and avoid errors. Furthermore, researchers often have hundreds of academic articles stored on their computers as part of the literature review. Finding a particular article can become time consuming.

There are a number of software-based reference management tools that automate citations and bibliography creation when writing an article. They also organize references into a database, making it easy to sort and search. Most of these programs also offer the ability to search online academic databases, such as IEEE Xplore, CiteSeer, ArXiv, and PubMed.

EndNote is a popular reference management tool and ANU has an institutional license which allows staff and students to install EndNote on their office and home computers, including laptops\(^1\). The Information Literacy Program also runs training courses in EndNote.

Unfortunately, EndNote does not run on Unix and cannot manage BibTex bibliographies; therefore, LaTeX authors and Unix users can use JabRef, which is a free program, runs on all operating systems\(^2\). JabRef can also import and export BibTeX’s and EndNote’s database formats.

Zotero\(^3\) is a free and open-source bibliography management ‘add-on’ for the Firefox Web browser\(^4\). A stand-alone version of Zotero is in beta development. It runs on all operating systems and has support for exchanging bibliographic databases with both BibTex and EndNote. The Zotero bibliographic database can also be shared using access to ANU’s file space [see Section 5.2.1].

---

\(^1\) Endnote, www.endnote.com
\(^2\) JabRef, http://jabref.sf.net
\(^3\) Zotero, www.zotero.org
\(^4\) Firefox Web Browser, www.mozilla.com
4.1.2... File Transfers and Remote Access

Researchers collaborating on projects will often need to share primary data and preliminary results; hence, it is often necessary for them to transfer data between computer systems. Researchers may also wish to transfer data stored on their university computer from outside the university, such as when overseas.

The most common method for transferring files is with email attachments, but there are limits to the size of files that can be transferred. Removable data storage media, such as USB keys and portable hard disks can transfer large amounts of data, but require the researcher to physically carry the data to its destination.

Specialist services such as CloudStor\(^\text{15}\), provided by AARNET can be used to transfer files without needing to use portable media to transfer files. While primarily a service for the exchange of data between AARNET member institutions you can generate a voucher or magic code which you can send to a colleague at a non AARNET institution to allow them to use Cloudstor to either send file to you or receive data from you.

To assist good data management, the ANU provides local area network and Internet access to Homedrive, a central storage space on which each member of the University is allocated file space to store personal files (4.5GB for students and staff). Homedrive is accessible from any campus computer or computers outside campus\(^\text{16}\)[see Section 5.2.1].

ANU-provided Web applications, such as Alliance (see page 20), allow data to be accessed and sometimes modified with just a web browser. If the Web Application allows data to be modified, such as a wiki, then the data can be edited in-place on almost any internet connected computer.

4.1.3... File Synchronization

Often researchers will work on their university desktop as well as a laptop, and possible a home computer. Typically files are just copied back and forth between the computers. This is the most obvious method but has a number of drawbacks.

- It is time consuming to manually copy files.
- You have multiple copies of data and you can easily lose track of which copy is the latest version.
- If both copies have been modified, then it is easy to overwrite some changes without knowing.

If you are synchronizing regularly or have lots of files to synchronize, then you should consider using file synchronization software. File synchronization software offers the following advantages over manual synchronization:

- Faster and requires less thought (usually just click a button).
- Automatically detects when two files have been modified and lets the user choose which one to keep. Some programs can also display the difference between the files.

One of the most popular file synchronization programs is WinSCP, which is primarily for SSH and FTP transfers, but can also synchronize data\(^\text{17}\).

\(^{15}\) https://cloudstor.aarnet.edu.au
\(^{16}\) https://myfiles.anu.edu.au/
\(^{17}\) rsync, en.wikipedia.org/wiki/Rsync
rsync is another widely used open-source utility for incremental file transfer and synchronization. It is cross-platform and can be used to generate ‘snapshots’ and regular backups.

The use of dedicated Version Control Software (see page 11) is another option for file synchronization.

Commercial, user-friendly, file synchronization services are becoming increasingly popular. **Dropbox**, for example, provides 2GB of data storage for free and also provides good user management tools to support collaborative work.

### 4.1.4... Collaboration

Many research projects are carried out collaboratively: between postgraduates and their supervisors; within departmental research groups; as cross-discipline research, and as inter-university research.

This is mutually beneficial as it improves access to funding; avoids repeating costly experiments; increases recognition through co-authorship; and can help lead to new research ideas.

For simple tasks this is usually done by transferring data by email, USB-key, or a network drive. Publications with multiple authors are often written this way – authors will take turns editing the document and email it to their colleagues, or the primary author will periodically email the latest version and their colleagues will reply with corrections and additions.

These methods are adequate for simple work and if there is only a small number of collaborators. It is worth considering using collaborative software tools, such as the ANU-provided **Alliance** (see Section 5.2.2) and/or version control software (see Section 4.1.5).

**Alliance** is a web-based service that allows ANU staff and students to easily set up collaborative project sites. Alliance provides a wide range of collaborative tools such as forums, chat rooms, calendars, and more as well as allowing users to share files and documents.

### 4.1.5... Revision Control

When the data is constantly being edited, especially by multiple users, it is a good idea to implement some form of version control to keep track of changes. This can be as simple as appending a number to the end of a file after each major edit. For example:

- Journal_v1.0.tex, Journal_v1.2.tex
- Journal_Feb12.tex, Journal_May5.tex
- Journal_Feb12_John_DRAFT_WithSallysEdits_NewDiagram.tex

Such conventions are good for simple work but quickly become unmanageable when you have multiple authors or make lots of edits.

The alternative is to use revision (or version) control software. These programs are used extensively for software development but are also excellent for documentation, such as writing a paper with several authors. Version control software also provides

---

20. **Alliance**, [alliance.anu.edu.au](http://alliance.anu.edu.au)
Managing Digital Research Data

access control, a collaborative work environment, synchronization between home/office/laptop computers, and a degree of data safety (although not as good as proper backups).

Such programs offer several advantages:

- The software requires you to input a description of the changes made, which makes it easier to pick up where you left off and for collaborators to see what you are doing
- You can be confident with making major changes as you can revert to an old version if you make a mistake. You can also easily compare two versions to help you find errors
- Useful for people who use more than one computer. It implicitly provides synchronization and is good for resolving conflicting changes

While the time required to learn the software may seem like a drawback, it is highly recommended for people in order to avoid regular problems with simple filename version control.

TortoiseSVN, for example, is a popular program that uses the Subversion system of version control\(^{22}\). It integrates with Windows Explorer making it one of the easiest version control programs to use.

While version control software is in some cases harder to set up, it provides more advanced version tracking. A distributed version control system like Bazaar\(^{23}\) can be used with Alliance to collaboratively manage documents and data.

Such tools also make it easier for any number of people to work on a document or code. It is also more efficient as everyone has access to the latest version and can make edits without conflicting with other people’s changes. The entire history of the document is also stored, making it easier to revert to an older version and for users to see what has changed they last looked at the data.

4.2 Data Administration

This section covers methods for protecting data and ensuring its quality. Many of these methods are necessary for compliance with the requirements of Data Management [see Section 3.3]. Some are policies, such as security and access, and others are practices, such as backups, quality control and documentation.

4.2.1 Backups

Making regular backups of data is probably the most important and, fortunately, one of the easiest tasks to manage.

Although most people are quite aware of the risk and cost of losing data through hard drive failure or accidental deletion, it is best to have a policy and schedule in place for maintaining data backups.

When considering your backup strategy, you need to know:

- How often will you make backups
- How long will backups need to be stored
- How much hard-drive space, or number of DVDs, will be required to maintain this backup schedule

---

\(^{22}\) TortoiseSVN, tortoisesvn.net/

\(^{23}\) Bazaar, bazaar.canonical.com/en/
Managing Digital Research Data

- If the data is sensitive, how it will be secured and (possibly) destroyed
- What backup services are available that meet these needs, if none, then what will be done
- Who will be responsible for ensuring backups are available

Backup security requires further mention. If the data is sensitive then it should not be stored on a computer that is connected to the internet, and preferably not connected to any network. If the data needs to be destroyed at the end of a project then consider what level is required – a hard drive will need to be overwritten several hundred times to ensure that no data can be recovered. Very high-level security institutions, such as Defence, require hard-disks to be physically destroyed and optical discs to be shredded.

The lifetime of backups should also be considered. Burned optical discs have average lifetime of two years, and five years if kept in a cool dark place.

If you are using a network drive then your data is probably already being backup up for you by IT staff. It is still a good idea to check with them to find out how often they backup, what is the maximum amount of data they can backup, and how long they keep old backups.

You may need to maintain your own backups if:
- There are no services available to you
- You have valuable data that you do not trust with other people
- You have sensitive data that you cannot store on unsecure computers (medical records, data for defence projects, etc.)

If you do maintain your own backups you need to ensure that all media is correctly labelled with the date and contents of the backup. Keeping a diary of what you backup and when can be helpful if you have to reconstruct a data set after a drive failure.

Remember that backups can sometimes fail due to faulty media, and it is important to keep more than one copy of important data.

4.2.2... Data Validation and Authentication

Your data will be used to obtain the results and conclusions of your research, so it is important to ensure its accuracy. Your data may also become an important dataset that is used by many others, so errors have the potential to hinder many research efforts.

It is therefore important to set up policies and practices to ensure the accuracy and authenticity of your data. This can include:
- Calibration of instruments.
- Use of Computer Assisted Interviews (CAI)\textsuperscript{24}
- Securing master copies to avoid accidental/intentional tampering.
- Data entry checks, such as two pass verification and range checking\textsuperscript{25}

4.2.3... Documentation

It is important to document the experimental or data gathering methods. Other researchers may question your results or want to repeat/extend your research, so it is important to document this. The sciences already have a culture of keeping good lab

\textsuperscript{24} Computer Assisted Interviewing, en.wikipedia.org/wiki/Computer-assisted_telephone_interviewing

\textsuperscript{25} Data Validation, en.wikipedia.org/wiki/Data_validation
Managing Digital Research Data

notes and the social sciences often record their survey methodology. This is often
done in a notebook, but you should also consider recording this information digitally
or converting it manually\textsuperscript{26}. This is important as notebooks are easily lost or put into
storage when an academic or postgraduate student leaves. This information is far
more useful if it is archived with the data it refers to. Scanners are available in most
ANU library buildings.

It is also valuable to document analytical methods. For example, if you write a
script/macro/program to help analyze the dataset by producing graphs or statistics
from your dataset.

4.2.4... Access Controls

Well-defined access controls help you comply with privacy and confidentiality policies
and help maintain data authenticity by limiting who can modify data. The access
controls may change throughout the life of the research project. Initially all data will
usually be restricted to the research group, when the results are published the data
may then be made available to other researchers.

Access controls can be defined on a per-user or per-data basis. When the data is
active and there are a small number of people using the data then you will probably
use per-user access permissions:

- None – has no access to the data
- Read – can read the data, but not modify
- Write – can read and modify data
- Administrator – has the ability to modify others’ access permissions

As an example, the principal researcher would have \textit{Administrator} permissions over all
data and may be the only one with \textit{Read} permissions of confidential survey data. Research collaborators would have no access to the confidential survey data, \textit{Read}
access to de-identified survey data, and \textit{Write} access to data analyses and
publications.

How you set access permissions varies between operating systems. In Windows they
are usually set by right-clicking on a file or directory and editing the security
properties. Some versions of Linux use a similar method.

4.2.5... IT Security

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

- Theft of valuable data
- Breach of confidentiality agreements and privacy laws
- Premature release which can void intellectual property claims
- Release of data before it has been checked for accuracy and authenticity

Security of digital research data is part of the issue of Information Technology
Security. The ANU has extensive range of policies and information related to IT
security\textsuperscript{27}.

\textsuperscript{26} You can scan a notebook and use software to extract the writing, but it may not work well unless you have
very good handwriting.

\textsuperscript{27} ANU IT Security, \url{security.anu.edu.au}
Managing Digital Research Data

The topic of IT Security is too large to cover here, but at the least you should install up to date antivirus software on your computer. ANU staff and students can install Sophos Anti-Virus on their office and home computers. If you have sensitive data that is 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. Encrypting data is a non-trivial exercise and there are currently no services at ANU to do this. It is important to remember that if you encrypt your data you need to be able to decrypt it and you need to ensure that you store the access code or password safely and securely.

The final issue to consider is physical security. A computer that is not connected to a network is still vulnerable to someone removing the hard-drive and installing it in their own computer where they can bypass passwords and access restrictions. For highly sensitive data you can use an external hard-drive and store it in a locked safe overnight.

### 4.3 Data Archiving and Sharing

Data archives are for long term preservation of digital data. Most digital storage media (optical discs, hard drives) have reliable lifetimes of only a few years. An archive ensures that data is preserved and maintained in file formats that are most likely to be useable in the future.

Data sharing is considered an important part of academic research that encourages open inquiry into research results and conclusions, as well as promoting data reuse and repurposing. Most archives facilitate data sharing and allow the data owner to maintain control over their data without needing to provide the facilities themselves. The benefits of data sharing are also covered in Section 3.1.

#### 4.3.1 Data Sharing Methods

Data dissemination is actively making your data accessible to others. Some researchers make their datasets available via their personal or group websites.

Data sharing is done in three ways:

- **Email request** – Interested researchers email and request the dataset. This is the most common way that data is shared.
- **Website** – Researchers place datasets on their website that anyone can download.
- **Archiving** – Researchers place their dataset in an archive.

Archiving is the preferred option as most archives have a search utility and are often indexed by the major web search engines, thus increasing the chances of other researchers using and crediting your datasets and publications. Archiving datasets also means the dataset owner does not need to maintain a website and can specify a wide range of access controls.

---

28 Sophos Anti-Virus Software, sophosupdate.anu.edu.au
29 Your Local IT Support Staff (LITSS) may be able to help.
Managing Digital Research Data

If your dataset is online, then including the link in your publications will greatly increase its use and exposure.

4.3.2... Copyright and Licensing

The owner of any original data holds copyright over that data from the time the data is created. In general, the ANU owns the copyright of material generated by staff in the course of their employment. The researcher, however, owns the copyrights on academic publications.

The owner is usually the creator, but some funding and research agreements require copyright to be handed over to another party.

Licenses grant permission for others to use the copyrighted data. Open Content Licenses are an easy way for researchers to license their data for others to use. Researchers can choose the most suitable license for their needs rather than develop a custom license themselves. The most notable open content licenses are:

- Creative Commons\(^{30}\) - most popular open content licenses.
- Science Commons\(^{31}\) - similar to Creative Commons but tailored for scientific data and publications.
- GNU Free Documentation License\(^{32}\) - used by Wikipedia.
- GNU General Publication Licence v3 – the most common open source software licence

4.3.3... File Formats and Standards

Before creating the data you should consider what formats and standards you should use as it is sometimes difficult to convert between file formats\(^{33}\). Using an inappropriate file format will also make your life more difficult in the long run.

Where possible, it is best to use preservation-friendly formats as they are more likely to be readable in the future and are easier to share with others [19]. It is usually safe to use a proprietary format if it is very widespread, as free programs will most likely exist to read these formats.

Some examples of preservation-friendly formats are:

- PDF – document format
- Open Document Format (ODF) – used by LibreOffice Writer among others
- PNG, TIFF, JPEG – Image formats

Your LITSS (see Section 5.1) and ANU Open Research@ANU staff (see Section 5.4) can give you advice on what file formats to use. For archiving, PDF (Portable Document Format) for documents and TIFF (Tagged Image File Format) for images are recommended. Note that most document and image formats can be converted to PDF and TIFF, respectively, but there may be some loss in quality.

\(^{30}\) Creative Commons, [creativecommons.org](http://creativecommons.org)
\(^{31}\) Science Commons, [sciencecommons.org](http://sciencecommons.org)
\(^{32}\) GNU Free Documentation License, [www.gnu.org/copyleft/fdl.html](http://www.gnu.org/copyleft/fdl.html)
\(^{33}\) File Format, [en.wikipedia.org/wiki/File_format](http://en.wikipedia.org/wiki/File_format)
4.3.4... **Access Restrictions**³⁴

When data is in a final state and ready for dissemination or archiving, you should define the *Access Restrictions* on each item of data:

- **Unrestricted** – Anyone can download.
- **Registered** – Users must give their name and affiliation so the data owner can track who is using their data.
- **Requested** – Users must submit a request outlining how they will use the data.
- **Closed** – No access (i.e., confidential data).
- **Metadata** (see section 4.3.6)

4.3.5... **Archiving**

Archiving of final research data is encouraged and in some cases required (see Section 3.3). Archiving your data ensures the data will not be lost, forgotten, or become unusable due to being stored in legacy file formats or storage media (see Section 3.1.5). Archiving also takes care of dissemination, access control and security.

Archives generally only accept final state data. The objective of the archive is to preserve the data and – if the data owner allows it – make the data available for further research. The owner of the data can specify a range of access restrictions (such as those described in Section 4.3.4); although, each archive will use different terminology. It is also possible to embargo data such that the data cannot be accessed until after a specified date. This is often done to give the data creators time to publish their results before making their data public.

An archive provides long-term storage of data and therefore prefers file formats that are unlikely to become obsolete. Most file formats can be converted to a suitable archiving format but some loss in quality (such as images or audio) or distortion (such as converting PowerPoint to PDF) may occur. Most archives are able to perform the conversion but it is best if the depositor does the conversion to ensure that they are happy with the result.

The time and costs associated with archiving are often underestimated. Each item of data deposited will need to have metadata written for it, which will be very time consuming if your data consists of several hundred images that were taken some years ago. It is therefore best to write metadata as the data is created and to archive data continuously rather than leaving it until the end of the project. It is recommended that you include the costs of archiving in your grant application.

4.3.6... **Metadata**

Metadata is often described as “data about data”³⁵, and helps people to discover, understand and re-use data. It can be used to describe physical items as well as digital items (files, documents, images, datasets, etc.). It includes descriptive metadata, such as title, time, author, keywords, relations to other data objects; administrative metadata such as ownership and use permissions; and history of changes to the data, versions etc. Good metadata is key for research data access and re-use.

Metadata are usually found in separate file with several text fields that describe the attributes of another piece of data, such as an experimental dataset, image, or video.

---

³⁴ Updated Access Restrictions: [http://services.ands.org.au/documentation/rifcs/1.6.1/vocabs/vocabularies.html#Access_Rights_Type](http://services.ands.org.au/documentation/rifcs/1.6.1/vocabs/vocabularies.html#Access_Rights_Type)

³⁵ [Metadata](en.wikipedia.org/wiki/Metadta)
Managing Digital Research Data

Sometimes, metadata is written into source data file itself, such as in jpeg files from digital cameras.

Metadata usually contains at least the following information about the data:

- Filename
- File size (kilobytes, megabytes, etc.)
- File type (LaTeX document, JPEG image, etc.)
- Date of creation
- Author or copyright holder
- Brief description
- Keywords

You can think of the metadata, in relation to the data it describes, as being analogous to the abstract or keywords of a paper – it is there to help people find your data and quickly decide if it is what they need. If you want people to find and reuse your data (and therefore help you by citing your work), then it is worth your while making good metadata in order to ‘sell’ your data.

Metadata is critical for archiving; most archives will not accept data that does not have adequate metadata. Creating metadata at the end of a project is also extremely difficult, as you may have to go through several hundred photographs or audio files. Metadata should therefore be made as the data is created.

Metadata schemas/standards include what are considered essential elements to describe any data. A list of commonly used disciplinary standards can be found in the ANU Data Management Libguide³⁶.

5 ANU’s Data Management Services

This section discusses the various data management services available at the ANU.

The ANU’s ITS and your Local IT Support Staff (LITSS) provide the majority of the data management services. They provide the day-to-day data management services like backups and shared drives.

Data archiving is more specialized and is performed by: ANU Data Commons, the ANU Supercomputing Facility (ANUSF), and the Australian Data Archive (ADA).

The IT Services website\(^ {37}\) contains extensive information on the ANU’s information services.

5.1 Local IT Support Staff (LITSS)

The majority of your data management needs will be provided by your Local IT Support Staff, or LITSS\(^ {38}\).

The services that they provide are usually determined by the head of college, so the services will vary between colleges and even schools. Most colleges will provide a file server and web hosting and employ staff as LITSS\(^ {39}\).

A file server is a computer that stores data and makes it accessible to your computer via a network connection. Normally it will appear as a mounted drive and behave the same as any other directory on your computer. It is recommended that you keep all your files on your mounted drive as the file server is automatically backed up at regular intervals and is less likely to fail. If you choose to keep your work on the hard-drive of your desktop/laptop, then you will need to perform the backups yourself.

If your data is sensitive then you should either store the data on your computer’s hard-drive or encrypt the data before placing it on the mounted drive. Your LITSS may be able to set up your account to automatically encrypt a directory within your mounted drive. If you forget the password it will be impossible to recover the data.

Most colleges have web servers for hosting faculty and department websites. They may also allow academic staff to use these web servers for personal or research group websites. This is an easy way to make your publications and datasets available online, but it is recommended to store the data in an archive, such as ANU Data Commons, and link to it from your website.

Your LITSS may also provide a mounted drive for collaborative work. Normally a file server provides a directory that only you can access, but your LITSS may be able to set up a mounted drive that several people can contribute to. Using a mounted drive creates problems such as keeping track of changes and simultaneous edits. If the data is being edited often or there are a large number of people using the data, then it is best to use version control software.

Finally, your LITSS are your first point of contact if you need software for data management, such as EndNote database tools, office suits, and conversion tools\(^ {40}\).

---

\(^ {37}\) ITS
\(^ {38}\) LITSS Information, its.anu.edu.au/litss/support-model.html
\(^ {39}\) Find LITSS, its.anu.edu.au/litss/contacts/
\(^ {40}\) Software Information, information.anu.edu.au/daisy/infoservices/8/29.html
5.2 Information Technology Services (ITS)

Information Technology Services group at the ANU are responsible for the undergraduate computer laboratories, the major mail servers, and more. In regards to data management, they provide the Homedrive fileserver, Alliance, Wattle and other webservers.

5.2.1... Homedrive

Homedrive is a personal network storage space where all research staff and students can store university related documents and files. Research staff and students are allocated 4.5GB of space. Home drives are for individual use only. All students and staff have a Homedrive. You can access your data from any campus computer. Outside campus you can access your home drive as well.

5.2.2... Alliance

Alliance is an online collaboration environment. It can be accessed through any web browser and provides tools such as a wiki, forums, and calendar. All ANU staff and students can log into Alliance and create a ‘project’. The creator of a project can make the project public or private and can give read or write access (called access and maintain) to individuals. It can be used to collaboratively write documents, discuss research on the forums or with the chat tool, and more.

Alliance is one of the easiest ways to share data between small groups of researchers. The data can be accessed from any internet-connected computer and researchers outside of ANU can be given a guest account. Alliance limits accepts individual file uploads up to 25Mb.

5.2.3... Webserver

The ANU IT services group manages over 800 webservers. If your school does not have a webserver for staff pages, you can request an account on the main ANU webserver. Any website that is a directory of www.anu.edu.au (such as www.anu.edu.au/polsci/ and www.anu.edu.au/music/ ) are hosted by the main ANU webserver. Personal websites are usually placed in a directory named after your University ID number – www.anu.edu.au/~u1234567/

41 https://myfiles.anu.edu.au
42 https://myfiles.anu.edu.au
43 alliance.anu.edu.au
5.3 Information Literacy Program

The Information Literacy Program (ILP) provides a variety of resources for training staff and students in Digital and Information Literacy (IT)\textsuperscript{44}. These include instructor-led training courses and online training courses\textsuperscript{45}.

Several training courses are offered that can assist in data management:

- **Alliance** – ANU’s online collaborative environment
- **LaTeX** and **BibTex** – Document writing software
- **Microsoft Office** – Word, Excel, PowerPoint/Prezi
- **EndNote** – Reference management software for Microsoft Word
- **NVivo** – Qualitative data organization tool
- **SPSS** – Statistical analysis software
- **Internet and Research** – How to use internet searches effectively
- **Search Strategies and Subject Databases/scholarly information online** – How to use online databases effectively

5.4 Open Research@ANU

Open Research@ANU, (formerly known as ANU Digital Collections) is the designated institutional repository of the University\textsuperscript{46}.

It provides long-term storage and dissemination of ANU research publications. Open Research@ANU has its own search tools (accessible via web interface) and is indexed by major search engines such as Google, thus increasing the likelihood of your research being re-used.

Open Research@ANU holds, for example, ANU Digital Theses, journal articles, conference papers and more. Contact Open Research@ANU staff at repository.admin@anu.edu.au if you would like to add your research to the online collection. Increasingly, funding agencies are requiring research results be made available in institutional repositories. For further information, contact repository staff on X59729.

5.5 ANU Data Commons

The ANU Data Commons provides a specialist data archiving service, providing long term storage of research data. Data can be uploaded in a number of ways and specialist upload facilities are available to both upload large datasets, large numbers of files, or data generated by research instruments.

Data stored within the Data Commons can be either public, private or available by request, ie people can search for and find the metadata description, but need to ask you directly for permission to access the data itself.

The Data Commons can create Digital Object Identifiers as an aid to data citation and can also publish your metadata record to Research Data Australia to aid discovery

Further information, as well as links to user documentation, is available online at http://itservices.anu.edu.au/research-computing/anu-data-commons/

\textsuperscript{44} ILP (Information Literacy Program), anulib.anu.edu.au/research-learn
\textsuperscript{45} ILP Training, training.anu.edu.au
\textsuperscript{46} Open Research@ANU, http://openresearch.anu.edu.au/
5.6 National Computational Infrastructure (NCI)

The National Computational Infrastructure (NCI\textsuperscript{47}) is Australia’s national research computing facility, providing world-class services to Australian researchers, industry and government.

NCI is home to the Southern Hemisphere’s most highly-integrated supercomputer and filesystems, Australia’s highest performance research cloud, and one of the nation’s largest data catalogues—all supported by an expert team.

The NCI provides:

- Compute facilities for intensive processing or analysis
- Mass Data Storage Facilities (MDSS) for projects with large data requirements (greater than 20Gb)
- Dataset hosting for very large datasets of national significance
- Consultancy and training to help people solve complex problems and to help them use the facilities

Use of NCI is usually free to ANU researchers upon application for an account\textsuperscript{48}. Resources are allocated through the National Computational Merit Allocation Scheme (NCMAS)\textsuperscript{49}, or the Australian National University Allocation Scheme\textsuperscript{50}. Applications are submitted online\textsuperscript{48}.

5.7 Australian Data Archive (ADA)

The Australian Data Archive (ADA) — now incorporating the Australian Social Science Data Archive — is a consortium of leading national Australian universities, managed by the Australian National University (ANU).

ADA employs a team of professional data archivists, advised by a panel of leading social scientists, provides both stewardship and outreach services to the Australian community. The archive:

- acquires, documents, preserves and disseminates data online to a broad range of social science researchers in the university, government, and other sectors
- provides the only comprehensive social science data collection in Australia, with a catalogue of over 2000 data sets
- holds data from Australian surveys, opinion polls and censuses and includes data from other countries within the Asia Pacific region
- provides specialist services within specific subject areas, including Indigenous studies, electoral behaviour, criminology and some humanities disciplines, and within specific data types, including quantitative, qualitative, time series and panel data, and historical statistics
- locates and manages access to overseas social science data sets required by Australian based researchers
- plays an important role on behalf of the Australian Research Council (ARC) through the management and dissemination of ARC funded data collections arising from Discovery and Linkage grants

\textsuperscript{47} NCI (National Computational Infrastructure), nci.org.au
\textsuperscript{48} my.nci.org.au/mancini
\textsuperscript{49} https://my.nci.org.au/anu/
\textsuperscript{50} https://ncmas.nci.org.au/2016/
5.8 Australian National Data Service (ANDS)

The ANU is a partner in the ANDS initiative, which engages in partnerships with the research institutions to enable better local data management that enables structured data collections to be created and published\(^5\). These connected data collections, together with the shared national infrastructure, form the Australian Research Data Commons.

- Data collections can be searched using the Research Data Australia (RDA) service\(^5\).
- ANU students and staff can publish their data into the RDA\(^5\) by contacting ANDS\(^5\).

\(^5\) ANDS (Australian National Data Service), www.ands.org.au
\(^5\) RDA (Research Data Australia), https://researchdata.ands.org.au
\(^5\) services@ands.org.au
6 Writing a Data Management Plan

This chapter gives an outline for a generic data management plan. Each project will be different and have different types of data, so some sections may not apply. Remember that a data management plan is a living document and should be reviewed and updated regularly, especially if unforeseen data is collected.

The recommended structure for a data management plan is as follows:

1. Project description
2. Survey of existing data
3. Data to be created
   - Data organization methods (optional)
4. Data administration issues:
   a. Funding and legislative requirements
   b. Data owners and stakeholders
   c. Access and security
   d. Backups
5. Data sharing and archiving
6. Responsibilities
7. Budget

Project description

Write a few paragraphs about the research project to give some perspective to the remainder of the plan. Use this section to introduce any terminology that will be used in the DMP.

6.1 Survey of existing data

Whilst not compulsory, it is good practice to see if there are existing data that could replace or augment the data you are planning to create. For example, in the UK it is a condition of Economic and Social Research Council (ESRC) grants that you conduct a review of the UK Data Archive to ensure that the data you are planning to create does not already exist.

- Have you searched the web and data archives for similar datasets?
- Are there any datasets that could assist with your research?
- How do the existing datasets fail to meet your requirements and therefore require new data to be created?

6.2 Data to be created

You should list all the data that will be created during the project. The remainder of the DMP then deals with how each item of data will be managed.

6.3 Data organization methods

Data organization methods are largely a matter of personal preference and will usually not be of interest to the recipients of the DMP. The exception would be if resources were required for IT infrastructure, software, or training.
6.4 Data administration issues

6.4.1 Funding and legislative requirements
List any relevant policies. Some policies (such as data archiving) are relevant to all research projects, whereas privacy will usually be associated with medical and social science projects.

- Does any of your data contain personal information that must be kept confidential?
- Does your funding agreement require data archiving?
- Are there any other data management requirements in your funding agreement?

6.4.2 Data owners and stakeholders
List the owners and stakeholders of the data. Also note who will own any intellectual property created by your research.

6.4.3 Access and security
List who will have access to the research data and what access permissions they will have for specific data. If the data will be distributed at some point, list the access restrictions and any embargoes that will be used.

Describe how the access permissions will be enforced and what IT security practices will be used. If you have sensitive data, describe any special measures used to store and backup this data.

- Is any data of a sensitive nature?
- What are the implications of unauthorized access to this data?
- Are there any safe measures warranted? (encryption, external hard-drive in locked cabinet/safe)

6.4.4 Backups
List what data will be backed up and what the backup schedule is. Also mention if any data will be kept under version control and how that will be implemented.

- Is there a backup service already available or will you need to do it yourself?
- How often will backups occur?
- Who will be responsible for performing backups?
- How will sensitive data be backed up?

6.5 Data sharing and archiving

6.5.1 File formats, standards and conventions
List what formats, standards, and conventions will apply to each data item. Justify the use of particular formats in terms of usability, longevity, suitability for archiving.

- Will other researchers be able to use this format?
- Will this format be usable in ten years time?
- Does your archive accept this file format or can you easily convert to an acceptable format?
6.5.2 Sharing
List what data will be made available for other researchers to use.
- What data will be shared?
- What facilities will be used/required to distribute the data?
- How will the data be licensed?
- What access restrictions will be placed on each item of data?

6.5.3 Archiving and disposal
Estimate the amount of storage space required for archiving, which archive you intend to use, and whether or not you have discussed your project with the archive manager. If the data is sensitive, describe how you will ensure the data will be safely disposed.
- Which archive service will be used?
- How long must you keep your data archived for?
- When do you plan to archive each item and will they have an embargo period?
- How much time and resources will be required for archiving?
- What metadata will be needed for each data type?

6.6 Responsibilities
List who will be responsible for ensuring each item in the data management plan is carried out. Also note who is responsible for reviewing and modifying the data management plan.

6.7 Budget
Now that the data management methods and responsibilities have been established, you can estimate the costs of data management for your project. Often the time involved in documenting, writing metadata, and archiving are underestimated. Make note of any costs associated with using data management services or purchasing equipment (such as fileservers, backup media, software, etc.) used for data management.
7 Glossary of Terms

**AIATSIS** Australian Institute of Aboriginal and Torres Strait Islander Studies  
**Alliance** Online collaborative environment for ANU staff and students  
**ANUSF** ANU Supercomputer Facility  
**APAC** Australian Partnership for Advanced Computing  
**ARC** Australian Research Council  
**Archive** Digital Archive  
**ASEDA** Aboriginal Studies Electronic Data Archive  
**ASSDA** Australian Social Sciences Data Archive  
**Backup** A copy of data kept for recovery in case of accidental deletion or loss of data  
**Bazaar** A distributed version control system for collaborative tracking of project history  
**BibTeX** A software tool and a file format for generating bibliographies in LaTeX  
**Binary file** A file that cannot be read without appropriate software, contrast with a text file which can be read with any text editor  
**BlueNet** Australian data archive for the marine sciences  
**CAI** Computer Assisted Interview - a computer program which helps direct an interview based on the responses given  
**CD** Compact Disc – the oldest form of optical disc, can store 700Mb  
**CiteSeer** Online scientific literature digital library  
**Client** A program used to interact with a server. For example, an FTP client is needed to download and upload files to an FTP server.  
**Creative Commons** An organization that provides generic licenses for freely distributed data  
**Data** Digital research data  
**Dataset** A collection of related data such as tables of numeric data or a group of related images  
**Data administration** Anything done to protect data or enhance the quality of data  
**Data management** Anything outside of using the data, such as organizing, protecting, sharing and archiving data  
**Data organization** Tools and techniques for working more efficiently with data  
**Data sharing** Actively making research data available for use by other researchers  
**Desktop** The most common type of PC (personal computer)  
**DocBook** An XML file format for documents  
**DVD** Digital Versatile Disc – an optical disc used for storing data, can store 4.7Gb or roughly 7 CDs  
**Encryption** Making data unreadable to anyone without the correct password or encryption key  
**EndNote** Software tool for managing bibliography databases and generating bibliographies in MS Word documents  
**Excel** Microsoft Excel, software tool for working with spreadsheets  
**External hard-drive** A device for storing data that can be easily connected to computers via a USB cable  
**Fileserver** A computer that stores data and allows authorized users to access their data. Fileservers are usually backed up every night and it is recommended to save all data on a fileserver rather than a computer’s hard disk  
**FTP** File Transfer Protocol. Simple method for transferring files over a network (such as the internet). An FTP server accepts connections from a software tool called an FTP client and allows data to be downloaded and uploaded.  
**Google Scholar** A popular search tool for finding citations and freely downloadable academic publications
Managing Digital Research Data

- **Homendrive** The main fileserver of the ANU
- **IMAP** Internet Message Access Protocol. Method of checking email from several different computers.
- **LaTeX** A markup language for writing scientific and technical publications. Widely used in academia.
- **LITSS** Local IT Support Services. Departmental IT Staff that maintain staff computers and IT infrastructure.
- **Lossless compression** Method of compressing a file without losing quality, such as zip, gif, png
- **Lossy compression** Method of compressing a file that results in a loss of quality, such as jpeg, mp3
- **Mac** Apple Macintosh computer
- **Markup language** A plain text file that can be processed into graphical format, such as LaTeX, html, docbook.
- **Metadata** A small data file that describes attributes of a data item, such as size, date, format.
- **Mounted drive** Remote data that appears as a directory on a computer
- **MS** Microsoft
- **NESSTAR** Software tool for statistical analysis of datasets
- **Network share** A device that can be accessed over a network. Similar to mounted drive but also includes printers.
- **Open Access** Providing free, electronic copies of academic articles online
- **Open Content** Data that is free for anyone to download and use
- **OpenOffice** A free software tool that is mostly compatible with Microsoft Office
- **Optical disc** A data storage medium such as a CD, DVD, and BluRay
- **PDF** Portable Document Format – common format for distributing documents
- **Server** See also: Client
- **Text file** See also: Binary file. A file that can be read with any text editor.
- **Ubuntu** An open-source Linux-based operating system for computers
- **Unix** An operating system used by many servers and high performance computers
- **USB** Universal serial bus – method for easily connecting devices (external hard-drives, digital cameras, printers, mice, keyboards) to a computer
- **USB key** A small device for storing data. Data is transferred to the device via USB and it can then be removed and connected to another computer.
- **Version Control Software** Software tool for methodically tracking changes to a file or files
- **Web Application** A software tool that can be used with any internet browser
- **Web Browser** A software tool for accessing web pages
- **Webserver** A server that stores web pages and other data that people connect to with a web browser
- **Wiki** A website that can be modified by anyone, such as Wikipedia. Some wikis only allow modification by registered users.
- **Windows** Microsoft Windows – the operating system used by most computers
- **Word** Microsoft Word – a software tool for writing documents (word processor)
- **XML** Extensible Markup Language – general purpose markup language that can be used for documents, metadata, databases
- **Zotero** Open-source Firefox extension to collect, cite and share research sources. Can be used with both LaTeX and Word/OpenOffice. Supports BibTex and EndNote databases.
8 References and Links

Alliance. ANU, alliance.anu.edu.au
ANDS (Australian National Data Service). www.ands.org.au
ANU IT Security, security.anu.edu.au
RCS (Australian Research Collaboration Service). www.arcs.org.au
Bazaar. bazaar.canonical.com/en/
Creative Commons. creativecommons.org
Data Validation. Wikipedia, en.wikipedia.org/wiki/Data_validation
Dropbox. www.dropbox.com
Endnote. www.endnote.com
Find LITSS. ANU, its.anu.edu.au/litss/contacts/
Firefox Web Browser. www.mozilla.com
FTP Explorer. www.ftpx.com/
GNU Free Documentation License. www.gnu.org/copyleft/fdl.html
ILP (Information Literacy Program). ANU, information.anu.edu.au/training
ILP Training. ANU, training.anu.edu.au
Information Services website, information.anu.edu.au
JabRef. jabref.sf.net
LITSS Information. ANU, its.anu.edu.au/litss/support-model.html
Metadata. Wikipedia, en.wikipedia.org/wiki/Metadta
NCI (National Computational Infrastructure). nci.org.au/
NCI National Facility. ANU, nf.nci.org.au/
Open Research@ANU. https://openresearch.anu.edu.au/
RDA (Research Data Australia). Australian National Data Service, services.ands.org.au/home/orca/rda/
Revision Control. Wikipedia, en.wikipedia.org/wiki/Revision_control
Sakai. sakaiproject.org/
Science Commons. sciencecommons.org
Software Information. ANU, information.anu.edu.au/daisy/infoservices/8/29.html
Sophos Anti-Virus Software. security.anu.edu.au/sophosinfo
TortoiseSVN. tortoisesvn.net/
Vislab. ANU Supercompting Facility, anusf.anu.edu.au/Vizlab/
WinSCP. Wikipedia, en.wikipedia.org/wiki/WinSCP
Zotero. www.zotero.org
Managing Digital Research Data

Other resources

Training notes
To access the Information Literacy Program’s training notes, visit the Research & learn webpage (anulib.anu.edu.au/research-learn) and select the skill area followed by the relevant course. You can register for a workshop and find other information.

Research & learn how-to guides
Explore and learn with the ANU Library’s how to guides (ql.anu.edu.au/howto). Topics covered are:

- Citations & abstracts
- Data Management
- EndNote
- Finding theses
- Finding books and more
- Finding journal articles and more
- Increasing your research impact
- NVivo
- Topic analysis
- Using Google scholar from off-campus

Subject guides
Find subject-specific guides (ql.anu.edu.au/subjectguides) and resources on broad range of disciplines. Such as:

- History, indigenous studies, linguistics and philosophy
- Criminal, human rights and taxation law
- Biochemistry and molecular biology, neurosciences and psychology
- Asia Pacific, Southeast Asia and East Asian studies
- Engineering
- Astronomy and astrophysics, earth sciences, mathematical sciences and natural hazards

Online learning
Online learning is available through ANU Pulse (ql.anu.edu.au/pulse), which can be accessed from both on and off campus by all ANU staff and students.

IT skills development modules available in ANU Pulse

- Microsoft Office (Access, Excel, OneNote, Outlook, PowerPoint, Project, Visio, Word)
- Microsoft Office (Mac)
- Adobe suite (Illustrator, Photoshop)
- Other IT (Concepts of IT, FrontPage, Internet Explorer, Type IT, Windows)

Training calendar
Select Events » near the bottom of the Library homepage to access our events calendar with upcoming training opportunities displayed day by day (ql.anu.edu.au/cal).

Feedback!

Please provide feedback about today’s workshop via an online feedback form (ql.anu.edu.au/survey)