# ANU Data Management Manual:

## Managing Digital Research Data at the Australian National University

Information Literacy Program
The Australian National University

Document Version – 1.01

August 5, 2008

This Manual was developed with the funding support of the Australian Partnership for Sustainable Repositories (APSR). APSR is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's *Backing Australia's Ability - An Innovative Action Plan for the Future*.

# Contents

# Chapter 1

# Introduction

Data Management is a necessary part of modern research. Almost all researchers will have digital data, whether it be measurements from instruments, survey records, multimedia, or documentation. Data management involves things 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.
- Raise awareness of the data management services at ANU.
- Be able to write a data management plan.

## 1.2 Data Management at ANU

| | |
|---|---|
| ILP DM Training | http://ilp.anu.edu.au/dm/ |

This website provides data management plan templates, links to websites for more in depth information, powerpoint presentations, workshop schedules, and the latest version of this manual.

## 1.3 Outline

This document is intended to be read in order. However, if you are only interested in Writing a Data Management Plan (see Chapter 6) or 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.

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

### Chapter 2 – Data Management

- All researchers have digital data. At the least they will have their publications, but may also have measurements, survey responses, multimedia, etc.
- Data Management can be loosely defined as "Anything outside of actually using the data". For example, organisation, 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.

### Chapter 3 – Benefits & Requirements

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

### Chapter 4 – Methods of Data Management

- Data Organisation: 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 & Sharing: Details of Data Archiving for preservation, and Data Sharing for exposure and open research.

### Chapter 5 – ANU's Data Management Services

- LITSS – provide your computer and software. May also provide a fileserver for backups and a webserver.
- Systems & Desktop Services – Manage ANU's central file server (Pebble) and webserver. Also manage Alliance, which is an online collaborative environment.
- Information Literacy Program – provide training in using software and general IT skills.
- Demetrius – ANU's Institutional Repository for long-term storage and dissemination of data.

- ANU Supercomputing Facility – High performance computing, visualisation, and large data storage.
- Discipline specific archives – ASSDA (Social Sciences), BlueNet (Marine Sciences), ASEDA (Indigenous Language Studies).

## Chapter 6 – Writing a Data Management Plan

- Recommended structure of a DMP

# Chapter 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.

## 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, analogue recordings, and non-digital images as they can be converted to digital images. Physical data such as biological specimens, soil samples, etcetera are not considered.

Some examples of digital research data 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: Sound data, interviews, wildlife recordings, language recordings.
- Digital Video: High-speed recordings, interviews.

## 2.2 Data Management

A very loose definition of data management is:

> Data Management is anything outside of actually using the data.

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

- Organising data into directories/folders and using meaningful filenames.
- Keeping backups of data in case you accidently 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.
- Synchronising data between desktop, laptop, USB key, etc.
- Maintaining a bibliography and electronic copies of relevant literature.

Data Management involves organising, 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 as 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 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.

# Chapter 3

# Benefits & 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.

- Section 3.1 covers the benefits of data management.
- Section 3.2 covers the benefits of data archiving and sharing.
- Section 3.3 covers funding and legislative requirements of data and data management, including:
  - ANU Responsible Practice of Research.
  - Australian Code for the Responsible Conduct of Research.
  - ARC Funding Agreement for Discovery Projects.

## 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 organisation, 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 collaboration with others and dissemination of results and publications.

### 3.1.1 Efficiency

Data management can improve the efficiency with which you work with your data. Typically organisation and documentation of data are only done when they are absolutely necesary. 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 recognise the risk associated with losing data through accidental deletion and equipment failure, theft or destruction. Multiple and backup copies are therfore 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. Innacurate data can invalidate results and conclusions resulting in lost time and damaging reputations.

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[13].

## 3.2 Benefits of Data Archiving & 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* [9]:

- 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 other's 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.
- Improves 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 & Legislative Requirements

There are some key funding and legislative requirements relating to data management. The current requirements are considered inadequate and are often unknown and not enforced. In the near future it is likely that grant applications will require a Data Management Plan (see Section 2.3 and Chapter 6) and that archiving of research data will be enforced.

The following sections summarise the policies relating to data management. The exact wording of the policies is given in Appendix A.

### 3.3.1 ANU Responsible Practice of Research Policy

See Appendix A.1.

- Data management should comply with the *Commonwealth Privacy Act 1988* [15].

- 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)* [14].

- Researchers should not unnecessarily enter into research agreements that limit access to information.

- Researchers are responsible for data security.

- Supervisers must ensure the validity of data gathered by their students.

- Research misconduct inlcudes: fabrication, falsification, and interference with data.

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

See Appendix A.2

- 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

See Appendix A.3

- Data from research in the social sciences should be archived with the Australian Social Sciences Data Archive (ASSDA, see Section 5.6) within 2 years.
- Data should be submitted to an appropriate subject and/or institutional repository within 6 months of project completion.

# Chapter 4

# Methods of Data Management

This chapter outlines some of the methods of data management.

- Section 4.1 presents methods for working more efficiently with data.
- Section 4.2 describes policies and practices that can be used to administer data.
- Section 4.3 describes data archiving and sharing.

At a minimum you should be doing backups, security, and archiving.

## 4.1 Data Organisation

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

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

### 4.1.1 Bibliography Management

| | |
|---|---|
| Reference Management Software | Wikipedia Entry[1] |
| Endnote | http://www.endnote.com |
| JabRef | http://jabref.sf.net |

[1] http://en.wikipedia.org/wiki/Reference_management_software

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 Reference Management tools that automate citations and bibliography creation when writing an article. They also organise 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 IEEEXplore, CiteSeer, ArXiv, and PubMed.

EndNote is the most popular Reference Management tool and ANU has an institutional licence which allows staff and students to install EndNote on their office and home computers, including laptops[2]. The Information Literacy Program also runs courses in EndNote.

EndNote does not run on Unix and cannot manage BibTeX bibliographies, so LaTeX authors and Unix users can use JabRef, which is a free program, runs on all operating systems, and can import and export BibTeX's and EndNote's database formats.

### 4.1.2 File Transfers & Remote Access

| | |
|---|---|
| FTP | Wikipedia Entry[3] |
| Mounting Pebble | Staff iGuide[4] |
| Connecting to Pebble via FTP | Staff iGuide[5] |

It is often necessary to transfer data between computers. Collaborating researchers will share primary data and preliminary results. 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 file that can be transferred. Removable media, such as USB keys and CDs or DVDs, can transfer large amounts of data, but require the researcher to physically carry the data to its destination.

Large files are usually transferred using FTP (File Transfer Protocol). FTP allows the user to download as well as upload, and access to files can be restricted by username and password. An *FTP Client* (such as FTP Explorer) is used to connect and transfer files, although most web browsers can access FTP servers by entering the URL in the location bar with `http` replaced by `ftp`. The ANU has FTP access to the Pebble server (see Section Section 5.2.1), which allows off-campus access to data.

All of the above options create multiple copies of the data, however. The best solution is to keep the data in one place, such as a fileserver, and edit the data in-place. Editing data in-place is usually achieved with a mounted drive, but can also be done with remote login or a web application. A mounted drive is

[2] http://ilp.anu.edu.au/endnote/how.html
[3] http://en.wikipedia.org/wiki/File_Transfer_Protocol
[4] http://staff-iguide.anu.edu.au/ConnectPebble.html
[5] http://staff-iguide.anu.edu.au/FTPdetails.html

the best option as the remote data appears as a directory on the users computer and any changes will be saved on the remote computer, thus avoiding managing multiple copies. For security reasons, it is usually only possible to mount a drive from within the university.

Web applications, such as Alliance (see Section 5.2.2), 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 Synchronisation

| | |
|---|---|
| File Synchronisation | Wikipedia Entry[6] |
| WinSCP | http://winscp.net/, Wikipedia Link[7] |

Often researchers will work on their university desktop as well as a laptop, and possibly 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 overwrite some changes without knowing.

If you are synchronising regularly or have lots of files to synchronise, then you should consider using File Synchronisation software. File Synchronisation software offers the following advantages over manual synchronisation:

- 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 synchronisation programs is WinSCP, which is primarily for SSH and FTP transfers, but can also synchronise data[8]. Version Control software (see Section 4.1.5) is another option for file synchronisation.

[6] http://en.wikipedia.org/wiki/File_synchronization
[7] http://en.wikipedia.org/wiki/WinSCP
[8] http://winscp.net/eng/docs/task_synchronize

### 4.1.4 Collaboration

| Collaborative Writing | Wikipedia Entry[9] |
|---|---|

A lot of research is carried out collaboratively: between postgraduates and their supervisors; within departmental research groups; as cross-discipline research, and as inter-univeristy research. This is beneficial as it: improves access to funding; avoids repeating costly experiments; increases recognition through co-authorship; and can help lead to new research areas.

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 are only a small number of collaborators. It is worth considering using collaborative software tools such as Alliance (see Section 5.2.2) or Version Control software (see Section 4.1.5). Version Control software is harder to set up, but provides more advanced version tracking. Alliance is an ANU web-based tool which allows ANU staff and students to easily set up collaborative project sites. Alliance provide a wide range of collaborative tools such as forums, chat rooms, calendars, and more.

Such tools 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 since they last looked at the data.

### 4.1.5 Version Control

| Revision Control | Wikipedia Entry[10] |
|---|---|
| TortoiseSVN | `http://tortoisesvn.net/` |

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 version control software. These programs are used extensively for software development but are also excellent for documentation,

[9] `http://en.wikipedia.org/wiki/Collaborative_writing`
[10] `http://en.wikipedia.org/wiki/Revision_control`

such as writing a paper with several authors. Version control software also provides access control, a collaborative work environment, synchronisation 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 synchronisation and is good for resolving conflicting changes.

The drawback is the time required to learn the software. It is therefore only recommended for people that regularly encounter problems with simple filename version control.

TortoiseSVN is a popular program that uses the Subversion system of version control. It integrates with Windows Explorer making it one of the easiest version control programs to use.

## 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

| | |
|---|---|
| Backup | Wikipedia Entry[11] |
| Pebble restore | http://pebble.anu.edu.au/ |

Keeping backups is probably the most important and, fortunately, one of the most widely practised data management methods. Most people are quite aware of the risk and cost of losing data through hard drive failure or accidental deletion. It is therefore best to have a policy for maintaining backups.

When considering your backup strategy, you need to know:

- How often will you make backups.
- How long will backups be stored.
- How much hard-drive space or number of DVDs will be required to maintain this backup schedule.
- If the data is sensitive, how will it be secured and (possibly) destroyed.

[11] http://en.wikipedia.org/wiki/Backup

- 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 phyiscally destroyed and optical discs to be shredded.

The lifetime of backups should also be considered. Burned optical discs have an average lifetime of 2 years, and 5 years if kept in a cool dark place.

If you are using a network drive then your data is probably already being backed 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, how long do 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.)

### 4.2.2 Validation & Authentication

| | |
|---|---|
| Computer Assisted Interviewing | Wikipedia Link[12] |
| Data Validation | Wikipedia Link[13] |

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 fo your data. This can include:

- Callibration of instruments.
- Use of Computer Assisted Interviews (CAI).
- Securing master copies to avoid accidental/intentional tampering.
- Data entry checks, such as: two pass verification and range checking.

[12] http://en.wikipedia.org/wiki/Computer-assisted_telephone_interviewing
[13] http://en.wikipedia.org/wiki/Data_validation

### 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 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[14]. This is important as notebooks are easily lost or put into storage when an academic or postgrad 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 analyse the dataset by producing graphs or statistics from your dataset.

### 4.2.4 Access Controls

| Windows File Permissions | Microsoft Link[15] |
|---|---|

Well-defined access controls help you comply with privacy & 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 usually 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 *Administrator* permissions over all data and may be the only one with *Read* permissions of confidential survey data. Research collaborators would have no access to the confidential survey data, *Read* access to de-identified survey data, and *Write* access to data analyses and publications.

Access Permissions are usually set by right-clicking on a file or directory and editing the security properties.

[14] You can scan a notebook and use software to extract the writing, but it may not work well unless you have very good handwriting.

[15] http://support.microsoft.com/kb/304040

### 4.2.5 Security

| ANU IT Security | http://security.anu.edu.au |
|---|---|

It is important to consider the security of your 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. IT Security is too large to cover here, but at the least you should install up to date anti-virus software on your computer. ANU staff and students can install Sophos Anti-Virus on their office and home computers[16].

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 for you[17], so it will not be covered here.

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 over night.

## 4.3 Data Archiving & 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 re-use and re-purposing. 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 3.2.

### 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 3 ways:

- Email request – Interested researchers email and request the dataset. This is the most common way that data is shared.

[16] http://security.anu.edu.au/sophosinfo/

[17] Your Local IT Support Staff (LITSS) may be able to help you.

- 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 serve the dual purpose of data preservation and dissemination. Their archives usually 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.

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

### 4.3.2 Copyright & Licencing

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 copyright on academic publications, however, are owned by the researcher.

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

Licences grant permission for others to use the copyrighted data. Open Content Licences are an easy way for researchers to licence their data for others to use. A researcher can choose the most suitable licence for their needs rather than develop a custom licence themselves. The most notable open content licences are

- Creative Commons[18] – most popular open content licences
- Science Commons[19] – similar to Creative Commons but tailored for scientific data and publications.
- GNU Free Documentation Licence[20] – used by Wikipedia

The ANU's institutional repository, Demetrius (see Section 5.4.1), has a copyright licence that can be used by depositors to give the archive permission to store and maintain the data, whilst leaving ownership of the data with the researcher. See Appendix B for more on licences.

[18] http://creativecommons.org/
[19] http://sciencecommons.org/
[20] http://www.gnu.org/copyleft/fdl.html

### 4.3.3 File formats & Standards

| | |
|---|---|
| File Format | Wikipedia Entry[21] |
| Digital Preservation | Wikipedia Entry[22] |
| Open File Formats | Wikipedia Entry[23] |

Before creating the data you should consider what formats and standards you should use as it is sometimes difficult to convert between file formats. Using an innapropriate file format will also make your life more difficult in the long run.

Where possible, it is best to use open formats as they are more likely to be readable in the future and are easier to share with others. It is usually safe to use a proprietry format if it is very widespread as free programs will most likely exist to read these formats. For example, almost all Microsoft Office documents can be read with Open Office.

Some examples of open formats are:

- PDF – document format.
- OpenDocument Format (ODF) – used by OpenOffice, similar to MS Word.
- PNG, TIFF, JPEG – Image formats.

Your LITSS (see Section 5.1) and Demetrius staff (see Section 5.4.1) 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.

### 4.3.4 Access Restrictions

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)

[21] http://en.wikipedia.org/wiki/File_format
[22] http://en.wikipedia.org/wiki/Digital_preservation
[23] http://en.wikipedia.org/wiki/Open_format

### 4.3.5 Metadata

| Metadata | Wikipedia Entry[24] |
|---|---|

Metadata is often described as "data about data". It is usually a file with several text fields that describe the attributes of another piece of data, such as an experimental dataset, image, or video. The 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 analagous 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.

### 4.3.6 Archiving

Archiving of final state 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. 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

[24] http://en.wikipedia.org/wiki/Metadata

archives are able to perform the conversion but it is best if the depositor does the conversion to ensure 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.

# Chapter 5

# ANU's Data Management Services

The majority of the data management services at ANU are provided by Systems and Desktop Services (SDS) and your Local IT Support Staff (LITSS). They provide the day-to-day data management services like back-ups and shared drives. Data archiving is more specialised and is performed by: Demetrius, the ANU Supercomputing Facility (ANUSF), and the Australian Social Sciences Data Archive (ASSDA). There is also BlueNet and the Aboriginal Studies Electronic Data Archive (ASEDA), which are not ANU services but are available to ANU staff and postgraduate students.

The Staff iGuide (http://staff-iguide.anu.edu.au) contains extensive information on the ANU's information services.

## 5.1 Local IT Support Staff (LITSS)

| | |
|---|---|
| Information | LITSS Support Model[1] |
| Find LITSS | Contacts[2] |
| Software for ANU Staff | Software[3] |

The majority of your data management needs will be provided by your Local IT Support Staff. The services that they provide are usually determined by the head of college, so the services will vary between colleges and even departments. Most colleges will provide a file server and web hosting.

A file server is a computer that stores data and makes it accesible 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 back-ups yourself.

[1] http://its.anu.edu.au/litss/support-model.html
[2] http://its.anu.edu.au/litss/contacts/
[3] http://information.anu.edu.au/daisy/infoservices/8/29.html

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 webservers 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 and link to it from your website.

Your LITSS may also provide a mounted drive for collaborative work. Normally a fileserver 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 suites, and conversion tools.

## 5.2 Systems and Desktop Services (SDS)

| | |
|---|---|
| Pebble | Pebble – Staff iGuide[4] |
| Alliance | http://alliance.anu.edu.au |
| Web hosting | Publishing to www.anu.edu.au[5] |

Systems and desktop services are responsible for the undergraduate computer laboratories, the major mail servers, and more. In regards to data management, they provide the *Pebble* fileserver, Alliance, and webservers.

### 5.2.1 Pebble

Pebble is the main university wide fileserver. All research staff and students have an account on Pebble, research staff and research students are allocated 250Mb of space. If your department does not have its own fileserver then you can use Pebble as your mounted drive.

Pebble is backed up nightly and the last 10 days are stored online. A user can request a backup of their Pebble account from the last 10 days. Older backups are stored on tape for roughly 2 years, but take longer to restore.

### 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

[4] http://staff-iguide.anu.edu.au/N10750.html
[5] http://information.anu.edu.au/daisy/infoservices/23/345/1232.html

(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 a small group of researchers. The data can be accessed from any internet connected computer and researchers outside of ANU can be given a guest account. Alliance has a quota of 25Mb, but this can be increased upon request.

### 5.2.3 Webserver

SDS manages over 800 webservers. If your department does not have a webserver for staff pages, you can request an account on the main ANU webserver. Any website that is a subdirectory 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/`.

## 5.3 Information Literacy Program

| | |
|---|---|
| ILP | http://ilp.anu.edu.au |
| Training Registration | http://training.anu.edu.au |
| Email | ilp@anu.edu.au |

The Information Literacy Program (ILP) provide a variety of resources for training staff and students in Information Technology (IT). These include instructor led training courses and online training courses. Several training courses are offered that can assist in data management:

- Alliance – ANU's online collaborative environment.
- LaTeX & BibTeX – Document writing software.
- Microsoft Office – Word, Access, Excel, PowerPoint.
- EndNote – Reference management software for MS Word.
- NVivo – Qualitative data organisation tool.
- Photoshop – Image editing and conversion software.
- Dreamweaver – Program for easily creating websites.
- SPSS – Statistical analysis software.
- E-publishing theses – How to publish a thesis to the ADT[6]
- WebWise – How to use internet search engines effectively.

[6] Australian Digital Theses

## 5.4 Digital Resource Services (DRS)

| | |
|---|---|
| Demetrius | http://dspace.anu.edu.au |
| User guide | http://sts.anu.edu.au/demetrius |
| Email | demetrius@anu.edu.au |
| Address | Hancock Building, Bldg 122[7] |

Digital Resource Services carry out projects related to the acquisition, storage, retrieval and dissemination of data. They created and manage the ANU's institutional repository – Demetrius.

### 5.4.1 Demetrius

Demetrius is the ANU's institutional repository. It provides long term storage and dissemination of research data. The data is stored securely and in formats that should not become obselete and unusable in the future. Demetrius 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 data being re-used.

Demetrius holds ANU Digital Theses, E-Press, ePrints

Demetrius is intended for final state data, you should not use it to store draft or incomplete work. It is not possible to modify data in the archive. If there is an error that must be corrected, then you must get the Demtrius staff to upload a corrected version.

The depositor can set access permission on the data they upload. The data can available to the public, require permission from the depositor, or not available at all. If the data is unavailable for download, the metadata will still be viewable.

To add data to Demetrius you must contact DRS and set up an account. You can then upload data using a simple web-based interface. All data uploaded must contain adequate metadata. If your data is unorganised and without metadata, then archiving can be time consuming. You should therefore write the metadata as the data is created, rather than leaving until just before archiving. You will also need to discuss with DRS what formats are suitable for archiving and convert files when necessary.

## 5.5 ANU Supercomputing Facility (ANUSF)

| | |
|---|---|
| ANUSF | http://anusf.anu.edu.au |
| APAC | http://www.apac.edu.au |
| APAC National Facility | http://nf.apac.edu.au |
| Email | help@anusf.anu.edu.au |
| Address | Leonard Huxley Building, Bldg 56[8] |

The ANUSF operates the Australian Partnership for Advanced Computing (APAC) National Facility. The APACNF is the fastest supercomputer in Aus-

[7] http://campusmap.anu.edu.au/displaybldg.asp?no=122
[8] http://campusmap.anu.edu.au/displaybldg.asp?no=56

tralia and was ranked 200th in the November 2007 TOP500. At the time of its last upgrade, June 2005, it was ranked 26th in the world[9]

The ANUSF provides:

- *Compute facilities* for intensive processing or analysis.
- *Mass Data Storage System* (MDSS) for projects with large data requirements (greater than 20Gb).
- *Dataset Hosting* for very large datasets of national significance.
- *Vizlab* for visualisation of complex data, such as 3D animations.
- *Consultancy & Training* to help people solve complex problems and to help them use the facilities.

Use of the APACNF is usually free to ANU researchers. Users are allocated a certain number of Service Units[10] (SU) and storage space. Resources are allocated every 6 months through the Merit Allocation Scheme (MAS), but small allocations (a few hundred SUs and several gigabytes) are possible at any time.

All applications are based on merit using the following criteria[11]:

- Research quality.
- Appropriateness to the National Facility.
- Reasonableness of the level of resources requested.
- Previous use of the National Facility.

Applications are submitted online[12]. You can view the forms without submitting if you would like to know what information you need.

## 5.6 Australian Social Science Data Archive (ASSDA)

| | |
|---|---|
| ASSDA | http://assda.anu.edu.au/ |
| Deposit Forms | Deposit Forms Webpage[13] |
| Email | assda@anu.edu.au |
| Address | 18 Balmain Lane, Bldg 66[14] |

ASSDA collects and preserves data relating to social, political and economic affairs and makes them available for further analysis. Data is obtained mostly from universities but also government organisations and market research companies. ASSDA is an Australia wide instituion with nodes in the University of Queensland, University of NSW, and more. ASSDA's central office is located at ANU and its archives are hosted by the ANUSF.

[9] http://www.top500.org/site/history/147

[10] A service unit is one hour of elapsed time on one processor, therefore 10 SUs could be 10 hours on 1 processor, or 1 hour on 10 processors.

[11] See http://nf.apac.edu.au/policies/MAC_criteria.php for more details

[12] http://nf.apac.edu.au/accounts/

[13] http://assda.anu.edu.au/deposit.html#deposit_forms

[14] http://campusmap.anu.edu.au/displaybldg.asp?no=66

ASSDA specialises in storing datasets from surveys and opinion polls. The datasets can be downloaded in a range of formats (SPSS, SAS, Stata, Statistica, excel) or processed online using the NESSTAR web application. NESSTAR WebView allows users to perform standard statistical analyses such as cross-tabulation, correlation & regression, recode variables, and produce graphs. As NESSTAR is running on the APACNF, users can manipulate large datasets that their computers could not handle.

ASSDA will accept data from almost any source in Australia, and sometimes from the Asia-Pacific if there are no local archives. To archive data with ASSDA, you need to fill out the depositor's forms and send the data to ASSDA. You should make sure to de-identify survey records and to include documentation on the data such as the survey questionairre. The forms help ASSDA to create metadata for your dataset and allow you to specify access controls on the data. The depositor can specify a dataset to have:

- Unrestricted Access – anyone can use.
- Notification of access – the depositor will be informed who their data is given to.
- Depositor Permission – use of the dataset requires the permission of the original depositor.

It is also possible to place an embargo on a dataset such that the creator of the dataset has exclusive use of the data for a limited time to allow them to publish.

The ARC Funding Agreement for Discovery Projects stipulates that social science datasets should be lodged with ASSDA within 2 years of completion of the fieldwork (see Section A.3 for more details).

Anyone wishing to access ASSDA's unrestricted datasets must register online. Registration is free and allows you to browse and perform online analyses of unrestricted datasets using NESSTAR WebView. To download datasets you must complete a *user undertaking form*. If you are affiliated with an ACSPRI Member Institution[15] then you can download datasets for free, otherwise you can be charged $1,000 per dataset. All ANU staff and postgraduate students have free access to ASSDA.

## 5.7 BlueNet

| | |
|---|---|
| BlueNet | http://www.bluenet.org.au |
| BlueNet Archive | http://bluenet.its.utas.edu.au/ |
| BlueNet @ ANU | http://ems.anu.edu.au/bluenet/ |
| Email | info@bluenet.org.au |
| Address | University of Tasmania, Private Bag 21, Hobart, Tasmania 7001 |

BlueNet is an Australia wide data network for the marine sciences. It provides:

- Data archiving.
- Data sharing.

[15] http://www.acspri.org.au/?page=members_listing

- Data manipulation through online tools.
- RSS feeds of newly added data.
- Interactive mapping

BlueNet is an online marine database which enables the discovery, access and online integration of multi-disciplinary marine science data on a very large scale. It is an extension of the nation's first on-line marine network - the Australian Ocean Data Network, which draws together research and data collected through government agencies and higher education institutions. The intention is to enable universities to access, utilise and add to data repositories creating a virtual national ocean data centre.

BlueNet uses the GeoNetwork[16] software for its Metadata Entry and Search Tool (MEST). As the name suggests, the MEST is used for entering metadata and also for searching the archive. An Oceans Portal is curently under development which will streamline this process.

Depositing data with BlueNet is currently performed through BlueNet Data Facilitators. To locate the nearest facilitator, users are requested to contact the BlueNet central office – info@bluenet.org.au. You can also contact Steve Eggins (steve.eggins@anu.edu.au) from the Research School of Earth Sciences (RSES) is also available to help ANU researchers with BlueNet.

## 5.8 Aboriginal Studies Electronic Data Archive (ASEDA)

| | |
|---|---|
| ASEDA | http://www1.aiatsis.gov.au/ASEDA/ |
| Forms | Forms Webpage[17] |
| Email | aseda@aiatsis.gov.au |
| Address | ASEDA, AIATSIS, GPO Box 553<br>Canberra, ACT 2601 |

ASEDA is the digital archive of the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS). ASEDA has data on Australian Indigenous languages – mostly dictionaries and grammars. The archive is for final state data and is intended for researchers wishing to share their work with others.

Some data is available for anyone to browse online, but most data requires you to submit a request via email to ASEDA and complete an access form. You must also state your affiliation and reasons for using the data.

AIATSIS is currently reorganising ASEDA and will be introducing a new policy in 2009. Any queries about depositing or obtaining data from ASEDA should be directed to: aseda@aiatsis.gov.au.

## 5.9 Summary

Backups, shared drives and webhosting will usually be provided by your LITSS, otherwise you can use the university wide services provided by SDS.

[16] http://geonetwork-opensource.org
[17] http://www1.aiatsis.gov.au/ASEDA/forms.html

Final state research data (including publications, datasets, multimedia, etc.) should be archived with Demetrius. If you have very large or complex data then you may need ANUSF to host your data. The ANUSF is also able to provide high performance computing, visualisation and consultancy.

You should also archive your data with the discipline specific archives:

- ASSDA – Quantitative social sciences.
- BlueNet – Marine sciences.
- ASEDA – Australian Indigenous language studies.

# Chapter 6

# Writing a Data Management Plan

This section 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 unforseen data is collected.

The recommended structure for a DMP is as follows:

1. Project description.
2. Survey of existing data.
3. Data to be created.
4. Data Organisation Methods (optional).
5. Data Administration Issues:
   (a) Funding & Legislative Requirements.
   (b) Data owners & Stakeholders.
   (c) Access & Security.
   (d) Backups.
6. Data Sharing & Archiving.
7. Responsibilities.
8. Budget.

## 6.1 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.

## 6.2 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. It is a condition of 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?

## 6.3 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.

## 6.4 Data Organisation Methods

Data Organisation 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.

## 6.5 Data Administration Issues

### 6.5.1 Funding & 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?

### 6.5.2 Data Owners & Stakeholders

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

### 6.5.3 Access & 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 unauthorised access to this data?
- Are any special measures warranted? (encryption, external hard-drive in locked cabinet/safe)

### 6.5.4 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?

## 6.6 Data Sharing & Archiving

### 6.6.1 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 10 years time?
- Does your archive accept this file format or can you easily convert to an acceptable format?

### 6.6.2 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 licenced?
- What Access Restrictions will be placed on each item of data?

### 6.6.3 Archiving & Disposal

Estimate the amount of storage space required for archiving, which archive you intend to use, and the 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 archiving 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.

## 6.7 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.

## 6.8 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.

# Glossary

**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

**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 organisation that provides generic licences 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 organising, protecting, sharing and archiving data.

**Data Organisation** Tools and techniques for working more efficiently with data.

**Data Sharing** Actively making research data available for use by other researchers.

**Demetrius** The institutional repository (archive) of the ANU

**Desktop** The most common type of PC (personal computer). Contrast with a laptop

**DocBook** An XML file format for documents

**DSpace** The software used by Demetrius

**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). Used for security.

**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 authorised 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.

**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 Blu-Ray

**PDF** Portable Document Format. Common format for distributing documents.

**Pebble** The main fileserver of the ANU.

**Server** See also: Client

**Text file** See also: Binary file. A file that can be read with any text editor.

**Unix** An operating system used by many servers and high performance computers. MAC OS X and Linux are derived from Unix.

**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 conneected 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 web site 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.

# Appendix A

# Data Management Policies

## A.1 ANU Responsible Practice of Research Policy

The *Archives Act* refers to the *Archives Act (1983)*[14]
http://info.anu.edu.au/Policies/_DRO/Policies/Responsible_Research_Practice.asp?tab=1
Section 4

> 4.1. Data management should comply with relevant privacy protocols. As the University is constituted under federal legislation it must conform to the *Commonwealth Privacy Act 1988 . In particular, researchers need to comply with the 11 Information Privacy Principles of the Act . Also relevant is the University's policy Privacy: Statement on the Collection, Use and Control of Personal Information (834/1994). Researchers must not use confidential information for their own personal advantage or that of a third party. Confidentiality may also be necessary for a limited period in the case of contracted research or of non-contractual research, which is under consideration for patent protection.*
>
> 4.2. Other than privacy or contractual requirements for confidential data, research results and methods in general should be open to scrutiny by colleagues within the institution and, through appropriate publication, by the profession at large. Non-confidential data related to publications must be available for discussion with other researchers.
>
> 4.3. Data (including electronic data) must be recorded in a durable and appropriately referenced form.
>
> 4.4. Data must be held for sufficient time to allow reference. For data that are published this may be for as long as interest and discussion persist following publication. The minimum period for retention shall be at least 5 years from the date of publication but for specific types of research, such as clinical research, 15 years or longer may be more appropriate. The retention period for data must comply

with legislative obligations for University information such as the *Archives Act*.

4.5. Wherever possible, original data must be retained in the department or research unit in which they were generated. Individual researchers should be able to hold copies of the data for their own use. Retention solely by the individual researcher provides little protection to the researcher or the University in the event of an allegation of falsification of data.

4.6. Confidentiality agreements to protect intellectual property rights may be agreed between the University, the researcher and a sponsor of the research. Where such agreements limit free publication and discussion, limitations and restrictions must be explicitly agreed (see the ANU policy, *Intellectual Property: Ownership, Protection and Commercialisation* ). *In general researchers should not unnecessarily enter agreements, which limit or prevent access to information.*

4.7. It is the obligation of the researcher to enquire whether *formal confidentiality agreements apply and of the Head of the Department or research unit to inform researchers of their obligations with respect to these provisions.*

4.8. All formal confidentiality agreements should be made known at an early stage to the Deputy Vice-Chancellor (Research) through the appropriate Dean or Director.

4.9. When the data are obtained from limited access databases, or via a contractual arrangement, written indication of the location of the original data, or key information regarding the database from which it was collected, must be retained by the researcher or research unit.

4.10. Researchers must be responsible for ensuring appropriate security for any confidential material, including that held in computing systems. Where computing systems are accessible through networks, particular attention to security of confidential data is required. Security and confidentiality must be assured in a way that copes with multiple researchers and the departure of individual researchers.

and Section 10

10.5. The supervisor must ensure, as far as possible, the validity of research data obtained by a student under his/her supervision.

and Section 11

11.3. Examples of research misconduct include, but are not limited to, the following:

(a) Fabrication of data: A researcher shall not claim data where none has been obtained.

(b) Falsification of data: A researcher shall not falsify data, including changing records.

⋮

(f) Interference: A researcher or reviewer shall not intentionally and without authorization take or sequester or materially damage any research-related property of another, including without limitation the apparatus, reagents, biological materials, writings, **data**, hardware, software, or any other substance or device used or produced in the conduct of research.

and section 1.3

1.3. The broad principles that guide research have been long established. Central to these are the maintenance of high ethical standards, and **validity and accuracy in the collection and reporting of data**. Researchers have an obligation to achieve and maintain the highest standards of intellectual honesty in the conduct of their research.

## A.2 Australian Code for the Responsible Conduct of Research

http://www.nhmrc.gov.au/publications/synopses/_files/r39.pdf
Section 2

### RESPONSIBILITIES OF RESEARCHERS

- 2.5 Retain research data and primary materials

  When considering how long research data and primary materials are to be retained, the researcher must take account of professional standards, legal requirements and contractual arrangements.

  - 2.5.1 Researchers should retain research data and primary materials for sufficient time to allow reference to them by other researchers and interested parties. For published research data, this may be for as long as interest and discussion persist following publication.
  - 2.5.2 Research data should be made available for use by other researchers unless this is prevented by ethical, privacy or confidentiality matters.
  - 2.5.3 Research data should be retained for at least the minimum period specified in the institutional policy.
  - 2.5.4 If the results from research are challenged, all relevant data and materials must be retained until the matter is resolved. Research records that may be relevant to allegations of research misconduct must not be destroyed.
  - 2.5.5 The institutional policy on the secure and safe disposal of primary materials and research data must be followed.

- 2.6 Manage storage of research data and primary materials

  Researchers must manage research data and primary materials in accordance with the policy of the institution. To achieve this, researchers must:

  - 2.6.1 Keep clear and accurate records of the research methods and data sources, including any approvals granted, during and after the research process.
  - 2.6.2 Ensure that research data and primary materials are kept in safe and secure storage provided, even when not in current use.
  - 2.6.3 Provide the same level of care and protection to primary research records, such as laboratory notebooks, as to the analysed research data.
  - 2.6.4 Retain research data, including electronic data, in a durable, indexed and retrievable form.
  - 2.6.5 Maintain a catalogue of research data in an accessible form.
  - 2.6.6 Manage research data and primary materials according to ethical protocols and relevant legislation.

- 2.7 Maintain confidentiality of research data and primary materials

  Researchers given access to confidential information must maintain that confidentiality. Primary materials and confidential research data must be kept in secure storage. Confidential information must only be used in ways agreed with those who provided it. Particular care must be exercised when confidential data are made available for discussion.

## A.3 ARC Funding Agreement for Discovery Projects

http://www.arc.gov.au/rtf/DP08_FundingAgreement.rtf
Schedule C, page 31, item C-6.

> Social Science Data Sets: Any digital data arising from a Project involving research relating to the social sciences should be lodged with the Australian Social Science Data Archive (ASSDA) for secondary use by other investigators. This should normally be done within two years of the conclusion of any fieldwork relating to the Project research. If a Chief Investigator is not intending to do so within the two-year period, he/she should include the reasons in the Project's Final Report.

Section 19.3, page 17

> The Administering Organisation shall consider the benefits of depositing the data and any publications arising from each Project in an appropriate subject and/or institutional repository wherever such a repository is available. If the Administering Organisation is not

intending to deposit the data from a Project in a repository either before, or within six months after, the completion of the Project the reasons for not doing so must be detailed in the Project's Final Report. Any research outputs that have been or will be deposited in appropriate repositories should be identified in the Final Report.

# Appendix B

# Copyright Licences

## B.1 Demetrius Copyright Licence

http://sts.anu.edu.au/demetrius/policy/rights_license.php

### Non-Exclusive Distribution Licence

In order for Australian National University (ANU) to reproduce, translate and distribute your submission worldwide, your agreement to the following terms is necessary. Please take a moment to read the terms of this licence.

By submitting this licence, you (the author(s) or copyright owner) grant to the ANU the non-exclusive right to reproduce, translate (as defined below), and/or distribute your submission (including the abstract) worldwide in print and electronic format and in any medium, including but not limited to audio or video.

You agree that ANU may, without changing the content, translate the submission to any medium or format for the purpose of preservation.

You also agree that ANU may keep more than one copy of this submission for purposes of security, back-up and preservation.

You warrant that the submission is your original work, and that you have the right to grant the rights contained in this licence. You also warrant that your submission does not, to the best of your knowledge, infringe upon anyone's copyright, or contain any libellous or defamatory material.

If the submission contains material for which you do not hold copyright, you warrant that you have obtained the unrestricted permission of the copyright owner to grant ANU the rights required by this licence, and that such third-party owned material is clearly identified and acknowledged within the text or content of the submission.

IF THE SUBMISSION IS BASED UPON WORK THAT HAS BEEN SPONSORED OR SUPPORTED BY AN AGENCY OR ORGANISATION OTHER THAN ANU, YOU WARRANT THAT YOU HAVE FULFILLED ANY RIGHT OF REVIEW OR OTHER OBLIGATIONS REQUIRED BY SUCH CONTRACT OR AGREEMENT.

ANU will clearly identify your name(s) as the author(s) or owner(s) of the submission, and will not make any alteration, other than as allowed by this licence, to your submission.

## B.2 Creative Commons

List of CC Licences – http://creativecommons.org/licenses/

### Version 3.0 Licences

- Attribution – Full Licence[1]
- Attribution-NoDerivs – Full Licence[2]
- Attribution-NonCommercial-NoDerivs – Full Licence[3]
- Attribution-NonCommercial – Full Licence[4]
- Attribution-NonCommercial-ShareAlike – Full Licence[5]
- Attribution-ShareAlike – Full Licence[6]

### Attribution

You let others copy, distribute, display, and perform your copyrighted work – and derivative works based upon it – but only if they give credit the way you request.

### No Derivative Works

You let others copy, distribute, display, and perform only verbatim copies of your work, not derivative works based upon it.

### Noncommercial

You let others copy, distribute, display, and perform your work – and derivative works based upon it – but for noncommercial purposes only.

### Share Alike

You allow others to distribute derivative works only under a license identical to the license that governs your work.

[1] http://creativecommons.org/licenses/by/3.0/legalcode
[2] http://creativecommons.org/licenses/by-nd/3.0/legalcode
[3] http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode
[4] http://creativecommons.org/licenses/by-nc/3.0/legalcode
[5] http://creativecommons.org/licenses/by-nc-sa/3.0/legalcode
[6] http://creativecommons.org/licenses/by-sa/3.0/legalcode

# Bibliography


[1] *Intellectual Property - Ownership, Protection and Commercialisation*, ANU Research Office, http://info.anu.edu.au/Policies/_documents/pdf/legal/517ippolicymay06.pdf

[2] *ANU Intellectual Property Manual*, ANU Research Office, http://info.anu.edu.au/Policies/_documents/pdf/research/ip/anu_ip_manual.pdf

[3] *Responsible Practice of Research*, ANU Research Office, http://info.anu.edu.au/Policies/_DRO/Policies/Responsible_Research_Practice.asp?tab=1

[4] *Australian Code for the Responsible Conduct of Research*, National Health and Medical Research Council (NHMRC) & Australian Research Council (ARC), http://www.nhmrc.gov.au/publications/synopses/_files/r39.pdf

[5] *Creating a Research Project Data Management Plan*, Belinda Weaver, University of Queensland, http://www.library.uq.edu.au/escholarship/BW_dmp.pdf

[6] *Importance of Data Management*, Deborah Terry, http://www.library.uq.edu.au/escholarship/DT_dmp.pdf

[7] *Building the Infrastructure for Data Access and Reuse in Collaborative Research: An Analysis of the Legal Context*, Anne Fitzgerald & Kylie Pappalardo, OAK Law Project, 2007, http://eprints.qut.edu.au/archive/00008865/

[8] *Guidance on Data Management*, Rural Economy and Land Use Programme Data Support Service, http://www.esds.ac.uk/news/publications/reluaug2006.pdf

[9] *Guide to Social Science Data Preparation and Archiving*, 3rd Edition, 2005, Inter-University Consortium for Political and Social Research (ICPSR), http://www.icpsr.umich.edu/ICPSR/access/dataprep.pdf

[10] *Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems*, Cornell University Library, http://www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html

[11] *Sustainability Issues for Australian Research Data: The Report of the Australian e-Research Sustainability Survey Project*, Markus Buchhorn & Paul McNamara, Australian National University, 2006, http://dspace.anu.edu.au/manakin/handle/1885/44304

[12] *Ten Major Issues in Providing a Repository Service in Australian Universities*, Margaret Henty, Australian Partnership for Sustainable Repositories (APSR), D-Lib Magazine, May/June 2007, http://www.dlib.org/dlib/may07/henty/05henty.html

[13] *Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals*, Stevan Harnad & Tim Brody, D-Lib Magazine, June 2004, http://www.dlib.org/dlib/june04/harnad/06harnad.html

[14] *Archives Act (1983)* http://www.comlaw.gov.au/comlaw%5Cmanagement.nsf/lookupindexpagesbyid/IP200401335?OpenDocument

[15] *Privacy Act 1983* http://www.comlaw.gov.au/ComLaw/Management.nsf/current/bytitle/32AA97DFE9AA8326CA256F7100071D25?OpenDocument&mostrecent=1