Data Management
Tame your research data!

Cameron McLean
Matthew Moore

Centre for eResearch
Learning Objectives

- Awareness of the storage services available at UoA
- Understand how to perform common tasks using figshare
- Awareness of the research data landscape
- Develop a strategy for capturing and organising research data
- Understand legal and ethical issues around research data
What is data?
What is your research topic?
What kinds of data will you deal with?

Where do you store your data?
How do you organise and label it?
How do you manage backups?
How do you describe or document your data?
Storage Options
Storage Options
Storage Options

IBM GPFS

https://www.flickr.com/photos/chrissamuel/
Storage Options
Storage Options

Google Drive

Keep everything. Share anything.
Storage Options
Publishing Data

Scientific Integrity
Funder requirements
Impact
Collaboration
Innovation and reuse
Preservation
Teaching
Public record
UoA Research Data Publishing and Discovery Service

auckland.figshare.com
Data Sharing and Management Snafu in 3 Short Acts

NYU Health Sciences Library

https://www.youtube.com/watch?v=N2zK3sAtr-4
Data Management Plans

NZ - based on NSF
Data collection

What kinds?
How much?  \[ \text{raw} + (\text{analysed} \times \text{no. analyses}) + (\text{backup} \times \text{redundancies}) \]
Will it grow?
Will it change over time?
What file formats?
How will you organise it?
Where will you store it?
How will you document and describe it?
How will you check it for errors?
Hierarchies
Hierarchies – tips

- Follow conventions from other host projects or communities if they exist.
- Try to avoid overlapping categories.
- Don’t let folders get too big, or too deep.
- Avoid using the same name for subfolders (or files in different subfolders).
Naming Files and Folders

- Project/grant name/number
- Date of creation YYYYMMDD
- Initials of creator
- Description of content
- Collection method
- Version number x.y
Filename Schemes

[investigator]-[method]-[specimen]-[yyymmdd].ext
cm-lcms-8887-20160126.dat

type of file]-[creator]-[subject]-[yyymmdd].ext
transcript-cm-fsgroup-2016060.md

[date]-[type]-[subject].ext
20130412-interview-recording-MDB.mp3
How does this compare with what you do now?
File Formats

All files
All images
ASCII art (*.txt, *.ansi, *.txt)
Alias Pix image (*.pix, *.matte, *.mask, *.alpha, ...)
AutoDesk FLIC animation (*.fli, *.flic)
C source code (*.c)
C source code header (*.h)
Colored XHTML (*.xhtml)
DDS image (*.dds)
Digital Imaging and Communications in Medicine image (*.dcm, *.dcm)
Encapsulated PostScript Image (*.eps)
Flexible Image Transport System (*.tif, *.flts)
GIF image (*.gif)
GIMP brush (*.gbr)
GIMP brush (animated) (*.gih)
GIMP compressed XJT image (*.xjl, *.xj gz, *.xj bz2)
GIMP pattern (*.pat)
HTML table (*.html, *.htm)
JPEG image (*.jpg, *.jpeg, *.jpe)
KISS CEL (*.cel)
MNG animation (*.mng)
Microsoft Windows icon (*.ico)
OpenRaster (*.ora)
PBM image (*.pbm)
PGM image (*.pgm)
PNG image (*.png)
PNM image (*.pnm)
PPM image (*.ppm)
Photoshop image (*.psd)
Portable Document Format (*.pdf)
PostScript document (*.ps)
SUN Rasterfile image (*.im1, *.im8, *.im24, *.im32, ...)
Silicon Graphics IRIS image (*.sgi, *.rgb, *.rgba, *.bw, ...)
TIFF image (*.tif, *.tiff)
TARGA image (*.tga)
Windows BMP image (*.bmp)
X BitMap image (*.xbm, *.icon, *.bitmap)
File Formats

01101100
01101111
01110110
01100101
File Formats

Open. Standardised. In wide use. Easy to datamine, transform, or re-cast.

What software do you expect to use? Are you collaborating or sharing with others?

Domain specific standards?

Consider fidelity or quality issues if using compression.
Metadata

Project
What is the study
Methodologies and instruments
Bibliographic references

File/Database
How files or tables relate
What formats
README.txt

Item or variable
Meaning or definition of variable terms
Metadata
Metadata

What metadata should I provide for single items?
What metadata should I provide for collections, and about the whole study?
Which standards will I use?
How long to keep data?
University Policy

Minimum 6 years

Clinical trial – 10 years (or until children turn 26)

Patent? – 21 years from date of filing

Ethics? – Check

Community or heritage value – indefinitely
Who owns your data?
HACKY HOUR @ AUCKLAND

Research - Tools - Code - Sharing
3-4pm @ Strata Cafe every Thursday
In the creation of the workshop we have taken inspiration and adapted some ideas and materials from a number of existing resources.

Research Data Management: File Organization
Katherine McNeill & Helen Bailey
(CC-BY-NC-SA)

Melbourne_MANTRA
University of Melbourne and University of Edinburgh
http://library.unimelb.edu.au/digitalscholarship/training_and_outreach/mantra2
(CC-BY)

Research Data Management: 101 The Lifecycle of a Dataset
Katherine McNeill
(CC-BY-NC-SA)

Escaping Datageddon - Dorothea Salo and Ryan Schryver - University of Wisconsin
(CC-BY)

http://www.data-archive.ac.uk/media/2894/managingsharing.pdf
(CC-BY-NC-SA)

(CC-BY)

Tidy Data http://vita.had.co.nz/papers/tidy-data.pdf