www.wanala.org

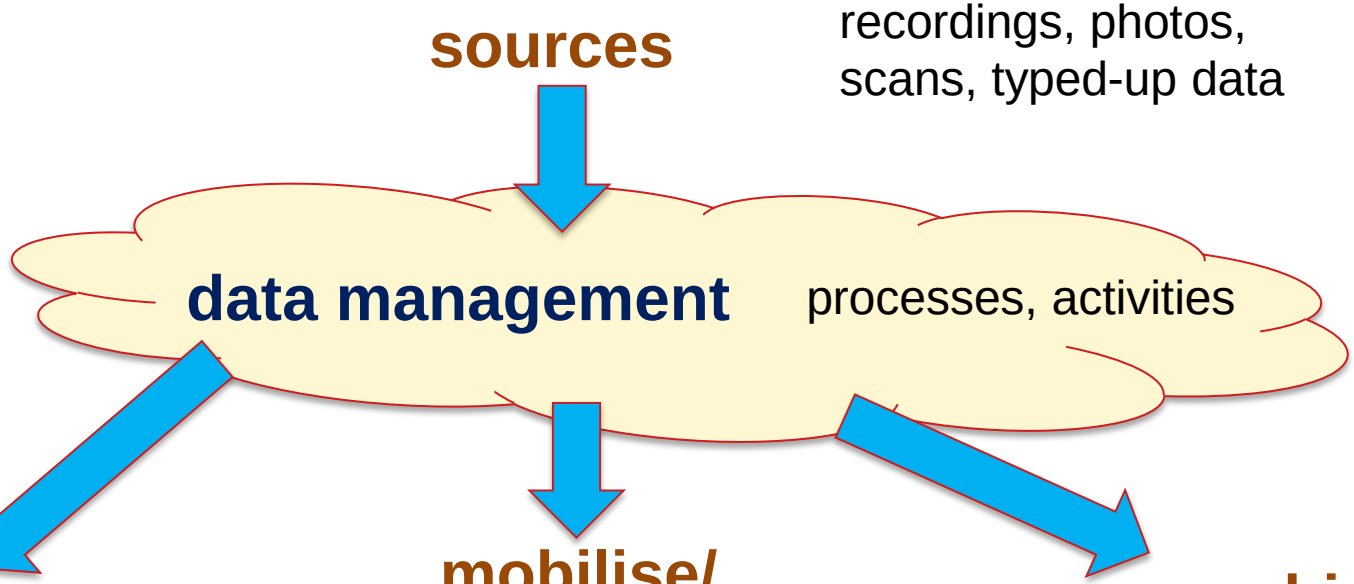
WESTERN AND NORTHERN
ABORIGINAL LANGUAGES ALLIANCE

Data management

Dave Nathan, Batchelor Institute NT

Batchelor Institute

From reality to data (and back!)



share/ collaborate

- research
- return to community
- work with publishers
- collaborate to mobilise

mobilise/ repurpose

- teaching/learning resources
- media - radio, mobile, CD ...
- web
- apps

archive

Batchelor Institute

Good (digital) data management =

- **content** clearly expressed and formatted appropriately
- computer **files**:
  - suitable **formats**
  - suitable **filenames**
  - **portable**
- **organisation** – file system, metadata, catalogue, etc
- **preservation** - backup and archiving

Batchelor Institute

Content

- images – is the object of attention clearly visible, well lit, in focus etc (and do you have caption or metadata for each, including permissions & contacts)?
- audio – is it the sound you want clear, and the noise level low (and metadata, permissions etc)?
- video – is the image well framed and well lit, does it capture the subject and activity, and is the audio clear (and metadata, permissions etc)?

Batchelor Institute

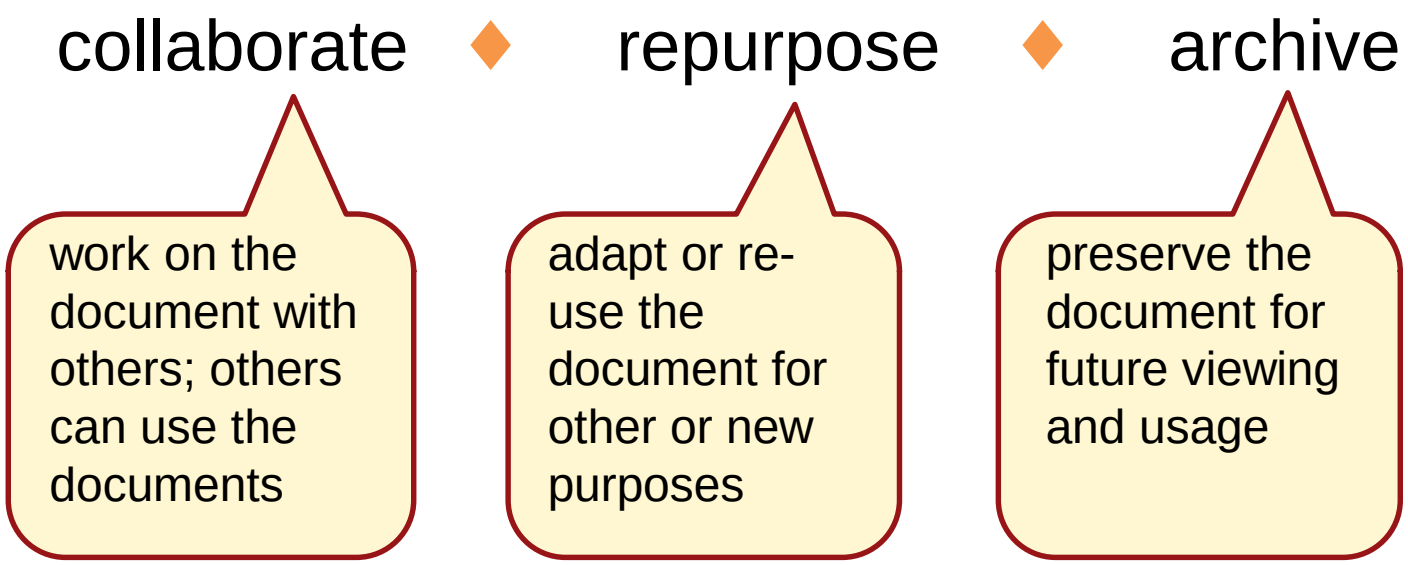
Content

- document examples
  - is it a list? – use a table or spreadsheet (and consider each column) metadata example
  - if possible, try to convey information clearly and in explicit text or images, *not* by using formatting that won't work as plain text example plain text better
    OR
    if formatting ***is*** important, save as PDF

Batchelor Institute

File formats

- file formats are how the information content is packed into the computer file. Let's distinguish "full" file formats from distribution formats.
- "full" formats make it most possible to:



Batchelor Institute

File formats

- but other formats may be better, or necessary, for particular purposes
- in some cases you need particular software to create a particular (“distribution”) format
- so what are these “full” formats (sometimes called archival or preservation formats) ?

Batchelor Institute

“Full” file formats for images

- use TIFF (_.tiff) or PNG (_.png) – only use JPG (_.jpg) if that is the original (e.g. from a camera)
- there are other options/settings (for images, video and audio, called “resolution”):
  - bits per inch, dots per inch, or pixels per inch (bpi, dpi, ppi)
    - the density of an image – how tightly the little dots are packed together
    - for printing: at least 250 dpi
    - archival or high quality printing: 600 dpi
  - colour depth, bit depth, or number of colours
    - how accurately each dot represents a colour
    - in general, use “millions of colours”, or *at least* 8 bit (16 or 24 are better)

Batchelor Institute

180 dpi
bit depth 24

Batchelor Institute

72 dpi
bit depth 24

Batchelor Institute

Batchelor Institute

“Full” file formats for audio

- use only WAV (_.wav, also known as PCM)
- resolution
  - good: 16 bit, 44.1KHz (kilohertz), stereo
  - best (but not practical for most): 24 bit 48 KHz stereo
- alternative formats:
  - OGG
  - other formats, eg MP3, are acceptable only if they are the original digital recorded format (i.e. you don’t have anything else)

Batchelor Institute

“Full” file formats for video

- video formats are problematic - in constant change, varying by camera brand
- the concept of “full” archival video format doesn’t apply in practice because:
  - original video files from high quality cameras are too large to be effectively processed and stored by organisations of our scale
  - the files produced by consumer video cameras are smaller and compressed, but can be processed and stored
    - these formats are among the set of formats known as AVCHD, H264, or MPEG4 (MP4)

Batchelor Institute

“Full” file formats for documents

- if original is MS Word, then

  is formatting necessary to convey content?
  - if no, then
    - convert to plain text (_.txt)
  - if yes, then
    - save both original MS Word (_.docx) and an archivable version such as PDF
      - strictly, PDF/A; also ok are ODF (Open Office/Libre Office) or ODT

Batchelor Institute

“Full” file formats for documents

- if original is MS Excel, then

  is formatting necessary to convey content?
  - if no, then
    - convert to CSV “comma separated values” (_.csv)
  - if yes, then
    - save both original Excel (_.xslx) and an archivable version such as
      - strictly, PDF/A; also ok are ODF (Open Office/Libre Office) or ODT

Batchelor Institute

“Full” file formats for documents

- if original is plain text (_.txt) then

  congratulations, you are using best practice and your document is already archive-ready

Batchelor Institute

… for scanning physical documents

- use a good quality scanner or high quality camera with stand, lights etc
- file formats: TIFF, BMP, PNG
  - note: for scans and images of predominantly text content, ***do not*** use JPEG - use TIFF, BMP, or PNG
- scan in colour (or greyscale), ***never*** ‘black and white’
- resolution
  - 24 bit colour depth (aka "millions of colours")
  - for printing: at least 250 dpi
  - archival or high quality printing: 600 dpi

Batchelor Institute

A couple of other format things ...

- you may also need to consider “text encoding” if you use special characters (i.e. ones that aren’t on a normal Australian keyboard, such as ā, ŋ or 語)
  - always use UTF-8 (= Unicode, ISO-10646)
- there are other formats for files created by special software for language documentation and research, such as Miroma, Toolbox, ELAN. These are generally OK for archiving, although their content is specialised to the particular software that created them.

Batchelor Institute

Example - au

- record as: WAV file,
  44KHz (kilohertz),
  6 bit
  stereo

Batchelor Institute

Example - audio

- but for web, mobile, some apps, you often need MP3 (compressed)
- typical is 128kbs CBR (constant bit rate)

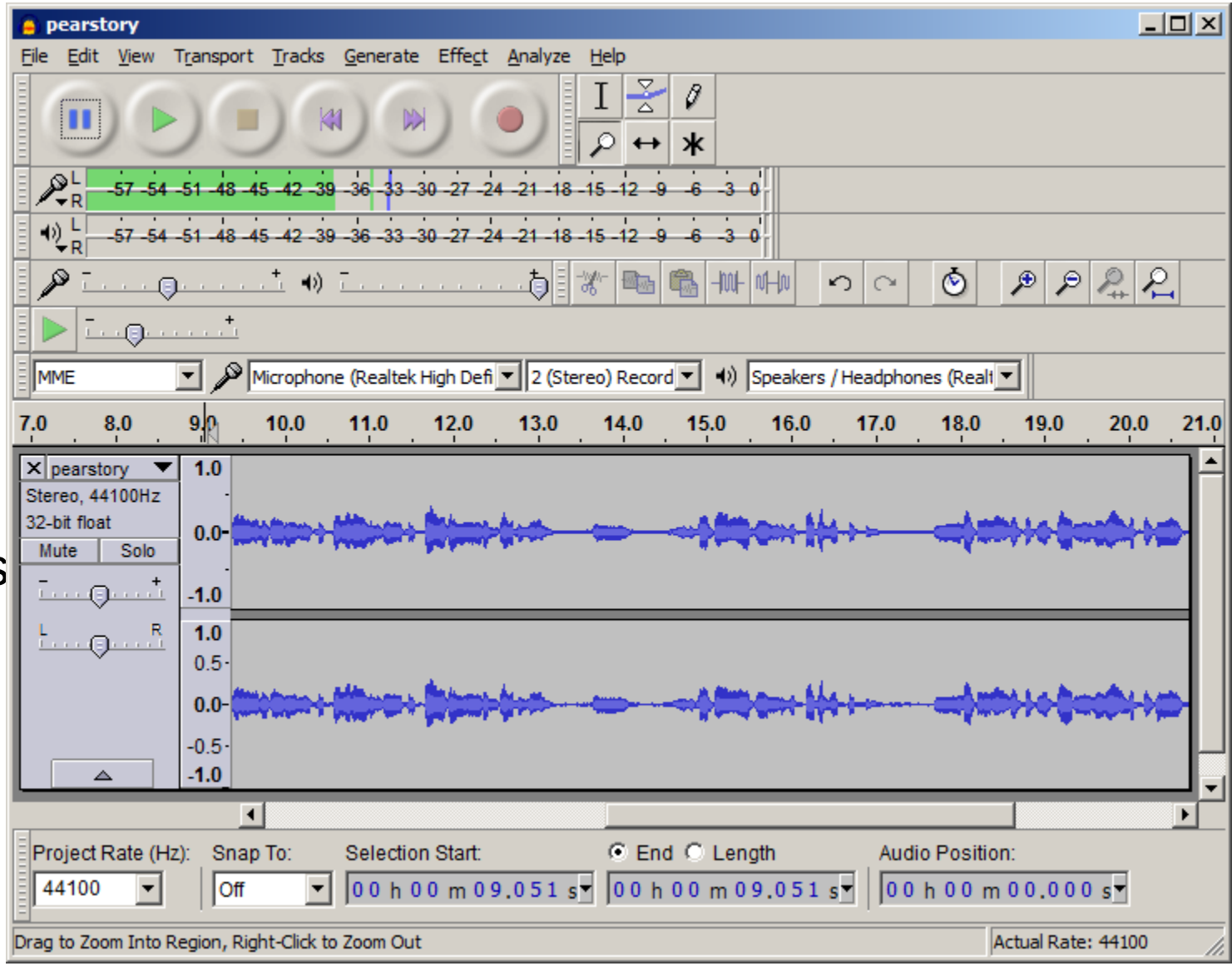
Batchelor Institute

Example - audio

- use software (eg Audacity) to convert between formats

Batchelor Institute

- File >
- Export audio
- Save as type: MP3 files



Batchelor Institute

Filenames

- getting file names (and folder names) right is the key to successful file management
- you may have a lot of files already – consider copying and renaming the whole lot
- in any case, develop a **file naming system**, suited to the types and content of files you expect to hold, **document** it, and **stick to it**

Batchelor Institute

Filenames – do this

- do not make filenames too long or try to stuff too much information into them
- naming rules:
  - use only letters, numbers, hyphens (-) and underscores (_)
- ***do not*** use:
  - any characters *except* a-z, A-Z, 0-9, hyphen (-) or underscore (_)
  - spaces
  - fullstops (except for the one before the extension, i.e. as in document**.**txt)

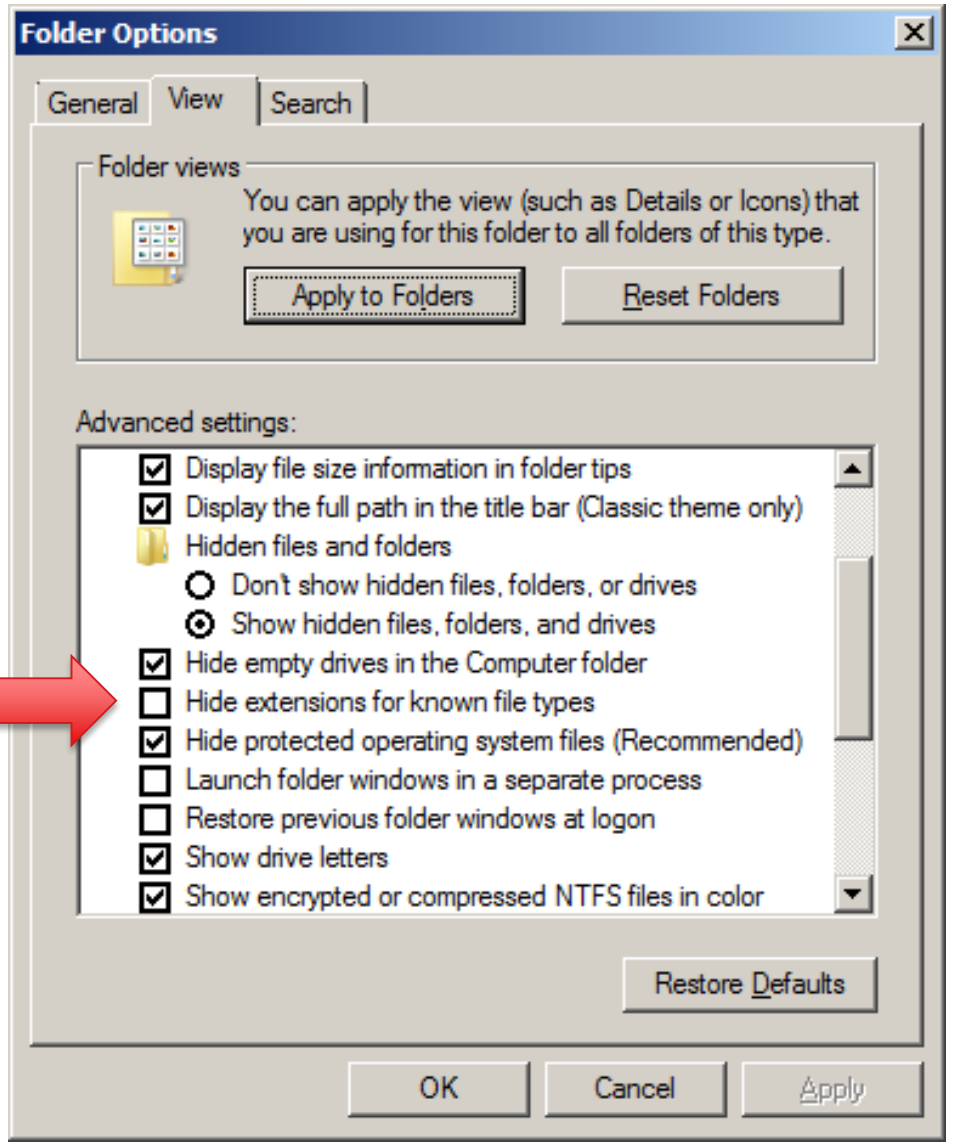
Batchelor Institute

Filenames – do this

- ensure file name has the right extension (e.g. .txt, .jpg, .wav etc).
  - some Windows computers “hide extensions for known file types”
  - extensions are also required on Macs

Batchelor Institute

TURN OFF
“Hide extensions for known file types”



Batchelor Institute

Filenames – do this

- ensure file name has the right extension (e.g. .txt, .jpg, .wav etc).
  - some Windows computers “hide extensions for known file types”
  - extensions are also required on Macs
- create and maintain a catalogue, inventory, spreadsheet, or database to help you manage your files and store additional information about them (metadata)

Batchelor Institute

example example example

How about these file names?

1. ready.audio.wav ✗
2. ReAlLyOdDtOReAd.txt ✓
3. éclair.jpg ✗
4. éclair_fr.jpg ✗
5. e'clair.jpg ✗
6. french-cake.jpg ✓
7. french-cake.jaypeg ✗
8. dictionary-master ✗
9. ɘɬIŋʰ.eaf ✗
10. ice cream.doc ✗
11. OBAMA.TXT ✓
12. Obama.txt ✓

Batchelor Institute

Metadata

- metadata is generally defined as: *data about data*
- you may want to think about it as a catalogue, inventory, database, or even a fancy list
- the big picture:

  metadata provides the **contexts** for **managing** and **understanding** files
- metadata can be in many forms, commonly using spreadsheets (e.g. Excel), or specialised software and formats (e.g. IMDI/Arbil, Miromaa, Saymore
- metadata is a complex topic and needs its own sessions

Batchelor Institute

Example system for recordings

pattern: aaa_bb_cc_yyyy-mm-dd_nnn.wav

aaa = village code*

bb = (main) speaker code*

cc = genre/event code*

yyyy-mm-dd = date (why this order?)

nnn = optional number (e.g. 001) (why?)

.wav = correct extension for file content type

example file: trk_lm_st_2006-07-18_003.wav

Batchelor Institute

* Make a table to explain your **codes**

The big picture

- ask these questions to find out if your digital data is likely to be preservable and shareable:
  - is the data or file **portable**, i.e able to be transferred in various ways (network, hard disk, memory stick) to different computers and operating systems and without modification – *and then still work*? AND
  - can the data or file be opened and read by various software (especially ones that are free, open source, and with a long lifetime) – *and retain its content true to source*?
    e.g. a plain text file can be opened by any text editor, browser etc

Batchelor Institute

Thank you

- contact me at

  david.nathan@batchelor.edu.au

Batchelor Institute