Formatting and organising research data

Research Data Management Support Services
UK Data Service
University of Essex

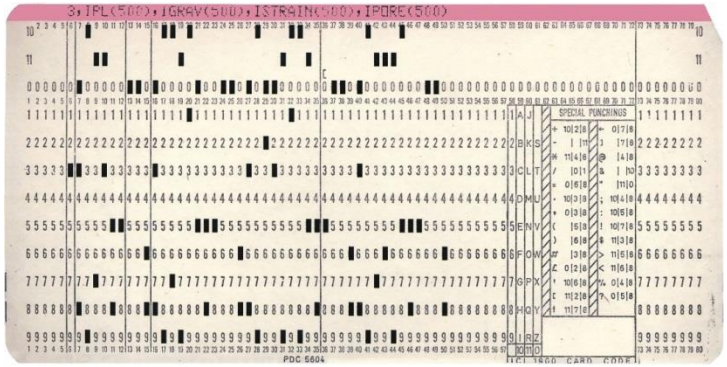
April 2014

UK Data Service

Overview

- File formats
- File conversions
- Organising files and folders
- File naming
- Version control and authenticity

Can you understand/use these data?



File formats

Digital data can take countless different form(at)s…

A file format is a specific way of structuring information so that a machine, and therefore a person, can understand it

- should be readable by as many types of system as possible
- without compromising the purpose of the data

File formats

Choice of software format for digital data:

- planned data analyses
- software availability/cost
- hardware used – e.g. audio capture
- discipline-specific standards and customs

Digital data is software dependent, so endangered by obsolescence of software/ hardware

Best formats for long-term preservation:

- standard, interchangeable, open
- *e.g. tab-delimited, comma-delimited (CSV), ASCII, RTF, PDF/A, OpenDocument format, SPSS portable, XML*
- UK Data Archive optimal file formats for various data types
- Digital Preservation Coalition guidance on preservation formats

File format conversions

Convert data for preservation or back-up:

- export
- save as
- scripts

Beware of conversion errors or losses:

- loss of internal metadata
  e.g. convert mp3 audio to ogg
- loss of editing, formatting, formulae
  e.g. convert DOCX to RTF; XLSX to CSV
- truncation or loss of values
  e.g. string variables lost in SPSS – Stata conversion; MS Access memo fields truncated in conversion to CSV

Check for errors and changes after conversion

Example: format conversion

	A	B	C	D	E	F
1		Timber volumes in m3				
2	Year	1994	1995	1996	1997	1998
3	Date recorded	20/01/1995	23/01/1996	11/01/1997	16/01/1998	14/12/1998[1]
4	Logging private land	20346.345	47005.223	26001.754	11468.897	0.000
5	Logging forest reserves	4060.567	1777.783	804.997	0.000	3329.653
6	Logging state land	0.000	1200.000	559.162	2077.567	358.935
7	Total	61119.912	87065.006	64802.913	51354.464	5686.588
8						
9		Data missing				
10		Estimate				
11						
12	[1] temporary volumes					

MS Excel (.xlsx) format

	A	B	C	D	E	F
1		Timber volumes in m3				
2	Year	1994	1995	1996	1997	1998
3	Date recorded	20/01/1995	23/01/1996	11/01/1997	16/01/1998	14/12/19981
4	Logging private land	20346.345	47005.223	26001.754	11468.897	0
5	Logging forest reserves	4060.567	1777.783	804.997	0	3329.653
6	Logging state land	0	1200	559.162	2077.567	358.935
7	Total	61119.912	87065.006	64802.913	51354.464	5686.588
8						
9		Data missing				
10		Estimate				
11						
12	1 temporary volumes					
13						

Formatting change

Loss of annotation

Tab-delimited text format

UK Data Service

Example: format conversion

Different formats store date values in different ways, and format conversion can wreak havoc with these.

e.g. 21:55 on the 21st April 2013 can be stored as:

- **1366581312**

Unix time - seconds elapsed since midnight 1 January 1970

or

- **2013-04-21T21:55:12Z**

ISO 8601 time - and international standard for representing time and date stamps

Quality assurance

Quality assurance procedures should be undertaken throughout the research process, ensuring data are:

- clean
- verified
- validated

Depending on the type of data, you may be able to automate aspects of this process using:

- statistical software to check e.g. frequencies on quantative data
- consistency checking with data manipulation tools like OpenRefine

Qualitative data collectors in for a harder time – manual proofreading

Can you understand these data?

SrvMthdDraft.doc 
SrvMthdFinal.doc 
SrvMthdLastOne.doc 
SrvMthdRealVersion.doc 

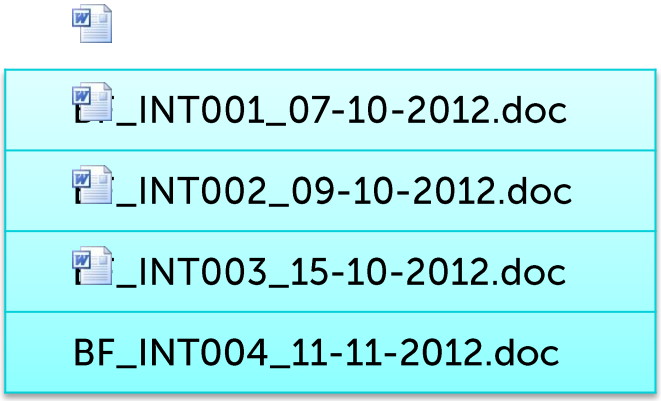
Organising data

Plan in advance how best to organise data

Use a logical structure and ensure collaborators understand

Examples

- hierarchical structure of files, grouped in folders, e.g. audio, transcripts and annotated transcripts
- measurement data – original, processed, analysed etc.
- interview transcripts – individual well-named files



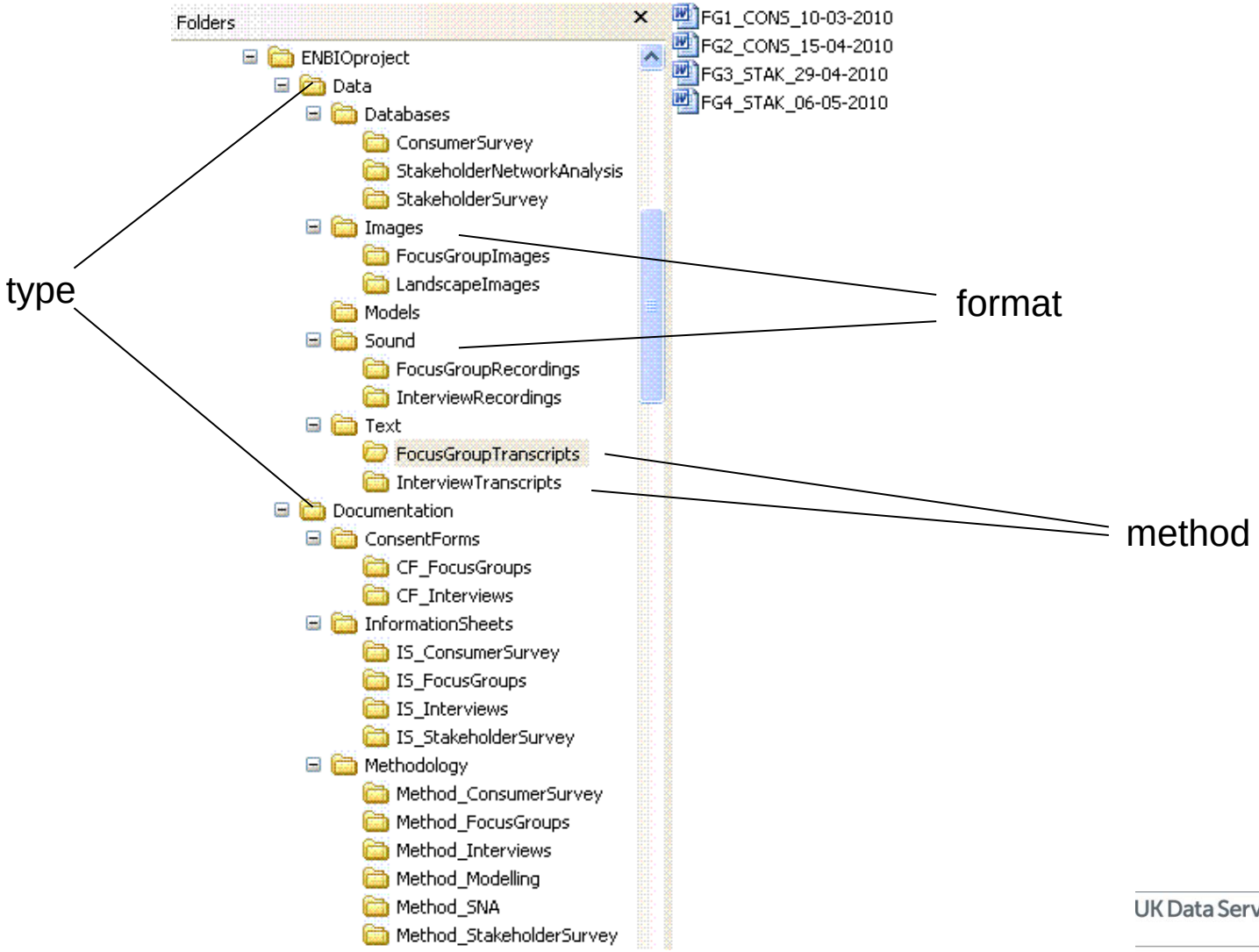
File naming

- file name = principal identifier of file
- use logical naming i.e. easy to identify and retrieve the file
- naming provides organisation, context & consistency
- name elements: version number, date, content description, creator name

Best practice

- name independent of location
- relevant to content
- no special characters, dots or spaces
- for separation use underscores _
- versioning via filename: ascending, decimal version numbers
- avoid very long file names

Directory structure



Version control

- Keep track of different copies or versions of data files
  - Useful for files kept in multiple locations
  - Or which have multiple users
  - A way to safeguard against accidental changes

- File names are a good way to do this
  - Unique descriptive names for files
  - Include date and/or version number in name
  - Indicate relationships between files

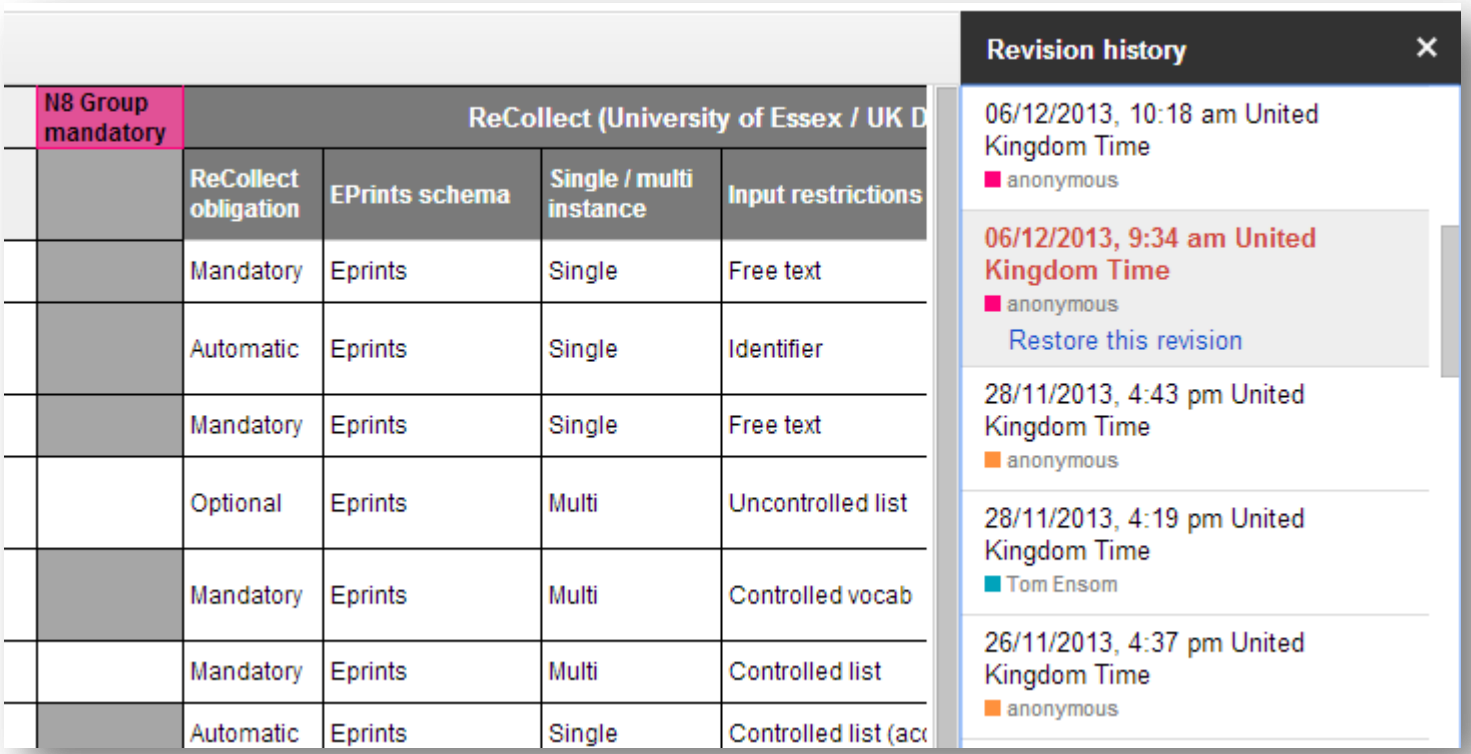
e.g. *FoodInterview_1_draft; FoodInterview_1_final; HealthTest_06-04- 2008; BGHSurveyProcedures_00_04*

Example: version control table

Title:		Vision screening tests in Essex nurseries	
File Name:		VisionScreenResults_00_05	
Description:		Description of the data files ……	
Created By:		Chris Wilkinson	
Maintained By:		Sally Watsley	
Created:		04/07/2007	
Last Modified:		25/11/2007	
Based on:		VisionScreenDatabaseDesign_02_00	
Version	**Responsible**	**Notes**	**Last amended**
00_05	Sally Watsley	Version 00_03 and 00_04 compared by SW	25/11/2007
00_04	Vani Yussu	Entries checked by VY, independent from previous	17/10/2007
00_03	Steve Knight	Entries checked by SK	29/07/2007
00_02	Karin Mills	Test results 81-120 entered	05/07/2007
00_01	Karin Mills	Test results 1-80 entered	04/07/2007

Example: Google Drive version control

- Collaboratively edit documents in ‘the cloud’ while tracking version history



N8 Group mandatory	ReCollect (University of Essex / UK D			
	ReCollect obligation	EPrints schema	Single / multi instance	Input restrictions
	Mandatory	Eprints	Single	Free text
	Automatic	Eprints	Single	Identifier
	Mandatory	Eprints	Single	Free text
	Optional	Eprints	Multi	Uncontrolled list
	Mandatory	Eprints	Multi	Controlled vocab
	Mandatory	Eprints	Multi	Controlled list
	Automatic	Eprints	Single	Controlled list (acc

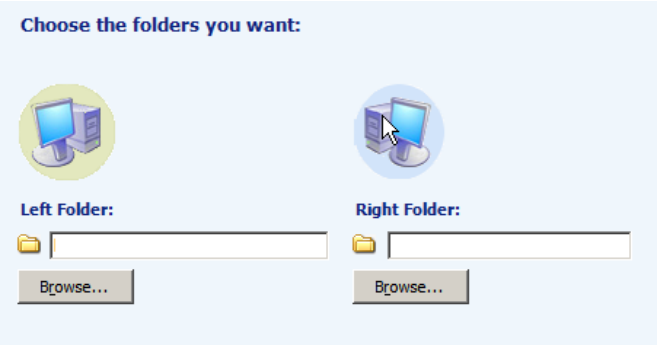
Version control

Multiple users of data files

- control rights to file editing: read/write permissions
  e.g. Windows Explorer
- versioning/file sharing software: check files out/in
  e.g. SharePoint, CMS, Google Docs, Amazon S3
- manual merging of multiple entries/edits

Synchronise files

- *software*
  e.g. MS SyncToy
- *command line*
  e.g. robocopy, rsync
- *web-based*
  e.g. DropBox, Google Drive



Digitisation of data

Non-digital data can (and should!) be digitised.

Approach dependent on situation

- e.g. type of data, resources availably, purpose of digitisation

Some general notes:

Photographs

- scan and save as TIFF

Maps

- scan, georeference using GIS software, and save as GeoTIFF

Audio e.g. audio recording

- capture as WAV

Video

- video formats complex, take care when digitising/converting to avoid degradation and errors

Digitising textual data

Text – more complex, with tiers of digitisation:

Create image file

- scan (or photograph) and save as TIFF image file
- used for poor typeface, handwritten materials, text with tables & graphs

Create searchable PDF

- collate TIFFs and convert to PDF
- bookmark PDF file for navigation: contents page, headings & metadata

Create rich text using Optical Character Recognition (OCR)

- automatically convert TIFF to RTF format
- requires rigorous proof reading and checking

Transcribe manually

- represent the original material as closely as possible
- avoid using formatting in data files

Data transcription

- translation between forms

- all transcription is:
  - representational
  - selective – can be multiple-perspective for video
  - interpretive
  - theoretical

Transcription template

Should:

- possess a unique identifier
- adopt a uniform layout throughout the research project
- make use of speaker tags - turn-taking
- carry line breaks
- be page numbered
- carry a document header giving brief details of the interview: date, place, interviewer name, interviewee details, etc.

Other considerations:

- cover page
- compatibility with import features of Computer Assisted Qualitative Data Analysis Software (CAQDAS)

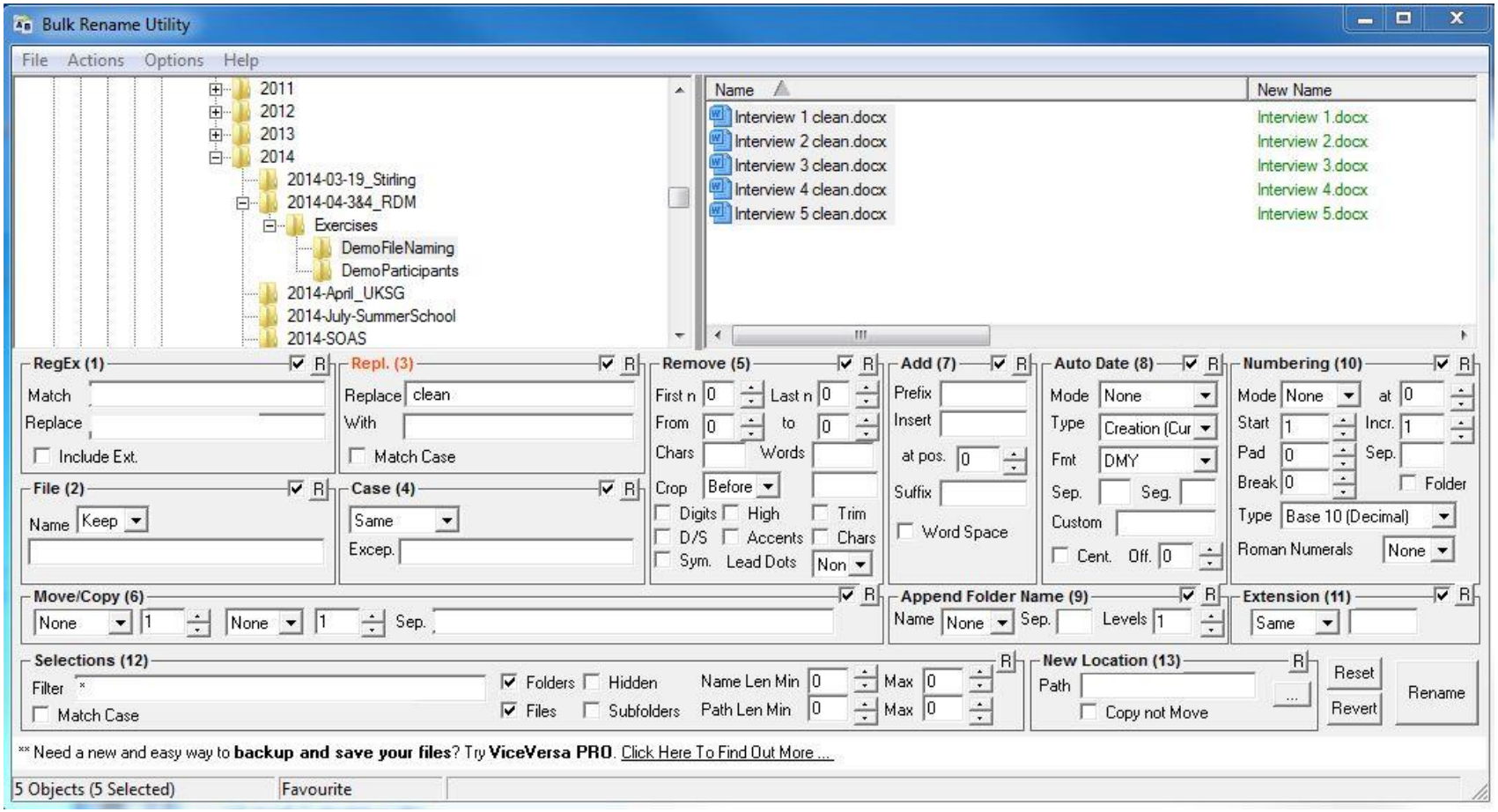
Transcription issues

- what to transcribe?
  - verbal and non-verbal?
  - turn-taking?
  - ‘interruptions’
- who does it – researcher, service?
- implications of technologies – video, multiple camera, screen capture, webcams

Transcription and data sharing

- added issues to consider when transcribing for data sharing
- in what format will the transcript be accessed?
  - paper
  - digital file
  - web
  - standalone or part of collection
- who will be reading the transcript?
  - need for more/different contextual information (‘metadata’) for secondary users?
  - exposes the researcher’s practices

Demo: Bulk Rename Utility



UK Data Service

Contacts

Collections Development team
UK Data Service
University of Essex
datasharing@ukdataservice.ac.uk