

Documenting and Organising Research Data for Archiving and Reuse

Scott Summers

UK Data Service

University of Essex

Creating Shareable Research Data: Managing and
Archiving Social Science Research Data

28th and 29th November 2017

UK Data Service



Overview

A crucial part of making data user-friendly, shareable and with long-lasting usability is to ensure they can be understood and interpreted by any user. This requires clear and detailed data description, annotation and contextual information.

Areas to be covered

- What is documentation?
- Why documentation is important
- What information should be captured?
- Study-level documentation and context
- Data-level documentation
- Quality control of data
- Metadata
- Versioning, naming and organising data



What is documentation?

- Data does not mean anything without documentation
 - A survey dataset becomes just a block of meaningless numbers
 - An interview becomes a block of contextless text
- Data documentation might include:
 - A survey questionnaire
 - An interview schedule
 - Records of interviewees and their demographic characteristics in a qualitative study
 - Variable labels in a table
 - Published articles that provide background information
 - Description of the methodology used to collect the data
 - Consent forms and information sheets
 - A ReadMe file



Why document your data?

- Enables you to **understand and interpret** data when you return to it
- It is needed to make data independently understandable and **reusable**
- Helps **avoid incorrect use or misinterpretation**
- If using your data for the first time, what would a **new user** need to know to make sense of it?
- The UK Data Archive uses data documentation to:
 - supplement a data collection with documents such as a user guide(s) and data listing
 - ensure accurate processing and archiving
 - create a catalogue record for a published data collection



What information should be captured?

Contextual information about the project and data

- background, project history, aims, objectives and hypotheses
- publications based on data collection

Data collection methodology and processes

- who collected the data and when
- data collection process and sampling
- instruments used – questionnaires, showcards and interview schedules
- temporal/geographic coverage
- data validation – cleaning and error-checking
- compilation of derived variables
- secondary data sources used
- what data manipulations (if any)

Any useful documentation such as:

- final report, published reports, user guide, working paper, publications and lab books



What information should be captured?

Information on **dataset structure**

- inventory of data files
- relationships between those files
- records and cases...

Variable-level documentation

- labels, codes, classifications
- missing values
- derivations and aggregations

Data confidentiality, access and use conditions

- anonymisation carried out
- consent conditions or procedures
- access or use conditions of data



Documentation should be considered early on

- Start documenting data early
- Good data documentation and metadata depends on what you as the creator can provide
- Start gathering meaningful information from as early on in the research process as possible
- This consideration forms an important part of data management planning



Quantitative study

- Smaller-scale study – single user guide may contain compiled survey questionnaire, methodology information
- Example from Understanding Society, a bigger study - many documents presented separately:

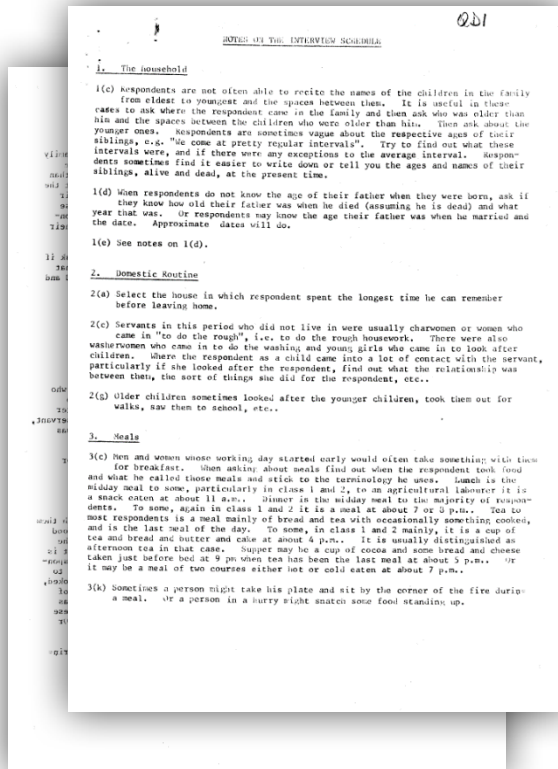
DOCUMENTATION

Title	File Name	Size (KB)
Cognitive Ability Measures	6614_cognitive_ability_measures_v1-1.pdf	348
Revisions November 2013	6614_ukhls_2013_revisions.pdf	375
Wave 1 Adult Main Questionnaire	6614_understanding_society_wave1_questionnaire.v04.pdf	2802
Wave 2 Adult Main Questionnaire	6614_understanding_society_wave2_questionnaire_v04.pdf	3726
Waves 1-3 User Manual	6614_usermanual_wave1to3_v1-1.pdf	883
Wave 3 Youth Self-Completion Questionnaire (GB)	6614_w3_youthquestionnaire_gbritain_annotated.pdf	1469
Wave 1 Consent Package	6614_wave1_consent_package.pdf	645
Wave 1 Adult Self-Completion Questionnaire	6614_wave1_main_adult_sc_questionnaire.pdf	429
Wave 1 Youth Self-Completion Questionnaire	6614_wave1_main_youth_sc_questionnaire.pdf	750
Wave 1 Project Instructions for Interviewers	6614_wave1_project_instructions_interviewers.pdf	2426
Wave 1 Showcards	6614_wave1_showcards.pdf	199



Qualitative study

- A user guide could contain a variety of documents that provide context: interview schedule, transcription notes and even photos



In practice: transcript format

Study Name:
Depositor:
Interviewer:

Interview number:
Interview ID: Firstname Lastname
Date of interview:

Information about interviewee

Date of birth:
Gender:
Geographic region:

Marital status:
Occupation:

Y=Interviewee

I=Interviewer

Y: I came here in late 1968.

I: You came here in late 1968? Many years already.

Y: 31 years already. 31 years already.

I: (laugh) It is really a long time. Why did you choose to come to England at that time?

Y: I met my husband and after we got married in Hong Kong, I applied to come to England.

I: You met your husband in Hong Kong?

Y: Yes.

I: He was working here [in England] already?



Qualitative study – data listing

- Data listing provides an at-a-glance summary of interview sets

Study Number 5407

Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001

Mort, M.

The panel respondents for the study were divided into six population groups. The data list for the diary and interviews has been colour-coded accordingly for clarity, using the depositor's original colours:

Group 1: Farmers	Group 2: Rural Business	Group 3: Agricultural related occupations	Group 4: Frontline Workers	Group 5: Community	Group 6: Animal / Human Health Professionals
------------------	-------------------------	-------------------------------------------	----------------------------	--------------------	----------------------------------------------

1. Interviews

Respondent ID	Population Group	Date of Birth	Gender	Occupation	Interview summary	Place of Interview
PM02	Group 6: Animal / Human Health Professionals	1975	M	Veterinary Surgeon	Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	North Cumbria, respondent's home
PM03	Group 6: Animal / Human Health Professionals	1966	F	Veterinary Surgeon	Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	North Cumbria
PM07	Group 6: Animal / Human Health Professionals	1964	F	Veterinary practice manager	Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	North Cumbria, respondent's home
					Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	

UK Data Service



Data-level documentation

- Aim to embed this documentation in your data file:
- Some examples:
 - SPSS: variable attributes documented in Variable View (label, code, data type and missing values)
 - MS Excel: document properties, worksheet labels (where multiple)
- Qualitative data/text documents:
 - interview transcript speech demarcation (speaker tags)
 - document header with brief details of interview date, place, interviewer name, interviewee details and context



Data-level documentation: variable names

- All structured, tabular data should have cases/records and variables adequately documented with names, labels and descriptions
- Variable names might include:
 - question number system related to questions in a survey/questionnaire
e.g. Q1a, Q1b, Q2, Q3a
 - numerical order system
e.g. V1, V2, V3
 - meaningful abbreviations or combinations of abbreviations referring to meaning of the variable
e.g. oz%=percentage ozone, GOR=Government Office Region, motoc=mother occupation, fatoc=father occupation
 - for interoperability across platforms – variable names should be max 8 characters, without spaces and not start with a number, question marks, exclamation marks or special characters (these are reserved for specific purposes in software applications)

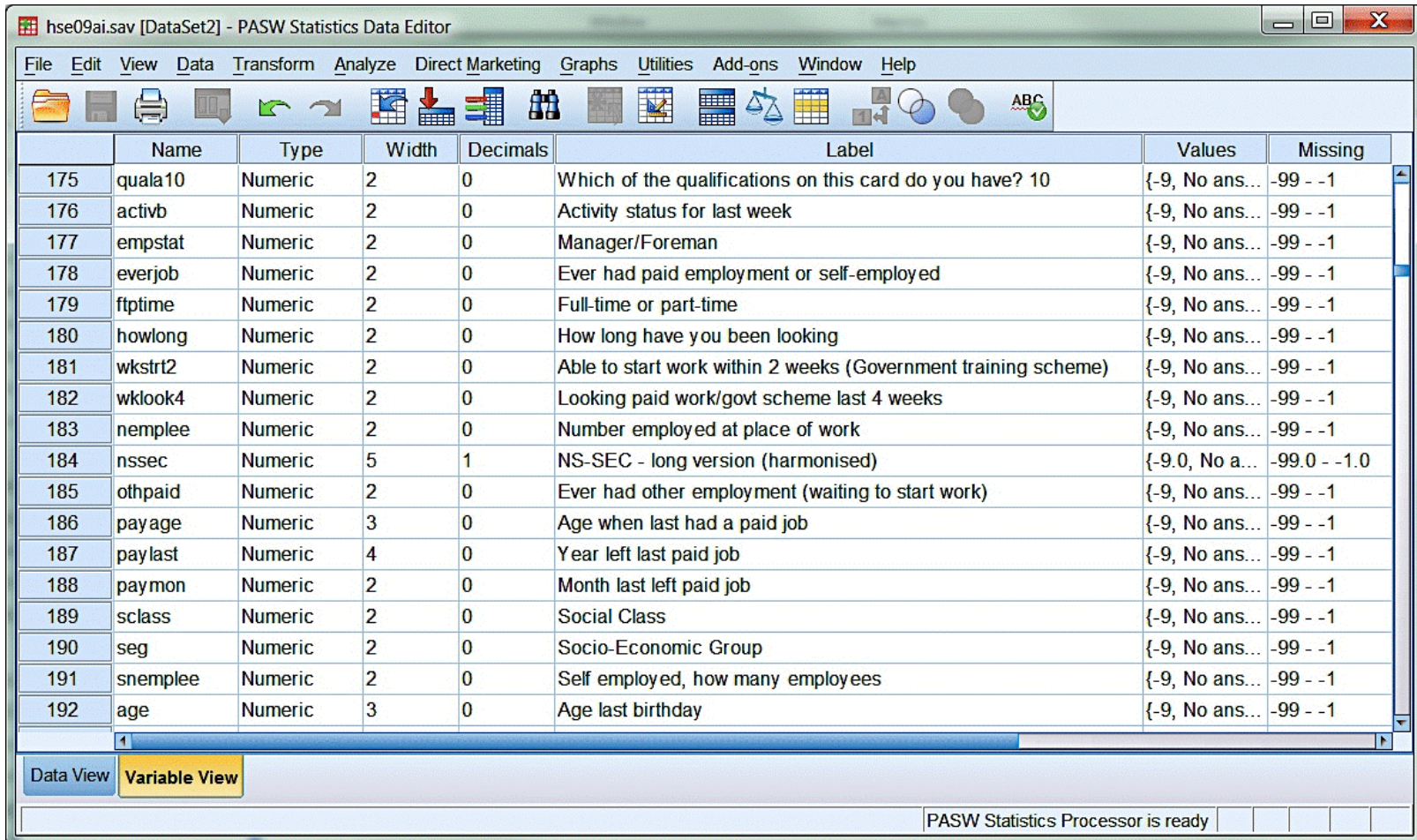


Data-level documentation: variable labels

- Similar principles for variable labels:
 - be brief, maximum of 80 characters
 - include unit of measurement where applicable
 - reference the question number of a survey or questionnaire
e.g. variable 'q11hexw' with label 'Q11: hours spent taking physical exercise in a typical week' - the label gives the unit of measurement and a reference to the question number (Q11b)
- Codes of, and reasons for, missing data
 - avoid blanks, system-missing or '0' values
e.g. '99=not recorded', '98=not provided (no answer)', '97=not applicable', '96=not known', '95=error'
- Coding or classification schemes used, with a bibliographic ref
e.g. Standard Occupational Classification 2000 - a list of codes to classify respondents' jobs; ISO 3166 alpha-2 country codes - an international standard of 2-letter country codes



Embedded data-level metadata in an SPSS file



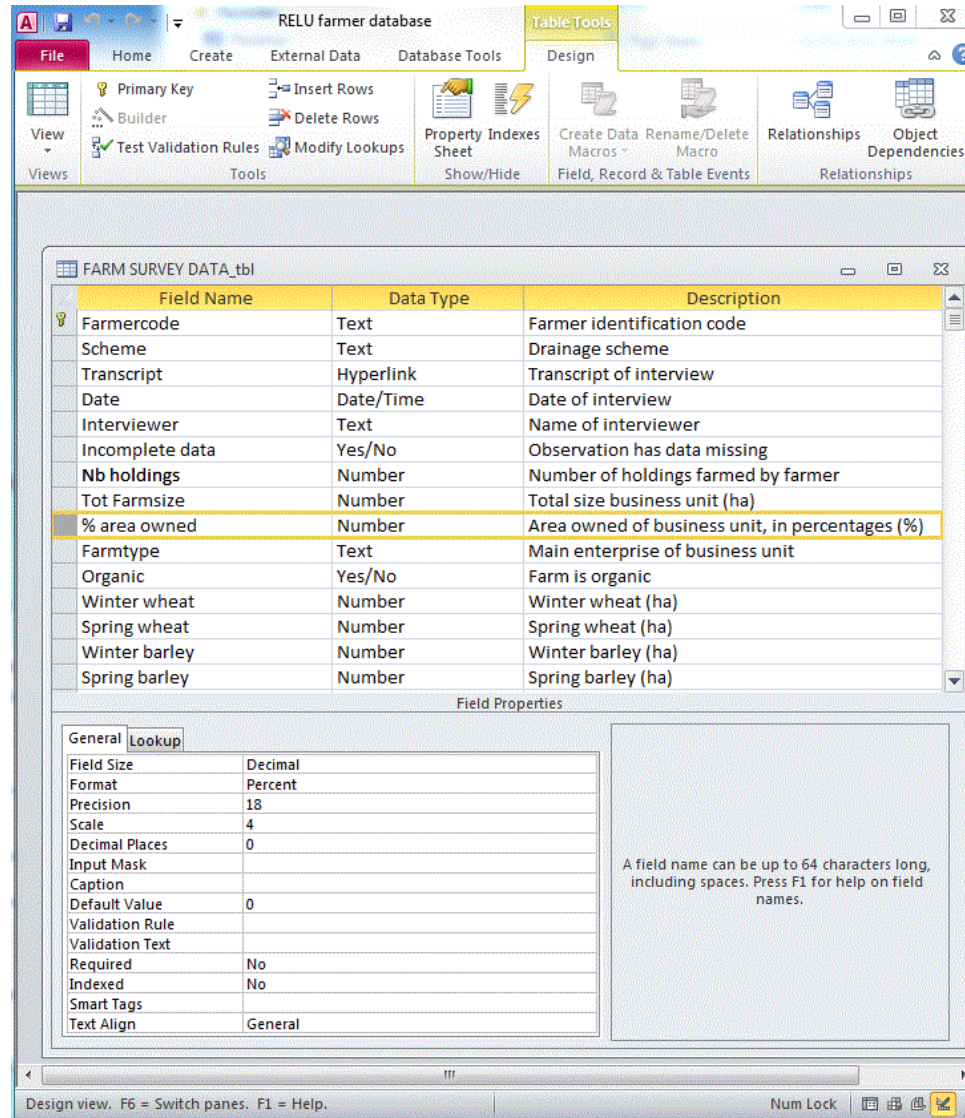
	Name	Type	Width	Decimals	Label	Values	Missing
175	quala10	Numeric	2	0	Which of the qualifications on this card do you have? 10	{-9, No ans...	-99 - -1
176	activb	Numeric	2	0	Activity status for last week	{-9, No ans...	-99 - -1
177	empstat	Numeric	2	0	Manager/Foreman	{-9, No ans...	-99 - -1
178	everjob	Numeric	2	0	Ever had paid employment or self-employed	{-9, No ans...	-99 - -1
179	ftptime	Numeric	2	0	Full-time or part-time	{-9, No ans...	-99 - -1
180	howlong	Numeric	2	0	How long have you been looking	{-9, No ans...	-99 - -1
181	wkstrt2	Numeric	2	0	Able to start work within 2 weeks (Government training scheme)	{-9, No ans...	-99 - -1
182	wklook4	Numeric	2	0	Looking paid work/govt scheme last 4 weeks	{-9, No ans...	-99 - -1
183	nemplee	Numeric	2	0	Number employed at place of work	{-9, No ans...	-99 - -1
184	nssec	Numeric	5	1	NS-SEC - long version (harmonised)	{-9.0, No a...	-99.0 - -1.0
185	othpaid	Numeric	2	0	Ever had other employment (waiting to start work)	{-9, No ans...	-99 - -1
186	payage	Numeric	3	0	Age when last had a paid job	{-9, No ans...	-99 - -1
187	paylast	Numeric	4	0	Year left last paid job	{-9, No ans...	-99 - -1
188	paymon	Numeric	2	0	Month last left paid job	{-9, No ans...	-99 - -1
189	sclass	Numeric	2	0	Social Class	{-9, No ans...	-99 - -1
190	seg	Numeric	2	0	Socio-Economic Group	{-9, No ans...	-99 - -1
191	snemplee	Numeric	2	0	Self employed, how many employees	{-9, No ans...	-99 - -1
192	age	Numeric	3	0	Age last birthday	{-9, No ans...	-99 - -1

Data View Variable View

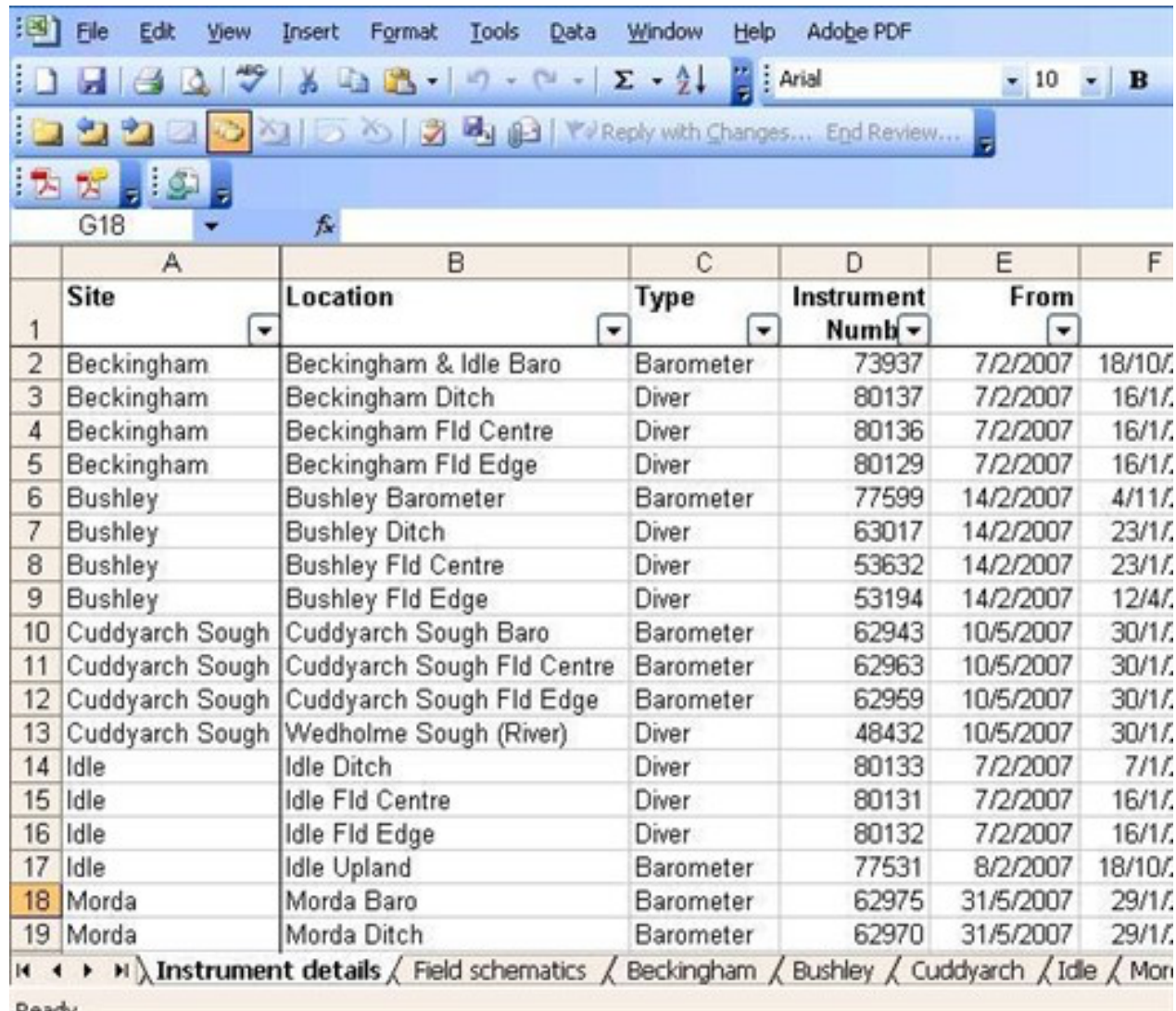
PASW Statistics Processor is ready



Embedded data-level metadata in an MS Access database file



Embedded data-level metadata in a MS Excel file



The screenshot shows a Microsoft Excel spreadsheet with a menu bar (File, Edit, View, Insert, Format, Tools, Data, Window, Help, Adobe PDF) and a toolbar. The active cell is G18. The spreadsheet contains a table with 6 columns (A-F) and 20 rows (1-19). The table has embedded metadata in the first row (row 1), which is highlighted in light blue. The metadata includes dropdown arrows for 'Site', 'Location', 'Type', 'Instrument Number', and 'From'. The data rows (rows 2-19) contain the following information:

	A	B	C	D	E	F
1	Site	Location	Type	Instrument Number	From	
2	Beckingham	Beckingham & Idle Baro	Barometer	73937	7/2/2007	18/10/.
3	Beckingham	Beckingham Ditch	Diver	80137	7/2/2007	16/1/.
4	Beckingham	Beckingham Fld Centre	Diver	80136	7/2/2007	16/1/.
5	Beckingham	Beckingham Fld Edge	Diver	80129	7/2/2007	16/1/.
6	Bushley	Bushley Barometer	Barometer	77599	14/2/2007	4/11/.
7	Bushley	Bushley Ditch	Diver	63017	14/2/2007	23/1/.
8	Bushley	Bushley Fld Centre	Diver	53632	14/2/2007	23/1/.
9	Bushley	Bushley Fld Edge	Diver	53194	14/2/2007	12/4/.
10	Cuddyarch Sough	Cuddyarch Sough Baro	Barometer	62943	10/5/2007	30/1/.
11	Cuddyarch Sough	Cuddyarch Sough Fld Centre	Barometer	62963	10/5/2007	30/1/.
12	Cuddyarch Sough	Cuddyarch Sough Fld Edge	Barometer	62959	10/5/2007	30/1/.
13	Cuddyarch Sough	Wedholme Sough (River)	Diver	48432	10/5/2007	30/1/.
14	Idle	Idle Ditch	Diver	80133	7/2/2007	7/1/.
15	Idle	Idle Fld Centre	Diver	80131	7/2/2007	16/1/.
16	Idle	Idle Fld Edge	Diver	80132	7/2/2007	16/1/.
17	Idle	Idle Upland	Barometer	77531	8/2/2007	18/10/.
18	Morda	Morda Baro	Barometer	62975	31/5/2007	29/1/.
19	Morda	Morda Ditch	Barometer	62970	31/5/2007	29/1/.

The bottom of the spreadsheet shows a navigation bar with tabs: 'Instrument details' (selected), 'Field schematics', 'Beckingham', 'Bushley', 'Cuddyarch', 'Idle', and 'Morda'.

Quality control of data

- Quality control of data
 - Integral part throughout the research project
 - During data collection
 - Data entry
 - Data checking
 - Data collection and entry
 - Calibration of instruments
 - Taking multiple measurements
 - Using standardised methods and protocols
 - Data checking
 - Checking inputted correctly
 - Checking data completeness
 - Adding variable and value labels



Metadata – data about data

- Similar to documentation in that it provides context and description, but is much more **structured** and facilitates the cataloguing and discovery of data
- Machine readable
- Standard data collection metadata includes:
 - Components of a bibliographic reference
 - Core information that a search engine indexes to make the data findable
- International standards/schemes
 - Data Documentation Initiative (DDI)
 - ISO19115 (geographic)
 - Dublin Core
 - Metadata Encoding and Transmission Standard (METS)
 - Preservation Metadata Maintenance Activity (PREMIS)



Versioning files

- Version control of files:
 - How many versions to keep? How long for?
 - It can be difficult to identify the correct version of a file if no standard naming practice is implemented
 - Major revisions vs minor revisions
 - 02-00
 - 02-01



Naming files

- Naming of files:
 - Version
 - Dates – YYYY-MM-DD (e.g. 2017-11-28)
 - Creator
 - Description of content
 - Spacing, special characters and dots – (e.g. Interview Transcript 01)
 - Interview20%Transcript20%01
 - Underscores – (e.g. Interview_Transcript_01)
 - Avoid very long names
 - Bulk file renaming

- 20130311_interview2_audio.wav
- 20130311_interview2_trans.rtf
- 20130311_interview2_image.jpg

Organising data

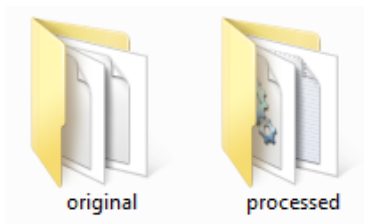
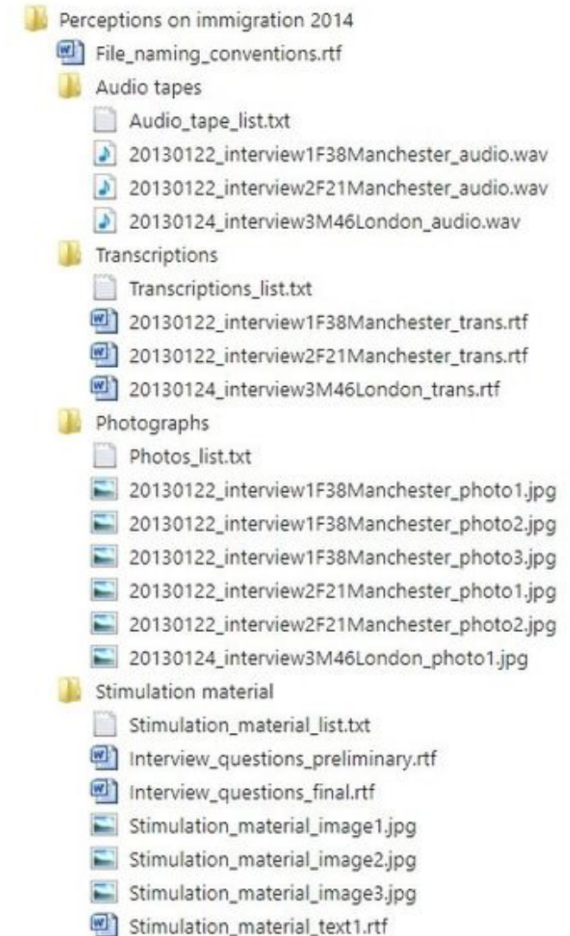
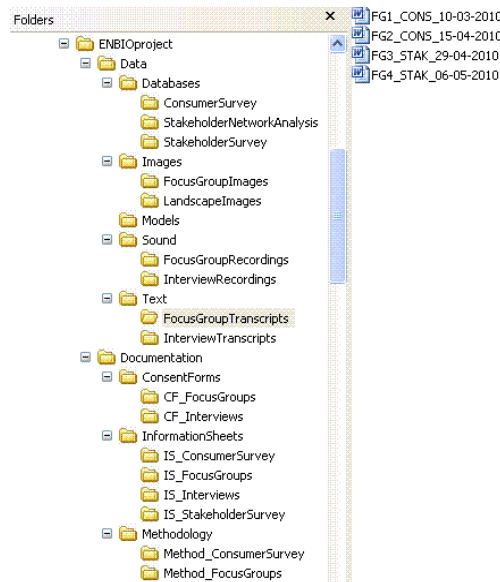
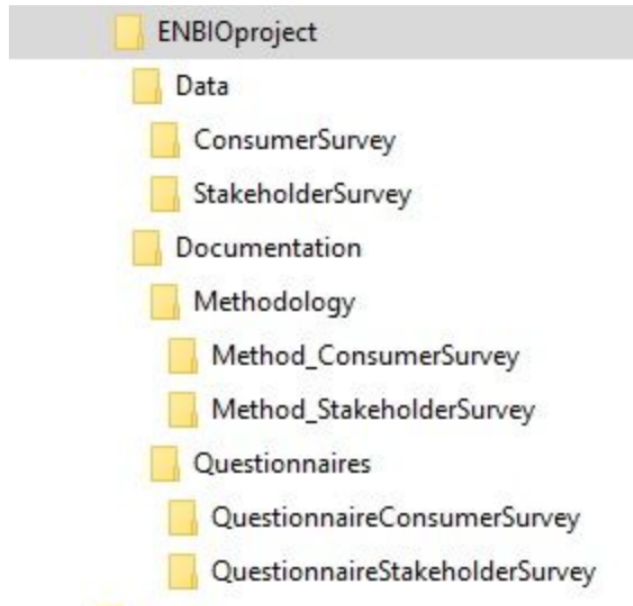
- Plan in advance how best to organise data
- Use a logical structure and ensure collaborators understand

Examples

- hierarchical structure of files, grouped in folders, e.g. audio, transcripts and annotated transcripts
- survey data: spreadsheet, SPSS, relational database
- interview transcripts: individual well-named files



Organising data examples



Recommended file formats

Documentation and scripts	Rich Text Format (.rtf) PDF/UA, PDF/A or PDF (.pdf) XHTML or HTML (.xhtml, .htm) OpenDocument Text (.odt)	plain text (.txt) widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHTML 1.0	Image data	TIFF 6.0 uncompressed (.tif)	JPEG (.jpeg, .jpg, .jp2) if original created in this format GIF (.gif) TIFF other versions (.tif, .tiff) RAW image format (.raw) Photoshop files (.psd) BMP (.bmp) PNG (.png) Adobe Portable Document Format (PDF/A, PDF) (.pdf)
Textual data	Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	Hypertext Mark-up Language (.html) widely-used formats: MS Word (.doc/.docx) some software-specific formats: NUD*IST, NVivo and ATLAS.ti			

