Documenting and contextualising your data

Research Data Management Support Services
UK Data Service
University of Essex

April 2014

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 of coverage

- Why documentation is important
- Study-level documentation and context
- Data-level documentation
- Metadata
- Context debate

What is documentation?

Data doesn’t 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 provides background information
- description of the methodology used to collect the data

Why document your data?

- Enables you to understand/interpret data when you return to it
- Needed to make data independently understandable i.e. reusable
- Helps avoid incorrect use/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 Service 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 should be captured?

Any useful documentation such as:

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

Information on dataset structure

- inventory of data files
- relationships between those files
- records, cases…

Variable-level documentation

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

What should be captured?

Contextual information about project and data

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

Data collection methodology and processes

- data collection process and sampling
- instruments used - questionnaires, showcards, interview schedules
- temporal/geographic coverage
- data validation - cleaning, error-checking
- compilation of derived variables
- weighting: factors and variables, weighting process
- secondary data sources used

Data confidentiality, access and use conditions

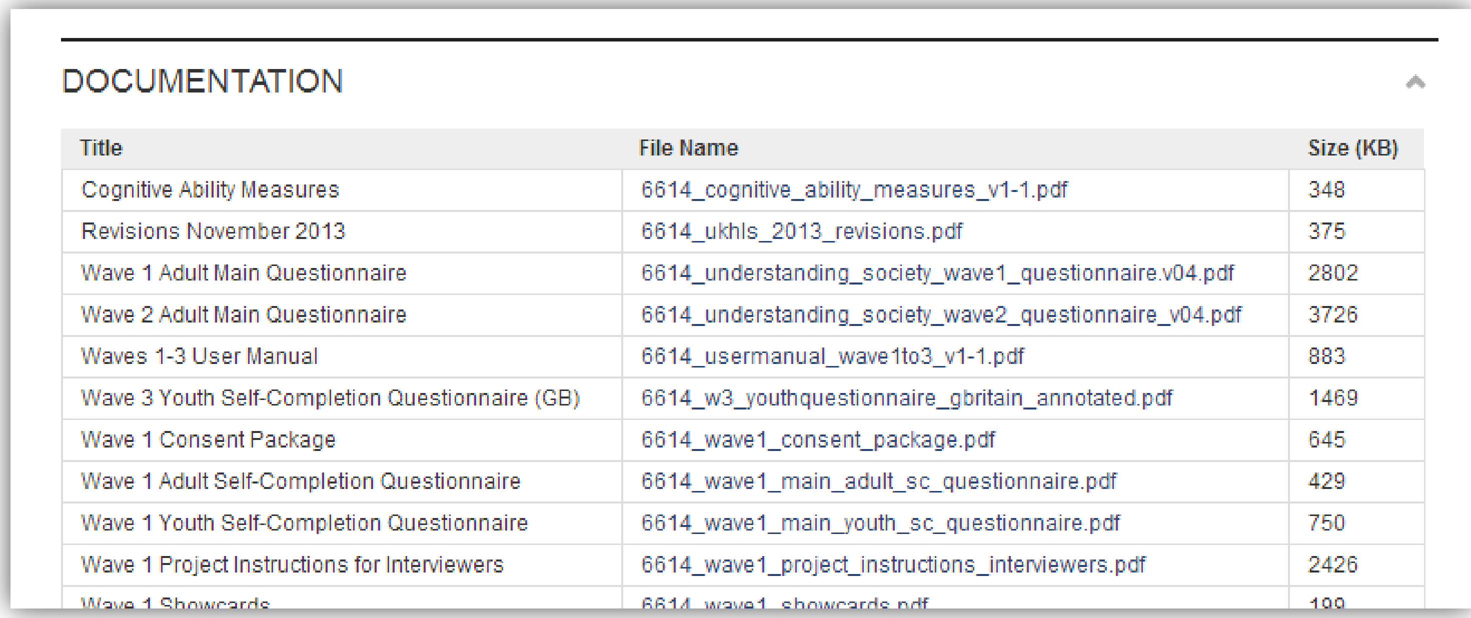
- anonymisation carried out
- consent conditions/procedures
- access or use conditions of data

Consider documentation early on

- 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 (which you will hear more on later in the course)

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:

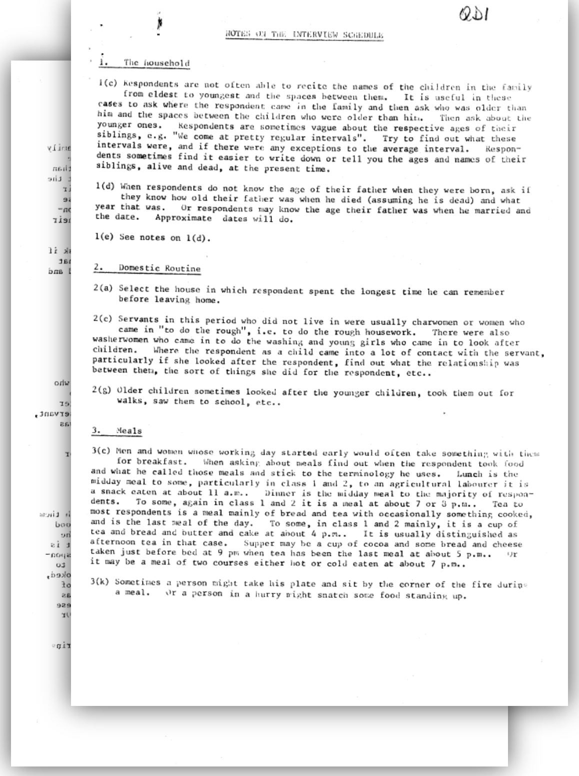


DOCUMENTATION

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 – user guide and doc

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



QD1

NOTES ON THE INTERVIEW SCHEDULE

1. The household

1(c) Respondents are not often able to recite the names of the children in the family from eldest to youngest and the spaces between them. It is useful in these cases to ask where the respondent came in the family and then ask who was older than him and the spaces between the children who were older than him. Then ask about the younger ones. Respondents are sometimes vague about the respective ages of their siblings, e.g. "We come at pretty regular intervals". Try to find out what these intervals were, and if there were any exceptions to the average interval. Respondents sometimes find it easier to write down or tell you the ages and names of their siblings, alive and dead, at the present time.

1(d) When respondents do not know the age of their father when they were born, ask if they know how old their father was when he died (assuming he is dead) and what year that was. Or respondents may know the age their father was when he married and the date. Approximate dates will do.

1(e) See notes on 1(d).

2. Domestic Routine

2(a) Select the house in which respondent spent the longest time he can remember before leaving home.

2(c) Servants in this period who did not live in were usually charwomen or women who came in "to do the rough", i.e. to do the rough housework. There were also washerwomen who came in to do the washing and young girls who came in to look after children. Where the respondent as a child came into a lot of contact with the servant, particularly if she looked after the respondent, find out what the relationship was between them, the sort of things she did for the respondent, etc..

2(g) Older children sometimes looked after the younger children, took them out for walks, saw them to school, etc..

3. Meals

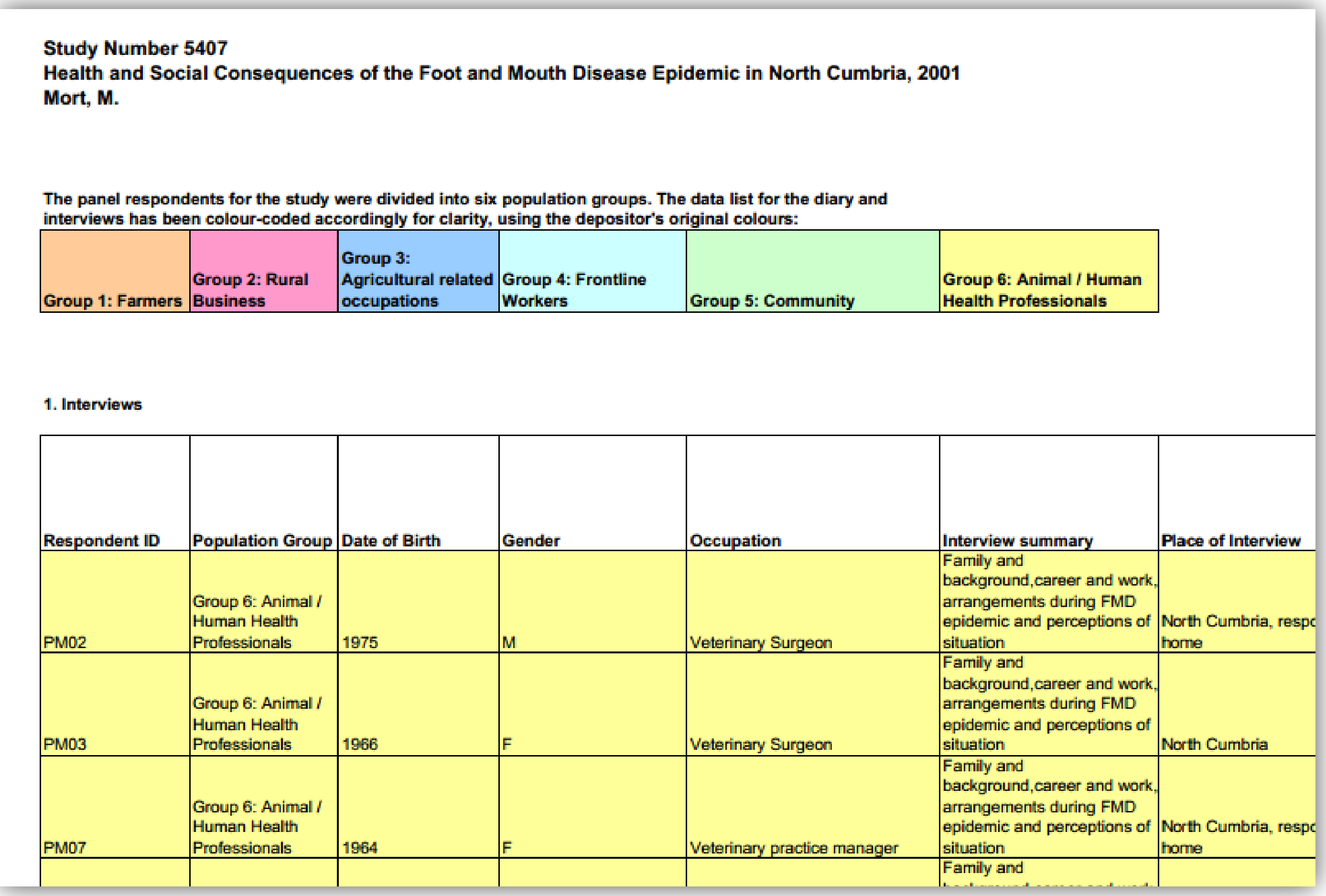
3(c) Men and women whose working day started early would often take something with them for breakfast. When asking about meals find out when the respondent took food and what he called those meals and stick to the terminology he uses. Lunch is the midday meal to some, particularly in class 1 and 2, to an agricultural labourer it is a snack eaten at about 11 a.m.. Dinner is the midday meal to the majority of respondents. To some, again in class 1 and 2 it is a meal at about 7 or 8 p.m.. Tea to most respondents is a meal mainly of bread and tea with occasionally something cooked, and is the last meal of the day. To some, in class 1 and 2 mainly, it is a cup of tea and bread and butter and cake at about 4 p.m.. It is usually distinguished as afternoon tea in that case. Supper may be a cup of cocoa and some bread and cheese taken just before bed at 9 pm when tea has been the last meal at about 5 p.m.. Or it may be a meal of two courses either hot or cold eaten at about 7 p.m..

3(k) Sometimes a person might take his plate and sit by the corner of the fire during a meal. Or a person in a hurry might snatch some food standing up.

UK Data Service

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

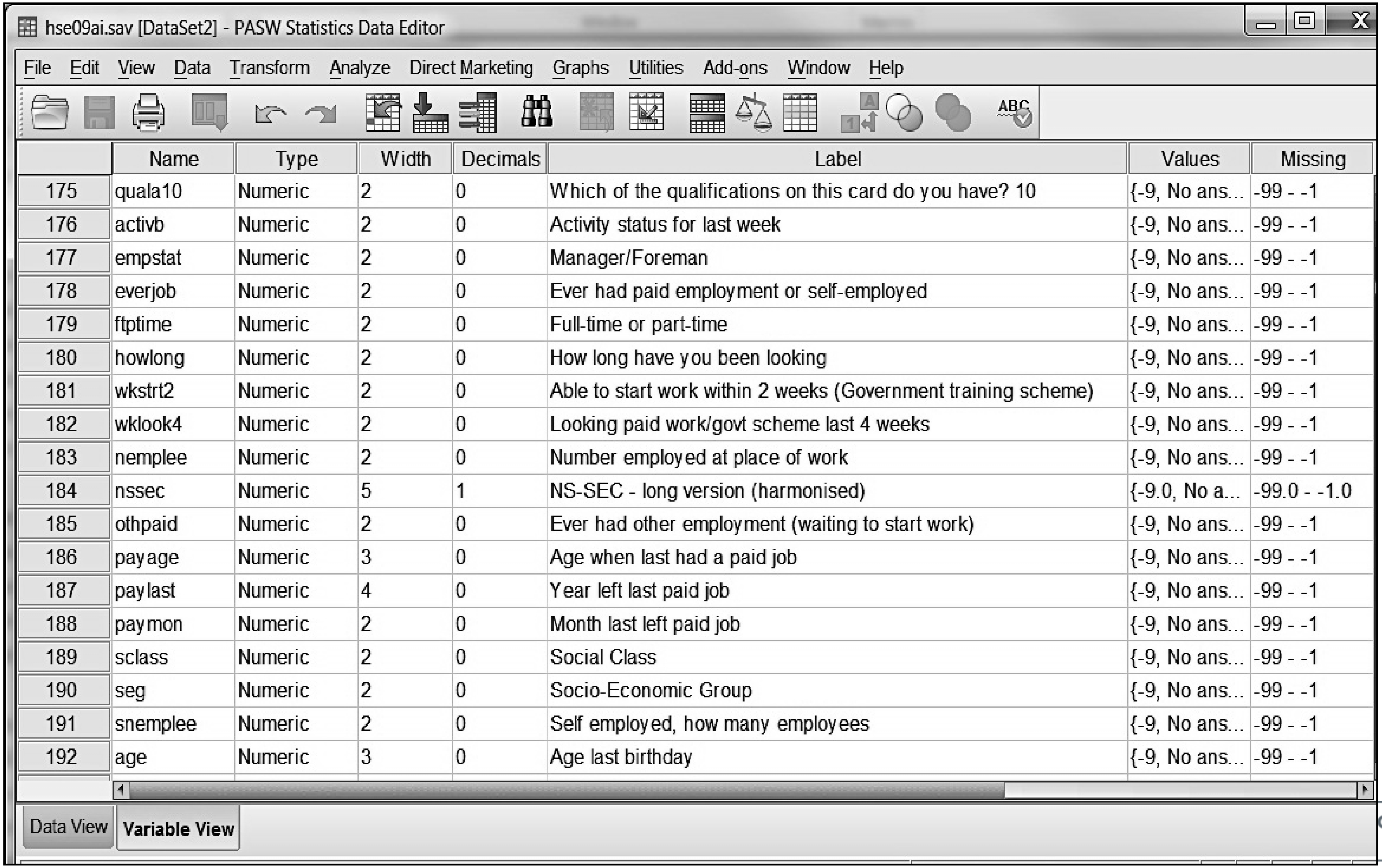
Respondent ID	Population Group	Date of Birth	Gender	Occupation	Interview summary	Place of Interview
PM02	Group 6: Animal / Human Health Professionals	1975	M	Veterinary Surgeon	Family and background,career and work, arrangements during FMD epidemic and perceptions of situation	North Cumbria, respo home
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, respo home
					Family and background,career and work,	

Data-level documentation

- Certain types of data file may contain important information which should be preserved:
  - variable/value labels; document metadata; table relationships and queries in relational databases; GIS data layers/tables

- Some examples:
  - SPSS: variable attributes documented in Variable View (label, code, data type, missing values)
  - MS Access: relationships between tables
  - ArcGIS: shapefiles (layers) and tables in geodatabase; metadata created in ArcCatalog
  - MS Excel: document properties, worksheet labels (where multiple)

Embedded data-level metadata in 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-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, moocc=mother occupation, faocc=father occupation
  - for interoperability across platforms - variable names should be max 8 characters and without spaces

Data-level documentation: variable labels

- Similar principles for variable labels:
  - be brief, max. 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

Data-level documentation: transcripts

- Qualitative data/text documents:
  - interview transcript speech demarcation (speaker tags)
  - document header with brief details of interview date, place, interviewer name, interviewee details, context

Metadata – data about data

- In some ways, just another kind of documentation
- But much more highly structured

- 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
  - Dublin Core
  - Metadata Encoding and Transmission Standard (METS)
  - Preservation Metadata Maintenance Activity (PREMIS)

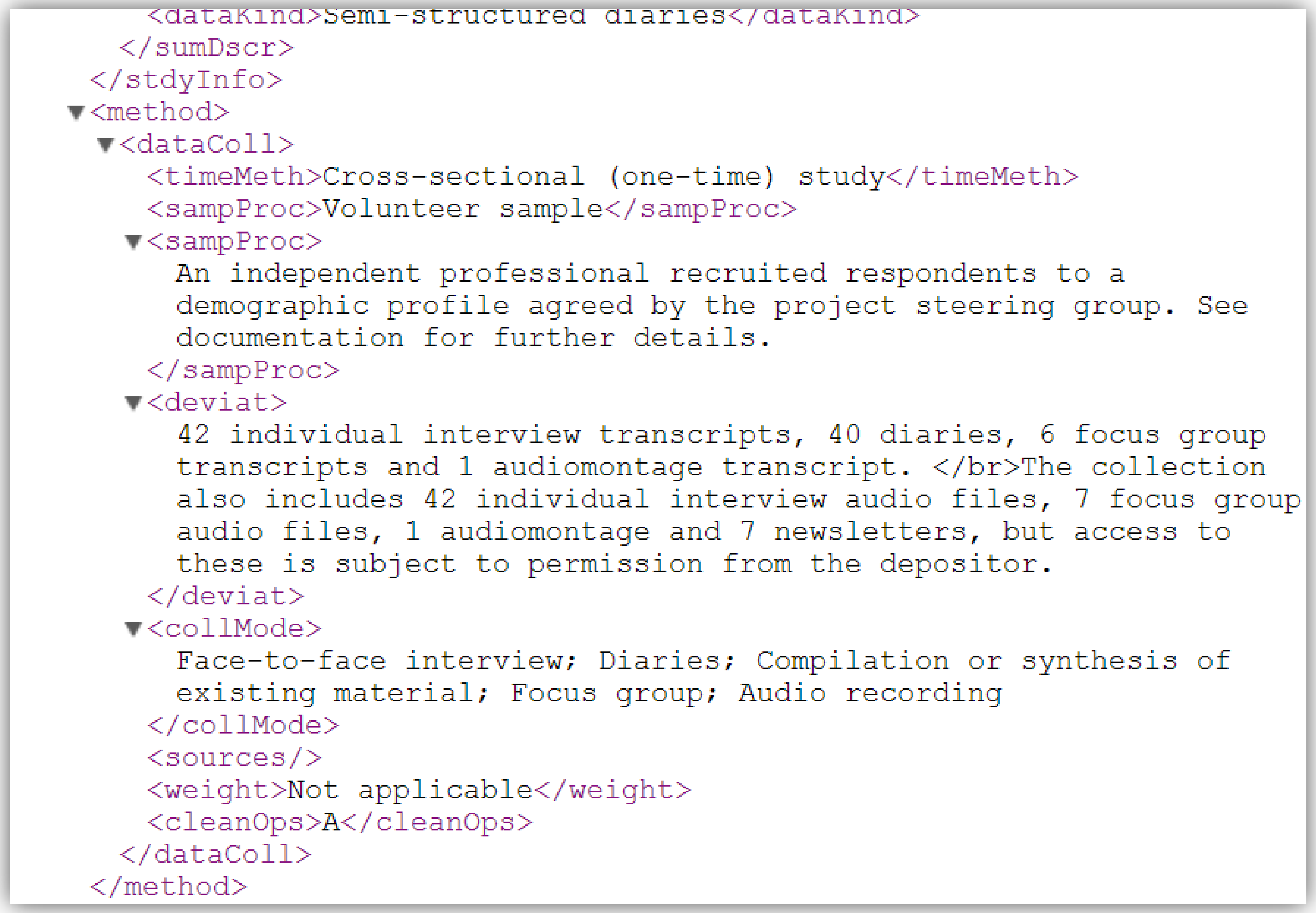
Metadata at the UK Data Archive

- Metadata for archived datasets at should include:
  - Core fields: title, abstract, details of data owner/creator
  - Administrative: Funding information source and award number, copyright holder
  - Detailed descriptive info: temporal coverage (data collection start and end dates), geographic coverage (country, region, longitude/latitude), keywords and subject categories
  - Methodological: sample size/units, methodology
  - Data availability/access conditions
  - Publications/references
  - Digital Object Identifier (DOI)

- Created from data deposit form/tool and information/documentation submitted by data owners/researchers
- UK Data Service: DDI metadata, rich detailed content

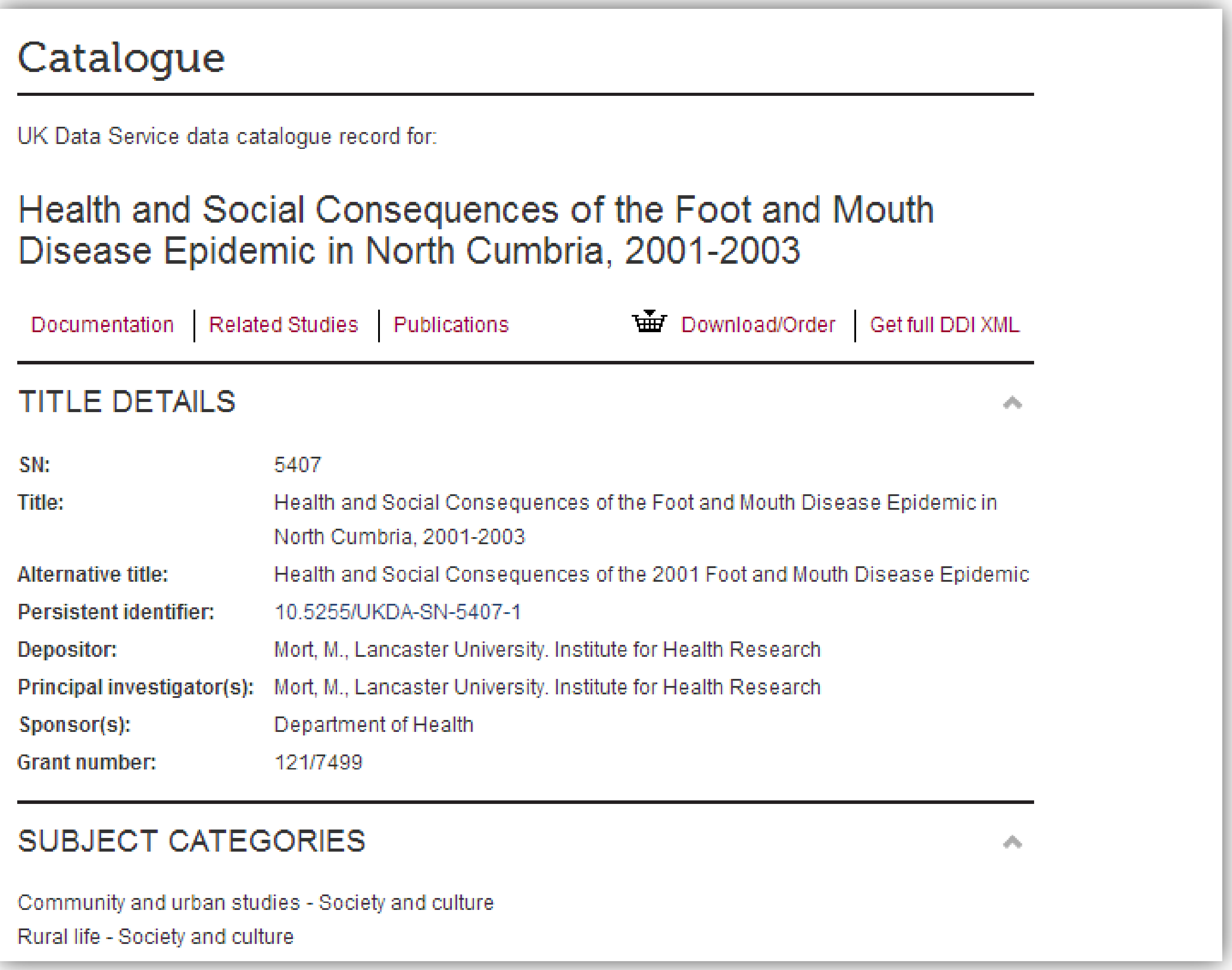
http://ukdataservice.ac.uk/manage-data/document/metadata.aspx

Study DDI XML metadata



```
      <dataKind>Semi-structured diaries</dataKind>
    </sumDscr>
  </stdyInfo>
▼<method>
  ▼<dataColl>
      <timeMeth>Cross-sectional (one-time) study</timeMeth>
      <sampProc>Volunteer sample</sampProc>
    ▼<sampProc>
        An independent professional recruited respondents to a
        demographic profile agreed by the project steering group. See
        documentation for further details.
      </sampProc>
    ▼<deviat>
        42 individual interview transcripts, 40 diaries, 6 focus group
        transcripts and 1 audiomontage transcript. </br>The collection
        also includes 42 individual interview audio files, 7 focus group
        audio files, 1 audiomontage and 7 newsletters, but access to
        these is subject to permission from the depositor.
      </deviat>
    ▼<collMode>
        Face-to-face interview; Diaries; Compilation or synthesis of
        existing material; Focus group; Audio recording
      </collMode>
      <sources/>
      <weight>Not applicable</weight>
      <cleanOps>A</cleanOps>
    </dataColl>
  </method>
```

Study DDI catalogue record



Catalogue

UK Data Service data catalogue record for:

Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003

TITLE DETAILS

SN:	5407
Title:	Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003
Alternative title:	Health and Social Consequences of the 2001 Foot and Mouth Disease Epidemic
Persistent identifier:	10.5255/UKDA-SN-5407-1
Depositor:	Mort, M., Lancaster University. Institute for Health Research
Principal investigator(s):	Mort, M., Lancaster University. Institute for Health Research
Sponsor(s):	Department of Health
Grant number:	121/7499

SUBJECT CATEGORIES

Community and urban studies - Society and culture

Rural life - Society and culture

How to create metadata for data

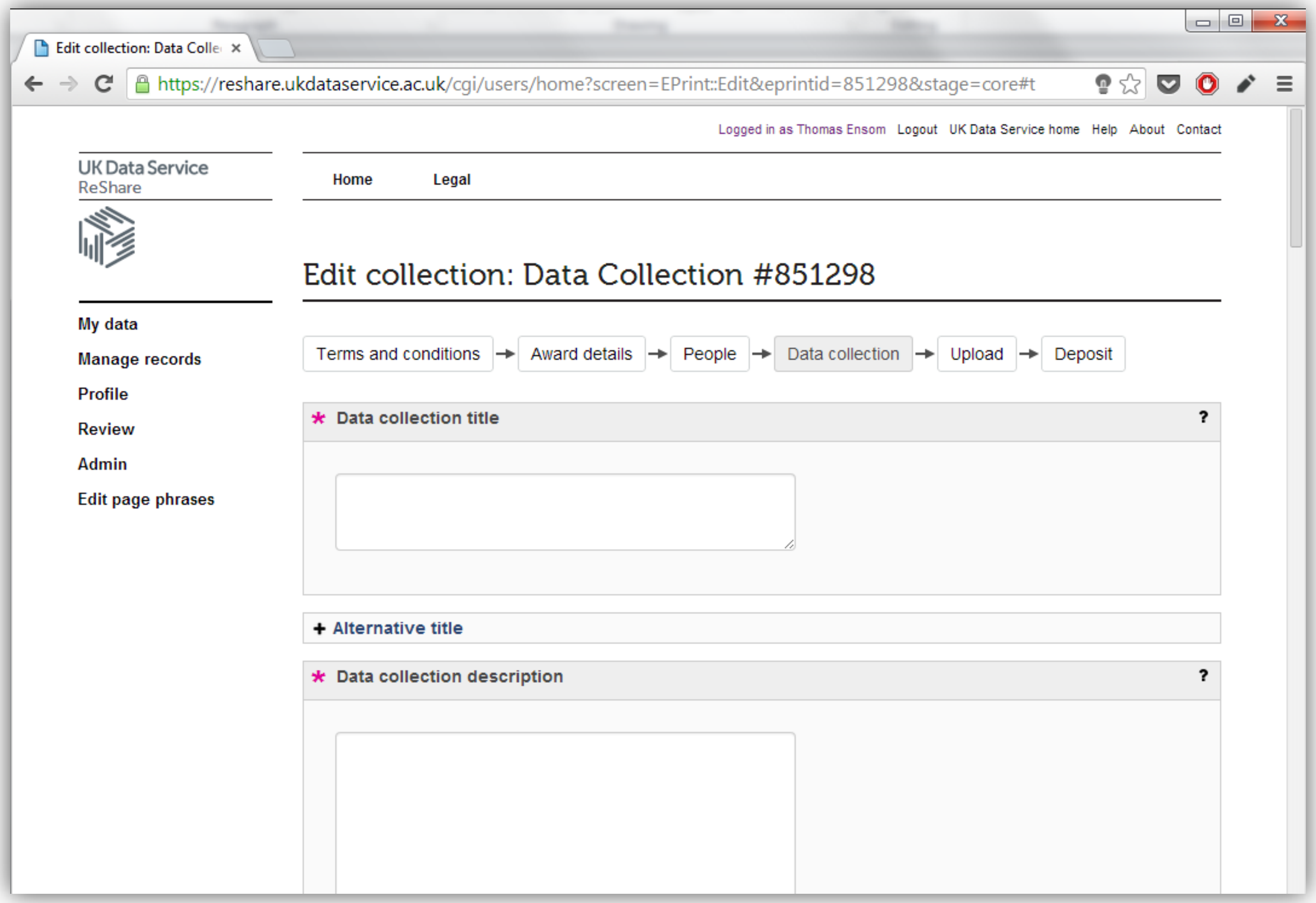
- Can be compiled using data deposit forms/tools
- Currently not many available that are user friendly and maintained
- May be better to take things into your own hands – create a spreadsheet!

- Data Documentation Initiative (DDI) documentation can be created in software packages using certain DDI tools: tools.ddialliance.org
  - Colectica Designer for survey data

    http://www.colectica.com/software/designer
  - Convert SPSS internal metadata to DDI using Nesstar Publisher

    www.nesstar.com/software/publisher.html
  - German Institute for Educational Progress (IQB) – educational data codebooks www.iza.org

Metadata entry for UK Data Service ReShare



Edit collection: Data Colle

https://reshare.ukdataservice.ac.uk/cgi/users/home?screen=EPrint::Edit&eprintid=851298&stage=core#t

Logged in as Thomas Ensom Logout UK Data Service home Help About Contact

UK Data Service
ReShare

Home Legal

My data
Manage records
Profile
Review
Admin
Edit page phrases

Edit collection: Data Collection #851298

Terms and conditions → Award details → People → Data collection → Upload → Deposit

* Data collection title ?

+ Alternative title

* Data collection description ?

UK Data Service

Exercise – how to document this data?

You carry out research on the public understanding of climate change and associated risks in the UK. Your data-generating research consists of:

- An online survey with 2000 invited members of the public in the UK to assess their understanding of climate change and climate change risks, as well as their sources of information.
- Interviews with 20 key stakeholders in climate policy and science communication.
- Qualitative content analysis of secondary data taken from newspapers and popular science journals, evaluating reporting about climate change in the media.

Data resulting from the online survey are transferred to SPSS for analysis.

Interviews are audio-recorded and transcribed into MS Word. Transcripts are imported into the NVivo software for content analysis.

Secondary textual data from newspapers and journals are also imported into NVivo for content analysis.

How would you document your resulting research data, to enable their future use by other researchers?

Exercise – some possible answers

For the survey data file, ensure that the SPSS file contains the full question text as variable label for each corresponding variable, or as detailed a description of the question text as possible. Variable names should consist of meaningful codes. Variable attributes are clearly defined, complete and without any abbreviations, explaining codes, categories and missing data values for each variable.

The online survey questionnaire is exported as a PDF file to complement the SPSS data file.

Each interview transcript contains an introductory paragraph providing the context and setting for the interview. The collection of 20 transcripts is accompanied by a data listing. Alternatively in NVivo a classification is created for all interviewees and interviews, capturing: for interviewees', relevant demographic and background characteristics (identifier or pseudonym) such as gender, age, profession, organisation and communication medium used; and for interviews, date, place and interviewer name.

A list or table of bibliographic references contains all sources of information used for the secondary analysis of media content.

A published article provides background information on the research methods used, sampling and so on.

Contacts

Collections Development team

UK Data Service

University of Essex

datasharing@ukdataservice.ac.uk

UK Data Service