Documenting and contextualising your data

Research Data Management Support Services
UK Data Service
University of Essex

April 2014
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:
Qualitative study – user guide and doc

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

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

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Width</th>
<th>Decimals</th>
<th>Label</th>
<th>Values</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>qua10</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Which of the qualifications on this card do you have? 10</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>activb</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Activity status for last week</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>empstat</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Manager/Foreman</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>everjob</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Ever had paid employment or self-employed</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>ftptime</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Full-time or part-time</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>howlong</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>How long have you been looking</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>wkstr2</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Able to start work within 2 weeks (Government training scheme)</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>wklook4</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Looking paid work/govt scheme last 4 weeks</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>nempllee</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Number employed at place of work</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>nsec</td>
<td>Numeric</td>
<td>5</td>
<td>1</td>
<td>NS-SEC - long version (harmonised)</td>
<td>-9.0, No a...</td>
<td>99.0 -1.0</td>
</tr>
<tr>
<td>othpaid</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Ever had other employment (waiting to start work)</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>payage</td>
<td>Numeric</td>
<td>3</td>
<td>0</td>
<td>Age when last had a paid job</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>paylast</td>
<td>Numeric</td>
<td>4</td>
<td>0</td>
<td>Year left last paid job</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>paymon</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Month last left paid job</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>sclass</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Social Class</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>seg</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Socio-Economic Group</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>snemplle</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Self employed, how many employees</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
<tr>
<td>age</td>
<td>Numeric</td>
<td>3</td>
<td>0</td>
<td>Age last birthday</td>
<td>-9, No ans...</td>
<td>99 - 1</td>
</tr>
</tbody>
</table>
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

<dataKind>Semi-structured diaries</dataKind>
</sumDscr>
</studyInfo>
▼<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>
<source/>
<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

DOCUMENTATION | RELATED STUDIES | PUBLICATIONS

Download/Order | Get full DDI XML

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 Collection #851298

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

Data collection title

Alternative title

Data collection description
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