Documenting and contextualising your data

Louise Bolger
UK Data Archive

Looking after and managing your research data:
An advanced course
18-19 June 2015
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
- Study-level documentation and context
- Data-level documentation
- Metadata
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
  • Consent forms and information sheets
  • ReadMe file
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 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 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
- secondary data sources used

**Any useful documentation such as:**
- final report, published reports, user guide, working paper, publications, lab books
What should be captured?

**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

**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:

<table>
<thead>
<tr>
<th>Title</th>
<th>File Name</th>
<th>Size (KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Ability Measures</td>
<td>6614_cognitive_ability_measures_v1-1.pdf</td>
<td>348</td>
</tr>
<tr>
<td>Revisions November 2013</td>
<td>6614_ukhls_2013_revisions.pdf</td>
<td>375</td>
</tr>
<tr>
<td>Wave 1 Adult Main Questionnaire</td>
<td>6614_understanding_society_wave1_questionnaire_v04.pdf</td>
<td>2802</td>
</tr>
<tr>
<td>Wave 2 Adult Main Questionnaire</td>
<td>6614_understanding_society_wave2_questionnaire_v04.pdf</td>
<td>3726</td>
</tr>
<tr>
<td>Waves 1-3 User Manual</td>
<td>6614_usermanual_wave1to3_v1-1.pdf</td>
<td>883</td>
</tr>
<tr>
<td>Wave 3 Youth Self-Completion Questionnaire (GB)</td>
<td>6614_w3_youthquestionnaire_gbritain_annotated.pdf</td>
<td>1469</td>
</tr>
<tr>
<td>Wave 1 Consent Package</td>
<td>6614_wave1_consent_package.pdf</td>
<td>645</td>
</tr>
<tr>
<td>Wave 1 Adult Self-Completion Questionnaire</td>
<td>6614_wave1_main_adult_sc_questionnaire.pdf</td>
<td>429</td>
</tr>
<tr>
<td>Wave 1 Youth Self-Completion Questionnaire</td>
<td>6614_wave1_main_youth_sc_questionnaire.pdf</td>
<td>750</td>
</tr>
<tr>
<td>Wave 1 Project Instructions for Interviewers</td>
<td>6614_wave1_project_instructions_interviewers.pdf</td>
<td>2426</td>
</tr>
<tr>
<td>Wave 1 Showcards</td>
<td>6614_wave1_showcards.pdf</td>
<td>100</td>
</tr>
</tbody>
</table>
A user guide could contain a variety of documents that provide context: interview schedule, transcription notes, even photos.
### In practice: transcript format

<table>
<thead>
<tr>
<th>Study Name:</th>
<th>Interview number:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depositor:</td>
<td>Interview ID: Firstname Lastname</td>
</tr>
<tr>
<td>Interviewer:</td>
<td>Date of interview:</td>
</tr>
</tbody>
</table>

#### Information about interviewee

<table>
<thead>
<tr>
<th>Date of birth:</th>
<th>Marital status:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender:</td>
<td>Occupation:</td>
</tr>
<tr>
<td>Geographic region:</td>
<td></td>
</tr>
</tbody>
</table>

### 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

<table>
<thead>
<tr>
<th>Study Number 5407</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001</td>
</tr>
<tr>
<td>Mort, M.</td>
</tr>
</tbody>
</table>

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

<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, responder 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>North 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, responder home</td>
</tr>
</tbody>
</table>
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, 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, context
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>quala10</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Which of the qualifications on this card do you have? 10</td>
<td>175, 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>176, 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>177, 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>178, 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>179, 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>180, No ans</td>
<td>99 - 1</td>
</tr>
<tr>
<td>wkstr12</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Able to start work within 2 weeks (Government training scheme)</td>
<td>181, 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>182, No ans</td>
<td>99 - 1</td>
</tr>
<tr>
<td>nemplee</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Number employed at place of work</td>
<td>183, No ans</td>
<td>99 - 1</td>
</tr>
<tr>
<td>nssec</td>
<td>Numeric</td>
<td>5</td>
<td>1</td>
<td>NS-SEC - long version (harmonised)</td>
<td>184, 90.0</td>
<td>99 - 1</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>185, 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>186, 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>187, 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>188, 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>189, 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>190, No ans</td>
<td>99 - 1</td>
</tr>
<tr>
<td>snemplee</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Self employed, how many employees</td>
<td>191, 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>192, 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, motoc=mother occupation, fatoc=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
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)
  • Dublin Core
Metadata at the UK Data Archive

- Metadata for archived datasets at UKDA 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

Metadata entry for UK Data Service ReShare
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
Questions?

Contact details:
Louise Bolger
UK Data Archive

datasharing@ukdataservice.ac.uk
http://ukdataservice.ac.uk/manage-data.aspx