Documenting and Providing Context for Data

Scott Summers
UK Data Service
University of Essex

Cancer Research UK Data Sharing and Management
24th May 2017
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
• Metadata
• ReShare metadata
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
• 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

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

• 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:

<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>3728</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_gbrazil_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>199</td>
</tr>
</tbody>
</table>
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:

|------------------|-------------------------|------------------------------------------|---------------------------|-------------------|---------------------------------------------|

<table>
<thead>
<tr>
<th>1. Interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Respondent ID</strong></td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>PM02</td>
</tr>
<tr>
<td>PM03</td>
</tr>
<tr>
<td>PM07</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 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
## 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>(-9, No answer)</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 answer)</td>
<td>-9 - 1</td>
</tr>
<tr>
<td>empslat</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Manager/Foreman</td>
<td>(-9, No answer)</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 answer)</td>
<td>-9 - 1</td>
</tr>
<tr>
<td>fptime</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Full-time or part-time</td>
<td>(-9, No answer)</td>
<td>-9 - 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 answer)</td>
<td>99 - 1</td>
</tr>
<tr>
<td>wkstrl2</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 answer)</td>
<td>-9 - 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 answer)</td>
<td>-9 - 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 answer)</td>
<td>-9 - 1</td>
</tr>
<tr>
<td>nssec</td>
<td>Numeric</td>
<td>5</td>
<td>1</td>
<td>NS-SEC - long version (harmonised)</td>
<td>(-9.0, No answer)</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 answer)</td>
<td>-9 - 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 answer)</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 answer)</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 answer)</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 answer)</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 answer)</td>
<td>99 - 1</td>
</tr>
<tr>
<td>snempllee</td>
<td>Numeric</td>
<td>2</td>
<td>0</td>
<td>Self employed, how many employees</td>
<td>(-9, No answer)</td>
<td>-9 - 1</td>
</tr>
<tr>
<td>age</td>
<td>Numeric</td>
<td>3</td>
<td>0</td>
<td>Age last birthday</td>
<td>(-9, No answer)</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’ and ‘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, 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' and '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

- Similar to documentation in that it provides context and description, but is much more 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 (geographic)
  - Dublin Core
  - Metadata Encoding and Transmission Standard (METS)
  - Preservation Metadata Maintenance Activity (PREMIS)
ReShare Metadata (1)
## Temporal coverage

Provide a start and end date of either the actual data collection period, or else the time period that is covered by the data collection.

**Collection period:** Start and end date for the data collection period.

<table>
<thead>
<tr>
<th>Date from</th>
<th>Date to</th>
</tr>
</thead>
</table>

**Temporal coverage:** The time period (start and end date) for this data collection, if it is different from the data collection period, e.g. for data representing a historical period.

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
</tr>
</thead>
</table>

## Geographical coverage

Provide information about the geographical area where your data were collected.

**Geographical area:** The name(s) of the location(s) or area(s) where your data were collected.

**Bounding area:** Define the geographical area covered by your data collection by four latitude/longitude coordinates in decimal degrees, as the smallest rectangular shape that totally encloses all of the data. Example: 60.85, 2.69, 49.84, -9.23 (bounding area for UK)

<table>
<thead>
<tr>
<th>North latitude</th>
<th>East longitude</th>
<th>South latitude</th>
<th>West longitude</th>
</tr>
</thead>
</table>

**Country:** The country or countries where the data were collected. When you start typing text, a controlled list of countries to select from will appear.

1. United Kingdom
2. 
3. 

[Add row]
Exercises and practical examples
Questions

Contact details:

Collections Development and Producer Relations team
UK Data Service
University of Essex

ukdataservice.ac.uk/help/get-in-touch