Elements of Data Documentation

Adam Mack
Education and Human Development Incubator
Social Science Research Institute
March 3, 2015
Why Is Documentation Important?

- Describe the contents of the data
- Explain context in which data was collected
- Explain any manipulations performed on the data
- Allow research data to be understood by people outside of the original project
Do I Need to Document?

Research:
Back in the day... ... and now.
Consequences of Insufficient Documentation

I FIND YOUR LACK OF DOCUMENTATION

DISTURBING
Consequences of Insufficient Documentation

• Data may be unusable
• May make inaccurate assumptions about data
  – May be unclear how to interpret contents of a variable
  – Manipulations performed on data may affect results of analyses
Elements of Data Documentation

• What are the most important elements to document?
Elements of Data Documentation

• What are the most important elements to document?
  – Data elements
  – Study elements
  – Processes and decisions
Elements of Data Documentation

• Who will be using the documentation?
  – Data managers
  – Statisticians
  – Researchers
  – Outside users
Elements of Data Documentation

• When should documentation be created?
  – Often, projects wait until data has been collected before creating documentation such as codebooks.
  – Creating documentation early in the project has numerous advantages.
Elements of Data Documentation

• How should these elements be documented? Potential forms that documentation may take include:
  – Codebook
  – Annotated version of instrument
  – More descriptive, less structured forms of documentation (data narratives)
Data-Level Documentation

• What are the most important elements to document?
  – Data elements
  – Study elements
  – Processes and decisions
Data-Level Documentation

Should include basic information needed to use the data, including:

• Structural information about variable
  – Name of variable
  – Label (if applicable)
  – Type of variable (numeric or character)
  – Length of variable
Data-Level Documentation

• Information describing variable contents
  – Question text (or text description of variable contents)
  – Valid values
  – Coding of values
Data-Level Documentation

• Scales/derived variables
  – Algorithm used to create variable
  – Procedures for handling missing data
Data-Level Documentation

• Question routing (if skip patterns used)
  – Identify number of participants asked each question/path through survey

• Error checking/validation
Data-Level Documentation

• Reliability of scales
  – Calculate Cronbach’s alpha for each scale included in the data
  – Compare values for your study to previously reported values in the literature
Types of Data Documentation

• Tabular codebook (Excel)
  – Good for organizing a large amount of information concisely
  – Sortable
  – Filterable
  – Customizable; can hide columns that may be needed but are not of interest to a general audience
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Variable Name</td>
<td>Measure</td>
<td>Type</td>
<td>Years present</td>
<td>Length</td>
<td>Values</td>
<td>Question verbatim</td>
</tr>
<tr>
<td>2</td>
<td>ID</td>
<td>Demographics</td>
<td>Numeric</td>
<td>Created by project</td>
<td>8</td>
<td>5-digit number. Cohort 1 IDs follow the format 1####, Cohort 2 IDs follow the format 2####</td>
<td>Student ID</td>
</tr>
<tr>
<td>3</td>
<td>race</td>
<td>Demographics</td>
<td>Numeric</td>
<td>Collected once</td>
<td>3</td>
<td>1= Black; 2= White; 3= Hispanic; 4= Multi-Ethnic, 5= Other, 6=Asian</td>
<td>Race/ Ethnicity</td>
</tr>
<tr>
<td>4</td>
<td>gender</td>
<td>Demographics</td>
<td>Numeric</td>
<td>Collected once</td>
<td>3</td>
<td>1= Male; 2= Female</td>
<td>Gender</td>
</tr>
<tr>
<td>5</td>
<td>cohort</td>
<td>Demographics</td>
<td>Numeric</td>
<td>Collected once</td>
<td>3</td>
<td>1 = Cohort 1, 2 = Cohort 2</td>
<td>Data cohort</td>
</tr>
<tr>
<td>6</td>
<td>ethident_01</td>
<td>Ethnic Identity</td>
<td>Numeric</td>
<td>3-5</td>
<td>1</td>
<td>1=Strongly Disagree; 2= Somewhat Disagree; 3= Somewhat Agree; 4=Strongly Agree</td>
<td>I have spent time trying to find out more about my own ethnic group, such as its history, traditions and customs.</td>
</tr>
<tr>
<td>7</td>
<td>ethident_02</td>
<td>Ethnic Identity</td>
<td>Numeric</td>
<td>3-5</td>
<td>1</td>
<td>1=Strongly Disagree; 2= Somewhat Disagree; 3= Somewhat Agree; 4=Strongly Agree</td>
<td>I am active in organizations or social groups that include mostly members of my own ethnic group.</td>
</tr>
<tr>
<td>8</td>
<td>ethident_03</td>
<td>Ethnic Identity</td>
<td>Numeric</td>
<td>3-5</td>
<td>1</td>
<td>1=Strongly Disagree; 2= Somewhat Disagree; 3= Somewhat Agree; 4=Strongly Agree</td>
<td>I have a clear sense of my ethnic background and what it means for me.</td>
</tr>
<tr>
<td>9</td>
<td>ethident_04</td>
<td>Ethnic Identity</td>
<td>Numeric</td>
<td>3-5</td>
<td>1</td>
<td>1=Strongly Disagree; 2= Somewhat Disagree; 3= Somewhat Agree; 4=Strongly Agree</td>
<td>I like meeting and getting to know people from ethnic groups other than my own.</td>
</tr>
<tr>
<td>10</td>
<td>ethident_05</td>
<td>Ethnic Identity</td>
<td>Numeric</td>
<td>3-5</td>
<td>1</td>
<td>1=Strongly Disagree; 2= Somewhat Disagree; 3= Somewhat Agree; 4=Strongly Agree</td>
<td>I think a lot about how my life will be affected by my ethnic group membership.</td>
</tr>
</tbody>
</table>
Types of Data Documentation

• Annotated instrument
  – Contains basic variable and value information in context
  – Easy to interpret
  – Difficult to integrate much additional detail; not useful for some forms of data
Annotated Instrument

**BRIEF SYMPTOM INVENTORY-18 (Derogatis, 2001)**

To what extent have you been bothered by the following over the past seven days?

<table>
<thead>
<tr>
<th></th>
<th>(0)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSI_01</td>
<td><strong>Faintness or dizziness</strong></td>
<td>Not at all</td>
<td>A little bit</td>
<td>Moderately</td>
<td>Quite a bit</td>
</tr>
<tr>
<td>BSI_02</td>
<td><strong>Feeling no interest in things</strong></td>
<td>Not at all</td>
<td>A little bit</td>
<td>Moderately</td>
<td>Quite a bit</td>
</tr>
<tr>
<td>BSI_03</td>
<td><strong>Nervousness or shakiness inside</strong></td>
<td>Not at all</td>
<td>A little bit</td>
<td>Moderately</td>
<td>Quite a bit</td>
</tr>
<tr>
<td>BSI_04</td>
<td><strong>Pains in heart or chest</strong></td>
<td>Not at all</td>
<td>A little bit</td>
<td>Moderately</td>
<td>Quite a bit</td>
</tr>
<tr>
<td>BSI_05</td>
<td><strong>Feeling lonely</strong></td>
<td>Not at all</td>
<td>A little bit</td>
<td>Moderately</td>
<td>Quite a bit</td>
</tr>
<tr>
<td>BSI_06</td>
<td><strong>Feeling tense or keyed up</strong></td>
<td>Not at all</td>
<td>A little bit</td>
<td>Moderately</td>
<td>Quite a bit</td>
</tr>
<tr>
<td>BSI_07</td>
<td><strong>Nausea or upset stomach</strong></td>
<td>Not at all</td>
<td>A little bit</td>
<td>Moderately</td>
<td>Quite a bit</td>
</tr>
<tr>
<td>BSI_08</td>
<td><strong>Feeling blue</strong></td>
<td>Not at all</td>
<td>A little bit</td>
<td>Moderately</td>
<td>Quite a bit</td>
</tr>
<tr>
<td>BSI_09</td>
<td><strong>Suddenly scared for no reason</strong></td>
<td>Not at all</td>
<td>A little bit</td>
<td>Moderately</td>
<td>Quite a bit</td>
</tr>
</tbody>
</table>
Study-Level Documentation

• What are the most important elements to document?
  – Data elements
  – Study elements
  – Processes and decisions
Study-Level Documentation

• Details about the source of the data
  – Study design and purpose
  – Collection method
  – Information about the research sample
  – Longitudinal time points (if applicable)
Study-Level Documentation

• Information about data files
  – File name/version
  – Date created
  – Number of records
  – Number of variables
  – Changes since last version of file
Study-Level Documentation

• Information about measures used
  – Description of measure
  – Description of scales
  – Source of measure, including references as appropriate
Study-Level Documentation

Programs used to process/manipulate data
- Documentation within program (comments)
- Documentation of what various programs do and in what order they are used
Study-Level Documentation

• Data narrative
  – Good for measure/study-level information
Study-Level Documentation

• Data narrative (continued)

Data Files:
Data files are available in SAS and SPSS formats. SAS files have a “.sas7bdat” extension, while SPSS files have a “.sav” extension.

<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
<th>Description</th>
<th>Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort1_t1_v7</td>
<td>6/18/2014</td>
<td>Cohort 1 time 1 data</td>
<td>Updated algorithm for calculating variableX</td>
</tr>
<tr>
<td>Cohort1_t2_v5</td>
<td>7/18/2014</td>
<td>Cohort 1 time 2 data</td>
<td>Merged admin data from source Y</td>
</tr>
<tr>
<td>Cohort1_t3_v1</td>
<td>8/23/2014</td>
<td>Cohort 1 time 3 data</td>
<td>New data from time 3</td>
</tr>
<tr>
<td>Long_cohort1_v3</td>
<td>9/1/2014</td>
<td>Longitudinal data: Cohort 1 time 1, time 2, and time 3 data</td>
<td>Incorporates changes from c1t1_v7/c1t2_v5; adds time 3 data</td>
</tr>
</tbody>
</table>

Study Design and Methodology:
This study was school-based—all paper and pencil questionnaires were completed by students at school during class time. Additionally, participants were recruited in their 9th grade science classes. The three surveyed samples came from the same school, and were recruited in consecutive years. The pilot sample was recruited from 9th graders in the 2010-2011 school year, Cohort 1 was recruited from the 2011–2012 9th grade students, and Cohort 2 students were recruited from the 9th grade in 2012–2013. Aside from the pilot data collection, the study has a longitudinal design, as the same students were followed from 9th grade through 12th grade. However, cross-sectional data sets from each data point were also created.
Decision and Process Documentation

• What are the most important elements to document?
  – Data elements
  – Study elements
  – Processes and decisions
Decision and Process Documentation

• By far, the least established area of research documentation.

• Due to individual differences between research projects, it can be difficult to identify a standard template.
Decision and Process Documentation

Elements to include in documentation:

- Scope (variables/measures)
- Time (if multiple time points)
- Describe purpose of process or situation requiring a decision being made
Decision and Process Documentation

Elements to include in documentation:

• Information from the data that describes or affects the decision or process

• A description of the process itself, including:
  – Any software or tools needed to complete the process
  – Any resources /references used
Decision and Process Documentation

• What sorts of decisions and processes should be documented with this level of detail?
  – Basic scales and processes that are commonly utilized may not require this much detail
  – Processes and procedures that are not well established or that deviate significantly from the standard method should be documented
Decision and Process Documentation

• Examples of processes that might need to be documented
  – Naming conventions for variables
  – Naming conventions for data files
  – Structure of data directories
  – Version information
Decision and Process Documentation

• Examples of decisions that might need to be documented
  – Resolving discrepancies in data obtained from multiple sources or at multiple time points
  – Data transformations that require interpretation
The survey contained a question asking students to categorize their race/ethnicity. This question was included for Cohort 1 at Times 1, 2 and 3 and Cohort 2 at Times 1 and 2.

Additionally, the project obtained information on students’ race from school records. For those students who were at the school for the initial data collection in ninth grade, the information was collected at that time. For students new to the school at each following data collection, the information was collected at that time point.

For a number of students, their self-report race/ethnicity changed over time. Additionally, some students self-report of race did not match the school’s report of their race/ethnicity.

This document highlights the differences found in self-report for fixing the self-report over time. The differences found between self-report and school records are as follows:

**Cohort 1: Frequency of discrepancies in self-reported race/ethnicity**

<table>
<thead>
<tr>
<th>Number of instances</th>
<th>Races/ethnicities given</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>White/Other</td>
</tr>
<tr>
<td>3</td>
<td>White/Hispanic</td>
</tr>
<tr>
<td>3</td>
<td>Black/Multi-ethnic</td>
</tr>
<tr>
<td>2</td>
<td>Multi-ethnic/Other</td>
</tr>
<tr>
<td>1</td>
<td>Black/White</td>
</tr>
<tr>
<td>1</td>
<td>Hispanic/Multi-ethnic</td>
</tr>
<tr>
<td>1</td>
<td>Black/Other</td>
</tr>
<tr>
<td>1</td>
<td>White/Multi-ethnic</td>
</tr>
<tr>
<td>1</td>
<td>Black/Multi-ethnic/Other</td>
</tr>
</tbody>
</table>

**Cohort 2: Frequency of discrepancies in self-reported race/ethnicity**

<table>
<thead>
<tr>
<th>Number of instances</th>
<th>Races/ethnicities given</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Black/Multi-ethnic</td>
</tr>
<tr>
<td>2</td>
<td>Black/Other</td>
</tr>
<tr>
<td>2</td>
<td>Multi-ethnic/Other</td>
</tr>
<tr>
<td>2</td>
<td>White/Other</td>
</tr>
</tbody>
</table>
Tools for Documentation

- Statistical software packages (e.g. SAS, Stata)
  - Variable information (PROC contents; describe)
  - Provides a good starting point for a codebook

- Database management systems
Tools for Documentation

• Data collection instruments
  – Paper forms
  – Electronic/online collection
PROC CODEBOOK (SAS)

PROC CODEBOOK is a SAS macro that creates a codebook based on a SAS data set.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Variable Label (VAR)</th>
<th>VAR Type</th>
<th>VAR Format</th>
<th>VAR Length</th>
<th>Mean</th>
<th>Range of Values</th>
<th>Frequency Category</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARM_CAT</td>
<td>ARM_CAT: Arm of study for participant: INCENT, USUAL, WPI, WEB (Y)</td>
<td>Char</td>
<td>$ARM_CAT</td>
<td>6</td>
<td></td>
<td>INCENT</td>
<td>Incentive Payment (IP)</td>
<td>314</td>
<td>27.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>USUAL</td>
<td>Neither (Usual Care)</td>
<td>256</td>
<td>22.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>WEB</td>
<td>WEB-Based Program (WEB)</td>
<td>288</td>
<td>25.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>WPI</td>
<td>Both WEB &amp; IP</td>
<td>280</td>
<td>24.60</td>
</tr>
<tr>
<td>B_HWT1</td>
<td>B_HWT1: 1. Completed consent form? 0=N/1=Y</td>
<td>Num</td>
<td>HWT_NOYES1F</td>
<td>8</td>
<td></td>
<td>.</td>
<td>Missing</td>
<td>21</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>Yes</td>
<td>1117</td>
<td>98.15</td>
</tr>
<tr>
<td>B_HWT10</td>
<td>B_HWT10: 10. FFQ Status: 1=Complete, 2=Incomplete, 3=Will mail</td>
<td>Num</td>
<td>HWT_STATUS_1F</td>
<td>8</td>
<td>1.603297</td>
<td>.</td>
<td>Missing</td>
<td>107</td>
<td>9.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>Completed</td>
<td>719</td>
<td>63.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>Incomplete</td>
<td>2</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>Will mail</td>
<td>310</td>
<td>27.24</td>
</tr>
</tbody>
</table>
PROC CODEBOOK (SAS)

• Requirements
  – Labels on variables and data set
  – Formats assigned to categorical values
  – Minimum of 1 categorical/2 numeric variables

• Optional elements
  – Ordering of variables (default is by variable name)
  – ODS formatting of title text
PROC CODEBOOK (SAS)

• Can be useful when dealing with data sets that include SAS formats

• If data set does not already have formats applied, may take as much time to add them as to create your own codebook (which has more flexibility)

• To download the SAS macro and access documentation, visit http://www.cpc.unc.edu/research/tools/data_analysis/proc_codebook
Documentation Standards

• Much of the documentation we’ve discussed serves to help users understand the data.

• How can we document the data in a way that helps potential interested parties find the data?
Dublin Core

- The Dublin Core Metadata Element Set includes 15 standard elements.
- Intended for describing a wide range of different web-based or physical resources.
- Each element is optional and repeatable.
Dublin Core


<dc:title>
  Carolina Abecedarian Project and the Carolina Approach to Responsive Education (CARE), 1972-1992
</dc:title>
<dc:creator>Ramey, Craig T.</dc:creator>
<dc:creator>Gallagher, James J.</dc:creator>
<dc:creator>Campbell, Frances A.</dc:creator>
<dc:creator>Wasik, Barbara H.</dc:creator>
<dc:creator>Sparling, Joseph J.</dc:creator>
<dc:subject>academic achievement</dc:subject>
<dc:subject>child development</dc:subject>
<dc:subject>cognitive functioning</dc:subject>
<dc:subject>early childhood education</dc:subject>
<dc:subject>intervention</dc:subject>
<dc:subject>marital status</dc:subject>
<dc:subject>poverty</dc:subject>
<dc:subject>socioeconomic status</dc:subject>

<dc:description>...</dc:description>
<dc:date>2004-12-08</dc:date>
<dc:type>survey data</dc:type>
<dc:identifier>4091</dc:identifier>
<dc:identifier>10.3886/ICPSR04091.v1</dc:identifier>

<dc:source>
  Interviews with parents, observations of the families, individually administered cognitive and academic tests,
</dc:source>
<dc:coverage>United States</dc:coverage>

<dc:rights>
  ICPSR metadata records are licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.
</dc:rights>
Data Documentation Initiative (DDI)

- An international specification for describing data from the social, behavioral, and economic sciences
- Supports the entire research data lifecycle
Data Documentation Initiative (DDI)
Data Documentation Initiative (DDI)

ICPSR data undergo a confidentiality review and are altered when necessary to limit the risk of disclosure. ICPSR also routinely creates ready-to-go data files along with setups in the major statistical software formats as well as standard codebooks to accompany the data. In addition to these procedures, ICPSR performed the following processing steps for this data collection:

- Checked for undocumented or out-of-range codes.
The Takeaway

• Good documentation is not just a product, it’s an approach
Resources

• Inter-university Consortium for Political and Social Research (ICPSR)
  – Guide to Social Science Data Preparation and Archiving

• Cornell Research Data Management Service Group
  – Guide to writing "readme" style metadata

• Duke University Libraries
Resources

- Dublin Core
- Data Documentation Initiative (DDI)
Questions?

• Ask away!

• If you would like to talk more about documentation for your own projects, contact us at ehdidata@duke.edu.

• Thanks for coming!