Cite as: Veerle Van den Eynden, Scott Summers and Stuart MacDonald (2018) Exercise: Data documentation. UK Data Service, UK

UK Data Service

Exercise: Data documentation

When as a researcher you want to reuse existing data, for example for new analysis or meta-analysis, it is important that you understand what exactly the data mean and how they were created. You also need to be able to assess whether the data are fit for your intended purpose. This requires the data to be well described and documented and well prepared.

Understanding how to best prepare and document your own research data when publishing them for future reuse, is easiest to appreciate from the perspective of the new user who is unfamiliar with the dataset.

For this exercise we will use the following datasets:

1. Malawi Household surveys for agricultural biodiversity assessment
2. Manufacturing growth and the lives of Bangladeshi women
3. All Ireland Traveller Health Study (AITHS)

Listed below are examples of key information that a researcher using data may need to know. Evaluate the example datasets to see whether you can find this information easily. Also indicate where you found the information, e.g. in the dataset descriptor, in the data file itself, in supplementary documentation, in a related publication. Consider whether this could have been described or presented better and whether additional information would help to reuse data.

	Did you find the information, and where?		
Key information needed for reuse of data (examples)	**1 Malawi HH survey**	**2 Manufacturing Bangladesh**	**3 AITHS**
Example: Number of respondents	*340; found in the dataset description*	*1395 households; found in the ReadMe file (part of the data zip bundle)*	*8492; found in the AITHS Technical Report 1*
Geographical area where the data were collected			
Is there sampling bias or is the sample random?			
Is there is a control group ?			
Were data collected directly in digital format or on paper and then submitted/transcribed into a database; if so was double entry or peer checking done to avoid errors?			

	1 Malawi HH survey	2 Manufacturing Bangladesh	3 AITHS
Which questions exactly were asked in the survey or interview (or which protocols used for measurements)			
Can you find the hypothesis or aims of the research that generated this dataset?			
How was consent gathered?			
Can the data be used for commercial purposes?			
What access conditions apply to the data?			
Can you find a publication that describes the findings of this dataset?			
Is it clear which respondents or interviewees are female?			
If there are missing data in the datafile, are they missing because the respondent did not respond or because the question was not asked to this respondent? (or missing because a measurement was not done or not relevant)			
Does the file format and structure of the data facilitate easy reuse?			
Are related datasets that use the same research protocol comparable to facilitate cross-analysis, e.g. same variable names, same coding structure, etc.			