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

UK Data Service

Answers: Data documentation

The table describes which information could be found and where it was found.

The assessed datasets are:

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

	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; 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	Malawi, Ntcheu District , areas Manjawira, Nsipe, Sharpevale and Tsangano; found in dataset descriptor	Bangladesh	Eire and Northern Irelan
Is there sampling bias or is the sample random?	Random sample	Unknown	Sampling bias; found in AITHS Technical Report 1
Is there is a control group ?	No	Unknown	No
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?	Unknown	Unknown	Data collected via paper questionnaire, checked and transcribed to a spreadsheet, and if required followed up with the study coordinators for Clarification; found in AITHS SUMMARY
Which questions exactly were asked in the survey or interview (or which protocols used for measurements)	Questionnaire is available as documentation	Unknown	Questionnaire is available as documentation
Can you find the hypothesis or aims of the research that generated this dataset?	No	Yes, in dataset abstract and in published paper	Yes in dataset descriptor

How was consent gathered?	Unknown	Unknown	As part of the questionnaire form
Can the data be used for commercial purposes?	Not clear, seemingly CC0 licence, so yes	No, CC-BY-NC licence; found in data descriptor	No, research and learning purposes only; found in dataset descriptor
What access conditions apply to the data?	Open access	Open access	Data available upon request
Can you find a publication that describes the findings of this dataset?	No	Yes, direct link from the dataset; but the paper requires payment/ subscription	Various reports included in documentation files
Is it clear which respondents or interviewees are female?	Yes, this can be seen in the datafiles	All respondents are female	Yes; found in AITHS SUMMARY
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)	Unknown; missing data are blank	Unknown; proprietary data, so cannot check	Yes, missing data information is available in the data dictionaries
Does the file format and structure of the data facilitate easy reuse?	Excel format	Stata format, not normalised	SPSS format, not normalised
Are related datasets that use the same research protocol comparable to facilitate cross-analysis, e.g. same variable names, same coding structure, etc.	Yes	N/A	Yes

Further discussion

These 3 dataset examples represent datasets that may be shared for different reasons: because a research institution wants to make their data available for further reuse (example 1), because a journal expects data to be available so research findings can be replicated (example 2), because a data repository wants to make valuable data resources available to the research community (example 3), because a research funder expects data sharing, etc. The examples show that the level of documentation for datasets can vary highly. Either way, once a dataset is available in a data repository or published in another form, it is important that sufficient documentation is available so the data can be reused by researchers that may find the dataset. It is also important that this documentation is openly available so users can judge whether the data files are worth accessing / downloading for their use, especially if the dataset is not openly accessible.

Whilst researchers usually do not have the time and expertise to provide documentation to the level of a professional data archive, it is important to provide sufficient documentation in an open format so interested users can judge the value of the data and its potential for reuse.

1. Malawi Household surveys for agricultural biodiversity assessment
This dataset has a good concise data descriptor (metadata record) and the research protocol and questionnaire form are available as extra documentation. Variable codes are explained in the data table.

2. Manufacturing growth and the lives of Bangladeshi women
Whilst this dataset is described well in the published paper it supports (summary methodology and variable descriptions are detailed), the paper is not openly available but hidden behind a paywall / subscription wall. The documentation that is openly available is very minimal. A user who finds the dataset needs access to the paper to be able to understand the data. A clear ReadMe file provides information about the structure and content of each data file.

3. All Ireland Traveller Health Study (AITHS)
This dataset is extensively documented through technical reports, a readme file, data dictionaries, questionnaires and details about anonymisation and content gathering (14 documentation files are available), created by the survey agency and the data archive.