UK Data Archive

Good Practice in Data Documentation

Revised Version
January 2002

UK Data Archive
University of Essex
Wivenhoe Park
Colchester, Essex CO4 3SQ
UK

GOOD PRACTICE IN DATA DOCUMENTATION

Why does the UKDA need good documentation?

Good documentation facilitates effective checking and preservation of a dataset and ensures that the research community will be able to use the data. It reduces the likelihood of misuse or incorrect use of the data. It can even help data creators should they return to the dataset for further analysis at some stage in the future.

What should be provided to the UKDA?

There are three main types of material that constitute ideal documentation for a dataset:

- **Explanatory material.** This is the material that is essential to the informed use of a dataset. Without this material no full understanding of the dataset and its contents can be achieved.
- **Contextual information.** This provides users with material about the context in which the data were collected and information about the uses to which the data were put. This also forms a more comprehensive historical record for researchers of the future. Whilst not essential, inclusion of this information is strongly recommended.
- **Cataloguing information.** This allows the UKDA to create a formal catalogue record, or study description, for the study. The study description serves two purposes. First, it serves as a bibliographic record of the dataset, which allows the dataset to be properly acknowledged and cited in publications that arise from secondary analysis. Secondly, it is the principal instrument used for resource discovery. It provides users with detailed searchable information by which they can identify the studies that are useful to their research or teaching. This material is gathered using the UKDA’s Study Description Form.

Explanatory material

This represents the minimum of material that should be created and preserved. It is comprised of the following:

Information about the data collection methods

This should describe the data collection process, whether a survey, collection of administrative information, transcription of a document source, experimental model etc., as well as the data collection method. Details on how the methods were developed are also useful. If applicable, details of the sampling design, sampling frame and sampling methods should be included. It is also extremely useful to include information on any pilot research or monitoring processes undertaken during the main data collection exercise, and details of any other quality controls using during the data creation stage. It should also detail the geographic and temporal coverage of the dataset.

Information about the structure of the dataset

This should include information about the relationships between individual files and records within a dataset. At a minimum, it should also include the number of cases and variables in each data file and the number of files comprising the dataset. For relational databases, a diagram showing the structure and the relations between the records and elements of the dataset should be constructed (i.e. an entity-relationship diagram).

Technical information

This should provide a record of the software and operating system on which the files were generated, the medium on which the data are stored and a complete list of all data files that make up the dataset.

Variables and values, coding and classification schemes

This should comprise a complete variable list that describes all the variables (or fields) in the dataset and full details of all coding and classifications used. It is helpful to identify variables to which standard coding classifications (e.g. SIC and SOC codes) apply and to record the version of the classification scheme used, preferably with a bibliographic reference to that code.

Information about derived variables

Many data producers derive new variables from the original data collected. These may be as simple as grouping interval age data (age in years) to groups of years or something more complicated created by a complex algorithm. When grouped or derived variables are created it is important that the logic of each derivation is made clear. Simple grouping, e.g. grouped age, can be explained by the variable and value labels but for more complex derivations, other means of recording the information are needed. The best method of describing these is to provide the logical statements that created these derived variables. If this is inherently comprehensible to the end user (e.g. an SPSS or STATA command file), then these commands should be supplied. If the command language is not inherently comprehensible, each derivation should be described using simple logical statements, and these should be sent in addition to the actual command files.

Weighting and grossing

Weighting and grossing variables need to be fully documented, explaining the construction of the variables with a clear indication of the circumstances in which they should be used. The latter is particularly important when different weights need to be applied for different purposes.

Data source

Details of the source from which the data are derived should be well documented. For example, where the data source is responses to survey questionnaires, the text of each question asked should be carefully recorded in the documentation. Ideally the question text will include a reference to the variable(s) it has generated. Details of the conditions under which a question has been asked, the number of cases to which a question applies, and summary response statistics are also very useful.

Confidentiality and anonymisation

It is important to record if the data contain any confidential information concerning individuals, households, organisations or institutions. Such information should normally be removed or anonymised prior to submission of the dataset to the UKDA. If the data have been anonymised to prevent identification of subjects, it is wise to record the procedures used and the resultant changes to the data. Such modification may restrict any subsequent analysis and an indication of this is useful.

Where secondary analysis requires confidential or otherwise sensitive information to remain in the dataset, agreements about any special access conditions to end users should be discussed with the Acquisitions section of the UKDA prior to submission of the dataset (e-mail: archive-acq@essex.ac.uk).

Validation and other checks
This should comprise details of any known data errors and any data checking and cleaning performed as part of the data collection or subsequent checks.

Contextual information

This type of information adds richness and depth to the Explanatory material already defined. It is comprised of the following:

Description of the originating project
This should comprise details of the history of the project or process that gave rise to the dataset, in terms of the intellectual, financial and organisational origins and developments over time. For example, it might detail why the data collection was felt necessary, the aims and objectives of the project, publications arising, policy developments to which it contributed, and any other relevant contextual information.

Provenance of the dataset
This information is useful in recording the history of the data collection process, changes and developments that occurred, both in the data themselves and the methodology, any adjustments made and so on.

Serial and time-series datasets, new editions
For repeated cross-section, panel or time-series datasets, additional information describing, for example, any changes in the variable content, question text, variable labelling or sampling procedures, is enormously helpful.

Cataloguing information

Most of the necessary information is provided by completion of the UKDA's Study Description Form, which has been designed for this purpose. The form gathers information such as the title of the dataset, principal investigator, sponsors, data collectors, dates of data collection, temporal and geographic coverage, methods of data collection, and sampling design and frames. Further information is taken from any other documentation accompanying the dataset or from related reports, publications and articles. Discrepancies found in the depositor forms or in the documentation are investigated. This form can be downloaded from:

http://www.data-archive.ac.uk/depositingData/depositForm.asp.