

Chapter 2

Organise & Document

Contents

| | |
|--|----|
| Main take-aways | 33 |
| 2.1 Designing a data file structure | 34 |
| 2.1.1 Organisation of variables | 39 |
| 2.2 File naming and folder structure | 43 |
| 2.3 Documentation and metadata | 47 |
| 2.4 Adapt your DMP: part 2 | 57 |
| Sources and further reading | 58 |

Main authors of this chapter

Jindrich Krejčí, Czech Social Science Data Archive (CSDA)

Johana Chylikova, Czech Social Science Data Archive (CSDA)

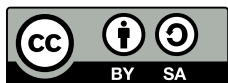
Katja Fält, Finnish Social Science Data Archive (FSD)

CITATION

CESSDA Training Team (2017 - 2019). CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC. DOI: 10.5281/zenodo.3820473

Retrieved from <https://www.cessda.eu/DMGuide>

LICENCE



The Data Management Expert Guide by CESSDA ERIC is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. All material under this licence can be freely used, as long as CESSDA ERIC is credited as the author.

Introduction



In this chapter¹, we provide you with tips and tricks on how to properly organise and document your data and metadata. We begin by discussing good practices in designing an appropriate data file structure, file naming and organising your data within suitable folder structures. You will see how the way you organise your data facilitates orientation in the data file, contributes to the understanding of the information contained and helps to prevent errors and misinterpretations.

In addition, we will focus on what counts as an appropriate documentation of your data. Development of rich metadata is required by FAIR data principles and other current standards promoting data sharing.

Main take-aways

After completing your journey through this chapter on organising and documenting your data you should:

- » Be aware of the elements which are important in setting up an appropriate structure for organising your data for intended research work and data sharing;
- » Have an overview of the best practices in file naming and organising your data files in a well-structured and unambiguous folder structure;
- » Understand how comprehensive data documentation and metadata increases the chance your data are correctly understood and discovered;
- » Be aware of common metadata standards and their value;
- » Be able to answer the DMP questions which are listed at the end of this chapter and adapt them to your own DMP.

¹ The content of this chapter was inspired by research data management manuals, guidelines, online courses and methodological texts published by several data organisations and experts, in particular the information provided by the UK Data Service (2017a), the “Guide to Social Science Data Preparation and Archiving” by the US-based data organisation ICPSR (2012), the online course Research Data MANTRA (EDINA and Data Library, University of Edinburgh, 2017), A guide into research data management by Corti, Van den Eynden, Bishop and Woollard (2014), Krejčí’s “Introduction to the Management of Social Survey Data” (Krejčí, 2014), Gibbs (2007) and Data Management Guidelines produced and published by the Finnish Social Science Data Archive (FSD-Finnish Social Science data Archive, 2017).

2.1 Designing a data file structure

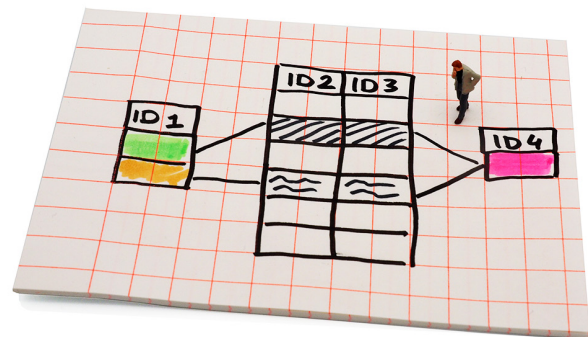


In an early stage of your research, you are faced with the question of what form your data files should take. Your initial decisions about the structure of your data files should be considered thoroughly.

Data file structure has a huge impact on the possible ways your files can be processed and analysed and once your structure has been filled with data, any changes to it are usually laborious and time-consuming.

File structure choice

Data files may have different internal structures and a research study may encompass several different data files in different relations to one another. The structure of the data file is also determined by the formatting of its content (e.g., types and organisation of variables). It provides information on relationships among different elements and parts of its content. An important part of the metadata is often embedded into the data file (e.g., in the form of variable names and variable and value labels, different kinds of notes and content of supplementary variables). So, the structure of your data also contributes to the clarity of your data documentation.



File structure choice often depends on the requirements of the software you are using and intended analysis. At the same time, your decisions about structure may define the possibilities of future data processing, choice of software and ways of data analysis.

When deciding on a data file structure, consider the following:

- » Units of analysis, possible analytical objectives and methods of analysis to be used;
- » Relations
 - » between different content items and parts of your data file;
 - » to sources of your data;
 - » to any other relevant external data and information and their structure.
- » Possibilities of building connections to other existing or future data files (future additions of new data or creation of cumulative data files);
- » Possible strategies for version control (see 'Data authenticity and version control');
- » Possible technical limitations, e.g. operability in relation to the size of the data file (consider that large and complicated structures may put high demands on both data management and computing capacities. Some software programs also have limitations with respect to the number of variables and cases they can manage);
- » The software you are going to use (this should be done also with respect to flexibility because of possible secondary analysis of your data in other software).

Designing qualitative data files



Qualitative data files emerge from many different types of research material. Such data files are texts (transcribed interviews or focus group sessions, various types of written texts, such as newspaper and magazine material, diaries etc.) or photographs, audio files (recordings of speech) or video files. Unlike quantitative data, qualitative data are not presented in the form of variables, numbers, data matrices etc. However, they must also be organised and stored in an exact manner so they are easily managed and available for use.

Usually, individual data collection events will be structured into individual files, e.g. one interview transcript, one image, or one audio recording each time make a single file. These single files are then organised into folders of similar files. Sometimes, qualitative information may also be organised into matrix structures, e.g. textual extracts from newspaper articles or diaries may be placed into a rectangular matrix, whereby further metadata and coding can be added alongside each entry.

Designing a qualitative data structure comes down to:

- » Thinking of ways to categorise data (see 'Qualitative coding');
- » Developing a file naming strategy (see 'File naming and folder structure');
- » Designing a comprehensive folder structure (see "File naming and folder structure").

Designing quantitative data files



In quantitative research, the content of the data often results from numerical coding in standardised questionnaires (see 'Quantitative coding'). In addition, full-text answers or textual codes can be recorded into specific types of variables in quantitative data files. Quantitative researchers may also store other material, i.e. administrative data, data from social media or various texts. In this chapter, however, when we speak about quantitative data, we usually mean survey data.

Below you will find a description of three types of file structures - flat, hierarchical and relational - which are commonplace in quantitative social science. Also, two examples which clarify the concepts are presented.

Flat file

Flat (rectangular) data files are organised in long rows, variable by variable. One row is dedicated to one subject of observation and/or analysis. An ID number usually comes first. If variable values are organised column by column, we obtain a rectangular matrix.

SPSS and STATA and similar software are often used for analysing flat files. Here the structure consists of one rectangular matrix with data, accompanied by variable and value labels. In this case, each record includes the same amount of information and has the same length as all other records in the data file. If the variable is not applicable for a particular observation, it is filled with blank spaces or missing values.

Hierarchical file

Hierarchical files consist of higher-order and lower-order records which are arranged in a hierarchical structure, i.e. several lower-order units may be linked to one higher-order unit and are contained in the same data file.

If there are different levels of units in your database the flat data file can be impractical because it may include a large number of blanks and put great demands regarding the size of the file. In addition, it may also reduce the operability and clarity with regard to differentiation of types of units of analysis. Database applications like e.g. D-base, MS Access or SQL, allow structuring your data in a hierarchical order.

Relational database

The relational database is a system of several data matrices and defined associations between them.

Different other database applications, e.g., D-base, MS Access or SQL databases, allow the structuring of your data in a hierarchical order. You may also split your data into several interrelated flat files, i.e. structure your data into a relational database, and retain the ability to use statistical software mentioned above.

Example 1: Structuring a database from a household survey

If you consider a database from a household survey, there are at least two different types of units of analysis, households and individual household members. However, you may structure such database in all three ways as follows:

Hierarchical structure

If you decide on a hierarchical structure, data on the household are recorded at one level and data on household members at another level.

You can download an example of a hierarchical file in *.sav [here](#).

Relational database

Another solution is to create a relational database. Information about household members is recorded in independent matrices that are interconnected by means of a household ID or a more complex parameter that represents not only the sharing of a household but also the type of family relationship between household members or similar. For instance, users can search for rows with equal attributes in this type of database. Relational databases may also serve as a basis for creating files adapted for individual exercises by combining information from different matrices.

Flat file

However there could be a situation where you may need to use complete household survey data in your analysis and your software requires a flat file. In this case, you can add a household ID variable and copy particular household data for each individual member of this household. This would create a set of individuals. Another possibility would be to organise records for all household members in long rows, which would create a set of households.

You can download an example of such a flat file in *.sav [here](#).

Example 2: SHARE - A complicated database of micro data on health, socio-economic status and social and family networks

The Survey of Health, Ageing, and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks. Surveys are organised bi-annually since 2004. SHARE currently covers 27 European countries and Israel. The SHARE database is easily accessible to the entire research community; data from the SHARE Waves 1 to 6 have been available since 2017.

Description

SHARE is targeted to individuals aged 50 or older and their households. The resulting database is quite complicated due to the following:

- » SHARE is an international survey and its data come from different countries;
- » SHARE is a panel survey repeating interviews with the same sample of households every two years, so the data come from different waves of the survey using questionnaires including both, repeated and new questions;
- » A systematic process of refreshment is implemented and new households are added to the panel at each wave. That is the reason there are two types of questionnaires: the baseline questionnaire for respondents who participate in a SHARE interview for the first time and the longitudinal questionnaire for respondents who participated in SHARE before;
- » There are different components of the survey with different sources of information collected under different data collection modes. The data collection modes include face-to-face interviews based on computer-assisted personal interviewing (CAPI) on household level, CAPI on individual level for different types of household members, paper and pencil (PAPI) drop-off questionnaires, so-called vignettes, i.e. questionnaires on respondents reactions to specific situations, physical tests, collection of dried blood spots (only in some countries), specific end-of-life questionnaires, interviewer observations and generated variables;
- » Different types of respondents answer different parts of an interview for household: (1) family respondent, (2) financial respondent, (3) household respondent; in addition, e.g., in case of physical or cognitive limitations of dedicated respondent, it is possible to organise a 'proxy interview' with another member of the household (proxy respondent);
- » The survey and its database are also structured by topics.

Unique identifiers

The features of the surveys which are described above, result in a complicated relational database.

Up to 30 different data modules are available for each wave of the survey, for each of participating countries and for databases combined across time and countries. Moreover, there are also two levels of data based on units of observation, households or individuals. Data from different modules and/or waves may be merged using the following set of unique identifiers:

- » For merging data on an individual level the variable "mergeid" provides a unique and non-changing person identifier for all waves. It has the format "CC-hhhhhh-rr" (e.g. "AT-070759-01"), where CC refers to the short country code (here: "AT" for Austria), "hhhhhh" are digits to identify the household, and "rr" is the respondent identifier within each household.
- » For merging data on the household level there is a set of variables hhid`w, where `w indicates the respective wave. hhid`w has the following format "CC-hhhhhh-S" (e.g. "AT-070759-A"), where "CC" refers to the short country code, "hhhhhh" is the household identifier, and "S" identifies possible split households, i.e. the household of a panel member who moved out of a previous household. In case of a household split there is not only an "A"-suffix but also "B", "C", etc.

In addition, there are several 'Special Data Sets', e.g. interviewer survey, country-specific projects to link SHARE data with selected administrative records and 'Biomarkers' (objective health measures or a retrospective panel about working life histories of SHARELIFE respondents (SHARE Wave 3)).

For purposes of analysis, SHARE provides a very extensive set of weights (See Weighting). Which weights to use really depends on the concrete research question, i.e. the cross-sectional or longitudinal nature of the study, the waves under investigation, the unit of analysis (household or individual), and the reason for weighting sample observations (SHARE, 2017: 34).

easySHARE for training purposes

Working with the SHARE panel data is very demanding. Thus, in addition to the standard SHARE database, also a longitudinal data set "easySHARE" has been created for training purposes. It contains only selected variables merged into a single data file, making it more user-friendly. However, for deeper analysis, a standard database is necessary.

Dive in deeper?

We have a subtopic prepared for you on organising variables. Here you will find tips on how to build the internal structure of quantitative data files by organising, naming and labeling variables.

Alternatively, you can proceed to the section on designing file names and folder structures.

2.1.1 Organisation of variables

Data file structure is supported by the organisation of variables. Variable names and labels contribute to the structuring of the data file, allowing to integrate part of the documentation into the data file and helping researchers to orient themselves in the structure of the data sets. At the same time, variable names should be short and should respect the usual requirements of standard software, because they are used as calling codes in software operations.

The position of variables in the data file, their names and labels should reflect the following:

» Relations between variables

E.g., sets of variables related to the same phenomenon (these should be placed together in a dataset, e.g. the age of all children in a household), original or derived variables (derived variables are created from other variables, e.g. the age in years is re-coded into broader categories)

» Links to elements of the study and sources of the data

E.g., different measurement instruments, different parts of the questionnaire, different source databases, different methods of observation, etc.

» Types of variables

E.g., identification variables and other supplementary variables with different specific roles, socio-demographic indicators, generated variables obtained by transformation of original information, etc.

Organising your data

Data files also include supplementary variables which facilitate orientation and management, ensure integrity, or are necessary to perform some analyses. As a rule, you should include a unique identifier (or set of identifiers) for cases (individual respondents) in the file. A unique identifier is an identification code for the case. These are usually numbers, for example, 0001, 0002, 0003 etc. To facilitate orientation, they are usually placed at the very beginning of the file.

List of variables:

| | |
|---------|---|
| SEX | Sex of respondent |
| AGE | Age of respondent |
| MARITAL | Marital status of respondent |
| COHAB | Do you live together with a partner? |
| EDUCYRS | Education I – years (of fulltime) schooling |



Other variables may help to distinguish between different sources of information, methods of observation, temporal or other links. Yet others may provide information about the organisation of data collection such as interviewer ID or interviewing date or distinguish cases which belong to various groups.

It is absolutely necessary for an analysis to distinguish data that result from overrepresentation sampling strategies, different waves of research, etc., especially if groups of cases distinguished by them are to be analysed in different ways.

For each variable in the data file, you should set the variable width, i.e. the number of characters or the length of the integer and fractional parts of a number. The set number of characters or digits for each variable is reserved for every case, even if they are left blank.

Naming variables

In the boxes below, basic rules for variable naming are given and an example is presented.

Basic rules for variable naming

The basic rules for variable naming are following:

- » Start with a letter. Do not start with a number, question or exclamation marks or a special character such as #, &, \$, @ (they are often reserved for specific purposes in software applications);
- » Variable names cannot contain spaces;
- » Variable names are also used as calling codes in software operations. For this reason, variables should be short and respect the usual requirements of a standard software. The standard is to not make variable names any longer than eight characters;
- » Do not use diacritics (marks above or below a letter) or national specific characters;
- » Make them meaningful (so they can be used for better orientation in the data files).

There are three basic approaches to naming variables:

- » Using numeric codes that reflect the variable's position in a system (e.g. V001, V002, V003...);
- » Using codes that refer to the research instrument (e.g. question number in a questionnaire: Q1a, Q1b, Q2, Q3a...);
- » Using mnemonic names that refer to the content of variables (e.g. BIRTH for the year of birth, AGE for respondent's age etc.). The word mnemonic means "memory aid".

Variable labels

Variable labels provide a short description of the variable name. These can be longer than the recommended eight characters for variable names. Although size limits are less strict here, it is advisable to keep variable labels rather brief and find an adequate compromise between clarity and the size of the label. Keep in mind that many analytical outputs are provided in tables. Thus, excessively lengthy labels can result in large and impractical tabulations. The size of the labels may also complicate format conversions. In some analytical outputs or after conversions, only a part of a lengthy label is kept. The loss of the remainder of the variable label may make the label incomprehensible.

Examples of variable labels include a short or full version of the question, or a question code if variable names are not constructed around them. E.g.:

- » The variable label is adapted from the number and question-wording from the questionnaire: "B10 - How old are you?";
- » The descriptive label is "Age of a respondent";
- » Schematically this becomes: "Respondent: AGE".

To reach the widest audience possible, the preferred language for variable naming is English.

Labels for variable values

Variables have two or more values (a variable with only one value is called a constant and in fact, is not a variable). Sometimes you must assign labels to values of variables. You do not need to assign labels to values of continuous variables like age (in years), height (in metres) or weight (in kilograms), because their units are generally known. This is different for nominal and ordinal variables. A nominal variable like gender has two values, usually represented by 0 and 1 in data. You should assign labels “male”/“female” to these two values, so you and another researcher who might use the data would know which value represents which gender. The same applies to ordinal scales, for example, agree-disagree scale with values 1, 2, 3, 4 and 5, where 1 represents “completely disagree” and 5 “completely agree”. You must label these values so you and others know what degree of dis/agreement the numbers represent.

Example

Two different concepts of variable naming and labelling in the data file from the International Social Survey Programme

The International Social Survey Programme (ISSP) is a continuing, long-term international programme of survey research on important sociological topics. It brings together pre-existing, social science projects and coordinates research goals, thereby adding a cross-national perspective to the individual, national studies. Established in 1984, it now has almost 50 member countries. The ISSP surveys are organised annually.

Each ISSP survey contains two international modules:

- » **ISSP thematic module**

A specific topic of the survey is selected for each year. There are about ten topics, which are repeated at regular intervals. However, sometimes a topic is skipped or replaced by a new one.

- » **ISSP background variables module**

These include a set of harmonised sociodemographic variables. This module is repeated every year. However, there are also frequent changes in this set of variables.

Two different concepts of variable naming and labelling are used for these two modules.

Table: Excerpt from the variable list of the international dataset from ISSP 2009 on 'Social Inequalities' (ISSP Research Group, 2017).

| Variable name | Variable label |
|--|--|
| <i>ISSP 2009 thematic module variables</i> | |
| V73 | Q24a Describe yourself: I work hard to complete my daily tasks |
| V74 | Q24b Describe yourself: I perform to the best of my ability |
| V75 | Q24c Describe yourself: I work hard to maintain my performance on a task |
| V76 | Q25a Describe yourself as <14-15-16> years old: I tried hard to go to school every day |
| V77 | Q25b Describe yourself as <14-15-16> years old: I performed to the best of my ability |
| <i>ISSP background variables</i> | |
| SEX | R: Sex |
| AGE | R: Age |
| MARITAL | R: Marital status |
| COHAB | R: Steady life-partner |
| EDUCYRS | R: Education I: years of schooling |
| DEGREE | R: Education II-highest education level |
| AR_DEGR | Country-specific education: Argentina |
| AT_DEGR | Country-specific education: Austria |
| AU_DEGR | Country-specific education: Australia |
| BE_DEGR | Country-specific education: Belgium |

In the table we see two approaches to variable labelling:

» **Simple variable names**

The first thematic part of the file contains simple variable names (numeric codes). The information on the numbers of the questions in the common international questionnaire is included in variable labels. It supports better user orientation in the data file. The question numbers are followed by a literal question, sometimes shortened adequately to remain comprehensible and follow the rule of keeping the variable label short. Some ISSP surveys allow alternative wording of questions – possible alternatives are bracketed in inequality signs. Similarly, after country specifics (e.g., country name, the currency used), general names come in inequality signs.

» **Mnemonic names of variables**

The second part contains background variables and uses mnemonic names of variables referring to their contents. These background variables are not directly linked to the wording of questions in the international questionnaire but are instead constructed from national versions of data. Their names refer to their contents and simultaneously to links between them (e.g., DEGREE = the education variable transformed into an internationally comparable form, XX_DEGR = education variables using original country-specific coding). Moreover, the set of mnemonic names of background variables is standardised across different ISSP surveys, which allows easier merging of ISSP data files across time and construction of time-series databases.

TIP! Mnemonic variable names may help to establish links between sets of variables within a data file. In addition, in repeated surveys, if the same naming convention of mnemonic names is used, it makes easier merging data over time.

2.2 File naming and folder structure

To enable you to identify, locate and use your research data files efficiently and effectively you need to think about naming your files consistently and structuring your data files in a well-structured and unambiguous folder structure.

File naming strategy

Two important starting points for your file naming strategy are:

» **A file name is a principal identifier of a file**

Good file names provide useful clues to the content, status and version of a file, uniquely identify a file and help in classifying and sorting files. File names that reflect the file content also facilitate searching and discovering files. In collaborative research, it is essential to keep track of changes and edits to files via the file name.

» **File naming strategy should be consistent in time and among different people**

In both quantitative and qualitative research file naming should be systematic and consistent across all files in the study. A group of cooperating researchers should follow the same file naming strategy and file names should be independent of the location of the file on a computer.

Below, best practice and examples of useful file names are given.

Elements in a file name

Common elements that should be considered (UK Data Service, 2017b) when developing a file naming strategy:

- » Version number (also see 'Data authenticity');
- » Date of creation (date format should be YYYY-MM-DD);
- » Name of creator;
- » Description of content;
- » Name of research team/department associated with the data;
- » Publication date;
- » Project number.

Best practice

According to the UK Data Archive (UK Data Service, 2017b), a best practice in naming files is to:

- » Create meaningful but brief names;
- » Use file names to classify types of files;
- » Avoid using spaces, dots and special characters (& or ? or !);
- » Use hyphens (-) or underscores (_) to separate elements in a file name;
- » Avoid very long file names;
- » Reserve the 3-letter file extension for application-specific codes of file format (e.g. .doc, .xls, .mov, .tif);
- » Include versioning of file names where appropriate.

File naming for qualitative data

Several aspects of naming that are particularly important for qualitative data (Finnish Social Science Data Archive, 2016):

- » If you have large numbers of files of different types, you should produce a document describing the file naming convention used for the research;
- » Background information about each item (individual interview, focus group, photograph etc.) is usually indicated in the file names. Nevertheless, you should always present background information in separate documents.

Consistency of naming

The benefit of consistent naming of data files is that it is easier to identify all files connected to one data collection event (e.g. one interview). The files related to one collection event (e.g. audio tape, its transcription and photographs that were taken by the interviewee) can be connected by the file name.

The most convenient way is to give all files connected to the same event an 'event identifier' in the beginning of the name, that is, in the first part of the name. The latter part of the name can be used to convey the specifics, for instance, whether it is an audio tape, transcription or a still image:

Example:

- » 20130311_interview2_audio.wav
- » 20130311_interview2_trans.rtf
- » 20130311_interview2_image.jpg

Documenting data file conventions

An example of how to document the data file conventions you use:

<date><type><ID1><gender><age><municipality><datatype><ID2>

where:

- » <date> is the date on which the data were collected (date format should be YYYY-MM-DD);
- » <type> specifies the type of event/data material;
- » <ID1> is the ID of the collection event;
- » <gender> is the gender of the interviewee;
- » <age> is the age of the interviewee;
- » <municipality> is the municipality of residence of the interviewee;
- » <datatype> specifies the type of data the file contains, for instance, "trans" means transcription, "audio" means audio recording, and "image" means photograph;
- » <ID2> is the ID number used to separate the images connected to the collection event.

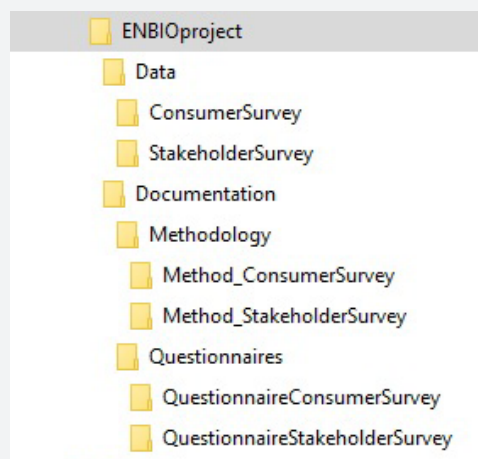
Folder structure

Structuring your data files in folders is important for making it easier to locate and organise files and versions. A proper folder structure is especially needed when collaborating with others.

The decision on how to organise your data files depends on the plan and organisation of the study. All material relevant to the data should be entered into the data folders, including detailed information on the data collection and data processing procedures.

Consider the best hierarchy of your files and decide whether a deep or shallow hierarchy is preferable. If you have several independent data collections, it is advisable to create a separate data folder for each collection. For inspiration, have a look at the examples below.

Survey data



For this survey, data and documentation files are held in separate folders. Data files are further organised according to data type and then according to research activity.

Documentation files are organised also according to the type of documentation file and research activity. It helps to restrict the level of folders to three or four deep and not to have more than ten items on each list.

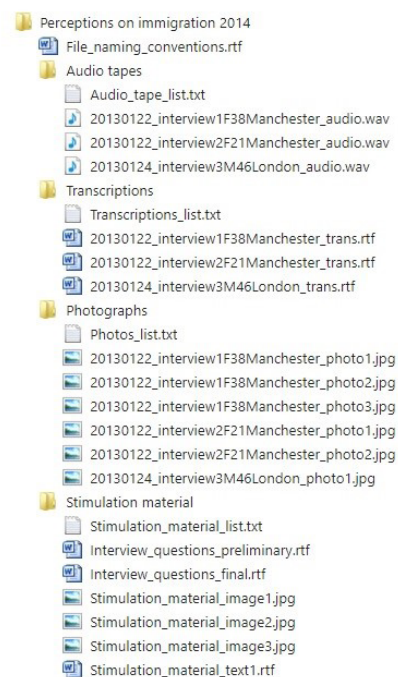
Qualitative data files

In this example, the data contain audiotapes of the interviews, interview transcripts, stimulation material shown to the research subjects, and photographs taken by the subjects.

Data files are files connected to the same interview event conducted on the 22nd of January 2013. The latter part of the name reveals the specifics of the file.

In this case, “audio” means audio tape and “trans” a transcription of the audio tape. However, background information must never be stored in the file name only.

This example was taken from UK Data Service (2017b).



TIP: Batch renaming of automatically generated files



Batch renaming is organising research data files and folders in a consistent and automated way with software tools (also known as mass file renaming, bulk renaming).

Batch renaming software exists for most operating systems. See the box below for examples.

Batch renaming tools

Examples of batch/bulk renaming tools include:

Windows:

- » Ant Renamer;
- » RenameIT;
- » Bulk Rename Utility.

Mac:

- » Renamer 5;
- » Name Changer.

Linux:

- » GNOME Commander;
- » GPRename.

Unix:

- » The use of the grep command to search for regular expressions.

It may be useful to rename files in a batch when:

- » Images from digital cameras are automatically assigned base filenames consisting of sequential numbers;
- » Proprietary software or instrumentation generate crude, default or multiple filenames;
- » Files are transferred from a system that supports spaces and/or non-English characters in filenames to one that doesn't (or vice versa). Batch renaming software can be used to substitute such characters with acceptable ones.

How to ... use Bulk Rename Utility

Follow the steps in this video to use Bulk Rename Utility to batch rename your files.

2.3 Documentation and metadata



I have never documented my data before. I have both qualitative and quantitative data and I work on a collaborative project. Where do I start?

How to start?

1. Do not panic. Much documentation is simply good research practice, so you are probably already doing much of it.
2. Start early! Careful planning of your documentation at the beginning of your project helps you save time and effort. Do not leave the documentation for the very end of your project. Remember to include procedures for documentation in your data management planning.
3. Think about the information that is needed in order to understand the data. What will other researchers and re-users need in order to understand your data?
4. Create a separate documentation file for the data that includes the basic information about the data. You can also create similar files for each data set. Remember to organise your files so that there is a connection between the documentation file and the data sets.
5. Plan where to deposit the data after the completion of the project. The repository probably follows a specific metadata standard that you can adopt.
6. Document consistently throughout the project. Data documentation gives contextual information about your dataset(s). It specifies the aims and objectives of the original project and harbours explanatory material including the data source, data collection methodology and process, dataset structure and technical information. Rich and structured information helps you to identify a dataset and make choices about its content and usability.

TIP: Use English for documentation. It increases the chance your data are understood and reused.

Systematically documented research data is the key to making the data publishable, discoverable, citable and reusable. Clear and detailed documentation improve the overall data quality. It is vital to document both the study for which the data has been collected and the data itself. These two levels of documentation are called project-level and data-level documentation.

Project-level documentation

The project-level documentation explains the aims of the study, what the research questions/hypotheses are, what methodologies were being used, what instruments and measures were being used, etc. In the following boxes, the questions that your project-level documentation should answer are stated in more detail.

1. For what purpose was the data created

Describe the project history, its aims, objectives, concepts and hypotheses, including:

- » The title of the project;
- » Subtitle;
- » Author(s)/creator(s) of the dataset;
- » Other co-workers and their roles (person, research group or organization that participated in the study and their roles);
- » The institution of the author(s)/creator(s);
- » Funders;
- » Grant numbers;
- » References to related projects;
- » Publications from the data.

2. What does the dataset contain?

Describe what is in a dataset:

- » Kind of data (interviews, images, questionnaires, etc.);
- » File size (in bytes), file format of the data files and relationships between files;
- » Description of data file(s): version and edition, structure of the database, associations, links between files, external links, formats, compatibility.

3. How was the data collected?

Describe how the data was acquired:

- » The methodology and technique used in collecting and creating the data;
- » Description of all the sources the data originate from (What is the subject of study? E.g. periodicals, datasets created by others?) together with an explanation of how and why it got to the present place (provenance);
- » The methods/modes of data collection (for example):
 - » The instruments, hardware and software used to collect the data;
 - » Digitisation or transcription methods;
 - » Data collection protocols;
 - » Sampling design and procedure;
 - » Target population, units of observation.

4. Who collected the data and when?

Describe the:

- » Data collector(s);
- » Date of data collection;
- » Geographical coverage of the data (e.g. Nation).

5. How was the data processed?

Describe your workflow and specific tools, instruments, procedures, hardware/software or protocols you might have used to process the data, like:

- » Data editing, data cleaning;
- » Coding and classification of data.

6. What possible manipulations were done to the data?

Describe if and how the data were manipulated or modified:

- » Modifications made to data over time since their original creation and identification of different versions of datasets;
- » Other possible changes made to the data;
- » Anonymisation;
- » For time series or longitudinal surveys: changes made to methodology, variable content, question text, variable labelling, measurements or sampling.

7. What were the quality assurance procedures?

Describe how the quality of the data has been assured:

- » Checking for equipment and transcription errors;
- » Quality control of materials;
- » Data integrity checks;
- » Calibration procedures;
- » Data capture resolution and repetitions;
- » Other procedures related to data quality such as weighting, calibration, reasons for missing values, checks and corrections of transcripts, transformations.

8. How can the data be accessed?

Describe the use and access conditions of the data:

- » Where the data can be found (which data repository);
- » Permanent identifiers;
- » Access conditions such as embargo;
- » Parts of the data that are restricted or protected;
- » Licences;
- » Data confidentiality;
- » Copyright and ownership issues;
- » Citation information.

Data-level documentation

Data-level or object-level documentation provides information at the level of individual objects such as pictures or interview transcripts or variables in a database. You can embed data-level information in data files. For example, in interviews, it is best to write down the contextual and descriptive information about each interview at the beginning of each file. And for quantitative data variable and value names can be embedded within the data file itself.

Quantitative data

Variable-level annotation should be embedded within a data file itself. If you need to compile an extensive variable level documentation, you can create it by using a structured metadata format.

Data-level documentation for quantitative data

For quantitative data document the following information is needed:

- » Information about the data file
Data type, file type, and format, size, data processing scripts.
- » Information about the variables in the file
The names, labels and descriptions of variables, their values, a description of derived variables or, if applicable, frequencies, basic contingencies etc. The exact original wording of the question should also be available. Variable labels should:
 - » Be brief with a maximum of 80 characters;
 - » Indicate the unit of measurement, where applicable;
 - » Reference the question number of a survey or questionnaire, where applicable.

Example of a variable and variable label

Variable: 'Q11eximp'

Variable label: 'Q11: How important is exercise for you?'

Value labels: 1: Very unimportant. 2. Unimportant. 3. Neutral. 4. Important. 5. Very important.

The label gives the unit of measurement and a reference to the question number (Q11).

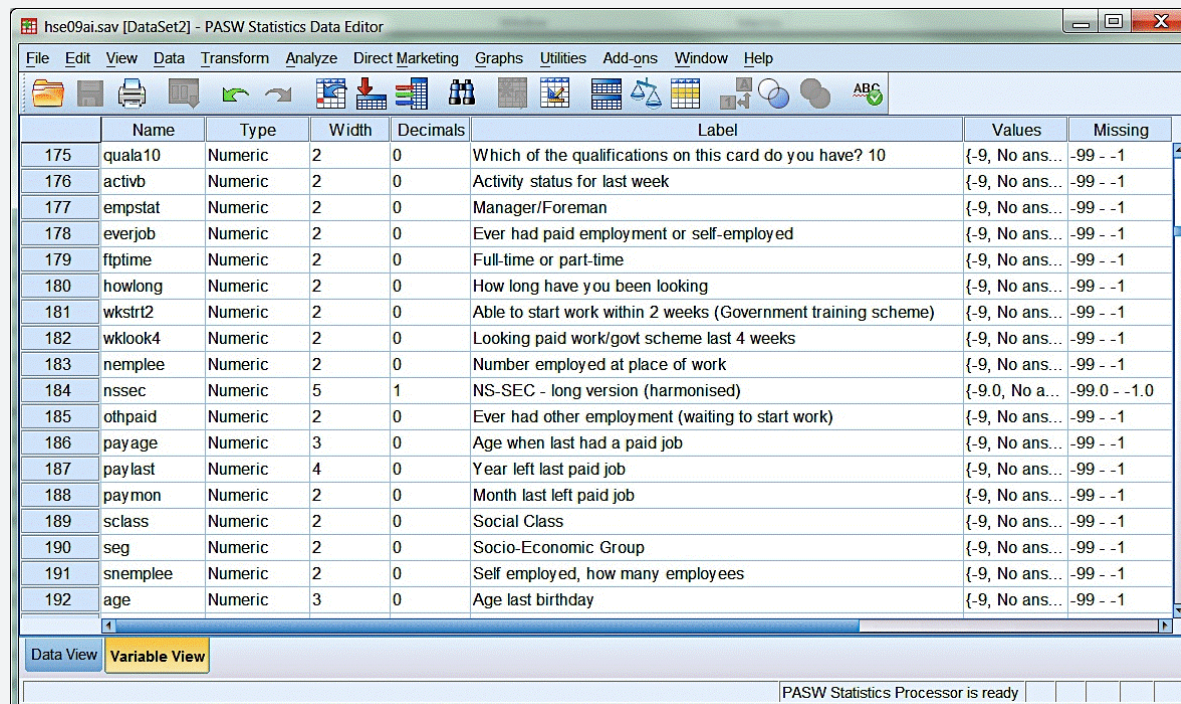
- » Information about the cases in the file
- » A specification of each case (units of research like e.g. a respondent) if applicable.
- » Names, labels and descriptions for variables, records and their values
- » Description of the missing values at each variable
- » Description of the weighting variable
- » Explanation or definition of codes and classification schemes used

Storing documentation

Whenever possible, embed data documentation within a file. See the following example.

Example of embedded data documentation

In this example from the UK Data Service (2017c), you see two SPSS tabs: Data view and Variable view, the tab which is visible right now.



| | Name | Type | Width | Decimals | Label | Values | Missing |
|-----|-----------|---------|-------|----------|--|----------------|--------------|
| 175 | quala10 | Numeric | 2 | 0 | Which of the qualifications on this card do you have? 10 | {-9, No ans... | -99 - -1 |
| 176 | activb | Numeric | 2 | 0 | Activity status for last week | {-9, No ans... | -99 - -1 |
| 177 | empstat | Numeric | 2 | 0 | Manager/Foreman | {-9, No ans... | -99 - -1 |
| 178 | everjob | Numeric | 2 | 0 | Ever had paid employment or self-employed | {-9, No ans... | -99 - -1 |
| 179 | ftptime | Numeric | 2 | 0 | Full-time or part-time | {-9, No ans... | -99 - -1 |
| 180 | howlong | Numeric | 2 | 0 | How long have you been looking | {-9, No ans... | -99 - -1 |
| 181 | wkstrt2 | Numeric | 2 | 0 | Able to start work within 2 weeks (Government training scheme) | {-9, No ans... | -99 - -1 |
| 182 | wklook4 | Numeric | 2 | 0 | Looking paid work/govt scheme last 4 weeks | {-9, No ans... | -99 - -1 |
| 183 | nemplee | Numeric | 2 | 0 | Number employed at place of work | {-9, No ans... | -99 - -1 |
| 184 | nssec | Numeric | 5 | 1 | NS-SEC - long version (harmonised) | {-9.0, No a... | -99.0 - -1.0 |
| 185 | othpaid | Numeric | 2 | 0 | Ever had other employment (waiting to start work) | {-9, No ans... | -99 - -1 |
| 186 | payage | Numeric | 3 | 0 | Age when last had a paid job | {-9, No ans... | -99 - -1 |
| 187 | paylast | Numeric | 4 | 0 | Year left last paid job | {-9, No ans... | -99 - -1 |
| 188 | paymon | Numeric | 2 | 0 | Month last left paid job | {-9, No ans... | -99 - -1 |
| 189 | sclass | Numeric | 2 | 0 | Social Class | {-9, No ans... | -99 - -1 |
| 190 | seg | Numeric | 2 | 0 | Socio-Economic Group | {-9, No ans... | -99 - -1 |
| 191 | sneemplee | Numeric | 2 | 0 | Self employed, how many employees | {-9, No ans... | -99 - -1 |
| 192 | age | Numeric | 3 | 0 | Age last birthday | {-9, No ans... | -99 - -1 |

Qualitative data

Background and contextual information and participant details of interviews, observations or diaries can be described at the beginning of a file as a header or summary page.

Data-level documentation for qualitative data

For qualitative data document the following information is needed:

- » Textual data file (for example, interview)
- » Key information of participants such as age, gender, occupation, location, relevant contextual information);
- » For qualitative data collections (for example image or interview collections) you may wish to provide a data list that provides information that enables the identifying and locating of relevant items within a data collection:
- » The list contains key biographical characteristics and thematic features of participants such as age, gender, occupation or location, and identifying details of the data items;
- » For image collections, the list holds key features for each item;
- » The list is created from an initial list of interviews, field notes or other materials provided by the data depositor.

Example of data level documentation of textual data

For textual data, background data are systematically entered at the beginning of each data unit (e.g. interview transcript) in a standardised manner.

The following example from the Finnish Social Science Data Archive presents a typical transcript of an interview with only one interviewee. The transcript of each interview in the data has been saved in a separate file, often in .rtf or .doc(x). Background data fields are entered in the following manner at the beginning of each transcription file.

Beginning of the transcript file

Interview date: 08.02.2013 [=8 February 2013]

Interviewer: Matt Miller

Pseudonym of interviewee: Ian (not the real first name of the interviewee)

Occupation of interviewee: Journalist

Age of interviewee: 32

Gender of interviewee: Male

Audiovisual data files

For some types of data (image, audio or video files) the file format does not always allow recording background information in the beginning of the data file. In such cases, the best practice is to store background information in a manually created data list or a separate text file: a data list which accompanies the data collection.

- » Provide the following information on each image: creator, date, location, subject, content, copyright, keywords, equipment used;
- » Some image files have embedded technical metadata (You may use tools to extract technical metadata from images, such as ExtractMetadata.com (n.d.)).

Example of a data list

In this case - shown on the site of the Finnish Social Science Data Archive (2016) - the background data fields are manually entered in table form using Excel (or Open Office Calc program). The data collected were video-recorded interviews. The data list contains background information related to the interviewee and the interview event as well as information on the model and brand of the camera used and the length of the video (in minutes).

See also another data list example from the UK Data Service (2017c)

| 1 | Interview videos 2012 | | | | | | | | |
|---|-----------------------|----------------|-------------|--------------------|-----|--------|------------|---------------------------|-----------------------|
| 2 | File name | Interview date | Interviewer | Interviewee's name | age | gender | occupation | Camera used for the video | Duration of the video |
| 3 | Peter_1.avi | 12.4.2012 | Matt Miller | Peter Herald | 37 | Male | Barkeeper | Panasonic HC-V10 | 2:45 |
| 4 | Peter_2.avi | 12.4.2012 | Matt Miller | Peter Herald | 37 | Male | Barkeeper | Panasonic HC-V10 | 5:05 |
| 5 | Lisa_1.avi | 17.4.2012 | Matt Miller | Lisa Smith | 43 | Female | Author | Canon XF305 | 10:12 |
| 6 | Mary_1.avi | 22.4.2012 | Matt Miller | Mary Davies | 42 | Female | Teacher | Panasonic HC-V10 | 6:56 |
| 7 | Pablo.mpg | 24.4.2012 | Matt Miller | Pablo Neftali | 76 | Male | Poet | Canon XF305 | 4:32 |

Periodicals, magazines, journal articles

Among materials you use for qualitative data analysis, there may be online periodicals, magazines or journal articles. The information about all such resources must be kept in separate files:

- » Material collected from online periodicals: save references to web resources, like URLs, and do not forget they may change over time. To be sure information is not lost, articles should be copied into a word processing program;
- » Materials from periodicals: When articles, photographs and other material are collected from periodicals for research purposes, bibliographic information should be carefully detailed (author(s), title, date of publication etc.);
- » When you analyse articles, make a list of them, sort them alphabetically or chronologically in the order they were analysed in the course of research.

Storing documentation

- » Write the documentation into a separate, well-structured file, and associate that with the data file. You may use the same filename stem in order to strengthen the file-metadata association. For example: 20130311_interviews_audio, 20130311_interviews_trans, 20130311_interviews_image, 20130311_interviews_metadata. The latter part of the name can be used to convey the specifics of the file. In this case “audio” means audio tape and “trans” a transcription of the audio tape;
- » Data-level documentation can be embedded within a data file. For example, in interviews, it is best to write down the contextual and descriptive information about each interview at the beginning of each file;
- » If you have a large amount of metadata or large amounts of data that will need metadata you can use a standard specific database for this purpose (such as the DDI Codebook (DDI Alliance, 2017a)).

Metadata: machine readable data documentation

Metadata or “data about data” are descriptors that facilitate cataloguing data and data discovery. Metadata are intended for machine-reading. When data is submitted to a trusted data repository, the archive generates machine-readable metadata. Machine-readable metadata help to explain the purpose, origin, time, location, creator(s), terms of use, and access conditions of research data.



Create machine-readable metadata

Check out The Dublin Core Metadata Generator ([dublincoregenerator](http://dublincoregenerator.org), n.d.) and see how metadata elements are converted into a machine-readable file in *.xml.

Also, if you enjoy working with *.xml schemas, get started in creating a codebook to accompany your dataset with the DDI codebook (DDI Alliance, 2017a).

Deposit data in a data repository

When you submit your dataset in a (trusted) data repository, machine-readable metadata will be added. See the chapter on ‘Archiving and Publishing data’ for a description of such data repositories.

In the boxes below we provide you with examples of:

» **Metadata templates** (for easy starting)

If you do not quite know yet what metadata you should generate (what fields are needed) have a look at the metadata templates provided. Some of them are very simple and can, therefore, help to create basic documentation.

» **Metadata standards** (for when you need your metadata to be very structured).

Metadata standards may at first look seem quite scary. They are used by data archives for enhancing discoverability, interoperability and reusability. When you submit your dataset to a trusted data repository, these standards are automatically applied.

Metadata templates

Metadata can, at its simplest, be stored in a single text file. However, you can also use a metadata template to help you structure your metadata or to see how your metadata appears in *.html.

Below we provide examples of metadata templates that you can use when compiling documentation. Or just for inspiration to take a look at typical fields which are often required. It is always possible to include additional documentation beyond what is suggested.

- » Create a codebook about your research to accompany the dataset (DDI Alliance, 2017a).
- » Download the York University (n.d.) Library Metadata Template, Dublin Core;
- » Have a look at the Georgia Tech Library (n.d.) Metadata Template;
- » Use the Dublin Core Metadata Generator (dublincoregenerator, n.d.);
- » Have a look at the Cornell University (n.d.) guide to writing “readme” style metadata (with downloadable template);
- » Use the ISO 19115-2 Metadata Editor (GRIIDC (2015)) web application.

Metadata standards

You may want your metadata to be very structured. For that purpose, you can choose a metadata standard or a tool (software that has been developed to capture or store metadata) to help you add and organise your documentation. Many standards are discipline-specific. These will help you to add metadata to the workflow as they have been created to suit the needs of research data.

Remember that you do not generally need to generate machine-readable metadata by yourself. The repository where you may want to deposit your data will do that for you. When you are depositing your data the repository will require a data documentation document from you and will convert the documentation into machine-readable metadata.

The recommended standard for research in the social sciences is the DDI metadata standard.

DDI for social sciences

DDI (Data Documentation Initiative) (DDI Alliance, 2017b) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences. Expressed in XML, the DDI metadata specification supports the entire research data lifecycle.

Common fields in the DDI include:

- » Title
- » Alternate Title
- » Principal Investigator
- » Funding
- » Bibliographic Citation
- » Series Information
- » Summary
- » Subject Terms
- » Geographic Coverage
- » Time Period
- » Date of Collection
- » Unit of Observation
- » Universe
- » Data Type
- » Sampling
- » Weights
- » Mode of Collection
- » Response Rates
- » Extent of Processing
- » Restrictions
- » Version History

MIDAS Heritage for historical sites

MIDAS Heritage (Historic England, 2012) is a British cultural heritage standard for recording information on buildings, archaeological sites, shipwrecks, parks and gardens, battlefields, areas of interest and artefacts.

VRA Core for images and works of art and culture

VRA Core (2015) is a standard for the description of images and works of art and culture.

ISO 19115 for geospatial data

ISO 19115 (DCC, 2017) is a schema for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, the spatial reference, and the distribution of digital geographic data.

Metadata standards for general research data

- » Dublin Core (DCMI, 2017);
- » DataCite Metadata Schema (Datacite, n.d.);
- » PREMIS (2017).

In its simplest form, the Dublin Core consists of 15 fields that basically describe all online resources:

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

Example of DDI

For an example of how to apply the metadata standard DDI, have a look at a dataset in the Finnish Social Science Data Archive (Galanakis, Michail (University of Helsinki): Intercultural Urban Public Space in Toronto 2011-2013 [dataset]. Version 1.0 (2014-02-13). Finnish Social Science Data Archive [distributor]. <http://urn.fi/urn:nbn:fi:fsd:T-FSD2926>).

The machine-readable XML file looks like this.

For a visually formatted example of a DDI record, see the online version of this chapter:
<https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata>

Metadata for new data types – new standards still under development

To provide metadata for social media data and transaction data with metadata, the metadata standards by the Data Documentation Initiative (DDI) should serve as the guiding framework. Importantly, however the DDI standard is “insufficient to document all the details required for reproducibility of a social media dataset” (Kinder-Kurlanda et al 2017: 3). For example, the DDI format does not allow describing biases caused by data mining interfaces of social media platforms, changes in data availability and formats, explanations about code and scripts used in collection, cleaning and analysis etc. Such information can be described only in an unstructured manner as an additional comment in the standard’s form.

Together with other CESSDA partners, GESIS is currently developing recommendations for the provision of metadata for new data types (esp. social media data).

2.4 Adapt your DMP: part 2



This is the second of six 'Adapt your DMP' sections in this tour guide. After working on this chapter, you should be able to plan for organising and documenting your data.

To adapt your DMP, consider the following elements and corresponding questions:

Document data type and size

- » What type(s) of data will be collected?
- » What is the scope, quantity, and format of the material?
- » What is the total amount of data collected (in MB/GB)?

Data organisation

- » How will you organise your data?
- » Will the data be organised in simple files or more complex databases?
- » What is your process for quality assurance? What are your quality measures?
- » Are there specific quality standards or quality management models you plan to apply?

File naming and folder structure

Are there any specific requirements for compatibility and comparability of your data?

Are there specific standards that you want to implement, e.g. naming conventions or standardised coding structures?

Data documentation and metadata

- » Will you be creating separate files accompanying the data?
- » Will you be using a database?
- » Are the data produced and/or used in the project discoverable with metadata?
- » What metadata will you use? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.
- » If you already know in which data repository you will publish your data, what metadata standard do they use?

For easy reference, we have put together a list of DMP-questions for all chapters in this tour guide. You can view and download the checklist as pdf (CESSDA, 2018a) or editable form (CESSDA, 2018b), and keep them as a reference while you are studying the contents of this guide.

Sources and further reading

Please see the online version of this guide.