Chapter 3

Process

Contents

The online version of this chapter is available at:

cessda.eu/DMEG

Main authors of this chapter

Jindrich Krejčí, Czech Social Science Data Archive (CSDA)

Johana Chylikova, Czech Social Science Data Archive (CSDA)

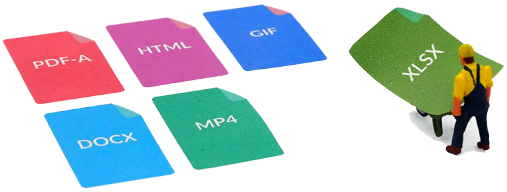
CITATION

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

Introduction



In this chapter[1], we focus on the data operations needed to prepare your data files for analysis and data sharing. Throughout the different phases of your project, your data files will be edited numerous times. During this process, it is crucial to maintain the authenticity of research information contained in the data and prevent it from loss or deterioration.

However, we will start with the topics of data entry and coding as the first steps of your work with your data files. Finally, you will learn about the importance of a comprehensive approach to data quality.

Main take-aways

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

- Be familiar with strategies to minimise errors during the processes of data entry and data coding;
- Understand why the choice of file format should be planned carefully;
- Be able to manage the integrity and authenticity of your data during the research process;
- Understand the importance of a systematic approach to data quality;
- 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 (Finnish Social Science data Archive, 2017).

3.1 Data entry and integrity

> *Data integrity means assurance of the accuracy, consistency, and completeness of original information contained in the data. At the same time, the authenticity of the original research information has to be preserved (see 'Data authenticity').*

The integrity of a data file is based on its structure and on links between data and integrated elements of documentation. From the moment that data is being entered, data integrity is at stake.

Data entry procedures have changed over recent years. Operators entering data into a computer manually are being replaced by automated computer technologies, while the universal distinction between the three phases of data collection, data entry, and data editing/checking is often becoming obsolete. In general, greater automation of processes generally prevents some types of errors, but at the same time, it produces other types of errors. For example, errors in scripts during computer-assisted interviewing may cause systematic shifts in data and to be able to detect such deviations in automated forms of data entry requires different kinds of checks in comparison to manually entered data.

Minimising errors in survey data entry

In the boxes below, a summary of recommendations on minimising errors in survey data entry is given (UK Data Service, 2017a; ICPSR, 2012; Groves et al., 2004).

Check the completeness of records

Check if your data files contain the correct number of records, number of variables or length of the records, etc.

Reduce burden of manual data entry

Manual data entry requires routine and concentration. Operators should not be burdened by multiple tasks. Tasks such as coding and data entry should be implemented separately.

Minimise the number of steps

The data entry process should include a smaller rather than a larger number of steps. This reduces the likelihood of errors.

Conduct data entry twice

When you have paper questionnaires, the data entry can be processed electronically by scanning questionnaires or manually entering data by a person responsible for data entry. If data are entered by scanning, execute the process of data entry twice and compare values. If data are entered manually, a portion of questionnaires should be entered twice by two different persons. For example, the Czech Association of Public Opinion and Market Research Agencies (SIMAR) recommends 20 percent of questionnaires be re-entered.

Perform in-depth checks for selected records

At least some randomly selected records, e.g. 5–10% of all records, should be subjected to a more detailed, in-depth check to verify the procedures and identify possible systematic errors. The cases should be selected by chance. Be sure to document the changes you make and keep the original data so you can restore them at all times.

There are multiple methods for logical and consistency checks, including the following:

- Check the value range (e.g. a respondent over the age of 100 is unlikely);
- Check the lowest and highest values and extremes;
- Check the relations between associated variables (e.g. educational attainment should correspond with a minimum age, the total number of hours spent doing various activities should not exceed 100% of the available time);
- Compare your data with historical data (e.g. check the number of household members with the previous wave of a panel survey).

Automate checks whenever possible

Specialised software for computer-assisted interviewing (CAPI, CATI, etc.) or data entry software allows to set the range of valid values for each category and to apply filters to manage the data entry or the entire data collection process. These automatic checks:

- Prevent meaningless values from being entered;
- Help to discover inconsistencies that arise when some values are skipped or omitted;
- Make the interviewer's work substantially clearer and easier;
- Reduce the number of errors that interviewers make.

The software can distinguish between permanent rules that cannot be bent and warnings that only notify the operator when entering an unlikely value.

CAPI software is used by the data collectors and it is usually expensive and therefore individual researchers cannot afford to buy it. In case you collected your survey data by yourself, you must write your own program/syntax to check your data for discrepancies.

An example of an SPSS syntax to check your data

Logical check of income - the household income cannot be SMALLER than individual income

The syntax search for respondents who indicated the household income as well as their individual income, while the household income was smaller than individual income.

Variable names:

ide.10 - household income

interval variable, income in Euros, with special values 8 - refused to answer; 9 - don´t know

ide.10a - individual income

interval variable, income in Euros, with special values 8 - refused to answer; 7 - doesn´t have income

Syntax (SPSS):

```
USE ALL.
COMPUTE filter_$=(ide.10a ne 0) and (ide.10 ne 0) and (ide.10a ne 7) and (ide.10a ne 8) and (ide.10 ne 8) and (ide.10 ne 9) and (ide.10 < ide.10a).
VARIABLE LABELS filter_$ '(ide.10a ne 0) and (ide.10 ne 0) and (ide.10a ne 7) and (ide.10a ne 8) and (ide.10 ne 8) and (ide.10 ne 9) and (ide.10 < ide.10a) (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
FREQUENCIES VARIABLES=CD
/ORDER=ANALYSIS.
FILTER OFF.
USE ALL.
```

In cases of errors ...

What to do with error values?

You can either delete or try to correct error values. Simple data entry errors can be easily corrected based on comparison with respondents' original answers. However, you should bear in mind that inconsistencies can also be generated by the respondents themselves, and a correction should make a minimum or no changes/reductions to their original answers. Any replacement of originally measured values must be planned for and done in conformity with your research concepts.

Entering data directly into the MS Excel sheet or data list sheets of statistical software packages is a source of frequent errors. It is easy to skip the column or row and then it is difficult to identify all the errors and correct them. However, even in MS Excel, it is easy to set up a form for purposes of entering the records one by one (see the video by United computers, 2013) video and set up some simple checks if you have at least basic programming skills. Using MS Access for this purpose would be easier. It is also possible to use suitable data entry freeware, which is widely available from the web.

Considerations in making high-quality transcriptions of qualitative data

The most common formats of qualitative data are written texts, interview data and focus group discussion data. In most cases, interview and discussion data are firstly digitally recorded and then transcribed. Transcription is a translation between forms of qualitative data, most commonly a conversion of audio or video recordings into text. If you intend to share your data with other researchers, you should prepare a full transcription of your recordings (Bucholtz, 2000).

There are several basic rules and steps in the process of making and checking a high-quality transcript from audio/video (Kuckartz, 2014):

Prevent mistranscription by recording high-quality data

The quality of interview data gathered by means of recorded interviews depends on both the skills of the interviewer and the quality of the audio-visual equipment. Taking steps to create audio recordings of good quality increases their usefulness. Good quality sound recordings should prevent mis-transcription and reduce the chance of sections of an interview remaining untranscribed due to poor sound quality. When recording an interview, consider the following (Bucholtz, 2000):

- The level of sound or picture quality needed;
- The budget available for equipment and related consumables;
- How quickly the technology being used will become redundant;
- Whether consent is in place to allow the fullest use of recordings;
- How the data created will be used;
- Whether data or information not allowed by consent can be excluded from recording;
- Whether the equipment will be simple to operate in the field.

Determine the transcription method

Transcription methods depend upon your theoretical and methodological approach and can vary between disciplines. Three basic approaches to transcription are (Bucholtz, 2000):

- **Focus on the content**
  This is also called the denaturalised approach, most like written language. The focus is on the content of what was said and the themes that emerge from that. This approach is used in sociological research projects.
- **Focus on what is said and how it is said**
  This approach is called the naturalised approach, which is most closely to speech. A transcriber seeks to capture all the sounds they hear and use a range of symbols to represent particular features of speech like the length of pauses, laughter, overlapping speech, turn-taking or intonation. This approach is usually employed in projects using conversation analysis.
- **Focus on emotional and physical language**
  In this approach detailed notes on emotional reactions, physical orientation, body language, use of space, as well as the psycho-dynamics in the relationship between the interviewer and interviewee are detailed. This approach is usually used in psycho-social research.

Choose between manually transcribing or with the help of speech recognition software (SRS)

SRS must "get used" to a speaker and can only be used if a high-quality recording is available. Gibbs (2007) recommends checking the utility and functionality of SRS software before using it.

When transcribing manually, you may sometimes hear something other than what an interviewee actually said. Listen carefully.

Determine the rules

- Determine a set of transcription rules or choose an established transcription system that is suited for the planned analysis;
- In setting up the rules, consider compatibility with the import features of QDA (Quality Data Analysis) software. For example, document headers and textual formatting, such as italics or bold, may be lost when transcripts are imported into software packages, and text formatted in two columns indicating speakers and utterances may also be problematic;
- All members who are doing the transcription should first agree on these rules;
- Write transcriber instructions or guidelines with required transcription style, layout and editing.

Transcribe

Transcribe the texts (or part of the texts) on the computer.

Check the transcription

Proofread, edit and modify the transcription, if necessary.

Protect your participants

- Anonymise data during transcription, or mark sensitive information for later anonymisation (see ‘Anonymisation’);
- When you assign the task of transcription to somebody else, make sure to take care of personal data protection before sending audio recordings and transcripts that contain personal or sensitive information. Draw up a non-disclosure agreement with the transcriber and encrypt files before transfer.

Choose a QDA-compatible file format

Format the transcription in such a way that your QDA (Qualitative Data Analysis) can be used optimally and files can be imported into the QDA software.

Choose a file format for long-term preservation

Save and archive the transcription in long-term preservation ready files such as *.rtf or *.pdf files (see ‘File Formats’).

3.2 Quantitative coding

Quantitative coding is the process of categorising the collected non-numerical information into groups and assigning the numerical codes to these groups. Numeric coding is shared by all statistical software and among others, it facilitates data conversion and measurement comparisons.

Closed-ended questions

For closed-ended questions in survey questionnaires, the coding scheme is often incorporated directly into the questionnaire and data is entered numerically. This process is automated in computer-assisted interviewing (CAPI, CATI, etc.), where an answer and its code are saved immediately into a computer in the course of data collection. Answers can also be coded on paper questionnaires when coders record codes in a designed spot of the questionnaire before they are digitalised. If the numberical codes are not incorporated in your questionnaire, set up a detailed procedure of how to code the different alternatives.

Open-ended questions and other textual information

More complex coding exercises, e.g. for textual answers in survey questionnaires, require an independent coding process with a clearly defined design: a coding structure and a procedure and schedule of exercises if there are several coders.

Documentation

The meaning of codes must be documented. Specialized analytic software (SPSS, SAS, STATA, etc.) lets the user assign labels directly to the codes. For the principles of the construction of labels, please, see the sub-section 'Organising variables'. If the software does not allow you to assign code labels directly to data, you have to document the codes in a separate document as part of the metadata.

Coding recommendations

In the boxes below you find coding recommendations which are inspired by ICPSR (2012).

Include identification variables

All identification variables should be included at the beginning of your data file. Identification variables usually include a unique identification of your study/data file, unique ID numbers of cases in your data file (e.g. ID of the respondent, ID of his/her household, etc.) as well as the identification of other characteristics essential for analysis (e.g. identification of different methods of data collection or sources, identification of the over-sample, etc.).

Make code categories exclusive and coherent throughout the database

Code categories should be mutually exclusive, exhaustive, and precisely defined. Ambiguity will cause coding difficulties and problems with the interpretation of the data. You should be able to assign each response of the respondent into one and only one category.

Preserve original information

Recording original data, such as age and income, is more useful than collapsing or bracketing the information. With original or detailed data, secondary analysts can determine other meaningful brackets on their own rather than being restricted to those chosen by others.

Document the coding schemes

Responses to closed-ended questions should retain the original coding scheme to avoid errors and confusion. For open-ended questions, investigators can either use a predetermined coding scheme or construct a coding scheme based on major categories that emerge in survey responses. Any coding scheme and its derivation should be reported in study documentation.

Check verbatim text data for data disclosure risk

Responses recorded as full verbatim (word for word) must be reviewed for disclosure risk and if necessary treated in accordance with applicable personal data protection

Check coding

It is advisable to verify the coding of selected cases by repeating the process with an independent coder. This provides means for verification of both the coder's work and the functionality of your coding scheme.

Distinguishing between major and lower level categories

If a series of responses require more than one field or if the response is very complex (for example a detailed description of one´s occupation), it is advisable to apply a coding scheme distinguishing between major, secondary and any possible lower level categories. The first digit of the code identifies a major category, the second digit can distinguish specific responses within the major categories, etc.

The International Standard Classification of Occupations (ISCO) (International Labour Organisation, 2016) is an example of such a hierarchical category scheme. An example of its use is given below.

Standardised coding schemes

The use of standardised classifications and coding schemes brings many advantages, e.g.:

- Economic and quality benefits as a result of adopting an existing structure which has a solid basis and has been verified in many studies;
- Comparability with data from other studies using the same concept;
- Comprehensibility for researchers who work with these concepts.

A disadvantage lies in the necessity to adapt your research intentions in line with the concept of the coding scheme.

Several standardised classification and coding schemes exist that you can use. For coding occupations it is the International Standard Classification of Occupations (ISCO) (International Labour Organisation, 2016), for coding education it is the International Standard Classification of Education (ISCED) (Unesco, 2011), for geographic territories it is the Nomenclature of territorial units for statistics (NUTS) (Eurostat, 2013), for economic activities it is the Statistical classification of economic activities (NACE) (Eurostat, 2008), for languages it is ISO 639.2 (Library of Congress, n.d.), for disease it is the International Classification of Diseases (ICD) (World Health Organisation, 2016), etc.

Example

Occupational classifications such as such as the International Standard Classification of Occupations (ISCO) (International Labour Organization, 2010) are examples of widespread standard coding schemes. ISCO is an example of a hierarchical category scheme.

Occupational information has several dimensions and in questionnaire surveys, these need to be collected in detail. This is, as a rule, done by means of one or more open-ended questions.

The current ISCO-2008 uses four-digit codes. In the table below you see some examples.

- 2 Professionals
- 21 Science and engineering professionals
- 211 Physical and earth science professionals
- 2111 Physicists and astronomers
- 2112 Meteorologists
- 2113 Chemists
- 2114 Geologists and geophysicists
- 212 Mathematicians, actuaries and statisticians
- 2120 Mathematicians, actuaries and statisticians
- 213 Life science professionals
- 2131 Biologists, botanists, zoologists and related professionals
- 2132 Farming, forestry and fisheries advisers
- 2133 Environmental protection professionals
- 214 Engineering professionals (excluding electrotechnology)
- 2141 Industrial and production engineers
- 2142 Civil engineers
- 2143 Environmental engineers
- 2144 Mechanical engineers
- 2145 Chemical engineers
- 2146 Mining engineers, metallurgists and related professionals
- 2149 Engineering professionals not elsewhere classified

Source: International Labour Organization (2016).

For an example of a recommended methodology of collection of information on occupations see Ganzeboom (2010).

Coding missing values

Not all the questions in a questionnaire are answered by all respondents, which results in missing values on a variable level in the data file (so-called item non-response). It is crucial for data integrity to distinguish at least the situations when values are missing, because the variable is not applicable to the particular respondents.

Furthermore, it is often useful for analyses to identify whether the value is missing because the respondent did not know the answer, refused to answer or simply did not answer or consider other reasons for missing values (see the example below). The information on missing values is always an important part of your documentation and promotes transparency of your research work. However, bear in mind that possibilities to differentiate between many different types of the missing values in analysis can be limited by the abilities of your software.

It is advisable to establish a uniform system for coding missing values for the entire database. Typically, negative values or values like 7, 8, 9 or 97, 98, 99 or 997, 998, 999, etc. (where the number of digits corresponds to the variable's format and the number of valid values) are used for numeric coding of missing values. The coding scheme for missing values should prevent overlapping codes for valid and missing values. For instance, whenever the digit zero is used for missing values, we should bear in mind that zero may represent a valid value for many variables such as personal income.

Example

Respondents in surveys sometimes do not answer all questions in a questionnaire. It is advisable to distinguish between various reasons that data went missing (ICPSR, 2012). The following situations are distinguished in survey research (frequently used acronyms are bracketed):

- No answer (NA): The respondent did not answer a question when he/she should have;
- Refusal: The respondent explicitly refused to answer;
- Don't Know (DK): The respondent did not answer a question because he/she had no opinion or did not know the information required for answering. As a result, the respondent chose 'don't know', 'no opinion' etc. as the answer;
- Processing Error: The respondent provided an answer but, for some reason (interviewer error, illegible record, incorrect coding etc.), it was not recorded in the database.
- Not Applicable/Inapplicable (NAP/INAP): A question did not apply to the respondent. For example, a question was skipped following a filter question (e.g. respondents without a partner did not answer partner-related questions) or some sets of questions were only asked of random subsamples.
- No Match: In this case, data are drawn from different sources, and information from one source cannot be matched with a corresponding value from another source.
- No Data Available: The question should have been asked, but the answer is missing for a reason other than those above or for an unknown reason.

Training coders to prevent coder variance

Coders may vary in the way they assign codes to variable values, i.e. each of them uses the same coding scheme in a slightly different way. This results in so-called "coder variance". Coder variance is a specific source of non-sampling error (i.e., error additional to the statistical "sampling" error) and may cause systematic deviations of the sample.

Coding of textual information is a complicated cognitive process and the coder may pose a significant influence on the information that appears in the database, as well as become a source of systematic error. That is why the implementation of complicated coding schemes often requires the construction of a theoretically and technically well-founded design and requires specific coder's competencies and training.

3.3 Qualitative coding

> *Coding is a way of indexing or categorizing the text in order to establish a framework of thematic ideas about it | Gibbs (2007).*

In qualitative research, coding is "how you define what the data you are analysing are about" (Gibbs, 2007). Coding is a process of identifying a passage in the text or other data items (photograph, image), searching and identifying concepts and finding relations between them. Therefore, coding is not just labeling; it is linking of data to the research idea and back to other data...

The codes which are applied enable you to organise data so you can examine and analyse them in a structured way, e.g. by examining relationships between codes.

Approaches to coding qualitative data

A basic division between coding approaches is concept-driven coding versus data-driven coding (or open coding). You may approach the data with a developed system of codes and look for concepts/ideas in the text (concept-driven approach) or you can look for ideas/concepts in the text without a preceding conceptualisation and let the text speak for itself (data-driven coding). Investigators can either use a predetermined coding scheme or review the initial responses or observations to construct a coding scheme based on major categories that emerge.

Both methods require initial and thorough readings of your data and writing down which patterns or themes you notice. A researcher usually identifies several passages of the text that share the same code, i.e. an expression for a shared concept.

An example

A code in a qualitative inquiry is most often a word or short phrase. In the table below an example (Saldaña, 2013) is given.

Raw data	Preliminary codes	Final code
The closer I get to "retirement age" the faster I want it to happen. I'm not even 55 yet and I would give anything to retire now. But there's a mortgage to pay off and still a lot more to sock away in savings before I can even think of it. I keep playing the lottery, though, in hopes of dreams of early winning those millions. No retirement luck yet.	* retirement age* financial obligations dreams of early retirement	RETIREMENT ANXIETY

Expert tips

> *Any researcher who wishes to become proficient at doing qualitative analysis must learn to code well and easily. The excellence of the research rests in large part on the excellence of the coding | Strauss (1987).*

Tip 1: Document the meaning of codes

The meaning of codes must be documented in a separate file. Make short descriptions of the meaning of each code. It is helpful to you and also to other researchers who will have access to your data/ analysis. What you need to know about your codes (Gibbs 2007):

- the label or name of the code
- who coded it (name of the researcher/coder)
- the date when the coding was done/changed
- definition of the code; a description of the concept it refers to
- information about the relationship of the code to other codes you are working with during the analysis.

Tip 2: Prevent coder variance

Coding textual information is a complicated cognitive process and the coder is necessarily a significant influence on the coding process. For each study coding procedures must be carefully planned and a specific coding design and guidelines must be established. Coders must undertake a training, where they are instructed about the specific coding design and coding rules. A part of coding procedures is concerned with reviewing the quality of the coding process. According to Gibbs (2007) several techniques to control coder reliability exist:

- **Checking the transcription**
  An independent researcher goes through coded texts and considers the degree to which coders differed from each other.
- **Checking for definitional drift in coding**
  If you code a large dataset the data at the beginning may be coded slightly different than material coded later. Check the the whole dataset for the definitional drift. Have good notes with descriptions of individual codes.
- **Working in a team**
  If there are multiple people working in a team, individual members can check each other´s coding.

3.4 Weights of survey data

When conducting a survey, having a representative sample of the population is of paramount importance. But in practice, you are prone to over-sample some kinds of people and under-sample others. Weighting is a statistical technique to compensate for this type of 'sampling bias'. A weight is assigned to:

- Reflect the data item's relative importance based on the objective of the data collection;
- Take into account the characteristics of sampling design;
- Reduce bias arising from nonresponse when the characteristics of the respondents differ from those not responding;
- Correct identifiable deviations from population characteristics.

Each individual case in the file is assigned a certain coefficient – individual weight – which is used to multiply the case in order to attain the desired characteristics of the sample.

Different types of weights and their different purposes

Several types of weights have different purposes and a different impact on data analysis.

An answer to the question whether or not to use weights is not straightforward. For particular methods of analysis (e.g., estimating associations, regressions, etc.) using weights may be dysfunctional. There are also general theoretical and methodological issues which discourage some researchers from using weights. However, different types of weights are useful for different purposes. In some situations, it is necessary to take an appropriate weight into account in your analysis (see several types of weighting below).

In all cases, if there are any weights in your data file, the rationale and calculation of the weights must be detailed in the data documentation.

Design weights

Design weights are constructed in order to mutually adjust individual units' probabilities of being sampled, which are normally not equal when complex sampling procedures combining multiple methods (stratification, group sampling) in several stages are implemented. For example, we want to adjust the probabilities of being sampled for all respondents in households. While individuals are the sampling units, households are sampled in the first stage. Therefore, respondents' probabilities of being selected depend on the number of household members.

To solve these differences in sampling probabilities we have to compute design weights. The design weights are equal to the inverse of the probability of inclusion in the sample. The sum of all design weights should be equal to the total number of units in our population.

Non-response weighting

During the implementation of a survey, we are normally not able to get a response from some of the targeted respondents we sampled due to:

- Their refusal;
- Our failure to contact them;
- Other administrative reasons.

Response rates differ between various population groups and those inequalities can be compensated for by weighting.

Post-stratification weighting

The way certain characteristics such as sex, age, and education of your sample population are distributed may differ from the way it is distributed in the actual population. For example, your sample may consist of 66 percent men when they make up only 48 percent of the population. Post-stratification weighting is done in order to achieve a distribution equal with that of such known characteristics of the population. It is called a post-stratification weight because it can only be computed after you have collected all of your data. Stratification comes from the various known strata (such as age group or sex distribution) of the population.

Population size weighting

Different groups may be represented in the database in different proportions than they are in reality. Such discrepancies are normally compensated through weighting. For example, international data files combine data from various countries. However, similarly, large surveys are usually implemented in each of these countries, although their total populations are radically different in size. If we want to analyse data about large populations, such as in Europe, then we have to adjust the proportions in the representation of individual European countries.

Combined weighting

The data file may include several different types of weights for different purposes. Subsequently, they are combined into a final, combined weight.

An example: Comparison of weighted and non-weighted data

Source: Data files from the ESS, round 8, Czech Republic (European Social Survey, 2016).

Variable name: netusoft

Question: How often a respondent uses internet

In the first column, no weight was applied.

In the second column, the Design Weights (DWEIGHT) are adjusted for different selection probabilities.

	No weight		Design weight	
	Frequency	Valid Percent	Frequency	Valid Percent
1 Never	244	10,8	187	8,2
2 Only occasionally	162	7,1	155	6,8
3 A few times a week	302	13,3	284	12,5
4 Most days	384	16,9	379	16,6
5 Every day	1177	51,9	1271	55,8
Total	2269	100	2277	100
System missing	31		23	
Total	2300		2300	

Distribution of weights

If the weight of a case equals 1 then the values measured are not adjusted. In the case of post-stratification weights both high or low numbers indicate either large deviations of the sample from the target population, poor quality of the weight or both. It is desirable that the large part of values of the weighting variable is close to 1.

Weights constructed by others

Is there any weighting variable in your working data file? If yes and you are not the author of the weight, never use it without knowledge of its origin and purpose. You should always thoroughly explore the distribution of the weighting variable and its impact on distributions of other selected variables from the data file.

An example: Using weights in European Social Survey data

The following table provides an illustration of using weights in the data from the European Social Survey (n.d.) (ESS). There are three different weights available in the ESS Source Main Questionnaire data file (see European Social Survey, 2014):

1. The design weight takes into consideration the different probabilities of being sampled given the sampling methods implemented in individual countries;
2. The post-stratification weight corrects for the differences of the sample from selected population characteristics caused by other sampling and non-sampling errors;
3. The population size weight corrects the fact that the individual countries' sample sizes are very similar while there are large variations in the size of their actual populations.

Different types of data analysis then require the use of different weights or their combinations. When analysing data from one country alone or comparing data of two or more countries, only the design weight or the post-stratification weight needs to be applied. When combining different countries, design or post-stratification weights in combination with population size weights should be applied.

	Example – voter turnout (% of respondents voting in the last election)	Weights to be used	
		Design weight / Post-stratification weight	Population weight
To examine data from a single country – whether a single variable or a cross-tabulation	Voter turnout in Germany	X	
	Voter turnout in Germany by age and gender	X	
To compare results for two or more countries separately – without using totals or averages	Compare voter turnout in France, Germany, and the UK	X	
To combine countries – whether on a single variable or via a cross-tabulation	Voter turnout in Scandinavia	X	X
	Voter turnout in the EU	X	X
	Voter turnout across all countries participating in the ESS	X	X
	Compare voter turnout between EU member states and accession countries	X	X
	Voter turnout by age group across all ESS participating countries	X	X

Source: European Social Survey, 2014.

3.5 File formats and data conversion

We use software for creating text documents, websites, databases, photos, 3D models, and movies. Software developers regularly release new versions of their products. It is not self-evident that the new software supports the use of files created with earlier software versions (compatibility). And some software packages even disappear completely from the scene. Conversions of file formats may be costly or result in loss of information or a reduction of data quality. This is exactly why the choice of file formats should be planned carefully.

Short-term data processing: file formats for operability

File format choice depends on your research phase. Choices for short-term data processing may differ from the choices you make for long-term data preservation.

For the reasons of short-term operability, it is advisable to choose a file format that is associated with the specific software that you intend to use for data analysis. Following discipline-specific standards and customs is generally the way to go. However, you should take into consideration how widespread these standards are and to what extent they will allow data processing by others than peers in your own discipline.

Proprietary file formats are owned and copyrighted by a specific company. Their specifications are usually not publicly available and their future development results from decisions and situation of their owner. Thus, the risk of obsolescence is high. However, some proprietary formats, such as Rich Text Format (*.rtf), MP3, MPEG, JPG, MS Excel (*.xls), SPSS (*.sav, *.por), STATA (*.dta) are widely used and you may assume that they will be useful for a reasonable time.

Learn more about suitable file formats for short-term data processing

Below we give an overview of the data analysis packages/file formats which are used most and which are suitable for short-term data processing.

Quantitative (statistical) data analysis packages

MS Excel (*.xls), SPSS (*.sav, *.por), R and STATA (*.dta) are widely used and you may assume that they will be useful for a reasonable time.

Some software also provides so-called portable formats which allow easy transfer of data between different versions of the software of the same brand, often including versions for different platforms (MS Windows, Mac, Linux...). For example, SPSS system files with the *.sav extension and SAS files with the *.sd7 extension (SAS Version 7 or 8 data file) are associated with the concrete version of the SPSS or SAS software. Instead of them, you may use “portable” SPSS files with the *.por extension or “transport” SAS files, which are compatible with different versions of this software running on different platforms.

Qualitative data analysis packages

Qualitative research data like transcribed interviews of focus group sessions, audio recordings, still images, photographs, ethnographic diaries and various types of written texts are usually transcripted into one of the following types of formats: *.docx, *.rtf, *.pdf, *.mp3, *.wav, *.jpeg and many others.

For the purposes of qualitative data analysis (QDA), textual data may be analyzed in special QDA software packages such as NVivo, ATLAS-ti, and MAXQDA. In such packages researchers are allowed to code their textual data, i.e.indicate parts of text related to same concepts, create a structure of concepts etc. In the process of coding, a “coding tree” emerges along other pieces of information, for example, notes and memos. Common QDA packages have export facilities that enable a whole ‘project’ consisting of the raw data, coding tree, coded data (Also see ‘Coding qualitative data’), and associated memos and notes to be saved.

Long-term data preservation: file formats for the future

Standard, open and widespread formats are advisable for long-term storage as they typically undergo fewer changes. Contrary to proprietary formats (see above) specification of open formats is publicly available. Some of them are standardised and maintained by a standards organisation and we may assume that their readability in the future is ensured. Examples of open formats are PDF/A, CSV, TIFF, ASCII, Open Document Format (ODF), XML, Office Open XML, JPEG 2000, PNG, SVG, HTML, XHTML, RSS, CSS, etc.

Learn more about file formats for long-term preservation

Quantitative data preservation

Long-term preservation of quantitative data is typically best off with simple text (ASCII) formats accompanied by a structured documentation file with information about the variables included, their position in the file, formats, variable labels, value labels etc.

In terms of location of variables in the file, we distinguish between fixed and free formats.

Fixed format In a fixed format, variables are arranged in columns and their exact positions, i.e. the start and end of each variable, are known.

Free format In a free format data for each variable is separated by blanks or specific characters, e.g. by tab space or a dash. If the character separating variables is used within an item, then it needs to be formatted specifically and separated from the surrounding text (as a rule, by quotation marks).

There exist several extensions for simple text formats, e.g. *.txt., *.dat and *.asc are used for both fixed and free formats, *.csv. is used for fixed format.

Qualitative data preservation

Qualitative data analysis software packages such as NVivo, ATLAS-ti, and MAXQDA have export facilities that enable a whole 'project' consisting of the raw data, coding tree, coded data, and associated memos and notes to be saved. For archiving such data, the raw data, the final coding tree, and any useful memos should be exported (UK Data Service, 2017)

Digital versions of documents are usually kept in the PDF/A format. This is an official archiving version of the PDF format as defined by the ISO 19005-1:2005 standard. It guarantees independence from the platform and includes all display information (including fonts, colours, etc.). XMLP format is a widespread standard for metadata. Structured textual documentation should, again, be saved in a simple text format, with tags and in line with a standard structure (e.g., DDI).

For audio files the recommended longterm format is WAV, video files are advised to be stored in MXF (Material eXchange Format) and JPEG2000 (Fleischhauer, 2010).

A very useful tool for searching an appropriate format for different types of data is provided by the UK Data Service (2017b) in the table of Recommended file formats.

Data conversion and possible data loss

Data files, depending on the nature of the data, are based on either text or binary encoding or both. Binary encoded information can be read only by specialised software, text information is universal and can be read by a wide range of different software including text editors.

It is advisable to store your data for use in the future, which means converting them from a current data format to a long-term preservation format. Most software applications offer export or exchange formats that allow a text-formatted file to be created for importing into another program. A typical example is Microsoft Excel, which through the 'Save As' command, can save spreadsheet data in comma delimited format (*.csv or comma separated values). The structure of the rows and columns is preserved through commas and line returns. However, multiple worksheets must be saved as separate *.csv files and any text formatting or macros in the native format will be lost on conversion.

During the process of data conversion, important pieces of information may be lost:

- In the conversion of a statistical dataset (i.e. survey data), parts of the dataset may be lost, same as missing data definitions, decimal numbers, changes in data formats (e.g., numerical into string data type), data also may be truncated;
- In case of texts, i.e. transcriptions of speech, editing such as highlighting, bold texts, headers, footers may be lost;
- In case of images a reduction of resolution, loss of layer, colours may be lost;
- In converting audiovisual data file conversion may reduce sound quality;
- Some file formats are constructed specifically to save space. However, this is done by a reduction of information and data quality. For example, .jpg removes details from images, while .tiff bears full information. Similarly, .mp3 is a lossy format for audio data, while .wav keeps detailed information.

For this reason, the conversion itself should be done by a researcher familiar with the data, so he or she can check for potential undesirable changes in the data that occurred as a result of the conversion.

Due to differences in national character sets you should pay attention also to character coding. Some coding systems (e.g., Windows 1250) do not cover all character sets at the same time. As a result, an adequate language environment (Central European languages) has to be set to ensure correct display, which cannot be done at all times. Other coding systems (e.g., UTF 8) allow correct display of symbols of several character sets simultaneously.

TIP: Plan ahead to simplify data publication

Different data archives have different preferred formats. Knowing about these preferred formats in advance can save you time later when you want to archive and publish your data. Usually preferred formats are frequently used, independent of specific software, and have open specifications (see for instance information by DANS (n.d.) on preferred formats).

3.6 Data authenticity

Processing and analysis of data inevitably result in a number of edits in the data file. However, it is necessary to preserve the authenticity of the original research information contained in the data throughout the whole data lifecycle.

There are many possible types of changes in the data:

- Data cleaning procedures may be implemented;
- Errors are often found and corrected;
- New variables may be constructed;
- New information may be added from external sources;
- File formats may be changed;
- New data may be included;
- The data file structure may be changed for the purpose of increasing operability, etc.

As a result of above-mentioned data management processes, several different **versions** of the data file are usually created. They are important, as they allow you to step back to versions before particular changes were made. Versions may be used simultaneously for different purposes or replace one another. When data files are being published to make them widely available, the treatment of errors, inclusion of new data and/or changes in a data file structure may result also in the publication of new **editions** of the same data file which may substantially differ in their content (e.g. when new country data are included into an international data file).

Best practices for quality assurance, version control and authenticity

Version and edition management will help to:

- Clearly distinguish between individual versions and editions and keep track of their differences;
- Prevent unauthorised modification of files and loss of information, thereby preserving data authenticity.

Best practices

The best practice rules (UK Data Service, 2017a; Krejčí, 2014) may be summarised as follows:

- Establish the terms and conditions of data use and make them known to team members and other users;
- Create a 'master file' and take measures to preserve its authenticity, i.e. place it in an adequate location and define access rights and responsibilities – who is authorised to make what kind of changes;
- Distinguish between versions shared by researchers and working versions of individuals;
- Decide how many versions of a file to keep, which versions to keep (e.g. major versions rather than minor versions (keep version 02-00 but not 02-01)), for how long and how to organise versions;
- Introduce clear and systematic naming of data file versions and editions;
- Record relationships between items where needed, for example between code and the data file it is run against, between data file and related documentation or metadata or between multiple files;
- Document which changes were made in any version;
- Keep original versions of data files, or keep documentation that allows the reconstruction of original files;
- Track the location of files if they are stored in a variety of locations;
- Regularly synchronise files in different locations, such as using MS SyncToy (2016).

Version control

Version control can be done through:

- Uniquely identifying different versions of files using a systematic naming convention, such as using version numbers or dates (date format should be YYYY-MM-DD, see 'File naming');
  - Record the date within the file, for example, 20010911_Video_Twintowers;
  - Process the version numbering into the file name, for example, HealthTest-00-02 or HealthTest_v2;
  - **Do not** use ambiguous descriptions for the version you are working on. Who will know whether MyThesisFinal.doc, MyThesisLastOne.doc or another file is really the final version?
- Using version control facilities within the software you use;
- Using versioning software like Subversion (2017);
- Using file-sharing services with incorporated version control (but remember that using commercial cloud services such as the Google cloud platform, Dropbox or iCloud comes with specific rules set by the provider of these services. Private companies have their own terms of use which applies for example to copyrights);
- Designing and using a version control table. In all cases, a file history table should be included within a file. In this file, you can keep track of versions and details of the changes which were made. On the following page is an example which was taken from the UK Data Service (2017c).

Example of a version control table

Title:	Vision screening tests in Essex nurseries
File Name:	VisionScreenResults_00_05
Description:	Results data of 120 Vision Screen Tests carried out in 5 nurseries in Essex during June 2007
Created By:	Chris Wilkinson
Maintained By:	Sally Watsley
Created:	04/07/2007
Last Modified:	25/11/2007
Based on:	VisionScreenDatabaseDesign_02_00

Version	Responsible	Notes	Last amended
00_05	Sally Watsley	Version 00_03 and 00_04 compared and merged by SW	25/11/2007
00_04	Vani Yussu	Entries checked by VY, independent from SK	17/10/2007
00_03	Steve Knight	Entries checked by SK	29/07/2007
00_02	Karin Mills	Test results 81-120 entered	05/07/2007
00_01	Karin Mills	Test results 1-80 entered	04/07/2007

Versioning new data types

Generally, the goal of version management is to enable reproducibility and support trustworthiness by allowing all transformations in the data to be traced. But difficulties emerge connected to versioning of "new data" as these data are (compared to "traditional data") more frequently or even continuously updated. A good example are collections of Tweets (e.g., for a certain hashtag) as individual posts may be modified or deleted. As the contents of these data are continuously changing and if archived data are expected to reflect such changes (e. g. deleting posts from data set if they were deleted from platform) the result is an increasing number of versions. Consequently, it is necessary to develop a systematic plan to create and name new versions of constantly changing datasets, or find new solutions for streaming data.

Both researchers and repositories can learn from the fields where versioning of dynamic data is already established, such as the field of software development. The most common version control software in software development is Git. Some of the established repositories, such as Zenodo and FigShare or the Open Science Framework, now offer integration with GitHub, so that every version of data sets in those repositories can be recorded through it. A new project called Dolt is developing version control specifically for data which is particularly interesting for dynamic data sets, such as social media data.

To identify the exact version of a dataset as it was used in a specific project or publication, the Research Data Aliance (RDA) suggests that every dataset is versioned, timestamped, and assigned a persistent identifier (PID). In the case of Big Data, however, the RDA warns against excessive versioning: "*In large data scenarios, storing all revisions of each record might not be a valid approach. Therefore in our framework, we define a record to be relevant in terms of reproducibility, if and only if it has been accessed and used in a data set. Thus, high-frequency updates that were not ever read might go - from a data citation perspective - unversioned.*"

3.7 Wrap up: Data quality

> *The quality of a survey is best judged not by its size, scope, or prominence, but by how much attention is given to [preventing, measuring and] dealing with the many important problems that can arise | American Association for Public Opinion Research (2015) (AAPOR).*

How you organise, document and process data has a clear impact on data quality and thus also on reliability and adequacy of research findings. In scientific research even small things matter. Data organisation, documentation and processing procedures are not an exception and this is true for quantitative as well as qualitative research. A systematic approach and punctuality in data management (Krejčí, 2010) are awarded by the following:

- Preventing errors and false findings;
- Smooth course, time efficiency and transparency of your own research work;
- Establishing assumptions for effective re-use of research data outside of the original research team.

While in quantitative research the quality is closely linked to standardization and control over the research situation the prevailing approach in qualitative research is different.

> *In qualitative research, discussions about quality in research are not so much based on the idea of standardization and control, as this seems incompatible with many qualitative methods. Quality is rather seen as an issue of how to manage it. Sometimes it is linked to rigour in applying a certain method, but more often to the soundness of the research as a whole | Flick (2007).*

A complex approach to data quality

In previous chapters, you have become familiar with a number of procedures and rules for the development of an appropriate data file structure, development of rich metadata and ensuring the data integrity and authenticity. At the same time, however, we should bear in mind that the data management is always an integral part of much more complex research work.

The quality of the outcome is achieved through the quality of the production process (Krejčí, 2010). Scientific research is not an exception. Thus the quality stems from professionalism based on continuous improvement. Data management is one important part of such processes. As such it is interconnected and influenced by other processes within the system and should contribute to a common long-term objective of continuous improvement of a research work within the research organisation.

> *The mechanical quality control of survey operations such as coding and keying does not easily lend itself to continuous improvement. Rather, it must be complemented with feedback and learning where the survey workers themselves are part of an improvement process | Biemer & Lyberg (2003).*

In addition, quality always involves a number of different dimensions. Quality is often defined as "fitness to use". However simple this sounds, it provides a point of departure for a comprehensive approach to data quality. The results must not only be accurate but must be delivered in time, understandable and clear, and meet other potential users' needs, e.g. comparability and coherence with other databases. Moreover, it must be also cost-efficient.

See the section on the next page for an example of how total quality management is handled by the European Statistical System (Eurostat, 2017).

Total Quality Management of the European Statistical System (ESS)

So-called models of Total Quality Management (TQM) recognise multiple dimensions of quality. They:

- Set the required characteristics of a final product;
- Define partial goals;
- Elaborate the individual dependable processes to achieve them;
- Identify and treat problematic points;
- Specify control points;
- Have procedures for quality monitoring, learning processes, and feedback-loops for ensuring continuous improvement.

Some of the guiding principles on which the Quality Assurance Framework of the European Statistical System (n.d.) is based are stated in the table below.

Data management procedures are an important part of such TQM models, building on similar principles and having the same goals.

Cost-effectiveness	Resources are used effectively.
Relevance	European Statistics meet the needs of users.
Timeliness and Punctuality	European Statistics are released in a timely and punctual manner.
Accuracy and Reliability	Source data, intermediate results, and statistical outputs are regularly assessed and validated.
Coherence and Comparability	European Statistics are consistent internally, over time and comparable between regions and countries; it is possible to combine and make joint use of related data from different sources.
Accessibility and Clarity	European Statistics are presented in a clear and understandable form, released in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance.

3.8 Adapt your DMP: part 3

This is the third of seven 'Adapt your DMP' sections in this tour guide. After working on this chapter, you should be able to define the processing that will be done to your data during the project.

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

Versioning

- How will you version your data files (and scripts) during the project?
- Will you create and/or follow a convention for versioning your data?
- Who will be responsible for securing that a "Masterfile" will be maintained, documented and versioned according to the project guidelines?
- How can different versions of a data file be separated?

Interoperability

In order to be able to link your work to other research, it might be useful to build on established terminologies as well as commonly used coding and soft- and hardware wherever this is possible.

- Which software and hardware will you use? How does this relate to other research?

If applicable:

- Will established terminologies/ontologies (i.e. structured controlled vocabularies) be used in the project? If not, how does yours relate to established ones?
- Which coding is used (if any)? How does this relate to other research?

Data Quality

- How will data quality be evaluated?
- What data quality control measures will be used?"

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.