Data Processing and Documentation: Generating High Quality Research Data in Quantitative Social Science Research

Netscher, Sebastian (Ed.); Eder, Christina (Ed.)

gesis Leibniz-Institut für Sozialwissenschaften

GESIS Papers

2018|22

Data Processing and Documentation

Generating High Quality Research Data in Quantitative Social Science Research

Sebastian Netscher & Christina Eder (eds.)

GESIS Papers 2018|**22**

Data Processing and Documentation

Generating High Quality Research Data in Quantitative Social Science Research

Sebastian Netscher & Christina Eder (eds.)

GESIS – Leibniz-Institut für Sozialwissenschaften 2018

GESIS Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Postfach 12 21 55
68072 Mannheim
Telefon: (0621) 1246 - 0
Telefax: (0621) 1246 - 100
E-Mail: christina.eder@gesis.org

ISSN: 2364-3781 (Online)
Herausgeber,
Druck und Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

Content

1 DATA PROCESSING AND DOCUMENTATION

Sebastian Netscher & Christina Eder

For researchers in empirical social sciences hardly any day passes without handling (research) data, i.e. information on individuals, institutions, countries, situations or events. Indeed, structured data, in terms of a well-defined and clean dataset, are at the heart of empirical research. It is a prerequisite for successfully realizing the research aim of origin. For that purpose, research data must be accessible, understandable as well as interpretable, enabling data analyses and generalization of research results achieved. Furthermore, research data and the process of data generation must be made transparent, enabling to evaluate data-quality as well as to reproduce data. This is also a criterion of good scientific practise: Well-documented and clean research data ensure reproducible research results, as required, e.g., by academic journals. Likewise, it also supports data sharing, as required nowadays by more and more research funders in the context of *Open Access* and FAIR principles of data (Force11; Wilkinson et al. 2016).

Research data must therefore be well-processed and documented. *Data processing* refers to activities involving data generating, i.e. composing a well-structured, consistent and anonymized dataset that can be (re-)used in data analysis. It maintains data-quality and ensures interpretability as well as reproducibility of research data. Data processing directly relates to data documentation, because all steps undertaken in processing research data must be documented in great detail. More generally, *data documentation* summarizes the description of research data, their content and structure, the context of data collection, data processing as well as the research data themselves. It tells a story about the research data, increases transparency in the research project and ensures data reproducibility.

The present document bases on a training concept, developed by members of the *Data Archive for the Social Sciences* at *GESIS – Leibniz-Institute for the Social Sciences*, to support researchers in the social sciences producing research data of high quality. Structured in three chapters, this document offers guidelines and best-practices for generating a well-processed and documented dataset. In the first chapter, we briefly introduce the purposes of data reproduction and data sharing. We discuss different reasons to care about data processing and documentation in the context of good scientific practices. In the second chapter, we then investigate relevant steps of data processing. In particular, we debate the composition of well-structured datasets, the definition of variables and codes, the validation of as well as consistency checks on the data at hand, and the anonymization of so-called personal information. In the third and final chapter, we focus on data documentation, discussing the role and relevance of metadata, the description of the research project and its data on the study level as well as on the variable level.

2 DATA SHARING AND REPRODUCIBILITY OF SCIENTIFIC RESEARCH

Kerrin Borschewski & Thomas Ebel

Data sharing and reproducible research seem to be on everyone's lips nowadays for good reasons. The goals of reproducible research and data sharing are to enable the comprehensibility of research results and methodology and in general to simplify the workflow, e.g. for researchers, research teams, re-users and reviewers. The advantages of data sharing and reproducible research are manifold as we will see in the next section. They concern primary researchers as well as the scientific community as a whole and society in general. Let's start by asking the question: "What is data sharing and what is reproducible research"? Data sharing means the practice of making data available to others. Sharing research data often serves the purpose of enabling other researchers to reproduce and verify research results and allow data to be re-used in other contexts and/or answer other research questions than the ones that led to the original data gathering. Reproducibility in the social sciences refers to the reproducibility of research results. In order to achieve this goal, data and scripts must be provided and documented in a sufficient way for other researchers to recreate the findings (Gandrud 2015).[1]

Definition for *Reproducible Research*

"The idea [of reproducible research] is that the final product of research is not only the paper itself, but also the full computational environment used to produce the results in the paper such as the code and data necessary for reproduction of the results and building upon the research."

(Xie 2015: 5)

To render data sharing and reproducibility of research possible, certain requirements must be met. These requirements concern many different aspects. To give an idea of the different issues and angles we need to consider for data sharing and re-use, here are some examples (please note that the following list is not exhaustive): Our research materials (such as data, scripts, questionnaires and according metadata) need to be archived, ideally for the long-term, and must be made publicly available. The methods we apply must be standardized. All our research materials and research processes must be documented in detail. In most cases we need informed consent from study participants before archiving and sharing their data. The file formats of our materials must be appropriate for data sharing and data re-use.

2.1 Why Data Sharing and Reproducible Research

Now, we might ask ourselves "why should we share our data and make our research reproducible?". That is where yet another concept comes into play, the one of Open Access. The idea of Open Access is becoming more and more prominent among all scientific disciplines. Open Access refers to the open accessibility of publications, methodologies, data, metadata and anything concerning the research process. In general, to render re-use of data possible, research data and corresponding research materials must be registered with a persistent identifier (PID), made accessible and open for exchange. This fosters access to scientific data, helps improve the quality of research results, fosters collaboration,

[1] Please note, there is an abundance of – often contradictory and ambivalent – definitions for the terms data re-usability, replicability and reproducibility. In this document, we decided to use all of these terms synonymously.

ensures greater efficiency (duplication of efforts can be avoided), speeds up innovations, improves transparency of scientific processes and boots researchers' reputations.

Open Access to existing primary research data and associated metadata is seen as good scientific practice. For safeguarding good scientific practices, the German Research Foundation (DFG) proposes that "primary data as the basis for publications shall be securely stored for ten years in a durable form in the institution of their origin" (DFG 2013: 74), or in an institution that is qualified to secure the data (and referring code). The unrestricted flow of knowledge is the basis for excellent and innovative research and serves therefore not only the scientific community, but also society in total. Hence, research funding institutions demand that scientific data are made open to access and for re-use by other scientists. To give you an example of those demands, we would like to provide you with citations from guidelines of EU Horizon 2020 Programme (EC 2016: 9f.):

> "participating projects [...] must deposit the research data [...] preferably in a research data repository [...] as far as possible, projects must then take measures to enable third parties to access, mine, exploit, reproduce and disseminate (free of charge [...]) this research data."

If we start thinking now "well, our funding agency has none of those demands, we don't need to worry about those issues", we will be disappointed. Not only research funding institutions have increased their demands to the accessibility of data and according research materials, more and more journals also require authors to share/publish their data, scripts and according metadata. In Germany, for example, the *Zeitschrift für Soziologie* cooperates with *datorium,* a data repository run by the *Data Archive for the Social Sciences* at *GESIS – Leibniz-Institute for the Social Sciences,* concerning the publication of research materials. And, on an international scale, we would like to point out the JETS (Journal Editors' Transparency Statement), which was signed by 27 journals, among them the *American Journal of Political Science*:

> "Require authors to ensure that cited data are available at the time of publication through a trusted digital repository [...] Require authors to delineate clearly the analytic procedures upon which their published claims rely, and where possible to provide access to all relevant analytic materials. If such materials are not published with the article, they must be shared to the greatest extent possible through institutions with demonstrated capacity to provide long-term access." (DA-RT 2014).

Recap: Advantages of data sharing and reproducible research

for the Scientific Community

- transparent and credible research
- wide-ranging access to data
- valid and reproducible results
- enables cumulative research and encouragement of collaboration
- efficient use of public funds
- hampering of scientific misconduct and fraud
- advancement of innovations
- avoidance of redundancy

for collaborating researchers/teams and individual researcher

- economizing
- veracity of results (working with the proper, latest versions of data/scripts)
- easing of workflow (e.g. by having a proper versioning system)
- proof of transparent and valid conduct
- easing of re-using one's own data
- credit for the researcher's work

2.2 How to research in a shareable / reproducible fashion

In order to do research in a sharable and reproducible fashion, a number of aspects must be considered. Within this document, the whole scope of these aspects cannot be discussed in detail. Therefore, we restrict ourselves to introducing some strategies to achieve reproducible research. For ideas and an in-depth look we recommend Christensen and Soderberg (2015) or Gentzkow and Shapiro (2014).

In general, research materials such as data, scripts and study documentations should be made publically available. This way, other scientists can perform replication attempts or their own research using those materials. There are, however, reasons not to publish materials in some context. These opt-out reasons include, but are not limited to, data privacy and copyright issues (Corti et al. 2014).

Even if publication is not feasible or desired, it is crucial to document the research process thoroughly. Imagine that we would like to re-use our own scripts or datasets a few years after the project ended, possibly for a similar project for which some of the materials could be re-used, or in order to replicate our results to prove they were computed correctly. How hard or easy would it be to understand the code? Which steps need to be taken to get from the input data to the final research results? Were there any manual actions or do the scripts run automated (i.e. without the user having to do intermediary steps outside the script)? Good documentation helps to make sense of research materials, be it the primary researcher or any other person wanting to understand the research process.

The more automated the workflow is, the less error-prone is reproduction. If all research results are produced by simply running a script with little or no adjustment to it, re-users can focus on understanding the code and, as a matter of course, also the data and the method(s) used instead of trying to get rid of error messages. One aspect of re-usability is to ensure that our scripts run on different platforms and in different environments. Are we using the base version of our statistical software or did we use special software, libraries, or packages? If so, we need to make sure we document which additional software, libraries or packages are necessary to process the code. Did we use absolute paths in our scripts? If that's the case, we must document this and the lines where adjustments need to be made, in case someone else uses the code. In general we need to write code in a fashion that allows others to re-use our materials with as little effort as possible.

Recap: Methods to foster reproducibility and shareability

- scientific data and scripts should be made available, wherever possible
- all materials should be documented
- workflows should be automated as much as possible (reducing the amount of manual/undocumented interference)
- reproducibility of workflows and scripts should be ensured for other working environments (e.g. the computer system of a third party)

2.2.1 Workflow

Once we have collected the data that form the basis of your research the resulting dataset should remain unchanged. Else we risk losing our raw data through unintended actions. We would no longer be able to prove (to ourselves or others) that our results are a genuine outcome of our analyses and not mere artefacts resulting from data manipulation or coding errors. To avoid this, we must store the original dataset as 'read-only' and continue to work with separate copies. Sometimes it even makes sense to store different versions of datasets after certain manipulations have been made.

Ideally, all changes should be performed through scripts. Scripts can be re-run (ideally without much ado), unlike user actions performed with mouse clicks through Graphical User Interfaces (GUI). However, sometimes we might simply not know how to conduct a specific action through script commands or even generally within our statistical software and resort to workflows we are familiar with. Lack of time, skills or motivation might prevent us from performing these steps in a (swiftly) reproducible fashion. In that case it is imperative to document our doing to our best ability, to ensure that re-users have a chance of imitating it.

When our scripts become longer and more complex, we need to consider splitting them up into several files. This improves clarity and forces us to write code that is abstracted from the specific context it is used in (like variable names, absolute paths, undocumented and/or unchecked assumptions about previous code) and consequently helps us and others to re-use functionality in other contexts.

Recap: Workflow improvement

- never alter the original dataset
- perform all changes via script (i.e. not through a GUI)
- avoid non-documented work steps (steps that are not included in and run by the script)
- the script must reproduce all relevant results
- document all materials
- reasonable structuring of scripts and functions

2.2.2 Versioning

We recommend the use of a versioning technique. This ensures we keep track of different versions of our files. Nothing is more frustrating than losing an important version of our files because we tried to improve them but failed and now going back to the previous state is impossible, for whatever reason. The easiest and fastest system is naming files with prefixes or suffixes that indicate the version, e.g. the current date or a version number – example: *MicroDataset_20170731_v1*. However, a fully fledge versioning system – like Git – has many advantages over simple file naming conventions. This technical report cannot possibly describe these advantages in detail. We recommend to have a look at the Software Carpentry's course "Version Control with Git" (Software Carpentry) and invest a few hours to learn about Git. If we are part of a cooperation project, we need to make sure everyone uses the same versioning and naming technique.

Recap: Versioning guidelines

- reasonable, consistent naming and versioning of all files that are subject to changes over time
- ideally, using a versioning software (e.g. Git) when coding
- versioning is immensely helpful in cooperative situations

2.2.3 Documentation

One of the most important aspects in making research data re-usable is the documentation of research materials and processes. We might understand our scripts at the very moment we are using them but – given the complexity of our code - we will probably not understand them so easily six

weeks (months, years) later, if we did not include comments on the code and take notes on our research steps. Neither will any other person. We should, therefore, make sure that every research step is documented.

We recommend to put the most important information (which we refer to as 'metadata') – i.e. which changes to apply to the script before it is executable, where to get the data from, etc. - at the beginning of the main script. There, the chances are best that a re-user will notice them immediately. Additionally, reasonable and consistent naming of variables helps us and others to understand what they refer to.

The dataset(s) should be referenced with a persistent identifier – e.g. a Digital Object Identifier (DOI). Persistent identifiers help to identify different versions of the data. This way, re-users have a chance of finding the correct version(s) of the dataset(s), as discussed in section 4.1.3.

A good example for data documentation is the dataset of the study *Managing resource-related conflict* by Roy Vita (2016). You find all information at *datorium*, the repository, where the dataset and information about the study are deposited:

- information (metadata) on study level (see section 4.2) and
- information (metadata) on variable level documented in the codebook (see section 4.3).[2]

Recap: Documentation guidelines

- essential metadata put in head of main script
- documentation of code
- original data referenced unambiguously, ideally with PID
- reasonable, consistent names and labels for variables and values
- documentation of all relevant changes to the script (path to data, user input, additional packages/libraries)

2.2.4 Archiving of Data and Scripts

The DFG, the EU Horizon 2020 Programme and other funding organizations as well as journals have stated that archiving research materials and making them available to the scientific community is part of good scientific practice. Funding and/or acceptance of our articles/research might depend on our willingness and capability to archive our research materials (datasets, scripts, questionnaires, metadata etc.).

Archiving of research materials in appropriate institutions comes with a multitude of benefits which include, but are not limited to, keeping data safe in the long run by applying consistency checks and professional backup workflows, increasing the visibility of publications by using standardized metadata and offering access points for other scientists.

[2] For the metadata field in datorium please refer to the documentation of Zenk-Möltgen and Linne (2014).

Recap: Archiving of research material

- good scientific practice requires the availability of data, scripts and according metadata, wherever possible
- ideally, research material is archived at an institution such as a data archive or repository. The advantages of using such archives or repositories include:
  - professional backup strategies and consistency checks
  - (possibly) quality checks of ingest material
  - increased visibility of research
  - standardized documentation
  - expertise in data privacy
  - long-term archiving

Submissions of research materials to archives (or other institutes/repositories) should include all relevant files to produce the output, i.e. the results, plus documentation. We must make sure our scripts execute with as few alterations (such as changing paths) as possible. If our data are well documented and finally stored and published by a data repository or archive, metadata standards are usually met. One repository (easy to use and free of charge) that stores different kinds of research materials is *datorium*.

Recap: Specific requirements for archiving code

- submission of the code that produces the (final) output
  (rather than submission of output itself)
- script should be executable on other computers with minimal changes (e.g. paths);
  operability of a script depends, inter alia, on software in specific versions, paths to data, IDE
- replication server (i.e. a cooperation between *datorium and social science journals*)
  - repository for social science data and scripts
  - user-friendly, fast and secure archiving of social science data and scripts (free of charge)
  - data depositors decide themselves on access restrictions and can use licenses to specify how submissions should be re-used
  - ingest control of submissions
  - authors submit their data and code in order to make published results reproducible

2.2.5 Avoid Redundancy and Static Code

Avoiding redundant code helps keeping scripts short and easy to understand. It decreases frustration as well as error-proneness when changes need to be applied. Keep your code DRY (Don't Repeat Yourself). Redundancy is decreased by re-factoring often used code as functions and splitting scripts into separate files which are abstracted from their current context as much as is reasonable to allow using them in other contexts too.

Recap: Redundancies in code

- redundancy equals effort, error proneness and frustration when changes become necessary
- (re-)write redundant code as functions
- write (modular) code that continues to work even if changes are made to other parts of the script, e.g. use variables instead of concrete digits

2.3 Wrap-Up

For most quantitative empirical studies, research findings should be reproducible. Reproducibility means that researchers are able to independently work with the original research materials in a way that allows them to verify or falsify scientific results. To enable the reproducibility of research results and methodology and to simplify the workflow for researchers, research teams, re-users and reviewers, we need to share data. Data sharing means that research material is made available to others. Open Access to data and data sharing are part of the principles for good scientific practices. To ensure reproducibility and data sharing, research materials must be sufficiently documented, securely stored, and made accessible. Data sharing and reproducibility, among others, improve the quality of research results, foster collaboration and improve transparency of scientific processes.

2.4 Further Reading

American Journal of Political Science. 2016. *Guidelines for Preparing Replication Files*. https://ajpsblogging.files.wordpress.com/2016/05/ajps-replic-guidelines-ver-2-1.pdf. [September 18, 2017].

ANDS Guide. 2018. *Data Sharing Considerations for Human Research Ethics Committees*. https://www.ands.org.au/guides/Data-sharing-considerations-for-HRECs. [August 10, 2018].

Christensen, Garret and Courtney Soderberg. 2015. *Manual of Best Practices in Transparent Social Science Research*. August 11, 2015. https://github.com/garretchristensen/BestPracticesManual/blob/master/Manual.pdf. [September 18, 2017].

Deutsche Forschungsgemeinschaft (DFG). 2017. *Replicability of Research Results. A Statement by the German Research Foundation*. http://www.dfg.de/en/research_funding/announcements_proposals/2017/info_wissenschaft_17_18. [September 18, 2017].

European Commission. 2016. *Guidelines on Data Management in Horizon 2020*. Version 2.1. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf. [September 18, 2017].

Gandrud, Christopher. 2015. *Reproducible Research with R and RStudio*. 2nd eds. Chapman & Hall/CRC (The R Series).

Gentzkow, Matthew and Jesse M. Shapiro. 2014. *Code and Data for the Social Sciences: A Practitioner's Guide*. March 11, 2014. http://web.stanford.edu/~gentzkow/research/CodeAndData.pdf. [September 18, 2017].

Replicability Research Group. 2017. *Replicability vs. Reproducibility*. http://www.replicability.tau.ac.il/index.php/replicability-in-science/replicability-vs-reproducibility.html. [September 18, 2017].

Vita, Roy. 2016. *Managing Resource-Related Conflict. A Framework of Lootable Resource Management and Postconflict Stabilization*. doi:10.7802/1452.

Zenk-Möltgen, Wolfgang and Monika Linne. 2014. *Metadatenschema zu datorium – Data Sharing Repositorium*. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2014/TechnicalReport_2014-03.pdf. [January 19, 2018].

3 DATA PROCESSING

In the present context, reproducibility is primarily related to the data processing workflows. The aim is to document every step of data processing in such detail that other researchers or primary investigators themselves can at any point reconstruct how the raw data was manipulated and the dataset was structured, how new variables, for instance weights, were created, how different data was merged or which flaws in the data were corrected by using which technique. To achieve this goal, scripts must be employed that reproduce all relevant results. Section 3.1 takes the next step and shows how variables should be sorted, named and labelled and how to deal with different kinds of "missing values". In Section 3.2, we focus on data consistency and explain how to develop customized checking routines. Depending on the dataset, these might include missing values, wild codes and unlabelled values, filters, plausibility checks and checks of the weighting variables. Apart from explaining where these inconsistencies might come from and how to detect them, we also discuss different strategies of how to deal with them. The last section of this chapter, Section 3.3, is dedicated to data anonymization.

3.1 Variable Definition and Composition of the Dataset

Uwe Jensen

To start data processing, we first need to define all variables and to compose a structured raw dataset using a software program like SAS, SPSS, Stata, or R. By "defining variables", we refer to the formal or technical description of the information that allows for understanding the meaning of the (numeric or alpha-numeric) values of the research data surveyed with the questionnaire and re-presented by the variables. As explained in section 4, we thus create metadata on variables as part of the data management to document, process, analyse, find, understand and re-use these datasets.

In planning data collection and data processing, it is sensible to define all variables and their elements (like labels, filters etc.) as early and thorough as possible on the basis of the final (master or field) questionnaire and to document them in a code plan[3]. A structured dataset in the social sciences is most often composed by a rectangular data matrix where each row represents a case (e.g. a respondent) and each column represents a variable that contains information on the cases (e.g. the age of respondents). These variables might directly correspond to questions in the questionnaire; sometimes, however, one question might also be represented by several variables in the dataset, by e.g. grid questions, item batteries, multiple response questions etc.

The following sections describe essential steps, rules and conventions first, on how to define 'speaking' variables and second, on how to compose a well-structured dataset.

3.1.1 Variable Types and Variable Definitions

When defining variables, in a first step, we can distinguish two types: *administrative variables* are necessary to manage the data and to identify the context of the observation; *content variables* or *question based variables* are derived from the measurement instrument. The variables presented in the following represent a (non-exhaustive) spectrum of variable types, which can be used depending on the complexity of the study design and the underlying questionnaire.

Variable Type 1: Administrative Variables

As noted above, *administrative variables* help managing and identifying the dataset as a whole as well as information within the dataset. Thus, this group largely consists of different identifiers (ID variables) that are added to the dataset in the processing stage and, amongst others, document the context in which the observations have been made. When dealing with survey data of different types, e.g. cross-section, time-series, studies over space and time the need to include specific ID variables grows with its complexity. As a consequence, a dataset might have more than one identifier, referring to, for instance, to different contexts each single observation is related to (e.g. respondent in a region that is part of a nation state). Comparative surveys, for instance, require at least an ID variable for each respondent, the time of measurement, the geographical unit of observation and maybe even one for the language version of country specific questionnaire(s). The following paragraphs provide more detailed information on different types of such ID variables.

[3] A code plan transforms all questions into variables and unambiguously defines all necessary variables and their order in a dataset to prepare the processing of the data. It defines for each variable, the variable name and label and the coded value and value label for each characteristic of the measured item. If applicable the corresponding question no. should be mentioned in a first column for transparency. An extra column may describe specifics, like the unit of measurement, question types with specific answer conditions, filter information, coding procedure of open questions etc.

Respondent ID Variable

This variable is the core administrative variable for case identification in a dataset. A dataset requires at least one unambiguous identifier for each single observation. The data matrix should thus initially contain a placeholder for coding the identification number for each case or interviewee. The value of each of such an ID variable must be unambiguous in order to carry out partial corrections, additions, re-interviews (e.g. in a panel) etc. Thereby, multiple identifiers can be used, e.g. to identify observations, time point of measurement or geographical units. This information could also be combined into a single ID variable.[4] For example the integrated dataset of the European Values Study (EVS 2008, Variable Report) contains a *Variable id_cocas* structured as a 12 digit code: YYYY (year of wave), CCCC (country ISO 3166-1), NNNN (case number).

Interviewer ID Variable

Interviewer ID and interview log variables document information like interviewer number, interview date, start and end of the interview, the language in which the interview was conducted and so on. Such process data are created with the data collection. Couper (1998) called this type of information *paradata*. It is employed to control quality when conducting surveys (Gregory et al. 2009).

Fieldwork ID Variable

This variable includes year/month of the start and the end of the overall fieldwork.

Dataset ID Variable

Dataset variables are used to administer more complex surveys designs to cover e.g. data over time (each time point of measurement has an identifier), over space (different geographical units, e.g. each country included has an identifier), or different units of observation (like group of persons).

Questionnaire ID Variable

Such a variable is required for example to distinguish questionnaire splits, data from pre-tests or to differentiate between language versions of the field questionnaire(s) e.g. per country.

Dataset Version ID Variable

This kind of variable is a mean to uniquely identify and maintain the several versions of the dataset created during data processing and later data analyses. The specific purpose is to follow up on data modifications, corrections etc. by a unique identifier of the dataset processed. The version of the dataset should be assigned as part of the file name and as part of the data structure, i.e. as a variable, or better to say a constant, in the data matrix. We should decide on which versioning standard should be employed and use this standard throughout the whole project for all digital objects, such as data, documentations, reports etc., like the versioning standard of DDI[5], illustrated below.

---

[4] Creating ID variable(s), we should also think about anonymization, because too many information in the identifier might yield re-identification possible (see section 3.3.1).

[5] The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences.

The DDI-Standard to version a dataset consist of three parts

major.minor.revision

"1.0.0"

1. Position (default 1): major change in the dataset, e.g. adding a new sample / new variable;
2. Position (default 0): minor change with relevance to the data, e.g. changing a variable;
3. Position (default 0): change due to smaller revision, e.g. correcting typos in labels.

(DDI Alliance)

Weighting Variables

The definition of weights should be considered e.g. when results should be based on population figures instead of the sample size, such as demographic weights. A second example regards so-called sample weights to correct for unequal selection probabilities of the sample design.

Recap: General rules to define and manage administrative variables

- to place administrative variables at the beginning of a dataset, i.e. as initial variable(s), for quick access
- to define administrative variables as string (alpha-numeric) variables to avoid accidental re-coding
- compose ID variables in a similar fashion to warrant consistency and to ensure the same length, i.e. number of characteristics in variable values

Variable Type 2: Content and Question Based Variables

The major purpose of this second basic variable type is the definition of variables that represents the questions from the questionnaire(s). According to their complexity, questions can be distinguished by two types: the single response and the multiple response questions. Content variables (in contrary to the mentioned administrative variables) also comprise all variables which are constructed or derived from additionally fielded data.

Variables from Single Response Questions

This is the most straight-forward relationship between a closed question in the questionnaire and the corresponding variable in the dataset. A typical example is a question on gender which is transformed into a single variable with values 1 (e.g. for female) and 0 (e.g. for male) as well as categories for further responses and missing value(s).

Variables Related to Multiple Response Questions

Such cases refer to questions types which are item based or comprise multiple answer instances, where sub-questions, statements, or lists of questions can be created as dummy variables with dichotomous values of, e.g. 1 = applies and = 0 = does not apply for each individual question or answer category, respectively.

Example on closed questions with multiple answer instances (Q5)

"Please look carefully at the following list of voluntary organisations and activities and say ...

- *a) which, if any, do you belong to?*
- *b) which, if any, are you currently doing unpaid voluntary work for?*

...

Social welfare services for elderly, handicapped or deprived people"

(EVS 2008: Q5)

related Variable definition:

- variable name: "v10"
- variable label: "do you belong to: welfare organisation (Q5aA)"
  - value: 0 value label: *not mentioned*
  - value: 1 value label: *mentioned*
  - value: -1 value label: *don't know*
  - value: -2 value label: *no answer*

Variable Type 3: *Demographic Variables*

Quite often, a questionnaire contains open questions with numerical data (e.g. income in euros) which are captured as numerical values. Here, it is essential to document the unit of measurement in the variable label as well. Moreover, if such variables are categorized or aggregated, i.e. grouping ranges of values, we should do so by constructing additional variables. The underlying original variable (or other categorizations in the raw data) should be retained in the interest of the long-term usability of the data, as long as they comply with privacy regulations.

A typical example is a question on household income, which might apply a standard coding scheme, e.g. in terms of aggregating continuous information into quintiles. Where standardized definitions and coding schemes are available, we should apply them to make the data comparable with data from other surveys of interest. Prominent examples are coding schemes for occupation, such as the International Standard Classification of Occupations (ISCO), or for education, like the International Standard Classification of Education (ISCED).

Variable *Type* 4: *Derived or Constructed Variables*

Variables that are constructed, like harmonized variables of question based variables, should be inserted into the dataset right after the respective source or original variable (compare for instance the coding of the educational level at the EVS 2008. Variable Report: 35). The same applies to derived variables, like household typologies or income recoded into groups, which may be inserted in the sequence of the source variables. A core requirement in all these case is to describe the construction rules creating the new variable. The formal coding rules of each derived variable should be documented in the script and codebook (see section 4.3), while the methodological concept of a derived variable should be described in the methods report (see section 4.2).

3.1.2 Conventions for Defining Variable Names and Variable Labels

When defining variable names or groups of variables in the dataset, we should consider that they are the main object of a data generating project. To facilitate working with the data from data processing up to data analyses, variable names should be as simple, comprehensible, unambiguous and recognizable as possible. Additionally the Variable labelling should clearly indicate the question content and show the structural relationship between the question in the questionnaire and the corresponding variable in the dataset.

Options to Define Variable Names

Clear naming conventions simplify the analysis and more general the re-use of data. The four strategies described below can be used to define the variable names. However it is strongly recommend that variable names should always be used together with explanatory variable labels to clearly indicate the question content and show the structural relationship between the question in the questionnaire and the corresponding variable in the dataset.

1: Variable Names with Ascending Numbering

A common naming convention is the ascending numbering with the leading letter V (variable) e.g. *V01* to *Vnn* for data records with up to 99 variables and as well to apply correspondingly in case of a higher number of variables. This rule creates a simple order of the variables in the data record due to consecutively numbered variable names. However, the variables can't be distinguished according to content or type of the underlying question (question module, demography module, or derived variables).

2: Variable Names with Question or Item Number

The second strategy adds the question number to variable name, e.g. *V1q1*, *V2q2a*, *V3q2b* etc. This approach creates a direct reference of the variable to the original question and its order in the questionnaire. Questions that require the creation of multiple variables can be named according to the sequence of queried categories or items (*Q2.1*, *Q2.2* ... or *Q3a*, *Q3b* ...), even if this extended label, e.g. in the case of question batteries, is not specified in the questionnaire.

3: Mnemotechnical Variable Names

These artificial names, which are supposed to represent the essential content of the variables, can be a simple memorization aid. Such a procedure can be used in longitudinal analyses, if question modules are used repeatedly and the variables - used at different positions of the different questionnaires - should retain the same names, e.g. *R_Inc* for the respondent's income.

It should be noted that mnemonic definitions of variable names are not always understandable for third parties (within and outside the research project) and difficult to handle with large variable numbers. In this case, it might be helpful to provide a more meaningful value label, for example by including the question position in the questionnaire (question number) in the variable label, e.g. *Respondents income Q45*.

4: Variable Names with Different Characteristics

When naming variables of complex datasets, e.g. comparative surveys conducted in several countries like the EVS, combinations of prefix and suffix are often used to identify thematic or structurally different variable blocks, like demographic variables. As an example, one might think about a country-specific variable for party affiliation in Austria, named as *AT_PRTY* and labelled *Country specific party*

affiliation: Austria with the same variable for France would be "FR_PRTY" labelled "*Country specific party affiliation: France*. Further country specific variables should also start with the country prefix, e.g. *AT_*, "FR" and the suffix for respective question, for example on religious affiliation (*AT_RELIG, FR_RELIG*) or region (*AT_REG, FR_REG*).

Recap: Recommendations editing variable names and variable labels

The purpose of variable labels is to supplement the variable names with a description of the content of each variable that is as short, distinguishable, and as meaningful as possible. Additional information on the context or the distinction of the variables that could be included in the variable label are

- the question number from the questionnaire,
- the number of response categories, or
- notes on the type or specifics of the variable: e.g. whether it is newly formed or recoded, contains special filters, the variable is country or wave-specific or has deviations from the standard code scheme.

On a more formal level users are recommended to apply the following rules:

- variables names should not be longer than eight characters for practical reasons, although the maximal number of characters for variable names or any kind of labels depend on the applied software;
- For programming reasons, variable names should not contain blanks; if required, "_" should be applied instead;
- variable names should exclude special and linguistic characteristics like "ä" and employ "ae" instead.

3.1.3 Coding and Labelling of Valid Values

Variables that require solid background knowledge like coding of occupations, should apply quality assurance measures, e.g. coding by an experienced project member, double-checks etc.. From the perspective of long-term use of the data, it is also advisable to use nationally or internationally accepted classification systems such as ISO codes for languages (ISO 639), countries (ISO 3166), or ISCED for educational levels. The following examples are applied in the EVS (2008).

EVS Example: General rules to define and manage administrative variables

Variable	Variable label	Composition of values
v305b	respondents nationality	ISO 3166-1 (3 digits)
v307b	respondents country of birth	ISO 3166-1 (3 digits)
v336	educational level respondent: ISCED-code one digit (Q110)	
v336_2	educational level respondent: ISCED-code two digits (Q110)	

Conventions on Coding Numerical Variables

In the following we list basic conventions for the numerical coding of response categories:

- the (numeric or alphanumeric) codes must fully represent all answer categories of the question, and these codes must be mutually exclusive and well defined;
- similar 'events' should be coded employing a standardized coding schema;
- the assignment of values to response categories is executed in ascending order following the sequence in which the response options appear in the questionnaire.

If the code 0 has the meaning of none (e.g. number of children), it remains as it is.

For questions with multiple answers, the individual answer categories might be encoded as a dichotomous variable with values of 1, referring to mentioned, and 0, referring to not mentioned. The information no answer can be stored for each variable in an additional code or mapped in a new variable (none of the answer categories). This newly created category/variable must be distinguishable from comparable categories already specified in the questionnaire.

Notes on Defining Value Labels

The explanation of response categories by value labels should be considered as an exact as possible transfer of the text of the answer categories of the question. Such content rich value labelling facilitates the work with the data in the project as well as secondary analyses by other researchers. At the same time, such meaningful values are much easier to read during data analyses than to look up these meaning in the code plan or questionnaire again and again.

However, the value labels should not be too long (20 to 40 characters). If it is necessary to shorten long answer texts from the questionnaire, make sure that the meaningfulness and the selectivity of the answers are not lost.

Declaration, Coding and Labelling of Missing Values

When defining missing values (non-responses), we can distinguish between item- and unit-non-response, as discussed in greater detail in section 3.2:

- Unit-non-response are complete failures of survey units (cases), e.g. a target person from the sample could not be surveyed because participation was denied or the person was deceased, unreachable etc. This information is included in the method or field report as data on the response rate.
- Item-non-response or filter-non-responses, on the other hand, describes missing values due to misunderstandings in the interview, missing or denied answers, etc. In particular structural response deficits must also be taken into account in this context. Typical cases occur when questions were not presented to individual respondents (due to filter conditions) or, in the case of complex studies, in individual (sub-)surveys, countries, regions, or waves.

Missing values are assigned special codes in the respective variables. Their meaning needs to be documented by appropriate missing value labels. They can, but do not have to be, specified in the questionnaire. They could also be based on written interviewer notes which are then used for the data input or at data validation.

All missing values must be fully declared in the course of the data definition in order to facilitate structured data control procedures during the various steps of data processing and data analyses. We therefore need a coding concept for missing values that states, what each missing code stands for within the dataset. This concept further ensures that missing values are the same over the course of

the dataset. In the EVS (2008) a value of e.g. -5 represents other missing, such as system missing, wild codes, illogic answer-patterns in e.g. filters.

Through these procedures, empty cells, i.e. system missing can be excluded. The avoidance of system missing affects in particular integrated datasets, trend series or other complex data. In such cases it is recommended to explicitly document the reasons for the lack of incompleteness of the responses, e.g. omitting questions in individual countries or a single time point of measurement, in order to obtain a fully documented dataset.

General Coding Rules of Missing Values

For handling the missing values within the record, it is recommended to use

- either the highest numeric codes that are outside the respective valid range of values, as illustrated by Examples A and B, below,
- or to encode missing values with negative values as shown in Example C.

The advantage of coding missing values with negative values is the clear perceptible distinction from the positive valid values of respectively constructed scales. This simplifies work with control documents or adjustments of scripts and supports the presentations of results. Depending on the range of valid values, the missing categories should be coded uniformly for all variables of a dataset.

Example: Coding missing values

Missing *Example A*	*Values*	*Missing* *Example B*	*Values*	*Missing* *Example C*	*Values*
If value "7" is part of the valid value range of the variable, "97" is encoded.		If the code "0" has a meaning or if dichotomous variables are encoded with "0" and "1" "Does not apply" gets the code 9		Coding with negative values	
7 (resp. 97, 997) refused 8 (resp. 98, 998) don't know 9 (or 99, 999) no answer 0 does not apply		6 (resp. 96, 996) refused 7 (resp. 97, 997) don't know 8 (or 98, 998) no answer 9 (or 99, 999) does not apply		-1 do not know -2 no answer -3 not applicable -4 not asked in survey	

In addition a *not applicable* category can be used in cases where no answer category or (due to error) several categories have been marked, where values are outside of the defined value range of the variable, or where information is presumably incorrect and not reproducible.

If different categories, such as *no answer* and *refused*, are grouped together, we should document this in the missing value label. However, such categories can be meaningful for some research questions and might thus be kept separately for analysis. This can also apply to *do not know any more*, *cannot remember* or *denied*. If these categories are not to be defined as missing, they will receive a code immediately following the categories already defined. For example, a category *do not know* represents a content statement in case of knowledge questions and therefore retains its code as a valid value for this variable.

Special Coding Rules of Missing Values due to Non-Submitted Questions

Sometimes, we cause non-responses through structural conditions of the survey or as a result of filter questions. Structurally caused losses arise when questions are not asked in a wave and / or in a country. These cases are encoded according to the standard requirements of the survey, e.g. by the category not asked in survey and the code -4 (see Example C, above).

Filter follow-up questions that contain the category does not apply (not applicable) regard those respondents to whom the question was not submitted because it does not apply to this group (filter non-response). In such cases the code for not applicable might be used (see Examples A and C, above).

Multiple filtering conditions from questions sequences are to be considered at the end of the sequence (filter-sequence relationship). In general, the filtering conditions start from the valid value range of the initial filter question. The values defined as missing in this question, are also set to missing in the subsequent questions.

Moreover, when defining filter-follow-up questions, the category does not apply should also clearly state which previous encoding(s) of the question and category(s) it refers to, as illustrated by the example below. In the case of multiple filters connected in series, all filter-sequence relationships should clearly be defined and described in related missing value labels for all variables that are part of the sequence.

Example: Coding a filter-sequence relationship

Q.17 Have you ever been unemployed?		*F18 (if respondent was unemployed): How long were you unemployed?*	
1 Yes	n = 210 (> F18)	1 Under one year	n = 150
2 No	n = 1060 (> F19)	2 One year and longer	n = 50
9 No Answer	n = 10 (> F19)	9 No answer	n = 10
		0 Does not apply (F.17 code 2 or 9)	n = 1070

3.1.4 Dataset Structure and Grouping of Variables

The basic purpose of thinking about the dataset structure is to set a clear dataset sequence to easily identify all variables, to keep relationship to the questionnaire, as well as to facilitate script processing.

In order to compose a well-structured dataset, it is best-practice to group variables according to formal criteria as already demonstrated in the previous sections. The variable structure derived from the questionnaire is the result of the sequential order of the questions, respectively the question types (simple question, item battery), and the additionally constructed (technical, harmonized, etc.) variables, as illustrated in the Info-Box, below.

Recap: Variable definition and composition of dataset-structure

- identification and administrative variables
- weighting variables
- questions related variables (content variables)
- demography variables (content variables)
- survey and interviewer variables (paradata variables)
- integrate further administrative or content related variables (derived; constructed at later stage) after the referenced variable or close to the source variable

3.1.5 Further Reading

Akremi, Leila, Nina Baur and Sabine Fromm (eds.). 2011. Datenanalyse mit SPSS für Fortgeschrittene 1. Datenaufbereitung und uni- und bivariate Statistik. 3. Aufl., Wiesbaden.

Jensen, Uwe. 2012. Leitlinien zum Management von Forschungsdaten: Sozialwissenschaftliche Umfragedaten. GESIS-Technical Reports 2012, 07. GESIS - Leibniz Institut für Sozialwissenschaften. https://www.ssoar.info/ssoar/handle/document/32065. [August 14, 2017]

Couper, Mick P. 1998. Measuring Survey Quality in a CASIC Environment. https://ww2.amstat.org/sections/srms/Proceedings/papers/1998_006.pdf. [October 04, 2017].

Long, Scott J. 2009. The Workflow of Data Analysis Using Stata. Stata Press.

3.2 Data Consistency

Hannah Schwarz

When referring to data consistency, we mean the plausibility of (combinations of) values of survey items within one interview or within groups of interviews. Implausible (combinations of) survey answers can be indicators of problems in the survey process leading to lower data quality. Hence, data consistency checks can be helpful to assess and improve the quality of survey data. There is a great variety of possibilities for checking data consistency. The following should be read as suggestions that can help researchers to create their own checking routines, customized for their specific datasets. Generally, when finding inconsistencies in survey data, the aim is to trace back how they arose in order to make an informed decision about whether corrections are useful. Hence, having a comprehensive documentation of the data collection and data processing phases at hand is essential, as discussed in section 4.2. In the following, we present considerations of how inconsistencies might arise, approaches to detecting such inconsistencies and approaches for dealing with them.

3.2.1 Data Validation

Some steps of initial data validation should be conducted before checking the dataset for inconsistencies. Validation here implies an initial assessment of the raw data as well as an initial representativeness check as outlined below.

Firstly, one should check whether unit nonrespondents are present in the dataset. Such cases with missing values for all variables should be deleted. Furthermore, partial interviews might be identified in the dataset, resulting from break-offs, i.e., situations in which respondents refused to continue the interview at some point during the interviewing process. Such cases can be deleted but can also be maintained. The decision about this has to be made individually and should depend on considerations about the partial cases' eventual usefulness for analysis. In any case, deletions of cases have to be documented carefully.

A second step of data validation should be checking for duplicate cases. The common statistical software programs provide specific commands for this. One can check duplicates of the ID variable but also duplicates in terms of whole data vectors (rows) pertaining to one observation. We should, furthermore, examine the completeness of the data by comparing the variables in the deposited raw data with the original survey questionnaire(s).

An initial validation of the data's representativeness gives an indication about this important quality criterion. For this, the relevant study-level documentation should be consulted to understand what the target population is, what the sampling methods employed to draw the sample are, and to what extent the sample is aimed to be representative of the target population. For a sample survey to be representative of the target population, sampling has to be random. This means all elements in the target population must have a non-zero and known chance of selection at all stages of the sampling process. Where random sampling procedures are used and a representative sample is aimed for, the below steps are relevant.

A measure which is commonly used as an indicator of representativeness is a surveys' response rate. However, a lower response rate does not necessarily mean that data are less representative. Instead, this depends on the extent to which respondents are missing at random. Yet, high response rates are commonly seen as a reassuring indicator of representativeness while low response rates indicate that one might have to go deeper into considerations about the randomness of the missing observations, e.g. by checking for systematic drop-outs (Groves et al. 2009).

Therefore, as a third step, response rates should be calculated. They can be calculated in slightly different ways, but most often the kind of information contained in the table below serves as a basis for calculation. For guidelines on how to calculate response rates, which differ based on the kind of sampling frame used, we recommend to consult AAPOR Guidelines (2016).

Relevant data for calculating response rates

total number of objects in sample

number of valid objects
number of invalid (non-sample) objects
number of objects of unknown validity

number of completed interviews
number of partial interviews
number of refusals and break-offs
number non-contact (never contacted)

A last step that we recommend undertaking in order to get an indication about the representativeness of the data is comparing variables' sample and population distributions. For variables where population values of the target population are available, e.g. from a recent census, such comparisons are advisable. Most likely, the comparison will be limited to socio-demographic variables and hence, in this case, will give an indication of the representativeness of the data in socio-demographic terms.

If notable deviations of the sampling distributions from the population distributions are identified, the construction of weights is a common solution. Weights can approximate the sample distributions to the desired target population distributions. Whenever weights are constructed, one should carefully document how they are constructed and how one recommends using them. If weights are delivered by the data collection agency, they should always be checked for system missing or zero values. Such values should usually not be included in a weighting variable. If such values are identified, we recommend consulting the documentation of the dataset, to see if there is any explanation for this. For example, in the European Values Study (EVS 2008), value zero on the demographic weight variables indicates that information on the relevant demographic variable was missing. If zero or values for system missing are found on weighting variables and no explanation for this is delivered by the constructor of the weights, we recommend documenting the problem and contacting the person who created the weights for further consultation.

Recap: Data validation

- delete unit-non-respondents from dataset
- decide which partial interviews to keep
- check for duplicate cases
- calculate response rate
- check representativeness on basis of available population data
- consider constructing weights

3.2.2 Overview of Common Inconsistencies and Data Consistency Checks

After an initial validation of the data, we recommend taking some time to check for consistency. This can be done in a variety of ways. We aim to help research projects to develop their individual, customized checking routines by providing some examples of common inconsistencies. Furthermore, considerations about how inconsistencies likely arise will be provided. These considerations take a central role in deciding how to deal with detected inconsistencies and are therefore included in the discussion.

To detect inconsistencies, one can employ what we refer to as 'visual checks' in the following, which are checks working via sorting the data, 'eyeballing' through the dataset and inspecting cross-tabulations. Such checks can sometimes be sufficient. However, especially where datasets contain large numbers of observations, these visual checks become inefficient. In such cases, scripts containing code should be employed by means of which the dataset can be checked for inconsistencies more systematically and efficiently. It might also be useful in certain cases to not only obtain the information how many inconsistencies exist but also which observations are concerned. Lists of IDs of the cases concerned can be easily retrieved via a command in a script. In the following, we discuss different types of inconsistency checks, namely checks detecting missing values in certain groups, checks alerting to wild codes and unlabelled values, checks picking up inconsistent applications of filters as well as checks examining the theoretical compatibility of different answers.

Missing Values for Certain Groups

One common inconsistency is having only or largely missing values on a variable for specific groups. One reason for this inconsistency to arise is a processing error. Missing values in certain groups are particularly likely to arise due to variables being deleted or entirely overwritten when processing data for different groups separately, for example in international surveys. In such contexts, data collection is likely to have taken place separately and, hence, certain variables might not have been collected and deposited in the first place for certain groups, as illustrated by the following example.

EVS 2008 Example: Missing values for certain groups

European Values Study 2008
GESIS Study No. ZA4800 (v.4.0.0), doi:10.4232/1.12458 and ZA4799 (v.1.0.0), doi:10.4232/1.12483

Variable, Label
Question Text (English Language)

intno - interviewer number
Interviewer number

Finland, Norway:
Question not implemented in Field Questionnaire. Computed and coded as -4 'question not asked'.

(EVS 2008)

Another issue that (largely) missing values for certain groups could arise from are problems during data collection, especially systematic refusal to provide an answer among certain groups. An example

of such a situation would be a survey question that is sensitive in certain countries but not in others (e.g. ethnicity) and therefore generates high refusal rates in one place but not in another.

To check whether values are systematically missing for certain groups, we recommend writing code that creates flag variables[6] indicating for each relevant group whether a variable contains (almost) only missing values or not. This script can then generate a list of groups with missing values for a single variable.

Wild Codes and Unlabelled Values

Another data consistency check should look for wild codes and unlabelled values. Such problems likely arise due to processing errors at some point during the data entry, the recoding or the labelling stage.[7] Checking for wild codes and unlabelled values can, again, be done via visual checks, particularly by looking at the frequency distributions of values for each variable in the dataset and comparing the present values with those indicated to be in the valid range in the codebook or questionnaire.

This is illustrated by the example below. Here, values 95, 96 and 99 were obviously out of range for a variable containing the number of children of a respondent (EVS 2008, v321). They were found to stand for "don't know", "no answer" and "other missing", respectively and were thus recoded to the respective valid codes specified for these answer options in the codebook. If valid values have been assigned with labels, one can also apply the respective labelling file before inspecting frequency distributions. In this way, out of range values or values which are unlabelled for different reasons easily catch the eye.

Running a check for this kind of inconsistency via a script (e.g. Stata's scandata) can best be done by writing a command that lists all out of range values per variable. For string variables, we recommend employing a command that checks for the correct length of values. As ranges do not exist in string variables, checking whether all variables have the correct length, i.e., the same number of characters, can be a useful technique to identify out of range values in string variables.

Inconsistent Application of Filters

Inconsistencies are also likely to arise where filter questions are employed. Including such skip patterns into a questionnaire results in the risk of filters being disregarded or applied incorrectly, at least in non-automatized data collection. Where interviewers or respondents themselves (in self-administered questionnaires) are responsible for the application of filters, this might lead to respondents answering follow-up questions that they should have skipped and thus should contain a missing code, such as not applicable (see section 3.1.3). One can inspect such inconsistencies concerning filter questions and follow-up questions by simple visual inspection of cross-tabulations of the respective variables. For a more efficient check of this problem, we recommend writing code that creates a flag variable and lists all observations that take on non-missing values for follow-up variables even though the value they have on the filter variable indicates that this follow-up question should not have been asked.

---

[6] In the survey context, flag variables are variables used to mark an observation and to provide additional information about a specificity occurring in this observation. In the context of inconsistencies they most often take on the values 0 = consistent and 1 = inconsistent.

[7] As a matter of fact, mere processing errors can be the cause for all kinds of inconsistencies. Processing errors will therefore not be mentioned in particular as a reason for the inconsistencies discussed in the following.

EVS 2008 Example: Wild codes - variable v321

European Values Study 2008
GESIS Study No. ZA4800 (v.4.0.0), doi:10.4232/1.12458 and ZA4799 (v.1.0.0), doi:10.4232/1.12483

Variable, Label
Question Text (English Language)

v321 - how many children do you have (Q105)
Q105
<ASK ALL>
How many children do you have?
Write in: ...
<IF CODE 0: GO TO Q107>

-5 other missing
-4 question not asked
-3 not applicable
-2 no answer
-1 don't know
0 no children

Codebook entry EVS 2008 for variable v321

	before recoding			**after recoding**	
	Frequency	Percent		Frequency	Percent
0	18845	28.4	0	18845	28.4
1	11554	17.4	1	11554	17.4
2	21476	32.4	2	21476	32.4
3	8828	13.3	3	8828	13.3
			output omitted		
95	7	0.0	-1 don't know	7	0.0
96	592	0.9	-2 no answer	592	0.9
99	9	0.0	-5 other missing	9	0.0
Total	66281	100.0	Total	66281	100.00

(EVS 2008)

Theoretical Consistency/Plausibility

Contradictory Answers

One step that is useful for detecting errors in one's data is checking for theoretical consistency or plausibility. Note that respondents' answers are not always theoretically consistent and might sometimes directly contradict each other. Such situations are most likely to arise due to respondents mis-

understanding the survey question, misreporting due to other reasons such as social desirability or simply an erroneous recall of information. However, where such inconsistencies arise in high numbers, this could also point to a more systematic problem in the survey process, for example, a mistake in the questionnaire or an inconsistency across questionnaires used in different waves of the same survey.

We recommend checking for contradictory answers by looking at cross-tabulations of variables to see whether implausible combinations of answers exist. The example below depicts a cross-tabulation of two variables of the EVS 2008, resulting from respondents being presented with a list of organizations and activities and being asked to say which, if any, they belong to. Variable A064 is a binary indicator depicting whether a respondent affirmed belonging to a welfare organization. Variable A080 is a binary indicator showing whether or not a respondent mentioned the answer option *none* when asked this question. Affirming to be member of a welfare organization and then affirming to belong to none is incompatible. Yet, eight respondents are coded to have indicated this answer combination. Accordingly, they were marked in the EVS 2008 as inconsistent answers by means of a flag variable (*a080_f*).

EVS 2008 Example: Wild codes - variable v321

		Do you belong to: none (Q5a)		
		not mentioned	mentioned	Total
Do you belong to: welfare organisation (Q5aA)	not mentioned	21,998	35,297	57,295
	mentioned	3,133	8	3,141

(EVS 2008)

An alternative to a simple cross-tabulation consists in writing a script that lists observations taking on values for a certain variable which are theoretically not consistent with a response indicated on another variable.

Systematic Answer Behaviour

Another type of implausible answer combination is summarized under the term systematic answer behaviour. A prominent example for systematic answers is straight lining, i.e., a situation where respondents give the same value to all items on a question battery. This pattern is likely to arise due to satisficing, which refers to a respondent's strategy to lower the burden of survey taking to such an extent that the given answers are meaningless.

One can do a visual check for this by inspecting values across battery items per observation in the data browser of the statistical software. For going through a large number of observations, we recommend writing code that lists observations for which the answers are systematic, for example, all having equal values. This is easiest done by first creating a flag variable indicating for each observation whether or not the respondent answered systematically. The flag variable can then be used to check whether the prevalence of systematic answers is especially high for certain groups.

Implausible Correlations

A final type of theoretical inconsistency refers to implausible correlations. One prominent reason why implausible correlations appear is that answer scales were reversed at some stage of the survey pro-

cess, e.g., in the questionnaire, during data entry or data processing. However, they can also arise for substantive reasons, such as relationships actually being reversed in certain contexts or groups. For example, different correlations might be plausible in different country contexts. Therefore, if implausible correlations are found, we recommend ensuring that for the specific context the survey data stems from, this correlation is indeed unexpected.

A visual inspection of a cross-tabulation can give a first impression about whether a correlation in the expected direction is present in the data. However, to check for implausible correlations, correlation coefficients should always be calculated. This check should be conducted with correlations for which a clear expectation about the directionality exists. If the inverse of the expected relationship is found, the coding of the scales at all stages of the survey process should be checked carefully.

Recap: Common Inconsistencies and Recommended Data Consistency Checks

- missing values for certain groups
- wild codes and unlabelled values
- consistent application of filters
- theoretical consistency and plausibility of answer combinations, considering particularly: contradictory answers, systematic answer behaviour and implausible correlations

How to Deal with Detected Inconsistencies

When identifying inconsistencies, one's primary aim should be tracing their sources. Researchers are advised to assemble as much information as possible about how and why inconsistencies arose, as this information can be crucial for making a judgment on how best to handle inconsistencies. The most important source for tracing back inconsistencies is the documentation of the survey, including documents such as the questionnaire, the method report, the fieldwork report, and, if available, processing scripts showing all processing steps applied to the data before one received them (see section 4.2).

Another measure that one can take in order to collect more information about potential sources of inconsistencies is contacting the data collection agency or repositories. One can ask them for additional documentation materials or present them with the detected inconsistencies and ask for advice or potential explanations. Furthermore, in certain situations, it might also make sense to consult with other subject matter experts or scholars with expertise about the specific context the survey was conducted in. For example, as mentioned before, context-specific knowledge can be essential for understanding which correlations are plausible.

In the process of making a decision about how to handle inconsistencies in the data, certain considerations should be undertaken. Firstly, one should consider whether one can be certain that an inconsistency actually represents an error. Some answers that seem contradictory or implausible might simply represent the way respondents choose to answer the survey questions, for various reasons. One's certainty about the presence of an error will often depend on whether one managed to trace back the source of an inconsistency and hence could confirm the presence of a mistake at some point in the survey process or not. Another important consideration concerns the number of cases affected by an inconsistency. If the proportion of affected cases in the dataset is large enough to substantially impact data analyses, the need to investigate, document and potentially correct inconsistencies is naturally higher than if only a small proportion of cases is affected. We might want to take more far-reaching measures in order to prevent distortions in re-users later analyses of the data if a high proportion of cases in the dataset is affected.

There are different ways to handle data inconsistencies, and we recommend using them under specific circumstances as outlined in the following. The measure representing the strongest intervention is correcting inconsistent values or setting them to missing. Corrections can only be made where the correct value could be inferred from the inspected documentation with high certainty. For example, if we inspect the questionnaire and realize that the scale in one of the variables was coded reversely to what we expected, it seems justified to go ahead and recode the values of this variable. If we have no way of being sure about the correct value(s), but are certain that an inconsistency results from an error, most likely because we were able to trace the source of this error, a suitable option might be setting the inconsistent value(s) to missing. Where we recode variables, we should consider leaving the original variable in the dataset and creating an additional variable containing the corrected values. In this way, each data user can decide which version she considers the best choice for her particular analysis.

From our experience, there will unfortunately be many cases where it is not fully possible to understand how and why an inconsistency arose. For such cases, we strongly recommend documenting all findings to the later users of the data but not changing the data at all. Leaving inconsistencies in the data and only documenting them has the advantage that the decision of whether or not to include such cases in data analyses is up to the (re-)users. In this way, they can make their own conclusions based on the findings presented to them and based on considerations about the particular analysis they want to run with the data. Whichever way of handling inconsistencies we decide for, a comprehensive documentation of what was done and why and how it was done should always be provided in the documentation, most likely in the codebook (see section 4.3).

An additional way of documenting inconsistencies is by using flag variables that indicate for each observation whether it was affected by a certain inconsistency or not. Two examples for this are displayed below. Here, the flag variable a043_f indicates whether more than the five maximally permitted items of a battery were chosen by respondents. The relevant battery contained a variety of qualities that children can be encouraged to learn at home and respondents were asked to choose up to five of those qualities which they considered particularly important. However, there were observations for which more than five qualities were coded which the EVS (2008) handled by creating a flag variable on which these respondents were coded as 1 'limitation ignored'. In a similar vein, flag variable e001_f marks respondents that seemingly gave inconsistent answers to two consecutive questions asking for the first and the second most important aim that their country should give top priority to over the next ten years. There were cases in which respondents indicated identical aims in these two questions and the EVS 2008 handled these by coding them 1 'inconsistent' on the relevant flag variable.

EVS 2008 Example: Two examples of flag variables

Variable	Label	N*	total N	Values	Value Labels	Description
a043_f	flag variable: learn children at home	157993 7004	164997	0 1	Limitation followed Limitation ignored	If respondent named more than the allowed 5 qualities. if a027 to a043=1 > 5 then a043_f=1
e001_f	flag variable: aims of this country - most/second important	39127 101	39228	0 1	consistent inconsistent	If most important aim is identical with next most important aim. if e001=e002 then e001_f=1

EVS 1981-2008 Longitudinal File 2008, Variable Report

Another step that one should always take is defining and documenting rules about how inconsistencies are detected and dealt with in a research project and why. As discussed in section 2, these rules should be defined up-front and should be followed consistently in order to ensure transparency of the processing for later re-users of the data. An example for such a documentation of rules can be found in Appendix D (Data Cleaning Rules) of the EVS 2008 Guidelines and Recommendations (EVS 2010).

Recap: Dealing with detected inconsistencies

- attempt to trace source with help of documentation - questionnaire, fieldwork report, methodological report, ideally also processing script
- possibly contact data collection agency and/or subject matter experts with your findings for consultation
- way of handling inconsistency depends on certainty about error, ability to infer correct value and number of cases affected
    - correct inconsistent values if certainty about error high and correct values could be inferred
    - set inconsistent values to missing if certainty about error is high but not enough information is available for correction
    - document inconsistent values in codebook and/or flag variable if certainty about error low
    - always: set up project rules for consistent handling and document them

3.2.3 Further Reading

Brislinger, Evelyn et al. 2011. *EVS 2008 - Project and Data Management. GESIS-Technical Reports 2011, 14*. GESIS - Leibniz Institute for the Social Sciences 2011. https://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2011/TechnicalReport_2011-14.pdf. [October 04, 2017].

European Values Study. 2010. *EVS 2008 - Guidelines and Recommendations. GESIS-Technical Reports 2010, 16*. GESIS - Leibniz Institute for the Social Sciences 2010. https://dbk.gesis.org/dbksearch/file.asp?file=ZA4800_standards.pdf. [October 04, 2017].

Frick, Joachim R. and Jan Goebel. 2011. *Biography and Life History Data in the German Socio Economic Panel (SOEP, v27, 1984-2011)*. DIW Berlin. http://www.diw.de/documents/publikationen/73/diw_01.c.391273.de/diw_datadoc_2011-061.pdf. [October 04, 2017].

Jensen, Uwe. 2012. *Leitlinien zum Management von Forschungsdaten: Sozialwissenschaftliche Umfragedaten*. GESIS-Technical Reports 2012, 07. GESIS - Leibniz Institut für Sozialwissenschaften. https://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-07.pdf. [October 04, 2017].

Lohmann, Henning and Sven Witzke. 2011. *Data Documentation 58 - BIOEDU (Beta Version): Biographical Data on Educational Participation and Transitions in the German Socio-Economic Panel Study (SOEP)*. DIW Berlin. http://www.diw.de/documents/publikationen/73/diw_01.c.379412.de/diw_datadoc_2011-058.pdf. [October 04, 2017].

3.3 Data Anonymization

Marcus Eisentraut

One core activity in processing research data is anonymization. Whenever researchers collect, process, store or (re-)use so called personal information, a strategy to protect individuals' identities is mandatory. Otherwise, researchers risk to lose survey participants' trust, might harm their individual right on self-determination and thus violate legal requirements of data protection regulations. This chapter aims to provide a first, non-exhausting insight into anonymization and how researchers can protect their participants from harm. To do so, we must first of all check for (country-specific) data protection regulations and legal requirements on data anonymization. Second, we can consider how to best anonymize personal information in the context of processing research data.

3.3.1 Basics of anonymization

There are numerous reasons to anonymize research data. From an ethical point of view, we are responsible to protect our participants from harm or commercial interests. From a legal perspective, data protection regulations, such as the EU's General Data Protection Regulation (679/2016/EC), require anonymization whenever we deal with personal information, i.e. data.

> Personal information covers all data "relating to an identified or identifiable natural person (…) who can be identified, directly or indirectly, in particular by reference to an identifier such as a name (…)" (Art. 4.1 679/2016/EC).
>
> Anonymizing such data first of all implies, that we take "into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons" (Art 25.1 679/2016/EC)

Thereby, anonymization means to process "personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information" (Art. 5.4 679/2016/EC). That is, identifiers, i.e. information that enables re-identification of a natural person, must be separated, deleted or manipulated in a way that it can no longer be used to (re-)identify an individual participant in the data, as soon as possible.

Finally, we should keep in mind that anonymization is a legal requirement and an integral part of good research when dealing with personal data. However, as we collect well-defined data for a specific purpose, we should not harm this purpose by anonymization. That is, if anonymization would prohibit the use of data in our analysis, we have to keep data weakly or even un-anonymized. Needless to say that in such a situation data must be kept securely and protected e.g. by access-control, passwords and encryption (Art. 5 2016/679/EC; Elliot et al. 2016; Corti et al. 2014).

Recap: Anonymization

- personal information enables re-identification of a particular individual
- dealing with personal data, we have to protect individuals from harm
- data anonymization means to process data that can no longer be related to a particular individual

3.3.2 Anonymization Strategies

In the context of data anonymization and identifiers, we can distinguish two types: direct and indirect identifiers. Direct identifiers cover information on a natural person that points directly to this individual, like a participant's name, her contact information as well as pictures and records. Such information is in most cases not relevant for data analysis, but maybe so for later purposes. However, information that is not needed for data analysis or any other (future) purpose should be separated from or at least be replaced, i.e. pseudonymization, as soon as possible (Corti et al. 2014; Doorn 2010).

Separating information from our data does not automatically mean to delete it. Instead, we should ask ourselves carefully, if this information is really not needed anymore before destructing it forever. For example, in panel studies, contact information must be kept to ensure that participants can be re-contacted in the following waves. However, once we are sure that the information is not needed any longer, we should destruct it. First of all, it excuses us from caring about legal regulations in the context of this information. Second, getting rid of information not needed anymore does not simple mean to delete a digital file from our hard drive. Indeed we must destruct it, ensuring that it cannot be replicated anymore, on our personal hard drive as well as on all other hard drives the information as well as its back-ups had been stored.

Indirect identifiers are somewhat more complex. It is the combination of different pieces of information in the data that enable re-identification of a particular individual (Corti et al. 2014). We might for example think about data on the place of living, e.g. a rural, sparely populated area, on an individual's familiar background, such as being married and having four children, as well as on the individual's occupation, which might be relatively rare, like being mayor of a town. If we know this person and we know that she took part in the survey we got in hand, it would be a waltz to re-identify this person in the data and thus find out more about her, her attitudes and behaviour.

Indirect identifiers are sometimes hard to discover in the data. In a first step, we thus must carefully review our questionnaire of origin and consider where information might be so fine-grained that re-identification is possible. Second, we should look at the marginal distributions of core variables as well as of the cross-tabulation of such variables (for particular sub-groups in the data). If these marginal distributions become quite small, we should be aware of a high risk of re-identifiability.

Strategies to anonymize data according to indirect identifiers are manifold. Therefore, they cannot be entirely discussed here and instead some examples will be given in the following. First, we once again should carefully recapture if sensitive information included in the data is needed for data analysis. If this is not the case, we might separate such information, e.g. a complete variable, from the research data. Second, we might suppress particular values, e.g. replacing it by a pre-defined type of a missing value, restrict lower and upper ranges of (continuous) variables or aggregate information.

That is, for example, information on an individual's income can be critical, due to the fact that very high income earners are quite rare and thus have a higher risk of re-identification. We might consequently define a threshold, based on the distribution of our continuous measurement on income, and recode all values above this threshold into a single category, such as "€ 10.000 per month and more". Likewise, we might aggregate all values of our income variable into a pre-define number of categories, such as quintiles. The basic idea behind such restrictions and aggregation is known as k-anonymization. Manipulating our data, we aim to reach a fixed minimal number of cases that must be represented in each possible sub-group and for each possible combination of crucial variables in our research data (Elliot et al. 2016).

Such strategies are examples that might not always be an appropriate solution for our data and our research purpose. If anonymization is not possible and would harm our primary research aim, we can of course keep the data un-anonymized, as mentioned above. For data sharing, we can always seek to

obtain informed consent for sharing un-anonymous research data or control access and re-use of the data, e.g. by data licensing.

3.3.3 Wrap-Up

In sum, we have to be aware of legal requirements whenever we deal with personal information. Accordingly, we have to protect our participants from harm by taking appropriate measures, such as data anonymization, into account to ensure their confidentiality. That is, we must manipulate the data collected so that each particular individual can no longer be re-identified in our research data. Therefore, we might separate sensitive and personal information, restrict ranges of our variables or aggregate information. Last but not least, we should always keep in mind that there are experts on data anonymization, such as data security officers or employees in data archives that might provide help.

Recap: Anonymization strategies

- separate sensitive and personal information from the research data, whenever it is not needed in data analysis
- replace personal information, such as names, by pseudonyms or IDs
- check for indirect identifiers in the data
- aggregate or suppress values and do k-anonymization to prohibit indirect identification
- in the context of data sharing, accessibility of research data determines the degree of anonymization

3.3.4 Further Reading

Elliot, Mark, Elaine Mackey, Kieron O´Hara and Caroline Tudor. 2016. *The Anonymization Decision-Making Framework*. Manchester (UK): University of Manchester.

Lagoze, Carl, William C. Block, Jeremy Williams, John Abowd, and Lars Vilhuber. 2013. *Data Management of Confidential Data*. doi:10.2218/ijdc.v8i1.259.

Narayanan, Arvind and Vitaly Shmatikov. 2008. *Robust De-anonymization of Large Sparse Datasets*. https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf. [February 16, 2018].

4 DATA DOCUMENTATION

Data documentation relates to additional information (metadata) that should be provided together with a dataset. The documents we have in mind here are

- method or field reports that allow for assessing the survey quality as they list possible "errors" occurring during the survey research process based on the total survey error approach or that describe how and when survey data was gathered, how the field management was conducted, informs about response rates etc.; field reports could also be part of the method reports;
- survey instruments that include the original questionnaire and any other materials used during fieldwork processes;
- variable reports and codebooks that describe contents, structure, and layout of a dataset or data collection on the variable level.

We begin with a brief introduction to the mean of metadata and its relevance in Section 4.1. Section 4.2 examines study level documentation, covering characteristics of the first two documents. Finally, section 4.3 deals with data documentation on the variable level and the creation of codebooks and variable reports.

4.1 Metadata and Metadata Standards

Anja Perry & Uwe Jensen

Metadata is the description of data, meaning it is data about data. Metadata can be a single piece of information, such as the title of a study or the survey method used. It can also come in form of documents, such as a method report with detailed information on data collection and data processing (National Information Standards Organization, 2004).

4.1.1 What are the purposes of metadata?

We use metadata on the study level to make data findable and accessible (see section 4.2), for example through study descriptions, method reports or persistent identifiers. That way, we allow sharing and re-use of the data. On the variable level, metadata are necessary to understand, interpret, and replicate the data (see section 4.3), e.g. in terms of questionnaire(s), codebook(s), variable reports or script files. This is at the same time the primary reason why it is important to collect and document metadata. Also for curating and providing data on the long-run, storing it in a data archive or repository, and to allow for data citation, metadata is crucial. Last but not least, metadata helps re-users to assess the credibility of the data at hand (Day 2005; Gregory et al. 2009).

The scope of documentation depends on the project size, amount of data as well as personal and financial resources. In general, documentation is time consuming. It is therefore necessary that we create a data management plan and apply for sufficient grants for data documentation and processing work (CESSDA 2018).

4.1.2 Metadata standards

Metadata can be structured or unstructured. One the one hand, unstructured metadata do not follow a specific language, script or standard. These could be any kind of free-hand written documents like method reports, questionnaires, and codebooks. The aim is to provide a very detailed description of the research project. However, it is ambiguous regarding what is understood by different elements of unstructured metadata, making it difficult for (re-)users to navigate through it.

Structured metadata, on the other hand, make data findable and allows comparisons between data (Jensen et al. 2011), following certain standards. Relevant metadata standards in the social sciences are, for example:

- Dublin Core for catalogue metadata (http://dublincore.org/);
- DDI for variable and study documentation (http://www.ddialliance.org/);
- SDMX for official statistical data (https://sdmx.org/).

Further examples of structured metadata are controlled vocabulary (e.g., for variable names, such as *birth year* instead of *year of birth*) and coding standards such as the International Standard Classification of Education (ISCED). An important advantage of structured metadata are standardized guidelines of what to document, how to document and why to document a research project.

For smaller research projects metadata standards are not directly relevant. When handing over data to an archive or repository, usually some minimum of unstructured metadata, e.g., method report, questionnaire, and codebook are sufficient. Archives and repositories will then extract relevant metadata from these documents and feed them into a structured metadata standard.

4.1.3 Persistent Identifier

Metadata is used to make data findable. The metadata of one study is thereby linked to one persistent identifier (PID). PIDs are permanent, unambiguous identifiers that can be used to uniquely identify and cite data (GBIF 2011). One example of a PID is the Digital Object Identifier (DOI). The registration agency for social and economic data in Germany is da|ra, working in close cooperation with the Data Archive for the Social Sciences at GESIS – Leibniz-institute for the Social Sciences. Whenever researchers store their data at GESIS, these data will also be registered at da|ra. To do so, we need to provide few mandatory metadata which increases data findability and thus data visibility. To register a DOI name at da|ra the following information is required (Koch et al. 2017):

- *resourceType*, specifying the general resource type, e.g. *Dataset*, *Text* etc.;
- *title* of the resource;
- details on *creators*; at least the name;
- *publicationDate*;
- *availability* of the resource, e.g. *Download* or *Onsite*;
- *dataURL*, i.e. the URL of the landing page to which the DOI should be resolved.

Based on this information a unique DOI is created that refers to a landing page of the archive or repositories where the published research data are permanently stored.

The DOI consists of a suffix that always starts with 10 followed by a dot and the organization that owns the DOI. Each organization is assigned a number. This is followed by / and an ID that the owners of a DOI can determine themselves:

doi:10.ORGANIZATION/ID

A DOI can be resolved through the DOI resolver at the International DOI Foundation or by retrieving the URL. The PID can thus be used in the data citation to ensure that the data is uniquely identified.

EVS Example: Digital object identifier (DOI)

The European Value Study 2008 has the DOI:

doi:10.4232/1.12458

in which *4232* refers to GESIS – Leibniz Institute for the Social Sciences and the remainder is a self-chosen ID for this particular dataset. The DOI for EVS' first version (doi:10.4232/1.10059) was the very first DOI registered at da|ra.

Recap: Metadata and metadata standards

- meta data is data about data
- meta data standards can help to provide detailed information about a research project, but are only indirectly relevant for smaller research projects
- PIDs uniquely identify research data and can be used for citing data
- it is important to create a data management plan and apply for sufficient funds for data documentation

4.1.4 Further Reading

DataCite Metadata Working Group. 2016. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data*. Version 4.0. DataCite e.V. doi:10.5438/0012.

Gregory, Arofan and Pascal, Heus and Ryssevik, Jostein, Metadata (March 1, 2009). RatSWD (German Council for Social and Economic Data) Working Paper No. 57. http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_57.pdf. [January 19, 2018].

Lagoze, Carl, William C. Block, Jeremy Williams, John M. Abowd and Lars Vilhuber. 2013. Data Management of Confidential Data. In 8th International Digital Curation Conference (p. 15). Amsterdam. http://ecommons.library.cornell.edu/handle/1813/30924. [January 19, 2018].

Miller, Ken and Mary Vardigan. 2005. *How Initiative Benefits the Research Community - the Data Documentation Initiative*. First International Conference on e-Social Science. Manchester, UK. http://www.ddialliance.org/sites/default/files/miller.pdf. [January 19, 2018].

Plant, Richard R., 2012. *How to Add Metadata to Your Data so That You and Others Can Make Sense of It*. http://www.shef.ac.uk/polopoly_fs/1.158828!/file/Metadatav6.pdf. [January 19, 2018].

Vardigan, Mary, Pascal Heus and Wendy Thomas. 2008. *Data Documentation Initiative: Toward a Standard for the Social Sciences*. International Journal of Digital Curation, 3/1. Pp. 107–113. doi:10.2218/ijdc.v3i1.45.

Zenk-Möltgen, Wolfgang. 2012. *Metadaten und die Data Documentation Initiative (DDI)*. In: Altenhöner, Reinhard and Claudia Oellers (eds.). 2012. *Langzeitarchivierung von Forschungsdaten: Standards und disziplinspezifische Lösungen*. Berlin: Scivero. Pp. 111-126. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-46679-8. [January 19, 2018].

4.2 Metadata Documentation at the Study Level

Alexander Jedinger

Metadata documentation at the study level encompasses information of the general context in which data were collected and the methodological steps in generating the data at hand. This in turn should enable researchers to assess the quality and the analytical potential of the data in light of their own research purposes. It is common practice to document all necessary contextual and methodological information to assess data quality in a method report, also sometimes called technical or field report. The issue of survey data quality is related to possible "errors" occurring during the survey research process based on the total survey error (TSE) approach (Weisberg 2005; Biemer and Lyberg 2003). Survey errors can be divided into three categories:

- errors in representation, such as coverage error, sampling error, nonresponse error at the unit level;
- errors in measurement, such as respondent-related errors, interviewer-related errors, mode-related errors, nonresponse error at the item level;
- post-survey errors, such as adjustment error, processing error.

The key sections of a method report should address each of these potential threats to survey data quality. A methodological report consists of a front matter and a main body. The front matter should contain general information about the project like the study title, principal investigators and project team members. It should also specify the funding of the study and a recommended citation. The main body should include the main sections, discussed in the following (Jedinger and Watteler 2017).

4.2.1 Objectives and Design

In this part of the methodological report, explain from whom and for what purpose the data were collected. This includes the specific background for conducting the study, research objectives, and the overall research design (e.g. cross-section, trend, and panel).

EVS Example: Background Information on EVS

"The European Values Study (EVS) is a large-scale, cross-national, and longitudinal survey research program on how Europeans think about family, work, religion, politics and society. Repeated every nine years in an increasing number of countries, the survey provides insights into the ideas, beliefs, preferences, attitudes, values, and opinions of citizens all over Europe. Four waves of surveys were executed from 1981 to 2008. These surveys explore value differences, similarities, and value changes."

(EVS 2008, Method Report)

4.2.2 Target Population and Sampling

Describe how the respondents were selected. This includes the target population and eligibility criteria, the sampling frame, respondent selection at each sampling stage for multistage sampling procedures, and a description of any clustering and/or stratification. Also report survey outcomes such as contact rates, cooperation rates, response rates and refusal rates using the American Association for Public

Opinion Research (AAPOR 2016) standard definitions (see Section 3.2). Finally, give information about the total sample size.

EVS Example: Sampling Procedure Germany

"Persons 18 years or older who are resident within private households, regardless of nationality and citizenship or language. [...] The basis for the study is a random sample drawn from resident registers of German municipalities (a national resident register does not exist). Therefore the sampling has to proceed in two steps: (1) a random sample of municipalities had to be drawn and (2) random samples of the municipalities' resident registers. The sample design is disproportional and takes the distinction of East and West Germany into account where the East and West of Berlin are attributed to the respective parts of Germany. In order to realize the oversampling of East German population, the sample of municipalities is stratified according to the federal states and to 7 size-classes of the municipalities (in order to reflect the population of the municipalities in the sampling probabilities)."

(EVS 2008, Method Report)

4.2.3 Mode of Data Collection

Describe how the data were collected e.g. by self-administered vs. interviewer-administered, or computer-assisted vs. not computer-assisted survey modes.

EVS Example: Fieldwork documentation

"In all countries, fieldwork was conducted on the basis of detailed and uniform instructions prepared by the EVS advisory groups. The EVS questionnaires were administered as face-to-face interviews in the appropriate national language(s). As far as the data capture is concerned, CAPI or PAPI was used in nearly all countries. Exceptions are Finland (internet panel) and Sweden (postal survey)."

(EVS 2008, Method Report)

4.2.4 Survey Instrument

Describe what information was collected. That pertains to the topics of the questionnaire and the rationale for including specific question modules, the construction of scales (and their psychometric quality), any special instruments used, and results of traditional or cognitive pretests. Also include information on the translation of field questionnaires if necessary.

4.2.5 Fieldwork

Describe who has collected the data, when and where. The documentation should entail the survey organization, field dates, and if an interviewer-administered survey was conducted the number of interviewers, interviewer experience and sociodemographic profile interview duration, interviewer training and monitoring. This is also the right place to mention any peculiarities during the field period that might affect data quality.

4.2.6 Data Processing

Describe how the data were edited and processed. If the data contains open-ended questions explain the corresponding coding of answers if applicable. If the survey is based on complex sample designs or feature over-/undersampling of specific subpopulations report how design and post-stratification weights were constructed.

4.2.7 Data Protection and Ethical Issues

Finally, document whether basic principles of research ethics and data protection laws were respected. As survey interviews are based on voluntary participation of informed individuals document how informed consents was obtained from respondents. Also report all measures to protect respondent privacy, such as data anonymization, as discussed in section 3.3.

In addition to a detailed method report, the study documentation should include any material that is used during data collection, either as part of the report or as standalone documents. The survey materials include the original questionnaire(s) and any other materials used during fieldwork processes.

Basis for the documentation is the fieldwork questionnaire as used by the survey organization. Since questionnaires change very frequently over the course of a project, consider to use screenshots from the survey software to document differences between original questionnaire(s) and the questionnaire(s) in field. Another important but underrated aspect of documentation is any material that is used during data collection, e.g. advance letters to contact respondents, show cards to illustrate scales and response categories, consent forms, or audio-visual stimuli as long as they are not already part of the questionnaire documentation (e.g. pictures).

Recap: Study level documentation

- a methodological report consists of a front matter and a main body.
- the front matter should contain the study title, principal investigators, project team members, study funding and a recommended citation
- the main body should include the following information
- study objectives and survey design (e.g. cross-section, trend, or panel)
- target population, sample frame, sampling design, sample size and survey outcomes such as contact rates, cooperation rates, response rates and refusal rates
- mode of data collection
- construction of the survey instrument, including the main topics addressed and translation procedures (if applicable)
- details of the fieldwork such as the name of the survey organization and dates of the fieldwork
- for interviewer-administered surveys: number of interviewers, interviewer experience and sociodemographic profile, interview duration, interviewer training and monitoring
- details on data processing such as data preparation and cleaning, coding of open-ended questions and weighting procedures
- documentation of informed consent and anonymization procedures
- in addition, document any material that is used during data collection, e.g. the original questionnaire and showcards

4.2.8 Further Readings

Biemer, Paul P. 2010. *Total Survey Error: Design, Implementation, and Evaluation*. Public Opinion Quarterly, 74/5. Pp. 817-848. doi:10.1093/poq/nfq058.

Blasius, Jörg, and Victor Thiessen. 2012. *Assessing the Quality of Survey Data*. London: Sage.

Corti, Louise, Veerle van den Eynden, Libby Bishop and Matthew Woollard. 2014. *Managing and Sharing Research Data. A Guide to Good Practice*. London (UK): Sage Publications.

Groves, Robert M., Floyd J. Fowler, Mick Couper, James M. Lepkowski, Eleanor Singer and Roger Tourangeau. 2009. *Survey Methodology*. 2. eds. Hoboken, NJ: Wiley.

Inter-University Consortium for Political and Social Research (ICPSR). 2012. *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle*. 5th eds. Ann Arbor: Michigan University.

Vardigan, Mary B, and Peter Granda. 2010. *Archiving, Documentation, and Dissemination*. In: Marsden, Peter V. and James D. Wright (eds.). *Handbook of Survey Research*. Bingley: Emerald Group. Pp. 707-729.

4.3 Metadata Documentation at the Variable Level

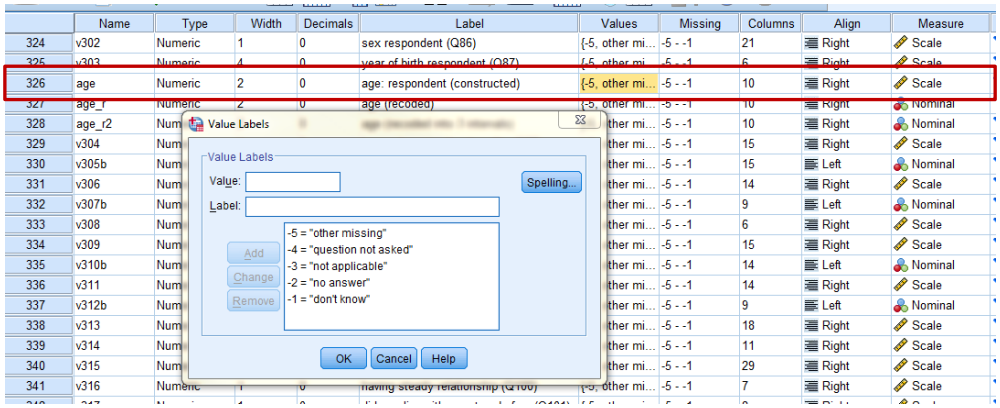
Karoline Harzenetter

In contrast to study level documentation, variable level documentation focuses on the description of the dataset itself, its structure, layout and content. A good variable level documentation should enable (re-)users not only to get a quick overview on a dataset's usability for secondary analysis. In addition, it should also provide detailed information on measurements to enable the reproduction of survey data and results and show characteristics and content of its variables. The dataset itself is object of documentation as well as every single variable within. Therefore variable level documentation can be summarized as the description of these entities: dataset(s) and variables (Corti et al. 2014). The aim of this chapter is to introduce variable or data level documentation and describe how a codebook can serve as a sufficient way of documenting data.

4.3.1 Forms of Variable Level Documentation

The variable level documentation is part of the dataset processed. If variables and values were labelled, missing values and variable types were defined and numbers were formatted during data processing, a dataset contains descriptive metadata on variable level. For example, in the European Value Study (EVS 2008) variable labels include the question number of the field questionnaire, all codes are labelled and non-responses are defined as missing, as illustrated below. Variable labels and some variable names are content related, width and numbers of columns are adjusted to their needed length. A variable that entails the information about a respondent's age, for example, has usually a maximum of 3 digits and is integer. In case of the EVS there is no respondent older than 99, therefore only 2 widths are needed to display age and the variable does not require a format with decimals.

EVS Example: Variable Documentation of Age Variable in SPSS File



(EVS 2008)

Although descriptive metadata can already be included in the dataset, too comprehensive information about a variable should not be displayed within the file itself. Sometimes researchers need descriptive information beforehand, for example in case of restricted accessibility due to privacy issues or if its usage requires registration and payment. Access to information before accessing the dataset can help to clarify if

- a dataset has a usable format,
- content of a variable is of interest,
- variables of interest have the scale of measurement that is needed for a specific statistical technique,
- there are sufficient valid answers for inferential statistical procedures,
- variables have been dropped or answer categories have been grouped due to reasons of quality and anonymization, and so on.

To avoid doing a blind bargain and to keep the dataset's structure clear and manageable, providing technical, structural and describing metadata along with the dataset is a useful addition for everybody who needs to work with the data.

Documentation of a dataset on the variable level can be provided in different forms, for example as a separate supplementary published text file (codebook) or as an online portal based on a dataset and its documentation (XML-file). The form of documentation follows its function. To browse and analyse a dataset online, a machine-readable form of variable level documentation with structured metadata is required whereas a manually created searchable PDF-codebook that is based on a text file with unstructured metadata is probably sufficient for human users to get a first impression of the data. Large scale survey programs, like the EVS, and data archives, such as the Data Archive for the Social Sciences at GESIS – Leibniz-Institute for the Social Sciences, often use a specific schema and structured metadata as fundament for their supplementary documentation. At GESIS, data documentation is based on the DDI metadata standard and large (inter-)national survey programs like Politbarometer, Allbus, ISSP and EVS are not only accompanied by codebooks (so called Variable Reports), they are also documented on the variable level in GESIS' social science data portal ZACAT, as illustrated below.

EVS Example: Documentation of Age Variable

Human Readable Codebook

European Values Study 2008
GESIS Study No. ZA4800 (v.4.0.0), doi:10.4232/1.12458

Variable, Label
Question Text (English Language)

v303 - year of birth respondent (Q87)
Q87
Can you tell me your year of birth, please?
19..

-5 other missing
-4 question not asked
-3 not applicable
-2 no answer
-1 don't know

Comparability:
Trend question: EVS 2008 and EVS 1999.
Comparable questions: WVS 1995 (v215), WVS 2000 (v224).

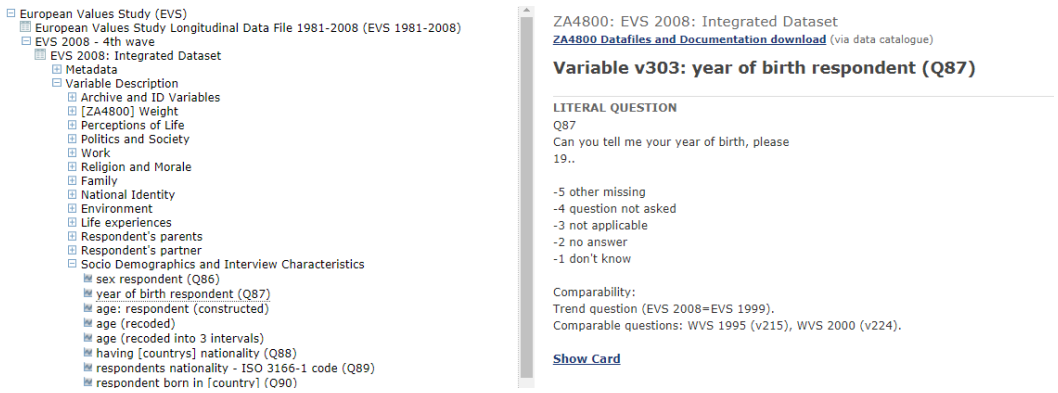
(EVS 2008 Variable Report)

Machine Readable DDI 2.0 Xml

```
<var ID="VAR323" name="v303">
    <labl>year of birth respondent (Q87)</labl>
    <qstn ID="SQ116" seqNo="Q87" sdatrefs="S1">
        <qstnLit>Can you tell me your year of birth, please</qstnLit>
        <qstnLit>19..</qstnLit>
    </qstn>
    <catgry ID="AV1466" missing="Y">
        <catValu>-5</catValu>
        <labl>other missing </labl>
        <txt ID="SA1466" sdatrefs="S1">other missing</txt>
    </catgry>
    <catgry ID="AV1467" missing="Y">
        <catValu>-4</catValu>
        <labl>question not asked </labl>
        <txt ID="SA1467" sdatrefs="S1">question not asked</txt>
    </catgry>
    <catgry ID="AV1468" missing="Y">
        <catValu>-3</catValu>
        <labl>not applicable </labl>
        <txt ID="SA1468" sdatrefs="S1">not applicable</txt>
    </catgry>
    <catgry ID="AV1469" missing="Y">
        <catValu>-2</catValu>
        <labl>no answer </labl>
        <txt ID="SA1469" sdatrefs="S1">no answer</txt>
    </catgry>
    <catgry ID="AV1470" missing="Y">
        <catValu>-1</catValu>
        <labl>don't know </labl>
        <txt ID="SA1470" sdatrefs="S1">don't know</txt>
    </catgry>
    <varFormat type="numeric" formatname="F4.0" schema="SPSS"/>
    <notes type="NoteNote">Trend question: EVS 2008 and EVS 1999.
Comparable questions: WVS 1995 (v215), WVS 2000 (v224).</notes>
</var>
```

(DDI-Codebuch for EVS 2008)

EVS Example: Web-based Documentation of Age Variable in DDI Format



(ZACAT for EVS 2008)

For smaller study projects, a human readable codebook is a sufficient and efficient way of documenting on the variable level. The decision of what information about datasets and variables is necessary and helpful for a third person's understanding goes hand in hand with the question in what form the documentation needs to be accessible and if this information is not already accessible elsewhere for example in a method report (see section 4.2). The UK Data Service (2018) recommends that variable-level notes should be embedded within a dataset when possible and more extensive variable level documentation should be created alongside in a structured metadata format such as XML (Corti et al. 2014). To keep dataset documentation simple and short we should ask ourselves before creating a codebook if it needs to be machine accessible and thus needs to be provided in a structured and standardized form like XML or if unstructured metadata is acceptable, and if there is information that is not already accessible and discoverable via other documenting material

4.3.2 Content and Structure of a Codebook

A codebook describes the contents, structure and layout of a dataset, lists its variables by names, labels and values and frequencies and provides additional details on this level in form of variable notes:

> "A codebook is an essential document that informs the data user about the study, data file(s), variables, categories, etc., that make up a complete dataset. The codebook may include a dataset's record layout, list of variable names and labels, concepts, categories, cases, missing value codes, frequency counts, notes, universe statements, and so on." (DDI 2018)

A codebook functions as a linking document between a dataset and its data creation/collection documents like the field questionnaire. It is a document connecting the raw data, information of data processing and the published dataset. It links question texts and answer categories of the field questionnaire with variable names and coding schemes of the dataset and additionally reports procedures on and issues of variables that occurred during data processing and are not displayed in e.g. the method report. If data manipulation and processing is well documented or the processed dataset is only a slightly modified version of the raw data, variable level documentation will not steal much of our time. We should also keep in mind that not all aspects of data processing are of interest and relevance for potential (re-)users and therefore need not be documented. Information about a variable, already

reported on a higher documentation level and included in, e.g., the method report, can also be ignored in the codebook.

Recap: Guiding questions to assure user friendly, exploitable and non-repetitive information in variable level documentation

- What kind of information on the dataset and data processing could be of interest for someone before accessing a dataset (e.g. deleted, aggregated variables, definition and range of missing values, applied standards for measurement?
- What information about a dataset and a variable is needed when using and preparing (e.g. recoding, number of valid answers, filter) data for analysis?
- Is there information about data processing and data evaluation that helps to interpret results after analysis and is not yet reported in the methodological report like for example translation errors in international surveys or inconsistent application of filters?
- Are there specific documentation standards, like the metadata scheme (DDI), or a specific publication form that need to be considered when preparing a codebook?

Front Page and Introductory Part

A codebook should consist of a front matter section and a main body. Its front matter section should be composed of a cover page with basic bibliographic information to enable user's citation of the document and a reference of the dataset the codebook is assigned with. Also a table of contents and a short list of all variables gives a quick overview of what to find in the dataset. Recommended information for the front pages are the codebook's title, the author name(s), the publisher, information for clear identification of the dataset, such as dataset name, study title, dataset's version[8], persistent identifier etc. The example below displays the content and structure of the EVS 2008 front matter.

Finally, setting citation standards on the front matter for the data set by naming the principal investigators responsible for data creation, the dataset's publication year, its title and version, the publisher(s) of the file and, if available, an object identifier (see section 4.1.3) is also useful. A citation recommendation simplifies paying credit to persons, who are responsible for datasets used in publications. There might be other topics noteworthy in the introductory part like purpose and format of a codebook or dataset restrictions. An overview of variables that correspond in question wording and coding within or across studies could also be notable information on variable level in such an introductory part of the document.

[8] If there are various versions of a dataset, it is useful to list changes between versions of a dataset in the codebook.

EVS Example: Content and Structure of Front Pages for the EVS

Tilburg University

gesis Leibniz Institute for the Social Sciences

VARIABLE *Reports* 2016|2

European *Values* Study

EVS 2008 - Variable Report
Integrated Dataset

Documentation release 2016/04/15

Related to the Integrated Datasets
GESIS Study-No. ZA4800, Version: 4.0.0, doi:10.4232/1.12458
GESIS Study-No. ZA4799, Version: 1.0.0, doi:10.4232/1.12483

European Values Study and
GESIS Data Archive for the Social Sciences

GESIS-Variable Reports

GESIS – Leibniz Institute for the Social Sciences
50667 Köln
Unter Sachsenhausen 6-8
Germany
Phone: +49/(0)221/47694-0
Fax: +49/(0)221/47694-199
E-Mail: evelyn.brislinger@gesis.org

ISSN 2190-6742 (Online)

Publisher: GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8, 50667 Köln
info@gesis.org, www.gesis.org

GESIS-Variable Reports No. 2016|2

EVS 2008 - Variable Report
Integrated Dataset

Documentation release 2016/04/15

Related to the Integrated Datasets
GESIS Study No. ZA4800, Version: 4.0.0, doi: 10.4232/1.12458
GESIS Study No. ZA4799, Version: 1.0.0, doi: 10.4232/1.12483

European Values Study and
GESIS Data Archive for the Social Sciences

Acknowledgements

The fieldwork of the 2008 European Values Study (EVS) was financially supported by universities and research institutes, national science foundations, charitable trusts and foundations, companies and church organizations in the EVS member countries.

A major sponsor of the surveys in several Central and Eastern European countries was Renovabis.

Renovabis - Solidarity initiative of the German Catholics with the people in Central and Eastern Europe: Project No. MOE016847 http://www.renovabis.de/.

An overview of all national sponsors of the 2008 survey is provided in the "EVS 2008 Guidelines and Recommendations", the "EVS 2008 Method Report", and on the website of the European Values Study at http://www.europeanvaluesstudy.eu/page/sponsors-of-the-2008-survey.html.

The project would not have been possible without the National Program Directors in the EVS member countries and their local teams.

Gallup Europe developed a special questionnaire translation system WebTrans, which appeared to be very valuable and enhanced the quality of the project.

Special thanks also go to the teams at Tilburg University, CEPS/INSTEAD Luxembourg, and GESIS Data Archive for the Social Sciences, Cologne.

GESIS – Leibniz Institute for the Social Sciences 2016

(EVS 2008, Variable Report)

Main Part with Detailed Variable Description

The detailed description of all variables of a dataset with variable name, variable label, question text, values, value labels, summary statistics, missing values, universe skip patterns (applied filters) and notes follows in the main body of the codebook. Contextual information like transformation proce-

dures for compiled, created or constructed variables can also be entered in the main part in form of annotations. References or citations, the function of a variable (e.g. splitting or filtering the sample) or further instructions for working with a variable can also be noted there (ICPSR 2017). In the EVS Variable Report's page header section, there is also by default a clear identification of the dataset the variable is part of (study number, study title and dataset ID number) entailed, followed by variable name, label and coding scheme in the body text. In case of a constructed or created variable, there is additionally a reference to the measurement the constructed variable is based on. There is also a note in case the variable is comparable with variables of earlier EVS wave, shown in the example, below.

EVS Example: Documentation of Aggregated Variable for Age

Continuous Variable for Age

European Values Study 2008
GESIS Study No. ZA4800 (v.4.0.0), doi:10.4232/1.12458

Variable, Label
Question Text (English Language)

v303 - year of birth respondent (Q87)
Q87
Can you tell me your year of birth, please?
19..

-5 other missing
-4 question not asked
-3 not applicable
-2 no answer
-1 don't know

Comparability:
Trend question: EVS 2008 and EVS 1999.
Comparable questions: WVS 1995 (v215), WVS 2000 (v224).

Constructed Variable for Age

European Values Study 2008
GESIS Study No. ZA4800 (v.4.0.0), doi:10.4232/1.12458

Variable, Label
Question Text (English Language)

age_r2 - age (recoded into 3 intervals)
Age of respondent - recoded (3 intervals)

Source variable: age

-5 other missing
-4 question not asked
-3 not applicable
-2 no answer
-1 don't know
1 15-29 years
2 30-49 years
3 50 and more years

Comparability:
Trend question: EVS 2008 and EVS 1999.

(EVS 2008, Variable Report)

How to create a codebook with the help of a software program is not part of this chapter but (online) manuals for creating a rudimentary codebook in statistical analyses programs like Stata or SPSS that can serve as starting point for a more elaborated variable and data documentation are listed in "Further Readings". There also exist tools for example by the DDI Alliance for creating machine readable files like pdf and XML.

Recap: Variable level documentation

- describes the contents, structure, and layout of a dataset or data collection
- describes the dataset on variable level (labels, values, ranges or frequencies)
- provides additional details on this level in from of notes (e.g. filter deviations, recodes, deviations from data processing rules,)
- it usually created for human readable purpose
- it consists of a front matter section and a main body

4.3.3 Further Readings

Colectica. 2017. *Create a PDF Data Dictionary or Codebook*. https://docs.colectica.com/designer/manage-content/data/create-data-dictionary/. [February 16, 2018].

Data Documentation Initiative Alliance (DDI). 2018. *Create a Codebook*. http://www.ddialliance.org/training/getting-started-new-content/create-a-codebook. [February 16, 2018].

Kent State University Libraries. 2017. *SPSS Tutorials: Creating a Codebook* . http://libguides.library.kent.edu/SPSS/Codebooks. [February 16, 2018].

Kent State University Libraries. 2017. *SPSS Tutorials: Defining Variables*. https://libguides.library.kent.edu/SPSS/DefineVariables. [February 16, 2018].

MAXQDA. 2018. *MAXQDA Online Manual - Creating a Codebook with Category Definitions*. https://www.maxqda.com/help-max12/reports/creating-a-codebook-with-category-definitions. [February 16, 2018].

Norwegian Centre for Research Data (NSD). 2011. *Nesstar Publisher v4.0 User Guide*. http://www.nesstar.com/help/4.0/publisher/download_resources/Publisher_UserGuide_v4.0.pdf. [February 16, 2018].

A FINAL REMARK ON DATA PROCESSING AND DOCUMENTATION

Sebastian Netscher & Christina Eder

Researchers working with data are increasingly facing questions on how to process and document these data so that others (and they) can make sense of them even years later; and so that the data are in good shape for dissemination via repositories and archives and eventually for secondary analyses. Data processing and data documentation have thus become essential in empirical social science research as these steps – taken carefully - increase data quality and at the same time ensure comprehensibility as well as interpretability of research results. High-quality data and related documents nowadays are therefore not only an asset; they are also a matter of good scientific practice as well-processed and documented data foster reproducibility of research findings as well as of research data. They also support researchers to face requirements of e.g. academic journals or research funders in the context of Open Access and FAIR principles of data.

However, it should not be denied that data processing and documentation can be an effort in a researcher's daily life. Furthermore, data management, data processing and data documentation are often neglected subjects in statistics or methods classes, so that most students never had the chance to learn about these issues at university. So when academic journals or research funders ask for these things, many – even experienced – researchers are left largely blank. This document was written to give first insights into these matters and provide readers with hands-on materials and real-life examples on how to meet journals' or funders' requirements. We also provided references to additional literature.

In sum, we like to emphasize that researchers should take their research purpose(s) and future plans for their research data (e.g. data sharing) into account from the beginning and plan their activities in this realm carefully. In this context, researchers should also consider asking for expert support to curate and disseminate data from data archives or repositories. These institutions ensure findability and accessibility of research data. They guarantee reproducibility of research findings and re-usability of research data, and thus foster researchers' reputation by linking primary investigators and their data. Moreover, (at least some) archives, like the Data Archive for the Social Sciences at GESIS, provide additional data services, such as consistency checks or data documentation according to (international) metadata standards like DDI, to unburden researchers and allow them to focus on their research aim(s) as well as training courses to introduce them to the world of data processing and documentation. .

REFERENCES

American Association for Public Opinion Research (AAPOR). 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 9. ed. Lenexa, KS: American Association for Public Opinion Research. http://www.aapor.org/Standards-Ethics/Standard-Definitions-(1).aspx [October 04, 2017].

Biemer, Paul P. and Lars E. Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, NJ: Wiley.

CESSDA. 2017. Expert Tour Guide on Data Management. https://www.cessda.eu/dmguide. [April 16, 2018].

Corti, Louise, Veerle van den Eynden, Libby Bishop and Matthew Woollard. 2014. *Managing and Sharing Research Data. A Guide to Good Practice*. London (UK): Sage Publications.

Couper, Mick P. 1998. *Measuring Survey Quality in a CASIC Environment*. https://ww2.amstat.org/sections/srms/Proceedings/papers/1998_006.pdf. [October 04, 2017].

da|ra 2018. *Registration Agency for Social and Economic Data*. http://www.da-ra.de/en/home/. [February 16, 2018].

DA-RT 2014. *The Journal Editors' Transparency Statement (JETS)*. https://www.dartstatement.org/2014-journal-editors-statement-jets. [September 18, 2017].

Data Documentation Initiative Alliance (DDI). 2018. *Create a Codebook*. http://www.ddialliance.org/training/getting-started-new-content/create-a-codebook. [February 16, 8018].

Day, Michael. 2005. *DCC | Digital Curation Manual. Instalment on "Metadata"*. https://www.era.lib.ed.ac.uk/handle/1842/3321. [January 19, 2018].

Deutsche Forschungsgemeinschaft (DFG). 2013. *Proposals for Safeguarding Good Scientific Practice*. Available at: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf. [December 18, 2017].

Digital Object Identifier (DOI). 2018. *The DOI System*. https://www.doi.org/. [February 16, 2018].

Doorn, Peter K. 2010. *Preparing Data for Sharing. Guide to Social Science Data Archiving*. DANS Data Guide 8. Amsterdam: Pallas Publications - Amsterdam University Press.

Elliot, Mark, Elaine Mackey, Kieron O´Hara and Caroline Tudor. 2016. *The Anonymization Decision-Making Framework*. Manchester (UK): University of Manchester.

European Commission (EC). 2016. *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020*. Version 2.1. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf. [December 18, 2017].

European Parliament. 2016. *Regulation 679/2916/EC of the European Parliament and of the Council on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*. http://eur-lex.europa.eu/legal-content/DE/ALL/?uri=CELEX%3A32016R0679. [February 16, 2018].

European Union. 2018. *Horizon 2020 Programme*. http://ec.europa.eu/programmes/horizon2020/en. [February 16, 2018].

European Values Study (EVS). 2010. *Guidelines and Recommendations*. GESIS-Technical Reports 2010, 16. GESIS - Leibniz Institute for the Social Sciences 2010. https://dbk.gesis.org/dbksearch/file.asp?file=ZA4800_standards.pdf. [October 04, 2017].

European Values Study (EVS). 2015. *European Values Study Longitudinal Data File 1981-2008 (EVS 1981-2008)*. GESIS *Data Archive, Cologne*. ZA4804 Data file Version 3.0.0, doi:10.4232/1.12253.

European Values Study (EVS). 2016. *European Values Study 2008: Integrated Dataset (EVS 2008)*. GESIS *Data Archive, Cologne*. ZA4800 Data file Version 4.0.0, doi:10.4232/1.12458.

European Values Study (EVS). 2016. *European Values Study. Method Report*. GESIS-Variable Report 2016/18. https://dbk.gesis.org/dbksearch/sdesc2.asp?no=4800. [February 16, 2018].

European Values Study (EVS). 2016. *European Values Study. Variable Report*. GESIS-Variable Report 2016/2. https://dbk.gesis.org/dbksearch/file.asp?file=ZA4800_cdb.pdf. [October 04, 2017].

Force11. 2017. *The FAIR Data Principles*. https://www.force11.org/group/fairgroup/fairprinciples. [March 06, 2018].

Gandrud, Christopher. 2015. *Reproducible Research with R and RStudio*. Second Edition. Chapman & Hall/CRC (The R Series).

GBIF. 2011. *A Beginner's Guide to Persistent Identifiers*. Version 1.0. Released on 9 February 2011. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/document/80575. [January 19, 2018].

GESIS – Leibniz Institute for the Social Sciences. 2015. *datorium*. https://datorium.gesis.org/xmlui/. [January 19, 2018].

GIT. 2018. https://git-scm.com/. [February 16, 2018].

Gregory, Arofan, Pascal Heus and Jostein Ryssevik. 2009. *Metadata*. RatSWD Working paper Series 57. http://www.ratswd.de/download/workingpapers2009/57_09.pdf [January 19, 2018].

Groves, Robert M., Floyd J. Fowler, Mick Couper, James M. Lepkowski, Eleanor Singer and Roger Tourangeau. 2009. *Survey Methodology*. 2. ed. Hoboken, NJ: Wiley.

Inter-University Consortium for Political and Social Research (ICPSR). 2012. *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle*. 5th eds. Ann Arbor: Michigan University.

Inter-University Consortium for Political and Social Research ICPSR. 2017. *What is a Codebook?* http://www.icpsr.umich.edu/icpsrweb/content/shared/ICPSR/faqs/what-is-a-codebook.html. [February 16, 2018].

International Standard Classification of Education (ISCED). 2018. http://uis.unesco.org/en/topic/international-standard-classification-education-isced. [February 16, 2018].

International Standard Classification of Occupations (ISCO). 2018. http://www.ilo.org/public/english/bureau/stat/isco/. [February 16, 2018].

International Organization for Standardization (ISO). 2018. ISO-3166, Country Codes. https://www.iso.org/iso-3166-country-codes.html. [February 16, 2018].

International Organization for Standardization (ISO). 2018. ISO-639, Language Codes. https://www.loc.gov/standards/iso639-2/php/code_list.php. [February 16, 2018].

Jedinger, Alexander and Oliver Watteler. 2017. *Improving the Quality of Methodological Reports in Survey Research: Practical Guidelines and a Content Analysis of Published Reports*. Paper presented at the Conference of the European Survey Research Association (ESRA), Lisbon, Portugal.

Jensen, Uwe; Alexia Katsanidou and Wolfgang Zenk-Möltgen. 2011. *Metadaten und Standards*. In: Büttner, Stephan, Hans-Christoph Hobohm and Lars Müller (eds.). *Handbuch Forschungsdatenmanagement*. Bad Honnef: Bock und Herchen. Pp. 83-100. http://opus.kobv.de/fhpotsdam/volltexte/2011/231/. [January 19, 2018].

Koch, Ute, Esra Akdeniz, Jana Meichsner, Brigitte Hausstein, and Karoline Harzenetter. 2017. *da|ra Metadata Schema: Documentation for the Publication and Citation of Social and Economic Data. Version 4.0*. GESIS Papers 2017/25. doi: dx.doi.org/10.4232/10.mdsdoc.4.0. [June 21, 2018].

National Information Standards Organization. 2004. *Understanding Metadata*. Bethesda, MD: NISO Press. http://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Met%E2%80%A6. [January 19, 2018].

Software Carpentry. 2017. *Version control with Git*. http://swcarpentry.github.io/git-novice/. [July 08, 2017].

UK Data Service. 2018. https://www.ukdataservice.ac.uk/. [February 16, 2018].

Vita, Roy. 2016. *Managing Resource-related Conflict. A Framework of Lootable Resource Management and Postconflict Stabilization*. doi:10.7802/1452

Weisberg, Herbert F. 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago: University of Chicago Press.

Wilkinson, Mark D., Michel Dumontier, IJsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton and Arie Baak. 2016. *The FAIR Guiding Principles for Scientific Data Management and Stewardship*. Scientific Data 3, 160018 EP. doi:10.1038/sdata.2016.18.

Xie, Yihui. 2015. *Dynamic Documents with R and knitr*. 2nd eds. Chapman and Hall/CRC (The R Series).

ZACAT. 2018. *GESIS Online Study Catalogue*. http://zacat.gesis.org. [February 16, 2017].

Zenk-Möltgen, Wolfgang and Monika Linne. 2014. *Metadatenschema zu datorium – Data Sharing Repositorium*. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2014/TechnicalReport_2014-03.pdf [January 19, 2018].