gesis Leibniz Institute for the Social Sciences

Survey Data Documentation

Bella Struminskaya[*], *Britta Gauly*[**†], *Jessica Daikeler*[**], *Julia Khorshed*[**], *& Alexander Jedinger*[**†]

March 2018, Version 1.0

[*] Utrecht University (Bella Struminskaya was a senior researcher at GESIS while preparing this survey guideline)
[**] GESIS – Leibnitz Institute for the Social Sciences
[†] Corresponding authors

Abstract

Documentation of research results is an essential process within the research lifecycle, which includes the steps of study planning and developing the survey instruments, data collection and preparation, data analysis, and data archiving. Primary researchers have to ensure that the collected data and all accompanying materials are properly documented and archived. This enables the scientific community to understand and reproduce the results of a scientific project. The purpose of this survey guideline is to provide a brief introduction and an overview about data preparation and data documentation in order to help primary researchers to make their data and other study-related materials long-term accessible. This overview will therefore help researchers to comply with the principles of reproducibility as a crucial aspect of good scientific practice. This guideline will be useful for researchers who are in the stages of planning a study as well as for those who have already collected data and would like to prepare it for archiving.

Citation

Struminskaya, Bella, Gauly, Britta, Daikeler, Jessica, Khorshed, Julia, Jedinger, Alexander (2018). Survey Data Documentation. Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines).

DOI: 10.15465/gesis-sg_en_024

1. Introduction

As is stated in the best practice guidelines for survey research of the American Association for Public Opinion Research, "excellence in survey practice requires that survey methods be fully disclosed and reported in sufficient detail to permit replication by another researcher and that all data be fully documented and made available for independent examination" (AAPOR, 2017).

The purpose of this survey guideline is to provide a brief introduction and an overview about data preparation and data documentation in order to help primary researchers to make their data and other study materials long-term accessible. In doing so, the guideline is based on the regulations of the GESIS data archive. This overview will help researchers to comply with the principle of reproducibility as a crucial aspect of good scientific practice. This guideline will be useful for researchers who are in the stages of planning a study as well as for those who have already collected data and would like to prepare these data for archiving.

The German Research Foundation (DFG) has developed general documentation guidelines for applicants submitting their proposals to be considered for funding.[1] Several of these guidelines involve handling the research data, including data documentation. At the stage of project planning, the DFG requires that researchers 1) conceptualize handling and long-term archiving of research data, 2) specify these processes in researchers' proposals, and 3) include the necessary costs associated with these processes in the budget planning. Furthermore, the data should be made accessible for secondary research, i.e. stored in a suitable repository.

The implementation of sufficient study documentation consists of two steps: (I) preparing the data for documentation and developing the study documentation materials and (II) data sharing and archiving. In the following we will describe the steps that can serve as a basis for data preparation and documentation.[2]

[1] http://www.dfg.de/en/research_funding/proposal_review_decision/applicants/submitting_proposal/research_data/

[2] For further information on the archiving process with GESIS, please consult the GESIS data archive www.gesis.org/datenservices.

2. Required Study Documentation

Several documents and deliverables are needed to document a study and prepare it for archiving. The following section provides information regarding the requirements for archiving a study in the GESIS data archive. This list of documents and detailed information about them can be used as a blueprint for documenting data. However, other archives might require different or additional information.

The documents needed for archiving a study in the GESIS data archive include:

- A methodological report
- Questionnaires and other study-related materials
- A codebook

In the following, a short description of each deliverable is provided.

2.1 Methodological report

The methodological report should include a front matter and a main body (Jedinger, 2017). The front matter should contain general information about the project such as the study title, names of the principal investigators and project team members. It should also specify the funding agency of the study and provide a recommended citation. The main body entails the study description as the essential part of the methodological report. The study description provides an overview of the study and encompasses the following information: research objectives and the overall research design (e.g., cross-section, trend, or panel), target population and details of the sampling method, information about the mode(s) of data collection, fieldwork period, and fieldwork agency. Furthermore, a study description should include information about the response metrics and a field pretest (in case it has been conducted), information about the number of contact attempts, and the id-number and qualification of the interviewers (in case of an interviewer-administered study).

Additional information pertaining to the description of the dataset can either be included into the study description or archived separately. This includes: syntax files, information about compliance with the data protection rules, information about applied data anonymization procedures as well as information about plausibility checks, consistency checks, and other error control processes.

Detailed check-lists and guidelines for writing study descriptions and methodological reports can be found in Watteler (2010) and Jedinger and Watteler (2017).

2.2 Questionnaires and other study-related materials

Questionnaires should provide question texts and answer options as well as information about the routing and skip logic. If data were collected in a self-administered mode, the questionnaire documentation will ideally resemble the "look and feel" of the original instrument (e.g., screenshots of the online questionnaire or a PDF-document of a mail questionnaire). Additional study-related materials that can be archived with the questionnaires include advance letters, reminder letters/emails as well as cards or other test materials used during face-to-face interviews.

2.3 Codebook

The codebook is a document that provides a connection between the data and the questionnaire. It can be static or interactive. The codebook should include variable names, question texts, answer options and value labels as well as codes for the missing values and their meaning. Every variable, either collected (e.g., year of birth) or generated (e.g., age from year of birth), must be included into the codebook. For every item of the questionnaire, the codebook must provide a reference, that is, a unique link between the item and the variable name. Often codebooks also include absolute and relative frequencies of variable distributions to provide a quick overview of the variable distributions as well as missing data. Sometimes central tendency measures/measures of dispersion such as a mean and a standard deviation are provided. If items are replicated from other studies, the item source can be documented in the codebook as well.[3] Detailed check-lists and guidelines for writing a codebook can be found in Harzenetter (2017).

If the dataset includes scales constructed by researchers who wish to archive their data, the scale handbook needs to be supplied for archiving as well. Information on the items that are used to build an index, the name of the construct, scaling method as well as scale values (mean, standard deviation, and reliability) should be provided. Furthermore, the scale book offers an opportunity to cite the sources of items that were used for scale construction. The documentation of the GESIS Panel, a large-scale panel survey of the general population in Germany, can serve as an example for extensive data documentation.[4]

[3] An example of the GESIS Panel codebook, which uses cross-sectional (within wave) and longitudinal (between waves) links between variables and provides sources where applicable, can be found here: https://dbk.gesis.org/dbksearch/download.asp?db=D&id=52375

[4] https://www.gesis.org/en/gesis-panel/documentation/

Codebooks can be created manually or with the help of special software. For studies such as the German General Social Survey (ALLBUS), DSDM (Dataset Documentation Manager) software[5] is used for the creation of codebooks. For datasets that need to be harmonized, such as cross-country surveys, Charmstats software[6] can be helpful.

3. Steps to Prepare the Data

Ideally, researchers should plan the documentation and data management processes before collecting data because the decisions about data archiving are closely connected to the decisions about study design and fieldwork (more about data documentation and data management as part of the research cycle can be found in Jensen (2012)).

Before the data gets archived, it has to be prepared by the researchers accordingly. This includes: (1) providing variable names and value labels, (2) coding the answers to open-ended questions, (3) checking that answers to open-ended questions comply with the data protection standards, (4) definition of missing values, and (5) anonymizing the dataset including the deletion of personal information and coding the regional information so that it is not possible to identify the respondents.

Furthermore, preparing the dataset for archiving includes performing plausibility checks of the values contained in the dataset, making sure that the routing is implemented correctly and ensuring that the dataset does not contain any implausible values. To find and correct implausible values, it is advisable to inspect frequency distributions of all variables. In order to find errors in the routing process, it is recommended to create cross-tabulations of the variables involved in the routing. If the dataset contains weights, the variables for the weights should be labeled. Furthermore, researchers should publish normalized weights or provide the code for the data users in order to enable them to perform the normalization procedure. For detailed information and check-lists about data preparation, see Ebel and Trixa (2015) and Schwarz (2017).

The researcher's data check should include the check for integrity of data, completeness of data, data format, correspondence between the survey instruments and data, examination of implausible values, coding errors, and duplicate cases as well as a plausibility check for the weighting variables. The variable labels should be checked for completeness and comprehensibility. Furthermore, variables

[5] https://dbk.gesis.org/software/
[6] http://www.gesis.org/en/services/data-analysis/data-harmonization/

should ideally be placed in an order that is consistent with the order of the questions in the questionnaire.

Researchers should assure that the relevant data protection provisions are respected (Eisentraut, 2017). In particular, datasets that contain individual-level data must be made anonymous in such a way that it is impossible to identify any individual (the so called de facto anonymity). For example, geographic coverage or occupational classification categories should be crude enough to be published. Furthermore, aspects of intellectual property and legality need to be clarified and verified before submitting the data for archiving.

4. Data Sharing

Depending on the nature of the data as well as the expertise and budget of the data provider there are several ways in which data can be shared.

4.1 Self-Depositing

Self-depositing solutions are primarily suitable for smaller individual studies, for which fewer resources are readily available, such as the distribution of additional material/software or syntax. The data are assigned a persistent identifier (DOI) in order for them to be clearly referenced and cited. In contrast to the archiving of large-scale studies, researchers can independently upload and edit their data, metadata, and related documents (e.g., background questionnaires, methodology report, and publications) to the platform and make them available to the academic community for further research. The advantage of this solution is that data is made available quickly and cost effectively. Disadvantages with this solution are the availability of an appropriate server and technical background to guarantee the long-term preservation of data (i.e. ensure readability and data integrity in the long-run). Furthermore, one has to fulfil the high demands of data protection regulations, missing data curation, the traceability of the data, and the control for data misuse.

4.2 Archiving by a Service Provider

The archiving process carried out for the researcher by a service provider is most suitable for large-scale cross-sectional and longitudinal surveys that allow comparisons over time and/or cross-country comparisons. Data and related documents are prepared and described following international meta-data standards, which are also used to enhance the traceability of data in professional data catalogues.

In addition, these service providers offer a range of professional data curation services, such as the creation of codebooks and variable reports, the preparation and harmonization of data, and long-term data preservation.

4.3 Data protection

Data depositors are responsible for compliance with the provisions of the European and German data protection regulations. If the collected data contains sensitive information (e.g., age, ethnicity, or health status) measures must be taken to protect the respondents from attempts to de-anonymize the data. Special cases include the archiving of process data (paradata), archiving of geo-referenced data, archiving of longitudinal datasets and studies with sensitive populations such as children or elites. In these cases additional measures have to be taken in order to guarantee the respondent's anonymity. There are dedicated solutions such as Secure Data Centers for data with a higher risk of disclosure. Further information about archiving in Secure Data Centers can be found in Kinder-Kurlanda and Watteler (2015) as well as in the General Data Protection Regulation (EU-GDPR) of the European Union.[7]

In the light of increasing data collection, researchers face challenges on data preparation, documentation, and sharing. The present guideline aimed to provide assistance in order to help primary researchers to make their data and other study materials long-term accessible.

In conclusion, researchers should already consider data documentation and archiving at the stage of study planning and data collection. This will ensure a proper study documentation which enables other researchers to re-use the data and examine further research questions as well as replicate research results to ensure a good scientific practice.

[7] www.eugdpr.org

References

American Association for Public Opinion Research (AAPOR) (2017). Best Practices for Survey Research. Retrieved from http://www.aapor.org/Standards-Ethics/Best-Practices.aspx

Ebel, T. and Trixa, J. (2015): Hinweise zur Aufbereitung quantitativer Daten. *GESIS Papers 2015-09*. Köln: GESIS – Leibniz Institut für Sozialwissenschaften.

Eisentraut, M. (2018). Data Anonymization. In S. Netscher & C. Eder, (Eds.), *Data Processing and Documentation: Generating High Quality Research Data in Quantitative Social Science Research (GESIS Paper)* (pp.39-42). Cologne: GESIS.

Harzenetter, K. (2018). Variable Level Documentation. In S. Netscher & C. Eder, (Eds.), *Data Processing and Documentation: Generating High Quality Research Data in Quantitative Social Science Research (GESIS Paper)* (pp.52-64). Cologne: GESIS.

Jedinger, A. (2018). Study Level Documentation. In S. Netscher & C. Eder, (Eds.), *Data Processing and Documentation: Generating High Quality Research Data in Quantitative Social Science Research (GESIS Paper)* (pp.48-51). Cologne: GESIS.

Jedinger, A. & Watteler, O. (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), 17.07.2017.

Jensen, U. (2012). *Leitlinien zum Management von Forschungsdaten: Sozialwissenschaftliche Umfragedaten.* GESIS Technical Report 2012-07. Köln: GESIS – Leibniz Institut für Sozialwissenschaften.

Kinder-Kurlanda, K. E., and Watteler, O. (2015): Hinweise zum Datenschutz: Rechtlicher Rahmen und Maßnahmen zur datenschutzgerechten Archivierung sozialwissenschaftlicher Forschungsdaten. *GESIS-Papers 2015/01.*

Schwarz, H. (2018) Data Consistency. In S. Netscher & C. Eder, (Eds.), *Data Processing and Documentation: Generating High Quality Research Data in Quantitative Social Science Research (GESIS Paper)* (pp.29-38). Cologne: GESIS.

Watteler, O. (2010). Erstellung von Methodenberichten für die Archivierung von Forschungsdaten. Köln: GESIS – Leibniz Institut für Sozialwissenschaften. Retrieved from https://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/Aufbau_Methodenbericht_v1_2010-07.pdf