



# FORS data management webinar series

# ata Documentation

**Dr. Marieke Heers**  
October 11<sup>th</sup>, 2022

# FORS webinar series

1.

**Informed consent** / September 27<sup>th</sup>

2.

**Data documentation**

3.

**Quantitative data anonymisation** / November 1<sup>st</sup>

4.

**Qualitative data anonymisation** / November 22<sup>nd</sup>

# Outline

1. What is documentation?
2. Documentation: Why, for whom, when?
3. Project-level documentation
4. Data-level documentation
5. Metadata
6. Documentation for data deposit

# What is documentation?

# What is documentation?

Documentation is any information that serves as a record of a research project and that renders data usable and meaningful.

- For members of the original research team
  - For secondary data users
  - For oneself

Documentation in the social sciences describes

- the context of the study
- the research process
- how the data were collected and manipulated
- the structure and the content of the data

**Why, for whom,  
when?**

# Why document?

- Without documentation, raw data are unusable.
- Documentation is vital information for anyone interested in conducting a secondary analysis of the original data.
- Systematically documented research data is the key to making the data publishable, discoverable, citable and reusable.
- Documentation allows for replication.
- Clear and detailed documentation improves the data quality.
- Documentation facilitates researchers' lives in the long run.

# What and how much to document?

The following questions can guide your reflection:

- What information would you need to work with the data?
- How long will the project last? Might you need the data in the future?
- Will the data be shared with others?
- Who might need the information?
- For what purpose(s) is the information needed?
- What might happen if the information was not available in the future – to you, to your team, to others?
- What resources are available for the work of documentation?



# Documentation for whom?

## The original research team vs. secondary data users

- Documenting your work throughout the research project helps keeping track of important decisions. This can be very useful during the project as well as beyond. For example, when revising an article.
- Proper documentation is also valuable when working in a team and dividing the work.
- While documenting for later data sharing might seem very different, it is also useful for your own reference. For example, in your own publications you can refer to a technical report that is also the basis for secondary data users.

# Documentation throughout the project

- Data documentation tends to be conducted at the end of the project.
- The descriptions of the earlier stages in the project are paid little attention to and the focus is on describing the data file that would be shared.
- Documentation should be part of the research project *from the beginning* and a constant component of the project.
- Rules on documentation should be fixed from the start of the project and updated on a regular basis.

# Documentation in the research lifecycle

Project phases to document:

- Planning
- Instrument development
- Data collection and preparation
- Data analysis
- Data archiving and sharing

## Two levels of documentation

- Documentation applies at the level of the research project and the data.
- Different approaches are applied to quantitative vs. qualitative data.

# Project-level documentation

# Project-level documentation

Project-level documentation includes information on:

- the purpose of the study
- the context of data collection: how, who and when?
- the content of the dataset and the documentation
- data instrument construction
- collection methods
- information on confidentiality
- consent
- anonymisation strategy
- data processing and manipulations
- quality assurance procedures
- information on access and use conditions

# Project-level documentation

- Project-level documentation can be summarized in a technical report, which can be written throughout the project and to which different team members can contribute.
- This report is a valuable reference for the project team and secondary data users.
- Much of the information needed for project-level documentation is produced as part of the research project or the proposal.

# Data-level documentation



# Data-level documentation

- Data-level documentation provides information at the level of individual objects such as pictures or interview transcripts or variables.
- Data-level information can be embedded in data files.
- Processes differ for quantitative vs. qualitative data.

# Data-level documentation – quantitative data

- Data files
- Cases in the file
- Names, labels and descriptions for variables
- Values for response categories
- Attribution of missing values
- Explanation of codes and classification schemes used
- Derived data created after collection, with code, algorithm or command file used
- Weighting variables and their application
- Syntaxes

# The data file(s)

Documentation should contain information about the data file

- Data type
- File type
- Format
- Size
- Cases in the file
- Data processing scripts (syntaxes)

# Variables

- Names, labels and descriptions for variables, records and their values
- Variable and value names can be embedded within the data file itself
- Description of the missing values
- Description of the weighting variable(s)
- Definition of codes and classification schemes (if applicable)

# Variable labels

- Be brief with a maximum of 80 characters
- Indicate the unit of measurement (where applicable)
- Reference the question number of a survey or questionnaire (where applicable)


# ch-x: The Swiss Federal Surveys of Adolescents

- 2016–2017 edition: life-course trajectories and mobility experiences
- Data collection during the recruitment procedure for a basic military service in which all Swiss men between age 18 and 20 have to participate in a procedure that assesses their potential fit for the military service.
- Additionally, data on a representative sample of women was collected.




# Missing variable labels

Variables ⌵ ⌵ ✕

 Filter variables here

Name	Label
Q001	
Q002	
Q003	
Q004	
Q005	
Q006	
Q007	
MRQ008_1	
MRQ008_2	
MRQ008_3	
MRQ008_4	
MRQ008_5	

Variables ⌵ ⌵ :

 Filter variables here

Name	Label
id	Numéro d'identification unique
lang	Langue du questionnaire
Q001	Rempli dans centre de recrutement
Q002	Age
Q003	Sexe
Q004	Lieu de naissance (CH-Etranger)
Q005	Canton de naissance
Q006	Pays de naissance
Q007	Age d'arrivée en Suisse
Q009M	Mère née en Suisse
Q041A	Séjour : Court séjour linguistique
Q041B	Séjour : Séjour prof./études

# Syntaxes as part of documentation

- Document all manipulations of the data.
- Describe the process of getting from the raw data to the final product.
- A master file might relate to several syntaxes that go through the full research process.



# Documenting qualitative data

- Embed data-level information in data files.
- Interviews: At the beginning of each file, add contextual and descriptive information about each interview, observation or diary.
  - For example, describe the interview setting: participants, location, time of day, etc.
- Grid with the main characteristics of cases, individuals or items studied: age, gender, occupation or location.
- Description of criteria and layout for transcriptions and anonymisation of interviews.

## **Beginning of the transcript file**

Interview date: 08.02.2013 [=8 February 2013]

Interviewer: Matt Miller

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

Occupation of interviewee: Journalist

Age of interviewee: 32

Gender of interviewee: Male

# Interview grid

Identifier	Gender	Age	Region	Interviewer	Participated in survey?
001	M	25	South	DL	Yes
002	F	65	West	WP	Yes
003	F	18	East	DL	No
004	M	19	North	WP	Yes
005	F	65	East	WP	No
006	F	45	North	DL	No
007	M	85	West	DL	No
008	M	63	South	WP	Yes
009	F	35	West	WP	No
010	F	38	North	DL	No
011	F	21	West	WP	Yes

# Metadata

# Metadata

- Metadata is data about data.
- Humphrey (2014): “a class of information that exists solely to describe other information systematically for discovery, retrieval, and reuse purposes”.
- Descriptive and structured record of the information produced throughout the research project. It facilitates cataloguing data and data discovery.
- Machine-readable metadata explain the purpose, origin, methods, time, location, creator(s), terms of use, and access conditions of data.
- Different systems can transfer the content, while its meaning is maintained.
- Metadata refer to the project- and the data-level.

## Metadata standards

- Different metadata standards exist.
- Within the social, behavioral and economic sciences the metadata standard of the Data Documentation Initiative (DDI) is mainly used.
- Your chosen repository ideally guides you through the process of providing the required information.

# Metadata in SWISSUbase

Dataset title

EN

Khoekhoegowab Lexical Study of Personality - Qualitative interview Responses

DOI	10.23662/FORS-DS-1216-1
Ref dataset	1216
Ref study	13355
Dataset language	English
Additional information	<p>The goals of the qualitative interviews relevant to the lexical study were (1) to determine the local attribution of meaning and common usage of targeted terms identified in the lexical study, and (2) to explore key etic traits that were absent from emic results. Terms were chosen for qualitative clarification where the English definition of a Khoekhoe word in the only Khoekhoe-English dictionary seemed surprising or incongruent given its association with other Khoekhoe terms loading on the same component in the results of the quantitative study. To explore the near-absence of certain content from the Khoekhoe lexicon, namely Extraversion and Openness, we sought individual interpretations of related terms. We also explored the use of personality trait terms adopted from other languages, given the multi-lingual context in which Khoekhoegowab speakers, like most Africans, live. Some descriptive quotes were provided in the manuscript for each theme identified in the responses to each question. Here we make available complete summaries, including all responses provided by participants. Data file includes only questions 5, 10, and 11</p>
Dataset version	1.0.0
Version notes	
Errata	
Publication date	10.11.2020
See variables on De Visu	
Bibliographical citation	<p>Amber Gayle Thalmayer, Sylvanus Job, Elizabeth N. Shino, Sarah L. Robinson: Khoekhoegowab Lexical Study of Personality - Qualitative interview Responses [Dataset]. Université de Lausanne - Faculté des Sciences sociales et politiques - SSP - Institut de psychologie. Distributed by FORS, Lausanne, 2020. <a href="https://doi.org/10.23662/FORS-DS-1216-1">https://doi.org/10.23662/FORS-DS-1216-1</a></p>

# Documentation for data deposit





# Find data and projects within Switzerland

11'832 studies

# Documentation for depositing data in SWISSUbase

Report which describes the following aspects:

- Project (context, objectives, hypotheses)
- Method (population, sampling, data collection mode)
- Data (cleaning, description of the constructed or recoded variables, weighting, anonymisation)

Documents on data collection:

- Information sent to research participants
- Instruments used for data collection (questionnaires, interview schedule, etc.)
- Materials presented to respondents during data collection
- Instructions to the interviewers

# Documentation for depositing data in SWISSUbase

If available, also include the following documentation:

- Publications and final reports
- Codebook
- Instructions for coding
- Syntaxes
- Data management plan

# Wrapping up

# Take home messages

- Careful planning of documentation at the beginning of the project helps save time and effort.
- Think about the information that is needed to understand the data.
- Plan where to deposit the data. The repository probably follows a specific metadata standard.
- Document consistently throughout the project.
- Documentation provides contextual information about the dataset(s). Rich and structured information helps others to identify a dataset and make choices about its content and usability.
- Using English for documentation increases the chance the data are discovered, understood, and reused.

## Want to learn more?

- Data archiving
- Data management support
- FORS guides
- SWISSUbase
- ...

[www.forscenter.ch](http://www.forscenter.ch)

[dataservice@fors.unil.ch](mailto:dataservice@fors.unil.ch)

[Marieke.Heers@unil.ch](mailto:Marieke.Heers@unil.ch)

# FORS webinar series

1.

**Informed consent** / September 27<sup>th</sup>

2.

**Data documentation** / October 11<sup>th</sup>

3.

**Quantitative data anonymisation**/November 1<sup>st</sup>

4.

**Qualitative data anonymisation** / November 22<sup>nd</sup>

# Resources

- CESSDA Training Team (2017 - 2022). *CESSDA Data Management Expert Guide*. Bergen, Norway: CESSDA ERIC. Retrieved from <https://dmeg.CESSDA.eu/>
- Humphrey, C. (2014). Metadata in the Social Sciences. In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 4002-4004). Dordrecht: Springer Netherlands.
- Struminskaya, B., Gauly, B., Daikeler, J., Khorshed, J., & Jedinger, A. (2018). *Survey data documentation*. Retrieved from Mannheim: [https://www.ssoar.info/ssoar/bitstream/handle/document/61699/ssoar-2018-struminskaya\\_et\\_al-Survey\\_Data\\_Documentation.pdf?sequence=3](https://www.ssoar.info/ssoar/bitstream/handle/document/61699/ssoar-2018-struminskaya_et_al-Survey_Data_Documentation.pdf?sequence=3)



# Questions?