2nd International Summer School in Uganda

Syllabus for course: “Research Data Management”

Lecturers:	Mahadia Tunga	Jessica Daikeler	Kabano H. Ignace
E-mail:	mahadiatunga@gmail.com	Jessica.daikeler@gesis.org	kabanoignace@gmail.com
Homepage:		https://www.gesis.org/en/institute/staff/person/jessica.daikeler	

Date: January 17-21, 2022
Time: 8:30-12:30 and 13:30-17:00

About the Lecturers:

Mahadia Tunga is a co-founder of the Tanzania Data Lab (dLab), which aims to strengthen the data ecosystem in Tanzania and Africa through capacity development. Mahadia is trained as a computer scientist and specialised in data science. She has vast experience in managing capacity development projects, gender-based and youth engagement programs with a special interest in young girls. Kindly visit sample projects in twitter pages (@dlabtz and @TungaMahadia) and through https://bit.ly/2K1kA3e and https://bit.ly/2lHPO2q. Since 2015, Mahadia has delivered a number of strategic consulting in research and capacity building programs on open data, data innovation, management, visualisation and analysis to several governments, non-government organisations and private entities. She has trained over 2000 individuals and 50 organisations in Tanzania, Uganda, Congo, South Africa, Egypt, and other countries.

Jessica Daikeler is a survey methodologist and works at GESIS in the Survey Operations and Survey Statistics teams in the Survey Design and Methodology department. She holds a PhD from the Statistics and Social Research Methods department in Sociology at the University of Mannheim, Germany. Jessica has already collected her own data and made it available for further processing and has a broad expertise to work with different research data sets. At GESIS she is involved in the application of evidence-based methods, in particular experiments and meta-analyses as well as in developing open science structures. Her research is currently focused on data quality in web and mobile surveys, data linkage and methods for the accumulation of evidence.

Kabano H. Ignace is a senior lecturer of Demography and Statistics in the department of Applied Statistics, School of Economics, College of Business and Economics and is the head of training at the African Centre of Excellence in Data Science at the University of Rwanda. He holds a Msc in Demography from State University of Groningen and PhD in Demography from Utrecht in the Netherlands. He has completed a Training of Trainers (ToT) on Data Management with Software Application from Harvard University (USA). With 18 years in academia, he has taught numerous courses related to Demography, Statistics, Economics and research methodology in both Social and Data sciences. He is an expert on resettlement action plans and has consulted public and private institutions, international organizations and local NGOs.

Short Course Description:

The course “Research Data Management” introduces participants to strategies, processes, and measures required to assure the quality, understandability, and (re)usability of research data from an Open Science and Open Data perspective. Not only is replicability of research data and research findings considered an integral part of good scientific practice. More and more research funders require active data management to ensure that data is of high quality and can be re-used by researchers for new research purposes. Participants will gain relevant information on openness in science and replicability ensuring that their research data is FAIR (*findable, accessible, interoperable,* and *reusable*). The course will cover a) researching data, b) data management plans, c) data collection, d) data processing, e) ethical and legal aspects of data sharing, f) documentation and metadata, g) data storing and archiving. The course will introduce the FAIR principles to guide researchers in creating re-usable research data, increasing transparency as well as replicability of research findings.

Each day will take six hours of classroom instructions, combining lectures in which the theoretical foundations of the literature are discussed, with discussions and practical examples, giving participants the opportunity to discuss their research projects and data.
The course is targeted at researchers and practitioners who produce qualitative or quantitative data and want to learn how to efficiently manage this data and ensure its reusability or how to work with quantitative data and want to understand how the FAIR principles can be implemented in their research. In the course, participants will develop a) familiarity with the idea of Open Science and the principles of FAIR data; b) an understanding of research data management; c) the skills to set up and implement a data management plan; d) the ability to efficiently handle research data; and e) the skills to prepare data in a way that makes it re-usable by other researchers.

Course Prerequisites:

- Participants should be experienced in working with quantitative research data
- Participants should be well-versed in using one of the main statistical software packages, such as Stata, SPSS, or R

Target Group:

Participants will find the course useful if:

- They are social science researchers at an early stage of study planning or data collection, working with quantitative data (principal investigators, researchers who are part of project teams, individual researchers and PhD students):
- They are faced with challenges related to data protection, data cleaning and documentation and have little experience in dealing with them so far;
- They aim to share their data for re-use after the end of the research project and/or want to learn how to ensure reproducibility of their research findings.

Course and Learning Objectives:

By the end of the course participants will:

- have gained a basic understanding of research data management in social science research within the larger data lifecycle;
- be familiar with techniques of data cleaning and data documentation, as well as preparing their data for re-use
- be aware of ethical and legal challenges to data sharing resulting from data protection regulations and intellectual property rights
- be familiar with applying re-use licenses to their data.

Software and Hardware Requirements:

Please bring your own laptops for use in the course. We will use R. For some exercises we might show something in Stata. Participants can follow us or take the package they are most familiar with.

Long Course Description:

Non-reproducible single occurrences are of no significance to science." - Popper (1956, p.66) With this quote Karl Popper already named in 1956 a highly relevant issue in science - the replicability of scientific studies. Particularly in the last decade, key results of many scientific studies in the social and life sciences have been difficult or impossible to replicate. Researchers have had trouble replicating their own work and this of others, this phenomenon is also known as the replication crisis (Baker 2016).

One consequence of the replication crisis is the increased global focus on open science. Open Science aims at increased research transparency, better possibilities for quality assurance of scientific work and an efficient exchange of knowledge. Open Science comprises the areas of Open Access, Open Data, Open Source and Open Educational Resources. The management of research data is thus a cornerstone of Open Science with direct influence on Open Data and Open Source.

What does research data management mean?

Research data management means the organization and administration of data, as well as all measures to maintain their usability. In the best case, you plan the management of your research data even before

the research process begins. A transparent structure in the management of research data is essential for research processes. The goal of sustainable data management is the collection, storage and documentation of research data and results based on standards, for instance the FAIR principle and their optional preparation for subsequent usage.

What are the advantages of managing research data?

Keep the overview.

You describe the data so that you and others can better understand it - even in a few years.

Name and structure your files.

When the research project is complete, you prepare research data for long-term archiving.

Keep your data.

You store your research data in a planned manner and make sure that it is backed-up regularly. You control changes to your files through versioning. Planning collaboration with project partners and giving them access to specific data prevents unforeseen difficulties and delays.

Ensure reproducibility.

Documentation as part of research data management makes it easier to reproduce your results.

Comply with third party requirements.

Funders, publishers and your institution require you to keep your data available. Structured data management before, during and after the project makes it easier for you to meet this requirement and you do not have to create order afterwards.

Bring science as a whole forward.

By being able to understand data sets even after years, you avoid duplication of work, save funding and speed up scientific progress overall.

In this workshop the most important aspects of research data management will be considered, such as

- Research data
- Life cycle of research data
- Research Data Policies
- Data management plan
- Structuring of data
- Anonymization
- Documentation
- Storage and backup
- Long-term archiving
- Access security
- Publication of research data
- Re-use of research data
- Legal considerations

On the first day, the course focuses on Open Science and the role of research data management in an Open Science context. You will learn about the FAIR principle in data management. We will also look at secondary data sources and deal with the re-use of research data.

On the second day of the workshop, the course focuses on research data management,ethical and legal aspects of data collection and sharing. You will learn about the research data management lifecycle, goals of research data management and the data management plan. On the ethical and legal aspects, you will learn about informed consent for data collection and sharing, the rights of your respondents and General Data Protection Regulation for the EU (GDPR).

The third day begins with a short excursion into primary data collection, the work of a survey designer is presented as well as the documentation of the fieldwork procedures.Then the first aspect of data preparation, data cleaning, is carried out. You will learn how to anonymize your research data and how to create a comprehensible data structure.

On the fourth day we will have a further focus on data preparation. You will learn data cleaning techniques such as how to deal with missing values and inconsistencies. You will further learn on documentation on the variable level and creating a Codebook.

The last day will be on data storing, archiving, and sharing. You will learn data security and protection, data protection tools, back-up strategy, file sharing and collaborative environments.

Teaching units and practical exercises are carried out alternately. The exercises are carried out with research data from the lecturers and submitted in the evening. The lecturers will provide detailed feedback on the participants' proposed solutions. At the end of the week the participants should be able to implement a research project from the data management plan to data archiving and data re-use according to the rules of Open Science.

Why do we need Research Data Management?

"Research Data Management" is the effective management of information created in the course of research. This effective management includes the handling, organization and structuring of research data. Effective planning and implementation of research data management not only has organizational advantages in archiving and publishing the data and raising third-party funds, but also has individual advantages for the researcher. On the one hand, valid research data management helps the researcher to reproduce the results even with a time lag and helps potential reviewers to better understand the data and analyses. Transparent data management can thus help the researcher to establish long-term continuity in research and reputation. To make your research as time-efficient, reproducible and secure as possible, it is important that your data management is well thought out, structured and documented. A good data management strategy considers technical, organizational, structural, legal, ethical and sustainability aspects. The time invested in creating a good data management strategy pays off when the time comes to reproduce your analyses and results.

What is the Research Data Management course about?

The aim of this course is to enable participants to cope with the technical, organizational, structural, legal, ethical and sustainability aspects of data management. Information about your data management needs to be easy to find and understand, not least if you are working on a project that will last several years and involves a large team of people. To simplify data management, a Data Management Plan (DMP) can be created early in the research process. A DMP is a formal document that provides a framework for the management of data during and after the research project. It is updated throughout the project to ensure that changes are tracked over time and to show the current status of your project. This course will empower its participants to create a DMP for their project and evaluate other DMPs. However, this course does not only focus on creating a DMP, but also integrates effective data management into the Open Science context. Furthermore, this course will take a short excursion on data collection and its documentation on the third day (see detailed schedule below).

Lecturing Format

The course will consist of two theoretical and two practical sessions each day. Feel free to ask questions at any time. We want to make this workshop as interactive as possible and all slides and other materials will be available
through a learning platform.

Requirements

The participants are always welcome to work on their specific data projects. However, we will also provide example data for usage and give exercises. The data will be shared through our learning platform. For data cleaning, we will use the Software package R, however participants are free to use their preferable Software.

Day-to-day Schedule and Literature:

Day	Topic(s)

1	Introduction to Open Science, the FAIR principles, and exploring existing data sources
	- Why is Open Science essential and what has Open Science to do with data management?
	  ○ Why do we need Open Science?
	  ○ Aspects of Open Science
	  ○ What can I do as data manager to implement Open Science – (the FAIR principle)
	- Primary and Secondary Data
	  ○ Pros and cons of reusing data
	  ○ Searching secondary data
	  ○ Assessing secondary data

	Compulsory reading (has to be read before the session): Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9. Struminskaya, Bella, Gauly, Britta, Daikeler, Jessica, Khorshed, Julia, Jedinger, Alexander (2018). Study Documentation. Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines). DOI: 10.15465/gesis-sg_en_24 Suggested reading (suggested, yet does not have to be read before the session): Bezjak, S., Clyburne-Sherin, A., Conzett, P., Fernandes, P. L., Görögh, E., Helbig, K., ... & Ross-Hellauer, T. (2018). *The open science training handbook* (https://hesso.tind.io/record/3298/files/Schneider_2018_open_science_training_handbook.pdf?version=1]. Fecher and Friesike (2014). Open Science: One Term, Five Schools of Thought. doi.org/10.1007/978-3-319-00026-8_2 Masuzzo and Martens (2017). Do you speak Open Science? Resources and tips to learn the language. doi.org/10.7287/peerj.preprints.2689v1 Watson (2015). When will 'Open Science' become simply 'science'?. doi.org/10.1186/s13059-015-0669-2

2	Research data management; ethical and legal aspects of data collection and sharing
	- Introduction to research data management
	  ○ Why should we care about research data management?
	  ○ Overview of the research data management lifecycle
	  ○ Goals of research data management
	  ○ The Data management plan
	- Ethical and legal aspects of data collection and sharing
	  ○ Why data protection?
	  ○ Informed consent for data collection and sharing
	  ○ The rights of your respondents: Protecting identities and regulating access
	  ○ General Data Protection Regulation for the EU (GDPR)
	  ○ Research ethics

	Suggested reading: https://www.ncbi.nlm.nih.gov/books/NBK321546/

3	Data collection and cleaning
	- Data collection
	  ○ An overview on data collection
	  ○ Toolbox for survey designers
	  ○ Documentation of the fieldwork process
	
	- Data cleaning I
	  ○ Anonymization of quantitative and qualitative data including audio visual
	  ○ Dataset structure and re-naming

	Compulsory reading:

	Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011) Inference and Errrors in Surveys - chapter 2; pp.39-63In: Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology* (Vol. 561). John Wiley & Sons. Schaurer, I., Kunz, T., & Heycke, Tobias (2020). Documentation of online surveys. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS - Survey Guidelines).DOI: 10.15465/gesis-sg_en_031 Sven Stadtmüller and Beuthner, Christoph (2020). Documentation of mail data collection. Mannheim,GESIS - Leibniz Institute for the Social Sciences (GESIS- Survey Guidelines).DOI: 10.15465/gesis-sg_en_032 Suggested reading: Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology* (Vol. 561). John Wiley & Sons. Schmidt; I. & Lechner,C. M. (2020). Documenting Measurement Instruments for the Social and Behavioral Sciences. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS - SurveyGuidelines).DOI: 10.15465/gesis-sg_en_033 https://www.gesis.org/en/gesis-survey-guidelines/operations

4	Data cleaning and preparing data for reuse
	- Data cleaning II
	  ○ Data preparation
	  ○ Missing values
	  ○ Consistency checks and re-coding
	  ○ How to deal with inconsistencies
	
	- Documentation on the variable level and creating a Codebook
	- Persistent identifiers
	  - Metadata Standard DDI

	Suggested reading: https://theodi.github.io/dlab2.learndata.info/management/#/

5	Data storing, archiving, and sharing
	- Data storing
	  ○ Data security and protection
	  ○ Technical and organizational data protection tools
	  ○ Back-up strategy
	- Archiving and Sharing
	  ○ File sharing and collaborative environments
	  ○ From closed to open data
	  ○ Licenses and Rights
	  ○ What an archive will do for you
	  ○ Overview on data archives
	- Take home message on data documentation

	Compulsory reading: tba Suggested reading: tba

Preparatory Reading:

O'Connell, Mary & Plewes, Rapporteurs. (2015). Sharing Research Data to Improve Public Health in Africa: A Workshop Summary (2015). 10.17226/21801.

Additional Recommended Literature:

de Jonge, Edwin & Loo, Mark. (2018). Statistical Data Cleaning with Applications in R. 10.1002/9781118897126.