

Data management and planning for Open Science

Faculty of Humanities and Social
Sciences
Zagreb, December 4, 2019

Outline

1. Presentation of FORS
2. Open science and data sharing
3. Introduction to data management
4. Data management planning
5. Day-to-day data management
 - Informed consent
 - Data anonymisation

1. Introduction

FORS main activities

Swiss Centre of Expertise in the Social Sciences

- Methodological research
- Large surveys
- Data and research information services (DARIS)

DARIS

FORSbase

Data archiving

New requirements

Long-term preservation

Enhance the value of research projects

Data access

Direct access to:

+ 500 datasets

+ 11'000 project descriptions

Data management

Training

Consultancy

Development of materials (i.e. guides)

Data management at FORS

Early days: Focus on DM from a data service point of view



- Early 2010's: increasing awareness of the importance of data management
- We follow the flow

Current days: Focus on DM from the researcher's point of view



- The ch-x experience
- Need for more concrete guidelines and solutions
- Pilot projects

Open science and data sharing

Open Science

“Open science is about the way researchers work, collaborate, interact, share resources and disseminate results. A systematic change towards open science is driven by new technologies and data, the increasing demand in society to address the societal challenges of our times and readiness of citizens to participate in research” (Amsterdam Call for Action, p.4)



Open data

“Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control”. [Wikipedia]

It follows three rationales:

- Ideological: ‘data as public goods’
- Scientific transparency: ‘data replication’
- Economic: ‘data reuse’

Some advantages to data sharing

On top of facilitating the reproduction and verification of research results as well as allowing data re-use, data sharing:

- makes research work and results more visible;
- increases the number of citations of scientific articles for which research data is also published;
- encourages new collaborations and new avenues of research;
- meets the requirements of some scientific funders and publishers.

New requirements

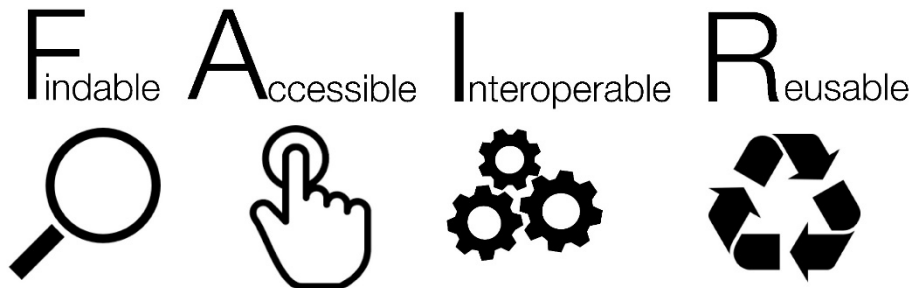
from funders:

- Data management plans (DMPs);
- Data sharing (in FAIR repositories)

from journals:

- Deposit of data used in publications
- Sufficient documentation

FAIR principles



In practice: ‘accessibility’ rather than ‘openness’

“This is a key, but often misunderstood element of FAIR. The ‘A’ in Fair does not necessarily mean ‘open’ or ‘free’. Rather, it implies that one should provide the exact conditions under which the data may be accessed. Hence, even heavily protected and private data can be FAIR” (www.go-fair.org/fair-principles)

Challenges to data sharing

Different levels of resistance

- across disciplines;
- across methodologies;
- across cultures.

Lack of know-how

- lack of discipline-specific guidance;
- ‘contradictory’ forces: data openness versus data protection

Data protection legal framework



European level: the GDPR applies from May 25, 2018

- ➔ any EU researcher who collects personal data about a person anywhere in the world;
- ➔ any researcher based outside the EU who collects personal data on EU citizens



National level: countries have their own legal frameworks (e.g. the UK Data Protection Act)



Region level: (e.g. Swiss cantonal laws)



Domain-specific laws (e.g. health domain)

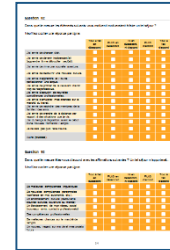
©iStock/AndreyPopo

Data management

Research data



written notes



QUESTION	YES	NO	NEUTRAL	OTHER
1. How satisfied are you with the service?	85%	10%	5%	0%
2. How easy was it to use the system?	70%	20%	10%	0%
3. How often do you use the system?	90%	5%	5%	0%
4. How helpful was the training?	60%	15%	15%	10%
5. How likely are you to recommend the system?	75%	10%	10%	5%

survey data



visually recorded data



observational data



human subject data



audio recorded data



pictures



institutional data

Definitions

Data management includes all activities associated with data other than the direct collection and use of the data.

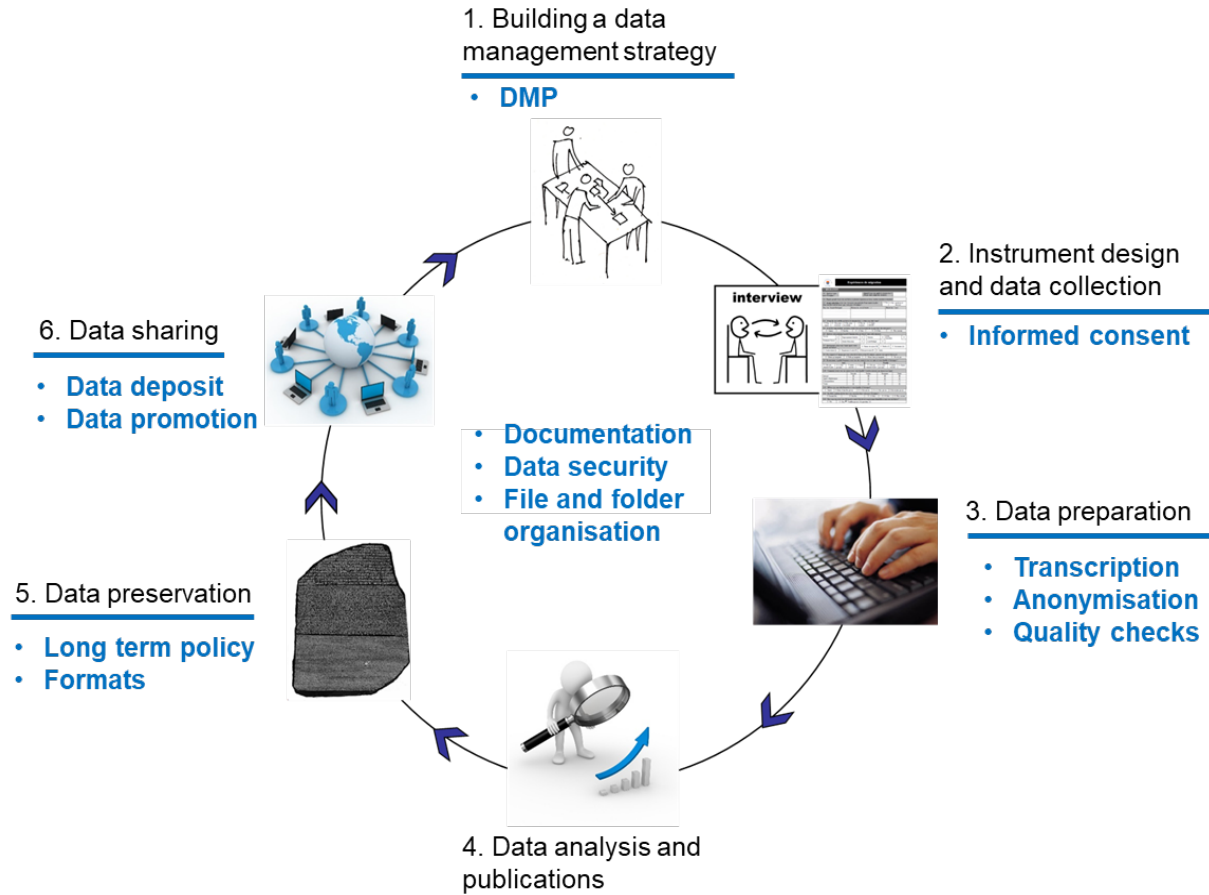
It covers all aspects of handling, organising, documenting and enhancing research data, and enabling their sustainability and sharing.

Advantages of good data management go beyond meeting open data requirements. It also saves significant time and improves data quality.

Some key data management skills for research:

- data and project planning;
- data collection considerations (e.g. informed consent)
- data preparation;
- documentation;
- anonymisation;
- data organisation;
- data storage and security;
- dissemination;
- copyright; and
- data sharing.

Data lifecycle



Planning versus implementation

Data management planning

vs

Day-to-day data management

- general overview
- generally rather brief
- intentions of good practices
- strong focus on data sharing
- expected 'problems'

e.g., DMP

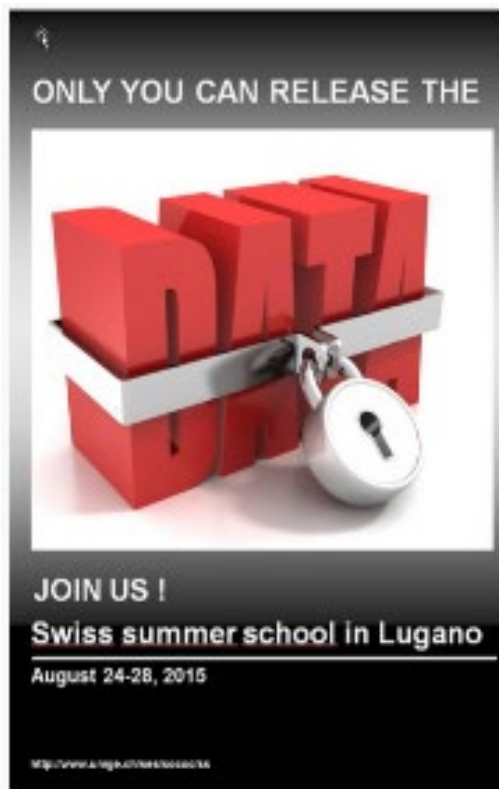
- applied data management
- detailed strategy
- clear rules
- focus on immediate needs throughout the life-cycle
- actual solutions

e.g., fixing rules; drafting a consent form

Data management planning

The example of Switzerland

2015



2019



A new funder's policy

- The SNSF implemented DMPs in October 2017
- DMPs concern all data collected and produced in the course of the research project
- DMPs are mandatory for each research proposal
- DMPs are not part of the review process
- DMPs are meant to be a living document
- At least the data underlying a publication must be shared
- The SNSF recognises exceptions to sharing

Content of the SNSF DMP

1. Data collection and documentation

- What data will you collect, observe, generate or re-use?
- How will the data be collected, observed or generated?
- What documentation and metadata will you provide with the data?

2. Ethics, legal and security issues

- How will ethical issues be addressed and handled?
- How will data access and security be managed?
- How will you handle copyright and intellectual Property Rights issues?

3. Data storage and preservation

- How will your data be stored and backed-up during the research?
- What is your data preservation plan?

4. Data sharing and reuse

- How and where will the data be shared?
- Are there any necessary limitations to protect sensitive data?
- [checkbox: I will choose digital repositories conform to the FAIR data principles]
- [Yes/No button: I will choose digital repositories maintained by a non-profit organisation]

Some key questions with respect to sharing:

- What data to share versus to preserve?
- How to handle ethical issues?
- How to ensure data security?
- Where to share data?

What data to share?

- Research data and related documentation

Research data can be defined as “*recorded factual material commonly retained by and accepted in the scientific community as necessary to document and validate research findings*” (Diaz, 2019)

They can be categorized in different ways:

- Data types (primary data, existing data, new data resulting from processing);
- Production conditions (observational, experimental, simulation, etc.)
- Format (audio, text, video, etc.)
- Legal status

- We need to distinguish active research data, the preservation of part of the data, and permanent archiving (sharing)
- Special attention needs to be given to personal and sensitive data

Personal data

Personal data fall under the data protection legislation. They consist in all information that can be traced to a person. Most common in research are:

- Names and background information;
- Information that can be traced online (social media);
- Information that is connected to a personal id (administrative data);
- Voice and video

Sensitive data (special categories of personal data)

Although definitions may change across cultures and legal frameworks, personal data are usually considered sensitive when they relate to the following topics:

- Racial or ethnic origin;
- Political opinions;
- Religious or philosophical beliefs;
- Trade union membership;
- Physical or mental health
- Sex life
- Criminal offences and court proceedings;
- Genetic data;
- Biometric data.

Under the GDPR:

- All processing of personal data requires a legal basis. The most common for research are (article 6):
 - a) *consent*,
 - e) necessary for the performance of a task carried out in the *public interest*
 - f) necessary for the purposes of the legitimate interests
- Special categories of data (article 9) are prohibited unless:
 - a) *explicit consent*,
 - e) personal data are manifestly made public by the data subject
 - j) necessary for archiving, scientific or statistical purposes (article 89)
- Anonymous data are not considered as personal data

How to handle ethical issues?

- The collection of personal and sensitive data is not a reason not to share (all) data;
- The DMP should reflect on the measures taken to protect research participants while making data shareable. This includes seeking informed consent and proceeding to data anonymisation.
- If data sharing is not possible, this has to be fully justified.

How to ensure data security?

To reduce the risks of destruction, damage, disclosure, falsification, loss, hacking, or data theft, various actions can be taken:

- development of a data security policy based on the degree of sensitivity of the data;
- securing premises and environments (computers, storage devices, servers, etc.);
- backups of all data, at regular intervals on at least three different media;
- regular updating of environments, software and computer programs;
- encryption;
- anonymisation or deidentification of confidential or sensitive data;
- data destruction.

Where to share data?

- FAIR and non-commercial repository
- Institutional repositories, discipline repositories (CESSDA service providers), project archives, online solutions (Dataverse).

Archives offer:

- Long-term preservation
- Checking of data and documentation
- Catalogue for discovery
- Dissemination
- Access control (accreditation, contracts, embargos, prior approval)

Informed consent

What is informed consent?

Informed consent is the process by which a researcher discloses appropriate information about the research so that a participant may make a voluntary, informed choice to accept or refuse to cooperate. Gaining consent must take into account any immediate or future uses of data.

Under the GDPR, it is a requirement that when personal data are collected or processed for research, participants must be informed about the purpose of the research and what will happen to their contribution.

Source: CESSDA Data management expert guide

To obtain consent, researchers should:

- Inform participants about the purpose of the research;
- Discuss what will happen to their contribution;
- Indicate the steps that will be taken to safeguard their anonymity and confidentiality; and
- Outline their rights to withdraw from the research.

Informed consent under the GDPR must be:

- **Free:** genuine choice
- **Specific:** clear information on the extent and consequences
- **Informed:** the content should be easily understood and accessible
- **Unambiguous:** silence, pre-ticked boxes and inactivity are not valid.

Information sheets:

A good information sheet usually addresses the following topics:

- The purpose of the research;
- What is involved in participating in the research;
- Details of the research;
- The procedures for withdrawing from the research project;
- The planned usage of the data during the research, dissemination, storage, publishing, and archiving of the data;
- The strategies for assuring ethical use of the data; and
- The procedures for safeguarding personal information, maintaining confidentiality, and anonymising data.

Consent procedures:

Consent procedures need to be tailored to the project (context, methods, nature of the data, planned data use, etc.)

- Consent can be written or oral (make sure language is adapted!)
- It may be given at different moments in time, all at once, or in several steps (e.g., before data collection, after data collection, retrospectively).
- Under the GDPR consent needs to be documented.

Data anonymisation

Anonymisation

The notion of anonymization refers to a process by which the elements allowing the identification of a person are definitively deleted from a dataset, a document, an interview transcript, etc. As a result, an individual cannot be identified without significant effort.

As such, it needs to be distinguished from pseudonymisation and confidentiality.

Pseudonymisation

Refers to the removal or replacement of identifiers with pseudonyms or codes, which are kept separately and protected by technical and organisational measures. The data remain pseudonymous as long as the original identifying information exists.

(<https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html>)

Confidentiality

Refers to ensuring that data “is only accessible and interpretable by authorized users in a specific context of use”.

<http://iso25000.com/index.php/en/iso-25000-standards/iso-25012/126-confidentiality>

How can identity be disclosed?

A person's identity can be disclosed through identifying information: a value may, possibly in combination with other values, lead to (re)identification. Identifying information may consist of:

- direct identifiers (e.g., name, address, telephone number, voice, picture, bank account number, social security number, ...);
- indirect identifiers – possible disclosure in combination with other information (e.g., occupation, geography, unique or exceptional values or characteristics, ...). Indirect identifiers may be strong or weak.



Direct identifiers

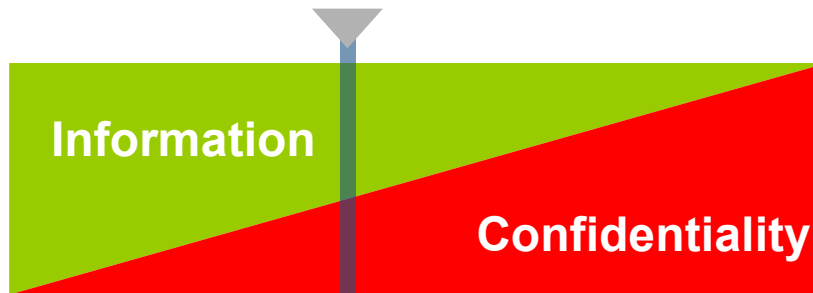
Unique combination of indirect identifiers

Rare combination of indirect identifiers

Soc. Sec.Nr	Gender	Age class	Town	Education	Profession	Income
1927384123	Female	40-55	Zurich (big town)	Higher	Civil Servant	80'000
1927384124	Male	30-40	Pully (small town)	Middle	Fisherman	50'000
1927384125	Male	55+	Vers-chez-les-Blanc (small town)	Higher	Politician	250'000
1927384126	Male	20-30	Yverdon (medium size town)	Lower	Plumber	70'000
1927384127	Female	55+	Lutry (small town)	Higher	Surgeon	150'000
1927384128	Male	30-40	Aubonne (small town)	Higher	IT consultant	80'000
1927384129	Male	55+	Zurich (big town)	Unknown	Surgeon	160'000
1927384130	Female	20-30	Cugy (village)	Middle	Violen maker	60'000
1927384131	Female	30-40	Neuchatel (medium size town)	Lower	House cleaner	55'000
...
...

How much anonymisation is enough?

Risk/utility balance



Some key steps:

- Remove direct identifiers (e.g., names, address, institution, photo).
- If necessary, limit the number of identifying variables.
- Reduce the precision / detail of a variable through aggregation and/or rounding (e.g., birth date, educational categories, replace a value with a less precise value).
- Restrict upper and/or lower ranges of a variable to hide outliers (e.g., income, age) (winsorization).
- Combine variables (e.g., create a non-disclosive rural/urban variable from place variables).
- Alter values (substitute values, suppress values).
- Generalize meaning of detailed text variables (e.g., occupational expertise)
- Watch out for open text (replace names with pseudonyms, recode/remove potentially identifying information).

Resources

CESSDA expert tour guide

Consortium of European Social Science Data Archives



cessda eric

About - Consortium - Projects - Research Infrastructure - Contact



Home / Research Infrastructure / Training / Expert tour guide on Data Management

Expert tour guide on Data Management



About this expert tour guide

This tour guide by CESSDA ERIC (the Consortium of European Social Science Data Archives European Infrastructure Consortium) aims to put social scientists like yourself at the heart of making their research data findable, understandable, sustainably accessible and reusable.

You will be guided by European experts who are - on a daily basis - busy ensuring long-term access to valuable social science datasets, available for discovery and reuse at one of the 17 CESSDA social science data archives. With this guide and the training events being held across Europe, we want to accompany and inspire you in your journey through the research data lifecycle.

Expert tour guide on Data Management

1. Plan
2. Organise & Document
3. Process
4. Store
5. Protect
6. Archive & Publish

Search this guide

Search

Target audience and mission

This tour guide was written for social science researchers who are in an early stage of practising research data management. With this tour guide, CESSDA wants to contribute to increased professionalism in data management and to improving the value of research data.

<https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide>

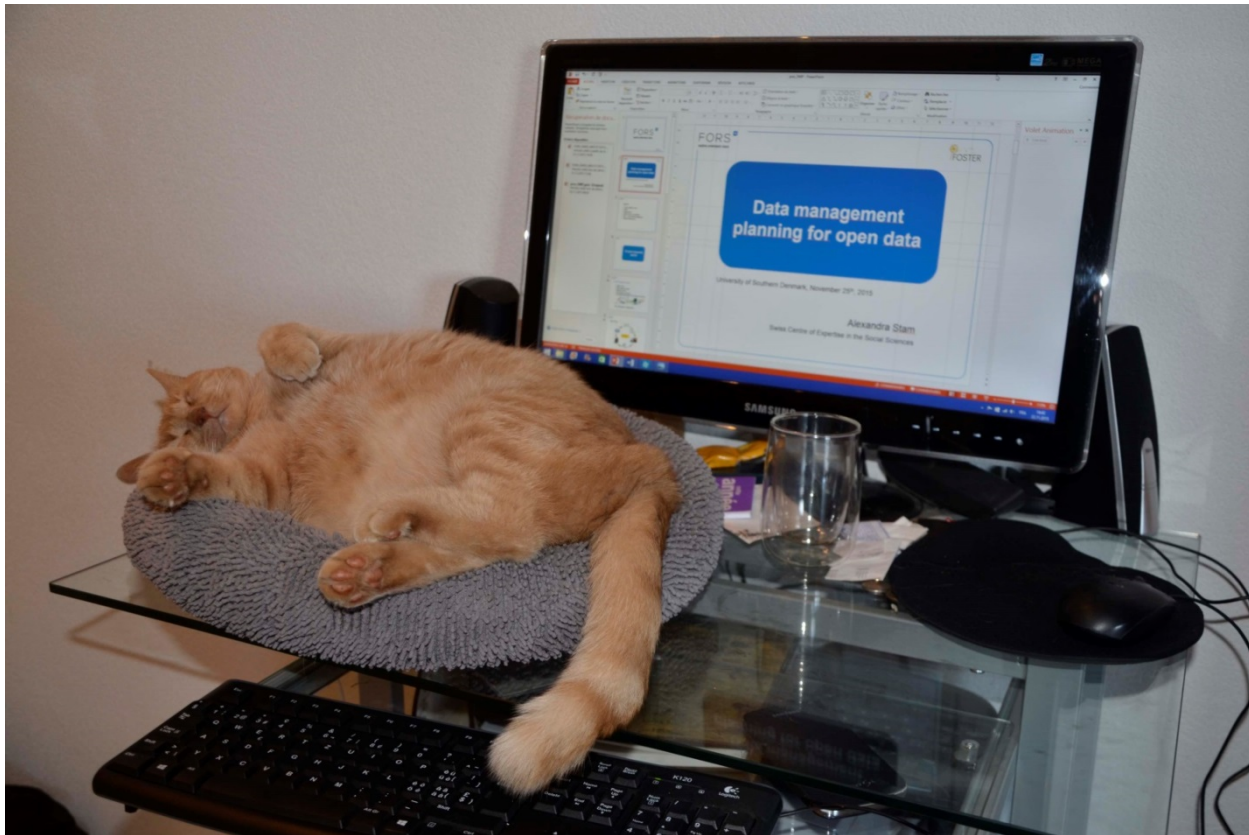
FORS guides:

FORS⁺ GUIDES
to survey methods
and data management



<https://forscenter.ch/publications/fors-guides/>

FORS⁺



Thanks for your attention

alexandra.stam@fors.unil.ch

FORS 

explore.understand.share.