TRAINING WORKSHOP on DMEG

RDM (Research Data Management)

DMEG (Data Management Expert Guide)

DMP (Data Management Plan)

FAIR Principles

Patrícia Miranda
Pedro Moura Ferreira

Outubro de 2019

Data Management Expert Guide

DMEG is open and freely available via CESSDA.eu/DMGuide

- Self-study for researchers (15-20 hours of content)
- Basis for interactive blended training by trainers or data stewards in (social sciences) research institutes to provide workshops on data management
- Train-the-trainer package

CESSDA and the DMEG

DMEG - FAIR Principles

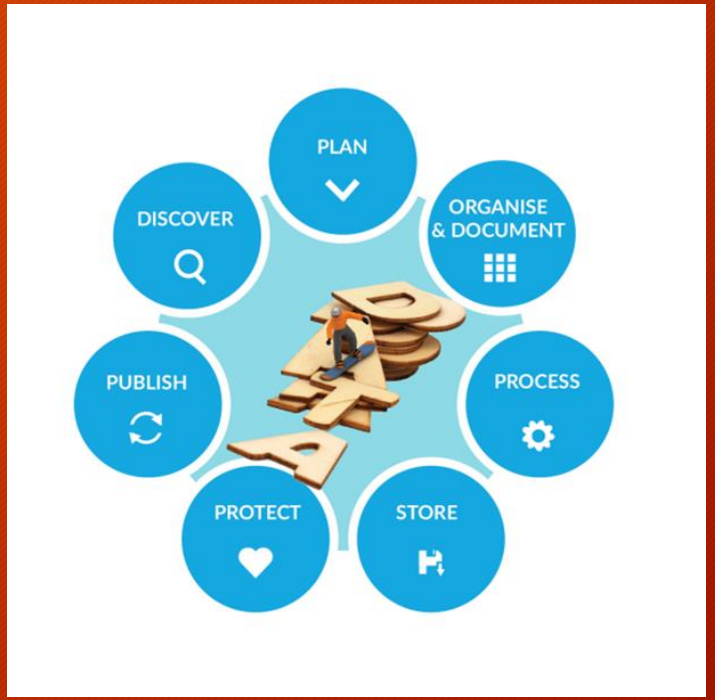
Data Management Expert Guide

This guide is designed by European experts to help social science researchers make their research data Findable, Accessible, Interoperable and Reusable (**FAIR**).

You will be guided by different 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 CESSDA social science data archives.

Data Life Cycle

If you follow the guide, you will travel through the research data lifecycle from planning, organising, documenting, processing, storing and protecting your data to sharing and publishing them.



Plan

CESSDA Expert Tour Guide

Gry Henriksen

14.05.19 Athens

www.cessda.eu\DMEG

cessda.eu @CESSDA_Data

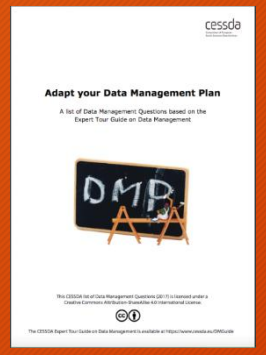
Cite as: Henriksen (2019), Jakobsson (2018). The CESSDA Expert Tour Guide – Plan RDM [presentation]. Athens TTT workshop, Bergen: CESSDA ERIC.

Data Management Plan

- Important tool – to handle, organize, structure and store research data
- It can be a formal document that outlines how to handle data during and after a project
- It is designed in accordance with the specific project

DMP checklist

- Adapt your DMP section at the end of every chapter
- Corresponding questions to each chapter
- The DMP checklist is downloadable


cessda

Adapt your Data Management Plan

A list of Data Management Questions based on the Expert Tour Guide on Data Management


Overview

Title of the project

Date of this plan

Description of the project

Origin of Data

Principal researchers

Collaborating researchers (if applicable)

Funder (if applicable)

Data producer

Project data contact

Data owner(s)

Roles

Costs

Adapt your DMP: Part 1

The Data Management Plan (DMP) is an important tool to structure the research data management of your project. After working on each chapter you should be able to answer part of the questions which make up a DMP.

This is the first of seven 'Adapt your DMP' sections in this tour guide. When you have finished the chapter on data management planning, you can start filling in the 'Overview of your research project' section. Below you can see what elements and corresponding questions are generally included in that section. You can select appropriate questions and answer them to adapt your own DMP.

For easy reference, we have put together a list of DMP-questions for all chapters in this tour guide. You can view and download the checklist as pdf (CESSDA, 2018a) or editable form (CESSDA, 2018b), and keep them as a reference while you are studying the contents of this guide.

- ⊕ Title of the project
- ⊕ Date and version of this plan
- ⊕ Description of the project
- ⊕ Origin of the data
- ⊕ Principal and collaborating researchers
- ⊕ Funder (if applicable)
- ⊕ Data producer
- ⊕ Project data contact
- ⊕ Data owner(s)
- ⊕ Roles
- ⊕ Costs

cessda

Organise and Document

CESSDA Expert Tour Guide

Gry Henriksen, NSD

14.05.19 Athens

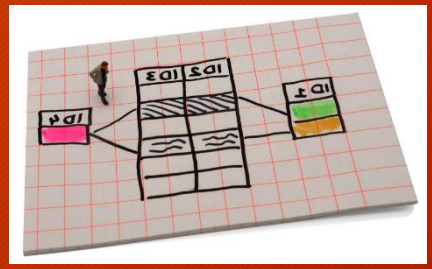
www.cessda.eu\DMEG

cessda.eu @CESSDA_Data

Cite as: Henriksen (2019), Krejci, J., J. Chylikova (2018). The CESSDA Expert Tour Guide. Chapter 2. Organise & Document [presentation]. Athens TTT workshop, Bergen: CESSDA ERIC.

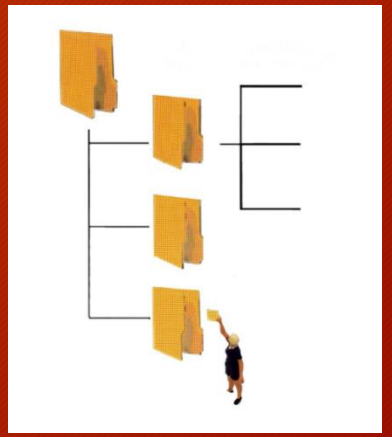
Designing a data file structure

Huge impact – the way data is processed and analysed



Qualitative data:

- Coding
- File naming
- Folder structure



Quantitative data, survey data:

Flat (rectangular) data files, hierarchical files, relational database

- Standard coding
- File naming
- Folder structure

Survey data

Variable names and labels:

- Contribute to structuring
- Allow researchers to integrate documentation into the data file
- Make the data file more FAIR compliant
- Follow basic rules and have a strategy

Documentation and Metadata

What is metadata?

- Information about your data, everything that can make data reusable – for you and others
- Interview guide, questionnaire (survey), instrument information, etc

Why document data?

- It makes your data Findable, Accessible, Interoperable and Reusable = **FAIR**
- It will give you as a researcher more relevance in your research field (and beyond)
- It will be easier to share your data
- Good metadata documentation gives your data added value

The FAIR Data Principles

- The **FAIR Data Principles** apply to metadata, data, and supporting infrastructure (e.g., search engines). Most of the requirements for findability and accessibility can be achieved at the metadata level. Interoperability and reuse require more efforts at the data level.

(in https://www.go-fair.org/fair-principles/)

In 2016, the '**FAIR Guiding Principles for scientific data management and stewardship'** were published in *Scientific Data*. The authors intended to provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

Process

CESSDA Expert Tour Guide

Gry Henriksen, NSD

14.05.19 Athens

www.cessda.eu\DMEG

cessda.eu @CESSDA_Data

Cite as: Henriksen (2019), Chylikova, J., J. Krejci (2018). The CESSDA Expert Tour Guide. Chapter 3.Process [presentation]. Athens TTT workshop, Bergen: CESSDA ERIC.

What is “Process”?

It includes all data operations needed to prepare your data files for analysis and data sharing

- Both by you and others in the project, and in the data archive.

Crucial to:

- Maintain the authenticity of data;
- Prevent loss or deterioration.

Data entry

Quantitative data:

- Automation – prevents some mistakes, produces others
- Coding
- Anonymize
- File format for long time preservation

Qualitative data:

- Transcription – record and transcribe
- Coding
- Anonymize
- File format for long time preservation

Data integrity

- Assurance of the accuracy, consistency and completeness of original information contained in the data.
- Integrity of a data file is based on its structure and on links between data and integrated elements of documentation.

Data entry and integrity

> *Data integrity means assurance of the accuracy, consistency, and completeness of original information contained in the data. At the same time, the authenticity of the original research information has to be preserved (see 'Data authenticity').*

The integrity of a data file is based on its structure and on links between data and integrated elements of documentation. From the moment that data is being entered, data integrity is at stake.

Data authenticity

- Preserve original research information/file
- Versions control
- Use best practice rules

Best practices for quality assurance, version control and authenticity

Version and edition management will help to:

1. Clearly distinguish between individual versions and editions and keep track of their differences;
2. Prevent unauthorised modification of files and loss of information, thereby preserving data authenticity.

Best practices

The best practice rules (UK Data Service, 2017a; Krejčí, 2014) may be summarised as follows:

- Establish the terms and conditions of data use and make them known to team members and other users;
- Create a 'master file' and take measures to preserve its authenticity, i.e. place it in an adequate location and define access rights and responsibilities – who is authorised to make what kind of changes;
- Distinguish between versions shared by researchers and working versions of individuals;
- Decide how many versions of a file to keep, which versions to keep (e.g. major versions rather than minor versions (keep version 02-00 but not 02-01)), for how long and how to organise versions;
- Introduce clear and systematic naming of data file versions and editions;
- Record relationships between items where needed, for example between code and the data file it is run against, between data file and related documentation or metadata or between multiple files;
- Document which changes were made in any version;
- Keep original versions of data files, or keep documentation that allows the reconstruction of original files;
- Track the location of files if they are stored in a variety of locations;
- Regularly synchronise files in different locations, such as using MS SyncToy (2016).

Data quality

Best Practices for Survey Research

- "The quality of a survey is best judged not by its size, scope, or prominence, but by how much attention is given to [preventing, measuring and] dealing with the many important problems that can arise." (American Association for Public Opinion Research (2015) (AAPOR) (CESSDA expert guide – ch3)

Store

CESSDA Expert Tour Guide

Gry Henriksen, NSD

14.05.19 Athens

www.cessda.eu\DMEG

cessda.eu @CESSDA_Data

Cites as: Henriksen (2019), Netscher, Sebastian (2018): RDM Expert Tour Guide: Chapter 4 Store [presentation]. Athens TTT workshop, 2019,Train the Trainers Workshop, 12-13 April 2018, Ljubljana (Slovenia).

Storage?

- Definition: “the action or method of storing something for future use”
- Data storage: “the retention of retrievable data on a computer or other electronic system” (Oxford Dictionary)
- Storage does not mean to simply push the storage button, i.e. to put something, somehow, somewhere for future use
- Storage is a systematic task

Why is it important?

- various storage solutions
- storage strategy:
  - what is stored and how
  - backup and disaster recovery
  - protect against unauthorized (mis-)use
- part of a systematic data management plan
- closely connected to other RDM activities, e.g.
  - organization (and documentation)
  - data protection
  - publication (and long-term preservation)
- requires systematic planning (early task)

❖ *Plan, organise and document, process and store, protect and publish*

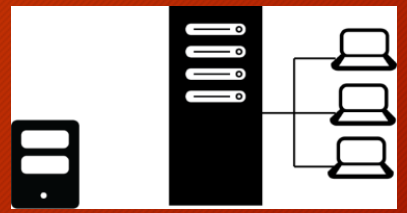
Storage strategy

- A Storage strategy contains:
  - storage solutions and media
  - backup strategy and disaster recovery plan
  - data protection
- …and it should be systematically implemented in a data management plan

Storage solutions:

Different solutions:

- Local storage
- Cloud storage
- Portable devices
- Network drives



Storage and media

- Optical (CD, DVD)
- Magnetic (hard drive)
- Portable flash drive (memory stick)
- Build in flash drive (hard drive)

Recommendations: *use at least two types of storage media, replace storage media (after 2-5 years) and carry out integrity checks, e.g. by checksum tool*

Backup

Various reasons for data loss:

- hardware failure
- software malfunction
- malware or hacking
- human error
- theft, natural disaster or fire
- degradation of storage media
- etc.

Developing a backup strategy

1. Institutional backup strategy
   - How does it work?
2. What has to be back upped
   - What needs to be copied?
3. When is it back upped and how often
   - Frequency and number of copies?
4. Where is it back upped
   - Storage solutions for copies?
5. Storage capacity needed for backups
   - Memory capacity needed for copies?
6. Tools that can be used to automate backups
   - Automate backup processes (e.g. cloud services?)
7. How long is it back upped and how will it be destructed
   - Storage period and destruction of irrelevant copies?
8. How will personal data be protected
   - Data protection strategy for copies?
9. Disaster recovery plan
   - How to access and (re-)use copies
10. Responsibilities
    - Who is responsible for backups?

Data Protection and Data Security

The Worst 10 Passwords of 2017

1. *123456*
2. *password*
3. *12345678*
4. *qwerty*
5. *12345*
6. *123456789*
7. *letmein*
8. *1234567*
9. *football*
10. *iloveyou*

Time Magazine (2018): The Worst 25 Passwords of 2017. Available at: http://tim.com/5071176/worst-passwords-2017/

- Passwords
- Encryption

Protect

CESSDA Expert Tour Guide

Gry Henriksen, NSD

14.05.19 Athens

www.cessda.eu\DMEG

cessda.eu @CESSDA_Data

Cita as: Henriksen (2019), Bezjak (2018) Athens TTT workshop 2019, Train the Trainers event April 2018, Ljubljana, Slovenia

Ethics and data protection

- Ethics and data protection
- Ethical review process
- Processing personal data
- Informed consent
- Anonymisation
- Copyright

Página no ICS relativa à Proteção de Dados:

https://www.ics.ulisboa.pt/info/protecao-de-dados

Ethics and data protection

When collecting, using and sharing research data, ethical considerations and legal obligations guide the way.

Ethics are an integral part of a research project, from the conceptual stage of the research proposal to the end of a research project. Within the EU the RESPECT project has drawn up professional and ethical guidelines (Institute for Employment Studies, 2004) for conducting socio-economic research. The RESPECT Code of Practice is based on three main guidelines:

⊕ 1. Upholding scientific standards

⊕ 2. Compliance with the law

⊕ 3. Avoidance of social and personal harm

Depending on the type of data you collect you will have to deal with different laws. Whereas Intellectual Property legislation applies to all data, the collection of personal data has its own laws to adhere to. Importantly, since 25 May 2018, the General Data Protection Regulation (GDPR; European Union, 2016a) applies to any EU researcher or researcher in the European Economic Area (EEA) who collects personal data about a citizen of any country, anywhere in the world.

Copyright

- Intellectual property right
- Diversity in copyright
- Licenses / re-use rights - Licensing your data (*in Archive & Publish*)

Copyright is an internationally recognised form of intellectual property right, which arises automatically as a result of original work such as research. It does not need to be registered to apply to a piece of work.

Copyrighted output from research could include spreadsheets (and other forms of originally selected and organised data), publications, reports and computer programs. Copyright will not cover the underlying facts, ideas or concepts, but only the particular way in which they have been expressed. The right will lie with the author of the work, or with their relevant institution—different universities will have different policies on intellectual property.

A copyrighted work cannot usually be published, reproduced, adapted or translated without the owner's permission.

Key copyright considerations for researchers

Whether you want to reuse someone else's data or if you are planning to archive and share your own, you should ask yourself who the copyright holder of the datasets is (also see 'Licensing your data'). Are you allowed to use them and in what way? Are you allowed to archive and publish them in a data repository? How do you answer the question who the copyright holder of a dataset is? Is it you, your employer, the data archive, fellow researchers? The answer depends on multiple factors, such as who had input into creating the research data, whether data were used from other datasets, and what the researcher's contract of employment stipulates.

Archive and Publish

CESSDA Expert Tour Guide

Gry Henriksen, NSD

14.05.19 Athens

www.cessda.eu\DMEG

cessda.eu @CESSDA_Data

Cita as: Henriksen (2019), Bezjak (2018) Athens TTT workshop 2019, Train the Trainers event April 2018, Ljubljana, Slovenia

Why archive research data?

- *Archiving and publishing your data properly will enable both your future self as well as future others to get the most out of your data.*

Archiving

Archiving data

- store your data in a suitable file format, with adequate documentation and keep your data safe on long term
- make sure you can read and access the data later on
- allow access to others for verification purposes

Publishing data

- Publicly disclose your research data:
  - To make the data findable;
  - To make them accessible, at least metadata;
  - To make them reusable;
  - To be in compliance with the FAIR principles.

What’s in it for them

- **Career benefits**

  increased visibility, reuse and citation and therefore recognition of scholarly work

- **Scientific progress**

  enabling new collaborations, new data uses and links to the next generation of researchers

- **Norms**

  openness of research data is at the heart of scientific ethics

- **External drivers**

  funders and publishers requirements

CESSDA archives

- (Trusted) domain specific data repositories

  For high-quality data with a potential for reuse, we recommend you to assure long-term access by publishing them with a trusted repository, like many of the CESSDA archives.

- Advantage of having expert help within research

  It helps you to increase the comprehensibility, visibility, findability…

Discover

CESSDA Expert Tour Guide

Gry Henriksen, NSD

14.05.19 Athens

www.cessda.eu\DMEG

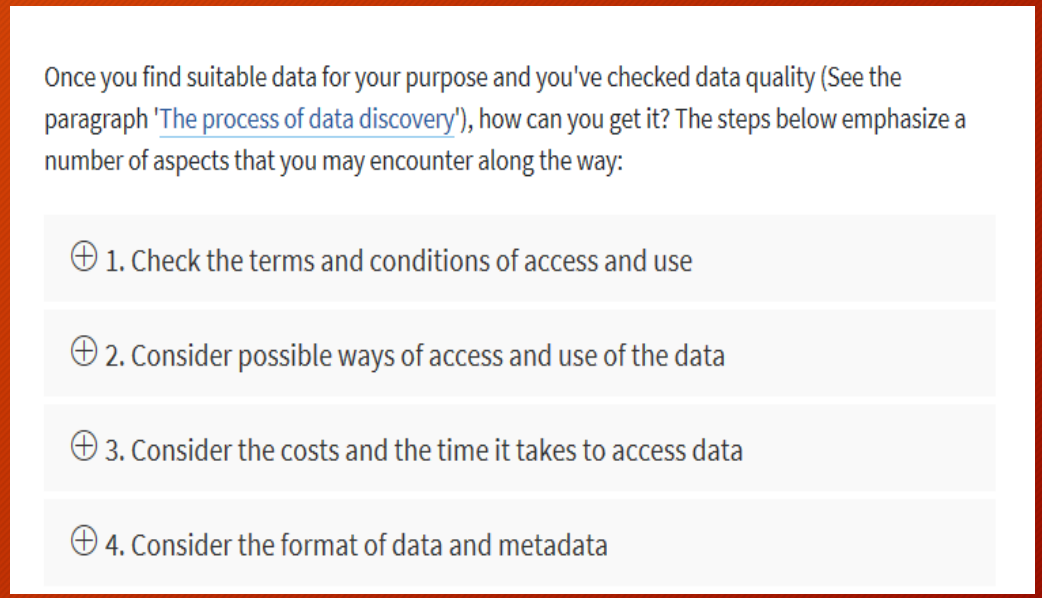
cessda.eu @CESSDA_Data

Cita as: Henriksen (2019), Bezjak (2018) Athens TTT workshop 2019, Train the Trainers event April 2018, Ljubljana, Slovenia

The process of data discovery

When you want to reuse or review research data shared by other researchers...

- Data repositories as data resources
  - International surveys
  - Other curated data sources
- Access, (re)use and cite data



PID - persistent identifier

Citing data

After you have used research data you may want to publish about the work you have done. In this case, you should always cite research data. Research data may be subject to intellectual property rights. However, citing data is usually included in the terms and conditions for the use of data. The obligation to properly acknowledge any research work, including the work invested into development of databases, also logically follows from research ethics.

Expert tip: Use a persistent identifier
In citation always use persistent identifiers (DOI – Digital Object Identifier) if available. It promotes findability and accessibility of data.

As social sciences tend to be more and more data intensive, data repositories must facilitate several ways of identifying and locating data. This development poses complex technical and organisational challenges to data providers. Persistent identification is becoming a prerequisite for sustained and reliable resource discovery and reuse. The use of PID is an important feature of a certified and trustworthy data repository. PID support access to data as well as referencing and citing data. They are an advertisement for data integrity – ultimately PID are part of the proof that an object which a repository has responsibility has not changed. Additionally, the use of PID helps data repositories to be compliant with the FAIR principles (Findable - Accessible - Interoperable - Reusable)[1] set by FORCE 11[2] and provides a future-proof plan in case of the relocation of its holdings.[3]

- PID - FAIR - metadata - DMEG - *connection between all phases of Data Life Cycle*

cessda

QUESTÕES
?

Em suma...

- O Plano de Gestão de Dados é uma ferramenta essencial para estruturar e organizar os dados dos projetos de investigação. Neste workshop focamo-nos nos materiais do CESSDA, mas há outros modelos de DMP, nomeadamente o difundido pela RDA; exemplo: Data Management Planning (DMP) themes, por DCC e UC3; vários *templates* disponíveis no site DMPTool: https://dmptool.org/public_templates.

- Relativamente à proteção de dados e ao **GDPR**, os investigadores principais (Principal Investigators – PIs) e/ou os responsáveis pelos dados (Data Controllers) devem ter em linha de conta as recomendações europeias em relação à *anonimização dos dados* e aos dados sensíveis (orientação política, orientação sexual, atos criminais, condições de saúde, entre outros).

- Adicionalmente, podemos referir que tornar os dados **FAIR** é cada vez mais uma responsabilidade conjunta dos investigadores e dos repositórios. Neste contexto, por forma a otimizar o ciclo de vida dos dados, incluindo o arquivo, a disseminação e a reutilização dos dados, o desenvolvimento de metadados ricos é um requerimento fundamental no âmbito dos princípios FAIR (Findable, Accessible, Interoperable and Reusable).

Para mais informação sobre os princípios FAIR, ver a *checklist* de Jones e Grootveld (2017).

Referências

- http://www.apis.ics.ulisboa.pt/
- https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide
- https://www.go-fair.org/fair-principles/
- https://fairsfair.eu/
- Jones, S. & Grootveld, M. (2017, November). How FAIR are your data? Zenodo. http://doi.org/10.5281/zenodo.1065991
- Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018. doi: 10.1038/sdata.2016.18
- https://dmptool.org/

to be continued...

Obrigada!