



Responsible Open Science

Fundamentals of Research Data Management for Social Scientists

Francesca Morselli & Ricarda Braukmann, DANS

14 and 17 June 2021



House Rules

- Keep mic muted during the presentations
- Raise your hand if you want to speak or write questions in the chat
- Don't take screenshots without the permission of the participants
- Presentations are recorded | Breakout sessions are not

Veertly Platform

- Google Chrome gives best performance
- Set language Veertly to English
- Agenda
- Main stage
- Break-out rooms (2nd day)
- Chat



Slack Channel

The screenshot shows a Slack workspace named "ResponsibleOpenSci...". The left sidebar lists several channels: #1-plan, #2-organise (selected), #3-process, #4-store, #5-protect, #6-archive, #7-discover, #general, and #random. The main content area displays the channel header "#2-organise" with a star icon and "Add a topic". Below the header, a system message states: "This is the very beginning of the #2-organise channel. @Ricarda Braukmann created this channel on 10 May. Discuss questions related to the 'Organise & Document' Chapter of the Data Management Expert Guide (cessda.eu/dmeg) Edit description". Action buttons include "Add people", "Share channel PREMIUM", and "Forward emails to this channel". A date separator indicates "Monday, 10 May". The message history shows three entries by Ricarda Braukmann: joining the channel at 11:12, setting the channel description to "Ask questions related to the 'Organise & Document' Chapter of the Data Management Expert Guide (cessda.eu/dmeg)" at 11:12, and setting the channel description to "Discuss questions related to the 'Organise & Document' Chapter of the Data Management Expert Guide (cessda.eu/dmeg)" at 11:14. A fourth entry shows "AJ" joining the channel along with Francesca Morselli at 18:18. At the bottom, there is a text input field "Send a message to #2-organise" with a rich text editor toolbar and a send button.



Open Data Infrastructure for Social Sciences and Economic Innovations

www.odissei-data.nl



- Open Data Infrastructure for Social Science and Economic Innovations
- Dutch national infrastructure
- Coordinates data collections (i.e. surveys and panels) operated by members
- Supports researchers to access to CBS microdata
- Supports researchers to use the LISS*-panel
- Provides funding (e.g. microdata access grants)
- Super computing facilities (hosted by SURF)

*Longitudinal Internet studies for the Social Sciences

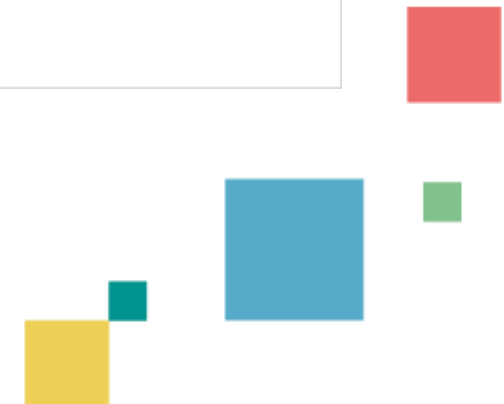
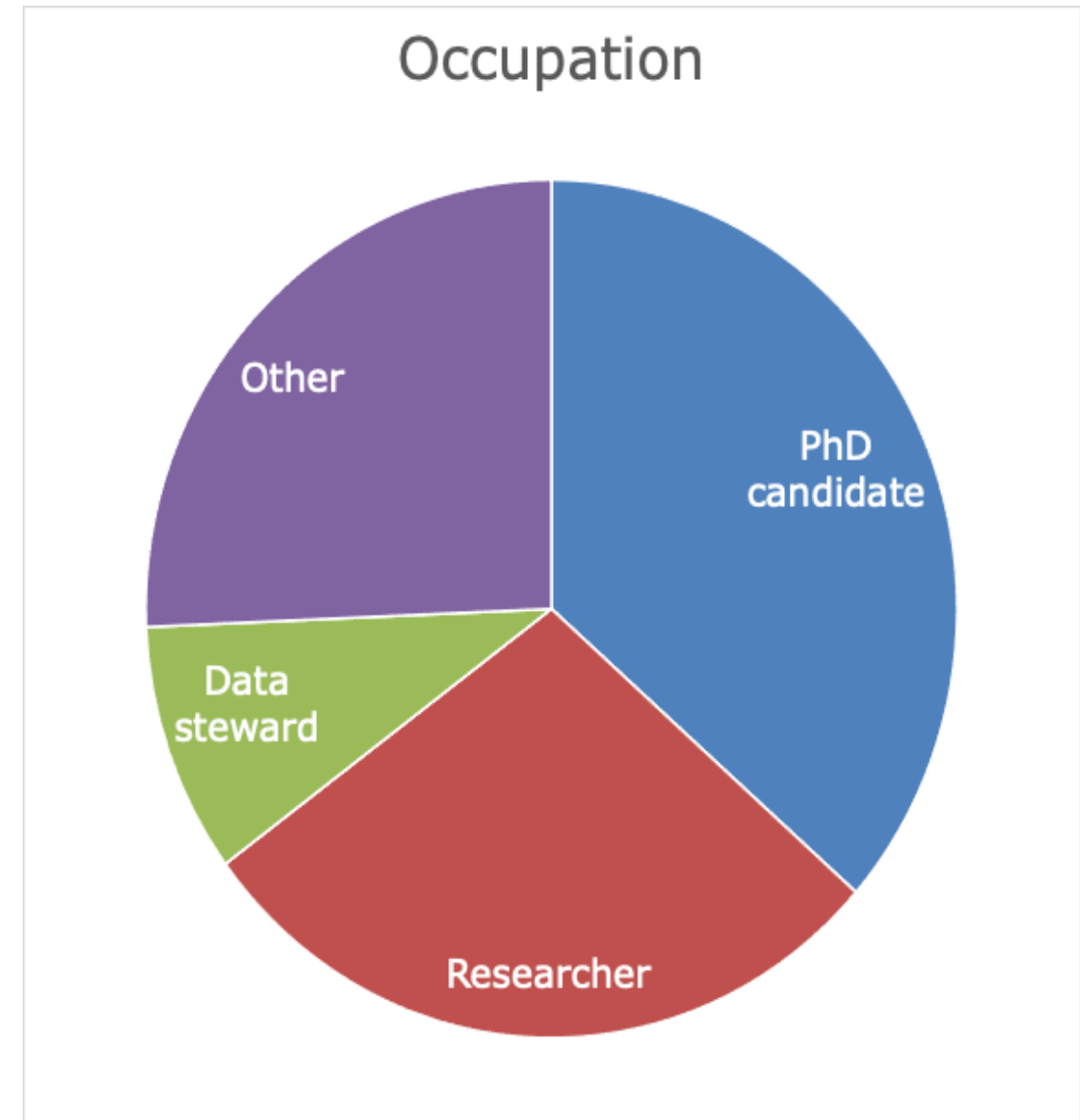
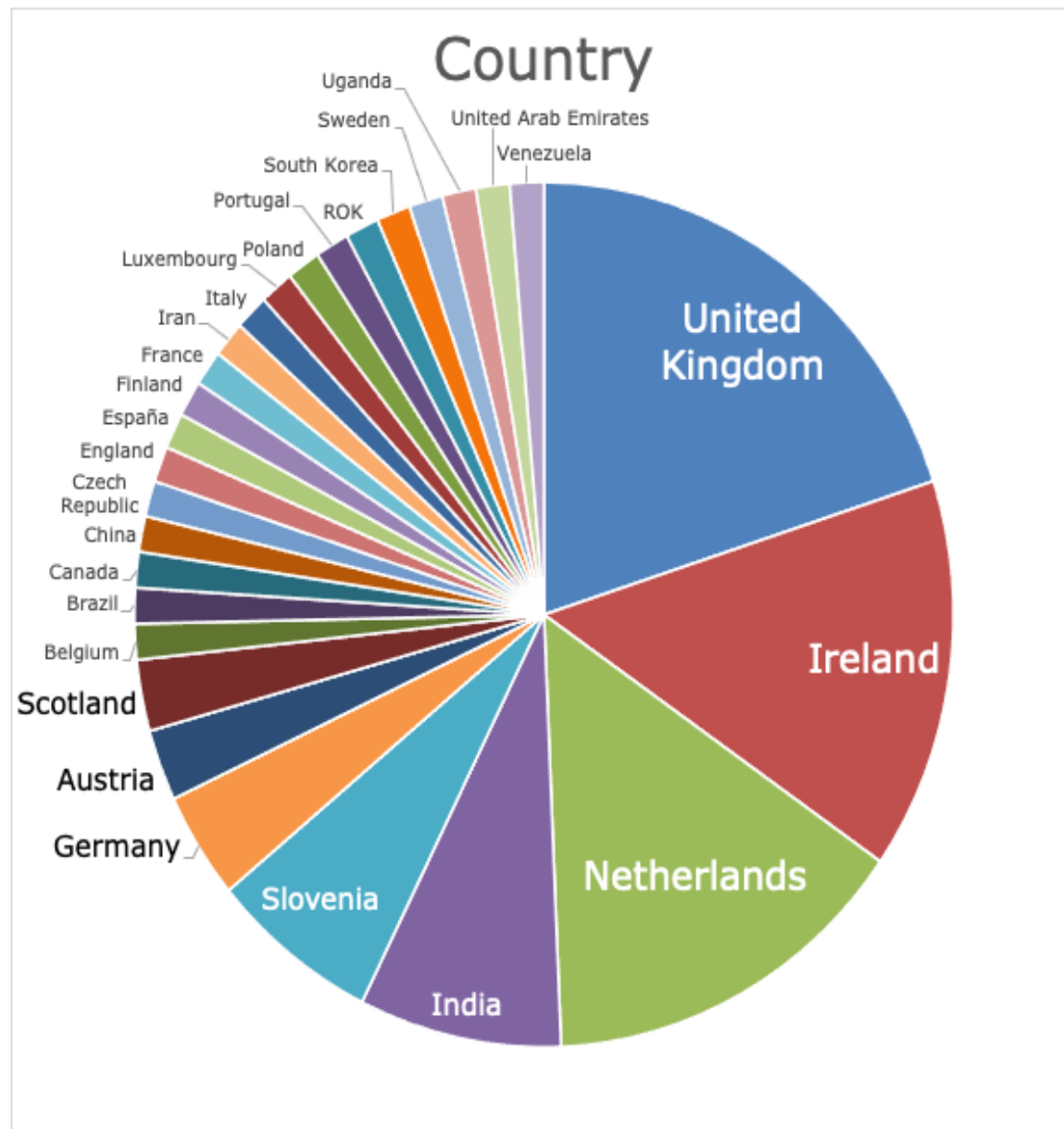


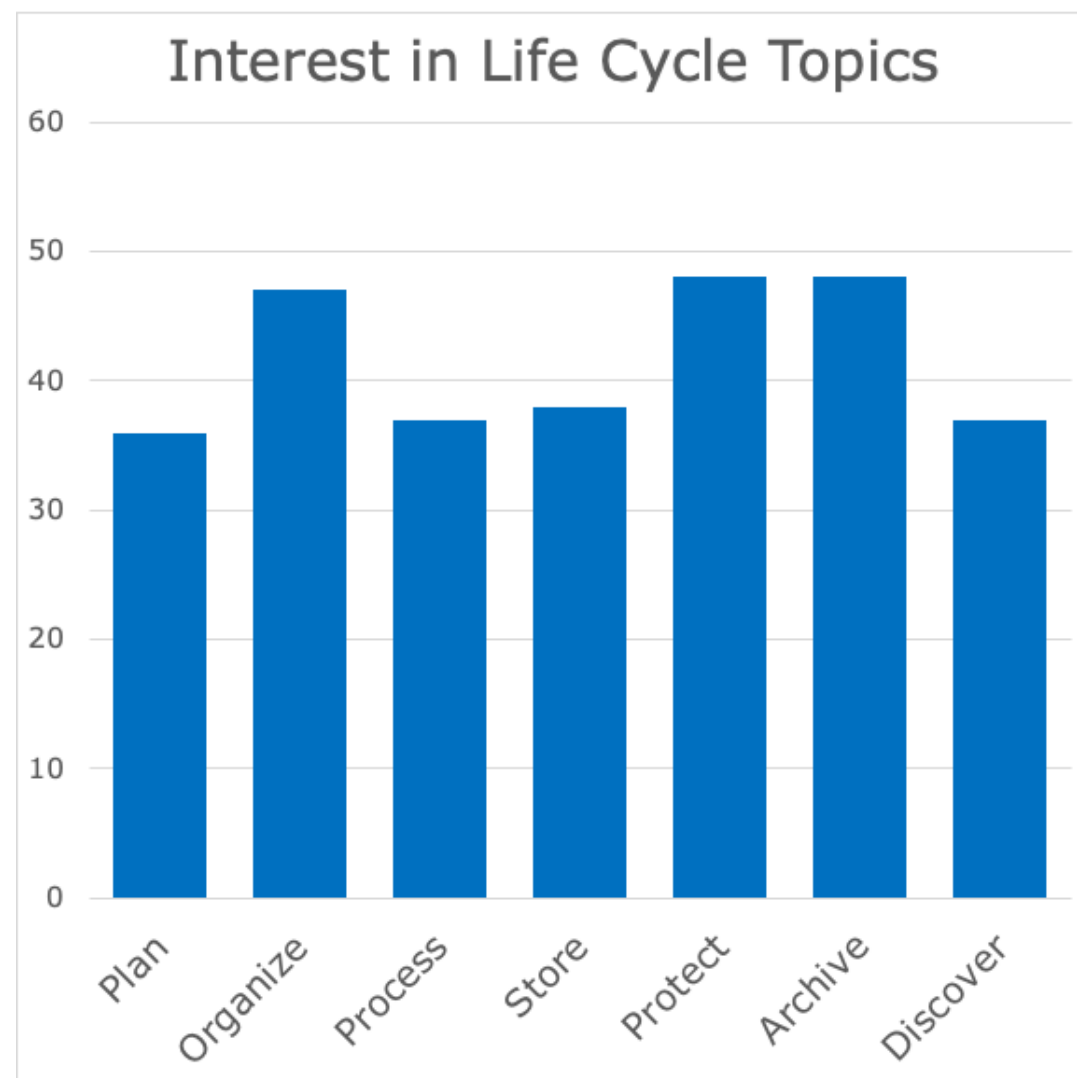
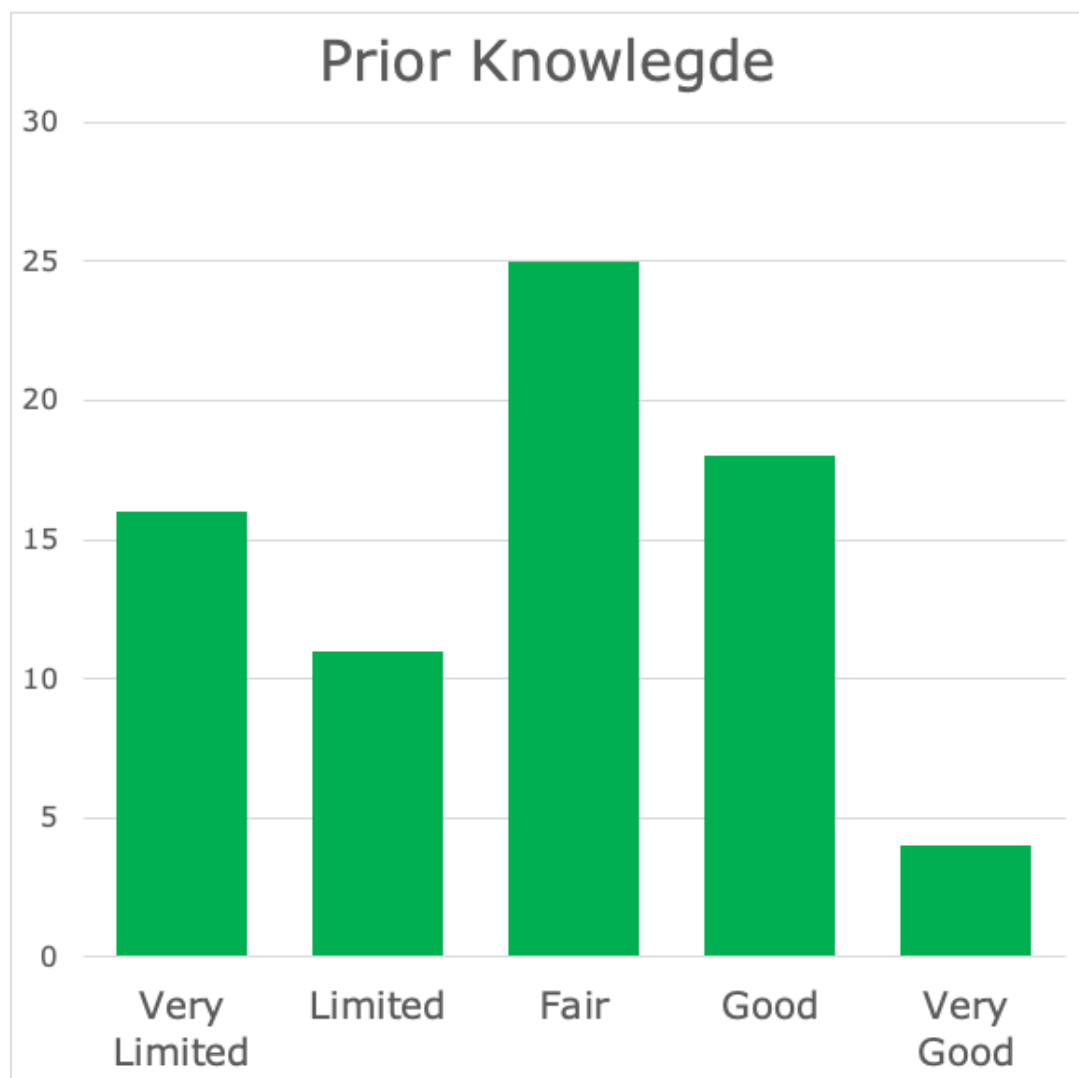


- Consortium European Social Science Data Archives
- European Research Infrastructure Consortium
 - DANS is the Dutch national service provider
- Providers data services to the social sciences
 - CESSDA Data Catalogue
 - CESSDA Training → Data Management Expert Guide

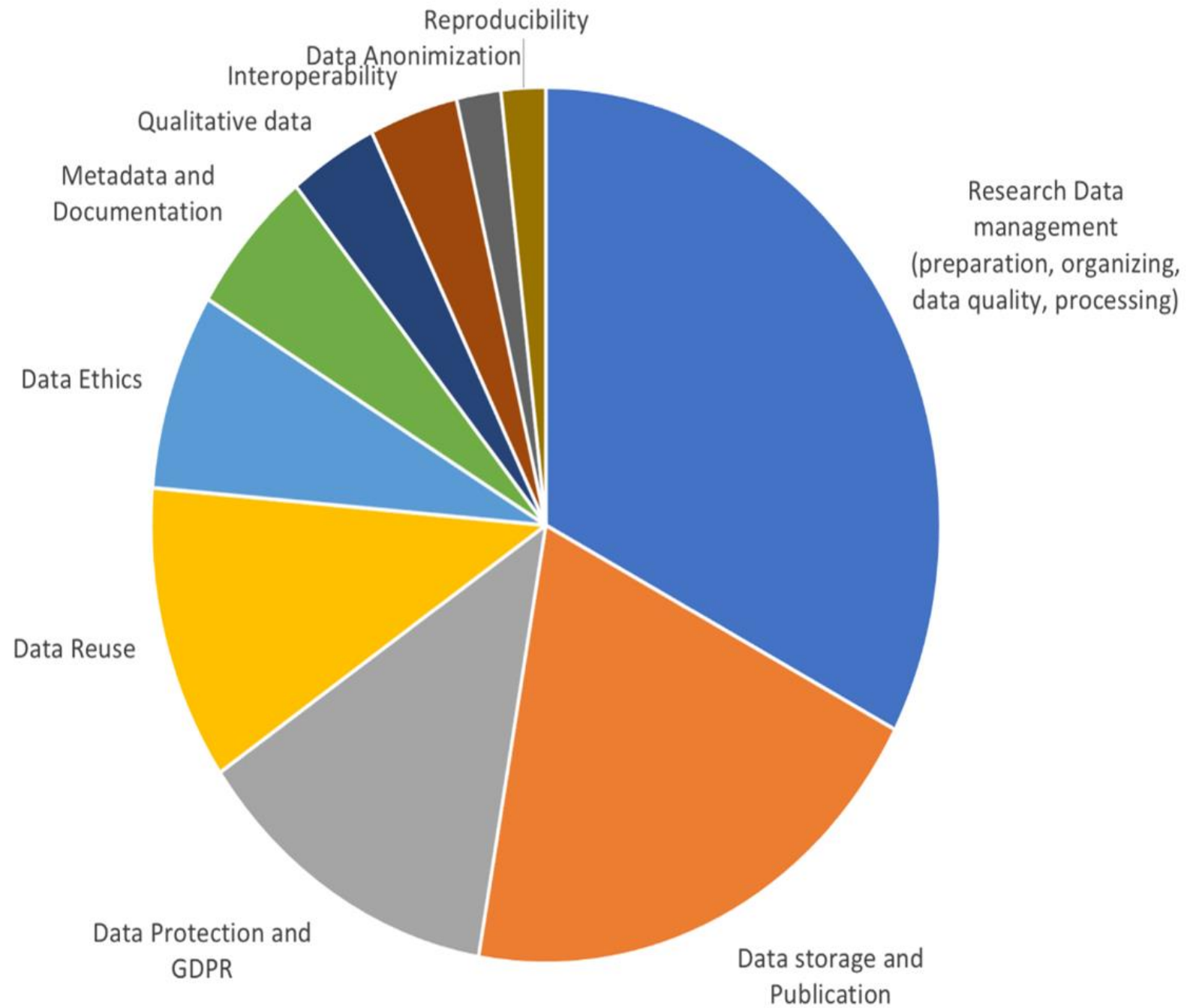


A bit about yourself

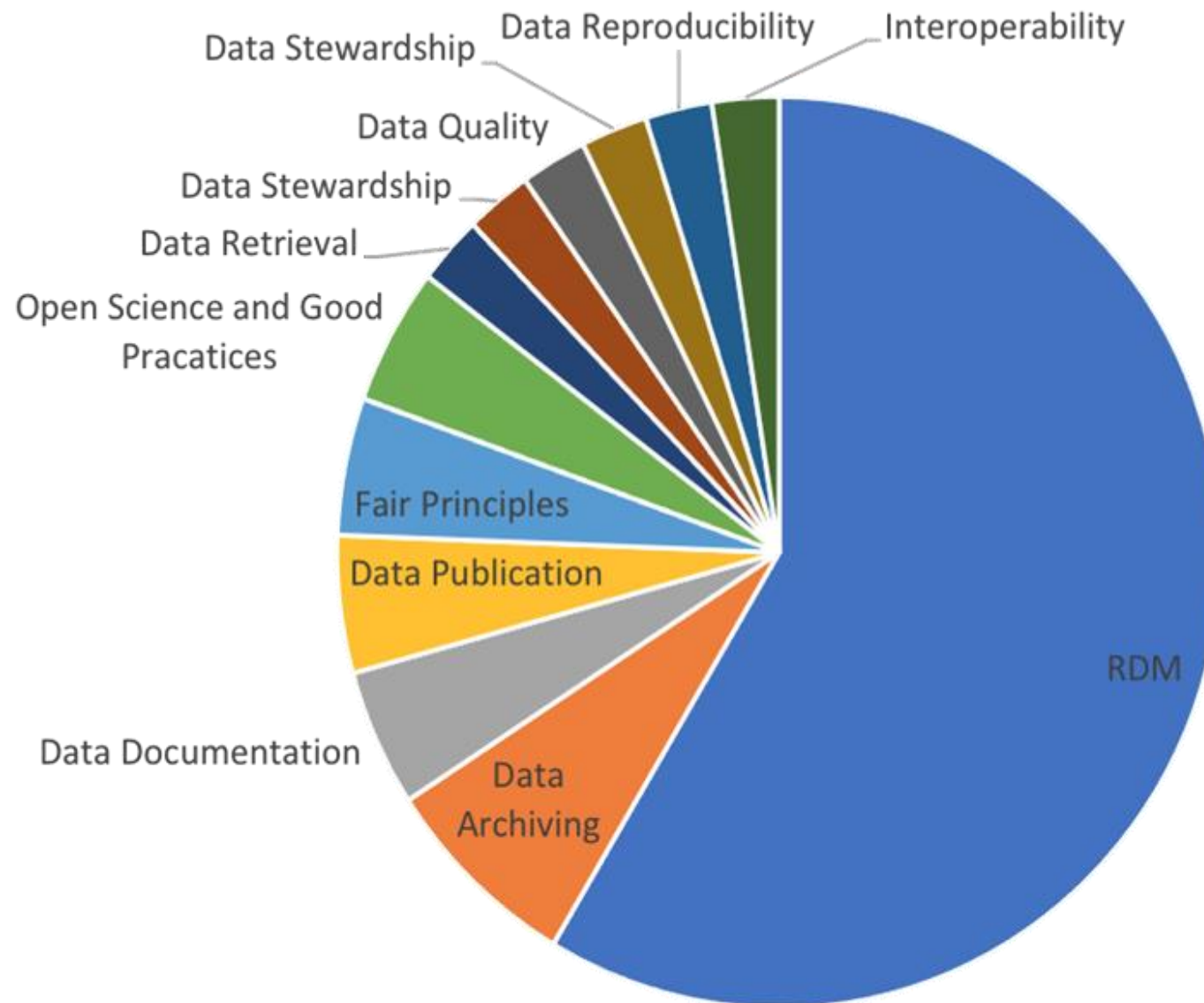




What are the challenges in RDM?



What would you like to learn?



What we won't cover

- Qualitative data (DMEG covers this aspect)
- The content of the data themselves (e.g. data completeness)



Workshop's programme

Monday 14th June 10:00-12:30 CEST

10:00 - 10:15 Welcome

10:15 - 11:00 An Introduction to Data Management: Organizing, processing and storing research data - Ricarda Braukmann (DANS)

11:00 - 11:15 Coffee break

11:15 - 12:00 Data protection and GDPR - Emilie Kraaijkamp & Ellen Leenarts (DANS)

12:00 - 12:15 Q&A

12:15 - 12:30 Introduction to the homework - Francesca Morselli (DANS)

In between, we will ask you to work on the [homework assignment](#).



Workshop's programme

Thursday 17th June 10:00-12:30 CEST

10:00 - 10:05 Welcome

10:05 - 10:35 Working with administrative data: an ODISSEI case study - Tom Emery & Kasia Karpinska (ODISSEI)

10:35 - 10:45 Coffee break

10:45 - 11:45 Group work in breakout sessions

11:45 - 12:15 An Introduction to Data Management: Archiving and discovering data - Ricarda Braukmann (DANS)

12:15 - 12:30 Closing



Introduction to Data Management

Organizing, processing and storing
data

Ricarda Braukmann

(DANS)
14th June 2021



Why Research Data Management?



Why Research Data Management?

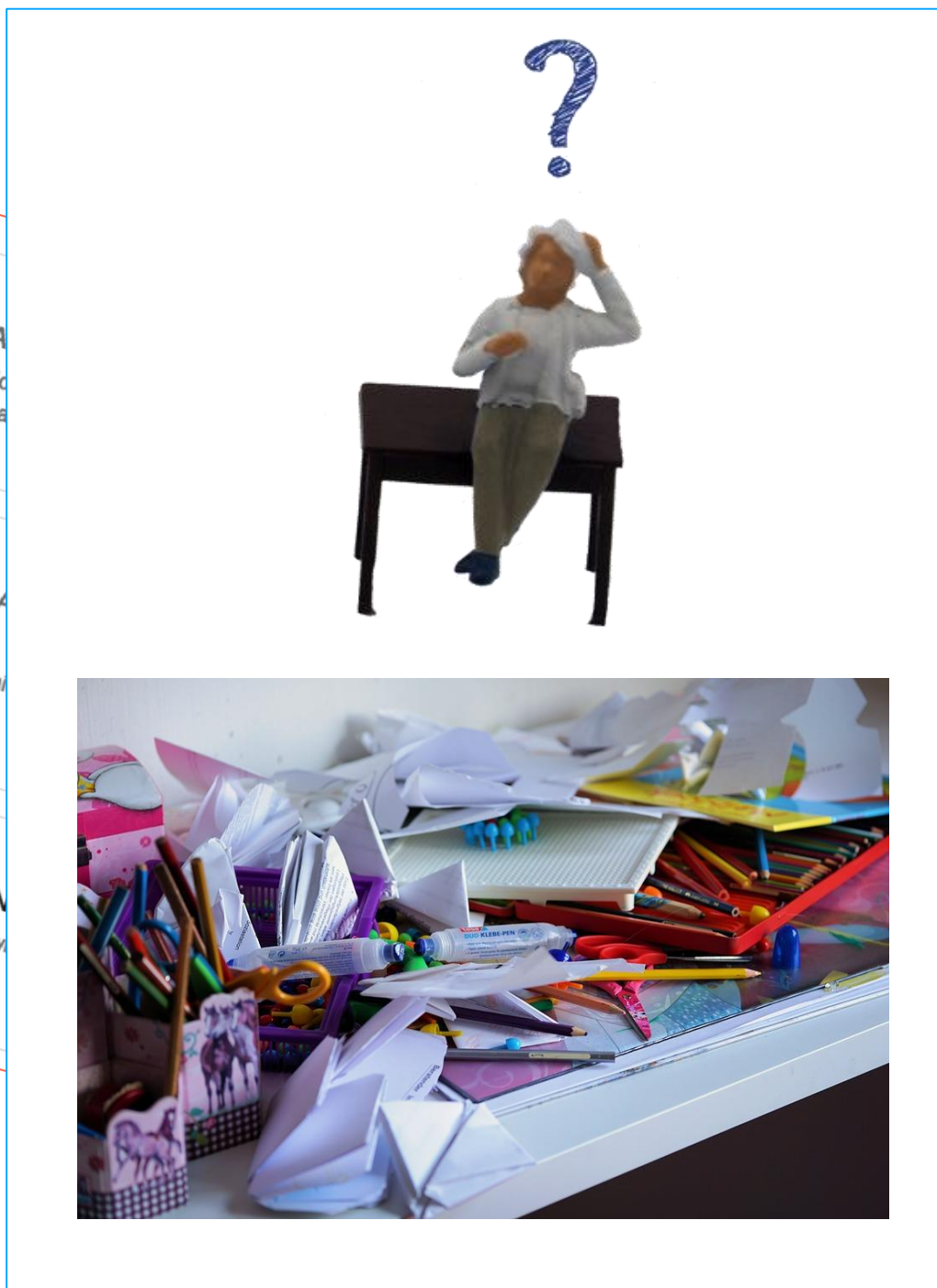
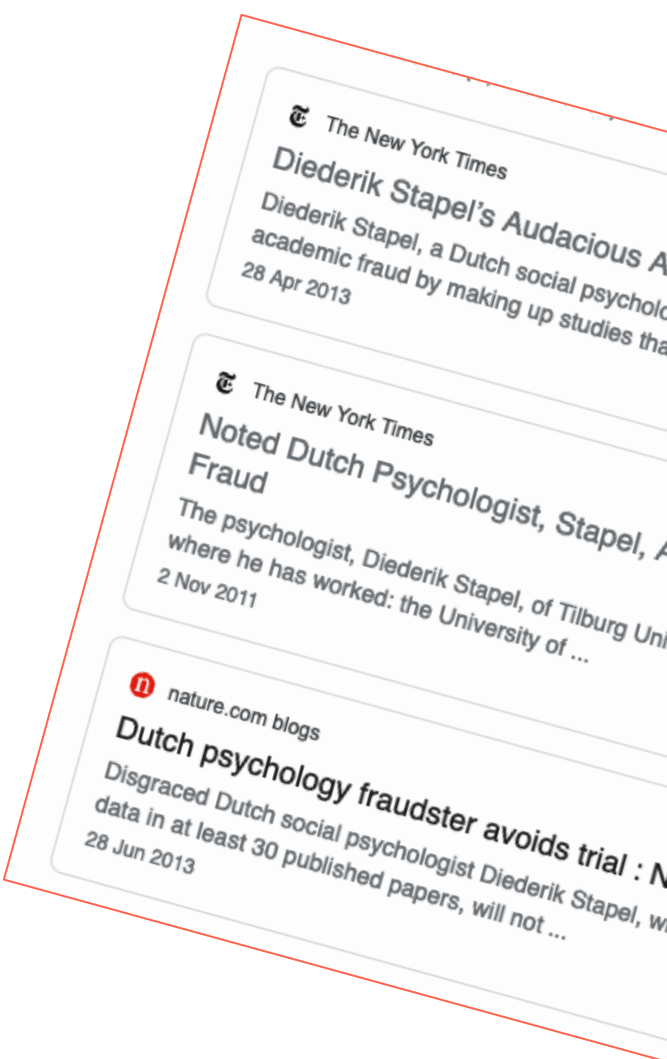
The New York Times
Diederik Stapel's Audacious Academic Fraud
Diederik Stapel, a Dutch social psychologist, perpetrated an audacious academic fraud by making up studies that told the world what it wanted ...
28 Apr 2013

The New York Times
Noted Dutch Psychologist, Stapel, Accused of Research Fraud
The psychologist, Diederik Stapel, of Tilburg University, committed ... by the three Dutch institutions where he has worked: the University of ...
2 Nov 2011

nature.com blogs
Dutch psychology fraudster avoids trial : News blog
Disgraced Dutch social psychologist Diederik Stapel, who in 2011 was found to have fabricated data in at least 30 published papers, will not ...
28 Jun 2013



Why Research Data Management?

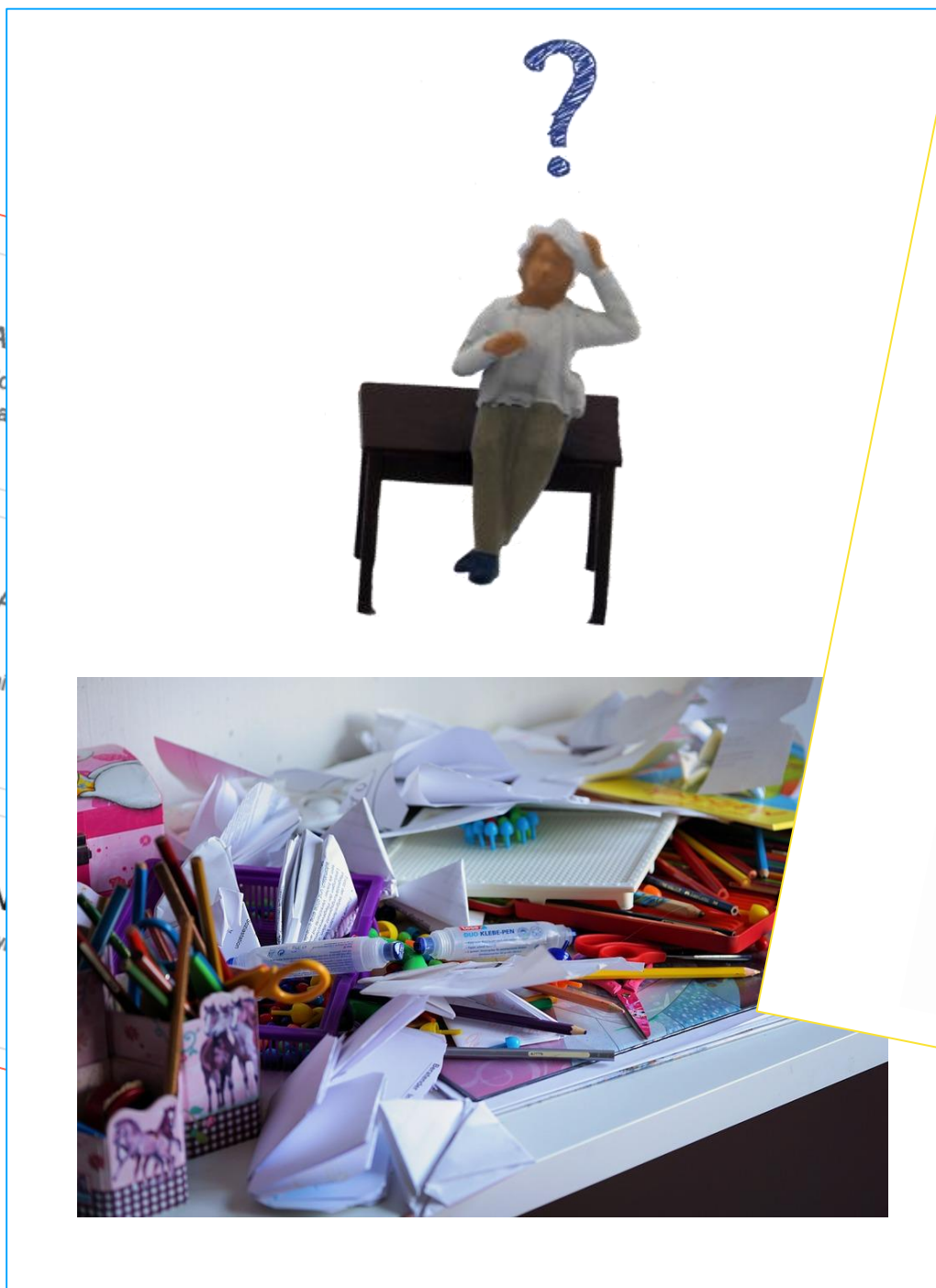


Why Research Data Management?

The New York Times
Diederik Stapel's Audacious Academic Fraud
Diederik Stapel, a Dutch social psychologist, is accused of academic fraud by making up studies that he conducted.
28 Apr 2013

The New York Times
Noted Dutch Psychologist, Stapel, Admits to Academic Fraud
The psychologist, Diederik Stapel, of Tilburg University, where he has worked: the University of ...
2 Nov 2011

nature.com blogs
Dutch psychology fraudster avoids trial : Nature
Disgraced Dutch social psychologist Diederik Stapel, who admitted to fabricating data in at least 30 published papers, will not ...
28 Jun 2013



Why Research Data Management?



Easily find and understand data

Increase impact

Make research reproducible

Increase reuse potential

Comply with funder mandates



PUBLICATIONS AND DATA



Why Research Data Management?



Findable

Accessible

Interoperable

Reusable



Why Research Data Management?



Findable

To aid automatic discovery of relevant datasets, (meta)data should be easy to find by both humans and machines and be assigned a persistent identifier.

Accessible

Interoperable

Reusable



Why Research Data Management?



Findable

To aid automatic discovery of relevant datasets, (meta)data should be easy to find by both humans and machines and be assigned a persistent identifier.

Accessible

Limitations on the use of data, and protocols for querying or copying data are made explicit for both humans and machines.

Interoperable

Reusable



Why Research Data Management?



Findable

To aid automatic discovery of relevant datasets, (meta)data should be easy to find by both humans and machines and be assigned a persistent identifier.

Accessible

Limitations on the use of data, and protocols for querying or copying data are made explicit for both humans and machines.

Interoperable

(Meta)data should use standardised terms (controlled vocabularies), have references to other (meta)data and be machine actionable.

Reusable



Why Research Data Management?



Findable

To aid automatic discovery of relevant datasets, (meta)data should be easy to find by both humans and machines and be assigned a persistent identifier.

Accessible

Limitations on the use of data, and protocols for querying or copying data are made explicit for both humans and machines.

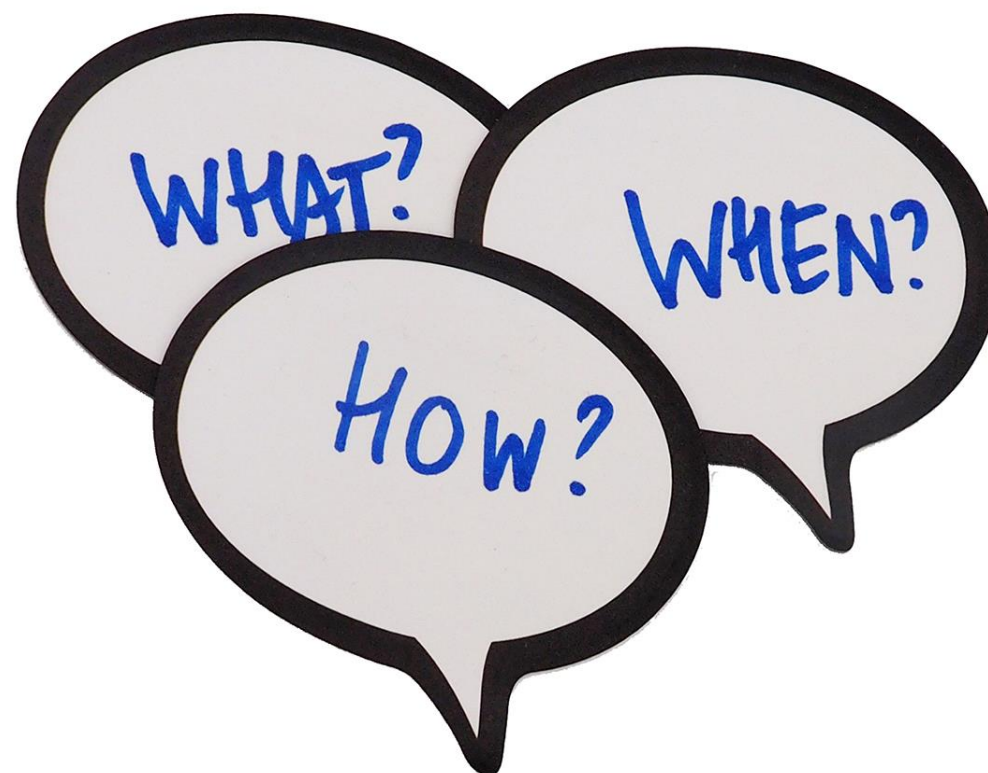
Interoperable

(Meta)data should use standardised terms (controlled vocabularies), have references to other (meta)data and be machine actionable.

Reusable

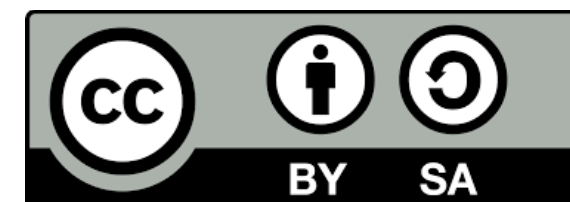
(Meta)data are sufficiently well described for both humans and computers to be able to understand them and have a clear and accessible data usage license.

The Data Management Expert Guide



The Data Management Expert Guide

- Guide on Research Data Management (RDM)
 - For early career researchers (in the social sciences)
- Provides useful information on RDM in one central place
- Created by CESSDA training team 2017-2020
- Free to use at www.cessda.eu/dmeg

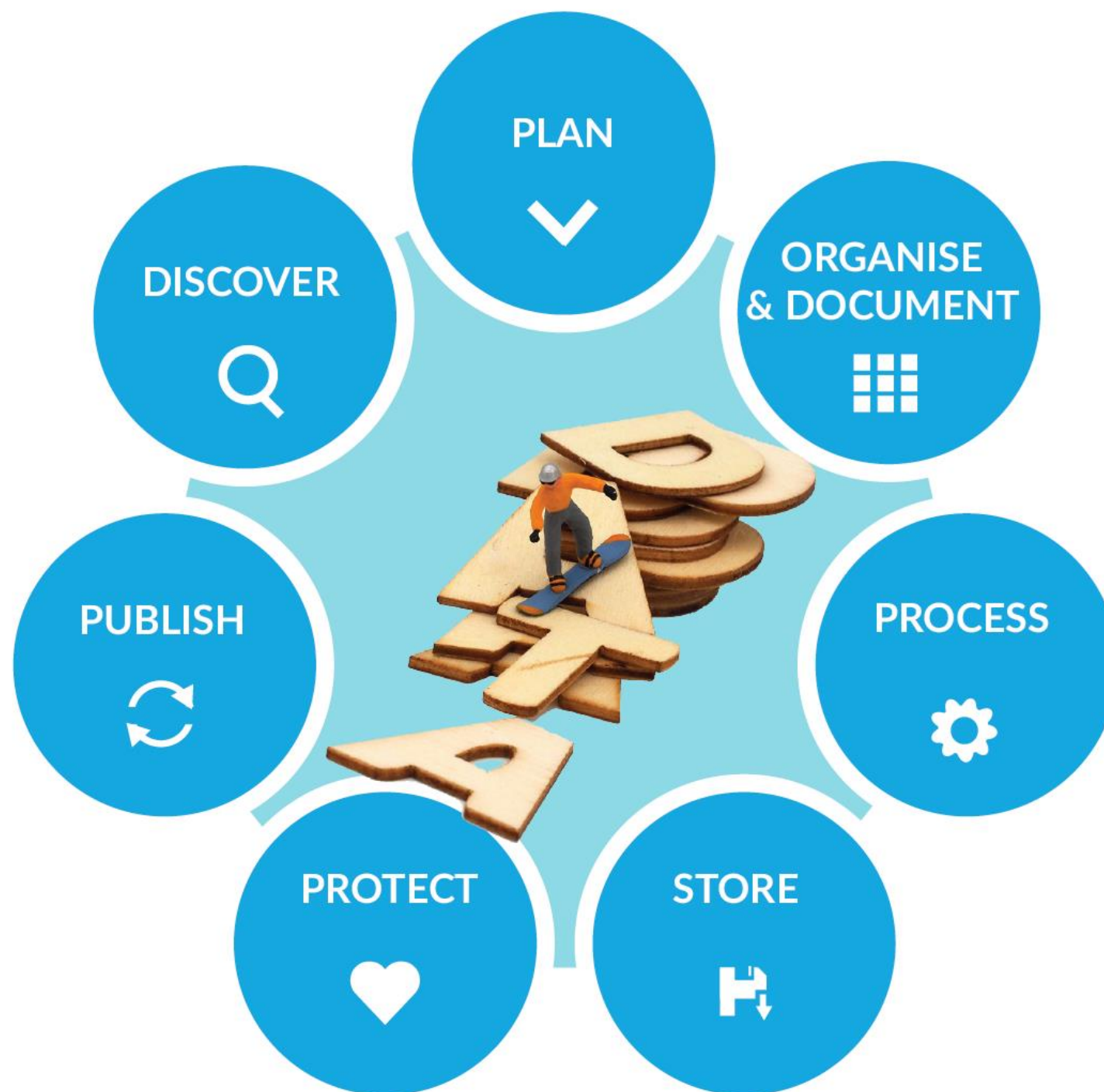


Helps you to create a DMP

- Data Management Plan [checklist](#) to answer questions for your own study



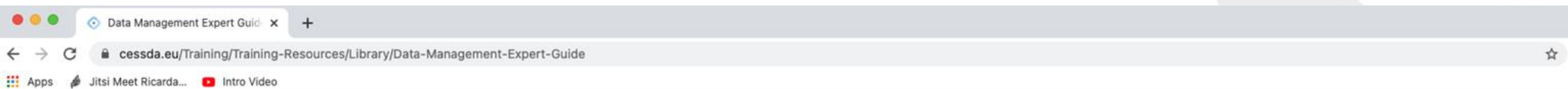
Following the Life Cycle



A closer look at the DMEG



A closer look at the DMEG



Consortium of European Social Science Data Archives



[EVENT CALENDAR](#) [TRAINING RESOURCES](#) [ABOUT](#)



Training / Training Resources / Data Management Expert Guide



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](#).

You can [download](#) the full DMEG for your personal study offline (DOI: [10.5281/zenodo.3820473](https://doi.org/10.5281/zenodo.3820473)). PDFs for every [single chapter](#) are also available for being printed as handouts for training.

Search this guide

Search



A closer look at the DMEG

Data Management Expert Guide ▾

- 1. Plan >
- 2. Organise & Document >
- 3. Process >
- 4. Store >
- 5. Protect >
- 6. Archive & Publish >
- 7. Discover >
- 8. Contributors >

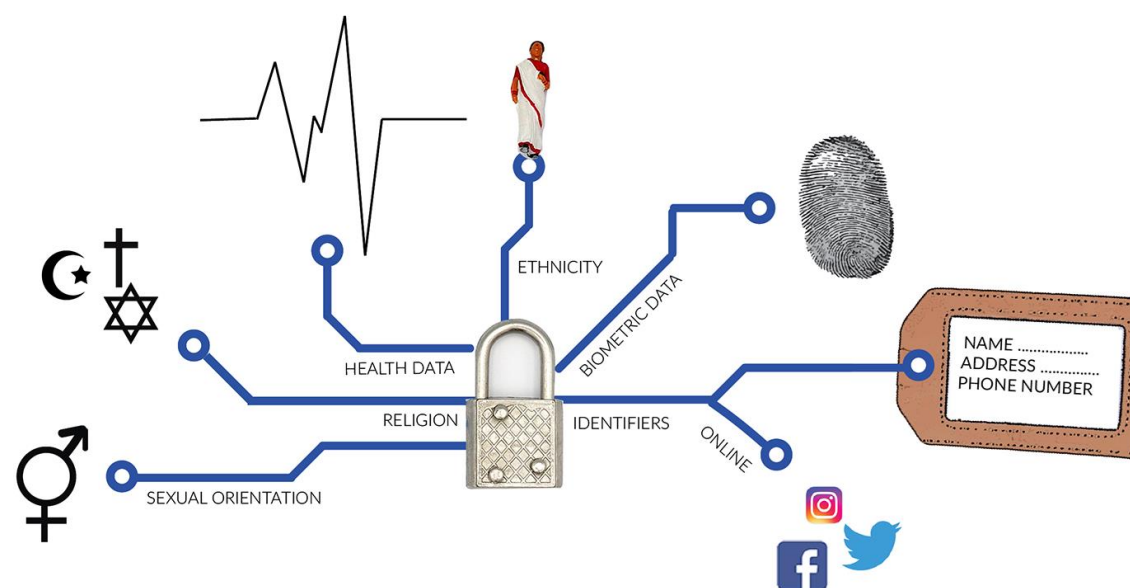
This presentation



A closer look at the DMEG

Data Management Expert Guide

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



A closer look at the DMEG

Data Management Expert Guide	▼
1. Plan	>
2. Organise & Document	>
3. Process	>
4. Store	>
5. Protect	>
6. Archive & Publish	>
7. Discover	>
8. Contributors	>

Thursday's presentation

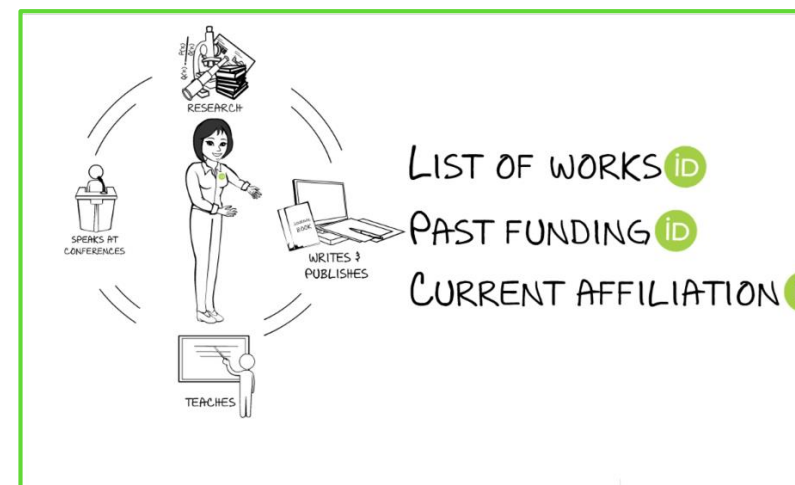


Plan



Plan

- Describe the project and the data you will collect
- Collect all basic information
 - Contact details + [ORCIDs](#)* of involved researchers
 - Grant information



*ORCID is a persistent identifier (PID) for researchers. Learn more about ORCIDs in [this video](#) and PIDs in general in [this video](#)



Plan

- Collect all basic information
 - Roles and responsibilities

⊖ Project data contact

Who can be contacted about the project during and after it has finished?

⊖ Data owner(s)

- Which organisation(s) own(s) the data?
- If several organisations are involved, which organisation owns what data?

⊖ Roles

- Who is responsible for updating the DMP and making sure that it's followed?
- Do project participants have any specific roles?
- What is the project time line?



Plan

- Collect all basic information
 - Think about the costs for data storage and management

⊖ Costs and Resources

- Are there costs you need to consider to buy specific software or hardware?
- Are there costs you need to consider for storage and backup?
- Are potential expenses and resources for (preparing the data for) archiving covered?
- What resources will be dedicated to data management ensuring that data will be FAIR?

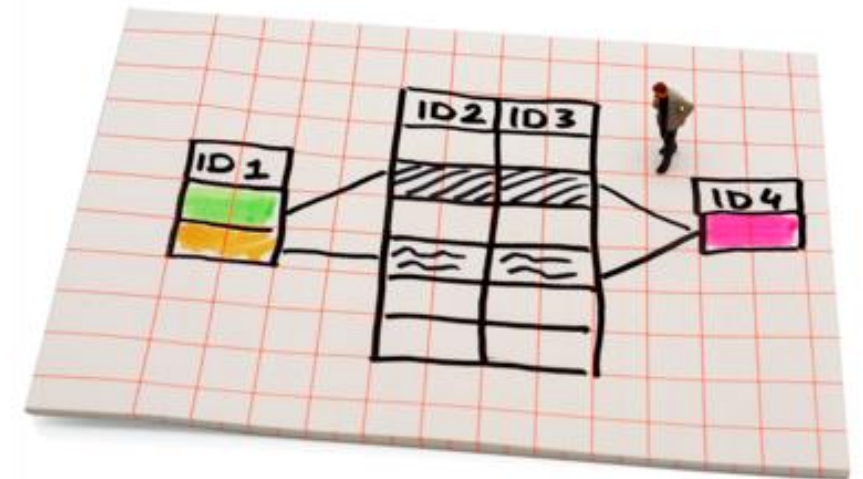


Organise & Document



Organize

- Think about file structures
 - Depends on
 - type of data
 - analysis software
 - Linking with other datasets
 - DMEG gives examples for qualitative and quantitative data



Organize

- File naming

Best practice

According to the UK Data Archive ([UK Data Service, 2017b](#)), a best practice in naming files is to:

- Create meaningful but brief names;
- Use file names to classify types of files;
- Avoid using spaces, dots and special characters (& or ? or !);
- Use hyphens (-) or underscores (_) to separate elements in a file name;
- Avoid very long file names;
- Reserve the 3-letter file extension for application-specific codes of file format (e.g. .doc, .xls, .mov, .tif);
- Include versioning of file names where appropriate.

Documenting data file conventions

An example of how to document the data file conventions you use:

`<date><type><ID1><gender><age><municipality><datatype><ID2>`

where:

- `<date>` is the date on which the data were collected (date format should be YYYY-MM-DD);
- `<type>` specifies the type of event/data material;
- `<ID1>` is the ID of the collection event;
- `<gender>` is the gender of the interviewee;
- `<age>` is the age of the interviewee;
- `<municipality>` is the municipality of residence of the interviewee;
- `<datatype>` specifies the type of data the file contains, for instance, "trans" means transcription, "audio" means audio recording, and "image" means photograph;
- `<ID2>` is the ID number used to separate the images connected to the collection event.



Organize

- File naming

Best practice

According to the UK Data Archive ([UK Data Service, 2017b](#)), a best practice in naming files is to:

- Create meaningful but brief names;
- Use file names to classify types of files;
- Avoid using spaces, dots and special characters (& or ? or !);
- Use hyphens (-) or underscores (_) to separate elements in a file name;
- Avoid very long file names;
- Reserve the 3-letter file extension for application-specific codes of file format (e.g. .doc, .xls, .mov, .tif);
- Include versioning of files where appropriate.

Documenting data file conventions

An example of how to document the data file conventions:

<date><type><ID1><gender><age><municipality><ID2>.ext

where:

- <date> is the date on which the data were collected (date format should be YYYY-MM-DD);
- <type> specifies the type of event/data material;
- <ID1> is the ID of the collection event;
- <gender> is the gender of the interviewee;
- <age> is the age of the interviewee;
- <municipality> is the municipality of residence of the interviewee;
- <datatype> specifies the type of data the file contains, for instance, "trans" means transcription, "audio" means audio recording, and "image" means photograph;
- <ID2> is the ID number used to separate the images connected to the collection event.

Do not include personal information in filenames!



Organize

- Folder structures

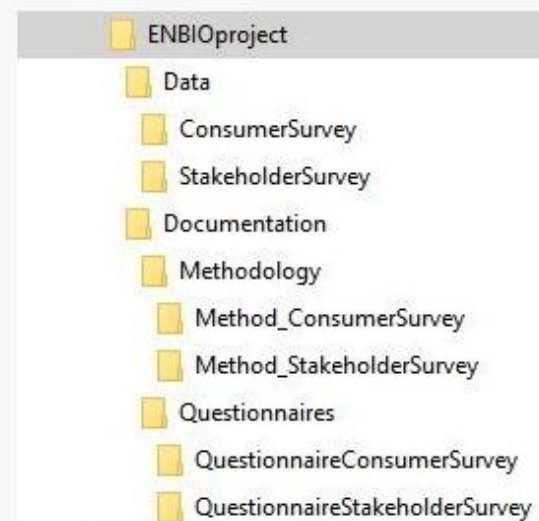
- Depends on

- Project structure
 - Data collected
 - Collaborations

- DMEG gives examples for qualitative and quantitative data

⊖ Survey data

For this survey, data and documentation files are held in separate folders. Data files are further organised according to data type and then according to research activity. Documentation files are organised also according to the type of documentation file and research activity. It helps to restrict the level of folders to three or four deep and not to have more than ten items on each list.



⊕ Qualitative data files



Organize

Metadata is defined as the data providing information about one or more aspects of the data

- Metadata and documentation

Project level details

- ⊕ 1. For what purpose was the data created
- ⊕ 2. What does the dataset contain?
- ⊕ 3. How was data collected?
- ⊕ 4. Who collected the data and when?
- ⊕ 5. How was the data processed?
- ⊕ 6. What possible manipulations were done to the data?
- ⊕ 7. What were the quality assurance procedures?
- ⊕ 8. How can the data be accessed?

Data level details

- **Information about the data file**
Data type, file type, and format, size, data processing scripts.
- **Information about the variables in the file**
The names, labels and descriptions of variables, their values, a description of derived variables or, if applicable, frequencies, basic contingencies etc. The exact original wording of the question should also be available. Variable labels should:
 - 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.

→ DMEG gives examples for qualitative and quantitative data



Organize

- Metadata and documentation → Crucial for FAIR



Findable

To aid automatic discovery of relevant datasets, (meta)data should be easy to find by both humans and machines and be assigned a persistent identifier.

Accessible

Limitations on the use of data, and protocols for querying or copying data are made explicit for both humans and machines.

Interoperable

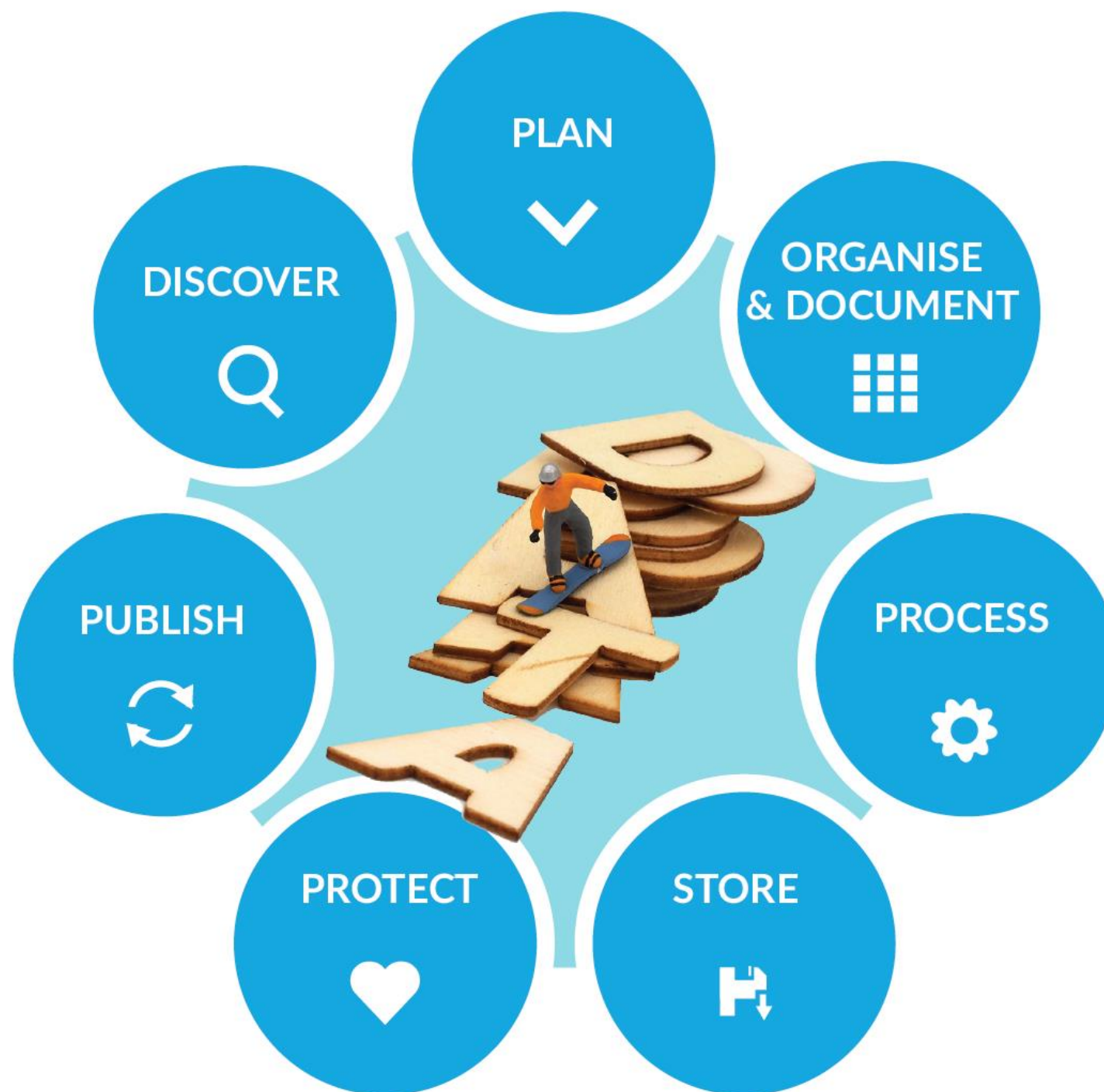
(Meta)data should use standardised terms (controlled vocabularies), have references to other (meta)data and be machine actionable.

Reusable

(Meta)data are sufficiently well described for both humans and computers to be able to understand them and have a clear and accessible data usage license.



Process



Process

- Minimize errors in data processing → DMEG examples
 - Information on coding
 - Information on weights

Minimising errors in survey data entry

In the accordion below a summary of recommendations on minimising errors in survey data entry is given (UK Data Service, 2017a; ICPSR, 2012; Groves et al., 2004).

- ⊕ Check the completeness of records
- ⊕ Reduce burden of manual data entry
- ⊕ Minimise the number of steps
- ⊕ Conduct data entry twice
- ⊕ Perform in-depth checks for selected records
- ⊕ Perform logical and consistency checks
- ⊕ Automate checks whenever possible

Process

- File formats

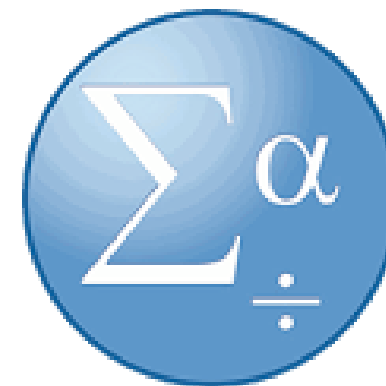
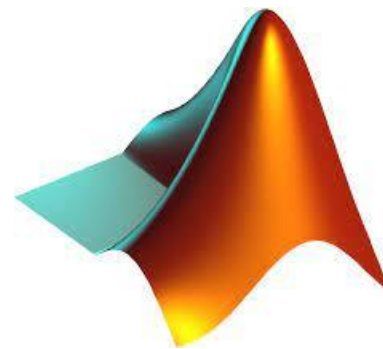
- Can I open this in 5 years from now?

- Some formats are more sustainable than others

- Data archives transform data for you!



vs



Process

- Data authenticity and versioning
 - Know what happened to the data to ensure high quality
 - Document changes
- DMEG gives tips on versioning and logging changes



Store



Store

- **Storage** is often organized by your institute → Find out your policies
 - How much space do you need?
 - Who needs access?

Portable devices	Cloud storage	Local storage	Networked drives
			
<i>Laptops, tablets, external hard-drives, flash drives and Compact Discs</i>			
Advantages	Disadvantages/Risks		Precautions for (sensitive) personal data

Store

- **Backup** is often organized by your institute → Find out your policies

CTRLHC

- ⊕ 1. Find out whether your institution has a backup strategy
- ⊕ 2. Determine what you want to back up
- ⊕ 3. Decide how many backups you will need and how frequently to back up
- ⊕ 4. Decide where backups will be stored
- ⊕ 5. Determine how much storage capacity will be needed
- ⊕ 6. Determine if there are tools you could use to automate backup
- ⊕ 7. Determine how long backups will be kept and how they will be destroyed
- ⊕ 8. Determine how personal data will be protected
- ⊕ 9. Devise a disaster recovery plan
- ⊕ 10. Assign responsibilities
- ⊕ Determine how to check the integrity of backed-up files

Store

- **Security** is crucial, in particular when working with (sensitive) personal data

⊕ Passwords

⊕ Encryption

⊕ Physical, network and computer security

⊕ Secure disposal

⊕ Organisational aspects



Store

How do you share (personal) data with your collaborators?





Thank you for your attention!

ricarda.braukmann@dans.knaw.nl

This workshop is organized by CESSDA-ERIC, ODISSEI and DANS.

