Research Data Mangement

CAS intervisiegroep dataverzameling

Neha Moopen
Research Data Manager

RDM Support

17-11-2022

Agenda

09:30 - 10:00

Welcome & Introductions

Some Definitions

10:00 - 10:45

DMPs

EXERCISE

Data Pipelining

EXERCISE

11:00 - 11:45

Metadata + Documentation

EXERCISE

Codebooks

EXERCISE

12:00 - 12:45

Data Storage & Archiving

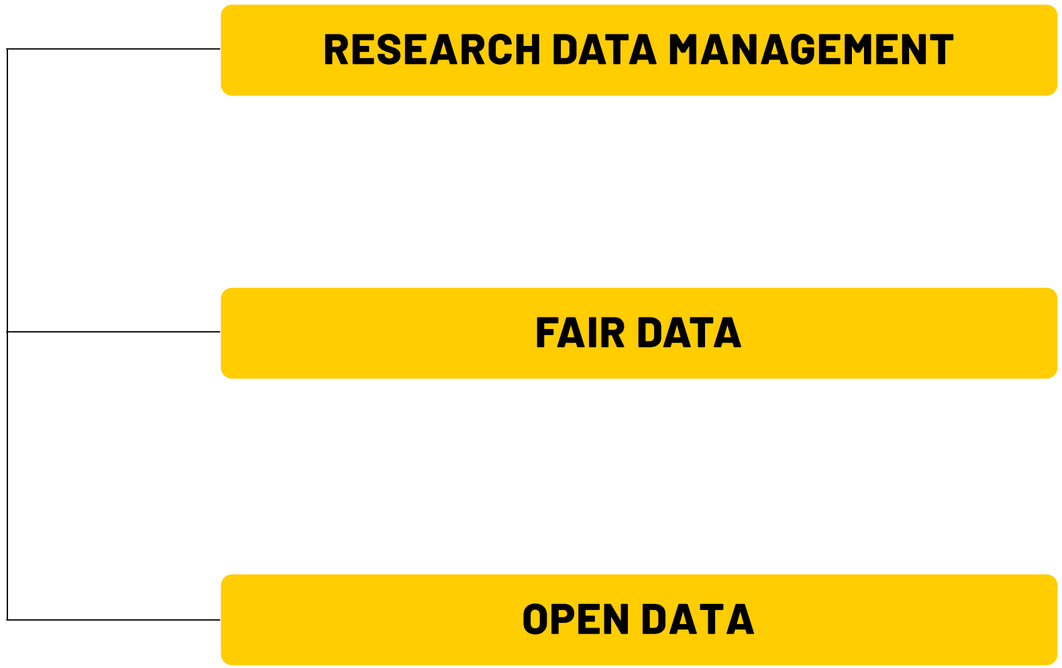
(Meta)data Publication

EXERCISE

Closing + Next Steps

Introduction

Definitions



Research Data Management

Research Data Management (RDM) refers to the **active organization** and **maintenance** of data created during a research project.

It is an **ongoing activity throughout the data lifecycle**, from initial planning to suitable archiving of the data at the project's completion.

FAIR Data

The FAIR Data Principles are **a set of guiding principles** to improve scientific data management and stewardship.

FINDABILITY makes it possible for others to discover your data (metadata, Persistent Identifiers, etc.).

ACCESSIBILITY makes it possible for humans and machines to gain access to your data, under specific conditions or restrictions where appropriate.

INTEROPERABILITY ensures data and metadata conform to recognized formats and standards which allows them to be combined and exchanged.

REUSABILITY requires lots of documentation, which is needed to support data and interpretation and reuse.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Open Data

Open Data is data that can be **freely used, re-used, and redistributed by anyone** - subject only, at most, to the requirement to attribute and share-alike (Open Data Handbook).

Your data does not have to be 'open' to be FAIR!

Make your data... **'as open as possible, as closed as necessary'** (European Commission).

Definitions

in context

RESEARCH DATA MANAGEMENT

RDM = an activity / practice

FAIR DATA

FAIR = principles that guide RDM activities / practices

OPEN DATA

data does not have to be 'open' to be FAIR!

Data Management Plans (DMPs)

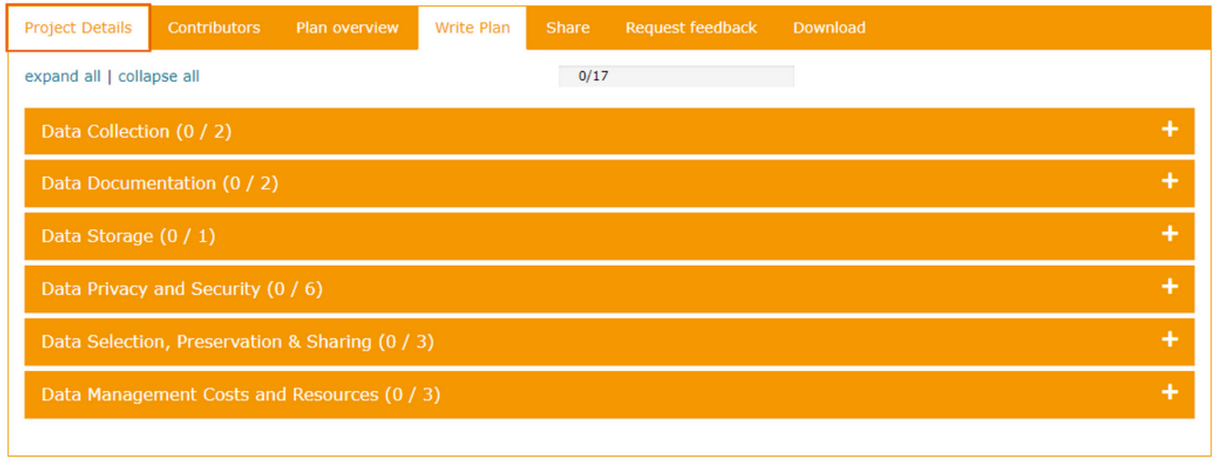
Data Management Plans

A Data Management Plan (DMP) is a formal document that:

- describes your data, and
- outlines all aspects of managing your data, both during and after your project.

It is also a *living* document, it can (and should) be continually edited and updated.

A DMP helps make your RDM activities more concrete and actionable. It will save you time, work, and potentially money too.



👆 A screenshot of the UU DMP template on DMPonline.

Did you know? When you complete a draft of your DMP, you can submit it to RDM Support for review and feedback! Alternatively, you can ask Neha to co-write a detailed DMP for your project!

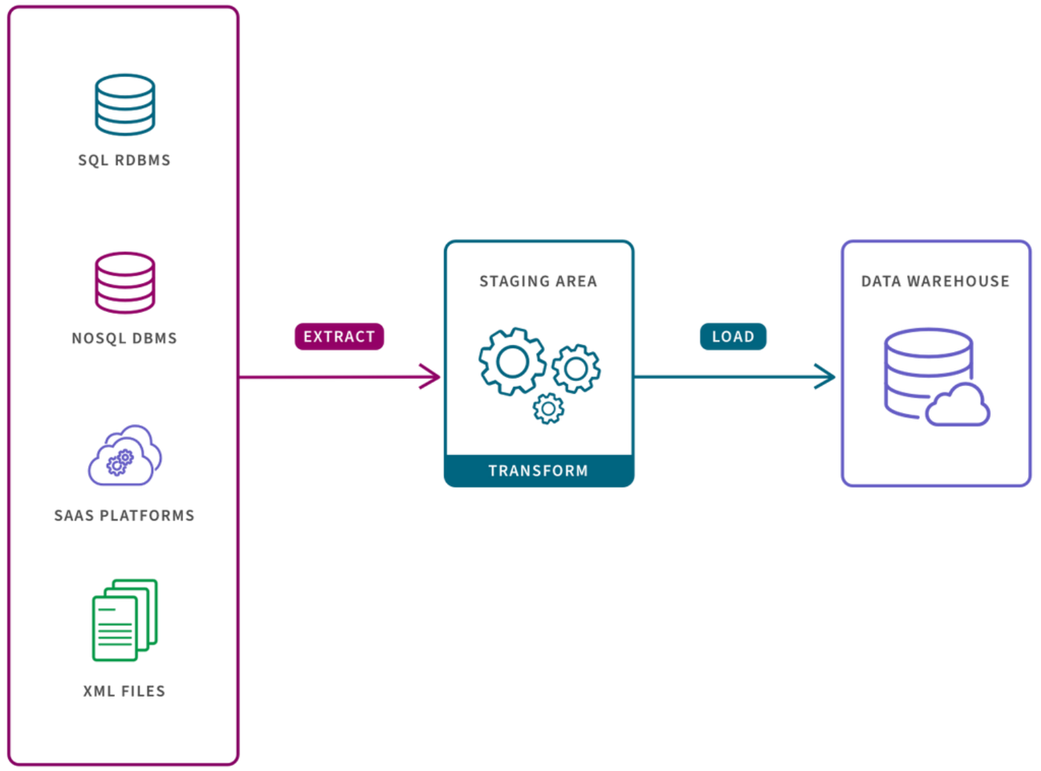
Data Pipelining

Data Pipelining

A data pipeline is a series of (automated) actions that ingests raw data from various sources and moves the data to a destination for storage and (eventual) analysis.

Benefits of a data pipeline include:

- Time saved by automating the boring stuff!
- Reduced mistakes.
- Tasks broken down into smaller steps.
- Reproducibility!



Data Pipelining

When do I need a data pipeline?

Here's a rule of thumb, just as an example:

If you have a task that needs to occur >= 3 times, you should think about automating it.

If automation is not possible, think about how you can make the task as efficient as possible.

Two examples of data pipelines inlcude:

ETL

ETL refers to an *Extract, Transform, Load* process for data. This involves applying some transformations to the raw data as soon as it is extracted and storing this (semi-)processed data (along with a copy of the raw data) until it's time to be analyzed.

ELT

ELT refer to an *Extract, Load*, *Transform* process for data. This involved extracting the raw data and immediately storing it, with the transformations applied as a later step - possibly on a case-by-case basis or closer to the analysis point.

Data Pipelining

How can I implement a data pipeline? Some examples for inspiration:

- If your data collection tools have APIs, they can be leveraged to extract data.
    - For example, Qualtrics has the qualtRics R package & QualtricsAPI Python library which contain functions to automate exporting surveys.
- If APIs are not available, you could use R/Python to automate the use of an internet browser using the RSelenium package / Selenium library. Imagine automating the clicks and typing of going to a specific website, logging in, clicking the download button.
- You can use Windows Task Scheduler / cron / taskscheduleR / cronR to schedule your scripts to run automatically, on a recurring basis as well (if needed).
- You can also send emails with R & Python! Consider if you've ever had to contact participants because you noticed something wrong with their incoming data. You could implement these data checks with a script and automatically draft and send emails (from a template) to those participants who were flagged as having issues with their data. For R users, the Microsoft365R and gmailR packages could be interesting.

EXERCISES!

Exercises

DATA MANAGEMENT PLAN

- Sign into DMPonline with your institutional credentials and create a DMP with the UU (or your institutional) template.

- Complete the first page on *Project Details*.

If you already have a DMP, you can ask Neha or Dafne questions about it.

No questions? You can (and should) later review your DMP to see if it's up to date.

DATA PIPELINING

- Go to https://app.diagrams.net/ and start drawing a data flow diagram for your project.

- Think about the following components:

    - sources of the data.
    - destination of the data.
    - (pre-)processing or transformations that needs to occur before analysis

- Think about where you can automate and make things more efficient!

Metadata & Documentation

Metadata

Metadata is **structured information** that describes one or more aspects of your research data.

In other words, metadata = 'data about data' 😁

Metadata is **machine-readable** and helps make your data findable and citable.

Metadata exists at different levels:

PROJECT-LEVEL METADATA

This type of metadata describes higher-order aspects of your dataset: the “who, what, where, when, how and why” ...

It provides context for understanding why the data were collected and how they were used.

DATA-LEVEL METADATA

This type of metadata is more granular and describes the data (variables) and dataset in detail.

Metadata

PROJECT-LEVEL METADATA

- Name of the project
- Dataset title
- Project description
- Dataset abstract
- Principal investigator and collaborators
- Contact information
- Dataset handle (DOI or URL)
- Dataset citation
- Data publication date
- Geographic description
- Time period of data collection
- Subject/keywords
- Project sponsor
- Dataset usage rights

DATA-LEVEL METADATA

- Data origin: experimental, observational, raw or derived, physical collections, models, images, etc.
- Data type: integer, Boolean, character, floating point, etc.
- Instrument(s) used
- Data acquisition details: sensor deployment methods, experimental design, sensor calibration methods, etc.
- File type: CSV, mat, xlsx, tiff, HDF, NetCDF, etc.
- Data processing methods, software used
- Data processing scripts or codes
- Dataset parameter list, including
    - Variable names
    - Description of each variable
    - Units

Documentation

Documentation refers to **contextual information** pertaining to your research data.

Documentation is meant to be **human-readable** and it is a crucial aspect of **interoperability** and **reusability**.

It accompanies (structured) metadata and guides users to understand and interpret your data and reuse it effectively.

EXAMPLES

- Grant / Study Proposals
- Study Protocol / Methodology
- Data Management Plan (DMP)
- Codebooks
- README files
- Lab Notebooks
- Legal / Policy / Administrative Documents

DOCUMENTATION CHECKLIST!

Here is a starter checklist (v0) that you can use to make an inventory of your documentation: https://tinyurl.com/documentation-checklist

Codebooks

A codebook is an example of data-level metadata.

The purpose of a codebook or data dictionary is to explain what all the variable names and values in your spreadsheet really mean.

Information to include in a codebook includes:

- Variable Names
- Readable Variable Name
- Measurement Units
- Allowed Values
- Definition Of The Variable
- Synonyms For The Variable Name (Optional)
- Description Of The Variable (Optional)
- Other Resources

See: https://help.osf.io/article/217-how-to-make-a-data-dictionary

HUMAN-READABLE CODEBOOK

These are common and you have probably created and/or used them. They are typically Word or PDF documents of several pages and tables.

See: https://osf.io/2hkrw

MACHINE-READABLE CODEBOOK

Codebooks in the form of a spreadsheet (csv or xlsx) that can be easily read by computers and used with programming tools, ideally to compare with other machine-readable codebooks.

See: Cookbook for a Codebook at https://osf.io/q3s4q/

Codebooks

MEASUREMENT-LEVEL

This provides a high-level overview of the data. It lists all the measurements in terms of what instruments, questionnaires/surveys, experiments are covered.

ITEM-LEVEL CODEBOOK

This provides a granular overview of the data. It dives into the measurements and describes every item/variable in detail, starting with the item name and going on to type of data, response options and their respective codes etc.

EXERCISES!

Exercises

DOCUMENTATION CHECKLIST

- Download / make a copy of the documentation checklist available via this link: https://tinyurl.com/documentation-checklist
- Complete the checklist as far as possible.
- Based on your checklist, go back to your DMP and update the *Data Documentation* section as far as possible.

CODEBOOKS

- Demo of the codebook R package!

Data Storage

Data Storage

WHAT DO YOU CONSIDER WHEN CHOOSING A STORAGE SOLUTION?

- STORAGE SPACE?
- PRICE?
- BACKUPS?
- UU-MANAGED?
- INTERNAL COLLABORATION?
- EXTERNAL COLLABORATION?
- SENSTIVE INFORMATION?
- REMOTE ACCESS?

Data Storage

OVERVIEW AND COMPARISON OF STORAGE SOLUTIONS

Storage Option					YODA	OneDrive	SURF DRIVE
Storage size	Varied	Varied	Varied	Varied	Varied	1TB	250GB
Price	NA	NA	Faculty	Faculty	TB €4/m	UU	UU
Back-up	✗	✗	✓	✓	✓	✓	✓
Controlled by UU	✗	✗	✓	✓	✓	✓	✓
Internal collaboration	✗	✗	✗	✓	✓	✓	✓
External Collaboration	✗	✗	✗	✗	✓	✓	✓
Sensitive Information	✗	✗	✓	✓	✓	✓	✓
Remote Access	✗	✗	✓	✓	✓	✓	✓

Data Storage

BEST PRACTICES IN STORING DATA

- I. Choose storage media wisely
- II. Manage versions and copies of your data carefully
- III. Structure names and folders
- IV. Find and understand your data by assigning metadata
- V. Use standard file formats
- VI. Secure your data files

For more information, see the RDM Support guide on Storing & Preserving Data.

EXERCISE!

Exercises

DATA STORAGE FINDER

- Go to the UU Data Storage Finder and see which storage tool might be most suitable for your project.
- Share the tool recommended for you in the chat!

DATA MANAGEMENT PLAN

- Based on the recommendations of the Data Storage Finder, go back to your DMP and update the *Data Storage* section as far as possible.

Data Archiving

Data Archiving

UNIVERSITY POLICY FRAMEWORK FOR RESEARCH DATA (UU)

Raw data from research must be kept available for a **minimum of ten years** for verification purposes…

…commencing from the date that the research results are published!

Research policy NWO › Themes › Open Science › Research data management

Research data management

NWO expects researchers to:

- Preserve these data for **at least ten years**, unless legal provisions or discipline-specific guidelines dictate otherwise;

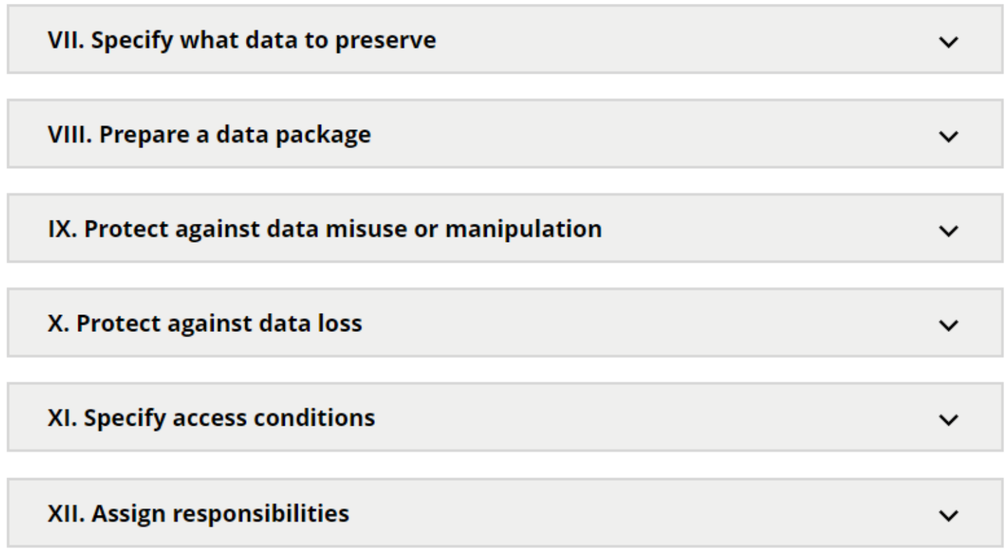
Data Archiving

Data Archiving refers to the **long-term preservation of research data**.

Unlike (Meta)data Publication, Data Archiving is not *directly* related to the FAIR Data Principles.

It is typically done for verification purposes / to check & maintain the integrity of the original research, whereas the FAIR Data Principles are about sharing and reusing the data.

Nonetheless, similar RDM practices are involved in both activities!
So archiving can always provide a basis for FAIRification…



For more information, see the RDM Support guide on Storing & Preserving Data.

(Meta)Data Publication

(Meta)data Publication

DATA ARCHIVING

You **archive** data for future reference.

- You create a data archiving package
- You make sure you can read and access the (raw) data later on.
- You ensure access to (specific) others for verification purposes.

V

S

(META)DATA PUBLICATION

You **publish** data for findability and reusability.

- You create a data publication package
- You share data (and information about the data) that can be used by others for their own purposes.
- You specify the terms and conditions for access and reuse.

REFLECT ON WHAT YOU NEED TO PRESERVE vs. WHAT YOU WANT TO PUBLISH

(Meta)data Publication

Data Publication Package

Part of documentation is also determining the contents and organization of the Research Project Package:

- Raw data
- Processed data
- Documentation and Metadata
  - Variable codebook
  - Lab journals
  - Metadata standard (At the data and descriptive level; if any)
  - Protocols and Standard operating procedures (SOP)
- Scripts & Software
  - Analysis
  - Processing
- Legal Documents
  - Data protection impact assessment (DPIA)
  - Licenses
  - Informed Consent form (template only)
  - Data transfer agreement (DTA)
- Administration
  - Ethical review
  - DEC approval (animal research)
  - Grant & Consortium: applications and agreements
  - Data management plan (DMP)

EXAMPLE

- If there are privacy-sensitive data involved, the processed data should be anonymized.
- Alternatively, the data files themselves can be placed under restricted access while the metadata and documentation are openly published.
- Remember, your data need not be 'open' to be FAIR!

(Meta)data Publication

WHAT DO YOU CONSIDER WHEN CHOOSING A DATA REPOSITORY?

4TU.Centre for Research Data

YODA

Data Archiving and Networked Services DANS

zenodo

DataverseNL

DRYAD

OSF

B2SHARE EUDAT

figshare credit for all your research

- **PERSISTENT IDENTIFIER (DOI)?**
- **SUSTAINABILITY?**
- **QUALITY CONTROL?**
- **SERVER LOCATION?**
- **COSTS?**

(Meta)data Publication

HOW TO CHOOSE A DATA REPOSITORY?

1 DOMAIN-SPECIFIC REPOSITORY

Use a trusted repository already established for your research domain.

2 INSTITUTIONAL OR RECOMMENDED DATA REPOSITORY

If a trusted domain-specific repository is not available, use an institutional research data repository.
The UU supports **Yoda** & **DataverseNL**.

3 GENERAL PURPOSE REPOSITORY

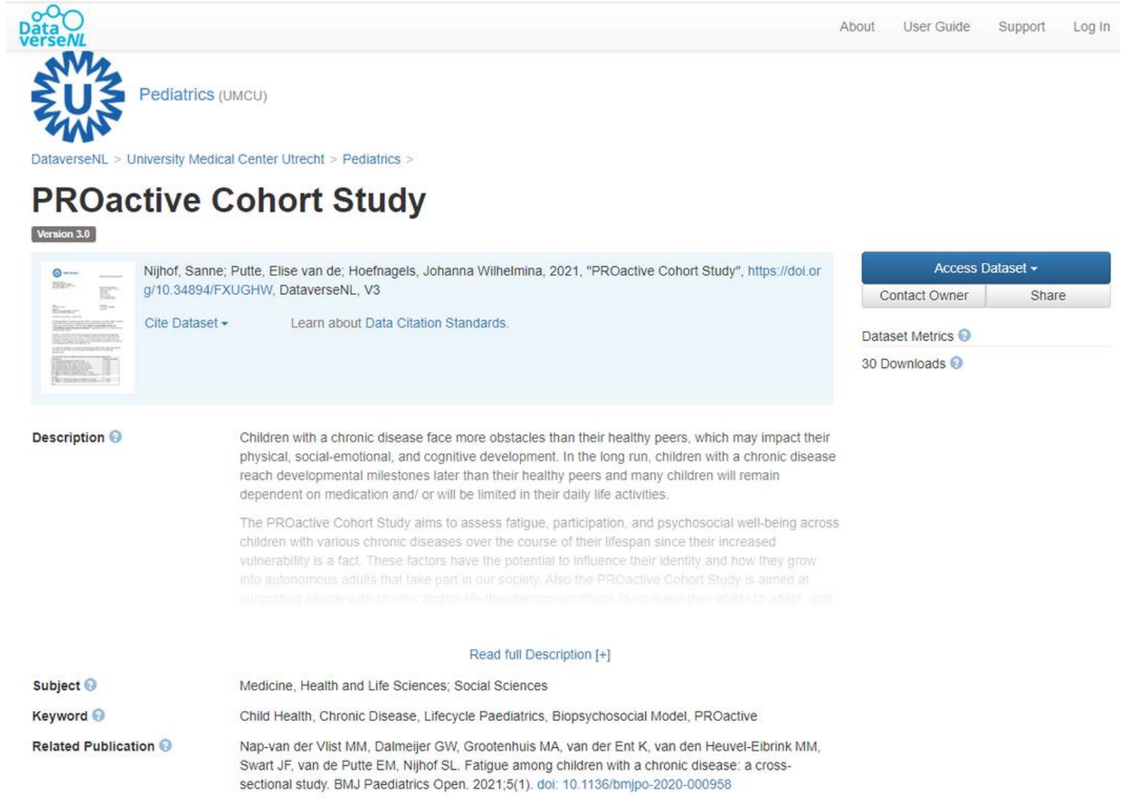
If neither of the above is available, use a general purpose repository like DANS EASY, Zenodo, 4TU.Center, Figshare.

(Meta)data Publication

AN EXAMPLE OF A PUBLISHED COHORT STUDY!

Nijhof, Sanne; Putte, Elise van de; Hoefnagels, Johanna Wilhelmina, 2021, "PROactive Cohort Study", https://doi.org/10.34894/FXUGHW, DataverseNL, V3

You'll see **the data itself has not been published**, but there is a **DOI** that makes the data **Findable and citeable**. There is also extensive **metadata and documentation** that outlines the **Accessibility** and ensures **Reusability**.



EXERCISES!

Exercises

DATA REPOSITORY FINDER

- Go to the UU Data Repository Finder and see which data repository might be most suitable for publishing your project.
- Share the tool recommended for you in the chat!

DATA SELECTION

- Reflect on what data & documentation you would like to publish for eventual citation and reuse.
- Share your thoughts in the chat!
- *Hint:* don't forget to check your project documentation checklist!

DATA MANAGEMENT PLAN

- Based on the above exercises, go back to your DMP and update the *Data Selection, Preservation & Sharing* question (5.2) as far as possible.

Thanks!

For questions or hands-on RDM support,
you can reach me at: n.moopen@uu.nl

I'd be happy to hear from you!

Neha Moopen
Research Data Manager

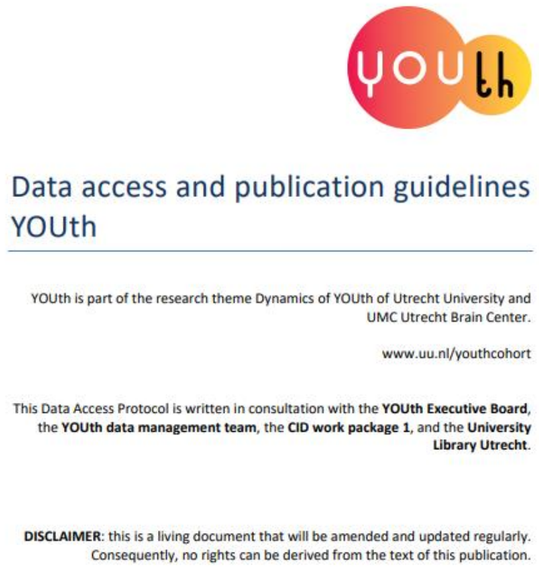
BONUS SLIDES

Data Access Protocol

Data Access Protocol

When you're ready to start sharing your data, you can set up a detailed Data Access Protocol (DAP) that **outlines the data sharing procedure for yourself, your research team, and potential re-users**. This DAP will ideally be public and findable in your chosen repository.

See the YOUth Cohort Study's DAP here: https://www.uu.nl/sites/default/files/dataaccessprotocol_youth_191029.pdf



Data access and publication guidelines
YOUth

YOUth is part of the research theme Dynamics of YOUth of Utrecht University and UMC Utrecht Brain Center.

www.uu.nl/youthcohort

This Data Access Protocol is written in consultation with the **YOUth Executive Board**, the **YOUth data management team**, the **CID work package 1**, and the **University Library Utrecht**.

DISCLAIMER: this is a living document that will be amended and updated regularly. Consequently, no rights can be derived from the text of this publication.

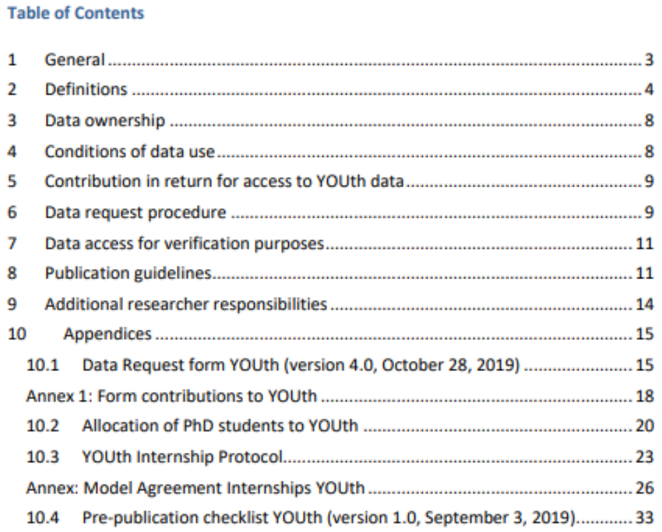
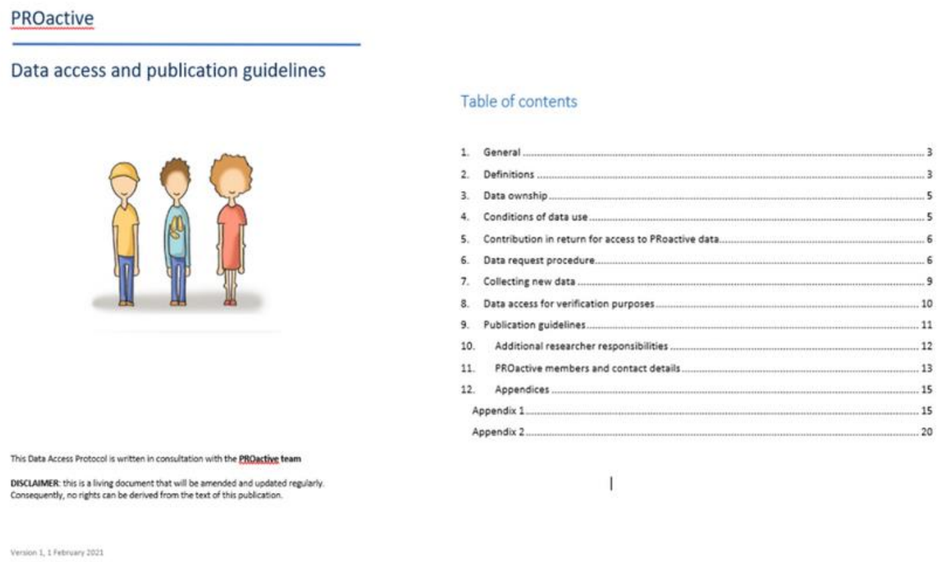


Table of Contents

Data Access Protocol

There are many topics within a DAP, it will require you (and/or the project team to come together) to decide on what is relevant and best for your data. This can include, for example, the terms & conditions for data reuse and the governance procedure.

See the PROactive Cohort Study's DAP here: https://dataverse.nl/file.xhtml?fileId=141206&version=3.0



PROactive

Data access and publication guidelines

This Data Access Protocol is written in consultation with the **PROactive team**

DISCLAIMER: this is a living document that will be amended and updated regularly. Consequently, no rights can be derived from the text of this publication.

Version 1, 1 February 2021

Table of contents

Data Access Protocol

DAPs do not have to be 20-30 pages long like YOUth's & PROactive's! 😈
You can make it as simple and elaborate as you like, it all depends on you and your project team.

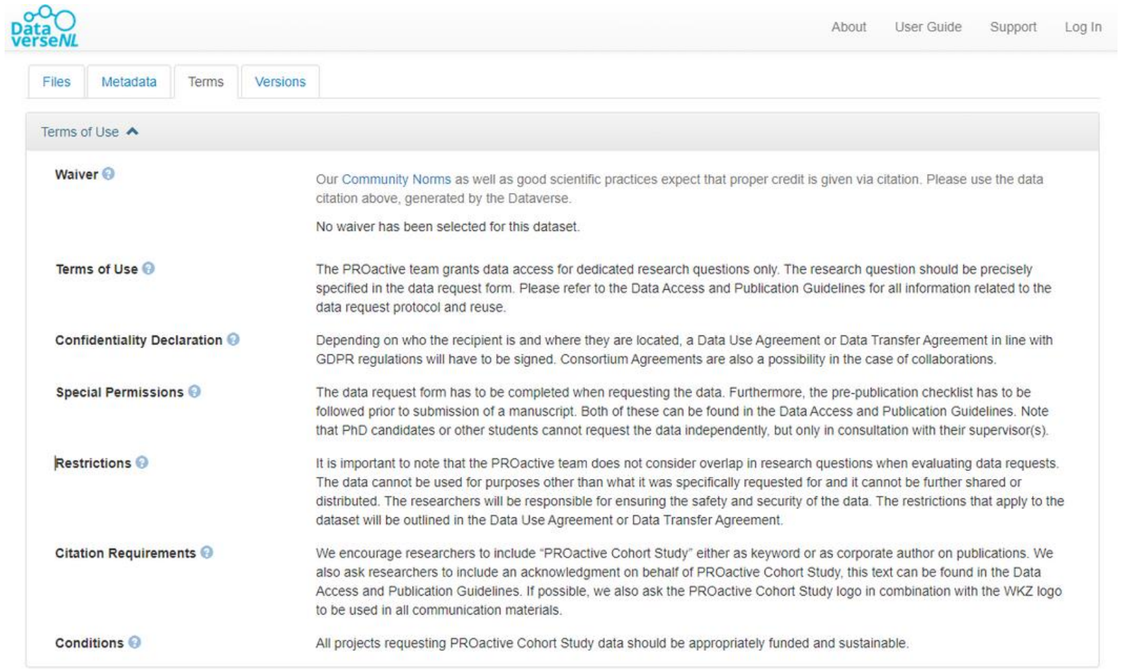
Reflect on:

- **What would you like to get out of sharing the data?** For example, citations/acknowledgments, co-authorship, collaboration? This should be specified in the DAP so the end-user knows their obligations.

- **What kind of time and effort can you and/or your team invest in the data governance?** For example, assessing incoming requests, preparing a datafile for sharing, maintaining a data sharing logbook.

Note: If there is privacy-sensitive data involved, even the simplest DAPs have to take some legal considerations into account!

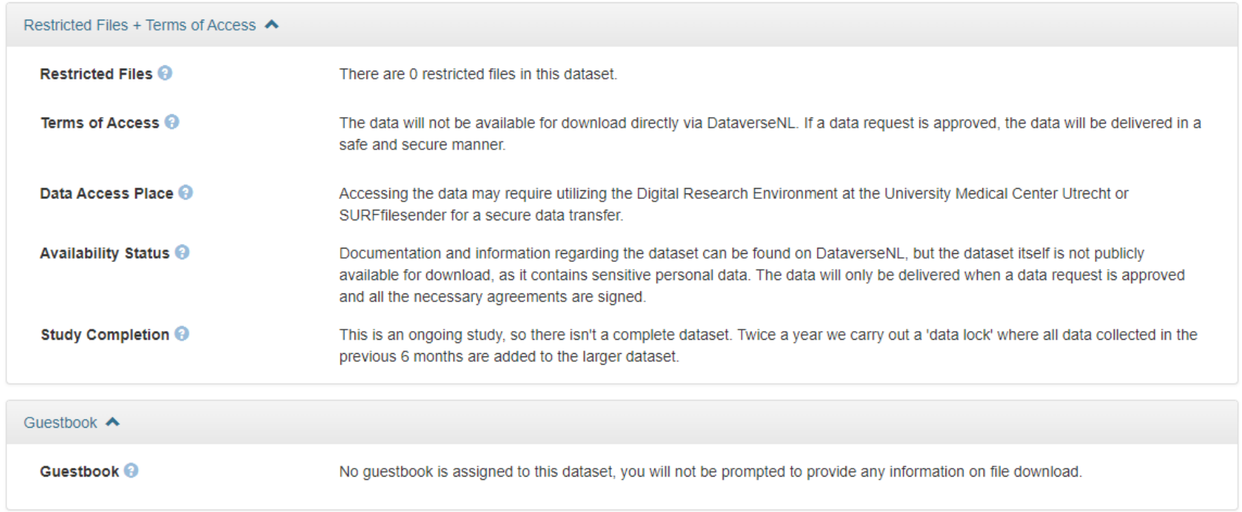
Data Access Protocol



Check under the *Terms* tab of the PROactive Cohort Study's publication: https://doi.org/10.34894/FXUGHW

You'll see the terms and conditions are specified more briefly.

This is an in-built form within DataverseNL and you'll find it in all repositories! It can be a start, in case you don't want to get into a 20-30 page document.



EXERCISE

Exercises

TERMS & CONDITIONS

- Imagine what a data sharing procedure could look like for your project.
- What are some terms and conditions you would like to specify for the end user?
- What investments (time & effort) would be required from you and your team for efficient data sharing
- If you like, share in the chat or discuss with one another!

LINKS

- YOUth Cohort Study's DAP: https://www.uu.nl/sites/default/files/dataaccessprotocol_youth_191029.pdf
- PROactive Cohort Study's DAP here: https://dataverse.nl/file.xhtml?fileId=141206&version=3.0
- PROactive Cohort Study's publication: https://doi.org/10.34894/FXUGHW (for the Terms tab)

Data Sharing

Data Sharing

TOOLS FOR DATA SHARING

SURFFileSender is a reliable tool to send data to another user.

Note: You should only transfer privacy-sensitive data once all the legal requirements are checked and met!

Instead of sending data to another user, you can use **Yoda** to add an (external) user to your research folder. **This does not always work in practice,** since the end-user could download the data to their computer and misuse it.

The IT teams are working on a **Virtual Research Environment** linked to Yoda, keep your eyes open for when it's available!

Data Sharing

IMPORTANT CONSIDERATIONS + DOCUMENTS / PROCEDURES (IN BRIEF)

When privacy-sensitive data is involved:

- The **ICF** should have clearly informed participants about data sharing and reuse + they should agree to it
- A **DPIA** may have to be carried out, this will reveal to what extent it is safe to share data (or not) and how that can be put into practice (for example, pseudonymization techniques)
- Any transfer of data outside the UU will require a **Data Transfer Agreement** in line with the GDPR, the complexity of the DTA will vary depending on the nature of the transfer (for example, transfer outside the EU).

INFORMED CONSENT FORMS

DMP

DATA PROTECTION IMPACT ASSESSMENT (DPIA)

DATA TRANSFER AGREEMENT (DTA)

DATA PROCESSING AGREEMENT (DPA)

Q&A!

Questions so far...

- What are the benefits of open science?
- What will a person do if he or she finds that his or her research idea has been done during pre registration?
- What if a person finds out that someone has done the same data collection as him or her?
- Right now I have the excel sheet for codebook. I've heard about programs like DublinCore to create codebooks, would you recommend using a system like this?
- What repository/storage would you recommend for research concerning data from three waves including interviews, questionnaires and audio files? What are considerations to keep in mind when choosing a repository?
- I'm new to the topic of FAIRification. Would you recommend specific e-learnings or guides?

Thanks!

For questions or hands-on RDM support,
you can reach me at: n.moopen@uu.nl

I'd be happy to hear from you!

Neha Moopen
Research Data Manager