Dynamics of Youth

DATA HANDBOOK

Neha Moopen

2025-10-15

Table of contents

Welcome!

Figure 1: This illustration is created by Scriberia with The Turing Way community. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807

Definitions

Before diving into the Handbook, it would be good to get familiarized with some data-related terms that are oftentimes misunderstood or used interchangeably.

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 (Wilkinson et al., 2016)

- 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.

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).

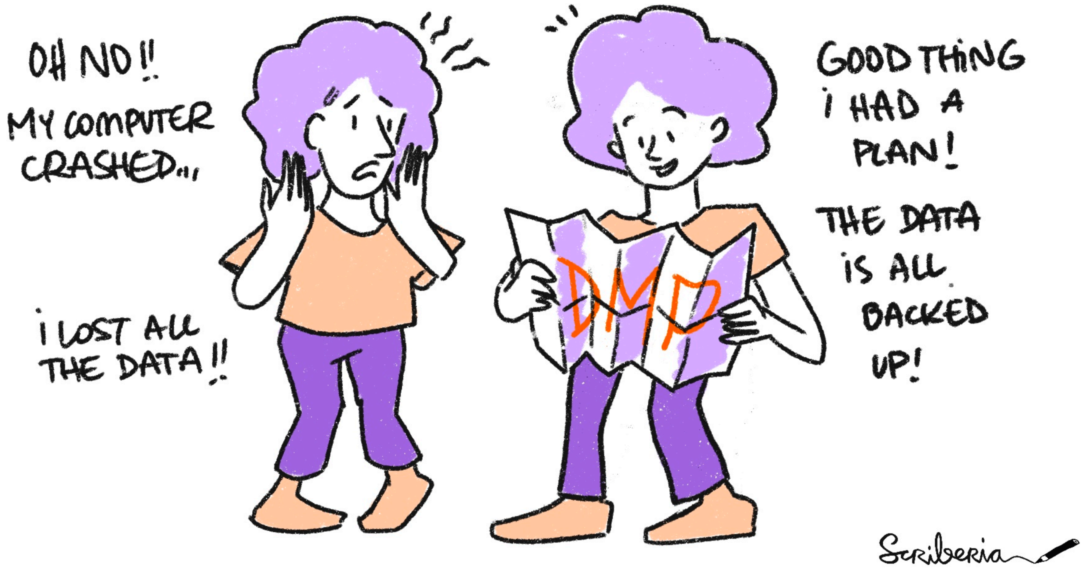
Note that your data does not have to be 'open' to be FAIR! Make your data… 'as open as possible, as closed as necessary' (European Commission).

Summary

In short,

- RDM = an activity/practice
- FAIR = principles that guide RDM activities/practices
- Open Data = data does not have to be ‘open’ to be FAIR!

Data Management Plans



The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807.

What Is A Data Management Plan?

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.

Moreover, it is a *living* document that can you can revise and update as needed.

Why Should You Write A DMP?

Writing a DMP provides an opportunity to reflect on your data, particularly how you organize and manage it. It nudges you to think about how to make your RDM more *concrete* and *actionable*. This creates efficiency and more value for your data.

When Should You Write A DMP?

Working on a DMP at the start of your project will ensure that you are better informed of best practices in RDM and prepared to implement them. That being said, you can also write a DMP can during the project or when it's completed.

DMPonline & DMP Templates

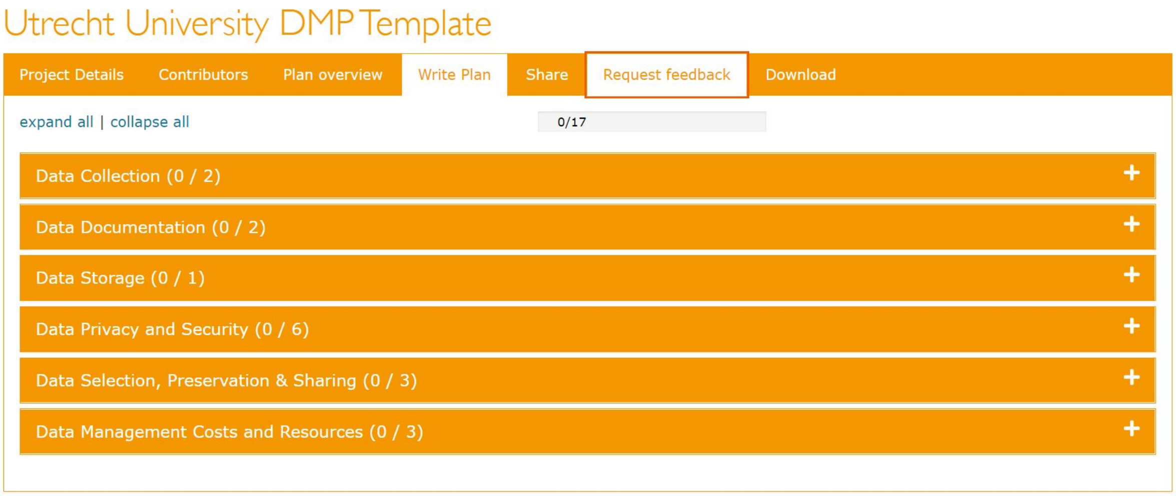
DMPonline is a tool that helps you create and maintain DMPs. With DMPonline, you can:

- register and sign in with your institutional credentials,
- write and collaborate on (multiple) DMPs,
- share DMPs or switch their visibility between private and public,
- request feedback from RDM Support,
- download DMPs in various formats.

DMPonline offers DMP templates from various institutions and funders, including:

- Utrecht University
- UMC Utrecht
- NWO
- ZonMw
- ERC
- Horizon 2020
- Horizon Europe

These templates also contain example answers and guidance.


Utrecht University DMP Template
Project Details
Contributors
Plan overview
Write Plan
Share
Request feedback
Download
expand all | collapse all
0/17
Data Collection (0 / 2)
Data Documentation (0 / 2)
Data Storage (0 / 1)
Data Privacy and Security (0 / 6)
Data Selection, Preservation & Sharing (0 / 3)
Data Management Costs and Resources (0 / 3)

Tips

!!! note "Tips"

```
- Contact your DoY data manager! They can (co)write your DMP and/or review it.
- If the DoY data manager is unavailable, you can still request feedback from RDM Support.
```

Resources

- Create your DMP online
- Data management planning
- Learn to write your DMP (online training)

References

1. https://www.uu.nl/en/research/research-data-management/guides/data-management-planning

2. https://www.kuleuven.be/rdm/en/faq/faq-dmp

3. https://rdm.uva.nl/en/planning/data-management-plan/data-management-plan.html

4. https://www.uu.nl/en/research/research-data-management/tools-services/tool-to-create-your-dmp-online.html

Data Flow Diagrams

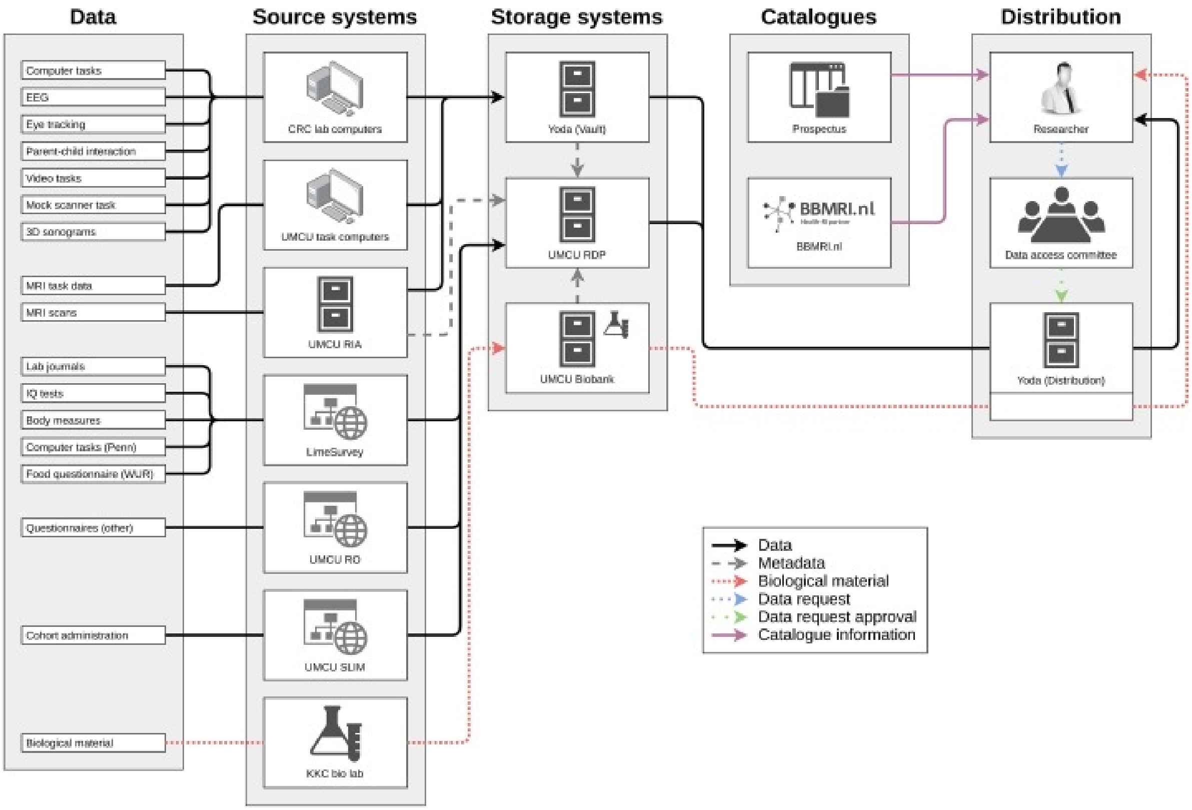
A data flow diagram (DPF) is a visual representation of the flow of data through a process or system. It provides an overview of incoming and outgoing data, as well as the processing and tools involved.

A DFD is valuable because it provides an outline of your data processes. It allows you to see how these processes interact and identify opportunities for improvement.

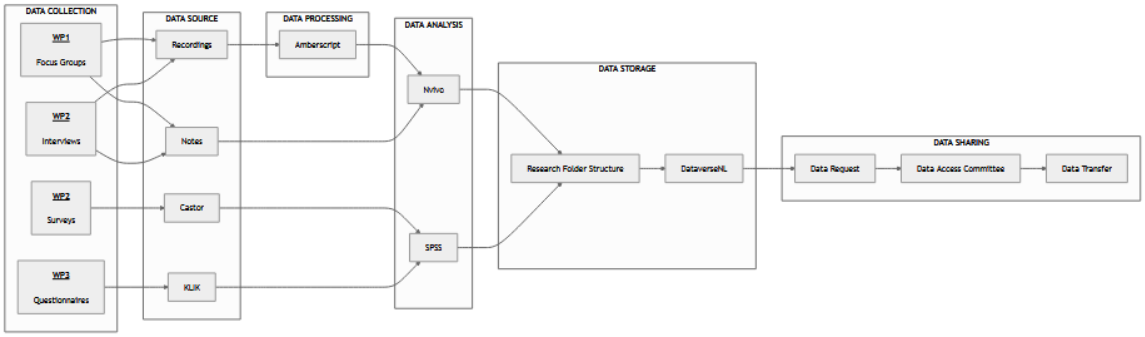
DFDs can be as simple as hand-drawn flowcharts on an A4 sheet of paper to elaborate flowcharts with different symbols and markers. It is recommended that you sketch a DPD while working on your DMP. It can also provide the basis for developing your data pipeline.

Examples

YOUth



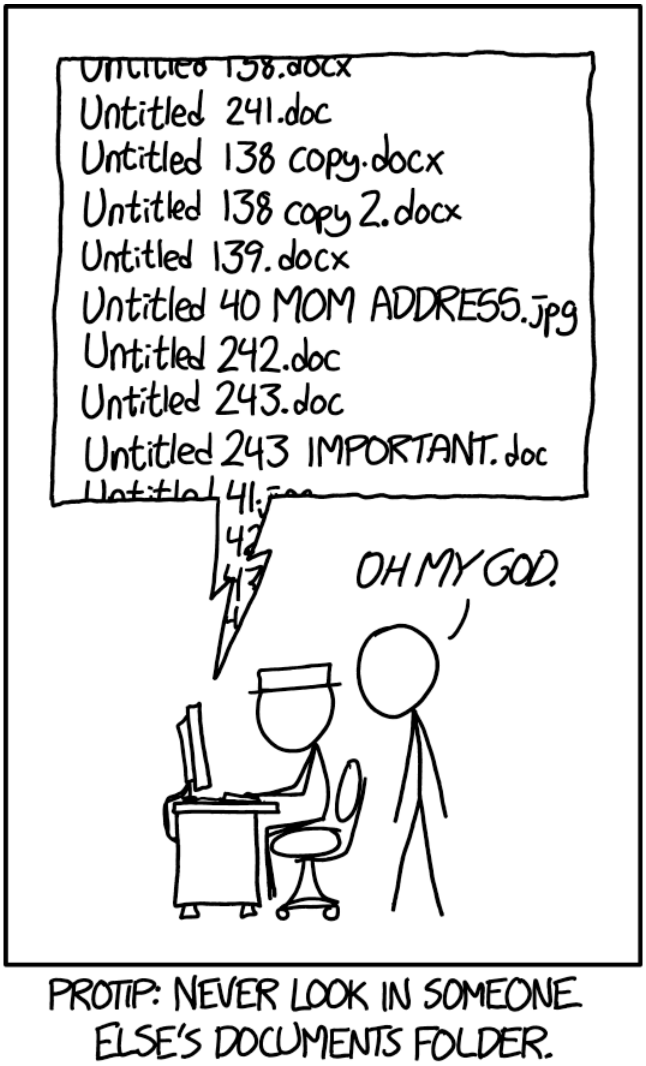
PFIC



Hear, Hear

Smart-Youth

Naming Conventions



Documents - xkcd. Used under a CC BY-NC 2.5 license.

What Is A Naming Convention?

A naming convention is a set of rules for naming things. You can apply it to things like folders, files, and variables.

Why Should I Apply A Naming Convention?

Names that are informative and useful for machines and humans are a step toward efficient data management and reproducible research. The more consistent and meaningful the name, the easier it will be to locate and identify things, understand what they contain, and (re)use them.

When Should I Apply A Naming Convention?

Aim to select and implement a naming convention at the beginning of a project. If you want to retroactively apply a naming convention, there are several tools for bulk renaming.

The entire research team should agree on and adopt a naming convention. Document the choice of naming convention in the DMP, so others can refer to and grasp it quickly.

Popular Naming Conventions

Instead of developing a naming convention from scratch, you can start with one that is already being used in programming and software development communities:

Naming Covention	Example	Description
original name	`an awesome name`	N/A
snake_case	`an_awesome_name`	All words are lowercase and separated by an underscore (`_`)
kebab-case	`an-awesome-name`	All words are lowercase and separated by a hyphen (`-`)
PascalCase	`AnAwesomeName`	All words are capitalized. Spaces are not used.

Naming Covention	Example	Description
camelCase	`anAwesomeName`	The first word is lowercase, the remaining words are capitalized. Spaces are not used.

Human-Readable Names

You can tailor naming conventions like `snake_case` and `PascalCase` to suit your project and workflow. Determine what information is relevant (or not) to create meaningful names and how you can string this information together. Don’t forget to document this in your DMP!

!!! note “Elements for Human-Readable Names”

```
Names should be =<25 characters long and can include:

- Date of creation/update (`YYYY-MM-DD` or `YYYYMMDD`)
- Description of content, like type of data
- Initials of creator/reviewer
- Project number or acronym
- Location/coordinates
- Version number (like `v2` or v2.2`)
```

Machine-Readable Names

When names are machine-readable, they can be efficiently processed by computers and software. This makes it easier to search for files and run operations that involve programming like extracting information from file names or working with regular expressions.

!!! note “Avoid”

```
- Spaces
- Special characters like `$`, `@`, `%`, `#`, `&`, `*`, `!`, `/`, `\`
- Punction characters like `,`, `:`, `;`, `?`, `'`, `"`
- Accented characters
```

A Note on Numbering, Dates, Versioning

- Append numbers to the beginning of a name to enable sorting according to a logical structure. Use multiple digits like `01` or `001`.
- Dates should follow the ISO 8601 standard which is either `YYYY-MM-DD` or `YYYYMMDD`. Append dates to the beginning of names to enable sorting in chronological order.
- Specify versions using ordinal numbers (1,2,3) for major revisions and decimals for minor changes (1.1, 1.2, 2.1, 2.2). Alternatively, you can specify versions with multiple digits like v01 and v02.

Renaming files

The following tools enable renaming in bulk:

- Bulk Rename Utility (Windows, free)
- Renamer (MacOS, paid)
- NameChanger, (MacOS, free)
- GPRename (Linux, free)

References

1. https://en.wikipedia.org/wiki/Naming_convention
2. https://help.osf.io/article/146-file-naming
3. https://rdm.elixir-belgium.org/file_naming.html
4. https://khalilstemmler.com/blogs/camel-case-snake-case-pascal-case/
5. https://dev.to/chaseadamsio/most-common-programming-case-types-30h9
6. https://rdmkit.elixir-europe.org/data_organisation http://dataabinitio.com/?p=987
7. https://dmeg.cessda.eu/Data-Management-Expert-Guide/2.-Organise-Document/File-naming-and-folder-structure
8. https://annakrystalli.me/rrresearchACCE20/filenaming-view.html

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!

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 could think about automating it.

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

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

- If you data collection tools have APIs, they can be leveraged to extract data.
- For example, Qualtrics has the qualtRics R package & pyQualtrics 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 / the taskscheduleR R package / 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.

QualtRics R package

```
library(readr)
library(qualtRics)

qualtrics_api_credentials(api_key = "YOUR-QUALTRICS-API-KEY",
                          base_url = "YOUR-QUALTRICS-BASE-URL",
                          overwrite = TRUE,
                          install = TRUE)

readRenviron("~/.Renviron")

surveys <- all_surveys()

survey_results <- fetch_survey(surveyID = surveys$id[2], # you can also replace surveys$id[2
                                verbose = TRUE)

write_csv(survey_results, paste0("path/to/folder/", format(Sys.time(), "%d-%m-%Y-%H.%M"), "_
```

taskscheduleR package

```
library(taskscheduleR)

scheduled_script <- "path/to/folder/myscript.R"

## run script once within 120 seconds

taskscheduler_create(taskname = "extract-data-once", rscript = scheduled_script,
                     schedule = "ONCE", starttime = format(Sys.time() + 120, "%H:%M"))

## Run every 5 minutes, starting from 10:40

taskscheduler_create(taskname = "extract-data-5min", rscript = scheduled_script,
                     schedule = "MINUTE", starttime = "10:40", modifier = 5)
```

```
## delete tasks

taskscheduler_delete("extract-data-once")
```

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.

• 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

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

Documentation

Documentation refers to contextual information pertaining to your research data. It accompanies (structured) metadata and guides users to understand and interpret your data and reuse it effectively.

Documentation is meant to be human-readable and it is a crucial aspect of interoperability and reusability. Some examples include:

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

Documentation Checklist

Here is a starter checklist 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

codebook R package

```
library(qualtRics)
library(readr)
library(dplyr)
library(codebook)
library(writexl)

surveys <- all_surveys()

survey_results <- fetch_survey(surveyID = surveys$id[2], # you can also replace surveys$id[2]
                               verbose = TRUE)

survey_results <- select(survey_results, -c(1:17))

# survey_questions() retrieves a data frame containing questions and question IDs for a surve
survey_questions <- survey_questions(surveyID = surveys$id[2])
survey_questions <- select(survey_questions, -c(1, 4))
survey_questions <- slice(survey_questions, -1)

# generate codebook

codebook <- codebook_table(survey_results)

codebook <- rename(codebook, qname = name)

codebook <- full_join(survey_questions, codebook, by = "qname")

write_xlsx(codebook, "documentation/codebook-demo.xlsx")
```

The `labelled` R package can also do something similar.

Data Storage

When discussing storage, we are considering the location of ‘active’ data is under use and subject to change during the research project. The related concepts of archiving and publishing refer to where the data will be saved or deposited after the project is completed.

When storing data, consider the following - choose storage media that is appropriate for the type of data you’re working with - implement reliable version control and backups - structure folders and organize files clearly - follow a naming convention - use preferred and sustainable file formats - secure data files

Data Storage Finder

The Data Storage Finder is a tool provided by IT help you decide which storage solution would be most suited to your needs.

Data Archving

Data Archiving refers to the long-term preservation of research data. It is typically done for verification purposes / to check & maintain the integrity of the original research.

There are varying policies on how long research data should be retained for verification purposes, a typical policy is 10 years for the preservation of raw data.

Archiving is not directly related to the FAIR principles, since the latter is focused on sharing and reusing the data. Nonetheless, the steps taken in archiving can provide a bsis for FAIRification, so the effort is never wasted!

Data Publication

When publishing (meta)data, you want to make it findable and reusable. The data (and information about the data) can be used by others for their own purposes. It's up to you to specify the terms and conditions for access and reuse.

Note that your data need not be 'open' to be FAIR! The data files themselves can be placed under restricted access (or retained internally) while the metadata and documentation are openly published. Once any data sharing agreements are signed, the data files an be transferred according to best practices.

When publishing (meta)data, you will receive a landing page for your dataset and a DOI (persistent identifier) that makes it findable and citeable. When you include your metadata and documentation, you improve accessibility and reusability.

Examples

- Nijhof, Sanne; Putte, Elise van de; Hoefnagels, Johanna Wilhelmina, 2021, "PROactive Cohort Study", https://doi.org/10.34894/FXUGHW, DataverseNL, V3
- Isabelle van der Linden; Henk Schipper; Sanne Nijhof; Kors van der Ent, 2024, "SMART-Youth: Data", https://doi.org/10.34894/FCBXSI, DataverseNL, V1

Tools

You can use the UU Data Repository Finder and see which data repository might be most suitable for publishing your project.

Data Governance

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

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 in terms of responsibilities and tasks of the team members.

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

Data Governance can be 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 Sharing

When sharing (personal) data with collaborators outside the university, there are a couple of important considerations:

- The participation letter and informed consent forms should have clearly informed participants about data sharing and reuse + they should agree to it.
- A Data Protection Impact Assessment 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).

Tools

SURFfilesender

SURFFileSender is a reliable tool to send data to another user. You can send large files securely and the option for encryption makes it more safe.

Virtual Research Environments

VREs, for example - AnDREa & ResearchCloud, is a temporary computing environment that is secure and contains the necessary tools and files for users to carry out some research activities.

FAIR Data Cheatsheet