Utrecht University

29-10-2024

CAS course Data Collection

Session 5: Data anticipation

Schedule

- 9:30 Introduction
- 9:45 Data management
- 10:45 Break
- 11:00 Data cleaning
- 12:15 Questions
- 12:30Course evaluation

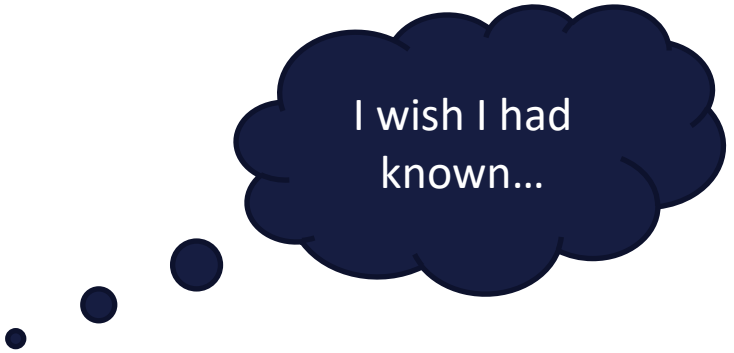
Recap

- Week 1: Introduction
- Week 2: Recruitment
- Week 3: Communication
- Week 4: Registration
- Week 5: Dealing with data

Recap

- Week 1: Introduction
- Week 2: Recruitment
- Week 3: Communication
- Week 4: Registration
- Week 5: ~~Dealing with data~~ Data anticipation

Our regrets

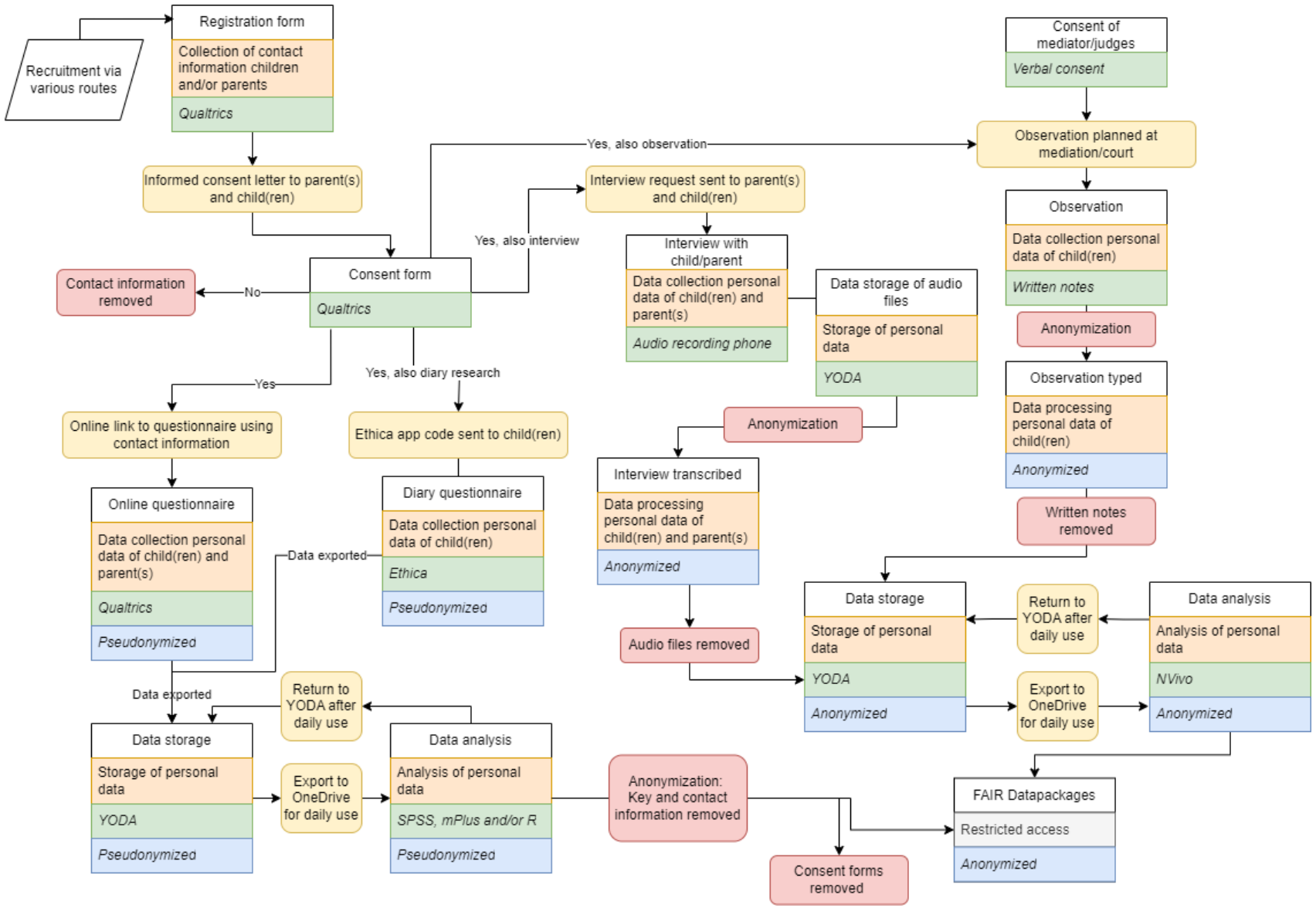


- Not using pilot data
- Overcomplicating (questionnaire) design
- Not automating data export + cleaning

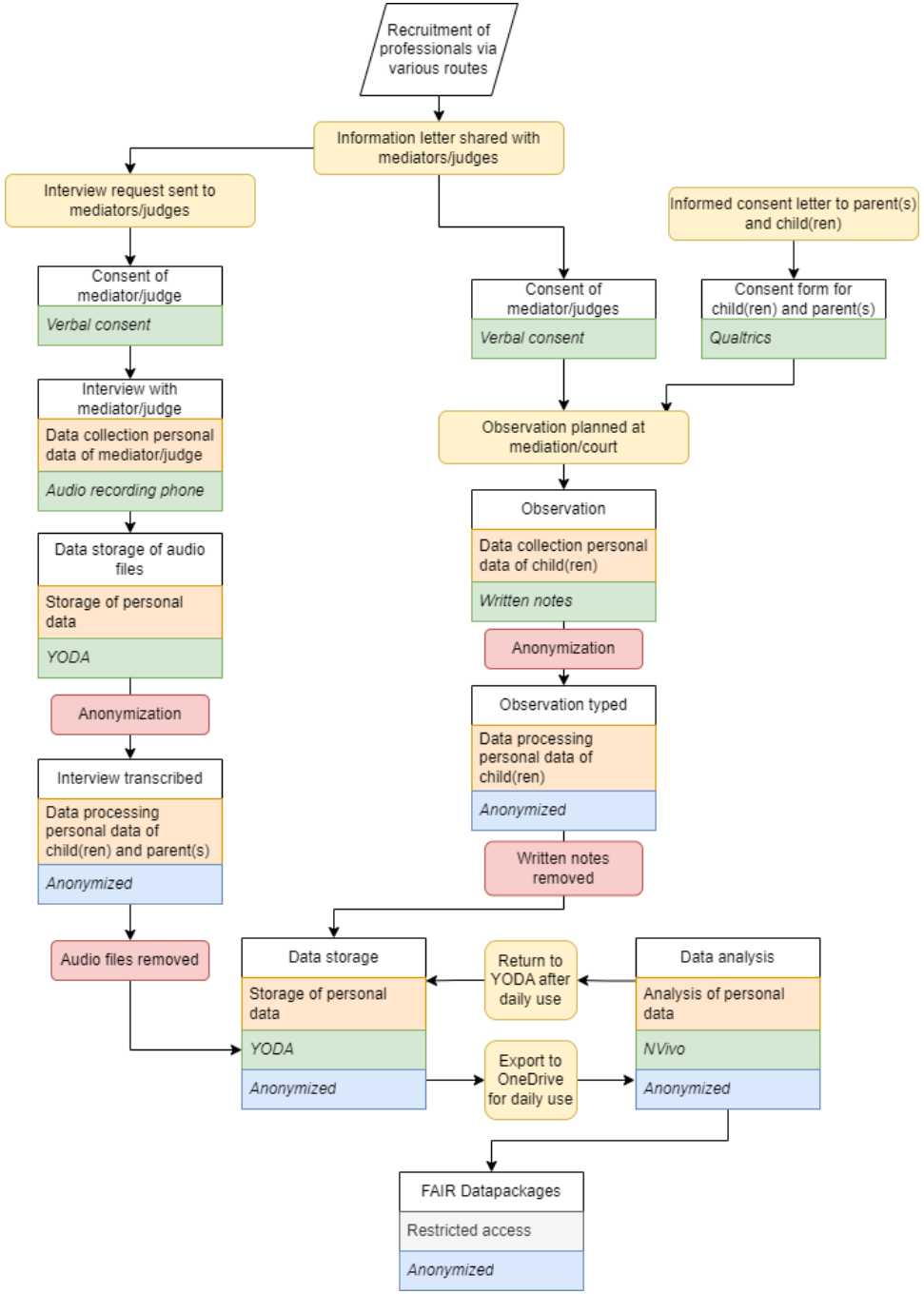
Data Management

by guest speaker Neha Moopen

LET'S GET STARTED!



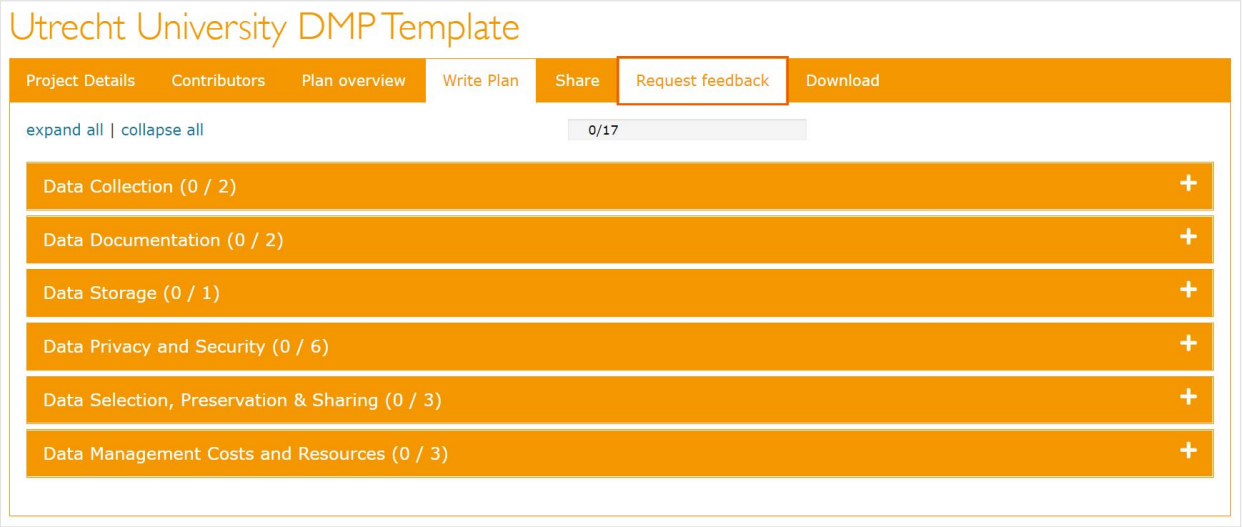
LET'S GET STARTED!



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.

EXERCISE

- Sign into DMPonline with your institutional credentials and create a DMP with the UU template.
- Complete the first page on **Project Details**.

DATA COLLECTION

Describe your data in terms of their:

- *Type* -> the kind of data you're working with
- *Format* -> think of preferred and sustainable formats
- *Volume* -> estimation

EXERCISE

- In your DMP template, type out a quick answer to question 1.2 about describing your data.

DATA DOCUMENTATION

- In order to make your data interoperable & reproducible, describe:
  - the ***documentation*** you will provide -> contextual, human-readable
  - the ***metadata*** you will provide -> structured, machine-readable
  - the ***file & folder structure*** you will utilize
  - the ***naming convention*** you will implement
  - the rules for ***version control*** you will follow

EXERCISE

- Download / make a copy of the documentation checklist available via this link: https://tinyurl.com/documentation-checklist. Complete the checklist as far as possible.

- Reflect on a suitable:
  - *file/folder structure*
  - *naming convention*
  - *version control rule*

- Go back to your DMP and update the Data Documentation section as far as possible.

DATA STORAGE

WHAT DO YOU CONSIDER WHEN CHOOSING A STORAGE SOLUTION?

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

EXERCISE

- Go to the UU Data Storage Finder and see which storage tool might be most suitable for your project.

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

PRIVACY & SECURITY

- Beyond the scope of today's session, but two important things to keep in mind:
  - Refer to the **Data Privacy Handbook**, your go-to resource (apart from the UU website) for all things privacy: https://utrechtuniversity.github.io/dataprivacyhandbook/
  - When in doubt, contact the **Privacy Officer** of your Faculty!

EXERCISE

- Bookmark the Data Privacy Handbook for future reference.
- Find out who is the Privacy Officer at your Faculty and where/how to find them if needed.

SELECTION, PRESERVATION, SHARING

This section has to do with:

- *archiving data* -> long-term preservation at institution
- *publishing data* -> making data FAIR/open in a repository
- *sharing data* -> how to share data in a practical way

EXERCISE

- Go to the UU Data Repository Finder and see which data repository might be most suitable for publishing your project.

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

COSTS & RESOURCES

- Not relevant for today's session, but a couple of things to remember:
  - Dynamics of Youth (DoY) has paid for a dedicated data manager from RDM Support to help DoY researchers with RDM & FAIRification – feel free to reach out to us!
  - The costs for using Yoda are covered by the Faculties at this point.

EXERCISE

- Remember Neha and make a note of where and how to find her for support with data management.

BREAK

Data cleaning

Researcher's perspective

Tips from a researcher

- Use syntax
- Lock raw data
- Work with your codebook
- Use help (if available)
- Take your time

Data cleaning: dataset level

- Merging (or splitting) datasets
- Check participant numbers
- Deleting redundant variables
- Renaming dataset
- Re-sorting variables

Data cleaning: variable level

- Renaming variables
- Adding variable labels
- Adding value labels
- Recoding/reverse coding
- Setting variable type
- (Missing data)

Data cleaning: scale level

- Check missing data
- Response tendencies
- Compute scale scores

Question: When to exclude participants

- Depends on your research question, design, team…
- Exclude participants when you have reason to believe they lower the quality of the dataset

Question: inconsistency across informants/waves?

- What types of data should be "cleaned", e.g., outliers, inconsistency (between different informants or across waves), etc.

Question: How to detect social desirability and regular patterns?

- Social desirability:
  - Administer social desirability scale
  - Check response patterns

- Response patterns:
  - Participants who score same on each item in scale

Example workflow

INTRANSITION live demo

Data cleaning

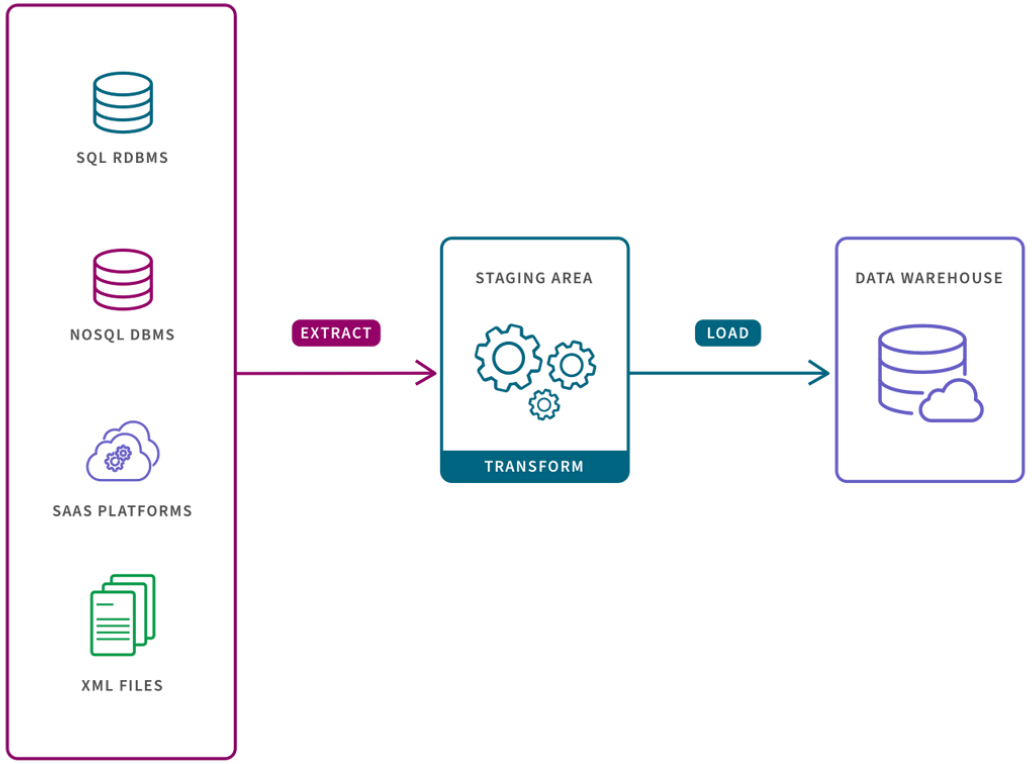
Data manager's perspective

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 تimplement 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.

Questions?

Question: What if a well-developed psychometric scale does not show acceptable features in the current sample?

- Depends how bad
- Some options:
  - Leave as is
  - Item deletion
  - Mean vs. factor analysis

- But mostly: ask others (methodological experts) for advice!

Question: What to include in codebooks?

- See our YouTube channel

Question: How to write a good code book in a big longitudinal project with multiple subjects?

- Include plenty of meta information!
- Examples:
  - Youth study
  - RADAR

Course evaluation

Reflect

What is the main take-away of this course for you?

Evaluation

What is the main take-away of this course for you?

Strengths of the course?

Suggestions for improvement?

Utrecht University

Sharing science,
shaping tomorrow