Team DataScienceAI-2
26.09.2024

Data Storage Protocol

Organizing and documenting data

Reuse, check and verification of data:
In order to make your research data available for reuse, check and verification research data needs to be FAIR (Findable, Accessible, Interoperable and Reusable). Making research data FAIR starts with proper documentation and organizing when storing the data:

[1] Documentation & metadata
[2] Filenaming & folder structure
[3] Version control
[4] Readme files

---

[1] Documentation & metadata

To make data usable for colleagues and other researchers/partners it is important to add documentation to your data. Data documentation describes the characteristics of a dataset. Proper documentation contains the following items:

A description of the data: what is the format (file-type) of the data. What software is needed to use the data
A description of the data collection process: which tools/instruments are used (for example: lab journal, logbook, diary etc.)

Metadata is data about data:
The metadata provides information about data that makes it findable, trackable and (re-)usable, because not only people can understand metadata, also computers can read and interpret the data. For example the metadata can be used to make your research data(set) findable and citable in a data-archive. There are several formats and standards for metadata.

To thoroughly document the data we collected, we have created two key files that describe the characteristics of our dataset. The first is a **codebook** that details all the quantitative variables from the survey. This codebook outlines the steps taken during the quantitative data collection process, as well as the qualitative data collection process, from the initial preparations through to data cleaning. It includes comprehensive information about the type of data collected, the format in which the data will be stored, and the software required to utilize the data. Additionally, the codebook describes the tools and instruments used in the data collection process,

such as the Qualtrics survey platform, Prolific for participant recruitment, Jupyter Notebooks for data cleaning and analysis, and Python for further data manipulation.

Once we receive the data from the Prolific platform, which facilitates survey distribution and data gathering, we will obtain a metadata file containing all the specifications related to the responses collected.

The second file contributing to our documentation efforts is the **Data Management Plan (DMP)**, which also provides a more general overview of the data collection process and the data itself.

By implementing these documentation practices, we aim to enhance the usability of our dataset for colleagues and other researchers, ensuring clarity and accessibility throughout the research process.

---

[2] Filenaming & folder structure

If you want research data to be findable and easy to interpret it is important to store the data in a structured and consistent manner. Elements of a file name can include:

A project acronym
Content description
File type information
Date : (make a choice how to write the date: for example DD-MM-YYYY or YYYY-MM-DD)
Creator name or initials
Version number
Use - or _ to separate elements in a file name

For example:
- 07022020_interview_BUaslibrary_audio.wav
- 07022020_interview_BUaslibrary_transcript.txt

Just as the file names having a meaningful folder structure makes it easier to find files. If you are working with a project group it is even more important to have an organised folder stucture. It helps to restrict the level of folders to 3 or 4 deep and not have to much files in each folder.

To ensure our research data is organized and accessible, our project will utilize a structured folder hierarchy within our group's GitHub repository. We will establish three main folders:

1. **DMP** - This folder will contain all files related to the data management plan.
2. **Data** - This folder will house all collected qualitative and quantitative data.
3. **Research Documents** - This will include various crafted research documents.

Folder Structure and Contents

DMP Folder:

- A **ReadMe.txt** file describing the contents of the folder.
- A **codebook** in either Markdown or Excel format, crucial for naming and detailing the variables early in the data collection process.
- The **Data Management Plan**, crafted using the NWO template.
- An application for a **BUas Ethics Review** for internal review by lecturing staff, submitted in PDF format.
- A **Research Information Letter** for participants, along with a corresponding questionnaire.
- A **Letter of Informed Consent**, presented to participants after the Research Information Letter.
- A **Data Storage Protocol** in PDF format, outlining stewardship procedures.
- A **Privacy and GDPR Checklist** for compliance.
- A **FAIR Checklist** for data findability and accessibility.

Data Folder:

- **Raw Data**: Both quantitative and qualitative datasets.
- A **Data Processing Script** for managing and cleaning the data.
- A **CSV file** of the cleaned survey data.
- **Analyzed Qualitative Data** for further insights.

Research Documents Folder:

1. A **Research Proposal** detailing the research plan, including the research question and study design.
2. A **Policy Paper** providing recommendation for the client based on research insights, including stakeholder analysis and an executive summary.
3. A **Digital Conference Presentation** in the form of an academic research poster, PowerPoint presentation, or an interactive report to convey research results effectively.

File Naming Conventions

For file naming, we will incorporate the following elements:

- Project group name and number
- Content description
- Version number or status
- File type information

Example File Name:
DataScienceAI2_Informed_Consent_Letter_final.pdf

To maintain clarity and consistency, we will use **-** or _ to separate elements or words in a file name. Given that all files relate to the same project and are categorized within corresponding subject folders, we will not include a project acronym in the file names. Additionally, individual names or initials of a file creator will be omitted to reflect the collaborative nature of our work, as multiple team members work on and review each of the files. The date will also not be part of the file names since our GitHub repository will document the dates of commits, ensuring we track changes effectively.

By adhering to this organized structure and naming convention, we aim to enhance the findability and interpretability of our research data.

---

[3] Version control

There are several ways to manage different versions of data / files:

- *Research Drive, OneDrive and Sharepoint have version control facilities. It is possible to see when the file is last modified, by who and get an overview of the history of the file.*
- *Record the date in the file*
- *Include a version number in the filename: for example 07022020_interview_BUaslibrary_transcript_v2.txt*
- *Include information about the status of the file: for example 'draft' or 'final'. But you have to be careful to not use confusing statuses in the filename like: 'final2' or 'final_revised'*

To ensure effective version control of our files and data, we will utilize a private Team GitHub repository named **"2024-25a-fai2-adsai-group-datascienceai-2-2,"** created by Breda University of Applied Sciences for our research team. This platform allows us to track changes efficiently, as each commit provides visibility into when files are added or last modified, who made the changes, and offers a comprehensive overview of the file's history and all its previous versions.

In addition to leveraging GitHub's version control features, we will document the creation date within each file. To further clarify the document's stage in the creation process, we will include either a version number or the document's status, such as **draft** or **final**, at the end of the filename, just before the file type. This approach will help us manage the evolution of our documents.

It’s important that we avoid ambiguous statuses, such as **final2** or **final_revised**, to prevent confusion regarding the document's stage in the creation process. By adhering to these practices, we will maintain clear and organized version control throughout our project.

---

[4] Readme files

Readme files must contain addition information to your raw data collection.
A readme file helps other researchers to interpret / reanalyze your datasets.
Plain text file is the recommended format for a readme text.

In accordance with the provided guidelines, we have created a **README** file that outlines the contents of the DMP folder. This file describes each of the documents contained within the DMP folder and includes functional links leading to the final versions of each file. To enhance readability and visual appeal, the README file is written in Markdown format. This approach ensures that other researchers can easily interpret and reanalyze our datasets while having quick access to relevant documents.