Data Protocol UvA Psychology - background information

Data storage at various points of the experimental research life cycle

Jaap Murre, Department of Psychology, University of Amsterdam

Introduction

This document is the outcome of a discussion between members of the Advisory Scientific Council (WAR) of the Department of Psychology of the University of Amsterdam and other colleagues at that department. The goal was to make an inventory of the type of data produced during normal research practice and to provide guidelines to improve the storage of these data. This document represents the official policy of the Department of Psychology of *Best Research Practice.*

Life cycle of data collection

During a research project data goes through various stages corresponding to the life cycle of a typical psychological experiment:

1. Study design
2. Data collection
3. Data analysis
4. Reporting

Storage goals and requirements evolve with a project. In the WAR, life cycles of several types of experiments were collected: longitudinal studies, reaction times experiments, social-decision making experiments, and EEG/MEG experiments. There are of course many more types but these types already had much in common, so we expect that other types of experiments will also share many storage characteristics. MRI (fMRI, DTI, VBM, etc.) studies may have additional storage requirements, though the MEG study type may already cover most of these.

On the basis of the above input, a synthesis is given here that can be adapted to specific types of experiments. The following scheme is based on input from Heleen Slagter, Carsten de Dreu, Reinout Wiers and Tim Janssen, Bruno Verschuere, and Eric-Jan Wagenmakers.

Four data storage goals

These are four examples of situations that the Department aims to prevent:

1. "Help, my hard disk crashed! Now I cannot get my Ph.D. because all my data was stored there."

2. "I would show you some of my amazing results that were recently published in *Nature* but unfortunately the dog swallowed the USB stick with my data."

3. "What the heck was VAR21 again? And were the women coded as 0s or 1s?"

4. "In response to your request for my data, please, find enclosed a zip file of the directory that contains my data from the 1986 paper. There are 37 folders and over 500 files, some of which may be corrupted or incomplete. I have no doubts that you can reconstruct the intricate system I used to name the files. Enjoy!"

These four examples correspond to four data storage goals. Which goals must be met depends on the guidelines for specific projects but Level 2 is very lowest that meets the Department's requirements. Full compliance with Level 3 is highly advised and is part of the so called *best research practice*. The Department aims to implement most of Level 3 in the coming years.

1. ***Safety.*** Data should be safe and secure. This means they should be **backed-up in at least two physically separate locations** (so that for example a fire cannot destroy both). Check: If your computer crashed right now, you should not lose any data? Backups should themselves be verified as they sometimes fail (i.e., there seems to be a backup procedure but there isn't).

2. ***Accountability.*** Researchers are accountable for their data. They know where it is and how it is organized. When asked for it, they can access it. The **Data Section on the website of the Ethical Committee has been completed**. The Department's research code makes it obligatory that at least **two researchers are accountable for each data set** (even if the research was done alone; in that case: find a 'data buddy'). Experimenters should keep a **lab log of ongoing events** (with dates and times).

3. ***Efficiency*** (and continuity). Data should be well-organized and documented. There should be **clear hierarchy of folders and files.** It is recommended to describe complex file structures in a **readme file** in the root directory. There should be a **code book** that documents variables, methods, analyses. Enough detail should be available to replicate an experiment, simulation, or analysis.

4. ***Sharing.*** Data may be placed in an ***online repository***, such as DANS or OSF (Open Science Framework, see https://osf.io/). This usually involves additional work in terms of uploading the data, possibly converting it into the right format, additional documentation or existing documentation in the repository's format. Access may vary from public (anyone may access) to private (nobody may access). For large data sets, it may be necessary to allow access rights per section of the data (e.g., for tests 1-10 but not tests 11-99; DANS does currently not support this).

Of these levels, each one higher one requires a little more work than the previous one.

At 'level 0', you have data spread over say your hard disk at work, some files on your laptop, a pile of consent forms in an old shopping bag, a used external hard disk that nobody wanted anymore, and then there are those files by that exchange student from Romania who went back and still has to E-mail them to me.

This is the situation that still occurs far too often and that we want to remedy. At level 1, you have made more efforts, for example: a high-quality hard disk on your computer, some form of cloud-backup (encrypted if you don't trust the NAS), perhaps a sturdy external hard disk. You have *all* the data, preferably in electronic format or else well-archived. At level 2, two researchers have looked at the data and know where it is. A lab log has been kept. Some data may have been 'frozen' (see note below). At level 3, thought and effort were spent in organizing and documenting the data. At level 4, additional efforts were put into moving the data into a repository.

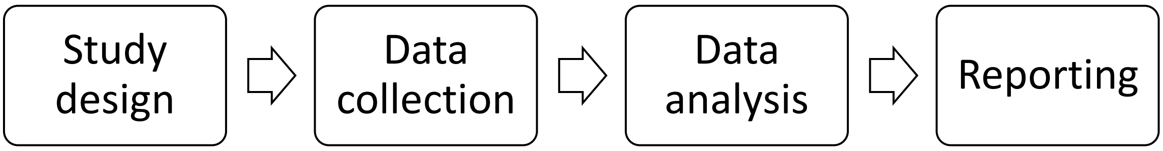
Though it appears that the levels also correspond to the three RDM 'silos' of *individual*, *group*, and *public*, this is not the case. It is the conviction of the WAR that data should be shared by a group at all times, where minimal group size is two researchers (hence, Level 1 above does not suffice). Eventually, data may be published (shared publically [EJ: publicly]), but even if this is not the case, it is possible it will be shared with trusted colleagues, possibly as part of a collaboration.

Table 1. Levels of storage summary (see text for details)

Level	Description	Implementation
1	Safety	Backup (accessible data storage in two physical locations)
2	Accountability	EC Approval (includes EC Data Section, lab log).
3	Efficiency and continuity	Clear directory and file structure, code book (variables, methods, analyses). Enough information to replicate the study.
4	Sharing	Repository (online access, correct formats, accessible description of contents)

Life cycle and workflow

There are different storage needs throughout the life cycle of an experiment, which is roughly as follows:



We will list here an steps to-be-taken at the end of each step. All points mentioned are part of the policy of the Department of Psychology unless marked as recommended. One of the goals of this policy is to store experimental data and results in such a way that the experiment and analyses can be replicated in detail.

Study design (and preliminaries)

1. For the duration of a research project, a **lead scientist** (PhD student, post-doc, tenured faculty) is identified.
2. The lead scientist creates a folder that is accessible to all participating researchers. Currently (early 2014), Dropbox is popular but there many several other options available, such as Google Drive, Microsoft Skydrive, Mozy Sync, etc. Very sensitive data may be encrypted to safeguard against breaches of security (TrueCrypt is free, safe, and easy to use

for this). The Department explicitly recommends using the website of the Open Science Framework (https://osf.io/) for the storage of documents and small data sets.

3. (recommended) A **readme file** is created at the root of the experiment folder. This gives a quick overview of the experiment and how it is documented. In case there is a complex folder structure, it should describe the project's organization in terms of folder hierarchy. Also some principal aspects of the experiment are be listed in a typical readme.txt. For example, a handy folder structured may be as follows:

   ```
   /Root
       readme.txt
       /Ethics
           protocol.pdf
           approval.pdf
           informed_consent.pdf
           information_brochure.pdf
           debriefing_brochure.pdf
           /informed_consent_forms
               form1.pdf
               form2.pdf
               form3.pdf
       /Design
           white_dog_experiment_design.pdf
       /Data
           /Raw
               white_dog_experiment_raw_frozen.zip
           /Cleaned
       /Analyses
       /Report
   ```

   It is recommended to save the readme.txt file in plain text format. If the **readme.txt** is in text format (e.g., with Markdown markup, see appendix), it can easily be read now and in the future and incorporated in other formats (e.g., DANS or other online repository). MS Word seems handy now, but may not be widely readable in 15 years from now (my oldest document files are in an ancient Wordstar format...). Ideally, the readme file is be kept up-to-date throughout the experiment.
   The readme file also contains other properties of the experiment in a concise format. The following list gives an idea of what properties to include:

   - Title of the research project
   - Principal investigator(s)
   - Geographic location(s) where the data were collected
   - Start and end date of data collection
   - Type of sampling (e.g., random sample)
   - Mode of data collection (e.g., lab experiment, survey)
   - Field of study
   - Abstract research report (update at end)
   - Keywords

   It is also worth it, I think, to put some thought into a *folder structure*, because any agreement we can achieve on its file naming and folder hierarchy will make it easier to inspect each other's data, work together in projects, and share data.

4. Prior to data collection, for each independent experiment, the lead scientist archives the following:

   i. **Ethics Protocol and Approval**, including informed consent form, information brochure, and debriefing brochure.
   All of this is already archived on the EC website, in immutable form. We plan to create a PDF export option that covers exports all info about the project from the EC website to PDF format, so that steps 4.i-v will be largely automatic.

ii. Overview of the **study's main hypotheses** (independent and dependent variables).
iii. A description of the **study design and procedure**. For a strict confirmatory design a detailed description must be provided, which can also be submitted to the EC and/or OSF. This also includes exact number of subjects aimed for or stopping criteria.
iv. A description of intended **recruitment** procedures. Exact wording of 'advertisements' for recruiting subjects. Inclusion/exclusion criteria for clinical studies. Known criteria for removing subjects from the data. Payment (credits, money, presents, criteria for receiving payment).
v. A description of all relevant data collection parameters
    1. Known criteria for removing outliers
    2. For EEG/MEG, e.g., sampling rate, filter settings, number of (scalp and external) electrodes. Sometimes this can be done with a data collection configuration file
    3. For MRI (and fMRI, DTI, etc.), appropriate configuration parameters

5. **Data collection**
6. Prior to data analyses, for each independent experiment the lead scientist archives the following materials:

i. The original, **raw data files** as they were collected. The nature of these will vary with the experiment. For MRI the original files can be stored and—if desired—frozen (also see Appendix). This also applies to EEG/MEG raw data (e.g., the Biosemi .bdf files). For behavioral experiments, this applies to the raw behavioral output files (e.g., the Presentation .log files).
ii. A brief report on subject recruitment achieved. Were the intended numbers per condition achieved? Which subjects were excluded? Did payment have to be increased to achieve goals? Did the wording of the 'advertisements' need to be changed?
iii. (recommended) **Informed consent forms** with subject IDs, preferably in scanned PDF format (as opposed to paper in a box).
iv. (recommended) Other subject-related materials, such as payment slips, medical details (if legal to store), etc.
v. The **experiment scripts** used (e.g.,, the Presentation or Eprime task code, with a clear mention of the version of the program that was used) or a PDF for paper-pencil questionnaires. If online survey software is used, it is often possible to export the questions in PDF, XML, or other format (PDF is preferred but XML is better than nothing). Also, it is important to include a description of the exact version and platform of the software used in the presentation (e.g., version 2.14 may have a timing bug that is solved in version 2.15 but reappears in version 2.17). For proprietary software (written expressly for an experiment, e.g., by TOP), a zip file with the software itself may be uploaded with installation instructions and a clear mention of the platform (e.g., Windows XP or higher) and machine (e.g., Esprima PC). It may also be relevant to include the brand and type of the screen used (e.g., with color or masking studies) and audio (e.g., Sennheisser S100 head phone, ...).
vi. (recommended) Intermediate, processed data, suitable for data analysis, e.g., in Excel or SPSS. This is only feasible if the intermediate files are not too large and cannot be easily regenerated with the saved processing scripts.
vii. A **code book** containing a description of all variable names and labels with sufficient detail to understand both the raw and processed data when inspected in the appropriate software (e.g., Excel, SPSS, Brain Voyager, etc.).
viii. File with the **lab log** with entries identified by date and experimenter; subjects identified by ID.

Data analysis

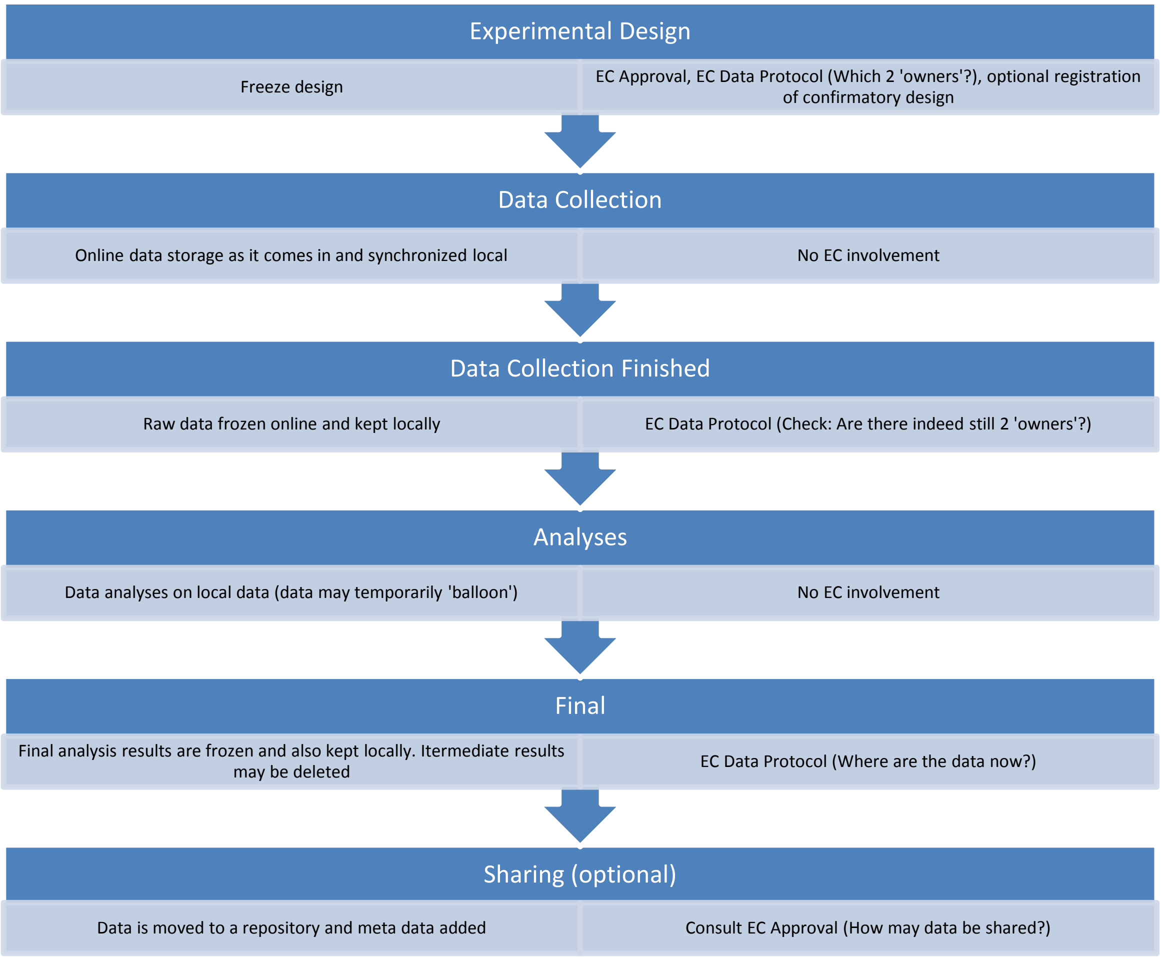
7. Following completion of data analyses, the lead scientist archives the following:

i. All **scripts and syntax-files** used to transform and/or analyze the behavioral data (e.g., Excel files, SPSS .sav and syntax files).
ii. For EEG/MEG: All scripts (syntax-files) used to analyze the EEG/MEG data specifying the preprocessing steps (e.g., recoding of event triggers, filtering, epoching, ICA analysis etc.) and further analysis steps taken (e.g., creating ERPs, time-frequency analyses, etc.), as well as the statistical analyses performed.
iii. A file noting details about the data analysis of specific subjects relevant for replication (e.g., for EEG analyses).
iv. A listing of dropped subjects plus reason for exclusion (e.g, could not keep eyes at fixation, EEG data of poor quality, etc.).
v. The **final preprocessed data** (e.g. file with Excel pivot tables, intermediate files filtered with R scripts or Matlab scripts, or the filtered EEG/MEG data cleaned of all artifacts) that formed the basis for all subsequent analyses. If raw data and preprocessed data are stored, as well as all further analysis scripts, other intermediate preprocessing data sets do not need to be stored as they can be reconstructed. This is also true for MRI analyses.
vi. (recommended) **Statistical output files** (SPSS, etc.). If these files are not too unwieldy, these may be added, for example, to verify that no typos were made in the manuscript. Keep in mind that some people do not have SPSS.
vii. (recommended) As part of the publication process, supplementary data may be placed in the experiment folder, or a full, unabridged report that contains all replications and analyses not reported in the final publication (e.g., because the editor of Science insisted you remove them).

Sharing

8. (recommended) If data is moved into a repository, **meta data** and files must provided in accordance with the repositories demands. If only a selection is shared, mention this fact in the meta data (readme file). Several repositories (e.g., DANS and OSF) have very few official documentation demands. If Level 1-3 guidelines have been followed this should also suffice for Sharing, though it may be a good idea to add an abstract, reference to publications, and additional roadmaps of your data.

Summary Figure. Typical data collection life cycle with to the left the data storage needs and to the right the role of the Ethical Committee (EC) with associated Ethical and Data protocol implemented at UvA Psychology Dept. The EC Data Protocol is obligatory and ensures there are always two responsible researchers and two storage locations of all data.



Experimental Design

Freeze design

EC Approval, EC Data Protocol (Which 2 'owners'?), optional registration of confirmatory design

Data Collection

Online data storage as it comes in and synchronized local

No EC involvement

Data Collection Finished

Raw data frozen online and kept locally

EC Data Protocol (Check: Are there indeed still 2 'owners'?)

Analyses

Data analyses on local data (data may temporarily 'balloon')

No EC involvement

Final

Final analysis results are frozen and also kept locally. Itermediate results may be deleted

EC Data Protocol (Where are the data now?)

Sharing (optional)

Data is moved to a repository and meta data added

Consult EC Approval (How may data be shared?)

Appendix

Note on freezing and time-stamping data

Some researchers attempt to 'freeze' their data by making files read-only. This, however, is useless, because it is quite easy to unset the read-only bit, change the file, and set it again. Also, any timestamps on the file (e.g., as appear in folder listings) can easily be changed with widely available tools. Furthermore, output formats like PDF are not immutable; there are many tools that allow editing of PDFs. If the goal is to make later tampering impossible, all of these efforts are wasted from the perspective of preventing fraude.

There are two ways in which real freezing can be achieved that meets Level 2 standards (if we deem freezing necessary):

1. Put the files in an external, trusted repository that does not allow changes (e.g., at Open Science Framework OSF, KWAW DANS, local repository). Or E-mail the to-be-frozen and time-stamped data via some cloud service such as hotmail, gmail, Yahoo mail, etc. Once it is in someone's inbox, it cannot be changed anymore (as long as it stays there). Research groups can even set-up a special E-mail account for data freezing. You can also send it to yourself, or the EC could set up a special E-mail address for this. A disadvantage of E-mailing is that cloud-services don't live forever and once they stop, you lose your freeze.

2. The option above may become particularly cumbersome for large files, especially the E-mail option. There is, however, a much easier alternative to freezing, namely check sums. A check sum maps the full contents of a file (i.e., all bytes) to a very long string of bytes. The mapping is unique and changing even a single byte in the original file will give rise to a very different check sum string. An example of such a string is:

   49B15A167EEEB9C5456E914222A5CBA702E62565164FC662C56AFE7DF2140CE0

   It is impossible to reconstruct the file from its check sum and it is exceedingly hard to change a file in such a way that the check sum stays the same (the file-to-check sum mapping is of necessity many-to-one).

So, in contrast to freezing files, it is much easier to freeze their check sums instead. You can then E-mailing these; they can be pasted right into an E-mail. Entire folders can be frozen by first zipping them and extracting the check sum. Of course, this will also preclude adding additional files to the frozen folder. Free check sum software can be found for PC and Mac by googling something like "checksum utility".

The work flow for freezing with check sum is:

1. Put all to be frozen files in a directory (folder)
2. Write the current 'freezing' date and time in the readme.txt file (which freezes the date and time of freezing)
3. Zip (or gzip, rar, etc.) the directory
4. Rename the file something like 'white_dog_exp_raw_data.zip'.
5. Extract check sum from file
6. Send E-mail to all co-authors (etc.) with subject: Frozen 'white_dog_exp_raw_data.zip' and with the check sum somewhere in that message. Or better: upload the check sum to a trusted, immutable and time-stamped repository (e.g., with UvA TOP, DANS, OSF, etc.).

If the zipped file is small enough, you can even include that with an E-mail message. It is not sufficient to create a small file with the checksum and add that to a Dropbox folder because it is always possible to change the file and generate a new check sum replacing the old one. Given that many E-mail messages online are currently immutable and time-stamped, such tampering is currently not possible. It is, however, good practice to include the check sum with the original file for later reference.

An even better way is to have an online storage system administered by the Department that automatically freezes and time-stamps data in various stages without additional efforts from the researchers. There are plans to create such a system, so that home-made freezing solutions will not be necessary any more.

Markdown

Markdown is a text format that uses plain text with a few formatting conventions to indicate header structure, lists, links, bold text, etc. The advantage is that markdown files can easily be converted into web pages, MS Word files etc. with the proper formatting (e.g., # Header changes into an MS Word Level 1 Header with # disappearing). See http://en.wikipedia.org/wiki/Markdown for more info, and http://daringfireball.net/projects/markdown/syntax for a full description. Example:

```
# Heading

## Sub-heading

Paragraphs are separated by a blank line.

Text attributes *italic*, **bold**, `monospace`.

A [link](http://example.com).

Items in a list are preceded by bullets (use * or -):

  - apples
  - oranges
  - pears
```