Managing the data during their lifetime

An example from experimental linguistics

Date: 30 June 2021

The cycle of scientific research encompasses the following elements: planning, data collection, data analysis, archiving and (data) publication. The following provides practical advice on each of these steps. Some of this applies in particular to research projects which involve (linguistic) experiments with participants but most of it can also be applied more widely in other types of projects.

*A. **Planning***

1. Folder structure at a storage location

Each project identifies a ***lead researcher*** or ***principal investigator***, who can be anybody from a PhD candidate to a full professor (but not a BA or MA student). The lead researcher creates a few ***folders*** at a suitable storage location that every team member has access to.[1] If personal data is stored, especially if it is sensitive data, then encryption need to be used, e.g. VeraCrypt.[2]

The storage of the project needs to include at least:

- the ***data management plan (DMP)***
- the ***readme.txt*** file, which could be a kind of logbook of the research and in that case should be kept up to date while the project proceeds; this file should be in plain text (for minimal mark-up you can use *markdown*)
- a folder for the ***raw data***

This is how a structure of a small research project could look like:

- *Research X* or *Project X*/ **DMP**
- *Research X* or *Project X*/ **readme.txt**
- *Research X* or *Project X*/**ethics**
  application.pdf
  information_brochure.pdf
  informed_consent.pdf
  approval.pdf
  informed_consent_scans

[1] Choose one of the UvA's cloud services, such as Research Drive, MS Teams, SURFDrive, or OneDrive. For more information see https://aihr.uva.nl/about-aihr/research-data-management/research-data-management-at-aihr-and-uva-figshare.html or https://rsp.uva.nl/en/execute/saving-data/saving-data.html. For advice contact the data steward.

[2] More information on encryption can be found here: https://rsp.uva.nl/en/training/data-management/data-management.html#anker-workshop-on-encryption and https://rdm.uva.nl/en/looking-after/security/security.html

- *Research X* or *Project X*/**design**
  my_experiment.pdf
  create_stimuli.praat
  run_test.eprime
- *Research X* or *Project X*/**stimuli**
  stimulus1.wav
  long_sound1.wav
- *Research X* or *Project X*/**raw data**
  eeg1.bdf
  eeg2.bdf
- *Research X* or *Project X*/**cleaned data**
  eeg1_downsampled.bdf
  eeg2_downsampled.bdf
- *Research X* or *Project X*/**analysis**
  statistics.Rmd
  statistics.pdf
- *Research X* or *Project X*/**publication**
  submission.pdf
  resubmission.pdf
  as_published.pdf
  supplementary_material.pdf

2. Ethics Review for projects with participants

If participants are involved in the research, it is required to make a folder for all the documents related to the proposal for the Ethics Committee and its approval. The ethics application written by the lead researcher may contain an overview of hypotheses, design and detailed analysis methods and settings, the recruitment procedure, the exact number of participants or the stopping criterion (crucial for RDM accountability), criteria for removing participants (before the data have been seen; equally crucial for RDM accountability), criteria for removing outliers (after the data have been seen), participant payment criteria. You will need approval by the Ethics Committee of the Faculty before data collections starts (for pure confirmative experiments).

Applications for an Ethics Review need to be handed in at this portal: https://rr.uva.nl/lab/ethics/pages/edmr_home

More information on the procedure can be found on the website: http://aihr.uva.nl/about-aihr/ethics-committee/ethics-committee.html

B. Data collection

For each research project, all the ***raw data*** files needs to be kept at the chosen storage location during the research project (for example: the .bdf files in an EEG experiment, the sound files in a speech production task, E-prime eye-tracking data, and so on). With this data, it should be possible to replicate the analyses done in the research project. If participants are involved in the research, the data has to be anonymized, or encrypted if anonymization is not possible, to protect the respondents.

Depending on the nature of the research, it is recommended to store specific information such as:

- intermediate data: the derived results will often turn up in a simple tab-separated data file with column headers, which is both human-readable and can be opened easily by R, Excel, Praat or SPSS. Even if the files can easily be generated by the scripts, it is impossible to be sure that the software version used by the researcher will be available 30 years from now, so it is recommended to archive them;
- processing scripts (E-prime scripts, Praat scripts etc.): to keep with the version number of the software (and operating system) you used them with (in the script itself, for instance);
- lab logs which contains experimenters, dates, participants by ID (i.e. anonymous from some point on; the connections to real people can be forgotten)
- consent forms scanned as PDF;
- participant payment receipts (uploading the PDF's of the originals for AC;
- a codebook if the data will be in code, with e.g. "M" for males and "F" for females in the column called "gender". Now, this example might be an easy one, but later confusion will easily arise with less obvious or even arbitrary examples, e.g. codings such as "0" or "1" for the two dialects investigated. A codebook helps to disclose all of this to potential future users of the data.

*C. **Data analysis***

In general, everything that is needed for a future researcher to replicate an analysis with the same or other data should be kept and stored after the analyses have been performed.

For ***derived data*** (data that has in any way been annotated or transformed by the researcher) an explicit description of the transformation process needs to be stored with the dataset. This does not imply that all transformation should be reversible, but it does imply that researchers need to track and show their process in transforming the data. Examples of this would be the generation of variables from items as well as the anonymization of data.

Depending of the nature of the research, the following could be stored:

- analysis scripts such as R scripts that do statistical analysis, Matlab scripts that do EEG and other analyses, Praat scripts that do EEG and other analyses, Excel files (which contain formulas) that do simple computations, SPSS output etc.;
- a list of dropped participants: participants can be dropped on the basis of pre-established criteria such as not reaching criterion on a control task or because of a failed recording;
- the output of statistical analyses: the output of the R scripts that perform the statistical analysis; if R-Markdown is used, the PDF output of the analysis script will contain its own statistical results, as well as a readable version of the explanation that the researcher wrote into the same script; the same goes, mutatis mutandis, for other software;
- unpublished parts of the report: articles are not only written by the authors; also journal editors and reviewers have a say; any data or analyses that the researcher could not publish, can be stored as supplementary material in the folders.

D. Data archiving and publication

It is recommended to use a free long-term repository such as UvA/HvA Figshare (recommended by the UvA and specially designed for UvA-research but not obligatory), DANS EASY or Zenodo. In all cases, the data steward needs to be informed where the data are stored.

A connection between Research Drive and UvA/HvA Figshare will be realized by the end of 2021. This means that data stored on Research Drive during the project can be easily transferred to Figshare for archiving and publishing. Connections with other archiving options are planned for the future.

Who has access to data?

Published data:
UvA/HvA Figshare (and other archives) allow the researcher to publish data online. Publishing data of finished projects of which results have been made public should always be the rule. Not publishing the data of published research should be the exception because the UvA's RDM policy aims for openness whenever possible. Published data can be available for all others to access and use, depending on the access granted by the researcher who publishes the data. Note that personal data can only be published if completely anonymized.

Publication of data means that a permanent identifier is attached to the dataset and that it is made publicly available for anyone to access and use it (within the limits of the chosen license). It is recommended to add a dataset specifically for publication to the project, in which all identifying information has been deleted and no intellectual property beyond the scope of the related published work is discernible. All authors and owners of the data should be made aware of and have signed off on publication of the data. All researchers should be aware that publication of the data is irreversible.

Unpublished data:
The researcher or lead researcher in a research group needs to give consent for access to unpublished data stored for a research project. The researcher (or lead researcher of the research project) is the first contact for access to data. For data archived on UvA/HvA Figshare, the data steward also has access to the data, but he or she will not use this opportunity without the consent of the researcher unless persistent unavailability of a researcher requires this, or in case of suspected fraud or unethical behavior.