UNIVERSITEIT VAN AMSTERDAM

Faculteit der Geesteswetenschappen

Research Data Management Protocol FGw

Datum 2017 september 13

Introduction

The Faculty of Humanities of the University of Amsterdam endorses the guidelines for Research Data Management (RDM) as established by the Executive Board of the University on December 15, 2014. All research data in the Faculty of Humanities shall be – in order of importance – **secure, compliant, reusable** and (where possible) **shareable**. The faculty strives to implement these four guiding principles fully before the year 2020.

The UvA guidelines require that every research institute creates a research data protocol in accordance with the UvA-wide policy on research data management. The present document serves as the FGw-wide implementation of the UvA-wide RDM requirements, supplementing the more general information on the UvA RDM website: rdm.uva.nl. The present document assumes that you are familiar with this UvA RDM website; if you are not familiar with this website, the present document will generate in your mind all kinds of questions about the background, purposes and concept of RDM, which are answered on the UvA RDM website.

This document starts with definitions, followed by a description of the responsibilities for data management of all people involved with research in the Faculty of Humanities, and concluded by a protocol for the management of data during its lifetime.

Definitions

A. Data

The term "data" in the RDM sense refers to any kind of material that is collected for or provides the basis for analysis in research projects and that can be stored in digital form. Data can either be in **raw form** (as collected) or in the form of **derived data**.

Applied to the FGw, the term *raw data* can refer to:

- texts that you recorded in your research or that you extracted from archives not (easily) accessible for research (as opposed to texts that you used from readily accessible bibliographically available sources or archives open to the public or researchers);
- photographs that you took for or during your research;
- videos that you recorded for or during your research;
- audio that you recorded for or during your research;
- 2D photographs and 3D scans created for or during your research;
- questionnaires that the participants in your research filled in;
- physiological recordings that you made for your research (e.g. EEG, MEG, TMS, fMRI);
- raw behavioral data that you recorded for your research (e.g. automated eye-tracking);
- structured data collected during your research.

The term *derived data* can refer to:

- formal annotations (perhaps as opposed to mnemonic notes) that you made to texts;
- video annotations that you made e.g. with ELAN;
- audio annotations that you made e.g. with Praat;
- selections from and adaptations of existing datasets;
- filtered and downsampled versions of your physiological recordings;
- behavioral data that you created by annotating a video (e.g. head movements).

In the humanities, the border between what counts as "data" (and therefore will have to be stored, organized, protected and perhaps made public) and what doesn't is subject to debate. Do notes scribbled in the margin of a book count as intermediate data about that book? Perhaps if counts have to be based on such notes...?

B. Data-owner(ship)

Usually, research data will be owned both by the researcher or group of researchers involved in the research project that collects or creates the data, and by the university whose name is used in the project's publications.

C. Data Steward

The *faculty data steward* is responsible on behalf of the AIHR for the practical management of data storage, information dispersion among research staff on RDM, and the management of access to the data. The faculty data steward is the contact for the researcher(s) in charge of the data management of their research group. He or she needs to have accurate knowledge of the protocol and should be able to advise researchers on data management

issues. He or she manages the research data management storage system on behalf of the researchers with regard to monitoring the use of the system and reporting on the use of the system.

The data steward's tasks are:

- making researchers and research groups aware of the RDM system and new developments and changes;
- leading discussions within the institute on its RDM protocol and RDMP template (see below under D), and initiating and supporting new policies about RDM;
- answering questions from researchers on the protocol;
- collecting questions and ideas from researchers and schools about the use of the system and linking it back to the administrator of the system (University Library);
- ensuring, once a year, that all publications in the yearly report have a registered RDMP in the system, with data if applicable[1];
- evaluating and reporting on the use by the researchers of the RDM system;
- supervising the storage of data, the quality of data management in the schools and the selection or destruction of data;
- assessing the validity of requests for extra storage space by researchers, and supplying that extra storage space;
- maintaining professional contact with officials of the faculty with responsibilities concerning RDM;
- maintaining professional contact with data service providers of the University Library (UB).

D. Research data management Plan (RDMP)

A research data management plan (RDMP) is a digital document in which the researcher or research group describes the workflow of the research project, the data that will be collected, the storage and management of the data, the usage of the data during and after the runtime of the research (project), and applicable metadata. It contains – in case it is relevant – the approval documents from the Ethics Committee[2].
Typical things to include in the RDMP are:

- overview of the experiment
- overview of the documentation, e.g. where the codebook(s) are
- the structure and contents of the files
- project data: title, names of researchers, location(s), time interval, sampling method, discipline, titles and abstract(s) of publication(s), keywords.

A RDMP should be written, assessed by and stored by the researcher or (the principal investigator of) the research group **before** data collection commences. RDM does not only concern the preservation of digital data, but also the way data are generated or measured. By the same token, analogue data (physical samples) and their relation with digital data should be characterized in the RDMP.

For each research (project), an RDMP needs to be written in which the researcher or research group states that the data to be stored will be in line with the qualities described by

[1] This can be realized fully only after 2020, because the implementation of RDM starting in 2017 is a work in progress.
[2] See the website of the AIHR: http://aihr.uva.nl/about-aihr/ethics-committee/ethics-committee.html

the UvA policy on RDM.

The UvA provides a generic template for the Data Management Plan on the RDM website: rdm.uva.nl/plannen/datamanagementplan.html. The filled-in template is then the data management plan. If this generic format for all types of research projects will turn out not to be feasible, the AIHR will develop a set of RDMPs that covers the complete spectrum of research activities of all schools of the Faculty. Templates that satisfy the requirements of research funders such as NWO for example or EU can be found on the RDM website.

E. Availability of data

In accordance with UvA policy data needs to be stored for a minimum of:

- 10 years for a research project resulting in a publication;
- 10 years for raw data associated with any project;
- indefinitely for research projects involving a PhD project.

When the researcher responsible for the collection of the data leaves the UvA, he keeps access to his data for 90 days after his departure, as is the case for his email account. After this grace period of 90 days, the data steward assumes the access to the data.

Responsibilities related to RDM

Both the Amsterdam Institute for Humanities research (AIHR) and the individual researcher bear responsibility for the management of research data.

First and foremost, the AIHR is responsible for providing and managing the resources needed for adequate data storage. In that sense, the research institute should provide researchers with an infrastructure that allows them to store data in accordance with this protocol. The institute is responsible for ensuring that this storage is safe (protected from threats, such as theft and technological malfunction) and managed properly by a data steward. Furthermore, the research institute is responsible for the distribution of the protocol among research-conducting employees.

The researcher as a member of a research school of the AIHR is responsible for proper data management of his or her research and therefore the practical adherence to this protocol. Proper data management of research has four goals, which are hierarchically arranged (from 1 to 4):

1. data should be kept safe and secure: the researcher should move his or her data as soon as possible to the UvA-wide service for archiving described below; if this is not immediately possible, the data should be backed up as soon as it has been generated, and temporarily kept in at least two physically separate locations until archiving is possible;
2. researchers are accountable for their data towards the scientific community and towards society, who may request to know where the data is, how it is organized and how it can be accessed;
3. data should be organized in a way that it is easy to retrieve and still to be understood after a few years by the researcher himself and other future researchers; the researcher should therefore describe this organization in a Research Data Management Plan (RDMP);
4. data should be shared as widely as possible, within the boundaries of generally acknowledged limitations of confidentiality and ethics; the site where the researcher stores his or her data for the long term should have an explicit long-term archiving function.

Even in cases where goal 4 cannot be achieved, perhaps because opening access to the data would violate privacy principles or copyright laws, it is still important to achieve goals 1 to 3. Thus, even if the data are copyrighted or sensitive, the researcher is still accountable for them, and organizing them is still important.[3]

[3] For instance, if your data are digitized works of art, then you can store the digitizations in a non-open archive. Or if your data are interviews with dissidents who live under a dictatorial regime, then you may even more rigorously want to organize those data neatly in a safe place. Video and audio recordings are in principle copyrighted property of the people who are seen and heard in these recordings. You can achieve goal 4 by asking the participants to transfer their copyright to the public domain. They would have to sign a statement to that effect, as is already common practice in the FGw. The condition of the participants' signature on the copyright transfer agreement is already enforced by the Ethics Committee of the FGw: if no signature, then no public use.

The UvA RDM organization has developed an UvA-wide service for archiving: Figshare (http://figshare.uva.nl) which has, in principle, unlimited storage space. For longer-term storage, KNAW's DANS (http://www.dans.knaw.nl) can (also) be used, although there are restrictions on file formats, which are very often impractical (they can – for instance – not store EEG files).

So the researcher is responsible for a research data management plan being present for each research project. The researcher is also responsible, during the storage of data at the UvA, for providing the data steward with accurate contact information so that he or she can be contacted when necessary. Finally, the researcher is responsible for the notification of third parties of this protocol and the associated data storage.

Managing the data during their lifetime

The cycle of scientific research encompasses the following elements: planning, data collection, data analysis, archiving and (data) publication.

A. Planning

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). A project that meets the requirements that make storage necessary needs to be generated in FigShare (or in another storage system). Please note that if the project is not generated in Figshare, the faculty data steward needs to be added as viewing collaborator to this project in the chosen system; for Figshare this happens automatically.

The lead researcher creates a few ***folders*** that every team member has access to[4]. If the data is especially sensitive, then it is recommended to use encryption, e.g. VeraCrypt.[5]

The storage of the project needs to include at least (for details see Appendix 1):

- the ***RDMP***
- 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***

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). The lead researcher has to follow the procedure as described 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 and stored in Figshare in accordance with the availability principles mentioned above (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 can be anonymized to protect the respondents.

[4] The UvA RDM organization is planning to provide an UvA-wide service for groups (connected to back-up facilities and archiving functions)

[5] The UvA RDM organization (University Library) will organize workshops on encryption.

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 in Figshare 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 Figshare (recommended by the UvA and specially designed for UvA-research but not obligatory), DANS or OSF (Open

Science Foundation). In all cases, the data steward needs to be informed where the data are stored. Options for storage should be considered in the following order of preference:

- shared storage on an (inter)national community basis,
- university storage (**Figshare**),
- shared storage on a local community basis,
- stand-alone storage.

Who has access to data?

Published data:
Figshare allows 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.

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 Figshare, 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. 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.

What will ultimately happen to the data?

In principle, the faculty's data steward may delete the data after 10 years, but only after consent by the researcher.

APPENDIX 1

The structure of a small research project generated in Figshare *could* look like this (Figshare allows only one level of folders):

- *Research X* or *Project X*/ **RDMP**
- *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
- *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