Institute for Logic, Language and Computation

A Manual for Managing Research Data

Peter van Ormondt

February 4, 2022

Contents

Introduction

Research data management (RDM) concerns the organization of data, from its entry into the research cycle through to the analysis of the data, the dissemination of the results and the archiving of the necessary and sufficient information by which these results were achieved. If set up in an intelligible fashion, it can help to make the research process more efficient. RDM is an integral part of the research process[1] and aims to meet expectations and requirements of the university, research funders, and legislation.

1 Before start of research

1.1 Do I need a Data Management Plan?

All research intended to generate results that will be or can be used in the research process and hence in the research community, must have a proper Data Management Plan (DMP) in place. This means that all research, whether funded by the university, or funded by second or third flow of funds[2], are subjected to at least basic RDM. The same holds for student internships and PhD research. RDM can be implemented at the level of an experiment or a coherent piece of research (= project). At which level you implement RDM is completely up to you and can be different for each part of your research.

In case you are a PhD candidate in the ILLC PhD programme you will need to submit a DMP together with your Training and Supervision Plan (TSP). See the website of the PhD Programme for more information. Note that the

[1] Cf. ILLC's Code of Scientific Integrity, section 3.1.

[2] The revenue of the Dutch universities can be roughly divided into three flows of funds. Cf. https://vsnu.nl/en_GB/funding-of-universities.html

Amsterdam Institute for Humanities Research (AIHMR) offers skill courses in RDM.

The best way to implement RDM is constantly during your research. In fact it is encouraged that once a DMP is set up a log of decisions is kept during the course of the project.

A DMP is required for all research projects that are concluded with a publication. What a publication is in this respect depends a bit on the particular research area. In some areas a conference paper is a first step to a journal paper that concludes a project. In other areas the concluding publication is a conference paper.

Publications that typically require a DMP are

- A journal article
- A PhD thesis
- A conference paper
- A manuscript

But it is important to note that it is the responsibility of the PI to judge which publication requires a DMP and what the DMP looks like.

2 Writing a Data Management Plan

A *data management plan* (DMP) consists of

1. Titlepage (Section 2.1),
2. A description of what data you will be collecting during your research project (Section 2.2),
3. A description of how the data will be stored and managed (Section 2.3),
4. A description of what will happen with the data after the project is concluded (Section 2.6).

See this document for a template provided by the University of Amsterdam.

DMPonline is an online tool for researchers to easily write, edit, share and store data management plans. It also offers several templates (some created by UvA researchers).[3]

[3] Cf. https://rdm.uva.nl/en/planning/dmponline/dmponline.html for more information.

The digital format of the document does not matter much and can be `docx`, `odt`, `txt`, LaTeX, etc.

2.1 Titlepage

Create a title page and at least provide

1. Name and email address of the PI,
2. Title of the project,
3. Planned start and end date of the project,
4. Abstract.

2.2 Description of the research data

In the DMP you will have to give a description of the data you intend to collect or generate. The idea is to give a description of the data in such a way that the purpose of the data and why it was collected is clear, also in the future. It is possible that this part of the DMP changes during the course of the research project. In fact it is encouraged to keep a log with decisions concerning the data and the dataset for future reference. A description of the data will consist of the following elements

- Provide context for the data by submitting the purpose of the research project and the methodologies used,
- Describe how the data will be collected or generated,
- Give details about the file formats in which the data will be stored.[4]
- In case copyright or intellectual property issues are relevant for the project, these should be specified in the DMP.

2.3 Description of storage and management of the data

Describe where you will store your data during the research. Depending on the type of the data and size of the dataset there are several options available. See Section 3.1 for more information.

See Section 2.5 in case you will be collecting personal data. Also see Section 3.2 on data security.

[4] Cf. `https://rdm.uva.nl/en/looking-after/file-formats/file-formats.html` for more information.

2.4 Responsibility and maintainer

Each researcher in ILLC is responsible for a proper RDM during his or her research. The overall responsibility on RDM lies with the Principle Investigator. In case of a PhD project the responsibility lies with the supervisor.

The ILLC Data Steward can assist and advice throughout all phases of RDM (drawing up Data Management Plans, metadata files, setting up archives).

2.5 Privacy and ethics

In some areas of research at ILLC personal data is used. Researchers are required by law to manage this information carefully. Personal data encompasses all details that either directly or indirectly lead to a specific person, for example, someone's name, occupation or age. This also includes the so called 'sensitive' personal data, which is data concerning someone's religion or beliefs, race, etc.

You must draw up a processing agreement if you have the details processed by a third party, for example, when you use an online application for surveys. Legal Affairs can help with drawing up a processing agreement.

By law every person has the right to protection of their privacy. When collecting personal data for research purposes you are therefore responsible for protecting the privacy of your participants. See this page for more information on the legal requirements regarding personal data.

See Sections 3.1 and 3.2 for more information about storage and encryption.

2.6 Description of the data after the conclusion of the project

How will the data be archived and shared? Who will be responsible for the data after the conclusion of the project? Note that it should anticipate that people may no longer be employed by the UvA.

3 During research

3.1 Data storage

For data storage there exist two possibilities ILLC recommends:

- The cloud storage service of SURFdrive offers you a secure option for saving and sharing personal files

- SURF Research Drive. Research Drive is a cloud-based shared-storage environment made available by SURF.

If the dataset is the only copy make sure to make backups, preferably, to a different network.

3.2 Data security

If the dataset contains sensitive materials, e.g., personal data, it is recommended to encrypt the dataset upon storage. If you are using an UvA device it is in fact required to encrypt this device, cf. the the security guidelines of UvA's ICTS department.

See this article for a general discussion of *data-at-rest encryption* and for an outline of the different concepts of encryption ranging from encrypting single files to encrypting complete devices including the boot loader.

There exist many ways for encrypting files, directories, file systems and complete operating systems. It depends in part on your platform on which you work (e.g., Linux, Mac or Windows) and the nature of the data what would be the best solution.

In any case it is recommended to look into Gnu Privacy Guard and to have a look at this page on security.

4 After research

4.1 Archiving

After the completion of your research, you are obliged to keep the raw research data for a suitable period and to make them available to other researchers upon request. You can do this yourself or you can elect to deposit your data in a data archive or repository.

Archiving means that you have to decide what data and metadata you need and want to store for at least 10 years[5] in order to comply with the objectives of good RDM. Archiving means making inventory and throwing irrelevant data and metadata away.

Because computational and experimental data can be very large, it would be too costly to hold all data on spinning disks during the archiving period. Therefore, in some cases it is advisable to transport the data to tape.

[5] In compliance with the Netherlands Code of Conduct for Research Integrity, UvA and AUAS request researchers to keep their raw research data for a suitable period. A commonly used period is a minimum of ten years.

4.2 Choosing a repository

Choosing a repository will most likely also determine how you will publish your dataset and how people will be able to access it. See Section 4.5.

4.2.1 UvA/HvA Figshare

The preferred repository of the UvA to store research data is *UvA/HvA Figshare*. All researchers at UvA have access to Figshare: a new system for safely storing, controlled sharing and publication of research data.

See this page for a description.

4.2.2 Zenodo

The OpenAIRE project, in the vanguard of the open access and open data movements in Europe was commissioned by the EC to support their nascent Open Data policy by providing a catch-all repository for EC funded research. CERN, an OpenAIRE partner and pioneer in open source, open access and open data, provided this capability and Zenodo was launched in May 2013.

In support of its research programme CERN has developed tools for Big Data management and extended Digital Library capabilities for Open Data. Through Zenodo these Big Science tools could be effectively shared with the long-tail of research.

The ILLC has a community on Zenodo and it is the preferred repository for the European Research Council.

4.2.3 Git server

Code is instrumental in the analysis and generation of datasets and should therefore be made public. Many proprietary git servers exist, i.e., `https://github.com/`, `https://gitlab.com/`. We try to collect all git repos of ILLC research in one place at `https://illc-uva.github.io/`. Please consider sharing you repo by sending an email to rdm-illc@uva.nl.

4.3 Registering your data collection in PURE

When you have created a dataset it is important for ILLC's research reports to register the dataset in the PURE database. PURE is UvA's database where all research output is registered. Moreover this system is the source for repositories such as UvA-DARE.

4.4 Publishing your DMP on the ILLC website

When your research project is completed, your data archived and your DMP is fixed, it is time to publish the DMP on the ILLC website. You can can submit the DMP at

https://www.illc.uva.nl/Research/Publications[6]

After it is checked by ILLC's data steward the DMP is published here

https://www.illc.uva.nl/Research/Publications[7]

Each DMP will be provided with a digital object identifier (DOI) so research will be able to link to it in a persistent manner.

Note that any additional material, e.g., technical reports, proofs, that are relevant for research projects can be submitted to ILLC's X series technical reports for online publication.

Information on how to submit ILLC reports may be found on the ILLC Prepublication Procedure page.

4.5 Sharing / making available

The copyright of the data is formulated in the Data Management Plan. In the Data Management Plan also possible restrictions on the data are mentioned. If a request arrives for data of a research project, the holder of the copyright or, if nothing is defined in this, the PI decides about the disposition of the data to the third party.

Publication of the data can be achieved via UvA/HvA Figshare. This way it will receive a DOI, it will be properly archived and it will be indexed easily by search engines.

Points of attention are

Anonymisation Since it is not allowed to publish personal data, datasets with personal data must be anonymised.

Description Data of which it is not clear when and how they are collected, are useless to others (and to your future self). That is why a dataset must be described.

Metadata In order to make sure that your dataset will appear as a search result, you must assign information to the dataset which can be read by search engines: metadata.

[6] This series and submission system are still under construction.
[7] This series and submission system are still under construction.

License By providing your dataset with a licence, it is made clear to others what they are or are not allowed to do with your dataset.

Identifier If you take care that your dataset is assigned a *persistent identifier*, others can simply and correctly refer to the (correct version of) your data.

Embargo Most data repositories allow you to place a (temporary) embargo on your data : during the embargo period the description of the dataset is published, but the data themselves are not available for reuse by others.

See this page for more on Sharing and publishing research data.

5 Procedure for people leaving the project and/or ILLC

If a researcher leaves the study and/or ILLC, the PI will appoint a researcher that will inherit the responsibility for the management of the research data that was produced by the departing researcher. All relevant research data and metadata belonging to this research data should be stored for at least 10 years after the date the research is presented. This means that after publication, after the presentation of a report of a student internship, after the defense of a dissertation the research data and metadata has to be curated, and moved from the environment in which the ongoing research is performed to a data archive.

If the PI is leaving the project and/or ILLC, the PI will, in consultation with ILLC's research director and ILLC's data steward, relay the responsibility of the data management to another ILLC member.

In some case, if researchers leaving the ILLC and UvA want to take data with them, this should be discussed with the director of the ILLC.

6 Using my RDM plan for grant applications

Funding agencies may have other or additional requirements with respect to research data management.

6.1 NWO

The Dutch Research Council (NWO) says on its page for RDM

> Responsible research data management is an essential component of good research practice. In addition to being safely stored and carefully curated, research data should be made available for

reuse as widely and as early as possible. The guiding principle in this respect is 'as open as possible, as closed as necessary.'

Every grant proposal will requires an answer to the following questions:

1. Will you be collecting new data?
2. How will the data be stored during the research?
3. What happens with the data when the research has been concluded?
4. Do you have everything needed for the above?

   If the proposal is granted the research will need to submit a DMP. If this is approved, the project can start. The UvA-template can be used for this as this template was approved by NWO.

6.2 European Research Council (ERC)

7 ILLC RDM policy and other UvA organisations

- The RDM website of the University of Amsterdam
- The RDM website of the Faculty of Science
- The RDM protocol of the Faculty of the Humanities
- The RDM protocol of the KdVI of mathematics
- The RDM protocol of the Informatics Institute

8 Contact

- Contact UvA's central support desk if you have questions about RDM. RDM Support is backed by an organisation-wide network of data experts. RDM Support is managed by the UvA/AUAS Library.
- Contact rdm-illc@uva.nl in case you have questions about ILLC's RDM policy, help with writing your Data Management Plan or to get in contact with ILLC's data steward.