Guide to develop a data management plan for doctoral students

(Doc.CO22/25) (4 RDM\Plans Gesto Dades\PlandeGestiodeDatos_Doctorandos-EN.docx, 23.11.22)

This document intends to assist doctoral students in the development of their Data Management Plans (DMP).

This document was prepared by the CSUC Working Group on Research Support, which is composed of representatives from the following universities: University of Barcelona, Autonomous University of Barcelona, Polytechnic University of Catalonia, Pompeu Fabra University, University of Girona, University of Lleida, Rovira i Virgili University, Open University of Catalonia, University of Vic, Central University of Catalonia, Ramon Llull University and University of the Balearic Islands.

Digital version: http://hdl.handle.net/2072/370509

NOTE: Point 6. Sensitive/personal data has been prepared by the University of Lleida and is necessary for the processing of data from UdL doctoral theses.

About your research

Name and email address	
Thesis director/s	
Working title for the thesis	
Describe your research	*Approximately 50 words*
Duration of your research	Start date: *DD-MM-YYYY* End date: *DD-MM-YYYY*
Linked project	*Is this a thesis related to a project? Which one?*
Funding	*Have you received funding to complete your dissertation? Whose?*

About this data management plan

Creation date	*DD-MM-YYYY*
Last update	*DD-MM-YYYY*
Version and date	*Make a new version every time there are significant changes (new datasets, significant changes in your research, or other factors)*

Sensitive/personal data	*If you work with personal or sensitive data, you have a legal obligation to process it according to applicable regulations. Personal data is any information that allows a person to be identified (name, address, location, etc.).* ☐ I'm not working with personal data ☐ I will work with personal data [see point 2]

1. Data collection

Describe the data you will create/collect

1.1 Will you use existing data during your search? If not, indicate the origin of the data you are going to use

☐ No

☐ Your own data or data from the research group in which you participate

☐ Academic collaborators

☐ Commercial collaborators

☐ Publicly available databases/files

☐ Commercial data providers

☐ Others (indicate which ones):

1.2 Data Description

Describe the data that you are going to create or the third-party data that you are going to reuse and specify:
- *if you are going to use protocols or standards used in your research area*
- *the tools, instruments, equipment, hardware, or software you are going to use*

If you reuse data from third parties, make sure to have the appropriate permissions and to be aware of the terms and conditions of the data.

1.3 Data type and formats

Keep in mind:
- *the type of data: for example, if you are going to work with measurements, simulations, observations, text (text, MS Word), images, audio-visuals or samples, statistics (spreadsheets), with computational models, with data from a qualitative survey (questionnaires), recordings (audio, video), software (code), etc.*
- *the longevity of the file formats: preferably use open standards so that the data can be read by multiple programs, facilitating preservation, and sharing with other users.*

1.4 Specify the data volume

❑ < 10 GB

❑ 10-30 GB

❑ 30-50 GB

❑ 50 GB-250 GB

❑ 250 GB-500 GB

❑ 500 GB-2 TB

❑ 2 TB

2. Data storage and security

Make sure that all your research data is securely stored and backed up regularly.

2.1 Specify any restrictions (commercial, ethical or confidentiality) that may affect your data

☐ Contractual obligations

☐ Legal obligations: protection of personal data (LOPDGDD, RGPD...) [see 4.1]

☐ Legal obligations: copyright, intellectual property [see 4.1]

☐ Ethical restrictions [see 4.1]

☐ Commercial aspects (p. ex. patentability)

☐ Formal security standards

☐ No obligations

☐ Other, specify:

Briefly explain the restrictions

2.2 Major data security risks

Identify the main risks, such as: accidental deletion of data, loss, or theft of data. Describe the consequences of potential data loss.

2.3 Measures to be taken to reduce the risk of data loss.

☐ Access restrictions

☐ Encryption

☐ Data processing

☐ Pseudonymization

☐ Anonymization

☐ Regular backups

☐ Other, specify:

Also specify the procedures you will use to guarantee the privacy of personal data.

2.4 Where will you store your data?

In case of processing and storing personal data, answer the point: 6.6 Where will you store the data in case of processing personal data.

☐ In the network of your department or research group

☐ In the university network

☐ Physical storage (e.g., USB, external hard drive)

☐ Cloud service (e.g., Drobox)

☐ Other, specify:

Briefly explain the storage and copying conditions.

3. Data documentation

Document data to facilitate comprehension and reuse

3.1 Name and structure of the files and the folders

Describe how your data files and folders will be organized and named

3.2 Version control

Describe how you will control the versions. Also, specify what you will do if you delete data.

☐ No version control (e.g., original files are overwritten)

☐ Software with version control, indicate it:

☐ Software with change tracking option

☐ Version number and date in the file or folder name

☐ Making a copy of the script in which the data is processed

☐ Other, specify:

3.3 Which metadata standards do you intend to use?

See "Disciplinary metadata standards" of DCC or "Metadata standards" at Wikipedia.

☐ I will not use any standard (specify the metadata needed to understand the data)

☐ Generic metadata schema (e.g., Dublin Core)

☐ Windows automatic metadata schema (e.g., from Word, Excel)

☐ Specialized thematic metadata schema, indicates:

☐ Another metadata schema, states:

Specify how they will be created (in a "readme" file, in a spreadsheet, embedded in the data) and what documentation you will produce to make the data understandable to others

4. Access, share and reuse the data

4.1 Do you have any restrictions on data sharing as regards the existing regulation (General Data Protection Regulations) or others (ethics, commercial, security, intellectual property, or copyright)?

Specify which ones.

4.2 Who are the potential users of your data and how are they going to find them?

Briefly describe who might be interested in your research and how you will distribute it (e.g. data repositories, website, conference publications, etc.)

4.3 Specify the licenses that you will apply to the data to enable maximum reuse

The use of Creative Commons licenses is recommended (CC - BY o CC Zero) or GNU

5. Deposit and conservation of the data

Keep in mind that all favourably evaluated theses must be published in the respective open access institutional repositories.

5.1 What criteria will you use when selecting the data for long-term preservation?

☐ Type of data (raw, processed) and ease of generation

☐ Relevance of content to others

☐ Ease of reuse of the formed by others

☐ Data linked to a post

☐ Investigation Verification

☐ Time available

☐ Available financial resources

☐ Others, specify:

5.2 How long do you intend to preserve the data?

Various international standards recommend a minimum of 10 years.

5.3 In which repository will you store your data?

Consider specific requirements in terms of format, metadata, size, cost, etc., that the repository may have to deposit data.

The repository for depositing data at the UdL is CORA-Repository of Research Data. It is a federated and multidisciplinary data repository that allows Catalan universities, CERCA research centers and other entities that conduct research to publish sets of research data in FAIR mode and following the EOSC guidelines.

☐ Institutional repository

☐ Thematic repository (international), specify:

☐ Multidisciplinary repository (e.g., Zenodo, Figshare, Dryad)

☐ Others, specify:

6. Sensitive/personal data

6.1 Will you work with personal data?

A personal data is that relating to an identified or identifiable natural person. In this sense, it can be considered data such as a person's name or ID that would allow them to be identified directly, as well as a set of data about a person that, despite not knowing their name, allows them to be identified.

For example, an IP address or an email address are considered personal data or knowing the date of birth, etc.

In case of working with specific personal data with which data.

□ No

□ Yes, with personal data of minors under 14 years of age

□ Yes, with personal data of minors between 14 and 17 years old

□ Yes, with personal data of adults

6.2 Will you work with specially protected data, such as health, biometric or genetic data, data of racial or ethnic origin, philosophical or religious convictions, political ideas or other data with a high risk for the affected individuals?

If so, indicate with which data.

6.3 In my research, is it essential to identify specific people?

□ No, because personal data are anonymous (see section 6.4)

□ Yes (see section 6.5)

6.4 Why are personal data anonymous?

□ Because no identifiers of the affected people are captured or saved (mail addresses, IP numbers of devices connected to the internet, mobile phone numbers, names and surnames or ID, etc.).

□ Because participation in the project only generates a code for each participant, which makes it impossible for the UdL research team to identify their identity.

□ Because under no circumstances are the voices or images of the affected people recorded.

□ Because it is not possible to deduce, in a reasonable way and without disproportionate efforts, from the results of the execution of the Projects (content of the surveys and interviews, content of the medical analytics, etc.), the identity of the people.

□ Because, although identifiers of the affected people are captured or saved, the necessary technical measures have been implemented to avoid re-identification ("anonymization" through data encryption or encryption).

6.5 Why do personal data identify specific people?

□ Because identifiers of the affected people are captured or saved and no technical measures have been implemented to prevent re-identification.

□ Because recordings are made of the voice or image of the people affected.

□ Because in a reasonable way and without disproportionate effort it is possible to deduce, from the results of the execution of the Project, the identity of the people.

□ Because personal data pseudonymization systems are used.

6.6 Confirm which tools you will use to manage personal data.

If you select the second option, you must answer the points:

6.7 Which non-UdL tools will you use to manage personal data? Why are they essential?

6.8 I confirm that despite using tools outside the UdL for the management of personal data, I will also store this data in a UdL environment.

☐ I confirm my obligation to use only tools from the UdL, or for which the UdL has acquired the license, under the control of my thesis supervisor.

☐ I confirm that, although I am aware of the obligation to use only tools from the UdL, or for which the UdL has acquired the license, it becomes essential for my research to use additional ones and I have the approval from my thesis director.

6.7 Which non-UdL tools will you use to manage personal data? Why are they essential?

For exemple specific plataforms for surveys, such as Google forms, EUSurvey, etc.

6.8 I confirm that despite using tools outside the UdL for the management of personal data, I will also store this data in a UdL environment.

☐ I confirm that, in any case, the data will remain stored in an environment (preferably electronic) of the UdL.