UNIT FOR RESPONSIBLE CONDUCT IN RESEARCH (URCR)

More information on data management[1]

A DMP is *a living* document outlining how research data collected or generated will be handled during and after a research project. The EC provides guidance on what to address in a DMP -- https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan-template_he_en.docx)

The Guidelines emphasise the goal to produce FAIR data: data that is Findable, Accessible, Interoperable and Re-usable

Findable: Make data discoverable with metadata and a standard identification mechanism (*e.g.* a DOI).

Accessible: Plan – and eventually provide – documentation and tools needed to access the data. If certain datasets cannot be shared, provide a clear reason why that is. For example, if you are using patients' data, you are not allowed to share it, unless it is properly de-identified and does not allow any tracking of patients' identity.

Interoperable: Allowing data exchange and re-use between researchers and institutions asks for standardised metadata and methodologies.

Re-usable: Releasing data with a license clarifies how data can be re-used.

In addition, DMPs should also address:

- Data volume and file formats
- Costs for FAIR data management and long-term preservation
- Data security
- Ethical issues

[1] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004525

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

How to set up a Data Management Plan?

You can find DMP templates at

DMPONLINE (https://dmponline.dcc.ac.uk/; https://dmptool.org/)

ARGOS, a tool initially developed by OpenAIRE and EUDAT from an open source software - OpenDMP.

https://argos.openaire.eu

Learn how to use it

https://argos.openaire.eu/splash/resources/user-guide.html

You can save your DMP anytime, share it, deposit in Zenodo, get a DOI and/or export it in various formats.

What about copyright and licensing?

Making research data openly accessible is best done using explicit licenses. For information about licencing see:

- https://opendefinition.org/guide/data/
- https://opendefinition.org/licenses/
- http://wiki.creativecommons.org/Data_and_CC_licenses

Where to store the data?

You can store your data in a data repository, a digital archive collecting and displaying datasets and their metadata. Many data repositories also accept publications and allow linking them to their underlying data.

An overview of repositories can be found at Re3data:

www.re3data.org

If you have difficulties finding a suitable repository, you can use OpenAIRE's catch-all repository, Zenodo. It is advisable to contact the repository of your choice when writing the first version of your DMP. Repositories may offer guidelines for sustainable data formats and metadata standards, as well as support for handling with sensitive data and licensing.

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

ZENODO:

Catch-all repository for EU funded research

- Up to 50 GB per upload
- Data stored in the CERN Data Center
- Persistent identifiers (DOIs) for every upload
- Includes article level metrics
- Free for the long tail of science
- Open to all research outputs from all disciplines
- Github integration
- Easily add EC funding information and report via OpenAIRE

Zenodo enables and encourages you to share your research as openly as possible to maximise use and re-use of your research results. However, one size does not fit all. Therefore, while open licenses are encouraged where possible, it also possible to upload under a variety of different licenses and access levels.

Tips for good data management plans

1. Identify the Data to Be Collected:

Types. A good first step is to list the various types of data that you expect to collect or create. This may include text, spreadsheets, software and algorithms, models, images and movies, audio files, and patient records.

Sources. Data may come from direct human observation, laboratory and field instruments, experiments, simulations, and compilations of data from other studies. Reviewers and sponsors may be particularly interested in understanding if data are proprietary, are being compiled from other studies, pertain to human subjects, or are otherwise subject to restrictions in their use or redistribution.

Volume. Both the total volume of data and the total number of files that are expected to be collected can affect all other data management activities.

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Data and file formats. Technology changes and formats that are acceptable today may soon be obsolete. Good choices include those formats that are nonproprietary, based upon open standards, and widely adopted and preferred by the scientific community (*e.g.*, Comma Separated Values [CSV] over Excel [.xls, xlsx]). Data are more likely to be accessible for the long term if they are uncompressed, unencrypted, and stored using standard character encodings such as UTF-16.

2. Define How the Data Will Be Organized

For many projects, a small number of data tables will be generated that can be effectively managed with commercial or open source spreadsheet programs like Excel and OpenOffice Calc. Larger data volumes and usage constraints may require the use of relational database management systems (RDBMS) for linked data tables like ORACLE or mySQL, or a Geographic Information System (GIS) for geospatial data layers like ArcGIS, GRASS, or QGIS.

Depending on sponsor requirements and space constraints, it may also be useful to specify conventions for file naming, persistent unique identifiers (e.g., Digital Object Identifiers [DOIs]), and versioning control (for both software and data products).

3. Explain How the Data Will Be Documented

Metadata—the details about what, where, when, why, and how the data were collected, processed, and interpreted—provide the information that enables data and files to be discovered, used, and properly cited. Metadata include descriptions of how data and files are named, physically structured, and stored as well as details about the experiments, analytical methods, and research context. It is generally the case that the utility and longevity of data relate directly to how complete and comprehensive the metadata are. The amount of effort devoted to creating comprehensive metadata may vary substantially based on the complexity, types, and volume of data.

First, identify the types of information that should be captured to enable a researcher like you to discover, access, interpret, use, and cite your data.

Second, determine whether there is a community-based metadata schema or standard (*i.e.*, preferred sets of metadata elements) that can be adopted.

Third, identify software tools that can be employed to create and manage metadata content (*e.g.*, Metavist, Morpho). In lieu of existing tools, text files

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

(*e.g.*, readme.txt) that include the relevant metadata can be included as headers to the data files.

The metadata recorded in the notebook provide the basis for the metadata that will be associated with data products that are to be stored, reused, and shared.

4. Describe How Data Quality Will Be Assured

It is good practice to describe the Quality Assurance and Quality Control (QA/QC) measures that you plan to employ in your project. Such measures may encompass training activities, instrument calibration and verification tests, double-blind data entry, and statistical and visualization approaches to error detection. Simple graphical data exploration approaches (*e.g.*, scatterplots, mapping) can be invaluable for detecting anomalies and errors.

5. Present a Sound Data Storage and Preservation Strategy

A common mistake of inexperienced (and even many experienced) researchers is to assume that their personal computer and website will live forever. They fail to routinely duplicate their data during the course of the project and do not see the benefit of archiving data in a secure location for the long term. Inevitably, though, papers get lost, hard disks crash, URLs break, and tapes and other media degrade, with the result that the data becomes unavailable for use by both the originators and others. Thus, data storage and preservation are central to any good data management plan.

Give careful consideration to three questions:

How long will the data be stored and accessible?
How will data be stored and protected over the duration of the project?
How will data be preserved and made available for future use?

Develop a sound plan for storing and protecting data over the life of the project. A good approach is to store three copies in at least two geographically distributed locations (*e.g.*, original location such as a desktop computer, an external hard drive, and one or more remote sites) and to adopt a regular schedule for duplicating the data (*i.e.*, backup). Remote locations may include an offsite collaborator's laboratory, an institutional repository (*e.g.*, your departmental, university, or organization's repository if located in a different building), or a commercial service, such as those offered by Amazon, Dropbox,

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Google, and Microsoft. The backup schedule should also include testing to ensure that stored data files can be retrieved.

Many universities and organizations also host institutional repositories, and there are numerous general science data repositories such as Dryad (http://datadryad.org/), figshare (http://figshare.com/), and Zenodo (http://zenodo.org/).

Alternatively, one can easily search for discipline-specific and general-use repositories via online catalogs such as http://www.re3data.org/ (*i.e.*, REgistry of REsearch data REpositories) and http://www.biosharing.org (*i.e.*, BioSharing).

It is often considered good practice to deposit code in a host repository like GitHub that specializes in source code management as well as some types of data like large files and tabular data (see https://github.com/).

Make note of any repository-specific policies (*e.g.*, data privacy and security, requirements to submit associated code) and costs for data submission, curation, and backup that should be included in the DMP and the proposal budget.

6. Define the Project's Data Policies

Explain how and when the data and other research products will be made available. Be sure to explain any embargo periods or delays such as publication or patent reasons. A common practice is to make data broadly available at the time of publication, or in the case of graduate students, at the time the graduate degree is awarded. Whenever possible, apply standard rights waivers or licenses, such as those established by Open Data Commons (ODC) and Creative Commons (CC), that guide subsequent use of data and other intellectual products.[2] The CC0 license and the ODC Public Domain Dedication and License, for example, promote unrestricted sharing and data use. Nonstandard licenses and waivers can be a significant barrier to reuse.

Explain how human subject and other sensitive data will be treated (GDPR; Clinical Studies Portuguese Law).

[2] See http://creativecommons.org/
http://opendatacommons.org/licenses/pddl/summary/

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

7. Describe How the Data Will Be Disseminated

There are passive and active ways to disseminate data. Passive approaches include posting data on a project or personal website or mailing or emailing data upon request, although the latter can be problematic when dealing with large data and bandwidth constraints.

More active, robust, and preferred approaches include:

(1) publishing the data in an open repository or archive (see tip 5 above);

(2) submitting the data (or subsets thereof) as appendices or supplements to journal articles, such as is commonly done with the PLOS family of journals; and

(3) publishing the data, metadata, and relevant code as a "data paper". Data papers can be published in various journals, including *Scientific Data* (from Nature Publishing Group), the *GeoScience Data Journal* (a Wiley publication on behalf of the Royal Meteorological Society), and *GigaScience* (a joint BioMed Central and Springer publication that supports big data from many biology and life science disciplines).

A good dissemination plan includes a few concise statements:

When, how, and what data products will be made available. Generally, making data available to the greatest extent and with the fewest possible restrictions at the time of publication or project completion is encouraged. The more proactive approaches described above are greatly preferred over mailing or emailing data and will likely save significant time and money in the long run, as the data curation and sharing will be supported by the appropriate journals and repositories or archives.

Keep in mind that the data will be more usable and interpretable by you and others if the data are disseminated using standard, nonproprietary approaches and if the data are accompanied by metadata and associated code that is used for data processing.

8. Assign Roles and Responsibilities

A comprehensive DMP clearly articulates the roles and responsibilities of every named individual and organization associated with the project. Roles may include data collection, data entry, QA/QC, metadata creation and management, backup, data preparation and submission to an archive, and

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

systems administration. For small to medium size projects, a single student or postdoctoral associate who is collecting and processing the data may easily assume most or all of the data management tasks. In contrast, large, multi-investigator projects may benefit from having a dedicated staff person(s) assigned to data management.

Treat your DMP as a living document and revisit it frequently (*e.g.*, quarterly basis). Assign a project team member to revise the plan, reflecting any new changes in protocols and policies. It is good practice to track any changes in a revision history that lists the dates that any changes were made to the plan along with the details about those changes, including who made them.

9. Prepare a Realistic Budget

Creating, managing, publishing, and sharing high-quality data is as much a part of the 21st century research enterprise as is publishing the results. Data management is not new—rather, it is something that all researchers already do. Nonetheless, a common mistake in developing a DMP is forgetting to budget for the activities. Data management takes time and costs money in terms of software, hardware, and personnel. Review your plan and make sure that there are lines in the budget to support the people that manage the data as well as pay for the requisite hardware, software, and services.

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt