FOUNDATION FOR RESEARCH
AND TECHNOLOGY – HELLAS
INSTITUTE OF MOLECULAR
BIOLOGY & BIOTECHNOLOGY

GUIDELINES FOR DATA MANAGEMENT

This document describes the best practices for acquiring and storing records of research activities at IMBB and aims to improve research efficiency and scientific cooperation by supporting data discovery, accessibility, interoperability and re-use.

A Data Management Plan (DMP) describes data that will be acquired or produced during research; how the data will be managed, described, and stored, what standards you will use, and how data will be handled and protected during and after the completion of the project. Each project should be accompanied by a DMP that covers the following topics.

A. GENERAL INFORMATION

Who is Responsible for the DMP?

The principle investigator of the research project is the main responsible for data management while the research is ongoing, but this responsibility can be delegated on others, lab manager, data steward).

What resources, for example financial and time will be dedicated towards data management?

It is important to pre-allocate the necessary resources for management of the research. That includes identifying the necessary time for data management and having a DMP for grant applications in order to budget for the associated costs of data storage and management for the duration of the project.

B. DATA MANAGEMENT

How will new data be collected or produced?

Describe in detail the methodologies/software that will be used for any data that are collected or produced. Additionally, explain how data provenance, the record trails that account for the origin of a piece of data together with an explanation of how and why it got to the present place, will be kept.

What kind of data will be collected or produced?

Detail the types of data that are going to be generated such as numeric (databases, spreadsheets), textual (documents), image, audio, video, and/or mixed media, and the data format that they will have. Give preference to open formats with accepted standards and/or with widespread usage within the research community when possible. Determine the approximate volumes of data in storage space that will be generated in order to be able to allocate the necessary resources.

What are the data that need to be retained/shared?

DMPs should specify specific criteria for the identification of **significant data** (data that will be shared and retained – see below). These should at least include:

- Data that substantiates published research findings.
- Unrepeatable observations.
- Experimental results that would be impossible or expensive to reproduce.

What metadata and will accompany the data?

Indicate which metadata will be provided to help others identify and discover the data. Indicate how the data will be organized during the project, mentioning for example conventions, version control, and folder structures. It is important to use community metadata standards where these are in place that will make research data will be easier to find, understand, and re-use.

Consider how this information will be captured and where it will be recorded and specify any other documentation that might be needed to enable re-use. The metadata information can reside in a database with links to each item, in a text file, in file headers, code books, or within lab notebooks. It is strongly suggested to use an electronic form of lab notebooks with standardized format and that supports backups. IMBB will provide the infrastructure for such a system. In general, a laboratory notebook should include the following information:

- The PI, the researcher and the relevant project under which this experiment was performed.
- As detailed description of the experiment as possible (i.e., what, when, and why something was done).
- Links to products/methods crucial for the repeatability of the experiment.
- Any digital data that can easily be inserted, such as images of gels, plots, etc.
- Data that cannot be digitized must be sufficiently described and their location should be identified.
- Brief description and interpretation of the experiment findings/results.

What data quality control measures will be used?

Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeated samples or measurements, standardized data capture, data entry validation, peer review of data, or representation with controlled vocabularies.

C. DATA STORAGE

How will data and metadata be stored and backed up during the research process?

DMPs should specify the manner in which data will be stored (file formats), annotated/documented (metadata) and shared (repositories).

All significant data and their metadata produced within the scope of research projects at IMBB should be stored on the institutes designated storage infrastructure, except in cases where legal obligations not allow it. Before data are transferred to the storage infrastructure or for non-significant data, describe the procedure that will ensure the safety of the data. The safety of data unsuitable for digitization, should also

be ensured along with proper cataloging that will be in the metadata of the project. **All data should be accompanied of description metadata and clear documentation, allowing its identification and effective reuse.**

How will the integrity of the data be ensured?

All additions, deletions and alterations to datasets shall be extensively documented and justified. Significant datasets will be content versioned and distinct persistent identifiers (PIDs) will be generated. Research results deriving from such data will reference the particular version of the dataset that was used, and any changes will result in distinct PIDs.

How will data security and protection of sensitive data be taken care of during the research?

Provide information about the protection and access restrictions that sensitive data with personal information might need. For the data that are not on the designated storage infrastructure of IMBB, describe their recovery in the event of an incident.

D. DATA SHARING

How and when and which data will be shared?

IMBB is committed to open and reproducible science and therefore abides by the FAIR principles (Wilkinson et al. 2016). Data (or any digital object including non-data research objects) & metadata (information about the data) should be:

Findable. Data & metadata should have a form that allows easy human-driven and machine-driven searching.

- Data are described with rich metadata to facilitate discovery.
- Data & their metadata are assigned a unique and persistent identifier (PID) that allows identification, enables searching & citing, and provides support for data versioning.
- Metadata clearly and explicitly include the identifier of the data it described.
- (meta)data are registered or indexed in a searchable resource.

Accessible. Users need to easily access the (meta)data.

- (meta)data are retrievable by their identifier using a standardized communications protocol.

- Data shall be made available preferentially in open file formats. In the cases in which this is not possible, software that can be used to access the data will be named (including where it can be obtained, and which version was used).
- Metadata are accessible, even when the data are no longer available.

Interoperable. The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- (meta)data include qualified references to other (meta)data.
- (meta)data must be kept in a form and quality that allows them to be retraced and reprocessed, with described with a broadly applicable language.
- available alongside the data, properly documented.
- Protocols used to collect the data shall be provided alongside the data.
- Any custom code for analysis or generation of the data.
- Listing of packages/libraries used, including version numbers.
- Data should be made available using the least restrictive license possible.

Reusable. To optimize the reuse of data, data should be well-described so that they can be replicated and/or combined in different projects.

- meta(data) are richly described with a plurality of accurate and relevant attributes.
- Meet meta(data) standards that are broadly accepted by the scientific community.

E. LEGAL & ETHICAL REQUIRMENTS

If personal data are used, how will compliance with legislation be ensured?

For all projects dealing with personal data or human samples, the DMP should be compliant with General Data Protection Regulations (GDPR) and therefore should consider anonymization or encryption of personal data for preservation and sharing and if managed access is required.

How will other legal issues, such as intellectual property rights and ownership, be managed?

Consider the type of property rights that your data will have before the collection process. The recommended license for the sharing of data is CC-BY-NC, allowing re-use for any non-profit purpose with attribution.

Are any data going to be reused?

Acknowledge the sources of the data used and abide by the terms and conditions under which they accessed the original data.

F. RESOURCES

- Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3.

- Science Europe: https://www.scienceeurope.org/our-priorities/research-data/