Research Data Management protocol RIMLS-FNWI

Data management refers to processes that guard and maintain the consistency and accuracy of collected data and facilitate the re-use of data. The RIMLS-FNWI institute highly values reproducibility of experiments, availability and re-use of experimental data by others. This document provides information about procedures of data management.

Responsibilities

- Researchers have the obligation to provide accompanying metadata. Supervisors (group leaders), together with the data steward, have the responsibility to check and enforce the archiving of appropriate metadata. According the Radboud University RDM policy[1)], the director of the institute is ultimately responsible.

  [1)] https://www.radboudnet.nl/onderzoek/onderzoek-visie-beleid-kwaliteit/onderzoeksbeleid/research-data-management/centraal-beleid

Included data

- Two categories of research data are to be distinguished:
  - Published data. Data belonging to research published in peer-reviewed journals
  - Unpublished data. Data belonging to unpublished research. This includes work-in-progress data but also data that was not selected for publication.
- Two types of data types are to be distinguished, regardless of data category:
  - Raw data
    - Depending on the platform used for generation of the data this includes data formats according to standards in the field:
    - FASTQ for raw DNA sequencing data
    - BAM for mapped DNA sequencing data (https://samtools.github.io/hts-specs/SAMv1.pdf)
    - RAW for mass-spectrometry data (Thermo RAW)
  - Processed data
    - This includes data resulting from any additional analysis of the raw data.
    - The format of processed data is inherently loosely specified, as it is specific to the analysis conducted. Formats are mostly non-binary (flat-text) files, such as tab-delimited or Microsoft Excel files.

Metadata

- Metadata should include experimental variables that are crucial to repeat the experiment, and to correctly interpret the results of the experiment. Metadata should include at least:
  - Source of biological material (cell line, tissue, organism)
  - Treatment (chemical, biological, compound)
  - Protocol by which the sample was prepared
  - Instrument settings

    - Date
    - Researcher
- Metadata for published data
    - Depending on the database used (see below) metadata is included in different ways:
    - GEO, dbGAP, ENA, and EGA explicitly specify which metadata must be included with data submission to their databases. This includes both experimental and technical ('machine') variables. Upon submission, this metadata is linked to the accompanying data files.
    - PRIDE accepts RAW files, which contain both the raw data as well as the metadata of the corresponding experiment, as such providing an 'in-file' metadata-data link.
- Metadata for unpublished data
    - Metadata accompanying DNA sequencing experiments is provided in Soladmin, a metadata management database that is available to all researchers within the departments. Soladmin runs on local infrastructure and a backup of its contents is stored on a second, independent server.
    - Metadata accompanying mass-spectrometry data is kept in a local database with monthly backup intervals to an independent location.

Data storage

- For published data, several public databases are used, according standards in our field:
    - GEO (https://www.ncbi.nlm.nih.gov/gds) for DNA sequencing and array data (USA), open access
    - dbGAP (https://www.ncbi.nlm.nih.gov/gap) for DNA sequencing and array data (USA), controlled access
    - ENA (https://www.ebi.ac.uk/ena) for DNA sequencing and array data (EU), open access
    - EGA (https://www.ebi.ac.uk/ega) for DNA sequencing and array data (EU), controlled access
    - PRIDE (https://www.ebi.ac.uk/pride) for mass-spectrometry data (EU), open access
    - Database choice is dependent on data-type (mass-spectrometry vs. DNA sequencing data), and per-project restrictions (e.g. EU/USA)
    - In publications, the respective data is referenced using unique persistent identifiers provided by the database. These identifiers link to the databases.
    - At the publishers' request, processed data is primarily provided either as supplemental data to the publication at the journal's website. However, some of the databases mentioned above also accept processed data. In case of large processed-data files, these are co-submitted with the raw data to the public database. For instance, GEO has a flexible policy for accepting various file formats (flat-text, Excel).
- For unpublished data, data is stored on the institutes' infrastructure. For integrity reasons, raw data is kept at a designated partition that is write-protected (only writable by the system administrator).

Data protection

To prevent data loss in case of technical failures, the institutes' data is stored:

- On one of the public databases mentioned above (published data only)
- At the institutes' local infrastructure
- In addition, raw sequencing data is mirrored at an independent physical location of the institutes local infrastructure.
- Raw data is write-protected (only writable by the system administrator).

Maximum retention period

- In accordance with the Radboud University RDM policy, both published and unpublished data is kept for a minimum of 10 years.
- The public databases mentioned above do not explicitly state a restriction on preservation time of submitted data. Therefore, we regard this as ‘permanent’.

Accessibility and re-use

- Published data
  - Raw and processed data is publicly available through one of the databases mentioned above. According current standards in our field, data of a published studies can be freely downloaded and re-used. Availability in databases is a prerequisite for acceptance of a manuscript; the journals request peer-reviewers to check the availability of both raw and published data.. Databases such as GEO provide functionality such as ‘reviewer links’ that allow for anonymized download/viewing of submitted data, prior to publication.
  - The use of standardized data formats for DNA sequencing and mass-spectrometry (FASTQ, BAM, RAW), allows for re-analysis of the original data by others.
- Unpublished data
  - Unpublished data is not publicly accessible, and is only available to researchers within the institutes’ departments.

Privacy of sensitive data

- Published data
  - Sensitive data containing identifiable information (DNA sequence data from donors, patients, healthy volunteers, etc.) is deposited under controlled access, depending on the informed consent of the corresponding project. Both EGA and dbGAP have controlled access mechanisms. As such they are appropriate databases for hosting sensitive patient data under secure standards. Access is controlled by a project-specific Data Access Committee (DAC) and applications are submitted to the Data Access and Compliance Office (DACO) or International Data Access Committee (IDAC). These ensure that potentially identifiable data will only be used by qualified scientists, taken into consideration access policies and restrictions on the purpose of data use. Procedures for identifyable data and controlled access are available via
    - https://www.ebi.ac.uk/ega/about
    - https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html

- Unpublished data
  - Sensitive data containing identifiable information is not publicly available and not accessible to anyone outside of our institute. Access restriction is technically implemented by C&CZ, user accounts are managed by supervisors, with the ultimate responsibility being with the director of the institute. For a subset of projects sensitive data is specifically restricted using group-based control, to be determined by the institute director and managed by the system administrator. However, for the high volumes of data we host, a highly manageable access control list (ACL) is favorable. For future research this needs to be implemented.