Data Management Plan (DMP) Template
NFDI4Microbiota Edition

Project Description

Project name and Acronym:

(very short and descriptive name that can be used as afl project title inline of further document texts. Example: "HumanParalogSeq", "Forschungsdaten Umfrage HU 2022")

Short project description:

(An abstract of the project that informs the reader about the projects scope in a few sentences.)

Contact person(s) in charge of data handling/storage/processing/archival:

(name one person or multiple with respective duties and contact address.)

Project/Funding/UseCase-ID:

(ID provided by the funder or institute, if available.)

Data collection and description of the research data

What data types will the project generate or make use of?

Options: (Meta)genomics / (meta)transcriptomics / (meta)proteomics / (meta)epigenomics, metabolomics / imaging / other ___

Will data from public sources be reused or will data be newly collected (or both)?

Options: Re-used from public sources / newly collected / both (please specify)

For the re-use of existing data, please specify the data sources including project or sample IDs.

Options: ENA / PRIDE / SRA / GEO / ArrayExpress / MGnify / other ___

For newly-collected data, how will they be collected and how will this collection be documented?

Options: paper lab notebook / word processing software, Note-taking app, Spreadsheet, Electronic Lab Notebook (ELN) (please specify which one(s)), Standard Operating Procedures (SOPs) (please specify which one(s)), protocols and standards (please specify which one(s))

What data formats will be used, intermediately and as final output? What is the anticipated final data volume (in GB or TB)?

Options: CSV, TSV, XLSX, RAW MS-Spectra, mzML, FASTA, FASTQ, BAM, image data, other (please specify)

How will data be organized in the group / institution during the project phase?

(Open question: e.g. collected data is saved along with a descriptive experimentalSetup.csv meta information file in a documented project folder on the institutes read-only file share. This file share is hosted on redundant machines which are backed-up nightly in increments and monthly as complete copy on tape by the central IT infrastructure. Processed data is stored on a writable file share and moved to the read-only project folder upon project completion.)

What file naming convention will be used in the group / institution?

(Open question: e.g. All primary data files use a predefined naming scheme: "projectname_experiment_condition_YYYYMMDD".)

Data documentation and quality

How will data be documented?

Options: README file, (electronic) lab notebook, sample sheet, data dictionary, codebook, article in a data journal, other ___

What metadata will be created/used?

Open question: e.g. Data is described according to the minimal meta data standards for the respective type as defined by NFDI4Microbiota. Any additional descriptive information especially to describe the varying experimental conditions is added in extra columns.

How will metadata be produced?

Open question: e.g. All generated meta data from instruments (settings, run information) is extracted and augmented, if necessary, with manual descriptions. Subsequent processing steps and statistical output is always stored with the data as provenance by NFDI4Microbiota workflows.

What metadata format(s) will be generated?

Open question: e.g. During collection and for on-site storage, accompanying CSV and README files describe the data and experimental provenance. The data will be uploaded to respective repositories with submission of the relevant meta data fields at the earliest time point possible.

What standards/vocabularies will be used?

Open question: e.g. For all genes, proteins, organisms and other ontological terms, fully qualified unique identifiers of referable databases (RefSeq, UniProt) are used. This effort is aided by the effort of NFDI4Microbiota (e.g. strain identifiers).

Data processing & analysis

What steps will be taken to ensure the quality of the data and metadata?

Open question: e.g. Quality checks of raw and processed data are conducted as part of the processing workflow and documented along with intermediate results. Metadata is recorded along each processing step as part of intermediate results.

What processing steps will be applied to the data during analysis?

Open question: e.g. sequencing reads are trimmed using... , mapped to the respective reference transcriptome/genome using... or assembled using ..., mass spectra compared with respective protein databases using..., microscopic images are screened for cell counts and shapes using...

What software and hardware will be used for data processing?

Open question: e.g. fastp v1.0.1, STAR v.2.0.4.11, MaxQuant v3.1.9, ImageJ v4.2.1 on the de.NBI OpenStack cloud computing infrastructure

Data sharing & publishing

Does your data contain sensitive information (about individuals, endangered species) or is it protected by legal rights of third parties? If so, how are potential issues handled during the data lifecycle?

Open question: e.g. the project deals exclusively with non-sensitive information that can be published along with all metadata without prior anonymization to the respective repositories.

How will the data be shared?

Options: (1) All raw, processed and meta data will be openly accessible at online repositories at the earliest time point. (2) Data will be shared openly accessible through online repositories upon completion of the project. (3) The following circumstances prevent data sharing...

How will data be shared with collaborators during the project phase? Is there an access management system in place to allow collaborators and reviewers (restricted) access to the data?

Options: Open Science Framework (OSF), Github, Gitlab. git-annex, datalad, other ___

Additionally, data will be...

Where will the raw and processed data be published?

Options: in a data repository, in an enhanced publication, in a data report, in a data paper, other ___

If data are deposited in a repository, which one?

Options: ENA, SRA, PRIDE, GEO, ArrayExpress, MGnify, other (please specify) (depending on previous selections)

Do you want to make this DMP publicly available?

Options: yes (where), no

Digital preservation

What steps will be taken to back up data at the research facility during processing (e.g. number of copies, media types, and physical locations)?

Open question: e.g. Physical storage of paper lab notebooks or the usage of ELNs and local backup of primary data on at least one computer and an external hard drive in a separate location.

Which data will be kept for how long?

Open question: e.g. raw primary data will be kept on institutional servers for at least ten years along with all metadata that trace its origin. Results will either be published in an open access journal or in case of intermediate results deposited at the university's data repository for five years. All workflows and code necessary to reproduce the (intermediate) results from the raw data are transparently published in a central git repository.

How is the secure and robust storage of data and associated metadata ensured during and after the project?

Open question e.g. The institute's storage cluster ensures triplicate storage redundancy across multiple machines and regular incremental backup at a second site for at least three years beyond the lifespan of the project. All data is transferred there as soon as possible along with all relevant metadata.