Exotoxins Lab - Data Management Plan

Dr. Luisa F. Jimenez Soto

Version from July 11, 2022

Contents

1 Credits for the template used for this document 3

2 General Information 4

3 Description of the project 4
3.1 Overview 4
3.2 Principal researchers involved 4
3.3 Collaborating researchers 4
3.4 Funding organization 4
3.5 Data producer 4
3.6 Project data contact 5
3.7 Data owner(s) 5
3.8 Roles 5
3.9 Cost and resources 5

4 Data description and collection or re-use of existing data 5
4.1 Origin of the data 6
4.2 Types of data created 6
4.3 Types of data re-used 6

5 Organizing and Documenting the data 7
5.1 Data collection 7
5.2 Data organization 12
5.3 Data type and size 13
5.4 Data format 13
5.5 Directory structure and names 13
5.6 File structure and names 14
5.7 Documentation 15
5.8 Metadata 15
5.9 Metadata standard (if applicable) 15

6 Processing your data 16
6.1 Versioning . . . 16
6.2 Interoperability . . . 16
6.3 Data quality . . . 16

7 Storing data and metadata 16
7.1 Storage . . . 16
7.2 Backup . . . 17
7.3 Security . . . 17

8 Protecting your data 18
8.1 Ethical Review (if applicable) . . . 18
8.2 Informed consent (if applicable) . . . 18
8.3 (sensitive) Personal data /confidential information (if applicable) . . . 18
8.4 Intellectual property rights (IPR) or Copyrights . . . 18
8.5 Agreements (if applicable) . . . 18
8.6 Restrictions (if applicable) . . . 18

9 Archiving and publishing your data 19
9.1 Archiving . . . 19
9.2 Data formats . . . 19
9.3 Access (if applicable) . . . 19

10 Discovering data 20
10.1 Identification of needs . . . 20
10.2 Search for data . . . 20
10.3 Gaining access to data . . . 20

A Creative Commons licenses 4.0 20
A.1 CC BY . . . 20
A.2 CC BY-SA . . . 21
A.3 CC BY-NC . . . 21
A.4 CC BY-NC-SA . . . 21
A.5 CC BY-ND . . . 22
A.6 CC BY-NC-ND . . . 22

B Agreement to follow the DMP 23

C FAIR data principles 24

D Practical approach to the data collection and storage 26
D.1 Documenting an experiment . . . 26
D.2 ... you think you need to order a new reagent . . . 27
D.3 Protocol design and documentation . . . 29
D.3.1 Title . . . 29

D.3.2 Purpose . . . 29
D.3.3 Materials . . . 29
D.3.4 Procedure . . . 29
D.3.5 Modifications . . . 29
D.3.6 References . . . 29
D.4 Experimental design . . . 30
D.4.1 What is the hypothesis that will be investigated? . . . 30
D.4.2 How will the crucial variables be operationalized?(Variables to be measured) . . 30
D.4.3 What is the source of the data included in the analyses? . . . 30
D.4.4 Are there any exclusion criteria for the data? . . . 30
D.4.5 What are the planned statistical analyses? . . . 30
D.4.6 What are the criteria for confirming and disconfirming the hypotheses? . . . 31
D.4.7 Have the analyses been validated on a subset of the data? If yes, please specify and provide the relevant files . . . 31
D.4.8 What is known about the data that could be relevant for the tested hypotheses? 31
D.4.9 Please provide a brief timeline for the different steps in the preregistration . . . 31
D.4.10 Transparency statement . . . 31
D.4.11 Conditional safeguards . . . 31
D.5 Linux / Unix File Naming . . . 32

E Sources **34**

1 Credits for the template used for this document

All sections and text explaining what should go in each section were copied directly from the CESSDA (https://www.cessda.eu/Training/DMEG) PDF guide called *Adapt your Data Management Plan*

4 Data description and collection or re-use of existing data

What kind of data will be used during the project? If you are reusing existing data: What is the scope, volume and format? How are different data sources integrated? If you are collecting new data can you clarify why this is necessary?

4.1 Origin of the data

Our project will be a mixture of bioinformatics resources and experimental assays. We will explain the types of data based on these two working areas. For a detailed description of the processes for Data collection, please go to Appendix D

4.2 Types of data created

Experimental Data produced in wet-lab will include:

- Bacteria strains (Part of the GenTSV and Infektionsschutz Gesetz)
- Plasmids or vectors (Part of the GenTSV requirements)
- Chemicals and reagents, their stock and working solutions
- Images (like from microscopy, or gels)
- Values from measurements (from machines)
- Protocols and/or SOPs (Standard Operating Procedures)
- The analysis from the measurements (statistics, analysis pipeline)

Informatics During the Data Science projects and the analysis of the experimental data we will have the following types of data

- Scripts for automated download of data from existing available resources.
- Scripts for parsing of downloaded data.
- Scripts for the submission of processes to the cluster.
- Code for data cleaning and building of predictors.
- DNA sequences from QC sequencing of vectors and plasmids (made/bought).

4.3 Types of data re-used

Reused data is defined as usage of data found in a database or repository and associated with a creator and/or publication. They will include (based on origin):

Experimental

- Ciliate strains
- Protocols
- Published and/or commercial vectors
- Commercial kits for DNA extraction, protein synthesis and/or RNA extraction.

Informatics

- Protein Sequences
- DNA sequences
- Packages for analysis of proteins
- Publications
- Python and/ or R libraries

All data types used will be documented using metadata (see Metadata section 5.8)

5 Organizing and Documenting the data

5.1 Data collection

For each type of data, the following will be listed here: How is the data collected? is there a specific software and/or hardware required? Who is responsible for the data collection? During which phase of the project will the data be collected? and where will the data be collected?

There are three classes of data we will gather during this project: experimental, bioinformatics and teaching material.

Experimental Experimental data includes the data related to physical objects in the lab (reagents, solutions, tubes, samples, etc) and to their digital recordings including documentation, properties, protocols, data analysis. For data analysis there will be always: raw data (data obtained from the machine measuring any physical or chemical property), exploratory data analysis (R or Jupyter notebooks used to explore the data, its properties and statistical analysis), and Code (Final code in R and python that can be run from the command line in linux). For the purpose of easy documentation and overview of results, all graphs or images obtained during the data analysis is considered "Figures" . This will be kept in the folder with the same name. For intermediate results that might be used for the analysis , like after data cleaning or wrangling, they will be stored in the directory "derivate".

Here are more details regarding each type of process in the project that will produce data and the plan on what to obtain it and store it.

Experiments This includes *wetlab* and notes taken for Data Science and Machine learning

- *Data Collection*: Daily entries shall be stored in the Electronic Lab Notebook (ELN) under their own project documentation, called "Experiment". Each entry has to start with the date (use the time stamp provided in the ELN. If files are developed for that entry, they are to be stored versioned attached to the Experiment (Do not overwrite the previous files). For all data, the position in the personal directory system should be added. For data is bigger than 1 MB, do not attach but only add the directory address and file name. The ELN allows for links between the database and other experiments. Use this property in each Experiment.
- *Hardware*: Each person has its own computer and access to the system and Research Group hard disk (AGJimenez). These will be used to enter the data in the ELN and personal disks. It will be recommended to each student to have an external harddisk for backups. Since the ELN is online accessible, LAN or WLAN are necessary to connect to the internet.

- *Software*:For the entries in the ELN only an internet browser is needed. The ELN has been successfully used with FireFox (MacOS and WinOS) and Chromium (Linux).
- *Responsibility for collection*: Each student or lab member or guest is responsible for the daily entries.
- *Project Phase*: All phases
- *Storage*:A daily backup of the ELN book will be performed at 6:00 am between Monday and Friday. The backup copy will be stored in the Institute's server. Its backup routine is not yet known.

Protocols

- *Data Collection*: Protocols shall be added to the Electronic Lab Notebook (ELN). In the ELN there is a "Database" section where protocols are added. The template at the moment (June 2022) is in LibreOffice write. The use of tags will allow to distinguish protocols for the wetlab (#wetlab) or computer related (#insilico)
- *Hardware*:Each person has its own computer and access to the system and Research Group hard disk (AGJimenez). These will be used to enter the data in the ELN and personal disks. It will be recommended to each student to have an external harddisk for backups. Since the ELN is online accessible, LAN or WLAN are necessary to connect to the internet.
- *Software*: For the templates, LibreOffice will be needed. The final version will be exported to PDF and this stored in the database with its corresponding entry.For ELN access a internet browser is needed. The ELN has been successfully used with FireFox (MacOS and WinOS) and Chromium (Linux).
- *Responsibility for collection*: Each student, lab member or guest is responsible for the upload of each protocol.
- *Project Phase*: All phases
- *Storage*: A daily backup of the ELN book will be performed at 6:00 am between Monday and Friday. The backup copy will be stored in the Institute's server. Its backup routine is not yet known.

Reagents and Solutions

- *Data Collection*: As soon as a chemical arrives, an entry shall be made in the ELN database, where the general storage information, company name and the lot number will be added. Their stock solution will be stored in the same entry together with the MSDS (Material Safety Data Sheet) attached to the file as PDF. The protocol for each stock and working solution has to be made while planning the solution and notes about solubility or unexpected changes in concentration shall be added. The placement of the chemical and its stock solutions will be tagged using #RT (for room temperature), #4°C (for fridge), #-20°C (for freezer) or #-80°C (for long term freezer).
- *Hardware*: All chemicals will be stored accordingly to their properties in the respective area in the lab. In order to take responsibility (not ownership!) of the chemical, all chemicals shall be labeled with **AG Jimenez**. The date of opening has to be added to the container, same with stock solutions.

- *Software*: For ELN access a internet browser is needed. The ELN has been successfully used with FireFox (MacOS and WinOS) and Chromium (Linux). For the solution protocol, LibreOffice Writer should be used.
- *Responsibility for collection*: Each student, lab member or guest is responsible for the upload of each protocol and chemical.
- *Project Phase*: All phases
- *Storage*: A daily backup of the ELN book will be performed at 6:00 am between Monday and Friday. The backup copy will be stored in the Institute's server. Its backup routine is not yet known. The physical storage has been explained under "Data Collection".

Plasmids and cloning

- *Data Collection*: All plasmids and cloning procedures will be stored in the ELN. The sequences shall be stored in GenBank with features format (*gb), together with the plasmid / vector map and the sequences used for the *in-silico* cloning. This needs to be done during the cloning planning phase. The sequences obtained from the verification of cloning products will be added in their raw form. Each plasmid shall be labeled p<Your Initials>_<consecutive number>. If they are not verified, the will have a * together with the clone number. Once they have verified, the * and clone number will be removed. If different versions are discovered with different properties, each will get a lower case letter for the variation. No plasmid should contain the gene or origin in their name. Only standardize names (as describe above) should be used. The images collected and created should be stored as SVG, jpg or png (in descending order of preference).
- *Hardware*: For the planning of plasmids and their sequences, a computer will be needed. For the laboratory work the following equipment is necessary: PCR machine, cloning material (recombinases or restriction enzymes), verification equipment (Gel chambers, agarose, DNA staining reagents, imager for UV light with camera),
- *Software*: Pending to find a proper software for mapping. A plain text editor is needed for entries of sequences and their features using the *gb format (see Data collection segment). A image viewer will be required for images collected. For alignments of sequencing results I am still looking for the software compatible with FAIR principles of data.
- *Responsibility for collection*:
- *Project Phase*: Phase II and Phase III
- *Storage*: A daily backup of the ELN book will be performed at 6:00 am between Monday and Friday. The backup copy will be stored in the Institute's server. Its backup routine is not yet known. The physical storage has been explained under "Data Collection".

Flow Cytometry - Toxicology assays

- *Data Collection*:

  Data obtained from flow cytometry assays will be collected via USB from the computer in fcs 3.0 format. The data will be transferred to the main hard disk of the group under its respective directory (see structure) under "raw data".
- *Hardware*: The flow cytometer (planning to use a Guava HTS) will allow us to collect the data.

- *Software*:

  The machine has its own software. The raw data will be stored in the software's format. However all data needs to be exported as fcs 3.0 for backup, together with its readme.md file (metadata). The data analysis will be done with R (FlowCore or python (FlowCytometryTools 0.5.1)
- *Responsibility for collection*: The experimentator
- *Project Phase*: Phase I (Exploratory), Phase II(Pilot experiments),Phase III (Main experiments).
- *Storage*: In the measurement computer and the Research group remote hard disk for the raw data (/AGJimenez/Projects/BMBF/). Analysis in the designated project directory with their versions in the ELN.

Fluorometer - Metabolic assay

- *Data Collection*:

  Data obtained from fluorometrics assays will be collected via USB from the computer in csv / tsv format. The data will be transfered to the main hard disk of the group under its respective directory (see structure) under "raw data".
- *Hardware*: Fluorometer - HTS in the lab
- *Software*: Fluorometer Software for the export (NAME HERE), R or Python for basic data wrangling, data analysis, and visualization.
- *Responsibility for collection*: The experimentator (Student, guest, lab member).
- *Project Phase*: Phase II and Phase III
- *Storage*: Raw data in its directory, and copy in the ELN entry if not over 1 Mb (usually the csv files will be just a hundred kilobytes). With the daily backup of the ELN the data will have a secure storage.

Microscopy - Confirmation assays

- *Data Collection*: Data will be copied in USB and transferred from any computer connected to a microscope with a installed camera (Epifluorescent or Confocal) and then transferred to the projects raw data folder. Each file hs to be named with the Experiment designation (usually initials followed by a number) used for the ELN documentation. Depending on the size of the files the data can be stored in the ELN or in the raw data folder of the group (see Storage locations). All data should be stored / transferred as *raw or *tiff.
- *Hardware*: The data will be produced using the epifluorescent and confocal microscopes. At the moment there are no other alternatives for imaging with digital recording.
- *Software*: Each microscope has its own acquisition software. For the analysis and/or visualization we will use ImageJ (or its fiji ImageJ version) which is Java based and platform independent. ImageJ has macros, so all processes dealing with the images should be store as a macro, or in a screen cast. Other software for labeling (if necessary) should be GIMP. In order to guarantee that the raw image has being modified only for illustration purposes, the whole process from raw data to final needs to be recorded with a screen cast and associated as metadata to the final result. If metadata is available from the microscopy session (not all microscopes have this information), this should be label as such and associated with the raw data.

- *Responsibility for collection*: The experimentator (students, guest, lab member)
- *Project Phase*: Phase II (Pilot experiments) and Phase III (toxicological analysis)
- *Storage*: Because of the big size of image files, the raw data needs to be saved in the computer used by the experimentator and a copy in the Research Group server storage.

Bioinformatics Bioinformatics applications will cover data for the Machine Learning and data used for predictions.

Benchmarking data Benchmarking is the step where our predictor, process or results are compared with the results of previously published ones. The data called "benchmarking data" is the one obtained from other researchers used for their analysis, or data collected based on their description of the analysis. Based on experience, the groups do not seem to have an organized view of where their used data are, so we have to be prepare for both cases.

- *Data Collection*: The publications will be used for the collection of the data and associated with it, so we can reference them when we do the benchmarking. All data will be considered raw data. Similar to all other processes, the structure of the files and the exploratory information (data wrangling) will be kept. All processing will be kept as code. No manual process will be done with the data after download. Names should be kept as the original publication. The location of the files as well the directory structure and content will be documented in the ELN.
- *Hardware*: A computer with LAN or WLAN connection to the internet. An external storage unit for the raw data (AGJimenez)
- *Software*: Internet browser and R and Python for processing of data and compression.
- *Responsibility for collection*: Operator responsible for data analysid, Principal Inverstigator if e-mail contact was successful.
- *Project Phase*: Phase I and Phase II
- *Storage*: In external hard drive until its use for evaluation. For processing, it will be stored in the computer of the member responsible for their analysis and curation. if the size allows it, it should be stored in the ELN entry. However, if it is too big (above 10 Mb), the directory address containing the data should be added to the ELN entry

Machine Learning Models

- *Data Collection*: The datasets ML compatible (Result of datasets processing processes) will be stored using the directory structure illustrated under figure 1. All code used for the ML model will be stored as exploratory Notebooks during the exploratory phase. The training code and its validation metrics will be stored as well. All data, files, processes, and directory addresses will be documented in the README.md file as well the ELN with the planning of the processes.
- *Hardware*: Computer with internet connection (WLAN or LAN), Cluster for calculations and training.
- *Software*:Python and R, Cluster software (no idea which one)
- *Responsibility for collection*: The experimentator or data analist (PhD student, Master Student, or PI)
- *Project Phase*: Phase I and Phase II

- *Storage*: For final storage it should be placed in the Server Hard drive for the research group provided. For cluster analysis, only temporary files will be stored there for the training processes. For publication, they will be submitted and stored in their final version (including raw data and its metadata) in a repository under CC BY-SA.

Test data Test data is part of the dataset used for the training and validation, in other words a subset of the main dataset. Its purpose is to be used with the final predictor to obtain a **final** and unbiased evaluation of the predictor you are building. For this reason, the test data is stored and isolated from the original dataset.

- *Data Collection*: The selection of the elements in the test data are to be chosen in a way that it guarantees a variability in the input when testing the models. It shall be selected before the model building starts, but it must have the same process of data wrangling as the data training and validation sets. The parameter for choosing the test need to be defined before the model training starts.
- *Hardware*: A computer and the external storage unit.
- *Software*: Text editors for code, web browser for ELN access for documentation.
- *Responsibility for collection*: The PI and the experimentator.
- *Project Phase*: Phase I and Phase II
- *Storage*: Test data shall be stored as raw data and as wrangled data in the AG Jimenez (for secure) and only transferred to the File storage during the final evaluation of the model.

Teaching material As aim of this project (and the lab in general) is to set the basics material for the teaching of Machine Learning for Life scientist, specially toxicologist. There are two target audiences: The PI and PhD student directly involved in the project, and the students that will attend the Computational Toxicology Summer School we will organize as part of the grant application.

5.2 Data organization

How will you organise your data? Will the data be organised in simple files or more complex databases? How will the data quality during the project be ensured? If data consists of many different file types (e.g. videos, text, photos), is it possible to structure the data in a logical way?

The project plan contains three predictors, three datasets for their training, several for benchmarking and hypothesis testing, and Teaching material. Each section of the project will be organized as shown in figure 1. Each Project directory will have a README.md file where the metadata for the directory will be located. It will have as well a Makefile that will allow to reproduce the data analysis. The directory *Data* will contain the raw data, as it was obtained from the machine or downloaded. The derived will contain all intermediate data obtained during its processing.

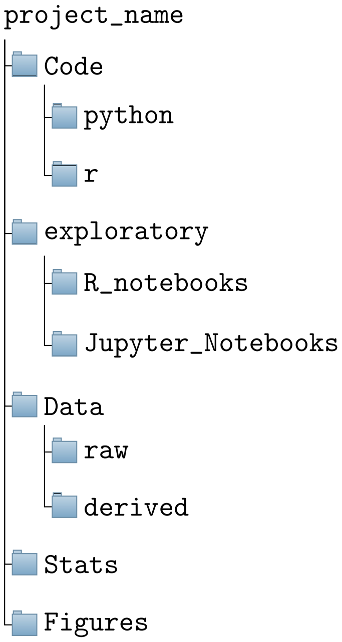

Figure 1: Directory Structure of a project

For compatibility with the FAIR data principles, the files should be the most simple and machine readable formats. For detailed information see section 9.2.

5.3 Data type and size

What type(s) of data will be collected? What is the scope, quantity and format of the material? After the project: What is the total amount of data collected (in MB/GB)?

For the data types, scope, and quantity see section **??**. For the format see section 9.2.

5.4 Data format

In what format will your data be? Does the format change from the original to the processed/final data? Will your (final) data be available in an open format?

The formats for data will be the same as it will be stored and archived (see section 9.2 for details of file format)

The raw data will be kept unchanged under the raw data directory and stored unmodified in the AGJimenez/Project_name/RawData/ file. The directory address, names, content and origin for each file collected will be documented in the ELN and in the README.md file in the raw file.

5.5 Directory structure and names

How will you structure and name your folders/directories? For the definition of each type of storage, see section 7.1

- Cloud Storage: Structure is given by the ELN.

- NSA Storage: Each Member will have its own directory where the backup of the data will be stored under AGJimenez/Project_name/LastName_Name and the structure of data as defined in figure 2.
- File Storage: This is the storage in the computer of each member. The Disk space project related should be structured as in figure 3.

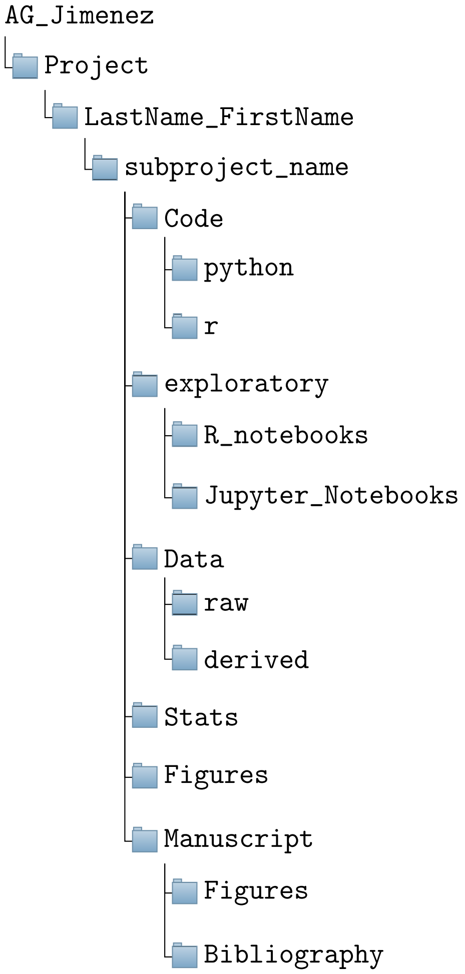

Figure 2: Directory Structure in the NAS (AG_Jimenez)

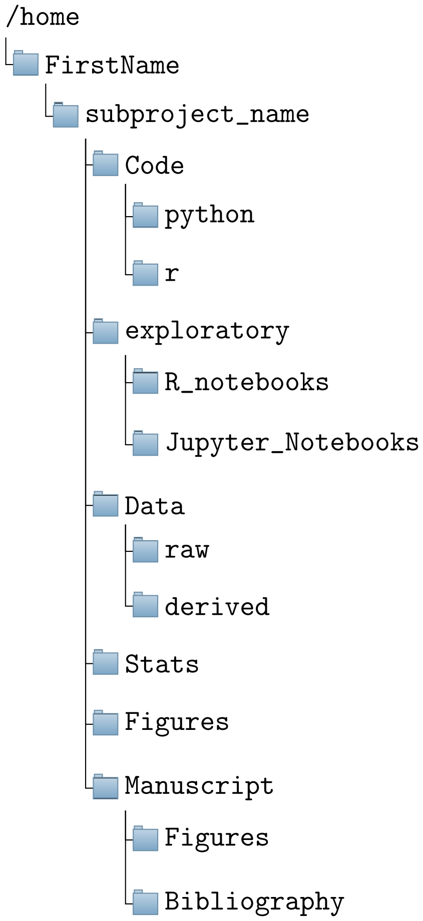

Figure 3: Directory Structure in the File Storage

5.6 File structure and names

How will you structure and name your folders?

Naming files and folders In order to access file names programmatically without the need of complicated Regular Expressions (RegEx), all files and directories should:

- Be descriptive but short

- For names with several words, they shall be attached to each other using a "_". Example: raw_data
- Although linux does not require the file extension, windows does. In order to keep the FAIR principles, you need to add the file extension.
- Keep the rules of file naming for linux. You will find them under Make your s

File structure The File structure can mean the structure in each file or the hierarchical organization of files and directories. Since I have already described the hierarchical organization of directories and what they need to have, I am describing here the file structure inside the file. There are three types of files: Text, Source and bits. I will explain the organization of the Text and Source files. Since at the moment I cannot see a purpose for storage of files in bits, there is no reason for its description.

Text files These include the formats fasta, genbank, markdown, text, comma and tab separated values. For fasta format details, see the corresponding wikipedia article. For genbank format , see the NCBI information website. Text files with extension *.txt are files that have been created in simple text editors as geany, vim, nano, or kate (linux). Comma and tab separated values (*.csv, *tsv) are a kind of text files that contain values separated by comma (,) or tab (\tab), and rows separated by new line character (\n).

Source files Under this types of files are all files that contain programing instructions and will be interpreted into machine language by an interpreter like python, R or Java. All source files should contain after the shbang

5.7 Documentation

What documentation will be created during the different phases of the project? How will the documentation be structured?

5.8 Metadata

What metadata will be provided with the collected/ generated/ reused data? How will metadata for each object be created? Is there any program that can be used to document the data? Can metadata be added directly into the files or will the metadata be produced in another program or document?

5.9 Metadata standard (if applicable)

What metadata standard(s) will you use?

6 Processing your data

6.1 Versioning

What is your strategy concerning versioning your data files (and scripts) during the project? Will you create and/or follow a convention for versioning your data? Who will be responsible for securing that a “Masterfile” will be maintained, documented and versioned according to the project guidelines? How can different versions of a data file be distinguished?

6.2 Interoperability

Will you make use of established software and hardware? If not, how does the software and hardware you use relate to other research?

If applicable: Will you make use of established terminologies/ontologies (i.e. structured controlled vocabularies) in the project? If not, how do your terminologies relate to established ones? Which coding is used (if any)? Will you build on established coding schemes? If not, how does your coding relate to other research?

6.3 Data quality

How will data quality be evaluated? What data quality control measures will be used?

7 Storing data and metadata

7.1 Storage

How and where will the (meta)data be stored during the project? For how long will the (meta)data be stored?

Data storage types (as found under RedHat) are:

- Software-defined storage

  Software-defined storage (SDS) uses abstraction management software to decouple data from hardware before reformating and organizing it for network use.
- Cloud storage

  Cloud storage is the organization of data kept somewhere that can be accessed through the internet by anyone—given the right permissions. You don’t need to be connected to an internal network (that’s known as NAS) and aren’t accessing the data from hardware directly attached to your computer.
- Network-attached storage

  Network-attached storage (NAS) makes data more accessible to internal networks by installing a lightweight operating system onto a server that turns it into something called a NAS box, unit, or head.
- Object storage

  Object storage, also known as object-based storage, is a flat structure in which files are broken into pieces and spread out among hardware.

- File storage

  File storage arranges data as hierarchical files that users can open and navigate from top to bottom. Since files are stored on back ends and front ends the same way, users can requests files by unique identifiers such as names, locations, or URLs.
- Block storage

  Block storage splits storage volumes into individual instances known as blocks. Each block exists independently, which gives users complete configuration autonomy.

In general all projects in the lab will use Cloud Storage for the ELN, NAS for the data with sporadic access, minimal manipulation, and high volume data types (raw data, images, sequences, datasets), and File storage in the members' computers for data that has to be used regularly during the analysis, experiment and project. This will include a copy of the raw data, the code, the exploratory files, the figures and stats (as shown in figure 1) Each directory shall have a README.md file with the description of each file found in the directory and the script that originated the file.

All metadata inside the files originated in other programs, has to be downloaded with the file and labelled with the identical name of their corresponding file and a "_meta" to be able to correlate the files later on.

The metadata and the data will all be stored at least 10 years in the Institute's long term storage, where it will be deposited as soon as the project is ended and/or submitted for publication and a repository for open access.

For long term storage, projects will be stored under a directory with the name of the person working on the project. Each directory will have a README.md explaining each subdirectory, which in itself there will be another README.md file with exactly the same purpose.

7.2 Backup

How, where and at what intervals will the (meta)data be backed-up? How will data be recovered in the case of a (meta)data loss incident?

For the cloud storage (ELN) the data will be backed up every day from Monday to Friday at 6:00 in the server and a copy placed in the internal NAS of the institute.

All data in the NAS is backed up to a backup server located in the Institute daily, and the backup server is backed up once a week to magnetic tape storage.

I will recommend each student to have a external hard disk inside the institute with their data backups of their laptops in case the computer will be stolen or damaged, This is necessary since the project requires regular visits to the collaboration partners in Garching increasing the probability of damage or loss.

7.3 Security

How will sensitive (meta)data be protected? (if applicable) How will (meta)data access be managed?

The data collected has no sensitive data, therefore it requires no extra security for access apart from data pirates or hackers.

All servers and data are behind the firewall of the LMU university and LRZ.

8 Protecting your data

8.1 Ethical Review (if applicable)

Does your project require approval by a local ethics committee? How will possible ethical issues be taken into account, and codes of conduct followed? Not that I am aware of. This has been already discussed with the biosafety office of the LMU (Dr. Ferdinand Neuberger) and the permits for the laboratory permits including generation of Genetic Modified Organisms (GMOs) containing candidate toxins has been started (June 22th, 2022, Dr. Isabel Müller).

8.2 Informed consent (if applicable)

Do you require informed consent for your project? If so, how will permission be obtained? How are consent files organised and stored?

This project does not include data that requires an Informed consent.

8.3 (sensitive) Personal data /confidential information (if applicable)

How will access to (sensitive) personal data during the project be controlled? How will collaborators be granted access to the data in a secure way? If the research project is going to have data that includes confidential information or information that requires informed consent, is there a requirement to notify a privacy officer? Is there any confidential information within the material that requires special treatment and/or limits the access to it during/after the project? How will the material be protected during/after the project? How will permissions and restrictions be enforced?

All projects do not include personal data and/or confidential information since we are using an open science policy.

8.4 Intellectual property rights (IPR) or Copyrights

Are there IPR or copyright issues to consider? Will permission be needed to collect/reuse the data? Will these rights be transferred to another organisation for data distribution and archiving?

8.5 Agreements (if applicable)

What are the agreements with other stakeholders?

This I need to discuss with Prof. Burkhard Rost.

8.6 Restrictions (if applicable)

Are there any other restrictions that need to be considered?

None that can be defined at the moment (June 2022)

9 Archiving and publishing your data

9.1 Archiving

How and where will the data be stored after the project's completion? Will you archive your data in a trusted data repository? Will the application of a persistent identifier to your data be ensured? The data will be archived internally in the Institute Long term storage which is guaranteed to store the data for 10 years. The data for publication and open access will be stored in the LMU repository Open Data LMU, which guarantees a life long storage and access to the data.

9.2 Data formats

What formats will you provide your data in for archiving (and sharing)? Will specific software be required to process your data? Can this software be deposited with the data? Since we will be doing all research based on open science principles, all data (for public acces or not) should be stored ins simple, non-propietary, easy machine accessible file formats. These include:

- comma separated values (*.csv)
- tab separated values (*.tsv)
- fasta files (*.fa, *.fasta)
- plain text files (*.txt)
- raw and tiff image files (*.raw, *.tiff, *.tif)
- genbank format with features (*.gb)
- compressed files in *.tar.gz, *.zip, *.h5
- code related (which are plain text, but have other extensions) like python 3 (*.py) and R (*.R).
- exploratory notebooks in R Markdown (*.Rmd) and Jupyter Notebooks (*.ipny)
- Documents in LaTex (*.tex) and Markdown (*.md)
- Data summaries and ELN in html (*.html)

9.3 Access (if applicable)

Will your data be available (Open Access)? Will all data or only parts of it be published? What licenses do you need for your data? How should your data be cited when reused? Will there be an embargo period for (all or some of) the data? Are there other agreements or restrictions (see above) that need to be considered? Are there any legal/ethical restrictions that prevents the publication of all the material? Will these restrictions mean that action must be taken before the material can be made available? Is there a risk of delayed publication/making data available (all or parts of)? If so what might be needed to do to avoid this?

All data and the code used for the processing of the data will be available in open access repositories (Open Data LMU), as long as no agreement has been reached with sources that limit their publications, as it might be the case with the benchmarking data. In that case, only the data processing and results will be available.

All data published will be licenced under CC BY-SA 4.0 (For more information about the current Creative Commons Licenses go to the Appendix A)

Data should be cited using the DOI given by the repository and the publication associated with the data.

Only on pre-registrations there will be an embargo of up to two years hoping that the publication has been published. The embargo will be lifted as soon as the manuscript has been submitted for revision.

At the moment there are no ethical or legal reasons to not to published the data.

10 Discovering data

10.1 Identification of needs

Do you plan to use existing data for your research? What is the purpose for which you need the data? What do you want to learn from the data? What type of data do you need?

The project will require the use of existing data stored in databases as fasta sequences. This includes bacterial proteins, bacteria genes and phage DNA sequences. The purpose of the data are to use them for training, validation and test of predictive models that will be built during the project.

10.2 Search for data

Do you know where the data may be located? How do you plan to search for the data? Evaluation of data quality What is the minimal required quality of the data (in terms of origin, contents, scope, size, methods, etc.)? How do you plan to evaluate data quality (evaluation of metadata, tests, analysis, comparisons)?

10.3 Gaining access to data

What are the (expected) terms and conditions for data access and use? What is the (expected) process for gaining access to the data? What is the (expected) time-span of the process for gaining access to the data? What are the (expected) costs for data access and use?

A Creative Commons licenses 4.0

This is the text currently published under the Creative Commons website (Access date: June 23rd, 2022)

The Creative Commons License Options There are six different license types, listed from most to least permissive here:

A.1 CC BY

Definition This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use. CC BY includes the following elements: BY – Credit must be given to the creator

Text for licence This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

A.2 CC BY-SA

Definition This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use. If you remix, adapt, or build upon the material, you must license the modified material under identical terms. CC BY-SA includes the following elements: BY – Credit must be given to the creator SA – Adaptations must be shared under the same terms

Text for licence This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

A.3 CC BY-NC

Definition This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator. It includes the following elements: BY – Credit must be given to the creator NC – Only noncommercial uses of the work are permitted

Text for licence This work is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

A.4 CC BY-NC-SA

Definition This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator. If you remix, adapt, or build upon the material, you must license the modified material under identical terms. CC BY-NC-SA includes the following elements: BY – Credit must be given to the creator NC – Only noncommercial uses of the work are permitted SA – Adaptations must be shared under the same terms

Text for license This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

A.5 CC BY-ND

Definition This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, and only so long as attribution is given to the creator. The license allows for commercial use. CC BY-ND includes the following elements: BY – Credit must be given to the creator ND – No derivatives or adaptations of the work are permitted

Text for license This work is licensed under the Creative Commons Attribution-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nd/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

A.6 CC BY-NC-ND

Definition This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, for noncommercial purposes only, and only so long as attribution is given to the creator. CC BY-NC-ND includes the following elements: BY – Credit must be given to the creator NC – Only noncommercial uses of the work are permitted ND – No derivatives or adaptations of the work are permitted

Text for license This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

B Agreement to follow the DMP

Munich, June 2022.

Dear Member / Associate Member of the Exotoxins lab,

Our lab is following the Open Science principles. You have been given the Data Management Plan (DMP) for this project. Please read it. Its purpose is to guarantee that we will have the same standards in data collection so that we can fulfill the FAIR data principles. In short, data must be Findable, Accessible, Interoperable, and Reusable.

These procedures have been created to guarantee that all your work and data can be used and reproduced in the future. This is essential in the advancement of science and part of Good Scientific Practices.

I am sure you see the important role Data plays in your work here. Sadly, as Data Steward, I need to place certain safeguards to make sure that everybody follows the Data Management Plan.

For students: Please follow the data management plan and its principle idea: Make your data FAIR. This will benefit you directly when it is time to write your report and thesis. I am afraid that failure to do so, will force me to retain any kind of certification you need for your studies, including revoking your Bachelor, Master, or PhD thesis.

For associated members: I know that the DMP might be new to you. This implies that getting used to other type of work typically done in your current or previous labs, will be difficult and you might struggle to follow the Plan. However, as Data Steward I need to guarantee the resusability and reproducibility of the data in future years. This is part of a proper scientific practices. I am afraid that if your data and work does not follow the DMP, I cannot include you or your data in any publication, since there is no way to guarantee that your data is reproducible. Any report regarding your work in our lab will reflect the lack of compliance with the data policies of the laboratory.

I hope you understand the reasons I have to place these safeguards, and you agree to them.

If you agree to follow the DMP and acknowledge the consequences of not doing so, please sign this form below and give it to me for documentation.

I am looking forward to work with you a create a stronger Data Management Plan that will guarantee a solid base for future research.

kind regards,

Dr. Luisa F. Jiménez-Soto

Hereby I, _______________________ , acknowledge that I have read and understood the DMP from the Research group under PD. Dr. Luisa F. Jiménez Soto. I am aware of the consequences it will bring if I do not follow the procedures placed in order to guarantee Good Scientific Practices regarding Data and its reproducibility.

Signed on: Munich, July 11, 2022
Signature:

C FAIR data principles

Taken from the Leibniz Informationszentrum (TIB, 2022) and modified it for clarity. Other sources can be found in section E

FAIR means the data shall be...

Findable

In short words...

Any data you produce has to be identifiable with an unique id, in a online resource, and with its correspondent description

F1. Metadata and Data are assigned a globally unique and eternally persistent identifier This is the reason for DOI for papers, IDs for proteins and sequences, DOI for protocols.

F2. Data are described with rich metadata Each data produced has to have very descriptive metadata that can help to understand and/or reproduce the data.

F3. Data and Metadata are registered or indexed in a searchable resource Data and its metadata have to be registered in an accessible source like repositories or publications.

F4. Metadata specify the data identifier The metadata is not only a generic description. It must describe exactly unique data and reflect its uniqueness.

Accessible

In short words...

Data and metadata have to be archived for long-term storage and made available so that machine and human can retrieve it.

A1. Data and Metadata are retrievable by their identifier using a standardized communications protocol In a world with internet, any standard communication protocol like ftp or https should be capable to access the data. However, it must be clarified before hand what kind of access will be possible and how it can be done.

A1.1 The (access) protocol is open, free and universal implementable It cannot be a personalized protocol or proprietary. Everybody should have access with the standard access protocols.

A1.2 The (access) protocol allows for authentication and authorization procedure when necessary Certain data can be considered sensitive. For this reason, they must have clear authorization procedures to guarantee a free access to it.

A2. Metadata is accessible even when the data are no longer available In cases where data is too big or papers have been retracted, the metadata must persist associated with the identifier and must be accessible.

Interoperable

In short words...

Any standard open software or program should be able to access and use the data and it should not be dependent of commercial software or operating systems.

I1. Data and Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation Controlled vocabularies or ontologies (common concepts for classification) should be used to described the datasets have to be properly identified and documented. If a search is needed using a computer, it should be possible independent of the OS or software.

I2. Data and Metadata use vocabularies that follow FAIR principles The vocabulary used should fulfill the FAIR principles. If a new format was created, the parser and all what is needed to transform the data to standard data structures should be included.

I3. Data and Metadata include qualified reference to other data and metadata Of a dataset used another dataset as reference or source, this has to be documented and explained in the data and metadata. Any kind of scientific relation between datasets should be known and disclosed.

Reusable

In short words...

The data and its metadata have to be published in a way that it can be reusable by future scientist and machines.

R1. Data and Metadata have a plurality of accurate and relevant attributes The metadata has to reflect the process used for its creation or analysis. This might include machine used, software version used, etc. Whoever wants to use the data should understand the origin of it.

R1.1 Data and Metadata are released with a clear and accessible data usage license The licence and legal conditions of the use of the data have to be clearly defined and declared for machines and for humans.

R1.2. Data and Metadata are associated with their provenance How and who made a dataset. This should be reflected in both data and metadata.

R1.3. Data and Metadata meet domain-relevant community standards If standards or best-practices for data archiving have been created by the community, they must be fulfilled when posting the data. This includes the data format, data structures, information, etc.

D Practical approach to the data collection and storage

In this section I will try to explain the day-to-day practices in the lab and in the data analysis you will need to follow in order to guarantee FAIR data. As a reminder, FAIR means Findable, Accessible, Interoperable and Reusable. With this principles in mind, here are the main practices:

D.1 Documenting an experiment

- These check lists are to guarantee that you have a clear experimental design, with hypothesis declared, and in line with Open Science and good Scientific Practices.
- The format you will find in the ELN is on section D.4 of this document. it reflects the main documentation required in a pre-registration, which is part of the planning of an experiment. The check list unify the design and the processing of data in an experiment.

  □ Open a new experiment in the ELN

  □ Document the hypothesis or the purpose of the experiment in case it is exploratory.

  □ Describe the number of biological replicas and if needed the number of technical replicas.

  □ Describe the purpose of each of the experimental treatments or conditions for each sample. Keep in mind your positive and negative controls.

  □ Describe the parameters you will use to reject the experimental data (p.e. The controls did not work as supposed, the machine gave an error, the variability between technical replicas is over a defined threshold)

  □ Define the type of measurements you will obtain. Please, try to design your experiments in a way that **only quantitative data** is obtained.

  □ Create the directory structure for your data if it has not been yet done (see figure 1 of the DMP).

  □ Obtain a sample of the expected output data.

  □ With the sample data, prepare your exploratory data (RMarkdown, Jupyter Notebook)

    – Things you need to have in your exploratory data. All code chunks have to be properly documented with the purpose of the code and in consecutive order.

      □ Descriptive Statistics- Univariates. This verify that data integrity is given before the analysis starts.
      □ Data wrangling preparing the data for analysis
      □ Descriptive Statistics - Bivariates.
      □ Your analysis

  □ Plan the time schedule for the experiment.

  □ Define and document the protocol you will follow during the experiment. If a new protocol is being developed, enter the protocol in the ELN database (see check list for Protocol in section D.3).

  □ Verify that the machine you need is available, functioning and you have been instructed in its use and what to do in case something does not work.

□ Verify that you have all reagents, kits, equipment and biological organisms available.

□ Verify that you know and can apply the safety measures for each protocol or procedure the experiment requires.

D.2 ... you think you need to order a new reagent

... before ordering it

- This is to make sure that we do not order things we already have and that you know exactly what you are ordering.

  □ Check the storage conditions.

  □ Check that the chemical or reagent is not in the lab. Verify in the database and in the laboratory area designated for its recommended storage.

  □ Check that the storage conditions can be met in the lab.

  □ Check the Material Safety Data Sheet (MSDS). Evaluate the health and safety risks. If the risks for environment and health are too high, search for alternatives that are less dangerous.

  □ If the reagent has high safety risk for health and environment, and no alternative is possible, discuss with your PI the need of the purchase.

  □ Get approval of purchase.

  □ After ordering, document it in the Lab Purchases Book

... after ordering it

- This is to document all and have it ready for the arrival of the chemical / reagent.

  □ Download the MSDS and/or the manual as PDF.

  □ Start the Database entry in the ELN. Enter the name in English and use the shortest and most common name used.

  □ Add catalog number, company name and date of purchase. Attach the MSDS and Manual as files.

... when it arrives

- You need to make its storage safe. Since the arrival can happen when you are not here, ***these steps should be followed by anyone in the lab receiving any package.***

  □ Open carefully the package.

  □ Read the documentation delivered and on the container.

  □ Pay special attention to the warning signs and storage temperature.

  □ Label the container with ***AG Jimenez***

  □ Place the reagents in the designated storage temperature.

□ Look at the Database entry in the ELN to document the date of arrival and use your initials so we know who received it.

□ Check in the Lab Purchases book who made the order.

□ Inform the person who ordered it of its arrival and storage area.

D.3 Protocol design and documentation

Protocols have to be documented as following

D.3.1 Title

Make it short and to the point

D.3.2 Purpose

Name the main objective of the protocol and the variables to be considered and/or obtained

D.3.3 Materials

Make a list of the materials needed to perform the experiments using this protocol. Include reagents and machines and their accessories. Since the reagents should be already in the database, link them to make sure we can document the company, catalog number and lot used.

D.3.4 Procedure

Described the procedure step-by-step in a enumerated list.

D.3.5 Modifications

If following modifications were made, write your name, date, modification and result of the modification.

D.3.6 References

Enumerate all documentation you used to assemble the protocol.

D.4 Experimental design

The documentation has been described in the check list of section D.1.

All experimental designs have to answer the following:

D.4.1 What is the hypothesis that will be investigated?

State the hypothesis or hypotheses you will be investigating in this experiment.

D.4.2 How will the crucial variables be operationalized?(Variables to be measured)

Dependent variables Which are the dependent variables, what will they measure, and how will they be obtained? Will they need transformations?

Independent variables Which are the dependent variables, what will they measure, and how will they be obtained? Will they need transformations? Is co-linearity expected?

D.4.3 What is the source of the data included in the analyses?

Is it experimental? or acquired from data available already?

D.4.4 Are there any exclusion criteria for the data?

Clarify the criteria you will use to exclude the use of the data obtained. Be real clear about it. One example is:

Data will only be excluded if one of the following are observed:

- *The positive or negative controls show a variation between biological replicas of more than 30%.*
- *Technical replicas variance is higher than 10%*
- *A positive or negative control do not fit the model created with previous data.*
- *The results from equations 1, 2 and 3 variate more than 2% from the initial measurements (Control for machine light sources and detection, as well media control)*
- *Time points between measurements variate more than 30 min between experiments.*

D.4.5 What are the planned statistical analyses?

Before an experiment begins, the data analysis should be clear. Here is an example:

1. ***Data wrangling**: Measurements will be in 96-well plate format (obtained as comma separated values (*.csv)). A template for data transformation will be created to format the data to fulfill the three principles of tidy data (Wickham 2014): One variable per column, one observation per row and one subject per table. For this a parser in python will be used.*

2. ***Modeling of positive and negative controls**: All descriptive analysis as well creation of models will be done using the flexplot package in R (Fife et al 2021).*

3. ***Hypothesis testing**: Will be done using the flexplot package in R (Fife et al 2021) using Generalized Linear Models (GLMs) for exploratory and confirmatory assays (Tukey 1998).*

D.4.6 What are the criteria for confirming and disconfirming the hypotheses?

Describe the alpha value or other parameters you will use to confirm or disconfirm your hypotheses. As an example:

For the evaluation of hypothesis testing a P-value of less than 0.05 (alpha value) will be considered significant. For P values between 0.05 and 0.06 Bayes factor and AIC will be considered based on the documentation and criteria of the flexplot package for difference between models.

D.4.7 Have the analyses been validated on a subset of the data? If yes, please specify and provide the relevant files

D.4.8 What is known about the data that could be relevant for the tested hypotheses?

You can add here any data from previous or reference publications that are related to your experimental setup.

D.4.9 Please provide a brief timeline for the different steps in the preregistration

This is just to have an idea of how long an experiment should take including the data analysis.

D.4.10 Transparency statement

Following Open Science ideas, all members involved in this project are aware of the FAIR data principles and will follow the Data Management Plan of our Research Group. Information and training required to fulfill this, will be provided before the project starts. All processes will be documented in the Electronic Lab Book and in the code.

D.4.11 Conditional safeguards

Think on all things that can go wrong and explain here what you will do in case something like it happens. Here is an example I did for a project planned for pregrade students: *Since this project will be part of the training of students, there might be delays we cannot foresee at the moment. Therefore, priority will be given to the documentation and modeling of controls for asexual life cycle, followed by the sexual life cycle evaluation.*

D.5 Linux / Unix File Naming

From Center for Imaging Science document. Some modifications were made to adapt it to the context of this DMP.

File Names UNIX permits file names to use most characters, but avoid spaces, tabs and characters that have a special meaning to the shell.

Case Sensitivity uppercase and lowercase are not the same! These are three different files:
NOVEMBER November november
Length: can be up to 256 characters
Extensions: may be used to identify types of files
libc.a - archive, library file
program.c - C language source file
alpha2.f - Fortran source file
xwd2ps.o - Object/executable code
mygames.zip - Compressed file mygames.tar.gz - Compressed file

Hidden Files : have names that begin with a dot (.) For example:
.cshrc .login .mailrc .mwmrc

Uniqueness : as children in a family, no two files with the same parent directory can have the same name. Files located in separate directories can have identical names.

Reserved Filenames This are file names that you need to know (and you cannot use for your file names)
/ - the root directory (slash)
. - current directory (period)
.. - parent directory (double period)
 - your home directory (tilde)

Pathnames Specify where a file is located in the hierarchically organized file system. You must know how to use pathnames to navigate the UNIX file system

Absolute Pathname : tells how to reach a file begining from the root; always begins with / (slash).
For example: /usr/local/doc/training/sample.f

Relative Pathname : tells how to reach a file from the directory you are currently in (current or working directory); never begins with / (slash). For example:
training/sample.f
../bin
/projects/report.001

For example, if your current directory is /usr/home/johnson and you wanted to change to the directory /usr/home/quattro, you could use either of these commands:

cd ../quattro - relative pathname

cd /usr/home/quattro - absolute pathname

E Sources

1. FORCE11 (2017): The FAIR Data Principles. https://www.force11.org/group/fairgroup/fairprinciples

2. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: Scientific Data 3, 160018 EP -. DOI: 10.1038/sdata.2016.18.

3. Schultes et al. (2017): The FAIR Data Principles explained. https://www.dtls.nl/fair-data/fair-principles-explained/

4. Swiss National Science Foundation (SNF): Explanation of the FAIR Data Principles.http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf

5. Mertens, Gaëtan, and Angelos Miltiadis Krypotos. 2019. "Preregistration of Analyses of Preexisting Data." Psychologica Belgica 59 (1): 338–52. https://doi.org/10.5334/PB.493/GALLEY/458/DOWNLOAD/.

6. Miltenburg, Emiel van, Chris van der Lee, and Emiel Krahmer. 2021. "Preregistering NLP Research." In , 613–23. Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2021.naacl-main.51.

7. Casadevall, Arturo, Lee M. Ellis, Erika W. Davies, Margaret McFall-Ngai, and Ferric C. Fang. 2016. "A Framework for Improving the Quality of Research in the Biological Sciences." MBio 7 (4). https://doi.org/10.1128/MBIO.01256-16.

8. Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. "The Preregistration Revolution." Proceedings of the National Academy of Sciences 115 (11): 2600–2606. https://doi.org/10.1073/PNAS.1708274114.