HANDS-ON: WRITING A DATA MANAGEMENT PLAN

# Instructions

Below you will find a fictional data management plan inspired by real research projects. The DMP is based on the Flemish Standard Data Management template.

The classroom will be divided into five different groups. Each group chooses a representative who will present the conclusions of the group for the classroom. Each section of the DMP will be addressed by a different group:

* [GROUP 1: Ethical and legal issues](#_fwed9fl7l9yg)
* [GROUP 2: Documentation and Metadata](#_2nl97gj8nvxs)
* [GROUP 3: Data storage & back-up during the research project](#_iumptng066tt)
* [GROUP 4: Data preservation after the end of the research project](#_qtte6of9lsa0)
* [GROUP 5: Data sharing and reuse](#_z5jcdzrthd4s)

Exercise step by step:

1. With your group, read the first sections (project information and research data summary) to get an idea about what the objectives of the project are and what type of data will be generated during the project.
2. Then move into the section of the DMP which corresponds to your group and read the questions and answers provided. You may read other sections of the DMP, but you may only modify your group’s DMP section.
3. Discuss with the group what your opinion is about the provided answers, taking into account the characteristics of the data described in the research data summary section. Use the evaluation checklist at the end of this document to help you out.
4. Modify and/or elaborate on the answers as discussed with the group directly in this document.

After all the groups have gone through their specific section, we look at the results and the whole DMP together with the classroom.

# The Data Management Plan

## General Project Information

For many solid cancers a tissue biopsy is required to confirm diagnosis and to decide on treatment. However, collecting a tissue biopsy is invasive for the patient. Liquid biopsies, such as blood, offer a less invasive and cost-effective alternative. Moreover, several samples can be collected over time (longitudinal sampling) providing an easy way to monitor treatment response.

Blood plasma is a well studied type of 'liquid biopy'. Tumors shed (actively and passively) genetic material into the bloodstream, which can be used to molecularly characterize the tumor. However, only a small portion of the genetic material originates from the tumor. For that reason it is crucial that detection assays have a high sensitivity. In this project we focus on mutation detection using both cell-free DNA (cfDNA) and extracellular RNA.

Plasma will be collected from cancer patients with a known mutation, before treatment. cfDNA and extracellular RNA will be purified from the plasma followed by analysis of both. For this we will use digital PCR and deep DNA/RNA sequencing for targeted detection of genetic variants to assess whether this workflow has the potential to be implemented in routine clinic.

## Research Data Summary

*List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.*

*If you reuse existing data, please specify the source, preferably by using a persistent identifier   
(e.g. DOI, Handle, URL etc.) per dataset or data type:*

Only new data will be generated during the project.

**Clinical data:**

Clinical information of cancer patients, including:

* Baseline data: containing general demographic information about the patient
* Baseline tumor characteristics: clinical and pathological TNM staging and mutation information
* Treatment plan: the treatment patients receive at the moment they were consulted for this study

**Data:**  
DNA- and RNA-seq data: We start from digital bcl files returned from the Illumina sequencer and convert it to fastq files (raw data) based on the indexes. The fastq files are stored on the department’s network drive. The intermediate bam files will be temporarily stored.

Digital droplet PCR (ddPCR) data: Data is generated by the Bio-Rad QX100 Droplet Reader and results can be analyzed using the QuantaSoft software. For each run, the following files are generated:

* Raw data (.qlb) - unprocessed data from each well in a ddPCR plate.
* Results (.qlp) - used-defined plate layout settings and processed data for a ddPCR plate
* Comma-separated values (.csv) - analyzed data in a format that can be assessed using other programs, such as Microsoft Excel.

Experimental data files: Data generated in the lab in excel, pdf, jpeg, qlb, qlp, csv and other formats are stored on SharePoint. Electronic lab notebooks, analysis notes and quality control documents (containing excel-files, word documents, powerpoints and pdf files) are stored on SharePoint which includes version control

**Scripts:**

Python and R scripts will be generated for data analysis purposes. The scripts will be versioned and managed using Github.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset name** | **Description** | **New or reused** | **Data type** | **Data format** | **Volume** |
| Clinical data | Clinical information of cancer patients: baseline demographic information, tumor characteristics, treatment | New | Observational | eCRF REDCap database files | < 1GB |
| DNA- and RNA-seq data | cfDNA and extracellular RNA sequencing data obtained from plasma samples of cancer patients | New | Experimental | .bcl , .fastq , .bam , .vcf | < 5TB |
| ddPCR data | Digital droplet PCR data generated by Bio-Rad QX100 Droplet Reader on plasma samples of cancer patients | New | Experimental | .qlb , .qlp , .csv | < 100GB |
| Lab experimental data | Data generated in the lab when preparing and analysing samples; including data of quality control steps. | New | Experimental | .xlsx , .csv , .pdf , .jpeg , other | < 100GB |
| Documentation files | Electronic lab notes, analysis notes and quality control documents | New | Other | .xlsx , .docx , .pptx , .pdf | < 1GB |
| Scripts | Data analysis scripts | New | Software | .py , .R | < 1GB |

## GROUP 1: Ethical and legal issues

|  |  |  |
| --- | --- | --- |
| **Question** | **Answer** | **Your comments** |
| *Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.* | Yes, human subject data  Cancer patient samples will be processed and clinical information will be used.  We foresee to receive patient plasma samples from different groups:   * Cancer Clinic, UZ Ghent * Hanaa Biotechnology SAS, Lyon, France |  |
| *Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.* | Yes  Health data and demographic information of cancer patients. Necessary general clinical information relevant to the project will be available and will be stored in the REDCap database, including age, type of tumor, gender, etc.  Study participants will be informed and asked to agree to sign an ICF before inclusion in the study. |  |
| *Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.* | Yes, the method of identifying mutations in cancer diagnostics can be patented. |  |
| *Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use  (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.* | No |  |
| *Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.* | No |  |

## GROUP 2: Documentation and Metadata

|  |  |  |
| --- | --- | --- |
| **Question** | **Answer** | **Your comments** |
| *Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, “Electronic’ Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).* | Throughout the workflow, starting from plasma samples, electronic lab notebooks will be made with detailed information on the experimental work, sample labelling, dates, persons handling the samples and performing the protocols. All samples will have unique codes that are used to label the samples and to annotate them in the electronic lab notebooks. Sample specific information can be found in the REDCap database. |  |
| *Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used.*  *If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.* | Yes, the sequencing data is kept in standard and community recognized formats such as fastq and vcf. |  |

## GROUP 3: Data storage & back-up during the research project

|  |  |  |
| --- | --- | --- |
| **Question** | **Answer** | **Your comments** |
| *Where will the data be stored?* | On a network drive and on the project’s SharePoint site. All data will be linked to the corresponding documentation (eg lab notes) by means of a code in the file name.  Scripts will be in the repository for this project in my Github account. |  |
| *How will the data be backed up?* | Clones of the Github repository allow for redundant storage locations.  All information on SharePoint is automatically versioned. |  |
| *Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.  If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.* | Yes |  |
| *How will you ensure that the data are securely stored and not accessed or modified by unauthorised persons?* | Only project team members have access to the REDCap database, and the password to my Github account is confidential.  Institutional security guidelines will be followed. |  |
| *What are the expected costs for data storage and backup during the research project? How will these costs be covered?* | Storage for sequencing data is covered by project funding.  My Github account is for free. |  |

## GROUP 4: Data preservation after the end of the research project

|  |  |  |
| --- | --- | --- |
| **Question** | **Answer** | **Your comments** |
| *Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).* | All data will be kept for at least 5 years. |  |
| *Where will these data be archived (stored and curated for the long-term)?* | The data will remain on the network drive and project SharePoint site.  Any resulting article publications will contain source data for the images included therein. |  |
| *What are the expected costs for data preservation during the expected retention period? How will these costs be covered?* | No |  |

## GROUP 5: Data sharing and reuse

|  |  |  |
| --- | --- | --- |
| **Question** | **Answer** | **Your comments** |
| *Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.* | Sequencing data (DNA and RNA) will be made available.  Clinical data will not be made available for reuse. |  |
| *If access is restricted, please specify who will be able to access the data and under what conditions.* | No one can use the clinical data. |  |
| *Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.* | Yes, there is valorization potential and therefore all information is confidential until a patent has been filed. |  |
| *Where will the data be made available? If already known, please provide a repository per dataset or data type.* | Sequencing data will be deposited in the European Nucleotide Archive (ENA). |  |
| *When will the data be made available?* | At the time of publication, after patent filing. |  |
| *Which data usage licenses are you going to provide? If none, please explain why.* | A license will be agreed upon with parties interested in applying the patented method. |  |
| *Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.* | Sequencing data submitted to ENA will have an accession number for the study and each experiment within. |  |
| *What are the expected costs for data sharing? How will these costs be covered?* | Data submission to ENA is without cost.  Scripts available on Github are without cost. |  |

## Responsibilities

(We will not tackle these questions in the exercise)

*Who will manage data documentation and metadata during the research project?*

*Who will manage data storage and backup during the research project?*

*Who will manage data preservation and sharing?*

*Who will update and implement this DMP?*

# DMP Revision checklist

## GROUP 1: Ethical and legal issues

* Indicates whether collecting/generating or using the data involves ethical issues that need a formal ethical approval.
  + If so, clearly specifies which issues and which datasets.
* Indicates whether personal data will be used,
  + If so, refers to the specific datasets and provides details about what kind of personal data.
* Indicates whether the research has valorisation/tech transfer potential that will require restrictions on (some of) the data to protect intellectual property.
  + If so, clearly specifies which restrictions will be put in place, and for which data.
* Indicates whether third-party agreements restrict dissemination or exploitation of the data.
  + If there are third-party agreements imposing restrictions, clearly specifies what restrictions, and what data they apply to.

## GROUP 2: Documentation and Metadata

* For each dataset, describes what approach will be followed to capture the accompanying descriptive and contextual information necessary to keep data understandable and usable.
* Clearly indicates whether or not a standard (discipline-specific) metadata schema will be used to describe the data.
  + If (a) metadata standard(s) is/are used, clearly specifies which one(s).
  + If no metadata standard is used, specifies the approach taken to create metadata.

## GROUP 3: Data storage & back-up during the research project

* Clearly describes the location and storage medium that will be used for storing data during research.
* Clearly describes the locations, storage media and procedures that will be used for backing-up data
* Clearly indicates whether or not there currently is sufficient capacity for data storage & backup during the project, and justifies the answer.
  + If available storage and backup capacity is not sufficient, explains how this issue will be resolved.
* Clearly specifies what (if any) costs are expected to be incurred for storing and backing up data during the project (e.g. based on similar research/datasets and/or university policy).
  + States how these costs will be covered if applicable (e.g. by using part of the allocated grant/project budget).
* Clearly describes the measures (in terms of physical security, network security, and security of computer systems and files) that will be taken to ensure that stored and transferred data are safe.

## GROUP 4: Data preservation after the end of the research project

* Clearly describes which (subsets/versions of the) data and accompanying documentation will be retained for preservation and which will be destroyed, and explains rationale (e.g. legal or contractual reasons, practical issues…).
* Clearly describes what means, facilities etc. will be used to effectively preserve digital and non-digital data and accompanying documentation.
* Clearly specifies what (if any) costs are expected to be incurred for preserving data for a minimum of 5 years after the project’s end.
  + States how these costs will be covered,if applicable (e.g. by using part of the allocated grant/project budget).

## GROUP 5: Data sharing and reuse

* Clearly describes which (subsets/versions of the) data and accompanying documentation will be made available to others for reuse.
* Clearly describes who will be able to access (subsets of) the data, and under what (if any) conditions.
* Clearly indicates whether the external sharing of (some) data and accompanying documentation should be restricted or delayed.
  + If so, clearly specifies why.
* Clearly indicates where/how data will be made available to others for reuse.
* Clearly indicates when data will be made available (e.g. immediately after the end of the project, upon publication of the research results, or after an embargo period).
  + If an embargo period is indicated, specifies why and for how long.
* Clearly specifies which licence will be chosen for (subsets of) the data that will be made available.
  + If no licence is provided, clearly explains why.
* Clearly specifies what (if any) costs are expected to be incurred for sharing data.
  + States how these costs will be covered, if applicable (e.g. by using part of the allocated grant/project budget).