# Introduction

#if\_$\_EU $\_PROJECT is part of the Open Data initiative of the EU. ##endif\_$\_EU

To best profit from open data, it is necessary to not only store data but to make data findable, accessible, interoperable and reusable (FAIR). #if\_$\_PROTECT Open and FAIR data, however, considers the need to protect individual data sets. #endif\_$\_PROTECT

The aim of this document is to provide guidelines on principles guiding the data management in $\_PROJECT and what data will stored by using the answers to the EU questionnaire on Data Management Plan (DMP) as a DMP document.

The detailed DMP instructs how data is to be handled during and after the project. The $\_PROJECT DMP is modelled according to the H2020 Online Manual. #if\_$\_UPDATE It will be updated/its validity checked during the $\_PROJECT project several times. At the very least, this will happen at month $\_UPDATMONTH. #endif\_$\_UPDATE

# Data Management Plan EU Template

#if$\_DATAPLANT Currently data management relies on the annotated research context ARC. It is password protected, so before any data can be obtained or samples generated an authentication needs to take place. #endif$\_DATAPLANT

## 2.1. Data Summary

**What is the purpose of the data collection/generation and its relation to the objectives of the project?**

For $\_PROJECT, data collection, integration $\_VISUALIZATION #if$\_DATAPLANT through the DataPLANT ARC structure is absolutely necessary #endif$\_DATAPLANT as data is not only used to understand principles, but it is also used for analyzing data and in the end stakeholder involvement, which need to be informed about data provenance. It is therefore of importance that not only data is well generated, but also well annotated with metadata using *open standards* as it is laid out in the following section. As $\_PROJECT aims at $\_PROJECTAIM

**What types and formats of data will the project generate/collect?**

We foresee that the following data will be collected and generated at the very least: $\_PHENOTPIC, $GENETIC, $\_GENOMIC, $\_METABOLOMIC, $\_RNASEQ data about $\_STUDYOBJECT. In addition, derived data from the original raw data sets will also be collected. This is important, as different analytical pipelines might yield different results or include *ad-hoc* data analysis parts #if$\_DATAPLANT and these pipelines will be tracked in the DATAPLANT ARC #endif$\_DATAPLANT. Therefore, specific care will be taken to document and archive these resources (including the analytic pipelines) as well #if$\_DATAPLANT relying on the vast expertise in the DATAPLANT consortium #endif$\_DATAPLANT.

**Will you re-use any existing data and how?**

The project builds on existing data sets and relies on them. #if$\_RNASEQ For instance, without a proper genomic reference it is very difficult to analyze NGS data sets #endif$\_RNASEQ. It is also important to include existing data sets on the expression and metabolic behaviour of $\_STUDYOBJECT but of course also on existing characterization and the background knowledge #if$\_PARTNERS of the partners #endif$\_PARTNERS. Genomic references can simply be gathered from reference databases for genomes/sequences like the National Center for Biotechnology Information: NCBI (US); European Bioinformatics Institute: EBI (EU); DNA Data Bank of Japan: DDBJ (JP). In addition, prior “unstructured” data in the form of publications and data contained therein will be used for decision making.

**What is the origin of the data?**

Public data will be extracted as described in the previous paragraph. For $\_PROJECT, specific data sets will be generated by the consortium partners.

#if$\_RNASEQ Short read sequencing will be outsourced and raw data will be received. #endif$\_RNASEQ

#if$\_METABOLOMIC Metabolomic data will be generated using chromatography coupled to mass spectrometry and from enzyme platforms mostly #if$\_METABOLOMIC.

#if$\_PROTEOMIC proteomic data will be generated using an EU platform which is in line with community standards #if$\_PROTEOMIC.

#if$\_PREVIOUSPROJECTS data from previous projects such as $\_PREVIOUSPROJECTS will be considered. #endif$\_PREVIOUSPROJECTS

**What is the expected size of the data?**

We expect to generate raw data in the range of $\_RAWDATA of data. The size of derived data will be about $\_DERIVEDDATA.

**To whom might it be useful ('data utility')?**

The data will be useful for the $\_PROJECT partners to the scientific community working on $\_STUDYOBJECT or the general public interested in $\_STUDYOBJECT. Hence, $\_PROJECT also strives to collect the data that has been disseminated and potentially advertise it #if$\_DATAPLANT e.g. through the DATAPLANT platform or other means #endif$\_DATAPLANT, if it is not included in a publication anyway which is the most likely form of dissemination.

## 2.2. FAIR data

### Making data findable, including provisions for metadata

**Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?**

All data sets will receive unique identifiers and they will be annotated with Metadata. $\_PROJECT will rely on community standards plus additional recommendations necessary in the plant field adapted by e.g. using suggestions from the Minimum Information About a Plant Phenotyping Experiment (MIAPPE). These -unlike across domain minimal sets such as Dublin core which mostly defines submitter and what general type of data is being dealt with (e.g. images) - allow reusability by other researchers as it also defines properties of the plant (see preceding section). However, of course minimal cross domain annotations are part of $\_PROJECT. #if$\_DATAPLANT The core integration with DataPLANT will also allow to tag individual releases with a digital object identifier (DOI) #endif$\_DATAPLANT

**What naming conventions do you follow?**

Data variables will use standard names. For example, this is the case for genes, metabolites and proteins. These will also be linked to free biomedical ontologies where this is possible. In the case of datasets, the dataset names will also be made to be meaningful and human readable. In addition, traditional names will of course be included where necessary as synonyms.

**Will search keywords be provided that optimize possibilities for re-use?**

Keywords about the experiment and the general consortium will be included as well as an abstract about the data, where useful. In addition, certain keywords can be auto-generated from dense metadata and its underlying ontologies. #if$\_DATAPLANT Here, DATAPLANT strives to complement these with standardized PLANT ontologies that are supplemented where the ontology doesn’t yet include the variables #endif$\_DATAPLANT

**Do you provide clear version numbers?**

To maintain data integrity and to be able to re-analyze data, data sets will get version numbers where this is useful (e.g. raw data must not be changed and will not get a version number and is considered immutable). #if$\_DATAPLANT this is automatically supported by the ARC Git DATAPLANT infrastructure #endif$\_DATAPLANT

**What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.**

We foresee to use #if$\_RNASEQ|$\_GENOMICS e.g. Minimum Information About a Next-generation Sequencing Experiment MinSEQe for sequencing data and #endif$\_RNASEQ|$\_GENOMICS #if$\_METABOLOMICS Metabolights submission compliant standards for metabolites #issuewarning some Metabolomics partners considers Metabolights not an accepted standard #endissuewarning #endif$\_METABOLOMICS as well as MIAPPE for phenotyping data in the broadest sense, but will also be relying on specific SOPs for additional annotations #if$\_DATAPLANT relying on advanced DATAPLAN annotation and ontologies #endif$\_DATAPLANT

### Making data openly accessible

**Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.**

By default, all data sets from $\_PROJECT will be shared with the community and made openly available. This is however after partners have had the ability to check for IP protection (according to agreements and background rights). #if$\_INDUSTRY This applies in particular to data pertaining to the industry. #endif$\_INDUSTRY However, all partners also strive for IP protection of data sets which will be tested and due diligence will be given.

**Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.**

**How will the data be made accessible (e.g. by deposition in a repository)?**

Data will be made available via the $\_PROJECT platform using a user-friendly front end which allows data visualization. Besides this it will be ensured that data which can be stored in international specialized technical discipline related repositories (Sequencing at the national US center NCBI, SRA; Sequencing and sequence data for the EU center: EBI, ENA; Proteome database: PRIDE), will be stored and processed there as well.

**What methods or software tools are needed to access the data?**

#if$\_PROPRIETARY $PROJECT relies on the tool(s) $\_PROPRIETARY. #endif$\_PROPRIETARY

#if!$\_PROPRIETARY No specialized software will be needed to access the data, usually just a modern browser. Access will be possible via web interfaces. For data processing after obtaining raw data, typical open source software can be used. #if!$\_PROPRIETARY

#if$\_DATAPLANT Dataplant offers tools such as the open source SWATE plugin for Excel, the ARC commander and the DMP tool which while not necessary make the interaction with data more convenient #endif$\_DATAPLANT

**Is documentation about the software needed to access the data included?**

As no software is needed, no documentation needs to be provided. #if$\_DATAPLANT However, DATAPLANT resources are well described and their setup is documented on their github project pages. #if$\_DATAPLANT

**Is it possible to include the relevant software (e.g. in open source code)?**

As stated above, here we use publicly available open source and well documented certified software #if$\_PROPRIETARY except for $\_PROPRIETARY #endif$\_PROPRIETARY

**Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.**

As noted above specialized repositories like SRA /ENA, Pride /Proteomexchange are the most common ones and will be used when adequate. In the case of unstructured less standardized data (e.g. experimental phenotypic measurements) these will be metadata annotated and if complete given a digital object identifier (DOI) #if$\_DATAPLANT and the whole data sets wrapped into an ARC will get DOIs as well #endif$\_DATAPLANT.

**Have you explored appropriate arrangements with the identified repository?**

The submission is for free and it is the goal (at least of ENA) to obtain as much data as possible. Therefore, arrangements are neither necessary nor useful. Catch-all repositories are not required. #if$\_DATAPLANT For DATAPLANT this has been agreed upon #issuewarning please do so #endissuewarning

**If there are restrictions on use, how will access be provided?**

There are no restrictions, beyond the aforementioned IP checks which are in line with e.g. European open data policies.

**Is there a need for a data access committee?**

Consequently, there is no need for a committee.

**Are there well described conditions for access (i.e. a machine-readable license)?**

Yes, where possible e.g. CC REL will be used for data not submitted to specialized repositories such as ENA.

**How will the identity of the person accessing the data be ascertained?**

In case data are only shared within the consortium, if the data is not yet finished or under IP checks, the data is hosted internally and username and password will be required (see also our GDPR rules). In the case data is made public under final EU or US repositories, completely anonymous access is normally allowed this is the case for ENA as well and both is in line with GDPR requirements.

### Making data interoperable

**Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?**

Whenever possible, data will be stored in common and openly defined formats including all necessary metadata to interpret and analyze data in a biological context. By default, no proprietary formats will be used, however Microsoft Excel files (According to ISO/IEC 29500-1:2016) might be used as intermediates by the consortium #if$\_DATAPLANT and by some ARC components in form #endif$\_DATAPLANT. In addition, text files might be edited in text processor files but will be shared as pdf.

**What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?**

As mentioned above, we foresee to use #if$\_RNASEQ|$\_GENOMICS e.g. MinSEQe for sequencing data #endif$\_RNASEQ|$\_GENOMICS and Metabolights compatible forms for metabolites as well as MIAPPE for phenotyping like data. The latter will thus allow integrating data across projects and safeguards reusing established and tested protocols. Additionally, we will use ontology terms to enrich the data sets relying on free and open ontologies. In addition, additional ontology terms might be created and be canonized during $\_PROJECT.

**Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?**

Indeed, open biomedical ontologies will be used where they are mature. As stated in the previous question sometimes ontologies and controlled vocabularies might have to be extended. #if$\_DATAPLANT Here $\_PROJECT will build on the advanced ontologies developed in DATAPLANT #endif$\_DATAPLANT

**In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?**

Common and open ontologies will be used thus this question does not apply.

### Increase data reuse (through clarifying licences)

**How will the data be licensed to permit the widest re-use possible?**

Open licenses will be used whenever possible, such as CC.

**When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.**

In general, IP issues will first be checked. All consortium partners will be encouraged to make data available prior to publication openly and/or under pre-publication agreements #if$\_GENOMICS such as those started in Fort Lauderdale and set forth by the Toronto International Data Release Workshop #endif$\_GENOMICS

**Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.**

There will be no restrictions once data is made public.

**How long is it intended that the data remains re-usable?**

Data will be made available for many years #if$\_DATAPLANT and potentially indefinitely after the end of the project. #endif$\_DATAPLANT

In any case data submitted to specific community accepted public subject area repositories (as detailed above) e.g. ENA /Pride would be subject to local data storage regulation.

**Are data quality assurance processes described?**

Data will be checked and curated. #if$\_DATAPLANT In addition data will analyzed for quality control (QC) problems using automatic procedures as well as by manual curation. #endif$\_DATAPLANT

## 2.3. Allocation of resources

**What are the costs for making data FAIR in your project?**

The costs comprise data curation, #if$\_DATAPLANT Arc consistency checks #endif$\_DATAPLANT and maintenance on $\_PROJECT´s side.

Additionally, last level costs for storage are incurred by last-level repositories (e.g. ENA) but not charged against $\_PROJECT or its members but by the operation budget of these repositories.

**How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).**

A large part of the cost is covered by $\_PROJECT #if$\_DATAPLANT and the structures tools and knowledge layed down in the DATAPLANT consortium #endif$\_DATAPLANT

**Who will be responsible for data management in your project?**

The responsible will be $\_DATAOFFICER

**Are the resources for long term preservation discussed (costs and potential value, who decides and how/what data will be kept and for how long)?**

The $\_PARTNERS decides on preservation of data not submitted to last level subject area repositories #if$\_DATAPLANT or ARCs in DATAPLANT #endif$\_DATAPLANT after project end. This will be in line with EU, institute policies and data sharing based on EU and international standards.

## 2.4. Data security

**What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?**

Once data is transferred to the $\_PROJECT platform #if$\_DATAPLANT and ARCs have been generated in DATAPLANT #endif$\_DATAPLANT, data security will be imposed. This comprises secure storage and the use of passwords and usernames are generally transferred via separate safe media.

**Is the data safely stored in certified repositories for long term preservation and curation?**

Wherever there are certified repositories these will be used as last level repositories. #if$\_RNASEQ Transcriptomics data and gene sequence data will be also made available upon publication via the standards ENA/SRA #endif$\_RNASEQ. #if$\_METABOLOMICS Metabolite data in e.g. Metabolights (and/or Nationwide repositories like the German NFDI the French INRAe) #endif$\_METABOLOMICS #if$\_PROTEOMICS and Proteomics data in e.g. Pride/Proteomexchange #endif$\_PROTEOMICS. In addition, the national resource will maintain safekeeping of data also after the project ends. #if$\_DATAPLANT In addition databases like e.g. Proteomexchange do not support deep plant specific metadata hence ARCs will be maintained to ensure resuababiltiy #endif$\_DATAPLANT

## 2.5. Ethical aspects

**Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).**

At the moment, we do not foresee ethical or legal issues with data sharing. In terms of ethics, as this is *plant data* there is no need for an ethics committee, however diligence for plant resource benefit sharing is considered (🡺see Nagoya protocol). #issuewarning You have to check here and enter any due diligence here at the moment we are awaiting if Nagoya gets also part of sequenc information. In any case if you use material not from your (partner) country and characterize this physically e.g, metabolites, proteome, biochemically RNASeq etc this might represent a Nagoya relevant action unless this is from e.g. US (non partner), Ireland (not signed still contact them) etc but other laws might apply…. #endissuewarning

**Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?**

The only personal data that will potentially be stored is the submitter name and affiliation in the metadata for data. In addition, personal data will be collected for dissemination and communication activities using specific methods and procedures developed by the $\_PROJECT partners to adhere to data protection. #issuewarning you need to inform and better get WRITTEN consent that you store emails and names or even pseudonyms such as twitter handles, we are very sorry about these issues we didn’t invent them #endissuewarning

## 2.6. Other issues

**Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?**

Yes, $\_PROJECT will use common DMP tools #if$\_DATAPLANT and in particular resources developed by the NFDI of Germany #endif$\_DATAPLANT

# Annexes

## Abbreviations

**CC** Creative Commons

**DDBJ** DNA Data Bank of Japan

**DMP** Data Management Plan

**DOI** Digital Object Identifier

**EBI** European Bioinformatics Institute

**ENA** European Nucleotide Archive

**EU** European Union

**GDPR** General data protection regulation (of the EU)

**IP** Intellectual Property

**MIAMET** Minimal Information about Metabolite experiment

**MIAPPE** Minimal Information about Plant Phenotyping Experiment

**MinSEQe** Minimum Information about a high-throughput Sequencing Experiment

**NCBI** National Center for Biotechnology Information

**NFDI** National Research Data Infrastructure (of Germany)

**NGS** Next Generation Sequencing

**RNASeq** RNA Sequencing

**SOP** Standard Operating Procedures

**SRA** Short Read Archive

**ONP** Oxford Nanopore

**qRT** **PCR** quantitative real time polymerase chain reaction

**WP** Work Package