

UNCOVER AND PROMOTE TOLERANCE TO TEMPERATURE AND WATER STRESS IN CAMELINA SATIVA



THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON 2020 RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 862524



D6.3

Data Management Plan

Table of Contents

Document Summary	3
Abstract.....	4
List of Abbreviations	5
1. Introduction.....	6
2. DMP Model.....	9
2.1. Data Summary.....	9
2.2. FAIR data.....	10
Making data findable, including provisions for metadata.....	10
Making data openly accessible	11
Making data interoperable.....	13
Increase data reuse (through clarifying licences).....	13
2.3. Allocation of resources	14
2.4. Data security	14
2.5. Ethical aspects	15
2.6. Other issues.....	15

Document Summary

Deliverable number & title: **D6.3 – Data Management Plan**

Version & submission date: **V2 – 30th September 2022**

Lead Beneficiary: **FZJ**

Related Work Package: **WP6**

Author(s): **Björn Usadel (FZJ)**

Reviewer(s): **Richard Haslam (RRes), Juan Rodrigues (RTDS), Filoklis Pileidis (RTDS), Stéphane Bernillon (INRAE), Maider Goñi (INI), Camino Fábregas (INI), Dominik Grosskinsky (AIT), Claudia Jonak (AIT)**

Datasets underlined: **None**

Communication level:

☒ **PU** *Public*

☐ **CO** *Confidential, only for members of the consortium (including the Commission Services)*

Approved by:

☒ AIT (COO)	☒ CCE	☒ INRAE	☒ RRes
☒ RTDS	☒ UNIBO	☒ FZJ	☒ INI

Grant Agreement Number: **862524**

Programme: **SFS-30-2019 / Agri-Aqua Labs: Looking behind plant adaptation**

Start date of Project: **1st September 2020**

Duration: **5 years**

Project coordinator: **AIT**

Abstract

UNTWIST is part of the Open Data initiative of the EU. To best profit from open data it is necessary to not only store data but to make this findable, accessible, interoperable and reusable (FAIR). Open and FAIR data, however, considers the need to protect individual data sets.

The aim of this document is to provide guidelines on:

- Principles guiding the data management in UNTWIST
- End point repositories for data deposition
- Answers to the EU questionnaire on DMP as a DMP document

The Data Management Plan (DMP) details how data is to be handled during and after the project. The UNTWIST DMP is modelled around the H2020 Online Manual. It will be updated/its validity checked during the UNTWIST project several times. At the very least, this will happen at month 36 (D6.4) and at the project end (D6.6).

Disclaimer: The opinions expressed on this deliverable are those of the author(s) only and should not be considered as representative of the European Commission's official position. This public deliverable is a revised version of D6.3 Data Management Plan (submitted by M6), updated as requested by reviewers in the 1st project periodic assessment.

List of Abbreviations

AIT	AIT Austrian Institute of Technology GmbH
CCE	Camelina Company España S.L.
CC	Creative Commons
DDBJ	DNA Data Bank of Japan
DEC	dissemination exploitation and communication (for UNTWIST)
DMP	Data Management Plan
DOI	Digital Object Identifier
EBI	European Bioinformatics Institute
EMPHASIS	European Multi-Environment Plant Phenotyping And Simulation Infrastructure
ENA	European Nucleotide Archive
EU	European Union
FAIR	Findability, Accessibility, Interoperability, and Reusability
FAME	Fatty Acid Methyl Esters
FZJ	Forschungszentrum Jülich GmbH
GA	General Assembly (of UNTWIST)
GDPR	General data protection regulation (of the EU)
GRA	Grant Agreement
INI	Iniciativas Innovadoras, S.A.L.
INRAe	Institut national de recherche pour l'agriculture, l'alimentation et l'environnement
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
MS	Mass Spectrometry
NCBI	National Center for Biotechnology Information
NFDI	National Research Data Infrastructure (of Germany)
NGS	Next Generation Sequencing
PRIDE	Proteomics Identification Database
RNASeq	RNA Sequencing
SOP	Standard Operating Procedures
SRA	Short Read Archive
ONP	Oxford Nanopore
qRT PCR	quantitative real time polymerase chain reaction
RRes	Rothamsted Research
RTDS	RTDS Association
UNIBO	Alma Mater Studiorum - Università di Bologna
WP	Work Package
XC-MS	any (x) chromatography coupled to mass spectrometry

1. Introduction

UNTWIST follows an open access strategy. In addition to plant data, UNTWIST is dedicated to modelling a large set of multi-omics data, some of which are well defined and standardized in terms of raw output and easily comparable within this omics discipline (sequencing data), others are standardized for output (proteomics data) and the remainder have varying levels of standardisation for output (e.g. metabolite measurements or even plant phenology). Despite this variation, in all cases necessary standards and specialist repositories exist (e.g. ENA for transcript and genomics data). Hence UNTWIST will ultimately deposit large volumes of data in public repositories, besides keeping derived values and models in the UNTWIST Knowledge Hub. Standardized omics repositories have proven their reliability and perfect integration into the FAIR landscape.

2. Data Management

Data Management and sample tracking is very important for UNTWIST. Hence UNTWIST has a separate work package for data management, sample tracking which are leading to data for the knowledge hub. The main two things that are being tracked:

a) **physical** materials i.e.

plants (plant lots) that are phenotyped for (whole) plant traits

seeds which have been centrally produced in INRAE-IJPB and

samples that are generated from plants which are grown. These might be shipped to other locations to be analysed (e.g. leaf samples from RRES to INRAE - Bordeaux for metabolite profiling) or are analysed on the spot (e.g. RRES leaf samples to be subjected to lipidomic profiling). Here **samples** can get a code as determined on the sampling workshop, however in addition **samples** are mostly identified by the primary growth location. In any case these **samples** are linked to the **plants** that they came from.

After physical **samples** have been subjected to analysis (e.g. FAME profiles, carbon isotope discrimination or phenotypes such as plant height) these then result in

b) **measurements** such as

i) genomic information which is independent of the environment

ii) measurements dependent on the environment (e.g. field conditions; greenhouse-imposed treatments such as drought by withholding water) such as

- carbon isotope discrimination
- total redox status
- phenotypic data

iii) “high throughput omics” measurements

- FAME oil content
- detailed lipidomic data
- transcriptomic
- metabolomic
- epigenetic
- proteomic

- enzyme activity datasets
- detailed redox status

iv) measurements and/or descriptions describing the **environment**.

Here, the differentiation between ii and iii is often arbitrary but is reflected in the WP structure of UNTWIST. In any case, genomic data can thus be made a characteristic of a plant line for modelling in task WP4 and to be shared with the community (keeping in mind that heterozygous SNPs will get lost or fixed in individuals).

All measurements which are dependent on the environment (i.e. ii and iii) need to be also clearly linked against the **environment** they were grown in. As **samples** are grown in several **locations**, **measurement** data is linked to a **sample** which is described by specific **environmental** information and can feature a “field map” which describes where plants were grown in a simple map in a greenhouse and/or a field (which is derived from coordinates describing precise **location**).

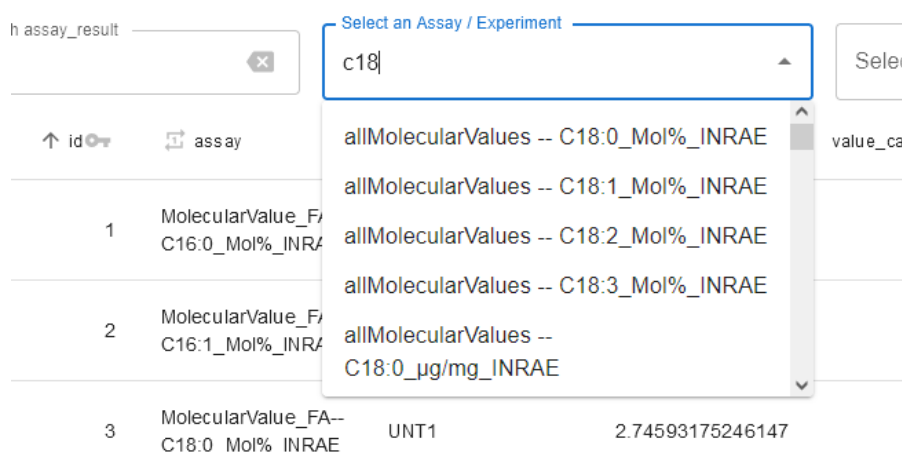
These data are first checked for consistency and completeness by INRAE and FZJ and are afterwards made available in a database. There the values can be looked at, plotted and downloaded at convenience.

Using the database ensures that once data is checked it is available and will not be lost.

Nevertheless, underlying data tables are also kept on a sftp server and are backed up regularly. Raw data (such as sequencing run fastq data) that is to be uploaded into other repositories is also kept there where necessary.

Bringing these data together in a database allows to link the different sets. For example, the sample metadata describing the sample can be derived by determining the plant (line) and the environment the plant was grown in. By anchoring samples to the primary producer of the plant the samples were derived from it is also possible to check and recheck environments.

Furthermore, the sample data base allows to search for plant and samples (See Figure 1) and allows plotting and comparing measurements within the database.



The screenshot shows a web interface for searching the UNTWIST database. At the top, there is a search bar with the text 'h assay_result' and a dropdown menu. The dropdown menu is open, showing a list of suggestions for the search term 'c18'. The suggestions are:

- allMolecularValues -- C18:0_Mol%_INRAE
- allMolecularValues -- C18:1_Mol%_INRAE
- allMolecularValues -- C18:2_Mol%_INRAE
- allMolecularValues -- C18:3_Mol%_INRAE
- allMolecularValues -- C18:0_μg/mg_INRAE

Below the search bar, there is a table with columns 'id', 'assay', and 'value'. The table contains three rows of data:

id	assay	value
1	MolecularValue_F... C16:0_Mol%_INRAE	
2	MolecularValue_F... C16:1_Mol%_INRAE	
3	MolecularValue_FA... C18:0_Mol%_INRAE	UNT1 2.74593175246147

Figure 1 search and autocomplete searching as of 8/2022 for C18 lipid measurements offers several data sets to select (in mol % as well as in μg /mg)

D6.3

Data Management Plan



Figure 2 plotting a measurement set and showing the values for the plot in the database. The example shows C18:2 lipid measurements

The location data allows to check for spatial effects but also allows to check easily if associations were correct and thus allows an additional way to check for data consistency.

The database is thus conceptually associating different plants, samples and environments as conceptualised in Figure 3.

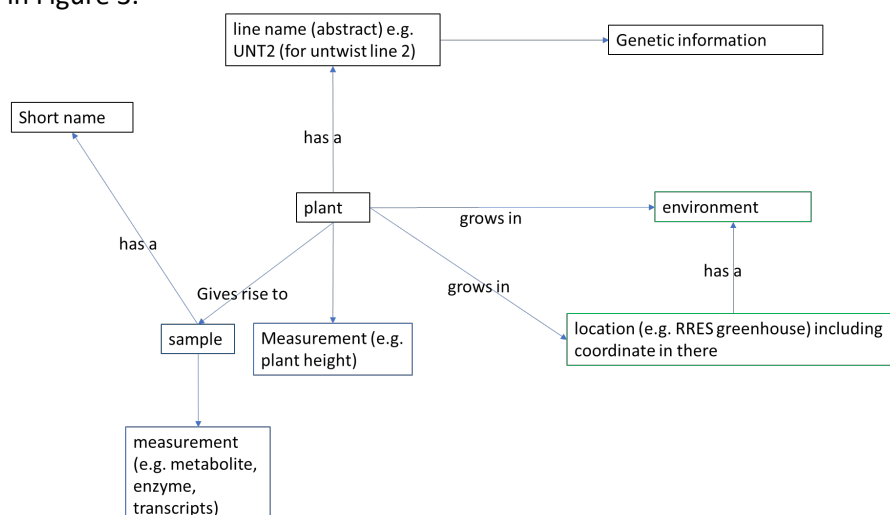


Figure 3 conceptual model of the UNTWIST database

The actual flow of samples is determined by plant producers (lysimeter, glasshouse, controlled chambers) and by the partners analysing the samples depending on whether samples are measured on site or shipped for measurements. The **location** of a sample is an important characteristic allowing to trace its provenience. Hence



the location is kept not only in the short sample name, but also usually in freezer boxes etc. In addition, often sample tables are accompanying samples in physical form (i.e. a print out) as an additional measure to make sure that samples are not confused and that data in the database can thus be rectified in an updated version. **This is particularly important for larger batches of shipments where only a simplified naming scheme is used on the actual sample vial to allow fast enough sampling in order to account for diurnal and other processes requiring fast sampling and based on the experience with “freezing resistant” stickers (please check revised version of 6.1). Hence, when entering data into the database the full sample code is being restored where applicable.**

Once all data is complete and correct it can then also be submitted as data sets to public domain specific databases such as the ones from the EBI.

3. DMP Model

The sample tracking platform now relies on a first working version of the UNTWIST. It is password protected, so before any data can be obtained or samples generated an authentication needs to take place. (please check MS7)

3.1. Data Summary

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

For UNTWIST, data collection and integration is absolutely necessary as data is not only used to understand principles, but it is also used for developing models in WP6, which need to be informed about data provenance. It is therefore of importance that not only data is well generated, but also well annotated using *open standards* and metadata as it is laid out in the following section. As UNTWIST aims at uncovering new principles across diverse camelina lines, culture conditions (countries, field vs contained) and as it crosses multiple omics disciplines experiments, good documentation, data keeping and integration are necessary.

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: Firstly, phenotypic data about camelina plants, secondly experimental descriptions and where necessary logging of the data, thirdly omics data including genomics, epigenomics, metabolomics, transcriptomics, lipidomic, enzyme activity profiles, and proteomics data sets stemming from different approaches such as next generation sequencing (NGS) and/or quantitative Real Time PCR-based approaches. 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. Therefore, specific care needs to be taken to document and archive these resources (including the analytic pipelines) as well.

Will you re-use any existing data and how?

The project builds on existing data sets and relies on them. For instance, without a proper genomic reference it is very difficult to analyze NGS data sets, even though better genomes will be produced in the course of the project. It is also important to include existing data sets on the expression and metabolic behavior of camelina and the background knowledge of the partners. Genomic references can simply be gathered from reference databases for genomes/sequences like the National Center for Biotechnology Information: NCBI¹ (US);

¹ <https://www.ncbi.nlm.nih.gov/>

European Bioinformatics Institute: EBI² (EU); DNA Data Bank of Japan: DDBJ³ (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 UNTWIST, specific data sets will be generated by the consortium partners. Short read sequencing will be outsourced and raw data will be received. For long read Oxford nanopore data (ONP) data, raw traces will be recorded as well; for enzymatic measurements these will be tracked according to best practices. Metabolomic/lipidomic data will be generated using chromatography coupled to mass spectrometry and from enzyme platforms mostly. Finally, proteomic data will be generated using an EU platform which is in line with community standards.

What is the expected size of the data?

We expect to generate raw data in the range of 10-100 TB of data. The size of derived data will be below 1 TB.

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

The data will be useful for the UNTWIST partners, to oilseed crop breeders and producers, to the scientific community working on oilseed crops, as well as to scientists trying to understand abiotic stress tolerance and adaptation. Furthermore, abstracted and summarized data will be interesting to “prosumers” and the general public interested in camelina breeding and composition. Hence, UNTWIST also strives to collect the data that has been disseminated and potentially advertise it e.g. through the hub or other means, if it is not included in a publication anyway which is the most likely form of dissemination.

3.2. 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 persistent unique identifiers and they will be annotated with Metadata. UNTWIST will rely on community standards for omics data plus additional recommendations necessary in the plant field adapted by e.g. Minimum Information About a Plant Phenotyping Experiment (MIAPPE⁴). These -unlike across domain minimal sets such as Dublin core⁵ 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 UNTWIST.

² <https://www.ebi.ac.uk/>

³ <https://www.ddbj.nig.ac.jp/index-e.html>

⁴ <https://www.miappe.org/>

⁵ <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

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. In the case of datasets, the dataset names will also encode the provenance.

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.

Do you provide clear version numbers?

To maintain data integrity and to be able to re-analyze data, data sets will get version numbers. However, some raw data will be “frozen” by making it immutable in the UNTWIST storage.

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 e.g. Minimum Information About a Next-generation Sequencing Experiment MinSEQe⁶ for sequencing data and INRAe-data and/or Metabolights⁷ submission compliant standards for metabolites as well as MIAPPE for phenotyping-like data, but will also be relying on specific SOPs for additional annotations.

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 UNTWIST 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 GA and DEC). This applies in particular to data pertaining to Camelina Company. 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.

Does not apply.

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

Data will be made available via the UNTWIST Knowledge Hub using a user-friendly front end and will create back-end storages versus the national research data infrastructure (NFDI DATAPlant⁸) of Germany where necessary and useful. It will be ensured that data which can be stored in international specialized technical discipline

⁶ <http://fged.org/projects/minseque/>

⁷ <https://www.ebi.ac.uk/metabolights>

⁸ <https://nfdi4plants.de/>

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?

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

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

As no software is needed, no documentation needs to be provided.

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

As stated above, software is only needed AFTER data has been obtained by a user in order to process and/or analyze the data. Here we use publicly available open source and well documented certified software.

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 SRA⁹ /ENA¹⁰, Pride¹¹ /Proteomexchange¹² 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).

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

There are no restrictions, beyond the aforementioned IP checks which are in line with EU 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 and ENA seems to still be bound to EU GDPR rules. However, this is beyond UNTWIST's influence.

⁹ <https://www.ncbi.nlm.nih.gov/sra>

¹⁰ <https://www.ebi.ac.uk/ena>

¹¹ <https://www.ebi.ac.uk/pride/>

¹² <http://www.proteomexchange.org/>

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) are used as intermediates by the consortium. 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 e.g. MinSEQe for sequencing data 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 UNTWIST.

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.

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, 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 using the DEC plan. All consortium partners will be encouraged to make data available prior to publication under pre-publication agreements such as those started in Fort Lauderdale and set forth by the Toronto International Data Release Workshop.

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 and potentially indefinitely after the end of the project via the Knowledge Hub and public databases. If necessary, data will be deposited in eIDAL, or the NFDI for perpetual public accessibility.

In any case data submitted to specific public subject area repositories (as detailed above) e.g. ENA /Pride would be subject to local data storage regulation. As of today data is stored “indefinitely”.

Are data quality assurance processes described?

Data will be analyzed for quality control (QC) problems using automatic procedures as well as by manual curation. This comprises automated checks for ranges as well as typical discipline specific checks.

3.3. Allocation of resources

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

The costs comprise data curation, setup of databases and long-term maintenance and storage including electricity on UNTWIST’s side. Additionally, last level costs for storage are incurred by the repositories but not charged against UNTWIST or its members. The costs are covered by the work described in WP6.

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 UNTWIST. However, some hardware costs etc. are based on FZJ contributions.

Who will be responsible for data management in your project?

Forschungszentrum Jülich (FZJ)

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 partner FZJ decides on preservation of data not submitted to last level subject area repositories. This will be in line with German law, institute policies and data sharing in the German national standards. This usually results in data stored and accessible for decades after the project ends.

3.4. 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 UNTWIST Knowledge Hub, data security will be imposed. This comprises secure storage and the use of encrypted connections to the data (https and sftp) where passwords and usernames are

generally transferred via separate safe media. Also, data blocks are tagged with access rights and thus require user authentication.

In terms of data recovery, data will be stored on a redundant based file system employing RAID like redundancy. In addition, regular backups and/or off-site data mirroring will be used.

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

Transcriptomics data will be also made available upon publication via the standards ENA/SRA. Metabolite data in Metabolights (and/or Nationwide repositories like INRAe) and Proteomics data in Pride/Proteomexchange. In addition, the national resource will maintain safekeeping of data also after the project ends.

3.5. 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 treated in D9.3. The consortium makes its best efforts to ensure data sharing, at the very latest after exploitation through papers and patents.

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 UNTWIST partners to adhere to data protection.

3.6. Other issues

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

Yes, UNTWIST will use tools and potentially resources developed by the NFDI Germany.



Horizon 2020
European Union Funding
for Research & Innovation

UNCOVER AND PROMOTE TOLERANCE TO TEMPERATURE AND WATER STRESS IN CAMELINA SATIVA



UNTWIST PARTNERS