UNTWIST

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

D6.4

Updated/confirmed data management plan

Table of Contents

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

Document Summary

Deliverable number & title: **D6.4 – Updated /confirmed data management plan**

Version & submission date: **V1- 31st August 2023**

Lead Beneficiary: **FZJ**

Related Work Package: **WP6**

Author(s): **Björn Usadel (FZJ)**

Reviewer(s): **Claudia Jonak (AIT), Filoklis Pileidis (RTDS), Maider Goñi and Camino Fábregas (INI), Jean-Denis Faure (INRAe)**

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**

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

Abstract

UNTWIST is part of the Research Data Pilot ORDP initiative of the EU, to make research data generated by Horizon 2020 projects openly available. 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 updated 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 has been checked during the UNTWIST project several times.

This document represents an update to the existing data management plan that has been revised in 2022. In addition, a sustainability plan following from the data management plan and first exploitation plans for the data encompassing the HUB are laid out.

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 describes D6.4 Updated/confirmed Data Management Plan and adds a first idea about longer term perspectives and exploitation plans from the view of the authors.

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

List of Abbreviations

AIT	AIT Austrian Institute of Technology GmbH
BnIR	*Brassica napus* multi-omics information resource
BnaGVD	*Brassica napus* Genomic Variation Database
BRAD	Brassicaceae Database
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	Findable, Accessible, Interoperable, and Reusable
FAME	Fatty Acid Methyl Esters
FZJ	Forschungszentrum Jülich GmbH
GA	General Assembly (of UNTWIST)
GDPR	General data protection regulation (of the EU)
GRA	Grant Agreement1
GWAS	genome wide association study
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
ONP	Oxford Nanopore
PRIDE	Proteomics Identification Database
qRT PCR	quantitative real time polymerase chain reaction
RNASeq	RNA Sequencing
RRes	Rothamsted Research
RTDS	RTDS Association
SFTP	Secure FTP
SNP	Single nucleotide polymorphism
SOP	Standard Operating Procedures
SRA	Short Read Archive
TAIR	The Arabidopsis Information Resource (a database)
UNIBO	Alma Mater Studiorum - Università di Bologna
WP	Work Package
XC-MS	any (x) chromatography coupled to mass spectrometry

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

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 is ultimately depositing 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.

b) 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 ***measurements*** such as:

1. genomic information which is independent of the environment
2. 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
3. high throughput omics" measurements
   - FAME oil content
   - detailed lipidomic data
   - transcriptomic
   - metabolomic
   - epigenetic
   - proteomic
   - enzyme activity datasets
   - detailed redox status
4. measurements and/or descriptions describing the ***environment***.

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

Here, the differentiation between the aforementioned categories 2 and 3 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., 2 and 3) 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 where 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. 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 database allows to search for plant and samples (*See Figure 1*) and allows plotting and comparing measurements within the database.

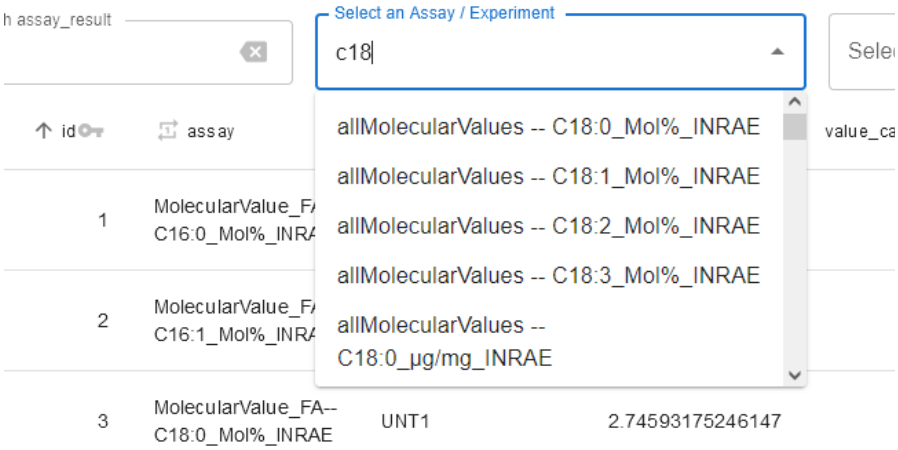

Figure 1 search and autocomplete searching as of 8/2022 for C18 lipid measurements offers several data sets to select (in mol % as well as in µg /mg)

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

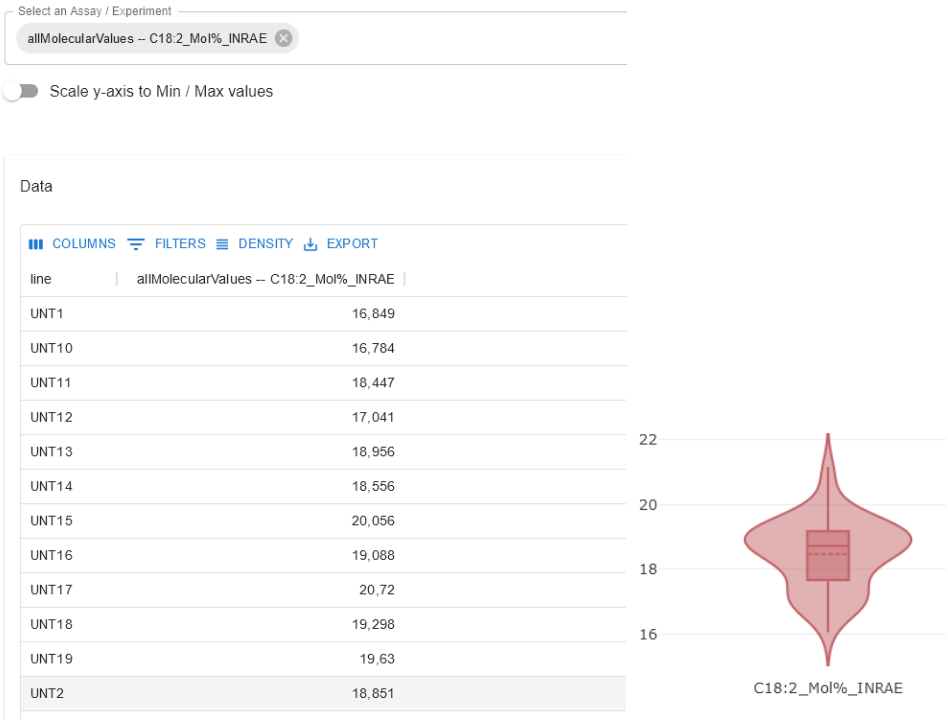

Figure 2 plotting a measurement set and showing the values for the plot in the database. The example shows C18:2 lipid measurements.

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 conceptionally associating different plants, samples and environments as conceptualised in Figure 3.

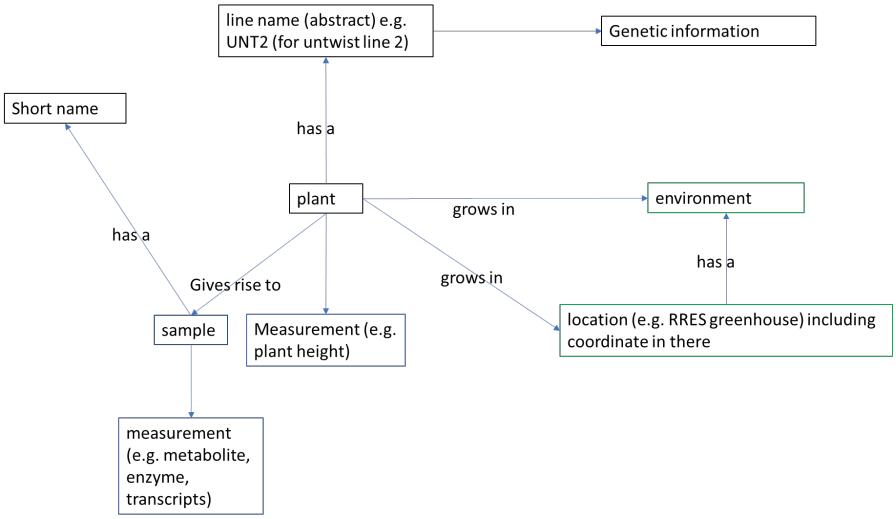

Figure 3 conceptional model of the UNTWIST database

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

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

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 are 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 (WP2), secondly experimental descriptions and where necessary logging of the data, thirdly omics data including genomics, epigenomics (WP1), metabolomics, transcriptomics (WP3), lipidomic (WP3), enzyme activity profiles, and proteomics data sets (WP3) 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 analyse NGS data sets, even though better genomes will be produced during 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: NCBI[1] (US); European

[1] https://www.ncbi.nlm.nih.gov/

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

Bioinformatics Institute: EBI[2] (EU); DNA Data Bank of Japan: DDBJ[3] (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 (MIAPPE[4]). These unlike across-domain minimal sets such as Dublin core[5] which mostly defines the 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/

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

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-analyse data, datasets 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 use e.g., Minimum Information About a Next-generation Sequencing Experiment MinSEQe[6] for sequencing data and INRAe-data and/or Metabolights[7] 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 Consortium Agreement). This applies in particular to data pertaining to Camelina Company. However, all partners also strive for IP protection of datasets 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 DATAPlant[8]) of Germany safeguarding integrative data reuse across data domains. It will be ensured that data that can be stored in international specialized technical discipline related repositories (Sequencing at the national US center NCBI, SRA;

[6] https://www.fged.org/projects/minseqe/
[7] https://www.ebi.ac.uk/metabolights
[8] https://nfdi4plants.de/

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

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 (such as trimmomatic, edgeR, R etc) 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 analyse 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 SRA[9] /ENA[10] , Pride[11] /Proteomexchange[12] 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 a username and password will be required (see also our GDPR rules[13]). In the case data

[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/
[13] https://www.untwist.eu/data-protection-policy/

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

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.

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 analyse 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.

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

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. To improve usability the NFDI DataPLANT will be used for long term public accessibility of multi-modal datasets.

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 analysed 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 storage and accessibility 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

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

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's 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 based on the model laid down down in D9.4 for the UNTWIST consortium.

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 resources developed by the NFDI Germany in particular the encapsulated structure.

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

4. Sustainable Data Availability Use and Exploitation

4.1 Sustainable Availability of Raw Data and Artifacts

An often-overlooked aspect with regard to project funding is sustainable availability and reuse of data in its context as meant by the original experimental question. This is because data integration is becoming more and more important and in the case of UNTWIST **an integrative multi omics data view is essential**. Multiple long term sustainable FAIR initiatives have and will safeguard sharing of data, relying on omics-discipline specific databases such as EBI-ENA[14] for e.g. quantitative RNAseq sequencing data for gene expression and e.g. epigenetic analysis of methylation data whereas metabolic data could reside in the EBI: MetaboLights database[15] and finally Proteomics data in e.g. Pride[16]. (See also the model section of the data management plan below).

However, these data on different omics disciplines can stem from the exact *same experimental unit (e.g., a plant) set* sharing the same growth condition. This is the case for UNTWIST that tries to get a holistic view on data making pairing of different data (e.g., RNASeq and epigenetics) of the same sample important. Therefore, UNTWIST has generated a dedicated database allowing (see above Figure 2) to query integrate data across the different disciplines and as the database uses derived and pre-processed data, it does not require the user to redo the steps taken by the UNTWIST consortium (e.g. quality control, data processing, pruning, normalisation etc). The latter could as well be addressed by uploading multi-omics derived datasets into e.g., Data Dryad[17] or Zenodo[18]. However, the data would be frozen and could not be integrated with newer datasets and would not be interactive.

By adding necessary metadata, auxiliary data (e.g. third-party data necessary to analyse a given data set such as the published camelina genome for a comparison to said genome) and workflows about data processing the whole dataset comprising experimental units will thus also be encapsulated into the NFDI DataPLANT ARC structure[19] developed by the German National Consortium for Research Data Management (with FZJ as co-speaker). In brief, this structure consists of: 1) a super set of metadata needed for annotating data sets for each of the omics data bases (Figure 4 for an example). 2) It comprises the raw data (e.g. expression, metabolite accumulation, enzyme activity etc) that would be submitted to e.g. ENA and 3) it also allows to describe computational workflows that were used to analyse raw data and potentially integrate this data with metabolic accumulation.

[14] https://www.ebi.ac.uk/ena/browser/home
[15] https://www.ebi.ac.uk/metabolights/
[16] https://www.ebi.ac.uk/pride/
[17] https://datadryad.org
[18] https://zenodo.org/
[19] https://github.com/nfdi4plants/ARC-specification

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

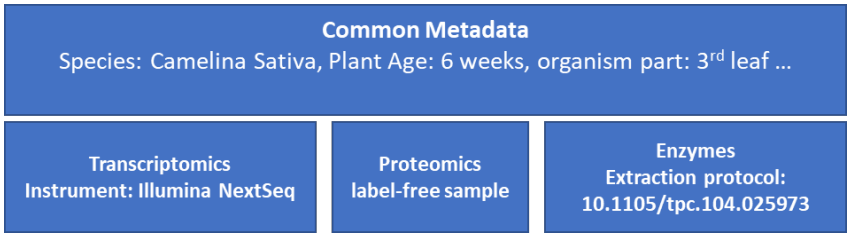

Figure 4 exemplary common and specific data set for typical UNTWIST datasets. Most biological metadata is common to all omics analysis.

Using this structure will thus 1) safeguard data integrity and will 2) allow to reproduce UNTWIST data more quickly foregoing the step of gathering data from different databases and 3) solving the problem of experiment matching as no unique experiment identifier is used across different databases in all cases (e.g. is sample 1 in metabolomics sample 1 in the transcriptomics experiment or does the same metadata mean the same experiment was performed twice or samples were used for two different omics analysis?). Furthermore, it will allow data integration through the NFDI DataPLANT after the UNTWIST end. Altogether these measures taken allow maximal reuse of UNTWIST data and thus facilitate scientific exploitation and citations.

4.2 Exploitation of data and Knowledge HUB

UNTWIST goes even one step further having realized that a) experiments are done with an overarching question in mind and might be linked to each other, b) data integration is key to best data analysis and exploitation and c) that a large user base wants to access data in a convenient and directly useful form potentially integrating with their own data. These considerations were matched to our preliminary post-project viability plan of the HUB.

Table 1 Use cases and their alignment with the preliminary post-project plan.

Use case	Expected Outcomes for Stakeholders
Using UNTWIST genomes to check for particular variants including structural genome variants and/or checking genes in regions that have been shown to be responsible for a particular trait	***Science*** for "genomic and post-genomics data"; ***Industry: Breeders*** for "access to more diverse germplasm" and "available natural variation for breeding"
Using UNTWIST and public (sequencing) data to identify variant data for genetic marker development or genome wide association studies (GWAS)	***Science*** to learn about "adaptive strategies of the abiotic stress"; ***Industry: Breeders*** to have "ideotype descriptions" and "access to novel recombinant camelina lines"
Identifying UNTWIST lines showing e.g. high drought tolerance using a particular measurement and identifying similar lines based on genetics.	***Science*** "adaptive strategies of the abiotic stress"; ***Industry: Breeders*** to learn about "mechanisms of plant performance", ***Farmers*** to optimally choose "yields of oilseed crops […] under drought-like conditions"; and ***Oil manufacturers*** to identify "how stable yields (quantity and composition) can be delivered"

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

Visualizing and comparing UNTWIST data against public data sets and building own hypotheses	***Science*** for analysing "adaptive strategies of the abiotic stress tolerant plant camelina"; ***General Public*** for understanding and "development of climate-smart sustainable agriculture"
Allowing to integrate different data from different species	***Science*** "strategies of the abiotic stress tolerant plant camelina"; ***General Public*** for "evidence-based solutions and field data to make educated decisions on development of climate-smart sustainable agriculture"
Selecting interesting Camelina lines based on phenotypic data	***Science*** for "genomic and post-genomics data sets to enable meaningful research progress" Industry: ***Breeders*** for a "to kick-start rapid expansion of climate-smart breeding of camelina;" and ***Farmers for "***details […] required to establish new crops" and ***Oil manufacturers*** to gain access to "oilseed crops to be adapted to more diversified markets"

The UNTWIST knowledge HUB is perfectly set up to allow answering these and multiple questions, which occur multiple times in the plant science domain. The UNTWIST HUB can be used to calculate and visualize data on demand in the users' browser. As an example, a genome wide (GWAS) analysis as well as correlating data to traits can be performed by simply selecting a stored trait (Figure 5). This allows to also associate future traits measured in the UNTWIST panel or subpanels to be analysed with the same tools developed for UNTWIST. As UNTWIST has also analysed and integrated public data on Camelina genomics, these data will also be made available through the database. This will allow to analyse data across more than 200 Camelina accessions and will provide a very useful tool for the Camelina research community similar to SNP Chip based tools that were developed for the model plant *Arabidopsis thaliana*. For a short-term exploitation, it is planned to fill the database with phenotypic values of public accessions to have immediate use cases there. (Currently, this is hampered by data privacy for the non-UNTWIST data). Furthermore, our new collection of reference genomes will be made available via the database as well relying on standard genome-based analysis toolkits already in place at FZJ. This will provide the consortium a unique scientific platform and settle its standing in the Camelina community. Further activities about exploitation and sustainability of the Knowledge Hub will be performed and reported in a later stage of the project (D6.6 and D7.6).

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

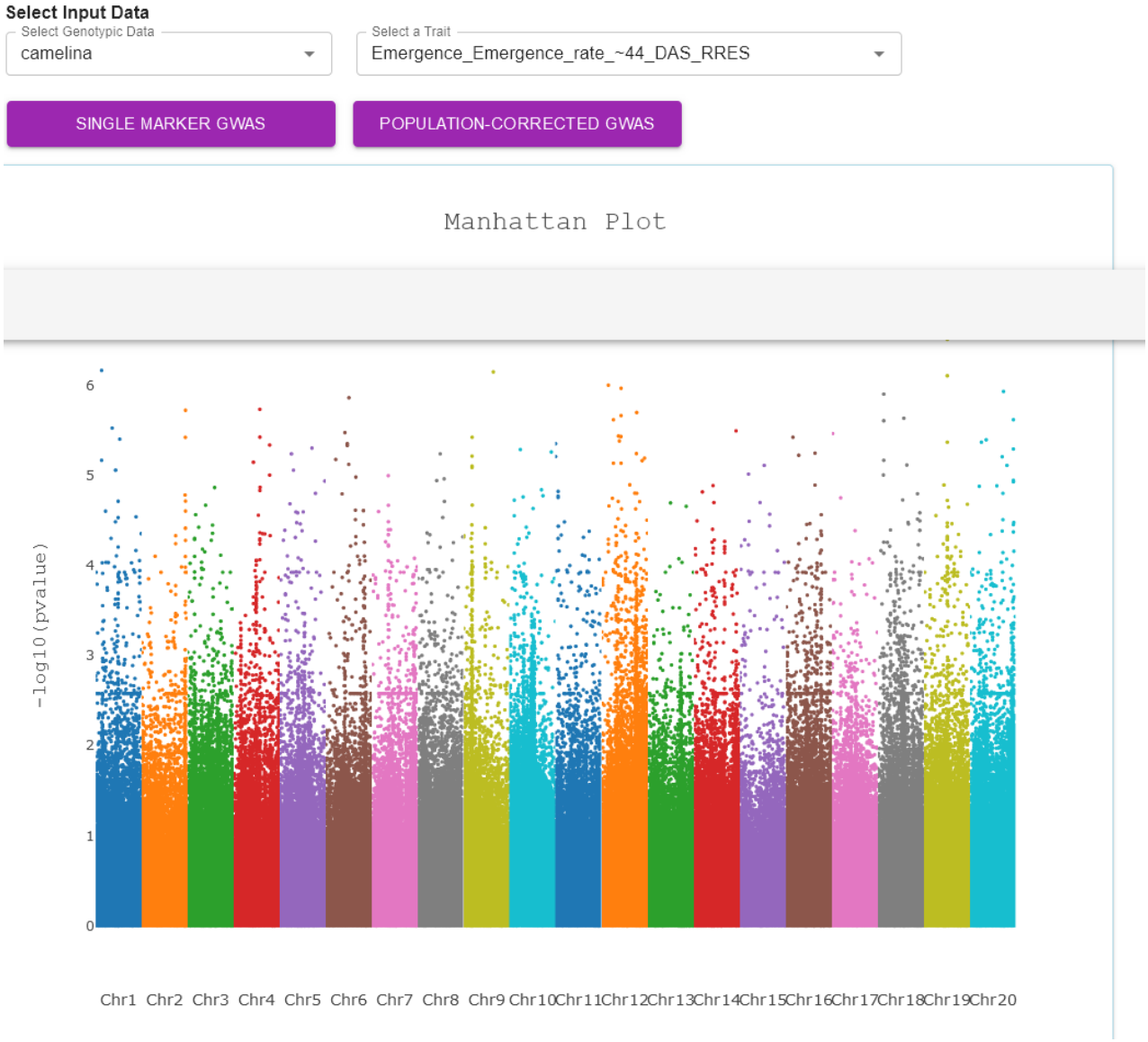

Figure 5 GWAS analysis performed upon user request within the UNTWIST database

To further exploit the data, FZJ is harvesting data from other non-model *Brassicaceae,* which will also be made available via the UNTWIST knowledge HUB and strengthen the position of FZJ. As the institute IBG-4 at FZJ, which is conducting the research, has the governmental mission to provide sustainable bioinformatics resources, the UNTWIST database will represent a valuable addition which we expect will be used and cited well, thus disseminating UNTWIST and its data. The UNTWIST database will be linked into European resources and infrastructure such as ELIXIR[20] and could be coupled to EOSC[21] resources as well in e.g. implementation studies.

At the same time, the underlying technology can be leveraged when joined with other data to a) generate novel knowledge for the company partner "Camelina Company" b) develop novel interactive breeding systems, especially after user feedback allowing to mature the database. This would finally give FZJ the opportunity to

[20] https://elixir-europe.org
[21] https://eosc-portal.eu/

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

spin off additional or improved service versions of the database (as the content per se is open access) and/or it would allow FZJ to develop new applications based on this core database idea which could be commercialized and will be further developed in D7.6.

To analyse the usefulness of the data HUB for exploitation, UNTWIST has performed an analysis of the area:

Key target groups for the database within the EU agricultural ecosystem would include

- Research institutions and universities
- Camelina and breeders of related crops (small-scale, medium)
- Agricultural associations and cooperatives
- Policy makers and regulators
- Camelina-specific communities and networks

Overlapping and/or competing database

- TAIR[22] the Arabidopsis information resource. model Arabidopsis only mostly annotation, genetics and genomics
- BnIR[23] a free multi omics database for rapeseed similar in design to the HUB
- BnaGVD[24]: A Genomic Variation Database of Rapeseed; rapeseed only and only genomic variation database
- BRAD[25] Brassica data mostly genomic focus

Exploitation and Dissemination of related databases

Usually, dissemination and exploitation of related databases are via scientific publications only. TAIR is a not-for-profit database that can levy some charges. Hence the UNTWIST knowledge HUB fits into a unique niche potentially fulfilling a similar role as BnIR for Camelina which was just recently published in the very high-profile journal BnIR. As the UNTWIST knowledge HUB strives to also integrate other related species datasets (similar to BRAD) but taking a holistic omics view and not a genetic view alone.

[22] https://www.arabidopsis.org
[23] https://yanglab.hzau.edu.cn/BnIR
[24] http://rapeseed.biocloud.net/home
[25] http://brassicadb.cn

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

Horizon 2020
European Union Funding
for Research & Innovation

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

UNTWIST PARTNERS

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

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