cousin

Crop Cousins, promise for the future

Deliverable 3.1 SOPs for data acquisition and storage. Establishment of procedures for pheno-, geno- and chemotyping and metadata

Acronym:	COUSIN - Crop Wild Relatives utilisation and conservation for sustainable agriculture
Project Number:	101135314
Call Topic:	HORIZON-CL6-2023-BIODIV-01-13 (RIA)
Start date:	1st January 2024
Duration:	60 months

Date of release	**30/09/2024**
Authors (Name, Organisation)	**Rick van de Zedde (WR),** **Marta Vasconcelos (UCP),**
Contributors (WPs involved)	**Christian Schöb, Luke Ramsay, Stephan Weise, José Felipe Ortega Soto**
Dissemination level (Sen/Public)	**Public**
Type (Report/Data/Dec/etc)	**Report**
Status (Version number)	**1.0**

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
0.5	19.09.2024	▪ Initial draft version
0.7	26.09.2024	▪ Second draft version, comments from co-authors
1.0	30.09.2024	▪ First final version

Contents

1 Introduction

The COUSIN project will generate insightful results and data, and project partners are expected to collect, compile and manage the data in a structured way according to COUSIN data protocols and standardised data collection templates. This document will focus on the procedures how to create the protocols and templates. To make plant genetic resources usable, extensive associated data is needed. These include descriptive data on plant material/ passport data, phenotypic observation data, for abiotic as well as for biotic stress, farm management details, and data on nutritional and health-related traits, genotyping data, and data on breeding material and cultivars. Additional metadata, e.g. definitions of traits, descriptions of observation methods, experimental set-ups, etc., will be essential for the reproducibility and interpretation of data. Therefore, Standard Operating procedures will be introduced.

1.1 What are SOPs for data acquisition and data storage?

A standard operating procedure (SOP) from the perspective of COUSIN is a set of step-by-step instructions defined by an expert to help fellow-researchers to consistently carry out routine (lab) operations (Wikipedia, 2024). SOPs aim to achieve efficiency when carrying out measurements, increase the quality output, and result in higher uniformity, while reducing human errors, inconsistencies, instruction miscommunication and other errors that have a negative impact on the output of these procedures. Standard Operating Procedures ensure that the activities for collecting, processing, and storing data are uniform, traceable, and replicable in line with the FAIR data requirements. Making data Findable, Accessible, Interoperable, and Reusable (FAIR) is essential for maximizing its impact. Metadata plays a key role in supporting the FAIR principles, enabling automated data discovery, retrieval, and integration across different platforms. This document outlines the key elements in defining and exchanging SOPs, including the capture of relevant metadata, including targeted examples and form the start of collecting and sharing these SOPs for applied end-use within the EU-project COUSIN and beyond.

1.2 What is the benefit of using SOPs

The establishment and exchange of Standard Operating Procedures (SOPs) for data acquisition and data storage is an especially critical component when carrying out multi-site, multi-year experiments in which multiple researchers are involved. Good quality data is essential, as partners within COUSIN will carry out numerous experiments in their laboratories, greenhouses, climate rooms and fields on different locations where performance and characteristics of the five flagship crops will be collected. Highly diverse datasets will be collected for phenotypic characteristics, genomic data different crops, and chemical analysis of samples. These activities involve generating large volumes of data that are complex to interpret and analyse. For consistent and reliable results across different experiments, sites and years well-defined SOPs are essential and form a practical solution to enable publications and reports to datasets that match FAIR data guidelines as set by Horizon Europe (EC Europe, 2024).

1.3 Metadata SOPs and operational SOPs

In the context of collecting phenotypic, genomic, and chemotypic data, capturing comprehensive metadata is as important as capturing the primary data. Therefor within this document we differentiate between metadata SOPs describing a procedure to collect contextual and general information needed for data interpretation, reproducibility, and integration across studies, and operational SOPs describing the step-by-step approach to carry out measurements and activities. The COUSIN milestone ‘SOPs for data acquisition available’ due in M9 is linked to this deliverable. It resulted in an easy-to-use SOP for data acquisition for the consortium that has been made available to all partners. This has been arranged by providing access to a clear Excel sheet which must be filled in by partners when collecting data within a dedicated experiment. These Excel-files are to be collected within the COUSIN data platform.

1.4 Traits and challenges for COUSIN

Priority traits as defined in the COUSIN EU-project plan are pest and disease resistance, tolerance to extreme weather for conditions like drought & heat, plus nutrient use efficiency (N/P). More specifically for the

individual flagship crops traits like resistance to pea root rot complex, to aphid, powdery mildew, broomrape, Ascochyta blight, and drought tolerance and for the introgression of rust resistances (e.g. stripe/ yellow and stem/black rust), weed suppression ability and other cultivar specific diseases in pot, on-station and on-farm trials. CWRs of lettuce will be evaluated for drought tolerance and nutrient (N/P) use efficiency. For barley the focus is on traits like root biomass, biotic and abiotic stress tolerance including drought/waterlogging and heat tolerance, nutrient deficiency tolerance, responses to varying pH and adaptation to marginal soils. For brassica the most relevant trait are drought stress evaluation. Chemotyping will be carried out on genotypes to analyze glucosinolates (GLSs) and polyphenols (PPs) content and profile. Brassica oleracea complex species will also be evaluated for the resistance to *Xanthomonas campestris* pv *campestris*. A postharvest phenotyping kit will be created and shared with partners to enable them apply trait extraction methods themselves.

The partners and researchers involved will need to capture the data in a standardised format as described in deliverable 'D.5.1 - Data Management Plan' to enable scoring the performance of the different genotypes of the selected crops, and their crop wild relatives. Next to that, the metadata is essential as set in the COUSIN milestone 'SOPs for data acquisition available'. Furthermore, operational SOPs are shared to describe how these measurements were carried out.

2 Methodology

2.1 Defining Standard Operating Procedures

SOPs are detailed, written instructions designed to achieve uniformity in the performance of a specific measurement. As phenotyping, genomics, and chemotyping, cover a broad spectrum of activities, this will also lead to a diverse set of SOPs, below we list the minimal requirements that should be present in such a document, ranging from sample collection and data acquisition to storage and dissemination. The focus is on minimal useful descriptors to ensure a way of working, and focus on differences and risks found in different protocols provided by COUSIN partners.

2.1.1 Structure of SOPs

A well-structured SOP typically contains the following sections:

- **Title Page:** Includes the title of the SOP, SOP number, version control, and date of issue.
- **Purpose:** Describes the objective and scope of the SOP.
- **Definitions and Abbreviations:** Clarifies terms and acronyms used in the document.
- **Responsibility:** Defines the roles and responsibilities of personnel involved in the procedures.
- **Materials and Equipment:** Lists the materials, reagents, and equipment required for the procedure.
- **Procedure:** Provides step-by-step instructions for each part of the process, including data acquisition, quality control, and storage methods.
- **Health and Safety Considerations:** Highlights any potential hazards and safety precautions.
- **References:** Lists any related documents, scientific publications, or standards used.
- **Appendices:** Includes templates, data sheets, and diagrams that support the SOP.

2.2 Essential elements in SOPs for Phenotyping, Genomics, and Chemotyping

The goal of COUSIN is to provide tools that generate a wide range of datasets which meet specific, local needs and constraints of the characterisation and breeding programmes. Regarding tolerances to extreme weather conditions – heat, drought and waterlogging – and nutrient use efficiency (N/P), wild relatives of flagship crops will be **phenotyped** in dedicated field, greenhouse and growth chamber experiments by breeders, SMEs and research organizations, using red-green-blue (RGB) and (Near) InfraRed (NIR/IR) reflectance and chlorophyll fluorescence, to determine plant stress status, photosynthesis parameters and plant growth over time. Mature plants will be assessed for plant architecture (3D imaging) and stress tolerance (RGB, IR reflectance) development. Based on the **genotyping** of the populations and the allele frequency patterns for certain traits, an improved selection process for all CWR-introgressed populations and breeding lines of each flagship crop will be achieved. These populations will be grown and phenotyped in multiple consecutive years in different

environments, and then pool sequenced to uncover any adaptations. RNA markers for bioactive compounds and drought and resistance will be implemented, next to implementation of microbiome-assisted screening assays, genetic markers and models addressing the specificities of breeding programmes and related SOPs. As SOPs can be very specific and linked to dedicated equipment and or local situations, we provide a short list of aspects that need to be considered to list in the SOPs in COUSIN:

- Phenotyping SOPs:
  - Metadata collection: protocols for sowing, planting, preparation of experiments, etc.
  - Measurement techniques: methods used to assess traits, use the equipment, setup drone flight, etc.
  - Data acquisition tools: instruments, software, and formats used to capture data.
- Genomics SOPs:
  - Sample preparation: DNA/RNA extraction protocols and quality control checks
  - Sequencing protocols: sequencing libraries, choosing sequencing platforms, and running bioinformatic pipelines.
  - Data storage and management: file formats, data standards, naming conventions, and storage locations
- Chemotyping SOPs:
  - Sample preparation: methods for sample extraction, purification, and analysis, etc.
  - Chemical analysis protocols: instructions for operating instruments, collect, process and interpret data.
  - Data storage, integration and comparison: how is chemotype data linked with other data and stored, etc.

2.2.1 Exchanging SOPs between partners.

Sharing SOPs within and between organizations is crucial for strengthening collaborative output and scientific reproducibility. SOP exchange can occur through several channels within COUSIN. The foundation of the collected information must be created via a centralized repository. Within the COUSIN project it is essential to setup an centrally accessible data platform (by work package 5) that enables data, metadata and protocol sharing in a structured way where researchers from partners can upload, share and update the content with version tracking. It's advised whenever possible to use standardized formats, which means the use of machine-readable formats (e.g., XML, JSON) for interoperability between systems. Here we would like to refer to the Minimal Information for a Plant Phenotyping Experiment guidelines for the phenotypic data collection (MIAPPE). MIAPPE is an open, community driven, data standard designed to harmonize data from plant phenotyping experiments. MIAPPE provides a specification including a checklist and a data model of metadata required to adequately describe plant phenotyping experiments (MIAPPE.org, 2024). Each document must have clearly findable version control information, which allows tracking changes in SOPs to ensure that users are always working with the most up-to-date version.

2.2.2 Minimal information for each SOP document

To ensure that SOPs can be effectively shared and understood, it is important to capture relevant information that helps the user how to benefit most from the SOP. Therefore, version information is essential, capturing the SOP version, date of last update, and a change log within the document. Next to that the author(s) and affiliation are essential identifying the contributors and their associated institutions. Wherever possible refer to applicable standards and guidelines (e.g., ISO, MIAPPE, etc.) that the procedures comply with. And to allow search processes and machine readability in a near future, a short list of keywords and used ontologies boost re-usability. Tagging SOPs with appropriate keywords and used ontologies helps to facilitate searching and categorization, and optionally ontologies can also support translation to local languages.

2.2.3 Several types of metadata

Capturing comprehensive metadata is as important as capturing the primary data. Metadata provides the contextual information needed for data interpretation, reproducibility, and integration across studies. We would like to illustrate below that there are multiple levels and types of metadata depending on the type of

SOP that is delivered. On the one hand we have the **descriptive metadata** that describes the basic characteristics of the dataset, including title, description, authorship, and keywords. For example: The name of the plant species in a phenotyping study, or the experimental setup in a chemotyping lab analysis. Next to that we have **technical metadata**, which details the technical aspects of the data acquisition process of a specific instrument. For example: The instrument settings for a mass spectrometry run, or the sequencing platform used. Then, there is also potentially relevant **provenance metadata**, which keep track of the origin and history of the data in a data processing pipeline. For example, this lists the sample source (e.g. accession number), collection date, and any preprocessing steps applied before analysis, and post-processing steps carried out on the raw data. Then there are **quality related metadata** aspects which captures quality indicators, such as error, events and warnings provided by either the equipment of personnel involved that might help in the data analysis steps and could provide an explanation of potential data anomalies, and next to that replicate numbers, read depths, and validation measurements, etc. For example: The number of quality-filtered reads in a genomic dataset or the signal-to-noise ratio in a chemometric data. And there are also potentially relevant **contextual metadata** elements, which provides information on the experimental design, including factors such as time points, treatments, and environmental conditions. For example, the weather conditions, soil characteristics, GPS-location for field experiments, and temperature, humidity and other settings and actuals during a controlled environment phenotyping experiment.

2.2.4 Best Practices for Metadata Management

Using established metadata standards and guidelines ensures compatibility and consistency across different datasets and systems. Relevant standards to be considered for each author of a COUSIN SOP to use are for **phenotyping:** MIAPPE (Minimum Information About a Plant Phenotyping Experiment), for **genomics:** Minimum Information about any (x) Sequence) from Genomic Standards Consortium (MIxS, 2024) and for **chemotyping:** ISA-Tab (Investigation, Study, Assay) framework for managing metadata in omics studies. Employing metadata schemas and controlled vocabularies/ontologies (e.g., Plant Ontology, Gene Ontology) enhances data findability and reusability. Preferable these metadata files should be structured in a way that allows integration into data management systems (JSON/XML formats). On the other hand, partners need to benefit from automated metadata capture and management tools that are available, as this reduces human error and improves efficiency. For example, electronic lab notebooks could support the generation of a SOP in a semi-automated way, and LIMS (Laboratory Information Management Systems) exports could help providing well-structured and automated metadata descriptors in real-time as part of the workflow. Below is an example of fully automated metadata export designed based on the MIAPPE guidelines from an advanced imaging system that is mounted on a small tractor, which has been purchased by the Netherlands Plant Eco-phenotyping Centre (NPEC).

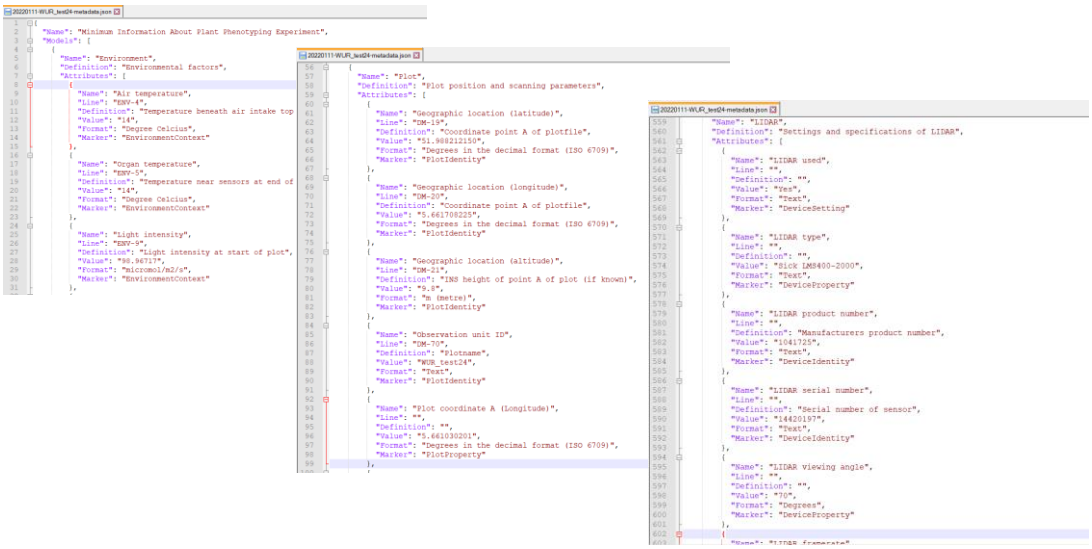


Figure 1: Illustrating the content of a JSON text file automatically generated by the software on the NPEC TraitSeeker device with climate data, GPS location of a plot and specific information of the sensors used incl. brand, product number and serial number.

2.2.5 Storage and accessibility of Data

COUSIN WP5 has provided a 4Tb data storage solution, which provides access to the consortium. Here all relevant documents, data and information generated over the life-time of COUSIN will be placed. Processed data, which means measured traits and all relevant metadata and used SOPs will be stored in this data platform, while access has been enabled for all COUSIN partners in a reliable and secure way. Data security is essential, therefore access control will be setup linked to individual COUSIN researchers involved, and regular backups will be carried out to prevent data loss. As described in the COUSIN data management plan, the data storage complies with relevant regulations (e.g., GDPR for personal data). The documents, datafiles etc will be stored in a nested folder format using the ISA-tab structure (ISA-Tab format, 2024), integrating the FAIRdomSeek principles which allows scalability, which is the ability to manage growing volumes of high-dimensional data (FAIRDOM Framework, 2024). On the other hand, 4Tb is not sufficient to place all raw data from all partners, to illustrate this – within the Netherlands Plant Eco-phenotyping since 2018 more than 1.200 terabytes of data has been collected which is placed in the affordable local tape archive access to online and cold storage is enabled by iRods (Glacier-Like Tier of Storage for Data-Driven Organizations, 2024). Copying these amounts of data to a different location would take weeks, and placing it in a commercial online cloud provider would costs >500k per year. Therefore the raw bulky data needs to be kept at each partner's data storage system, and a good description and a fixed link must be enabled from the COUSIN data platform, to enable access in a FAIR way.

Only at the moment when a publication or a public report will be generated referring to a specific subset of the collected data, a DOI needs to be generated. To explain – a Digital Object Identifier (DOI) is a string of numbers, letters and symbols used to uniquely identify an article or document, and provides a permanent web address (URL). This is in more detail described in the COUSIN Data Management Plan: all datasets uploaded to public data provisioning repositories and databases will be assigned a global unique DOI, which can be universally resolved to a particular digital object to be retrieved. And also software used should at that moment be cited in the document including the version (if unknown the date of access should be used) and an identifier, a DOI or a URL to where the software exists.

Within the context of software development, SOPs are also used to establish a standardized and secure approach to DevOps operations to ensure compliance with industry regulations, enhance security measures, to improve the reliability and resilience of data platforms, minimize downtime, and enable rapid scaling of applications and infrastructure, but this goes beyond the scope and capabilities of most of the COUSIN partners. We will set up a COUSIN organisation on GitHub to track technical assets generated in the project (source code, technical documentation, guidelines, tutorials, etc.). This way, we can coalesce all technical resources from COUSIN in one place without cluttering the main project website. Persistent IDs can be obtained for source code repositories and published for instance on Zenodo and other FAIR data sites, research articles, etc.

3 Results

In this chapter we will provide several examples of procedures that are either full standardized, or non-standardized to illustrate how relevant experiments linked to COUSIN are carried out. Within COUSIN systematic characterisation across CWRs and pre-existing CWR-based breeding populations of the five flagship crops will lead to the adoption of new or updated SOPs for data acquisition and storage, for biotic and abiotic stress resistance and adaptation traits (T3.1), and for (non-)nutritional and health traits (T3.3) where the upcoming example shown will serve as a starting point.

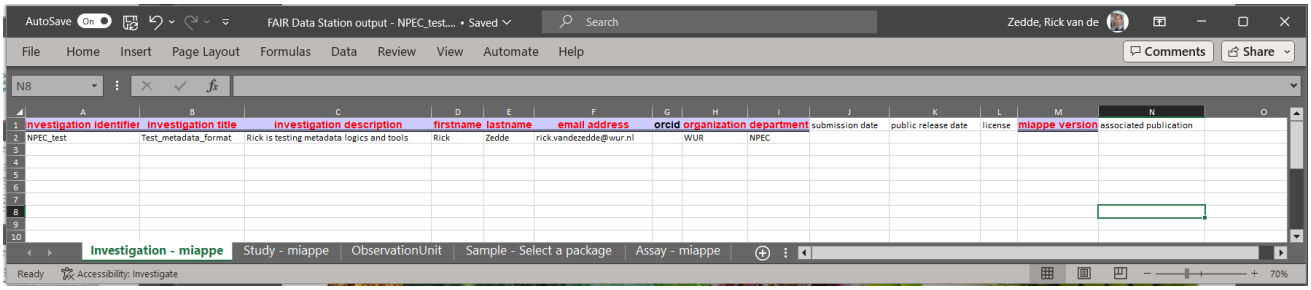
3.1 Dataset metadata example as presented in the COUSIN data management plan

Dataset number		*COUSIN internal ID*
Is it an open dataset (publicly accessible)?		*Yes/No*
If "No", please explain why you cannot open the data		
Open data repository for publication		URL
Description		Describe in 1-2 paragraphs
Intended use		Describe in 1-2 paragraphs
Keywords		Taken from GA (task description)
Data format		
Type of data		*Select an option from the list*
Format		*Examples:* • *Numerical data: CSV, XLSX, ODS* • *Textual data: PDF, HTML, TXT* • *Multimedia: JPEG, MP4, TIFF, PNG, MP3, WAVE* • *Geodata: SHP, GeoTIFF*
Dataset is		• *Fixed: Never changed after collection/generation* • *Growing: New data may be added, old data is never changed or deleted* • *Revisable: New data may be added, old data may be changed or deleted*
Discipline		
Is instrument specific?		*If "Yes", please answer the subquestions below*
Equipment		
Format		
Origin/lineage - *Define and describe data source/origin. Data can be gathered from different sources*		
Observational		*Data captured in real time - often not reproducible i.e. sensor readings, images, telemetries, sample data…*
Experimental		*Data from lab equipment, often reproducible, but with high costs - i.e. chromatograms, magnetic fields readings…*
Simulation		*Data generated by computational models where model and metadata are equally important to output data - i.e. climate models, economic models, materials models,…*
Derived/compiled		*Data coming from analysis or compilation. Reproducible but with high costs - i.e. the results of text and data mining, compiled databases…*
Reference or canonical (links)		*Collection or conglomeration of smaller (peer-reviewed) datasets published and curated - i.e. chemical structures, gene sequence databanks, spatial data portals…*
FAIR access		
Is there any specific license model?		Examples: CC-BY, CC0, etc.
Target audience/users		Researchers, practitioners, public, etc. Describe in 1-2 paragraphs
Possible applications		Describe in 1-2 paragraphs
Any metadata provided?		
Additional comments		

Table 3.1 Template describing a dataset created in the COUSIN project (original development by RTDI, available under a CC-BY-NC license).

3.2 COUSINE Milestone 9 – metadata collection

Within COUSIN an agreement has been arranged to capture metadata completely in line with the MIAPPE guidelines, and a practical approach has been chosen that each experiment needs to fill in an excel file in which the fields need to be filled in in 5 tab-pages. The main concern is the amount of hours a researcher needs to invest to fill in all these fields, plus to retrieve all the specific IDs and/ or understand which IDs and information is required. During the COUSIN experiments practical evaluations and easy-to-use tools will be developed to speed up this metadata collection process. An example of such a semi-automated data fill-in workflow can be found in this webpage/ linked to NPEC and the data scientists involved: https://data.m-unlock.nl/configurator based on the ISA-JSON specifications described here: https://isa-specs.readthedocs.io/en/latest/index.html

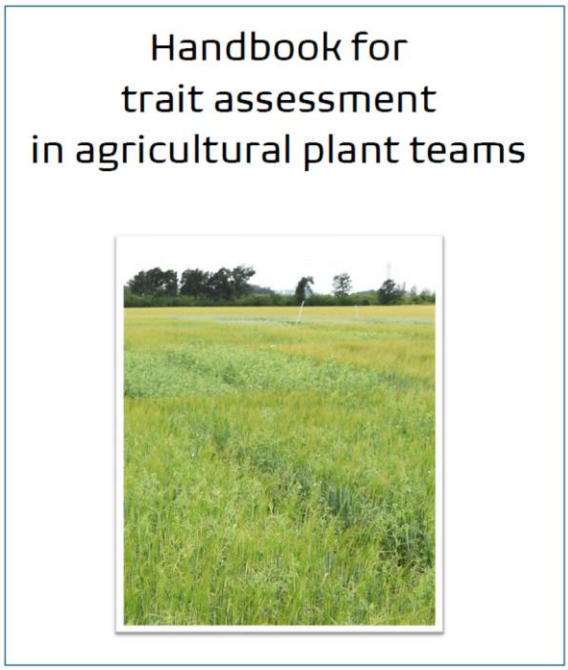


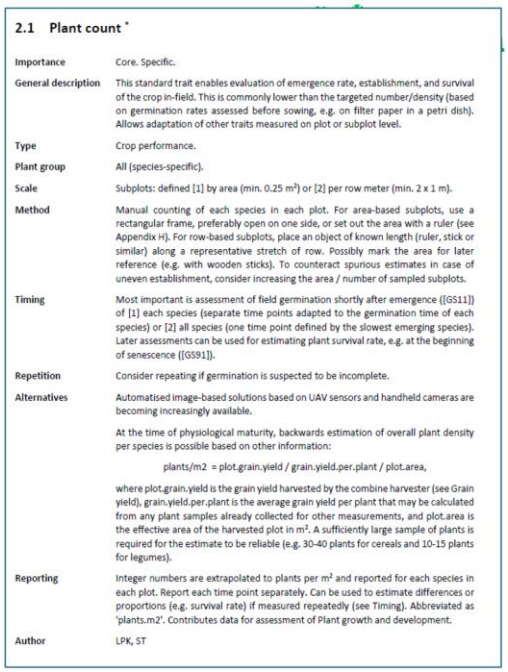
The tab pages and the fields that need to be filled in are, where the red fields are mandatory:

1. **'Investigation – miappe'** with the fields: investigation identifier, investigation title, investigation description, firstname, lastname, email address, orcid, organization, department, submission date, public release date, license, miappe version, associated publication.
2. **'Study – miappe'** with the fields: study identifier, study description, study title, investigation identifier, start date of study, end date of study, contact institution, geographic location (country), experimental site name, geographic location (latitude), geographic location (longitude), geographic location (altitude), description of the experimental design, type of experimental design, observation unit level hierarchy, observation unit description, description of growth facility, type of growth facility, cultural practices, map of experimental design.
3. **'ObservationUnit'** with the fields: observation unit identifier, observation unit name, observation unit description, study identifier, observation unit type, observation unit external id, spatial distribution, observation unit factor value, biological material id, biological material external id, organism, genus, species, infraspecific name, biological material latitude, biological material longitude, biological material altitude, biological material coordinates uncertainty, biological material preprocessing, "material source id (holding institute/stock centre, accession)", material source doi, material source accession number, material source accession name, material source institute code, material source institute name, material source other identifiers, material source latitude, material source longitude, material source altitude, material source coordinates uncertainty, material source description.
4. **'Sample - miappe'** with the fields: sample identifier, sample description, sample name, ncbi taxonomy id, scientific name, biosafety level, observation unit identifier, collection date, geographic location (country and/or sea), data file path.
5. **'Assay – miappe'** with the fields: assay identifier, assay description, protocol, facility, assay date, sample identifier, variable accession number, trait, method, method accession number, method description, reference associated to the method, scale, scale accession number, time scale.

3.3 Handbook for trait assessment in agricultural plant teams

Below one relevant example is chosen from the Handbook for trait assessment in agricultural plant teams (Kiær, 2020). This handbook was developed to help facilitate coordinated assessments across multiple field trials of plant teams, specifically, legume-cereal intercrops and grassland mixtures. In preparation of the H2020-funded project DIVERSify they identified the need for a handbook combining protocols for assessing plant traits as well as agronomic characteristics, filling a gap that other handbooks were not addressing.





2.1 Plant count *

Importance	Core. Specific.
General description	This standard trait enables evaluation of emergence rate, establishment, and survival of the crop in-field. This is commonly lower than the targeted number/density (based on germination rates assessed before sowing, e.g. on filter paper in a petri dish). Allows adaptation of other traits measured on plot or subplot level.
Type	Crop performance.
Plant group	All (species-specific).
Scale	Subplots: defined [1] by area (min. 0.25 m^2) or [2] per row meter (min. 2 x 1 m).
Method	Manual counting of each species in each plot. For area-based subplots, use a rectangular frame, preferably open on one side, or set out the area with a ruler (see Appendix H). For row-based subplots, place an object of known length (ruler, stick or similar) along a representative stretch of row. Possibly mark the area for later reference (e.g. with wooden sticks). To counteract spurious estimates in case of uneven establishment, consider increasing the area / number of sampled subplots.
Timing	Most important is assessment of field germination shortly after emergence ([GS11]) of [1] each species (separate time points adapted to the germination time of each species) or [2] all species (one time point defined by the slowest emerging species). Later assessments can be used for estimating plant survival rate, e.g. at the beginning of senescence ([GS91]).
Repetition	Consider repeating if germination is suspected to be incomplete.
Alternatives	Automatised image-based solutions based on UAV sensors and handheld cameras are becoming increasingly available. At the time of physiological maturity, backwards estimation of overall plant density per species is possible based on other information: plants/m2 = plot.grain.yield / grain.yield.per.plant / plot.area, where plot.grain.yield is the grain yield harvested by the combine harvester (see Grain yield), grain.yield.per.plant is the average grain yield per plant that may be calculated from any plant samples already collected for other measurements, and plot.area is the effective area of the harvested plot in m^2. A sufficiently large sample of plants is required for the estimate to be reliable (e.g. 30-40 plants for cereals and 10-15 plants for legumes).
Reporting	Integer numbers are extrapolated to plants per m^2 and reported for each species in each plot. Report each time point separately. Can be used to estimate differences or proportions (e.g. survival rate) if measured repeatedly (see Timing). Abbreviated as 'plants.m2'. Contributes data for assessment of Plant growth and development.
Author	LPK, ST

3.4 Standard phenotype definitions in an Excel sheet

Here a list of standard phenotype definitions aka SOPs as it is used for field characters in barley. They align with those used in a recent multi-site project 'BARN' as was used to described the genomic resources for a historical collection of cultivated two-row European spring barley genotypes. (Miriam Schreiber, 2024). These traits are handy to use for an overview, but the risk to result in differences between observers is large if no ring-test/ calibration training is provided.

Developmental traits	
Days to awn tipping	50% of the main tiller awns per plot have emerged up to 1 cm out of the flag leaf sheath. Recorded as days since sowing
Days to heading	50% of the main tiller spikes per plot have emerged halfway out of the flag leaf sheath. Recorded as days since sowing
Days to senescence	50% of the main tiller peduncles per plot are senescent (yellow). Recorded as days since sowing
Days from awn tipping to heading	Derived from days to awn tipping and days to heading
Days from awn tipping to senescence	Derived from days to awn tipping and days to senescence
Days from heading to senescence	Derived from days to heading and days to senescence
Growth habit (GH)	Visual evaluation using a scale of 1 (erect), 2 (intermediate) and 3 (prostrate). Recorded at the onset of stem elongation
Height and length traits	
Peduncle base height	Height of the base of the peduncle in cm
Flag leaf blade height	Height of the flag leaf sheath in cm
Culm height	Height of the base of the spike in cm
Plant height	Height of the top of the spike in cm
Awn tip height	Height of the tip of the awns in cm
Spike base to flag leaf	Calculated distance from base of spike to flag leaf sheath (auricle) in cm
Peduncle length	Calculated distance from base of spike to base of peduncle in cm
Awn length	Calculated distance from tip of awns to top of spike in cm
Spike culm ratio	Spike length divided by culm height

Screen capture of the content of the BARN project stand phenotype definitions, on which barley plants are scored.

3.5 Standard Operating Procedure for processing and analyzing UAV data with Photogrammetry

Usage of data processing tools also requires a standard way of working in which a list of steps, software button clicks need to be carried out to generate a certain identical result. In this unpublished document used internally within Wageningen University & Research, instructions are given to enable research to process and analyse their own Unmanned Aerial Vehicle (UAV) aka drone data using a processing technique called photogrammetry.

Processing and analyzing UAV data with Photogrammetry

Harm Bartholomeus[1], Niels Anders[2], Corjan Nolet[2],

[1]Laboratory for Geo-Information and Remote Sensing, Wageningen University

[2]Soil Physics and Land Management, Wageningen University (former employees)

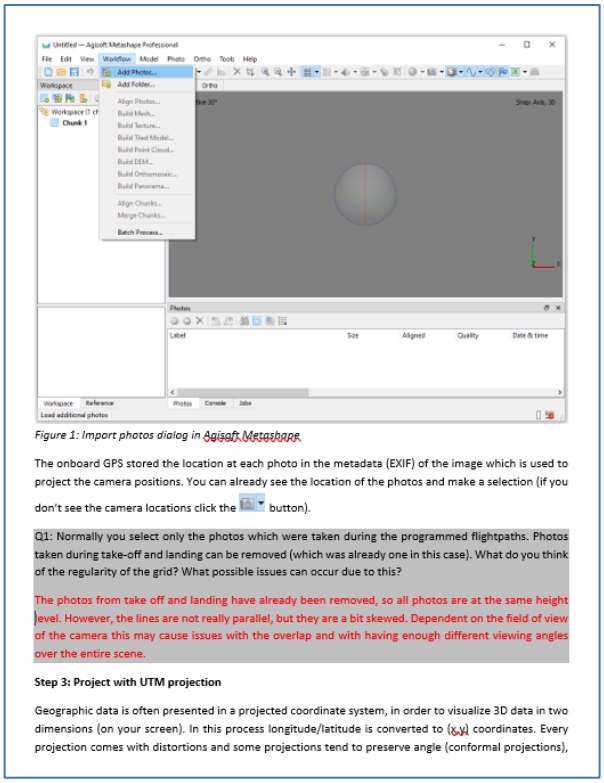


Figure 1: Import photos dialog in Agisoft Metashape

The onboard GPS stored the location at each photo in the metadata (EXIF) of the image which is used to project the camera positions. You can already see the location of the photos and make a selection (if you don't see the camera locations click the button).

Q1: Normally you select only the photos which were taken during the programmed flightpaths. Photos taken during take-off and landing can be removed (which was already one in this case). What do you think of the regularity of the grid? What possible issues can occur due to this?

The photos from take off and landing have already been removed, so all photos are at the same height level. However, the lines are not really parallel, but they are a bit skewed. Dependent on the field of view of the camera this may cause issues with the overlap and with having enough different viewing angles over the entire scene.

Step 3: Project with UTM projection

Geographic data is often presented in a projected coordinate system, in order to visualize 3D data in two dimensions (on your screen). In this process longitude/latitude is converted to (x,y) coordinates. Every projection comes with distortions and some projections tend to preserve angle (conformal projections),

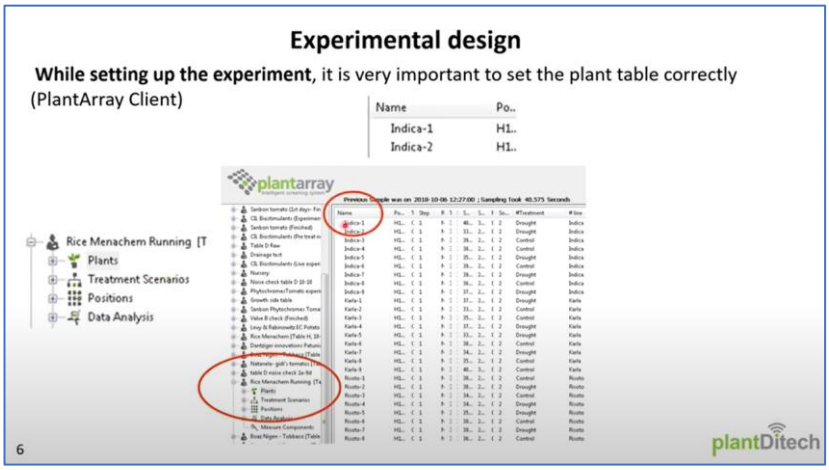
3.6 Operating procedure example in NPEC greenhouse for salt in quinoa

As is common in scientific papers, the Methods section is used to described the experiment, which also includes a substantial amount of the operating procedures to duplicate the results, to illustrate the content of such a section, we have added a small text which shows a non-standardized text from the quinoa trial carried out within the NPEC greenhouse on salt responses in quinoa genotypes (Viviana J. Roman, 2021)

The experiment was conducted between March and May 2019 at the Unifarm greenhouse facilities of Wageningen University & Research, The Netherlands. Plants were sown in trays filled with potting soil and transplanted to 4 L pots 16 days after sowing (DAS). The pots were filled with standard filtered sand (grain size 0.6–1.0 mm) and each pot contained 4 plants. To prevent evaporation, small PVC balls were put on the surface of the pots, surrounding the plants. The greenhouse air humidity was set to a minimum of 80% and the photoperiod to 16 h light. When the incoming shortwave radiation was below 200 Wm–2, additional lighting was supplied (100 Wm–2). The temperature in the greenhouse was set up to a minimum of 15°C during the night. During the day, ventilation was controlled so the temperature did not exceed 35°C. The plants were irrigated with half-concentrated Hoagland's nutrient solution. Salt stress treatment started 33 DAS with irrigation with 0.5 × Hoagland's solution plus 200 mM NaCl.

3.7 Tutorial video to transfer operating procedures

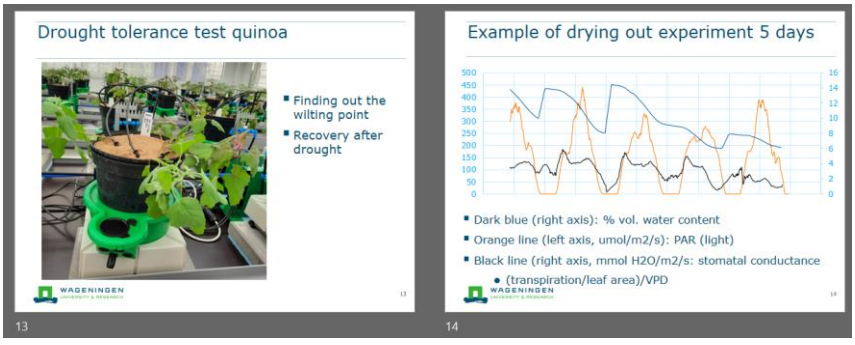
Another approach to hand-over detailed information that could be translated to a SOP is to explain this over a YouTube video, as was generated for an experiment in pots to accurately water the plants based on transpiration information per pot.



Prof. Moshelion describing the Plantarray system that was also used in the quinoa salt trial (Viviana J. Roman, 2021). Link: Screen capture from a scene in this YouTube video -https://www.youtube.com/watch?v=VMNd2ipSono

3.8 PowerPoint slides to share a semi-standardized way of working

Often methods are non-standardized shared in PowerPoints and other public outings, to at least give an impression of the protocols that are used. But of course preferred is a SOP which is also publicly shared with the researchers involved/ readers of the publication:



3.9 Chemotyping SOP Related to WP3 (nutritional analysis)

Below we provide an list of examples of clear lab protocols how to measurement certain nutritional components in samples. These are the most widely used protocols for all the samples/CWR for task 3.3 within COUSIN, and these will be carried out and should be able to generate identical results no matter in which lab these measurements were carried out. These SOPs for nutritional analysis will be evaluated during the COUSIN project, while other. For some CWR samples, we may specifically want to quantify other metabolites such as beta-glucans, nitrates, and glucosinolates, but these protocols need to be optimized first.

Condensed tannins quantification

Protocol Steps

1. *Add 1 mL of 80-100% methanol or ethanol to 100 mg of finely ground flour powder*
2. *Extract samples by sonicating for 30 min, followed by centrifugation at maximum speed for 20 min. Transfer the supernatant to a new microtube for use in step 3*
3. *In a 96-well microplate, add:*
    - *50 µL of the extracted sample or standard or extraction solvent*
    - *200 µL of 3% vanillin and 14% HCl in methanol*
4. *Incubate the mixture at room temperature in dark for 20 min.*
5. *Measure the absorbance at 500 nm.*
6. *Calculate the condensed tannin content using a calibration curve made with a catechin standard.*

Total Phenolics quantification

Protocol Steps

1. *Add 1 mL of 80-100% methanol or ethanol to 100 mg of finely ground flour powder*
2. *Extract samples by sonicating for 30 min, followed by centrifugation at maximum speed for 20 min. Transfer the supernatant to new microtube for use in step 3*
3. *In a 96-well microplate, add:*
    - *20 µL of the extracted sample, standard, or extraction solvent*
    - *100 µL of Folin–Ciocalteu reagent (previously diluted 2:10 (v/v))*
    - *100 µL of 7.5% (w/v) sodium carbonate in a 96-well plate*
4. *Incubate the mixture at room temperature in dark for 30 min.*
5. *Measure the absorbance at 750 nm.*
6. *Calculate the total phenolic content using a calibration curve made with a gallic acid standard*

Total Flavonoids quantification

Protocol Steps

1. *Add 1 mL of 80-100% methanol or ethanol to 100 mg of finely ground flour powder*
2. *Extract samples by sonicating for 30 min, followed by centrifugation at maximum speed for 20 min. Transfer the supernatant to new microtube for use in step 3*
3. *Add 275 µL of $NaNO_2$ into a microtube*
4. *Add 62.5 µL of the extracted sample or standard or extraction solvent*
5. *Vortex and incubate for 5 min*
6. *Add 37.5 µL of $AlCl^3$*
7. *Vortex and incubate for 6 min*
8. *Add 250 µL of NaOH*
9. *Vortex and transfer 250 µL (in triplicate) to the 96-well plate, then measure absorbance at 510 nm.*
10. *Calculate the total flavonoid content using a calibration curve made with a catechin standard*

Vanillin–Sulfuric acid assay for Saponins quantification

Protocol Steps

1. *Extract 75 mg of finely ground flour powder with 1 mL of 99.5% ethanol (analytical grade) and sonicate for 1h at RT.*
2. *Centrifuge the samples and transfer the supernatant to a new amber microtube (filter if needed).*
3. *In a new amber microtube:*
   -Add 100 µL of sample
   - Add 100 µL of vanillin 8% (w/v) in 99.5% ethanol (analytical grade)
   - Add 1000 µL sulphuric acid 72% (v/v) in water
4. *Incubate at 60 ◦C for 15 min with shaking.*
5. *Cool the tubes on ice*
6. *Vortex and transfer 250 µL (in triplicate) to a 96-well plate, then measure absorbance at 560 nm.*
7. *Calculate the total saponin content using a calibrating curve made with a diosgenin standard*

Coomassie (Bradford) Protein Assay Kit

Protocol Steps

1. *Extract 20 mg of finely ground flour powder in 2mL of TE buffer.*
2. *Vortex the samples for 15 min, sonicate for 15 min, and centrifuge at 15 000 rpm for 5 min.*
3. *Transfer the supernatant to a new tube and keep it on ice during the analysis.*
4. *Pipette 5 µL of each standard or sample into a 96-well microplate*
5. *Add 280µL of the Coomassie Reagent to each well and mix for 30 seconds.*
6. *Incubate plate for 10 minutes at room temperature.*
7. *Measure the absorbance at 595nm using a plate reader.*
8. *Calculate the protein content with a calibrating curve made with a BSA standard*

Microwave Digestion of Samples for Mineral Quantification (ICP-OES)

Protocol Steps

1. *Weigh 0.5g of finely ground flour powder into the digestion vessels*
2. *Add 10mL of HNO_3*
3. *Gently swirl the mixture and wait approximately 15 minutes before closing the vessel*
4. *Digest the sample using the following program:_*

Satge	*Temp (ºC)*	*Ramp (mm:ss)*	*Hold (mm:ss)*	*Pressure (psi)*	*Power (W)*	*Stirring*
1	*200*	*15:00*	*15:00*	*800*	*900-1800*	*off*

Transfer the clear resulting solution and dilute with water to a final volume of 50 mL

---

4 Discussion

It is essential to agree on species-specific traits, use uniform lists of traits and commonly agreed and standardized recording methods (i.e. SOPs), and several existing approaches such as IPGRI/Bioversity descriptor lists or UPOV descriptors could provide an important starting point for the development of a SOP. But although SOPs are commonly agreed upon as essential aspects of scientific experiments, it remains a challenge to define over-arching and generic procedures that can be implemented in other labs, other locations. Within COUSIN we have the ambition to overcome this situation by focusing on the five flagship crops, and carry our dedicated experiments.

To boost development and adoption, strategies need to be formulated. A first step which is already made within COUSIN is to generate easy-to-use templates. And to establish a process which allows improvements and discussions and therefore enabled regular involvement of partners. Subsequently, training is needed to show how to use the templates and the SOPs, and this needs to be pro-actively organised to promote

adoption. In the following paragraphs a number of dilemmas and concerns are presented that could hamper this update process.

4.1 Minimal or complete metadata SOPs

Too lengthy or too complicated SOPs will hardly be used, and therefore minimal and essential aspects of the operating procedure must be written down in a SOP, if it will take researchers too much time or if it is too specifically develop for a device/ unique situation the SOP will not be generally and widely used.

4.2 Dynamic and updates

The rational is that groups work in a way that is consistent and comparable. Where methods are different this should be reflected, but if the same trait or environmental factor is being measured, similar approaches should be used (especially for the same crop) or differences justified. Naturally, SOPs can be updated during the project.

4.3 Free format vs a structured SOPs lay-out.

The issue with all collected 'standardized' operating procedures is that they are developed and shared in different lay-outs, and in different information transition approaches. One solution could be, which will remain an open challenge, to explore if the effort of turning these SOPs into a machine-readable JSON format is worth the time investment.

4.4 Crop trait measurement manuals of previous projects towards SOPs

Based on existing crop trait measurement manuals of previous projects (for instance DIVERSify, ReMIX) COUSIN will need to select relevant traits to be included in the COUSIN Standardized Operating Procedures (SOPs) list for good quality data collection. This process needs to be driven by a need-to-have attitude.

5 Conclusions

Developing and exchanging SOPs for data acquisition and storage in phenotyping, genomics, and chemotyping requires a comprehensive and structured approach. Defining clear procedures, capturing detailed metadata, and adopting standardized practices are key steps toward achieving consistency, reproducibility, and collaboration across scientific communities. The integration of robust SOPs with effective metadata management ensures that datasets are not only useful in the context of their original study but also valuable resources for future research. By adhering to these principles, researchers can enhance data interoperability and contribute to the global research ecosystem. The main challenge is formulating correct and usable SOPs and promote adoption and usage in the daily activities of researchers.

6 References

EC Europe. (2024, 9 19). Retrieved from Guidelines on FAIR Data Management in Horizon 2020: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

FAIRDOM Framework. (2024, 9 19). Retrieved from https://fair-dom.org/fairdom_framework

Glacier-Like Tier of Storage for Data-Driven Organizations. (2024, 9 19). Retrieved from iRods: https://irods.org/2022/09/spectra-logic-and-irods-consortium-partner-to-provide-glacier-like-tier-of-storage-for-data-driven-organizations/

ISA-Tab format. (2024, 9 19). Retrieved from ISA Model and Serialization Specifications: https://isa-specs.readthedocs.io/en/latest/isatab.html

Kiær, L. &. (2020). *Handbook for trait assessment in agricultural plant teams.*

MIAPPE.org. (2024, 9 19). Retrieved from MIAPPE is an open, community driven, data standard designed to harmonize data from plant phenotyping experiments. MIAPPE provides a specification including a checklist and a data model of metadata required to adequately describe plant phenotyping experime

Miriam Schreiber, R. W. (2024). Genomic resources for a historical collection of cultivated two-row European spring barley genotypes. *Nature*.

MIxS. (2024, 9 19). Retrieved from Minimum Information about any (x) Sequence: https://fairsharing.org/262

Viviana J. Roman, R. v. (2021). High-Resolution Analysis of Growth and Transpiration of Quinoa Under Saline Conditions. *Frontiers in Plant Science*, DOI: https://doi.org/10.3389/fpls.2021.634311.

Wikipedia. (2024, 9 19). Retrieved from Standard Operating Procedures: https://en.wikipedia.org/wiki/Standard_operating_procedure

Annex 1: Glossary and acronyms

- *Dataset*: A collection of data obtained from a single source or combined from different sources that is intended to be used in a single study or analysis.
- *Data pipeline*: Sets of processes that usually consists of a task or a group of tasks and dependencies between them, organised in a sequential manner (e.g. linked graph), that move and transform data from various sources to a destination where a new value can be derived.
- *Digital Object Identifier* (DOI): Permanent link in the form of a globally unique alphanumeric code that identifies an electronic content unambiguously.
- *Embargo period*: In scientific publishing, a period during which access to datasets or other resources required to reproduce a study or analysis is not provided, to protect legitimate exploitation or commercial interests of authors and organisations who hold their IP/IPR.
- *Metadata*: Data providing details about the representation format, meaning and other useful information for finding, accessing, interpreting and reusing a dataset. This may be represented either in human-friendly, machine-friendly format or both.
- *Qualified references*: A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)
- *Repository*: An accessible information system providing storage and management services for datasets, source code, documentation and other digital resources pertaining to a project or study. These digital resources must be uniquely identified within the repository.
- *Repeatability*: The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation (same team, same experimental setup) (https://www.acm.org/publications/policies/artifact-review-and-badging-current).
- *Reproducibility*: The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts (different team, same experimental setup) (https://www.acm.org/publications/policies/artifact-review-and-badging-current).
- *Replicability*: The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that

an independent group can obtain the same result using artifacts which they develop completely independently (different team, different experimental setup) (https://www.acm.org/publications/policies/artifact-review-and-badging-current).

- *RESTful API*: Also known as REST API is an application programming interface that adheres to the constraints of the REST (Representational State Transfer) architecture. In a client-server network communication system, when a client requests a digital resource via a RESTful API, a representation of the state of the resource is provided to the requester or endpoint over the HTTP protocol, using a standard format (JSON, HTML, XLT, PHP, plain text, etc.).
- *Uniform Resource Identifier* (URI): A unique character sequence that identifies abstract or physical resources in a communication network.
- *Uniform Resource Locator* (URL): A unique reference that specifies the location of a digital resources on a computational device connected to a network.
- Workflow: Series of steps or tasks that must be conducted to obtain, process, transform, validate or analyse data to achieve a specific outcome or goal.

Acronym	**Description**
API	Application Programming Interface
BrAPI	Breeding API
CA	Consortium Agreement
CGIAR	Consultative Group of International Agricultural Research
CSV	Comma Separated Values
CWR	Crop Wild Relatives
DOI	Digital Object Identifier
FAIR	Findable, Accessible, Interoperable, Reusable
GA	Grant Agreement
HTTP	Hypertext Transfer Protocol
IP	Intellectual Property
IPR	Industrial Property Rights
JSON	JavaScript Object Notation
MCPD	Multi-Crop Passport Descriptors
MFA	Multi-Factor Authentication
MIAPPE	Minimum Information about a Plant Phenotyping Experiment
SOP	Standard Operating Procedures
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
XML	Extensible Markup Language

Table A.1 List of acronyms used in this document

Annex 2: SOP example in JSON format:

Below is an example of a Standard Operating Procedure (SOP) in a structured text-based JSON format for a very simple and for reporting purpose only plant phenotyping. This is a simplified version and includes essential sections such as procedure steps, metadata, equipment, and parameters. Generated with support of the

OpenAI tool ChatGPT. The advantage of a SOP in such a format is the machine-readability, software is able to extract useable elements and report/ share them in a user-interface in a structured way.

JSON format style:


```
{
  "sop": {
    "title": "Standard Operating Procedure for Plant Phenotyping",
    "version": "1.0",
    "date": "2024-08-20",
    "author": {
      "name": "Rick van de Zedde ",
      "affiliation": "Wageningen Plant Research ",
      "contact": "rick.vandezedde@wur.nl"
    },
    "purpose": "This SOP describes the process of collecting and analyzing plant
phenotypic data, focusing on morphological and physiological traits.",
    "scope": "The procedure applies to all experiments involving field-grown or
greenhouse-grown plants intended for phenotypic analysis.",
    "responsibilities": [
      "Researcher: Performs phenotyping tasks according to the SOP.",
      "Technician: Maintains equipment and prepares samples.",
      "Data Manager: Ensures data capture and storage according to metadata
standards."
    ],
    "materials_and_equipment": {
      "equipment": [
        {
          "name": "Digital Caliper",
          "model": "Mitutoyo 500-196-30",
          "calibration": "Quarterly"
        },
        {
          "name": "Leaf Area Meter",
          "model": "LI-3100C",
          "calibration": "Annually"
        },
        {
          "name": "High-Resolution Camera",
          "model": "Nikon D850",
          "calibration": "N/A"
        }
      ],
      "supplies": [
        "Sample labels",
        "Markers",
        "Data sheets",
        "Storage containers"
      ]
    },
    "procedure": {
      "steps": [
        {
          "step_number": 1,
          "description": "Prepare the plant samples in the greenhouse or field by
labeling and recording plant IDs."
        },
        {
          "step_number": 2,
          "description": "Measure plant height using the digital caliper,
ensuring measurement is taken from the soil level to the tip of the main stem."
        },
        {
          "step_number": 3,
```

```
          "description": "Capture images of the plant from multiple angles using
the high-resolution camera. Ensure consistent lighting and background."
        },
        {
          "step_number": 4,
          "description": "Use the leaf area meter to measure leaf area. For each
plant, measure three representative leaves and calculate the average."
        },
        {
          "step_number": 5,
          "description": "Record all measurements in the data capture system,
along with associated metadata (e.g., date, time, environmental conditions)."
        },
        {
          "step_number": 6,
          "description": "Store the collected data in the designated database,
following the naming conventions and file formats specified."
        }
      ]
    },
    "health_and_safety": [
      "Wear appropriate protective clothing when working in the greenhouse or
field.",
      "Ensure proper handling of equipment, following manufacturer guidelines."
    ],
    "metadata": {
      "descriptive_metadata": {
        "experiment_id": "EXP_20240820_001",
        "project_name": "Drought Tolerance in Wheat",
        "location": "Greenhouse 5, Institute of Plant Sciences",
        "species": "Triticum aestivum",
        "variety": "Drought-resistant line DR-01"
      },
      "technical_metadata": {
        "instrument_settings": {
          "camera_resolution": "45.7 MP",
          "leaf_meter_mode": "Automatic"
        },
        "sampling_conditions": {
          "temperature": "25°C",
          "humidity": "60%",
          "light_intensity": "500 µmol m-2 s-1"
        }
      },
      "provenance_metadata": {
        "sample_origin": "Greenhouse, Block A",
        "collection_date": "2024-08-19",
        "operator": "Technician A"
      },
      "quality_metadata": {
        "quality_checks": [
          "All images reviewed for focus and clarity",
          "Calibration of digital caliper verified prior to measurements"
        ],
        "validation_methods": [
          "Measurements repeated three times for accuracy",
          "Data cross-checked with previous phenotyping data"
        ]
      }
    },
    "data_management": {
      "storage_location": "/data/phenotyping/2024/drought_tolerance_wheat/",
```

```
      "file_naming_conventions":
"projectname_experimentid_sampleid_measurementtype_date.ext",
      "file_formats": [
        "CSV for measurement data",
        "JPEG for images",
        "JSON for metadata"
      ],
      "backup_strategy": "Daily backups to cloud storage with version control."
    },
    "references": [
      {
        "title": "Phenotyping Best Practices in Plant Science",
        "author": "Smith et al.",
        "publication_year": 2021,
        "doi": "10.1016/j.plantsci.2021.01.001"
      }
    ]
  }
}
```