INnovations in plant Varlety Testing in Europe

Deliverable D7.4
A unified ontology for DUS and performance testing data

This deliverable is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817970.

Technical References

Project Acronym	INVITE
Project Title	INnovations in plant VarIety Testing in Europe
Project Coordinator	François Laurens
Project Duration	60 months

Deliverable No.	D7.4 - A unified ontology for DUS and performance testing data
Dissemination level [1]	PU
Work Package	WP 7 - Database management and data interoperability
Task	T 7.2 – Project data & code repository
Lead beneficiary	Partner n°24 (ACTA)
Contributing beneficiary(ies)	Partner n°10 (Agroscope), partner n°5 (WU), partner n°3 (U Hohenheim), partner n°18 (GEVES), partner n°16 (NAKTUINBOUW)
Due date of deliverable	July 2022 (month 20)
Actual submission date	July 2022

[1] PU = Public
PP = Restricted to other programme participants (including the Commission Services)
RE = Restricted to a group specified by the consortium (including the Commission Services)
CO = Confidential, only for members of the consortium (including the Commission Services)

Document history

V	Date	Beneficiary	Author
1	29/08/2022	ACTA	Thomas Gomez
2	26/09/2022	ACTA	Thomas Gomez
3	21/06/2023	ACTA	Thomas Gomez
4	17/07/2023	ACTA	Thomas Gomez

Summary

The deliverable D7.4, as part of the work package 7, provides recommendations for the implementation of data exchanges good practices. The goal is to harmonize the data semantics within the datasets to guaranty their interoperability between partners. The challenge focuses on finding a common ontology for the two types of variety testing data: DUS (Distinctiveness, Uniformity and Stability) and VCU (Value for Cultivation and Use).

In this document, we provide recommendations to ensure data interoperability based on the metadata. The goal is to present a realistic method based on existing ontology and standard that can be implemented in the prototype developed in task 7.4.

The document starts by detailing the project's expectations in terms of data exchange and semantic. Then, it explains the difference between an ontology and other data standards. To finish, it gives guidelines to implement good practices within the INVITE project.

Table of content

1 Needs for plant variety description

2.1. Project goals

The aim of the INVITE project is to foster the introduction of new varieties with high resilience towards biotic and abiotic stresses, high adaptation to sustainable management practices, and high resource use efficiency (RUE), through improved variety testing and better information to stakeholders on variety performance under a range of contrasting production conditions. This will be exemplified by major crop species that represent the main features of propagation, food and feed uses, and exhibit significant breeding activity in the EU.

The variety tests conducted within the INVITE project generate a high amount of data. The diversity of crops (apple, perennial ray grass, sunflower, soybean, wheat, maize, potato, tomato, rapeseed, and lucerne) and tests performed (DUS and VCU) generate different data formats that are difficult to exchange between organisations and between EU countries. In addition, the project re-uses a large amount of historical data. The WP7 is dedicated to facilitating data interoperability and exchanges within the consortium.

The deliverable 7.4 aims at providing the consortium with guidelines based on a common ontology to ensure the reusability of the data produced and processed in the project. Following the guidelines will help partners from different organisations to exchange datasets and can be used as a baseline for the prototype of database that will be developed in Task 7.4. This prototype will consist of developing a common database to store phenotypic and genotypic variety data and provide a user-friendly interface for Examination Offices (EOs) and Post-Registration Organisations (PROs).

2.2. Difficulties

Collecting variety data is useful for an organisation only if the data is understandable and if comparison with reference data is possible. However, the exchange of variety testing data encounters difficulties at several levels:

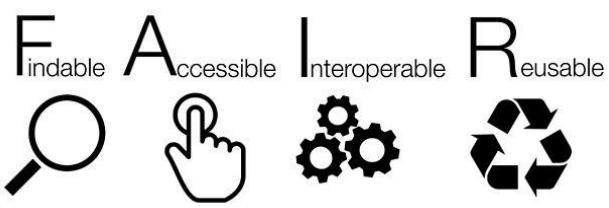
- Tests are performed on different varieties at the European level;
- Tests are closely related to the environmental conditions: weather, soil, etc;
- EU countries are selecting varieties based on different criteria;
- Data confidentiality blocks data exchange.

2.3. A long-term vision

Beyond proper collection, annotation, and storage; data management includes the notion of 'long-term care' of valuable digital assets. Those assets should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data.
Plant variety testing is a complex and time-consuming process that can be speed up by exchanging results between PROs and EOs. The project in anchored in a context of climate change where innovations in plant varieties are key to ensure food production and lower environmental impacts. The research led in the INVITE project is meant to pave the way to other data exchange project in the future.

2 A common ontology

3.1. F.A.I.R. data management



The FAIR (Findable, Accessible, Interoperable and Reusable) principles are the backbone of data management good practices. The data management strategy of the INVITE project must be compliant with the FAIR principles; it is therefore necessary to establish a framework for data sharing through a common ontology. Several points of attention were raised and need to be addressed to be consistent with the FAIR principles:

Findable

- Distributed model: no central phenomic data archive
- Same or compatible ID and Metadata policy
- Data portals

Accessible

- Complex life cycle: different types of data complexify the data management process

Interoperable

- Compatible standards
- Issues related to big data

Reusable

- Phenotype = Genotype x Environment x Cultural practice: there are different silos of metadata definition

- Provenance is complex to identify

International network for data interoperability already exists. It is key to understand the existing standards and to identify the ontology that best meets the needs of the INVITE project.

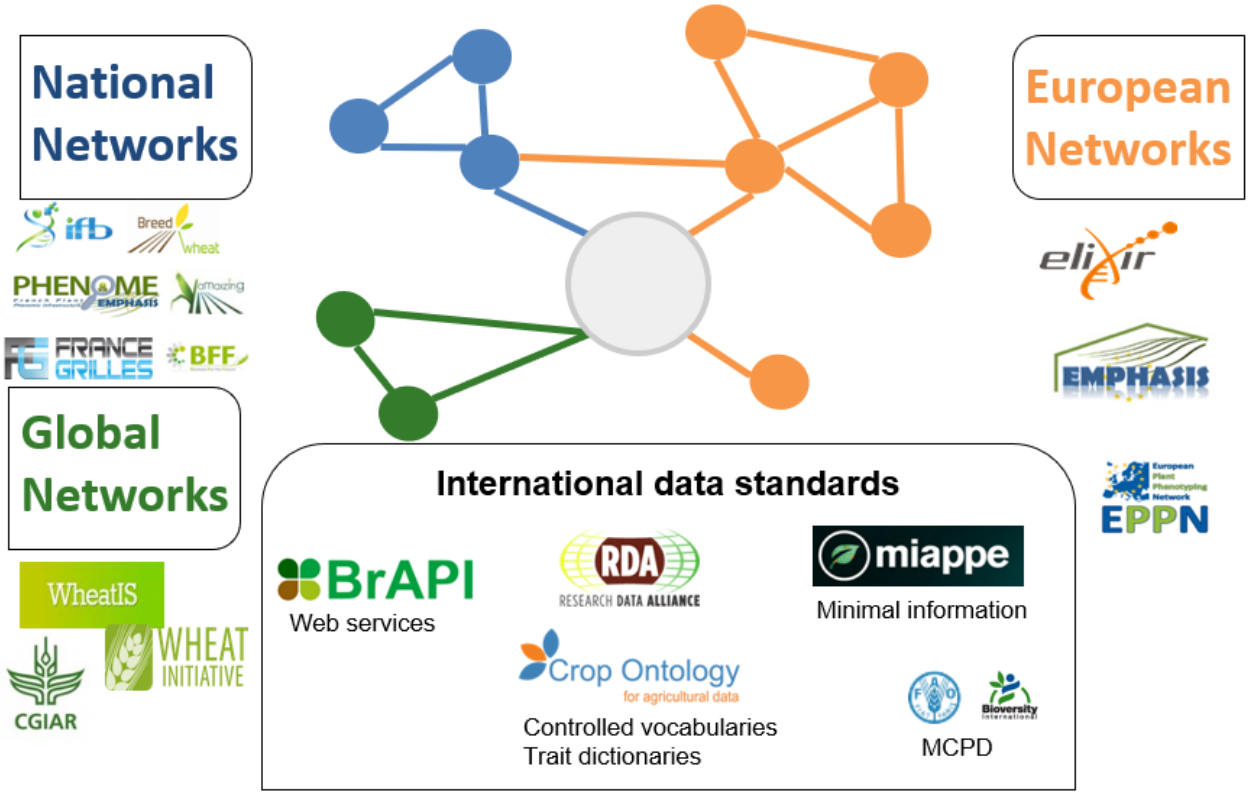


Figure 1: Example of interoperability in international network

Figure 1 above shows the diversity of existing networks at different levels: national, European and global. These networks bring together communities of scientists who produce and use data in their research. Data interoperability is at the heart of these networks to ensure the reusability of data. Part of their mission is to train and share good practice in data collection, storage and exchange. International data standards, such as the controlled vocabularies of the Crop Ontology project, are developed to harmonise data collection between network members. These international data standards are very detailed and diverse. They cover the vast majority of plant trait descriptions and should form a starting point for the common INVITE ontology.

The standard registry Fair Sharing[1] is used to have an overview of the existing standards.

[1] https://fairsharing.org/

3.2. What is an ontology?

In computer science and information science, an ontology encompasses a representation, formal naming, and definition of the categories, properties, and relations between the concepts, data, and entities that substantiate one, many, or all domains of discourse. More simply, an ontology is a way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject (Wikipedia, "Ontology", 2022).

In the INVITE project, the ontology is the vocabulary used to describe genotypic and phenotypic plant traits. Fortunately, there are already many existing ontologies dedicated to this academic subject.
However, the semantic aspect of the data collection is not the only standard used to ensure a FAIR data management strategy.

There are 3 different standard levels for FAIR data:

1. **Semantic**

- Description of the data
- Controlled vocabularies: term name and definitions
- Ontologies: semantic links between terms
- Biologist driven

2. **Structure**

- Formatting and organizing the data
- Data Models
- Standards : CSV, VCF, GFF, MIAPPE, etc.
- *Biologist & Computer scientist* driven

3. **Technical**

- Data integration and sharing
- Interoperability: tools and systems
    - GA4GH
    - Breeding API[2]
- *Computer scientist* driven

It is essential to keep in mind those 3 levels of standards to be able to exchange data between the project partners. Technical standards are interesting mostly to exchange data Machine to Machine and can be used for the prototype database development in Task 7.4.

[2] www.brapi.org

3.3. Reusing existing standards

The agronomic research led to the development of many different data formats. They are usually dedicated to a specific agronomic application. To select the right data standard, not only the fitness of the application should be taken in account, but also the size and involvement of the user ecosystem, the format flexibility, and the accessibility of the barrier to entry.
Two relevant standards were identified for the INVITE project: ICASA[3] and MIAPPE[4]. They both seem to answer to project requirements. However, interviews with researchers handling DUS and VCU data oriented the choice toward the ICASA standard. Therefore, it is the standard identified that suit best the INVITE project needs.

The MIAPPE standard is aimed more at biologists and computer scientists than at agronomists. It would be suitable for structuring DUS and VCU data, but it is not commonly used by agricultural researchers. Using the MIAPPE standard would require a significant amount of effort to adapt existing data and models to this format. In addition, the lack of experience in using MIAPPE for agronomic applications means that it may not correspond exactly to the needs of researchers.

3.3.1 ICASA: advantage for plant variety description

The use of ontologies can be complex due to their high level of detail and diversity. Describing the traits of a plant variety can require mobilising vocabularies from different ontologies: phenotypic traits of the plant, soil profile, climatic conditions, experimental protocols, etc.

We identified ICASA as a relevant guideline for the INVITE project. It stands for International Consortium for Agricultural Systems Applications. It was established to provide an effective linkage between advanced organizations that have experience in systems science and national research centers or agricultural universities and international agricultural research centers.

ICASA was created by members of the International Benchmark Sites Network for Agrotechnology Transfer (IBSNAT) from the Agricultural University in Wageningen (Netherlands). It is an historical initiative that has been created by and for agronomists, with a proven value over the years. The foundation of the standards is a master list variables that is organized in a hierarchical arrangement with major separations among descriptions of management practices or treatments, environmental conditions (soil and weather data), and measurements of crop responses (White J., 2013). ICASA is

[3] https://www.sciencedirect.chttps://www.miappe.org/om/science/article/abs/pii/0308521X95000284

[4] https://www.miappe.org/

an open community driven project that provide researchers and persons in charge of the experiments a kit with a list of variables to describe:

- Plant phenotypes
- Metadata about the trial
- Description of agricultural practices, description of soils, weather, etc.

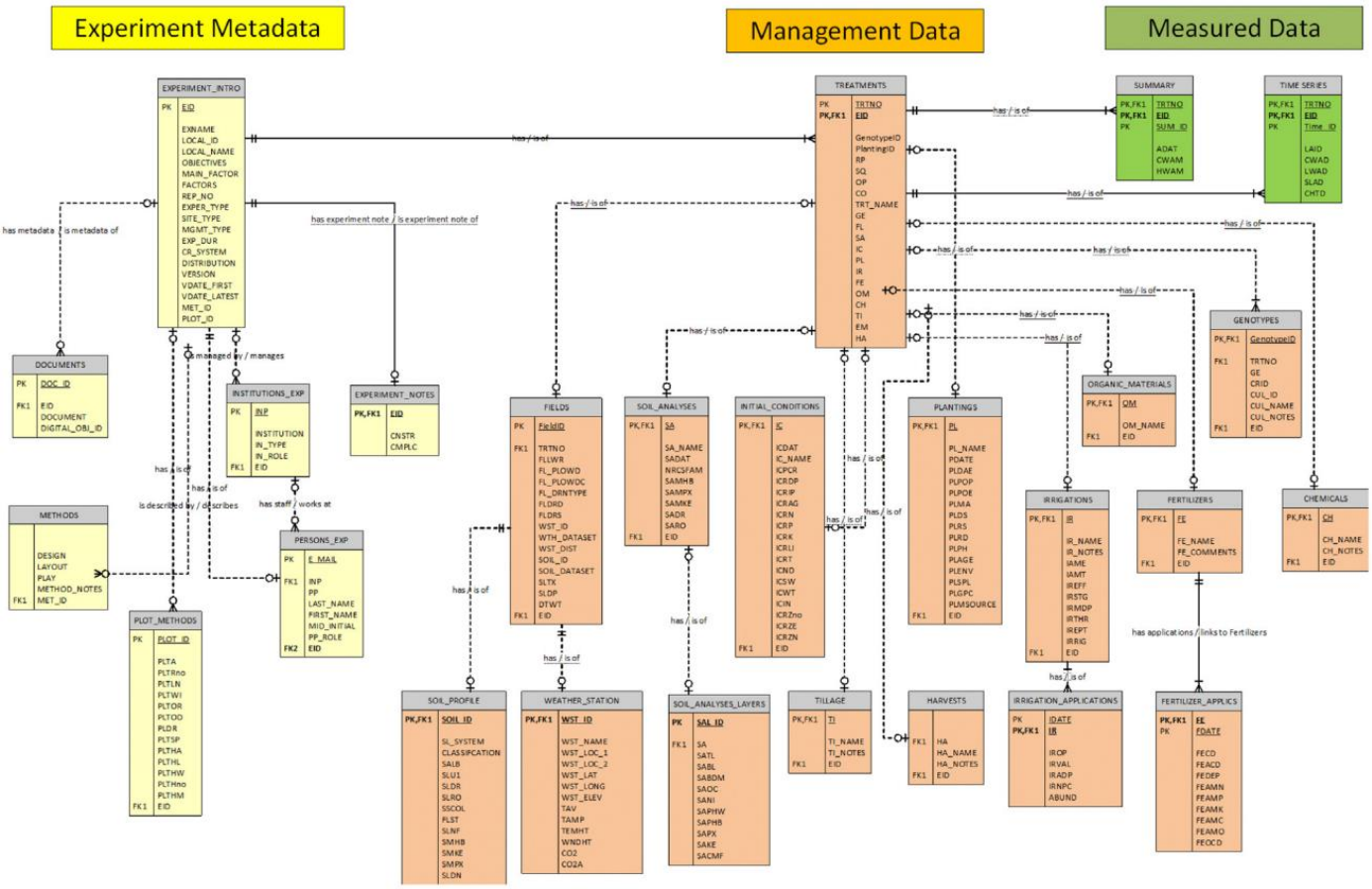


Figure 2: Entity-relation diagram (Pedersen, 2004) for the ICASA Version 2 standards.

The ICASA standard version 2.0[5] is currently used by many researchers modelling agricultural systems, such as INRAE experimental units. An on-going work on ontologies is being done in the context of a phenotyping platform with the creation of new template based on ICASA to describe sensor life records, interventions, etc.

[5] https://www.sciencedirect.com/science/article/abs/pii/S016816991300077X?via%3Dihub

DUS and VCU variety testing data have distinct purposes and defining a common ontology for both type of test is not realistic. However, the data structure proposed by ICASA offers a way to details the data collection with metadata. Using the ICASA standard would allow to:

- Make datasets understandable by anyone thanks to the metadata description and a common list of variables;
- Maintain the diversity of testing methods.

In addition, there are specific implementations of tools designed to support ICASA standard use and application, for example, the PHIS [6] an open-source information system for plant phenomics developed by INRAE or the Agricultural Model Intercomparison and Improvement Project (AgMIP[7]). Using the ICASA standard will pave the way towards data exchange automation by giving a clear framework for presenting DUS and VCU data.

3.3.2 Limit of the solution

The use of the ICASA standard would be very beneficial by offering a common format for experimental data, but shows some limitations:

- First, the definition of metadata for a specific dataset is very detailed and time consuming. It adds complexity to the data collection and extra work to the data provider.
- The metadata quality relies on the author. There are no guarantees or quality check before sharing data with other partners.
  A solution could be to use a convertor developed by INRAE to transform Excel files in JSON. The tool sends an error message in case of non-conformity which can help to guarantee data quality.
- Using ICASA vocabulary is complex and involves a training or providing pedagogical resources to help data providers to structure their data sets.
- The ICASA template provided to the partners might not fit exactly to the specificity of each testing protocols. It will require some extra work and a good understanding of the ICASA standard to convert the original protocol in the common format (using different scales, etc.).

[6] http://www.phis.inra.fr/

[7] https://agmip.org/

4 Use of the ICASA standard

4.1. Training material

The ICASA community shares information on the standard on the DSSAT[8] website and makes a template[9] viewable and downloadable.

The standards are being updated continuously to address the needs for adding new definitions for crop traits, environmental variables and other related crop model parameters. This list will be the main document to use in the INVITE project.

4.2. Use for the INVITE project

One of the aims of data standards is to digitise data, make it computer-readable and automate data exchange. There are different levels of data sharing, ranging from PDF to Linked Data. The ambition of the INVITE project is to target the intermediate level, which is to use a non-proprietary format for the data (CSV) to share data and metadata.

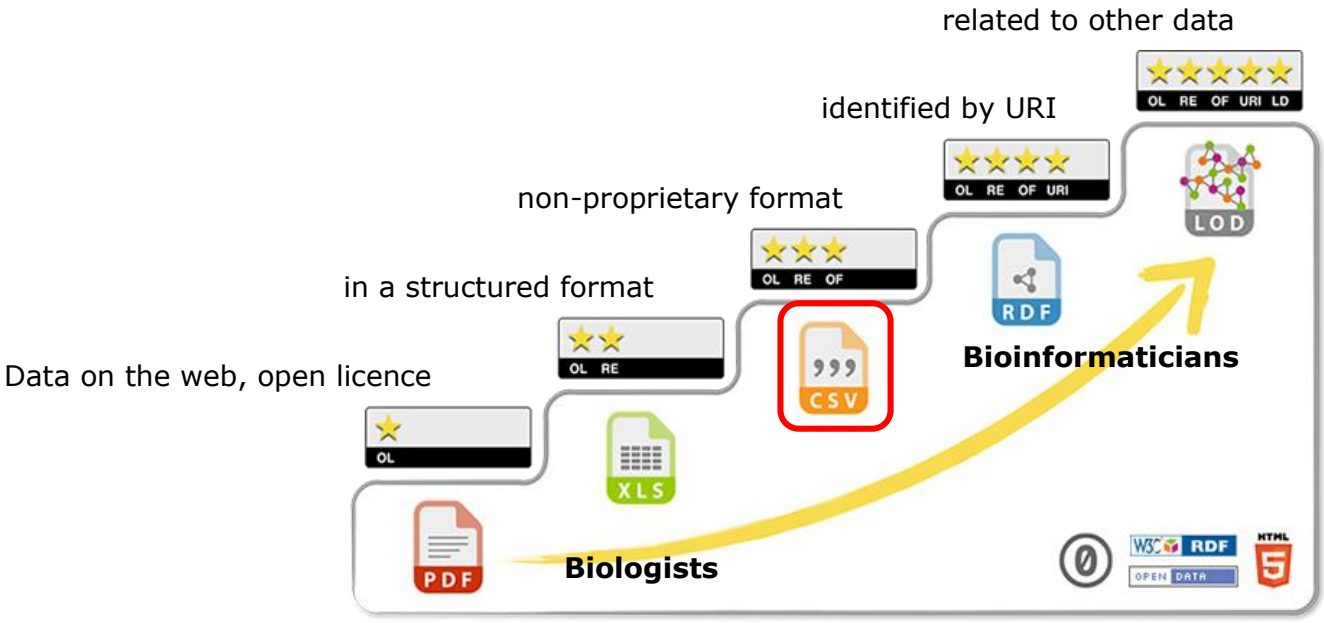


Figure 3: 5-star levels of data sharing

[8] https://dssat.net/data/standards_v2/

[9] http://www.tinyurl.com/icasa-mvl

In the long term, this initial work can be used as a research base for creating more advanced automated data sharing, such as RDF.

The metadata of the dataset will be the key element to ensure the FAIR principles. In a very practical way, each researcher will be asked to present their VCU data and metadata in an Excel file using the ICASA standard.

This data and metadata file will contain details on the general research organization (project, parties responsible), to plant materials and their provenance, experimental set-ups (time, place, purpose, layouts, independent variables), to phenotyping and environmental variables measured. The file will be based on the template provided by the ICASA community.

4.3. Sharing data and metadata

The NextCloud data sharing system developed for the INVITE project will be used to share dataset within the research partners. Original datasets and their associated ICASA standard file can be share as a pair and will both be uploaded in the dedicated folders.

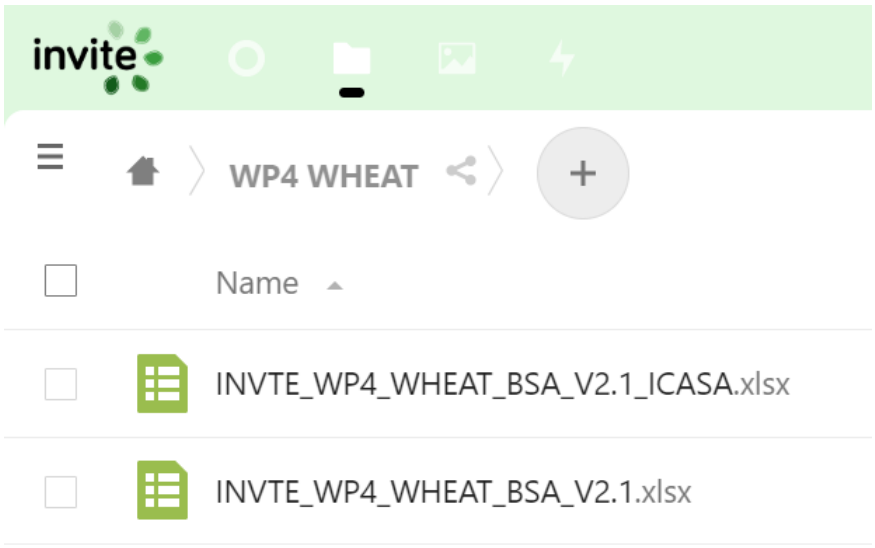


Figure 4: Example of data set and associated standardised ICASA file shared on NextCloud

The datasets available under the ICASA standard can be converted into JSON file and uploaded into the prototype of data sharing system that will be developed in the task 7.4.

5 Conclusion

The diversity of plant varieties, experimental protocols, soil qualities and weather conditions in Europe make the utilisation of a common ontology for DUS and VCU data a very difficult task. To address the issue of data interoperability in the INVITE project, we recommend focussing on an historical data format created to detail the dataset content and the method applied to collect these data. Such a standard already exists, with an active community using it: we recommend using the standard ICASA in the INIVTE project to describe datasets. This solution has many advantages:

- An existing method for describing plant phenotypes.
- It is designed to work with the diversity of experimental protocols.
- It is used and maintain by an active community of researchers and agronomists.
- It offers the possibility go further with data exchange automation using technological standards, by converting Excel file into JSON. The use of existing API standard such as BrAPI can be also applied to ICASA data format.

6 Annexe – links

ICASA standard

ICASA template with the master list of variables: http://www.tinyurl.com/icasa-mvl

Decision Support System for Agrotechnology Transfer (DSSAT) – Information on ICASA: https://dssat.net/data/standards_v2/

Agricultural Model Intercomparison and Improvement Project: https://agmip.org/

MIAPPE data format

CGIAR - Agricultural Ontologies in Use: The MIAPPE Standard: https://bigdata.cgiar.org/blog-post/agricultural-ontologies-in-use-the-miappe-standard/

Elixir - Training material: https://tess.elixir-europe.org/search?q=miappe

MIAPPE website: https://www.miappe.org/

MIAPPE metadata template: https://github.com/MIAPPE/MIAPPE/tree/master/MIAPPE_Checklist-Data-Model-v1.1/MIAPPE_templates