
# INnovations in plant Varlety Testing in Europe

# Deliverable D7.1
# Decision tree for data categorisation

**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. | D7.1 |
| Dissemination level [1] | Public |
| Work Package | WP 7 - Database management and data interoperability |
| Task | T 7.2 – Project data & code repository |
| Lead beneficiary | 24 (ACTA) |
| Contributing beneficiary(ies) | |
| Due date of deliverable | 01 December 2019 |
| Actual submission date | 19 November 2019 |

[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 | 28/11/2019 | Nesrine Mezghrani and François Laurens | Géraldine Hirschy |
| 2 | 29/11/2019 | Final deliverable Nesrine Mezghrani and François Laurens | Géraldine Hirschy |
| 3 | | | |
| 4 | | | |

# Summary

One of the objectives of the WP7 team is to facilitate the data interoperability and exchanges within the consortium all along the project.
Thus, a common **data sharing system** will be implemented and provided to developers, researchers and data providers and data users to manage all the data collected and generated in the project.

This data sharing system is a set of three tools that will be available in a webportal, hosted in the INVITE website.
In order to help the project partners to choose the right system according to their needs and restrictions, a decision tree is provided in this document and will be accessible online as soon as the three tools are released and properly set up (the due date is December 31th 2019).

# Table of content

# 1 Introduction

The objective of the WP7 is to facilitate the data interoperability and exchanges within the consortium and to set up both a prototype of a common database to store phenotypic and genotypic variety data as well as a user-friendly interface for Examination Offices (Eos) and Post-Registration Organizations (PROs).

This will be done by:

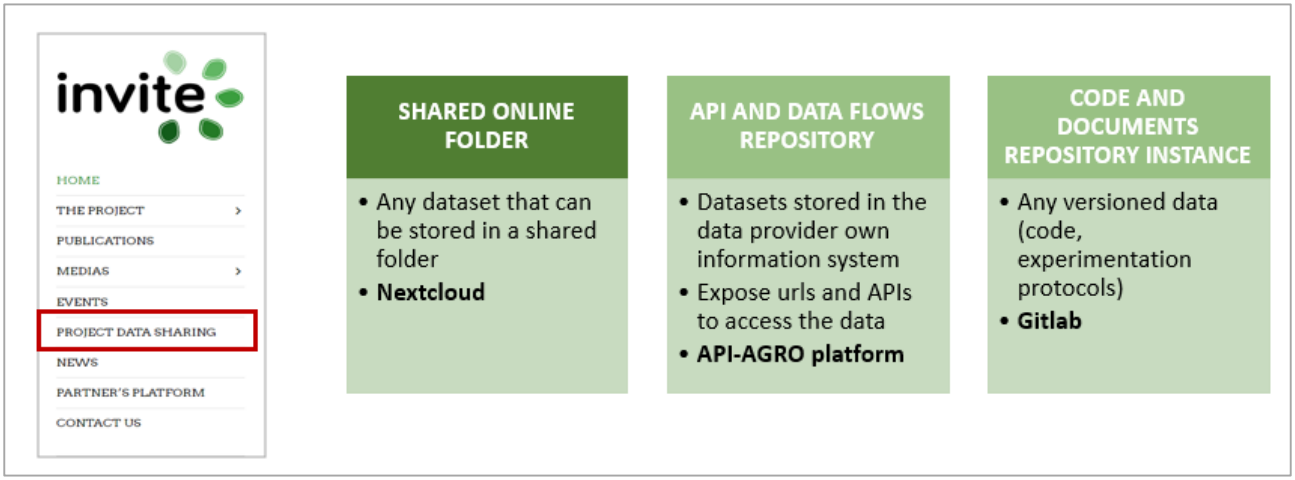
- Creating a data management plan - **Task 7.1**.
- Developing a common data sharing system composed of a mix of tools, adapted to developers, database managers and (re)users to manage data collected and generated by the project – **Task 7.2**.
- Creating an exchange group between data providers and users to define requirements for a common European database for variety testing - **Task 7.3**.
- Creating a database with unified semantic (ontologies) and associated user-friendly interface for DUS, VCU & genetic data storage with standard API that will help to build the basis for a future European database for variety testing - **Task 7.4**.

In order to ensure the success of this project, all the datasets (from historical datasets or generated in the tasks) and tools that will be used, generated and updated by the project's participants will be centralized in a single place, the **data sharing system**.
The WP participants will use and combine the datasets stored in this system in order to produce new results analysis and to validate the applicability of the developed tools.

As a reminder, the data sharing system will be a web portal (hosted in the INVITE website) including three tools:

1) A shared online folder (Nextcloud software, hosted exclusively in the EU) to store any data types and datasets.
2) An API and data flows repository (using API-AGRO platform) to expose urls and API where some datasets can be accessed. In this case, the data provider is responsible for the data storage and the data access management.
3) A code and documents repository instance (Gitlab software) hosted by Naktuinbouw for the prototype code and versioned datasets. It enables version control on codes and documents.

All three systems accessible via the web portal will be password protected. Content can only be shared within the project according to the access rights defined by the WP leaders. All WPs will input data to the listed systems by providing different datasets for the different crops. Any relevant data use limitations will be made clearly visible in the project data repository according to the access rights inventory made in Task 7.3.




In order to guide the project partners to choose the right system according to their use cases, a decision tree is provided in this document.
This decision tree will be used in the user interface of the web portal. It is essential that the decision tree remains easy to understand and simple to use, in order to provide clear and actionable information to the data providers and consumers.

# 2 Partners' use cases

We anticipate that the partners will have to main use cases regarding the data management:

- Either add new data on the data sharing system;
- Or consult and download data provided by other partners.

The criteria to categorize the data will remain the same regardless the use cases, but partners will be guided in different ways depending on the action they want to perform.

## 1.1 Add new data on the data sharing system

If a partner wants to add data on the data sharing system, he will need to define the following points:

- Will the data be stored in a shared repository or in an internal storage place?
    - Because the data storage is a sensitive topic, it is up to the data provider to decide how the data will be stored, according to their organization 's policy. They can either be stored in the data provider's information system or stored in a shared repository.
    - If the data provider wants to store the data in his information system, then he should expose urls or API in the API-AGRO platform.
    - If the data provider agrees to share the data in a shared repository, then the shared online folder (Nextcloud) and the code and documents repository instance (Gitlab) will fit the needs.

- Who can visualize the data?
  - The three tools allow to perform such action; in addition, they all include a permissions management feature for users' accounts.
- Who can edit the data?
  - The shared online folder (Nextcloud) and the code and documents repository instance (Gitlab) will allow the users that have the right permissions to edit the data.
- Who can download the data?
  - The three tools allow to perform such action; in addition, they all include a permissions management feature for users' accounts.
- What is the type of the data?
  - This point is detailed in the section 3.

## 1.2 Visualize and download the data provided by other partners

If a partner wants to visualize and/or download the data available in the data sharing system, first he needs to check the following points:

- Do I have the required permissions to visualize and/or download the data?
  - If the user has not the right permissions, he needs to contact the data provider to request them.
- What is the type of the data I am looking for?
  - This point is detailed in the section 3.

# 3 Decision criteria to categorize the data

## 3.1 Data storage

The data storage place is a very discriminant criterion to categorize the data.
If the data provider decides to store the data in his own information system, then he can either (i) choose to expose url, (ii) expose webservices or API. In both cases, the url and API will be exposed in the API-AGRO platform.
The API AGRO platform will expose all the existing url and API available, with filters based on the information provided by the data providers (data owner name, data types, metadata, conditions to visualize and download the data).

If the data can be shared on a file repository, then the users can use either Nextcloud or Gitlab to manage the data. Additional information on the data types and structuration are requested in order to assess which exact tool should be used.

## 3.2 Data type and format

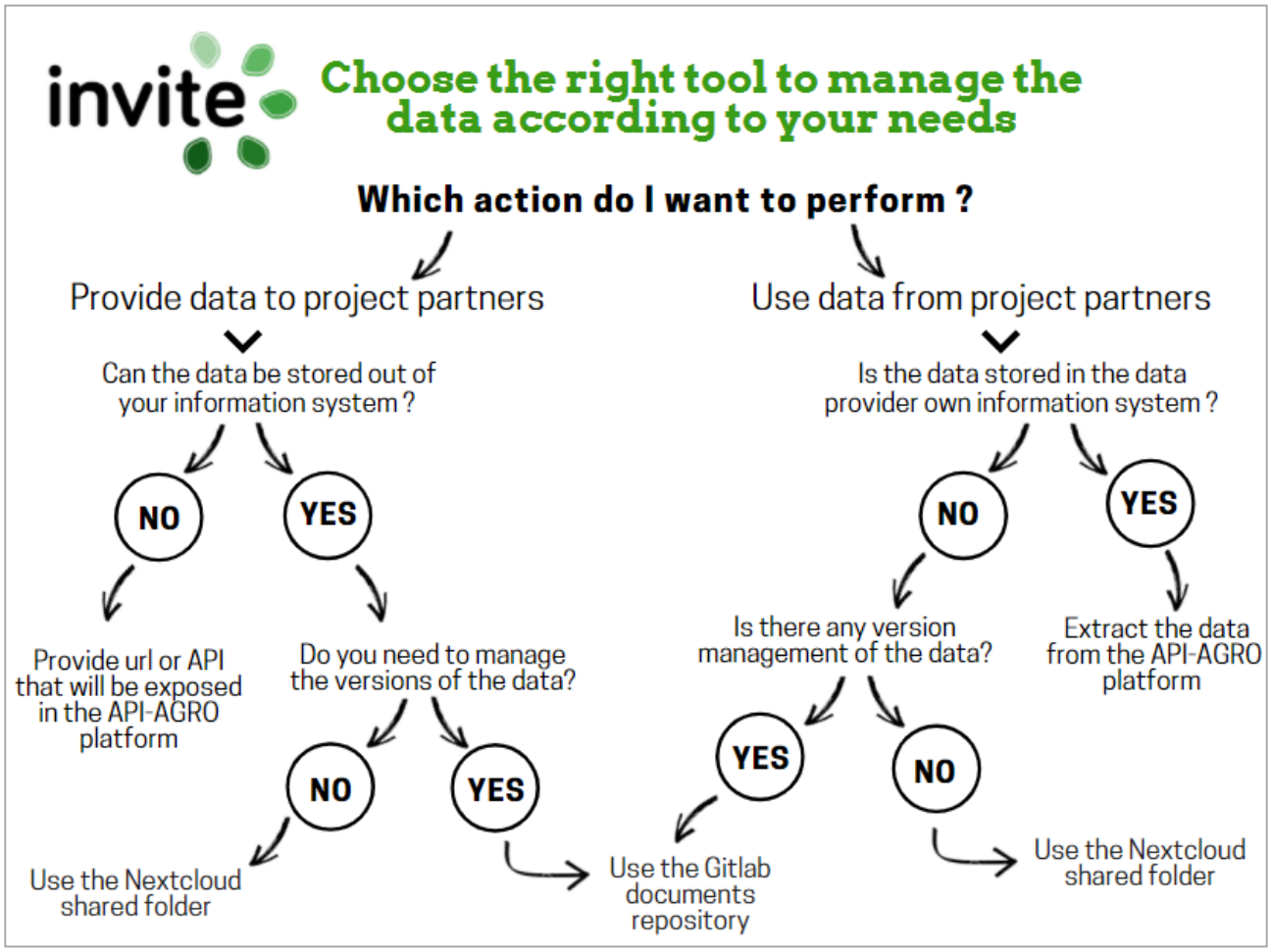
Several data types will be managed during the project. The table below summarizes the data types that are listed so far and the tools where they could fit.
This list is not exhaustive and will be updated along the project.

| Data category | Data type | Data format | Can this data be stored in… | | |
|---|---|---|---|---|---|
| | | | …the API-AGRO platform? | …the Nextcloud shared online folder? | …the Gitlab private code repository instance? |
| Phenotypic data | VCU (for a variety, location and given year) | • Raw measurement data<br>• Numerical score | Yes | Yes | Yes but not the best option |
| Phenotypic data | DUS (for a variety, location and given year) | • Raw measurement data<br>• Numerical score | Yes | Yes | Yes but not the best option |
| Phenotypic data | Morphological data (roots, aerial parts) | • Images from specific measurement tools<br>• Metrics | Yes | Yes | Yes but not the best option |
| Phenotypic data | Transpiration efficiency data | • Drones images<br>• Metrics from images | Yes | Yes | Yes but not the best option |
| Genotypic data | Molecular data Genome sequences | Genome sequencies (from bisulfite sequencing of DNA) | Yes | Yes | Yes but not the best option |
| Agronomic data | Disease score | | Yes | Yes | Yes but not the best option |
| Agronomic data | Number of spores in fields | Quantitative data | Yes | Yes | Yes but not the best option |
| Agronomic data | Grain yields | | Yes | Yes | Yes but not the best option |
| Agronomic data | Phenological stages | | Yes | Yes | Yes but not the best option |
| Environmental data | Weather data | | Yes | Yes | Yes but not the best option |
| Environmental data | Soil characteristics | | Yes | Yes | Yes but not the best option |
| Fields management data | As-applied nitrogen doses | | Yes | Yes | Yes but not the best option |

| | | | | | |
|---|---|---|---|---|---|
| Fields management data | Farming operations dates and types | | Yes | Yes | Yes but not the best option |
| Methodological data | Trials protocols | Word document, with several comments and versions. | No | No | Yes |
| Code | Code of the DB prototype | | No | No | Yes |

# 4 Decision tree

So far, the decision tree that will be displayed in the website to guide the users is the following:




Nonetheless, because the decisions on the datasets that will be used in the different work packages are not final yet, we anticipate that the list of data to manage and consequently, the decision tree, will evolve in the coming months.
An updated version will be provided both in a deliverable and in the data sharing system web portal.