This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 773139

Grant agreement N. 773139

DELIVERABLE N° 8.1

Title: Data Management Plan

Validation of diagnostic tests to support plant health

Due date:	M6
Actual submission data	M6
Start date of the project	01-05-2018
Deliverable lead contractor (organisation name)	ANSES
Participants (Partners short names)	ANSES
Author(s) in alphabetical order	Rolland M.
Contact for queries	Rolland M.
Level of dissemination	Public

Abstract:

VALITEST data will follow the "*FAIR*" principles, meaning "*Findable, Accessible, Interoperable and Re-usable*"). The data will be made findable and accessible within the Consortium, and to the broader research community, stakeholders and policy makers. Also, data will be compliant with national and European ethic-legal frameworks, such as the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679). The present data management plan (DMP) describes the data management life cycle for all data to be collected, processed and/or generated by the project. It includes information on the handling of research data both during and after the end of the project; the nature of the data, the methodology and standards applied, whether data will be shared or provided in open access, and how the data will be curated and preserved.

HISTORY OF CHANGES			
Version	Publication date	Authors	Change
1.0	30 Oct. 2018	Rolland Mathieu	Initial version

Table of content

Introduction

The VALITEST project [https://www.valitest.eu/] aims at improving the diagnostic of plant pests by i) producing validation data on the performance of the tests that are used in diagnostic, ii) harmonising further processes and iii) enlarging/triggering enlargement of the commercial offer for reliable detection and identification tests.

To achieve those objectives, a significant amount of data will be collected, processed and generated. According to the European Commission (EC), "*research data is information (particularly facts or numbers) collected to be examined and considered, and to serve as a basis for reasoning, discussion, or calculation*". In general terms, VALITEST data will follow the "*FAIR*" principles, meaning "*Findable, Accessible, Interoperable and Re-usable*". The FAIR principles will ensure sound management of data, leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse. The data will be made findable and accessible within the consortium, and to the broader research community, stakeholders and policy makers. Also, data have to be compliant with national and European ethic-legal frameworks, such as the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679), which is applicable since May 2018. This Data management plan (DMP) describes the data management life cycle for all data to be collected, processed and/or generated by the project. It includes information on the handling of research data both during and after the end of the project, the nature of the data, the methodology and standards applied, whether data will be shared or made open access, and how the data will be curated and preserved.

The DMP is intended to be a living document, and can be further modified or detailed during the project. The information can be made available on a finer level of granularity through updates as the implementation of the project progresses and when significant changes occur. Those changes might include new data or changes in consortium policies. At the very least, the DMP will be updated in the context of the periodic evaluation/assessment of the project, but the implementation of the DMP at project level will also be part of the annual reporting.

The VALITEST DMP is structured according to the H2020 templates. It includes 6 components:

- Data Summary
- FAIR data
- Allocation of resources
- Data security
- Ethical aspects
- Important topics requiring progress and/or update in future version of the DMP.

1 Data Summary

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

The VALITEST project [https://www.valitest.eu/] aims at improving the diagnostic of plant pests by i) producing validation data on the performance of the tests that are used in diagnostic, ii) harmonising further processes and iii) enlarging/triggering enlargement of the commercial offer for reliable detection and identification tests.

The project will include two rounds of test performance studies (TPS) to produce validation data, i.e. data concerning the performance of diagnostic tests while used by several laboratories on a panel of samples prepared to be as representative as possible of the potential samples.

To maximise the impact of the project, calls for interest will be organised to include in the validation programme, kits from suppliers outside the consortium and allow participation to the TPS of voluntary proficient laboratories.

For a better understanding of the demands for current and future testing options, identified stakholders will be contacted to collect their views.

Current harmonised procedures in Plant Health for validation and organisation of TPS will be improved by including appropriate statistical approaches and by adapting the process for new promising technologies.

The management of the project will also require the collection of data concerning the partners.

1.2 What types and formats of data will the project generate/collect?

Several kinds of data are foreseen to be collected or generated during the project:

- Data concerning the partners;
- Validation data;
- Data concerning the stakeholders and their views on the diagnostic market;
- Data concerning the interest of kit suppliers outside of the consortium;
- Data concerning reference material;
- Reports and publications.

Data formats will be selected with the view to facilitate data storage and transfer. Therefore, used data format will be machine-readable, but also human-readable using common software. Additionally, the management team recommends the use of non-proprietary formats, if possible.

1.3 Will you re-use any existing data and how?

The project will generate validation data but will also collect existing validation data in order to enrich an existing data-base dedicated to validation data. Existing data will not be processed during the project.

1.4 What is the origin of the data?

Data will be either generated during the project or collected in the context of surveys and of calls for interest. Validation data generated during the project will correspond to results obtained by several laboratories performing the same experiments. Data will be collected and analysed by the laboratory in charge of the test performance study.

1.5 What is the expected size of the data?

The actions planned during the VALITEST project should not require the storage and handling of big data sets. The exact data size will be evaluated by the Consortium partners during the course of the project.

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

According to the domain of expertise, data generated within the VALITEST project can be useful to:

- Scientific community;
- Industries involved plant health diagnostic;
- Inspection services, National plant protection services;
- Policy and decision makers, governmental authorities;
- International and regional organisations involved in plant health, such as EPPO, IPPC, APPPC, CAHFSA, …;
- Farmers/growers, landowners, agricultural advisors, breeders, …;
- Consumers and society.

It is the objective of the Consortium to provide most of deliverables to the widest public possible; however, restrictions in the use of data will apply especially for intellectual property or ethical reasons. When restrictions will apply, the rational for such restrictions should be provided.

2 FAIR data

Through the life cycle of the project, the FAIR principles will be followed as far as possible, while paying attention to the non-disclosure of data susceptible to compromise the quality trademark of SMEs and ensuring compliance with national and European ethic-legal framework. The FAIR component of the DMP still comprises points to clarify, which will be addressed during the course of the project.

2.1 Making data findable, including provisions for metadata

Data discoverability can be obtained by different means, which include:

- Providing data visibility through a communication system (e.g. social media, website);
- Providing online links between research data and related publications or other related data;
- Providing open access (e.g. open data repository);
- Providing data documentation in a machine-readable format;
- Using metadata standards or metadata models;
- Providing access through application;
- Providing online data visualisation/analysis tool for the data, to help researchers to explore data in order to determine its appropriateness for their purposes.

2.1.1 Discoverability of data

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)?

The number of databases to store data and documents is increasing with the expansion of the open access/open data approach. This brings the intrinsic problem that access to information is fragmented (different locations) which ironically has the counter-effect to hinder use and re-use of open data. In order to ensure visibility and accessibility of data, the Digital Research Object Portal (DROP) hosted within EPPO and maintained by Euphresco will be used to reference the open data and documents produced by the VALITEST consortium. The Portal constitutes a unique entry point to ease the retrieval of information and direct users towards the various infrastructures where the actual data (and documents) will be hosted. All the deliverables are also listed on the VALITEST website (https://www.valitest.eu/index). Once available, links will be available between the VALITEST website and the appropriate open repositories where the deliverables or datasets will be submitted. Some repositories, such as

Zenodo, provide also social media links. Scientific publication will be advertised using the website and social media, each will be identifiable and locatable by means of a Digital Object Identifier (DOI).

According to the EC, metadata is a systematic method for describing such resources and thereby improving access to them. In other words, it is data about data. Metadata provides information that makes it possible to make sense of data (e.g. documents, images, datasets), concepts (e.g. classification schemes) and real-world entities (e.g. organisations, places). Different types of metadata exist for different purposes, such as descriptive metadata (i.e. describing a resource for purposes of discovery and identification), structural metadata (i.e. providing data models and reference data) and administrative metadata (i.e. providing information to help management of a resource).

The Dublin Core Metadata Initiative is providing best practices in metadata. All the data collected or generated during the project will be described using the Dublin Core interoperable metadata standards. Furthermore, as stated in the grant agreement of the project, the metadata will always include all of the following:

- the terms “European Union (EU)” and “Horizon 2020”;
- the name of the action, acronym and grant number;
- the publication date, and length of embargo period if applicable, and
- a persistent identifier.

Concerning validation data, specific metadata have been developed by the European and Mediterranean Plant Protection Organisation Panel on Diagnostics and Quality Assurance, these will systematically be used.

Finally, some criteria will be ascertained to ensure best practice in metadata management:

- Availability: metadata need to be stored where they can be accessed and indexed so they can be found;
- Quality: metadata need to be of consistent quality, so users know that it can be trusted;
- Persistence: metadata need to be kept over time;
- Open License: metadata should be available under a public domain license to enable their reuse.

2.1.2 Naming conventions and clear versioning

Each deliverable can be identified by a unique number: D.work_package_number.deliverable_number. When applicable a versioning is used. Most documents made public will have a unique persistent identifier such as Digital Object Identifier. DOI are automatically provided by most repositories. If needed, data managers will purchase DOI numbers to reference some outputs of the project.

2.1.3 Will search keywords be provided that optimize possibilities for re-use?

To facilitate the queries by keywords, metadata of the digital objects generated during the projects must include the term “VALITEST”. Other keywords will belong to the harmonised vocabulary used in the domain[1].

2.2 Making data openly accessible

According to the H2020 online manual, open access refers to the practice of providing online access to scientific information that is free of charge to the end-user and reusable. In the context of research and innovation, 'scientific information' can mean: peer-reviewed scientific research articles (published in scholarly journals), or research data (data underlying publications, curated data and/or raw data). Open access to scientific publications means free online access for any user. The costs of open access publishing are eligible, as stated in the Grant Agreement. Open access to research data refers to the right to access and reuse digital research data under the terms and conditions set out in

[1] ISPM 5 - Glossary of phytosanitary terms; Produced by the Secretariat of International Plant Protection Convention, Adopted 2018, Published 2018; https://www.ippc.int/static/media/files/publication/en/2018/06/ISPM_05_2018_En_Glossary_2018-05-20_PostCPM13_R9GJ0UK.pdf

the Grant Agreement. Users should normally be able to access, mine, exploit, reproduce and disseminate openly accessible research data free of charge.

2.2.1 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, the data and metadata of VALITEST will be made openly available. However, three types of restrictions will apply:

- Open access is incompatible with rules on protecting personal data: protection of the personal right needs to be ascertained either by avoiding open access to sensitive and personal data, or by anonymising the data. Deliverable D9.1 (confidential) specifies the procedures implemented for personal data collection, storage, protection, retention and destruction.
- Open access is incompatible with the obligation to protect results that can reasonably be expected to be commercially or industrially exploited. The management board will identify project results with exploitation potential or commercial value or results that are costly and difficult to replicate. The management board will also assess the best options for exploiting these results (e.g. opt for disclosure and publication of results or protection through patent or other forms of IPR), consulting patent attorneys, IP specialist agents and officers at Knowledge Transfer Departments, if required. During every general assembly meeting, a special session on IP issues will be scheduled to enable open discussions and joint decisions on the best strategies for managing and exploiting the project results. Hence, protection of the interests of all the involved parties will be ensured. This will enable the exploitation strategy by all parties to be reviewed regularly and to make sure any relevant result is on route for exploitation (directly by the partners or indirectly by third parties) with appropriate terms and conditions for all project partners. Decisions relative to data management needed between meetings will be approved through electronic correspondence.
- Open access data may compromise the quality trademark of partners. Only the name of the best performing kit(s) will be made available. The performance characteristics of marketed tools providing unsatisfactory performance levels will be systematically anonymised before being made available.

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

Within the consortium, the deliverables are accessible on a restricted access platform. For the public and other stakeholders, the deliverables are listed on the VALITEST website. Some deliverables will be kept confidential, but most will be made publicly available. Public deliverables will be downloadable from the project website.

Validation data will be hosted by the European and Mediterranean Plant Protection Organisation on the Section 'validation data for diagnostic tests' of the 'EPPO Database on Diagnostic Expertise' (https://dc.eppo.int/). This database already includes validation data for diagnostic tests for regulated pests, generated by various laboratories in EPPO member countries. The validation data are presented according to a common format developed by the EPPO Panel on Diagnostics and Quality Assurance. Validation data can be submitted by any laboratory registered in the EPPO database on diagnostic expertise. At this point, this database does not comply with Open Data policy: it is not referenced and the data can only be visualised and downloaded as a PDF file. During the project, the database will be referenced and the data it contains will be made findable, accessible, interoperable and reusable.

Other data-sets will be uploaded in machine-readable format on one single OpenAire compliant repository (Zenodo or HAL, still to be determined by the management board – see more information on these repositories below) from which data can be found through a web browser and downloaded by a potential interested user.

Regarding peer-reviewed publications, the VALITEST partners will provide at the minimum a 'green' open access; they will archive the publications on an online OpenAire compliant repository and ensure open access within a maximum

of six months. The 'gold' open access should be preferred, in this case, the article is immediately provided in open access by the publisher.

2.2.3 What methods or software tools are needed to access the data?

Only standard software, e.g. web browsers, pdf-file readers, and text readers, or open licence free software, e.g. 'R' will be needed.

2.2.4 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.

Validation data will be integrated to the 'EPPO Database on Diagnostic Expertise' (https://dc.eppo.int/). During the project, the database will be referenced and the data it contains will be made findable, accessible, interoperable and reusable.

Concerning other data sets and 'green' open access scientifique articles, two options are considered at this stage: Zenodo (https://zenodo.org/) and HAL (https://hal.archives-ouvertes.fr/). Both are online OpenAire compliant repositories. The management board will have to decide which repository should be used.

2.2.5 Have you explored appropriate arrangements with the identified repository?

Arrangements already exist between the partner ANSES coordinating the project and the repository HAL (https://hal-anses.archives-ouvertes.fr/). Appropriate arrangements will be considered when repository will have been chosen by the management board.

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

As described in 2.2.1, three types of sensitive data (including personal data) are identified. These, will be managed following the GDPR requirements. Access to these sensitive data will be granted by the management board. The data processing will have to correspond to one of the uses announced to the subject during the collection of the data. Data will be transferred by electronic mail using a format including metadata. Deliverable D9.2 (confidential) provides more details on the procedures implemented for personnel data management.

2.2.7 Is there a need for a data access committee?

Data issues will systematically be discussed in the general assembly meetings. The access to sensitive data will be granted by the management board of the project.

2.3 Making data interoperable

2.3.1 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)?

All the data collected or generated during the project and made public will be available on machine readable format using common or open licence free software, described using appropriate metadata, except for the personal and sensitive data as described in deliverable 9.1 (confidential).

2.3.2 What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

For most of the data collected or generated during the project, used metadata will follow the Dublin Core interoperable metadata standards. For the specific validation data, used metadata will follow the standards

developed by the European and Mediterranean Plant Protection Organisation Panel on Diagnostics and Quality Assurance.

2.3.3 Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

To allow inter-disciplinary interoperability, all the documents and data generated during the project will use the standard vocabulary developed by the International Plant Protection Convention (IPPC) to provide a harmonised internationally agreed vocabulary associated with phytosanitary measures[2].

2.4 Increase data re-use (through clarifying licences)

2.4.1 How will the data be licensed to permit the widest re-use possible?

For public data, the reuse of the data will be possible through the open repositories where they will be stored.

2.4.2 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.

The specific decision on an embargo for research data will be taken by the management board. Scientific research articles should have an open access at the latest on publication if in an Open Access journal, or within 6 months of publication. For research data, open access should by default be provided when the associated research paper is available in open access.

2.4.3 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.

Most of the data collected or generated during the project will be available from open repositories, and therefore reusable by third parties, even after the end of the project. For ethical and legal reasons, personal or sensitive data will not be made public (deliverable 9.1 - confidential). Data concerning intellectual property will be discussed between relevant partners, and decision will be taken according to the European and national rules.

2.4.4 How long is it intended that the data remains re-usable?

Regarding data stored on an OpenAIRE compliant public repository, all files stored within the repository shall be stored after the project to meet the requirements of good scientific practice.

For data stored on other repositories, researchers, institutions, journals and data repositories have a shared responsibility to ensure long-term data preservation. Partners must commit to preserving their datasets, on their own institutional servers, for at least five years after publication. If, during that time, the repository to which the data were originally submitted disappears or experiences data loss, the partners will be required to upload the data to another repository and publish a correction or update to the original persistent identifier, if required.

2.4.5 Are data quality assurance processes described?

For the VALITEST consortium, it is essential to provide good quality data. This will be ensured through various methods. Firstly, partners have existing data quality assurance processes, which are described in their quality manual. Secondly, publications will be disseminated using peer-reviewed journals, and similarly, research data will be deposited on repositories providing curation system appropriate to the data.

[2] ISPM 5 - Glossary of phytosanitary terms; Produced by the Secretariat of International Plant Protection Convention, Adopted 2018, Published 2018; https://www.ippc.int/static/media/files/publication/en/2018/06/ISPM_05_2018_En_Glossary_2018-05-20_PostCPM13_R9GJ0UK.pdf

3 Allocation of resources

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

Costs directly associated to FAIR data management have been included within the description of the different tasks of the project.

3.2 How will these be covered?

Costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

3.3 Who will be responsible for data management in your project?

The management team (ANSES, valitest@anses.fr) will ensure best practices and FAIR principles in the data management of the project. Each partner will be responsible for managing the data it uses, processes or generates in the project. In relation with their data protection officer, each partner has appointed a data manager ensuring the respect of DMP principles and involved in personal data protection.

3.4 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)?

As an intergovernmental organisation responsible for cooperation in plant health within the Euro-Mediterranean region, EPPO will ensure the long term preservation and availability of the validation data on its own resources.

4 Data security

4.1 What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Partners will store and process personal data according to their internal procedures. The security of data will be ensured by means of appropriate technical and organisational measures.

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

Validation data will be stored on EPPO servers located in a French datacenter. The databases are located on servers independent of the web platform. Access to the servers is only possible from a secure network (dark fiber from EPPO HQ or VPN). Access to raw data is only possible from the accounts of EPPO IT Officers, through an authentication mechanism. The servers are monitored and supervised 24/7 by a service provider. The provider may under no circumstances access the data without the agreement of EPPO (the servers belong to EPPO).

Other public data will be stored on one single certified repository (HAL or Zenodo).

5 Ethical aspects

Deliverables D9.1 (personal data management) and D9.2 (ethical standards and guidelines in non EU countries) specifically deal with ethical aspects of the project. These documents include the management of personal and sensitive data and the management of data imported to, or exported from EU.

6 Important topics requiring progress and/or update in future versions of the DMP

- Decision on the repository used for long term storage of data made public
- Consider appropriate arrangements with the selected repository
- Arrange DOI assignment to all documents and data sets (even when not uploaded on a repository automatically providing the DOI)