



## Deliverable D6.12 – Activity 6.18, CNR-IPSP-FI

**FAIR access to centralized information systems and services for the Phenotyping Research Community, providing protocols on data set interoperability, advanced data-fusion and data-harmonization, including machine learning analytical approaches.**

**Authors: André P.M. Fabbri, Sabrina Mazzoni, Giulia Atzori, Mauro Centritto, Josè Saporita**

**April 30, 2025**



Deliverable number:	D6.12
Work package:	WP6 – Terrestrial Biosphere
Intermediate Objective:	IO1.1
Deliverable type:	X Document, report
	<input type="checkbox"/> Websites, patent filings, videos, etc.
	<input type="checkbox"/> Other: please specify .....
Dissemination level:	X Public
	<input type="checkbox"/> Restricted
Estimated delivery (bimester):	31/12/2014 (B13)
Actual delivery date:	30/04/2025 (B15)
Author(s) (Partner-OU):	André P.M. Fabbri, Sabrina Mazzoni, Giulia Atzori, Mauro Centritto, Josè Saporita
Reviewed by:	ITINERIS Executive Board
Note:	

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System - CUP B53C22002150006 (D.D. n. 130/2022)

Funded by EU - Next Generation EU

Mission 4 “Education and Research” - Component 2: “From research to business” -

Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

## Table of contents

<i>1</i>	<i>INTRODUCTION</i> .....	<i>5</i>
<b>1.1</b>	<b>Purpose of the document</b> .....	<b>5</b>
<b>1.2</b>	<b>Project partners</b> .....	<b>5</b>
1.2.1	EMPHASIS.....	5
1.2.2	ITINERIS HUB.....	5
1.2.3	D4Sciences.....	5
<i>2</i>	<i>REQUIREMENTS</i> .....	<i>6</i>
<b>2.1</b>	<b>Overall requirements</b> .....	<b>6</b>
<b>2.2</b>	<b>Services of the platform</b> .....	<b>6</b>
<b>2.3</b>	<b>Implementation of services</b> .....	<b>8</b>
<i>3</i>	<i>SPECIFICATIONS</i> .....	<i>9</i>
<b>3.1</b>	<b>Data Repository Service: Automated orchestration OpenSILEX</b> .....	<b>9</b>
3.1.1	Semantic data storage.....	10
3.1.2	Datafile storage .....	10
3.1.3	Tabular data storage .....	11
<b>3.2</b>	<b>GeoNetwork Metadata Export Service</b> .....	<b>11</b>
<b>3.3</b>	<b>Monitoring Service: dashboard</b> .....	<b>12</b>
<b>3.4</b>	<b>Network Manager Service: management tools</b> .....	<b>12</b>
<b>3.5</b>	<b>Assistance and training</b> .....	<b>13</b>
<b>3.6</b>	<b>Exit strategy</b> .....	<b>13</b>
<i>4</i>	<i>PROJECT PLAN</i> .....	<i>13</i>
<b>4.1</b>	<b>Activities</b> .....	<b>14</b>
4.1.1	Design Data Repository Service .....	14
4.1.2	Implementation of the orchestrator for Data Repository Service .....	15
4.1.3	Setting up other cloud services .....	15
4.1.4	Design GeoNetwork Metadata Export Service .....	15
4.1.5	Implementation orchestrator for GeoNetwork Export Service .....	15
4.1.6	Deploy of the full platform for a limited number of users .....	16

4.1.7	Consumption Monitoring.....	16
4.1.8	Deploy of the full platform .....	16
<b>4.2</b>	<b>Release.....</b>	<b>16</b>
4.2.1	Release 1 .....	16
4.2.2	Release 2 .....	17
4.2.3	Release 3 – First full platform release.....	17
4.2.4	Release 4 – Second full platform release. ....	17
4.2.5	Release 5 – Third full platform release. ....	17
4.2.6	Release 6 – Fourth full platform release. ....	17
4.2.7	End Cloud Provider support.....	17
<b>5</b>	<b>CONCLUSION.....</b>	<b>17</b>

## 1 INTRODUCTION

### 1.1 Purpose of the document

The ITINERIS project aims to strengthen the Italian Research Infrastructures (RI) working in the environmental scientific domains, including the development of PhenItaly, a dedicated RI to serve the Italian plant phenotyping community. With over six thousand active members, it is part of its European counterpart, *EMPHASIS* which already consists of several high-throughput research infrastructures. High throughput plant phenotyping generates large amounts of data with a high frequency and diversity. The Italian plant phenotyping community needs therefore a suitable computational infrastructure to handle and manage this big data on a long-term basis.

The activities 6.10 and 6.18 of the ITINERIS project propose a twofold hybrid infrastructure with local computer resources and an online cloud platform, thus ensuring scalability and continuity for its processes. The current document describes the main characteristics of the online cloud platform being developed, identified as the final deliverable of Activity 6.18. The technical and contractual requirements of the platforms and the process leading to its practical implementation are detailed in the following pages.

### 1.2 Project partners

A network of researchers, educational and research institutions, NGOs and national public bodies have contributed to the development of this Phenotyping Platform. Through their direct or indirect endorsements, their roles and specific contribution will be described below.

#### 1.2.1 *EMPHASIS*

*EMPHASIS* is a European initiative aimed at fostering collaboration within the European plant phenotyping community and stakeholders. Among other objectives, its main focus is on the *analysis of crop performance with respect to structure, function, quality, and interaction with the environment and exploitation of the crop genetic diversity required for enhancing plant productivity and plant breeding innovations.*

This cloud platform represents the first effort to create an integrated online platform for plant phenotyping at the Italian level. As such, this is to be considered a prototype for a limited usage that is by design scalable or easily replicable for other institutes working on plant phenotyping, provided the necessary resources are available.

#### 1.2.2 *ITINERIS HUB*

The ITINERIS project, is developing the *ITINERIS HUB*, which will host (directly or through virtual links) all the different RI cloud platforms it supports, whether already existing or being developed, such as ours. As such, this new platform needs to be developed following specific technical standards adopted by the ITINERIS HUB, facilitating access and integration.

#### 1.2.3 *D4Science*

*D4Science* is a not-for-profit entity, providing a range of digital infrastructure services (i.e. Virtual Research Environment), founded and managed by the *Istituto di Scienza e Tecnologia*

dell'Informazione<sup>1</sup>, itself a part of the Italian National Research Council research institutes. They have been developing Open Science Data Infrastructure services for the last 10 years, and has ongoing collaborations with most of the other ITINERIS Work Package partners (i.e. WP2)<sup>2</sup>.

## 2 REQUIREMENTS

This cloud platform offers a novel approach to this community introducing current Open Science concepts and standards. This section will first describe the overall requirements that must follow the platform throughout. Then we will describe in more details and appropriate terminology the platform's range of components, called *services*, each providing specific *features*. Not all the services detailed will be implemented or managed by our research team and in the last part we will detail the respective responsibilities of each service.

### 2.1 Overall requirements

All features and implementations for the platform need to meet the following minimum requirements:

- Guarantee security standards following Italian and European laws. In case of conflicts, the most conservative law will apply.
- Guarantee quality of service to allow partial failure of the system and provide load balancing strategy.
- Always privilege, wherever possible, the use of standards, and open source solutions.
- The actual data storage servers need to be located in Italy, whereas those dedicated to processing or replication can be in the UE.
- Provide all versions of the source code and configurations with the contractor following a General Public License to concord.
- Guarantee long-term stability of the platform facilitating its maintenance and extendibility using modular architecture.
- The cloud provider should provide a “Do No Significant harm” (DNSH) certificate to comply [Operational Guide for the DNSH principle](#), which includes in the [Annex](#) a specific checklist (Checklist 6) intended to verify DNSH requirements for IT suppliers for the provision of hosting and cloud services.
- Consider a configuration for 10 regular users that may eventually scale up to 50 users.
- Provide a design that eases the migration towards another cloud infrastructure (commercial cloud provider or on-premises) while maintaining datasets integrity.

### 2.2 Services of the platform

The cloud infrastructure needs to provide two main services for elaborating and sharing the research outputs. The first service is a FAIR data repository to harmonize and archive the experimental datasets. The second service is a collaborative platform to develop and share workflow and data processing pipelines. Other services support the integration and maintenance of the platform. We detail hereafter the services with their main requirements:

---

<sup>1</sup><https://www.isti.cnr.it/it/>

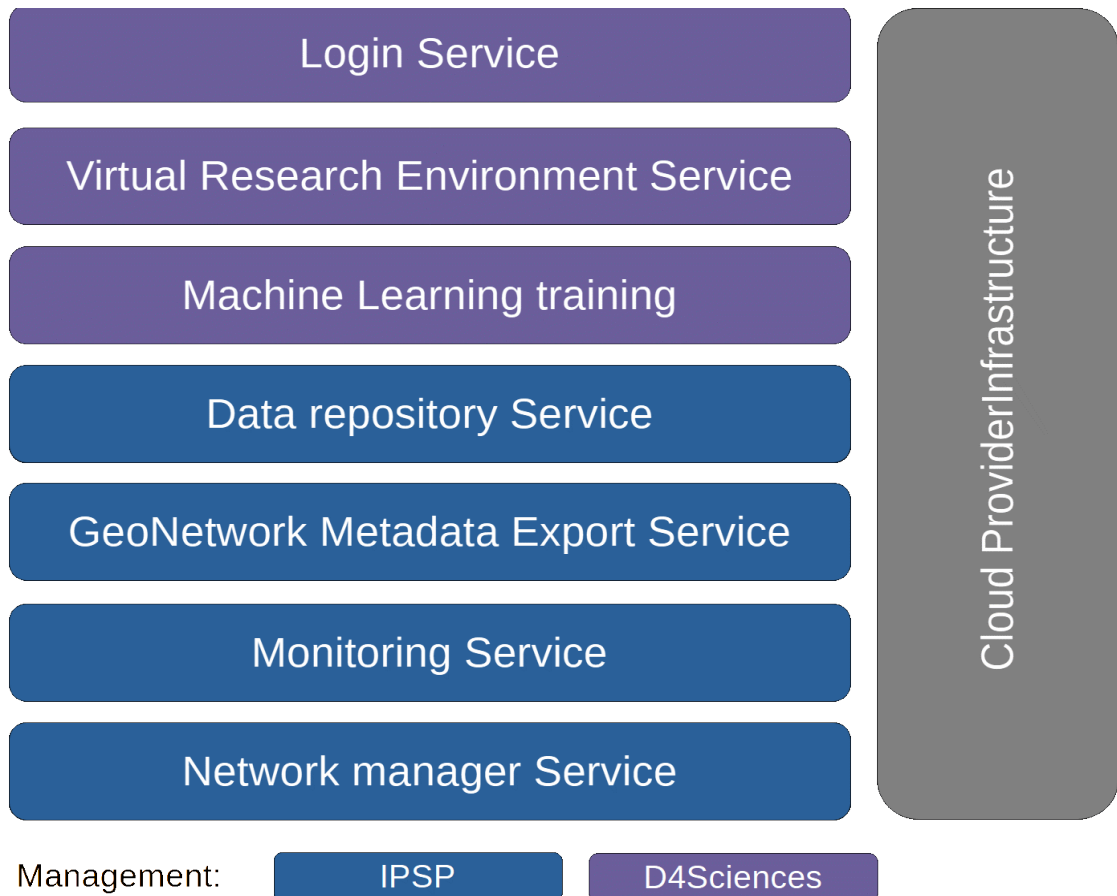
<sup>2</sup><https://itineris.d4science.org/>

- Login service: a gateway to access other services of the platform through a Single-Sign-On. This service should fulfil the common standard protocol such as OAuth 2.0 and UMA2 to interact with all services or other ID providers such as ORCID or institutional login.
- Virtual Research Environment: a workbench to process directly the data from the platform. A standard implementation of this service should provide statistics prone environment such as Jupyter notebook and the possibility to connect to the other services to enlarge the users' possibilities. This service also aims to create a community around the phenotyping data processes.
- Machine learning training: a simple environment to get access to extra computational resources for extensive training of machine learning. It should provide an API to easily upload script and launch the training with privileged access to the data.
- Data repository service: a portal to store and archive the data. This service will collect and organize all the data produced by the instruments. This services though should comply with FAIRness principles and offer API and protocol to allow automatized interactions such as harvesting or data retrieving.
- GeoNetwork Metadata Export Service: The ITINERIS HUB will harvest automatically data repository server that implements Catalogue Service for WEB (CSW) standard to integrate their metadata into its own catalog. OpenSILEX does not implement this standard, but the open source GeoNetwork <sup>3</sup>server does. This service will provide a REST API to export only the accessible metadata of OpenSILEX to an existing GeoNetwork instance or a containerized one.
- Monitoring services: a dashboard to control cloud service usage and resources consumption (storage and computational). In the context of a new platform with a set of newly purchased instruments, this service helps to adjust the services settings to meet the request.
- Network manager Service: a flexible and user-friendly control panel to manage the network settings and modify them as necessary. The platform may be administrated by staff that will not necessarily be IT experts; thus the main aspects of this management should be provided with adopting reasonably understandable and accessible tools.

---

<sup>3</sup> <https://geonetwork-opensource.org/>

Including elements of additional network specific training via the use of “hover over description or help options” should be provided, as current technology allows.



*Figure 1: Cloud platform services overview*

### 2.3 Implementation of services

A number of specific services are already being carried out by some of our partners as part of existing collaborations. Despite the extra efforts and dedication, not all the services herein describe can be implemented or managed by the current and very limited research team. As such, the following services will be contracted out, and our team will play a key supervisory role in their definition and implementation. This section will detail service by service their respective realization and management.

- Login service: This service is already provided by EOSC and implemented by D4Sciences to log in the VRE. As part of the same integrated platform the login will be delegated to EOSC service as a Single-Sign-On provider through D4Sciences.
- Virtual Research Environment: This service is exactly what D4sciences offers. This will be delegated to D4sciences.
- Machine learning training: This service is a feature of D4sciences called CCP. This will be delegated to D4sciences.
- Data repository service: The open-source FAIR data repository platform OpenSILEX meets the requirements and is already used by the plant phenotyping community across Europe. Though setup is made manually and does not scale easily. This service and the

deployment automation will be managed by our research institute, but the implementation and the setup are carried out with the help of an external subcontractor, identified by the appropriate public selection procedure.

- **GeoNetwork Metadata Export Service:** This service will interact with the Data repository service through the existing API and export only the metadata to a GeoNetwork server but remains separated from the Data repository service itself. This service will be managed by our research institute, but the implementation and the setup are carried out with the help of a contracting authority. The GeoNetwork instance could be whether the GeoNetwork catalog provided by D4sciences whether a standalone GeoNetwork containerized instance for the time of the harvest. The latter option should also be implemented by the external subcontractor as part of the setup when no existing GeoNetwork server is provided for the service.
- **Monitoring service:** This service monitors credit consumption with alerts and regular summary. This service will cover only the service directly managed by our research institute (i.e. data repository service, network manager service, monitor service) but the setup is carried out with the help of a third-part consultant.
- **Network manager service:** This service manages network settings to set up online the platform and eventually load-balancing functionalities. This service will cover only the service directly managed by our research institute (i.e. data repository service) but the setup is carried out with the help of a third-part consultant.

Despite this service being contracted out, we ask that attention is paid to “flexibility and user-friendliness” in the development of these services as noted in previous sections.

### 3 SPECIFICATIONS

The realization of a cloud platform requires additional specific expertise that is beyond the scope of our research institute. Therefore, the implementation and setting up will be assigned to a third part consultant under our supervision. As part of the process, members from the research institute will be trained to manage and eventually tweak the configured platform. This section details the elements of the projects that will be realized by the external subcontractor with respect to the service managed by our research institute and further assistance.

#### 3.1 Data Repository Service: Automated orchestration OpenSILEX

As previously mentioned, we will use OpenSILEX (<https://opensilex.org/>) as our Data repository service. This open-source software meets the requirements of the phenotyping research community (<https://github.com/OpenSILEX>) but lacks easy deployment (instances are installed manually on servers) and scalability (no load balancing strategy has been automatically implemented), as defined by the ITINERIS project. We would like to implement a container orchestration of OpenSILEX using Kubernetes to facilitate the deployment, scalability and the transferability towards other cloud infrastructure.

Minimal requirements for this part are the following:

- Provide a configurable and modular Kubernetes orchestrator to adapt strategy over the years without the need to reimplement all the modules.
- Ensure a minimal quality of service to cope with failure or overload.
- Prevent any loss of data using redundancy strategy.

The deliverables expected are the following:

- Define a testing procedure for the service (automated when possible)
- Define a scalability plan with respect to the number of active users (to limit the usage and eventually adapt the configuration)
- Define a consumption plan for a minimal number of users (to monitor the consumption and eventually adapt the configuration)
- Provide a “standalone” Kubernetes version of the orchestrator that deploys a local OpenSILEX with all the necessary sub-services using only open-source solutions. This
- Provide a “production” Kubernetes version of the orchestrator that deploys an OpenSILEX directly on the on the cloud infrastructure with the following configuration:
  - login through EGI Self-Check in portal of D4Sciences using Oauth 2.0 protocol.
  - redundancy of data storage.
  - continuity of service in case of failure using duplicate instance or automatic redeploy.
- Define a backup plan to recover locally all the data and transfer it transparently to another deployment.

OpenSILEX is a complex system that interacts with other IT components. OpenSILEX front-end stands on an Apache server and communicates with three different kinds of data storage described in this document as sub-service that are more detailed hereafter along with specific requirements.

### 3.1.1 *Semantic data storage*

Semantic data provides reference vocabulary to describe metadata of the datasets. RDF is the format used to store them efficiently. The connector of OpenSILEX to manage this kind of data storage has been realized with RDF4J (<https://rdf4j.org/>)

This data storage is rarely modified in writing: mainly during the setup of the OpenSILEX instance but some record could be added occasionally by the users. However, metadata are often read since they are always accessible for search. As an example of usage, an existing instance of OpenSILEX with 10 regular users use around 45MB of memory and is accessed between 7000 and 55000 times per day.

The minimal requirements for this subservice are the following:

- Use RDF data storage that is compatible with OpenSILEX. There are existing instances of OpenSILEX that already use the following RDF data storage: RDF4J, Virtuoso and GraphDB.
- Ensure data preservation (through redundancy or other strategy)

### 3.1.2 *Datafile storage*

Datafile are highly structured data complex that will be stored through OpenSILEX in scalable and unstructured storage spaces (data lakes). OpenSILEX provides the following datafile storage connectors:

- IRODS (<https://irods.org/>).
- Amazon S3 (<https://aws.amazon.com/it/s3/>).
- GridFS (<https://www.mongodb.com/docs/manual/core/gridfs/>).

These datafiles will be usually uploaded one time and never been modified afterwards and downloaded several times the year after the upload and then a couple of times for backup. As an example of usage, an existing instance of OpenSILEX with 10 regular users use around 900TB of

memory (considering the redundancy). In our case, we expect a minimal annual upload of 30TB for this pilot platform.

The minimal requirements for this subservice are the following:

- Use Datafile storage compatible with OpenSILEX considering eventually all the necessary setup behind to be operative. As an example, IRODS requires a dedicated data catalog server to work. Potentially, any S3 compatible storage may work but the only existing instances using this connector have been installed on an Amazon Cloud Provider. Any attempt on another Cloud Provider should consider eventually developing time to adapt it.
- Ensure data preservation (through redundancy or other strategy)
- Consider a steady upload of 30TB per year to reach 150TB after 5 years (without considering eventual redundancy)
- Rapid download times (couple of seconds) are required only for a limited time after the upload (around one year) then the download time may increase to reduce the cost of data storage.

### 3.1.3 Tabular data storage

All the data are referenced in this data storage (including the data described in previous sections). Suitable experimental data are only stored in this database where rows are observations and columns are features. More complex experimental data like images are stored in the Datafile storage and in tabular data storage a reference to its URI is written. OpenSILEX has been developed to interact with NoSQL MongoDB database.

This database is frequently modified in read and write. The volume of data is greater than the semantic data storage and they need to be correctly indexed. As an example of usage, an existing instance of OpenSILEX with 10 regular users use around 18GB of memory (including indexes) and receive between 600 and 1500 access per day (up to 8GB CACHE memory is generally necessary for this database).

Minimal requirements are the following:

- Use a MongoDB database.
- Configure the cluster with at least 2 shards
- Consider a steady accumulation of data up to reach 20GB of memory

## 3.2 GeoNetwork Metadata Export Service

This service should provide a simple REST API to export all the accessible dataset metadata of the Data repository service to a GeoNetwork instance.

On the one hand, OpenSILEX provides BrAPI<sup>4</sup> a generic API compatible with standards widely used by the Phenotypic community, through the MIAPPE<sup>5</sup> standard (Minimum Information About Plant Phenotyping Experiments). On the other hand, GeoNetwork provides its own API<sup>6</sup> the ITINERIS HUB will harvest API to store metadata following the ISO 19139<sup>7</sup> geographic format.

---

<sup>4</sup> <https://brapi.org/>

<sup>5</sup> <https://www.miappe.org/overview/>

<sup>6</sup> <https://docs.geonetwork-opensource.org/3.12/api/>

<sup>7</sup> <https://www.iso.org/standard/67253.html>

Both platforms have developed APIs (both R<sup>8</sup>, and in python<sup>9</sup>), for access between machines. The core of this service resides in the mapping between two data formats. Since harvesting won't, initially be carried out frequently, the service does not require a highly efficient response time at the moment.

Regarding the GeoNetwork instance, D4Sciences should provide a catalogue instance for metadata storage. Otherwise, a standalone procedure should be possible using a containerized version of GeoNetwork<sup>10</sup> to store locally the metadata for the time of the harvesting. For production use GeoNetwork requires 4GB RAM<sup>11</sup> and won't exceed the amount of tabular data, less than 20GB should suffice and this instance would be up daily up to 10 minutes.

An orchestrated version of the service may set up the REST export API and the standalone GeoNetwork the time of the request or execute the export and redirect following requests for the existing D4Sciences GeoNetwork server.

Minimal requirements are the following:

- Implement a REST API to export all accessible metadata from OpenSILEX to a GeoNetwork server.
- Implement an orchestrator that set up the REST API and interact with a GeoNetwork instance
- Provide a configurable script to use an existing GeoNetwork server or a standalone containerized one.

### 3.3 Monitoring Service: dashboard

This service helps the admin of the platform to monitor the consumption of resources credits and ensure that they will be enough over the year of execution. Public research institutes could not allow debit of resources and therefore the exit strategy should be implemented before depleting the credit. Another aspect is to monitor the quality-of-service of orchestrated services.

Minimal requirements are the following:

- Monitor consumption for any product provided by the cloud provider
- Consumption automatic alert at various threshold for manager of the platform
- Monthly report on credit consumption and status
- Monitor breakdown in the Data Repository service (the time when no instance is responding)
- Monitor the number of accesses to Datafile and database

### 3.4 Network Manager Service: management tools

This service should help the manager of the platform (not necessarily a network expert) to administrate the settings regarding the network in the perspective of future evolution of the platform. Regarding this platform, the network settings are meant essentially the Data repository service, part of these settings may whether be embedded in the orchestration or whether handled by

---

<sup>8</sup> <https://github.com/OpenSILEX/opensilex-r> and <https://cran.r-project.org/web/packages/geonapi/geonapi.pdf>

<sup>9</sup> <https://github.com/OpenSILEX/opensilexClientPython> and <https://docs.geonetwork-opensource.org/3.12/api/the-geonetwork-api/#connecting-to-the-api-with-python>

<sup>10</sup> As an example, [https://hub.docker.com/\\_/geonetwork/](https://hub.docker.com/_/geonetwork/)

<sup>11</sup> <https://docs.geonetwork-opensource.org/4.4/maintainer-guide/production-use/>

a dedicated provider specific product. Though the external subcontractor should provide the best compromise between cost and efficiency.

Minimal requirements are the following:

- Redirection of request (proxy, DNS, load balancing),
- SSL certificate.

### 3.5 Assistance and training

As this is a first implementation of a cloud platform for high throughput plant phenotyping, the consumption of resources should be monitored carefully, to be able to make needed adjustment to the platform settings. Better management of all resources such as type of storage, scalability and quality-of-service expected will ensure that the platform will be used sustainably, across time (up to 5 years) and increasing numbers of active users.

Gradual scheduled releases of the platform following a detailed analysis of resource use and consumption, will provide an incremental in feature, and contribute to the development of flexible and intuitive resource management settings, as described both previously and in the next section.

In case of inadequate delivery by the external subcontractor is expected to promptly intervene within a maximum of 7 days after receiving the report through certified electronic mail (PEC). Such an intervention is meant to provide immediate assistance to recover the platform services. In case this is not possible, assess the breakdown and identify the necessary fix.

As previously described, flexible and intuitive services and adequate training for should be provided to the researcher or technician tasked to “manage the platform’s services. The external subcontractor should provide at least 21 hours of training for the members of the institute managing the platform.

### 3.6 Exit strategy

Full documentation and a “hand-over” strategy needs to be developed by the contractor to ensure a continuity of the services, that is Open Source and follows the FAIR principles underlying the entire ITINERIS projects. Including, also a procedure that ensures all data and services offered by the platform being developed is stored and handed over and accessible on an actual physical hardware component or maintained *on the cloud* on a very low resource consuming and extra long-lasting storage.

## 4 PROJECT PLAN

The overall duration of the project is 54 months and includes the first full release of the platform within 6 months, followed by timely full release of the platform with adjusted settings to minimize resource consumption. The modification of settings is carried out with respect to a comprehensive monitoring report of consumption for the last period.

The realization of the platform outlined in the present document must at least go over the 8 activities mentioned in Table 1 and provide at least 3 full releases of the platform. Further details for activities and release are provided in the following sections. The timing for the phases and releases may be adjusted after consultation with the external subcontractor.

Activities/Month	M1	M2		M3		M4		M5	M6		M12		M24		M54	
Design Data Repository Service	X															
Implementation of the orchestrator for Data Repository Service	X	X		X		X										
Setting up other cloud services (Network, Monitoring)		X		X		X		X	X		X		X			
Design GeoNetwork Metadata Export Service			Release 1	X	Release 2	X	Full Release 3			Full Release 4		Full Release 5		Full Release 6		
Implementation orchestrator for GeoNetwork Metadata Export Service								X								
Deploy of the full platform for a limited number of users						X										
Deploy of the full platform in production condition								X	X							
Consumption Monitoring						X		X	X		X		X		X	
																End of Cloud Provider support

Table 1: Timeframe of the project

#### 4.1 Activities

This section details the activities performed by the external subcontractor under the regular supervision of the contractor.

##### 4.1.1 Design Data Repository Service

The purpose of this activity is the effective definition of the technical and functional specifications of the platform detailed in the present document, adapting those with solutions available on the market or alternative provided by the external subcontractor to comply with the requirements.

The following deliverables are expected at the end of this activity:

- Shared definition of the platform's specification
- Definition of the technical and functional specifications for the orchestrator of the Data Repository Service including the tests that verify them (End-to-End, Performance, Resilience).
- Definition of a standard configuration of orchestrator that meets the minimum requirements and their scalability limit.
- Estimation of the consumption of the platform related to reference indicators (accesses, volumes, calculation time).
- Definition of the functional specifications for any supplementary developments to be carried out.

#### *4.1.2 Implementation of the orchestrator for Data Repository Service*

The purpose of this activity is the complete implementation, testing, and validation of an orchestrator (Kubernetes) for the Data Repository Service (OpenSILEX) according to the specifications and methods defined in the previous activity.

The following deliverables are expected at the end of this activity:

- Project document detailing the choices and characteristics of the implemented solution in terms of quality of service.
- Automated tests to verify detailed functional specifications.
- Source code produced and released under the same license as the OpenSILEX project (AGPL) along with appropriate documentation.
- Configuration of the orchestrator to run a local standalone version of Data Repository Service under AGPL license.
- Configuration alternative of the orchestrator to run a cloud version of Data Repository Service integrate with other service (Network, Login etc) under another GPL license to concord.

#### *4.1.3 Setting up other cloud services*

The purpose of this activity is setting up the connection with other cloud services (Network, Monitoring, Login).

The following deliverables are expected at the end of this activity:

- Configuration of the service as files or inserted in a report.
- Training operators on their use.

#### *4.1.4 Design GeoNetwork Metadata Export Service*

The purpose of this activity is the effective definition of the technical and functional specifications of the platform detailed in the present document, adapting those with solutions available on the market or alternative provided by the external subcontractor to comply with the requirements.

The following deliverables are expected at the end of this activity:

- Shared definition of the GeoNetwork Metadata Export Service specification
- Definition of the technical and functional specifications for the orchestrator of the GeoNetwork Metadata Export Service including the End-to-End tests that verify them.
- Definition of two alternative configurations for the orchestrator to write metadata on an existing GeoNetwork or to set up one using the same pipeline.
- Definition of the functional specifications for any supplementary developments to be carried out.

#### *4.1.5 Implementation orchestrator for GeoNetwork Export Service*

The purpose of this activity is the complete implementation, testing, and validation of an orchestrator (Kubernetes) for the GeoNetwork Export Metadata Service according to the specifications and methods defined in the previous activity.

The following deliverables are expected at the end of this activity:

- Automated tests to verify detailed functional specifications.
- Source code produced and released under a GPL license to concord along with appropriate documentation.
- Configuration of the orchestrator to write Metadata on an existing GeoNetwork instance.
- Configuration alternative of the orchestrator to write metadata on a GeoNetwork instance set up on the fly and linked to the Network service.

#### *4.1.6 Deploy of the full platform for a limited number of users*

The purpose of this activity is to test the execution of the full platform on a selected number of users and identify any failure in their execution.

The following deliverables are expected at the end of this activity:

- Report on the eventual bug and fixes provided.
- Detailed report on resources consumption.

#### *4.1.7 Consumption Monitoring*

The purpose of this activity is to provide tools for effective monitoring of consumption and anticipate any complete depletion of the credits.

The following deliverables are expected at the end of this activity:

- Set up alert mechanism on consumption threshold (absolute and relative to last week's usage).
- Automatic monthly report.
- Detailed consumption report before each release suggesting settings adjustment to implement.

#### *4.1.8 Deploy of the full platform*

The purpose of this activity is to test the execution of the full platform for a longer period and with an extended number of users. During this activity, the GeoNetwork Export Metadata Service should be completed and integrated into the overall platform.

The following deliverables are expected at the end of this activity:

- Report on the eventual bug and fixes provided.
- Detailed report on resources consumption.
- Update of orchestrator and documentation.

## **4.2 Release**

This section details the release as timely meeting between the contractor and external subcontractor.

### *4.2.1 Release 1*

At the first release, a standalone version of the Data Repository Service is expected. This version will be tested on a local cluster hosted by our research team and identify any immediate failure.

#### 4.2.2 Release 2

At the second release, a cloud version of the Data Repository Service is expected. This version should be integrated with at least the Network Service and the Login Service.

#### 4.2.3 Release 3 – First full platform release.

At the third release, the first full version of the platform is expected including Data repository Service, GeoNetwork Metadata Export Service, Network Service, Monitoring Service and integration with partner services (Login Service, VRE Service, Machine Learning).

#### 4.2.4 Release 4 – Second full platform release.

At the fourth release, the second full version of the platform is expected including Data repository Service, GeoNetwork Metadata Export Service, Network Service, Monitoring Service and integration with partner services (Login Service, VRE Service, Machine Learning). This version will integrate correction considering the issues identified after a 2-month execution.

#### 4.2.5 Release 5 – Third full platform release.

At the fourth release, the third full version of the platform is expected including Data repository Service, GeoNetwork Metadata Export Service, Network Service, Monitoring Service and integration with partner services (Login Service, VRE Service, Machine Learning). This version will integrate correction considering the issues identified after an 8-month execution.

#### 4.2.6 Release 6 – Fourth full platform release.

At the fourth release, the fourth full version of the platform is expected including Data repository Service, GeoNetwork Metadata Export Service, Network Service, Monitoring Service and integration with partner services (Login Service, VRE Service, Machine Learning). This version will integrate correction considering the issues identified after a 20-month execution. This last version should be configured for a steady use of the platform in the next 3 years.

#### 4.2.7 End Cloud Provider support

This last meeting will state how to proceed with the cloud platform and eventually apply the exit strategy, if not applied before.

## 5 CONCLUSION

The last goal of the Activity 6.18 of the ITINERIS project is the deployment of the facilities and services needed for the Italian Plant Phenotyping community as part of the European research infrastructure EMPHASIS. To this end, the construction of a hybrid computational infrastructure of the ITINERIS project could be considered as “pilot” to foster the digitization, harmonization and fusion of experimental datasets. Due to administrative constraints, we expect to have a first release for the 15 bimester and the first full release before the end of the project. Though the specifications are detailed in the present deliverable.

Across the world, other plant phenotyping institutes have created online platforms to support their research. They generally choose implementations and architecture that are not publicized through scientific literature. As part of the ITINERIS’s project this platform will follow Open Science

standards, contributing to the development of Findable, Accesible, Interoperable and Reproducible Research Infrastructures, thereby ensuring also quality of services.

Taking advantage of our partners' experience, including bringing in external subcontractors with solid expertise in developing Open Science RI, we expect to obtain a set up tailored for our purposes, limit the duplication of services and offer maximum visibility of this platform.