D6.1 State of the Art and Community Needs Report from Use Cases

Lead Partner:	**CNRS**
Version:	**V1.1**
Status:	**Final**
Dissemination Level:	**PU**
Document Link:	10.5281/zenodo.3894676

Deliverable Abstract
The main purpose of this document is twofold: (i) describe the Use Cases and, (ii) analyze their technical requirements in compliance with EOSC framework.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 857650.

COPYRIGHT NOTICE

The EOSC-Pillar project is co-funded by the European Union Horizon 2020 programme under grant number 857650.

DELIVERY SLIP

	Name	**Partner/Activity**	**Date**
From:			
Reviewed by:			
Approved by:			

DOCUMENT LOG

Issue	**Date**	**Comment**	**Author**
V0.1	31/03/2020	Initial version	Alessandro Rizzo
V0.2	20/04/2020	First round of comments from internal reviewer	Federica Tanlongo
V0.3	14/05/2020	Second round of comments from internal reviewer	Fulvio Galeazzi
V1.0	11/06/2020	Final version	Alessandro Rizzo, Christelle Pierkot
V1.1	09/02/2023	Added Zenodo URL	Fulvio Galeazzi

TERMINOLOGY

https://eosc-portal.eu/glossary

Terminology/Acronym	Definition
API	Application Programming Interface
CINES	Centre Informatique National de l'Enseignement Supérieur
CMCC	Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici
CMIP6	Coupled Model Intercomparison Project (Phase 6)
CND3D	Conservatoire National des Données 3D
CNR	Consiglio Nazionale delle Ricerche
CNRS	Centre National de la Recherche Scientifique
DKRZ	Deutsches Klimarechenzentrum
ECAS	ENES Climate Analytics Service
EOSC	European Open Science Cloud
EOSC-Pillar	Coordination and Harmonisation of National and Thematic Initiatives to Support EOSC project
ESGF	Earth System Grid Federation
eTDR	European Trusted Digital Repository
FAIR data	Findable, Accessible, Interoperable, Reusable data
HAL	Hyper Articles en Ligne (Open Science repository)
HPC	High Performance Computing
INFN	Istituto Nazionale di Fisica Nucleare
INRAE	Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement
INRIA	Institut National de Recherche en Sciences et Technologie du Numérique
INSERM	Institut National de la Santé et de la Recherche Médicale
LOD	Link Open Data
NLP	Natural Language Processing
POC	Proof of Concept
RI	Research Infrastructure
SSH	Social Sciences and Humanities
UC	Use Case
URL	Uniform Resource Locator
VAP	Virtual Analysis Platform
VRE	Virtual Research Environment
WP	Work Package
WG	Working Group

Contents

Executive summary

This deliverable reports on the findings of the Use Cases analysis, with a special focus on the community requirements and on the opportunity to extend every single Use Case to other communities. The document contains a report on a first state of the art and community needs from each Use Case. The present document is the result of several months activities driven by the WP6 and centred on the analysis of the Use-cases submitted by the scientific communities involved in the EOSC-Pillar project. The main aim of the deliverable is to present “the state of the art” through a cross-cutting analysis in order to identify commonalities and synergies among the UCs with regards to technical requirements and needs. Finally, some options are suggested in terms of common actions and developments.

1 Introduction

It looks evident that each scientific community is inclined to develop tools and services adapted to its own requirements, that are often different from one community to another. This document aims to take stock of the current situation in terms of services already developed and implemented as well as existing infrastructures in order to identify their technical requirements in accordance with the EOSC framework.

Data from each Use Case have been gathered and analyzed in order to identify the gap between existing services and requirements to define future cross-cutting services and then to avoid duplication. Starting from the needs put forward by each Use Case, the main purpose here is to bring out technical requirements for future actions in coordination with WP5 and WP7 into the framework of the EOSC-Pillar project.

The UC templates, submitted by each Task, have provided a deep description of the UCs in all their dimensions but not in terms of possible interactions and common developments with other UCs. In the perspective of the bottom-up approach, on which the project has been built, it was essential that these ties could be identified and then links with the WP5 and WP7 strengthened. Then a cross-cutting analysis has been conducted during the last months with the explicit objective, when it was possible, to cluster UCs with regards to technical requirements in terms of data and/or infrastructures.

In the document we are going, firstly, to introduce the methodology adopted for the analysis, secondly, to present each UC and the major results of the analysis and, finally, to suggest actions that could be implemented to achieve a more effective coordination among all the UCs or some of them.

2 Methodology

The present state of the art is a preliminary stock. It should make possible to draw up a general overview of current knowledge and practices in each scientific community previously selected. Identifying questions, results and / or elements that are missing are all points that we wanted to address at this stage. The state of the art within the framework of the EOSC-Pillar project aims to deepen and clarify not only the knowledge but also the needs and requirements in terms of services and infrastructures in view of the structuring of EOSC. The results of this first activity will support the implementation of the project and its progress.

Firstly, it was essential to define a main framework for the state of the art as well as the indicators which would make it possible to consolidate it. For this purpose, we have developed a **specific template common to all the Use Cases**. This template therefore represents the enabling analysis matrix to synthesize inputs from Use Cases around two major categories: **(i) description**, and **(ii) perspectives**.

The achievement of the state of the art was structured on the contributions and analysis from each Use Case throughout this first phase of the project. To go forward with the templates analysis we decided to set up a working group composed by 6 project members involved in different WP and tasks.

However, it was essential that the state of the art could be fixed at a specific time. For this, we have decided to stop collecting data and gathering information in February 2020 to analyze and synthesize them. According to the recommendations of the “virtual F2F technical meeting” held in March it has been decided to go ahead with the analysis of the Use Cases to illuminate dynamics of interactions among the Use Cases or some of them.

The methodology adopted for the **cross-cutting analysis** consisted of the undertaking of the fifteen Use-cases submitted by the scientific communities participating at the project. Two main analysis criteria were deemed particularly relevant: **(i) collaboration and common developments with other UCs**, and **(ii) existing and possible links with WP5 (Data layer) and/or WP7 (Infrastructure layer)**.

In a first stage we have sent a survey to the Task Leaders in order to gather their perceptions and opinions concerning cross-cutting tools and possible collaborations with other scientific communities (https://repository.eosc-pillar.eu/index.php/f/27539). The survey had an overall response rate of 60%.

In a second stage we have scheduled and driven a round of meetings with each Task in order to deepen the UC analysis previously based only on the UC templates. These meetings have been held from April, the 9th, to April, the 24th, 2020.

In a third stage we have summarized feedbacks and inputs coming from the Task Leaders and partners. The main objective was to identify common demands for services and cross-disciplinary technical developments in the framework of EOSC. Finally, in this perspective some actions have been proposed and discussed with the Task Leaders.

2.1 The Template

The template used to collect data and information concerning the Use Cases is organized around two main axes, namely, (i) the description of the current implementation status and (ii) the identification of the main perspectives and therefore the definition of needs and requirements for their future development and implementation.

i. The template focuses on a detailed description of the Use Cases with a particular attention to **(i) users and stakeholders' profiles**, **(ii) current deployment status** both nationally and internationally, and **(iii) services involved in the current implementation** and other standard services already available.

ii. The template's objective was to inscribe the Use Cases perspectives into the framework of the structuring of EOSC and therefore in the EOSC-Pillar project in order to highlight the added value expected by the scientific communities. The identification of needs (services, infrastructure, resources, capacities) and technical requirements were considered fundamental for implementing and monitoring future actions.

3 The State of the Art

3.1 Use Cases description and analysis

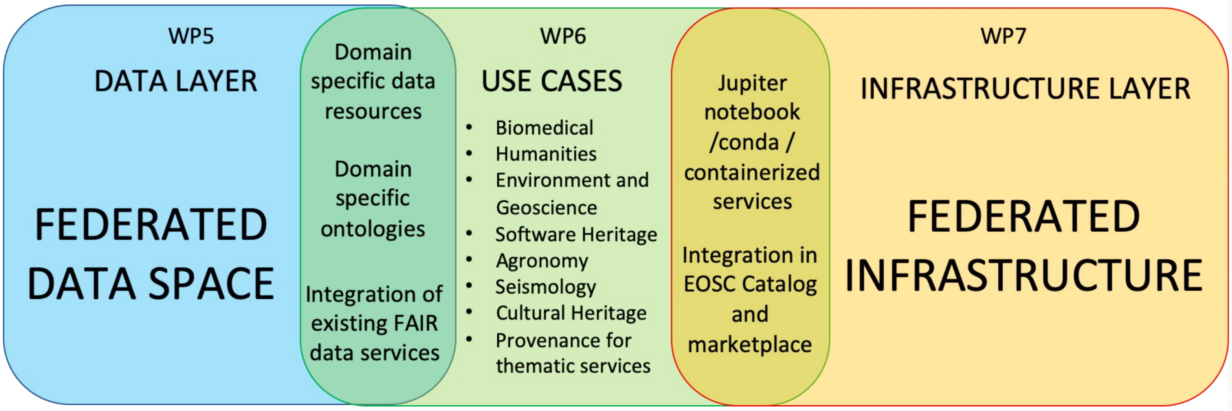
The WP6 aims at collecting Use Cases based on the requirements of scientific communities that are normally connected to real scientific production and intended to improve the usage of data in a FAIR perspective in view to i) gain experience and make UCs as templates for further activities, and ii) validate services (cross-community / cross-cutting) in a production environment.

Here below a brief description of the Use Cases as they have been defined by each scientific community as well as the main results of the UCs analysis.

Use Cases have been analysed by using the following characteristics:

1. **Data services for management:** This section highlights the set of services that need to be integrated in order to support the implementation of the Use Case. Services in this class include both existing services (such as community repositories) as well as forthcoming services (i.e. services to be developed and provisioned by the EOSC-Pillar consortium).
2. **Datasets type:** This property defines the domain of each Use Case, their datasets type or format if this information is available in the description of the Use Case. Moreover, their metadata information has been added as well, if available.
3. **Data services for storage:** This property emphasize the need to store data to a specific data storage, or the need for specific data services. This is also relevant when considering a possible re-instantiation of the service in a location from the self-contained data repository of the data provider.
4. **Need of Use Case in terms of compute power:** This requirement emphasizes the need of the Use Case in using computing resources for short/long periods of time to accomplish computational tasks, such as HPC/HTC systems offered already as ready-to-use services in WP7.
5. **Current scale:** This property defines the ability to scale the service up (to handle growth i.e., extend the service from national to international level) or down (e.g., minimum size, possibly meet needs of smaller research groups).
6. **FAIRness of data:** This requirement consists of a set of guiding principles in order to make data Findable, Accessible, Interoperable and Reusable. Making data FAIR is the responsibility of the Research Infrastructures and their data repositories.
7. **Current practices**: This property in the context of the Uses Case implementation identifies ongoing uses and constraints, if available.
8. **Types of stakeholders:** This property lists which users (data providers or consumers) have access to the data in their data repositories categorized by different profiles. These users might be humans but there are of course cases when the data is being accessed and manipulated by other services.
9. **Auxiliary applications:** This property consists of a set of other applications or dependent services that the service should consider to run in development/or production environment in compliance with EOSC framework.

10. **Objectives of the Use Case:** This section presents the objectives that the scientific community plans to achieve throughout the implementation of the Use Case within the framework of the EOSC-Pillar Consortium.
11. **Gap Analysis/Missing pieces:** This section focuses on the missing elements to reach the objectives identified starting from the current status.
12. **Cross-cutting analysis:** This section presents possible extensions to other communities by defining links and collaborations with other Use Cases.
13. **Services needed by WP5 and WP7:** This section highlights links with WP5 and WP7, using the schema here-below.


WP5
DATA LAYER
FEDERATED DATA SPACE
Domain specific data resources
Domain specific ontologies
Integration of existing FAIR data services
WP6
USE CASES
Biomedical
Humanities
Environment and Geoscience
Software Heritage
Agronomy
Seismology
Cultural Heritage
Provenance for thematic services
Jupiter notebook /conda / containerized services
Integration in EOSC Catalog and marketplace
WP7
INFRASTRUCTURE LAYER
FEDERATED INFRASTRUCTURE

Evidently, the cross-cutting analysis does not avoid any other development needed to implement the UC or any other cross-UCs cooperation. Here the aim is that disciplinary UCs make progresses through a structured interaction with other communities facing similar difficulties or questions and addressing similar technical issues.

3.1.1 T6.1: "Defining procedures/service to enforce data provenance for thematic communities and beyond"

Provenance management is a key component in order to guarantee scientific data discovery, reproducibility and results interpretation. Provenance management should be able to define a set of metadata able to capture the derivation history of any stored data, including the original raw data sources, the intermediate data products and the steps that were applied to produce them.

Use-Cases:

1. **The material science Use Case.** This UC is connected with NFFA-EUROPE and EUSM EU project. A data repository has been made available and several metadata associated with scientific data collected in a semi-automated way. Some data services are also currently developed and, in some cases, specific data-workflows have been defined. Data provenance associated metadata should be identified and automatically collected in such workflows.

2. **The climate science Use Case.** This UC relates to several classes of data analytics workflows (e.g. the computation of climate indicators). Data are preliminarily set up and made available on a dedicated data pool directly connected with the compute infrastructure. The user develops the analysis application and runs it on top of a compute facility. Provenance information needs to be captured during the application execution and properly stored for further (re-)use by multiple scientists.

User scenario (i):

End users run data analytics workflows on a compute facility and produce data outputs (e.g. final products, intermediate results) which are published and shared within the community. End users search for some data and retrieve the provenance information to have a better understanding of its data lineage as well as of the entire analytics process associated to it.

UC#1 Analysis

Data services for management:

- Material use case : NFFA Datashare : cloud platform for data management
- Climate use case : ECAS[1] (EOSC Hub thematic service) enables scientific end-users to perform data analysis experiments on large volumes of research data from multiple disciplines. ECAS is based on the Ophidia[2] big data framework which provides an array-based storage model and a hierarchical storage organisation to partition the data and distribute the workload across multiple nodes.

Datasets type : Climate datasets in NetCDF-CF[3] format from the CMIP6[4] project available through the Earth System Grid Federation[5] (ESGF).

Data services for storage: *Not specified*

Need of UC in terms of compute power: No.

Current scale: national & International (ESGF, CMIP6, EOSC-hub).

FAIRness of data: Need to enhance FAIR approach for material science by means of extension of metadata schema and precise data acquisition workflow. For climate data, there is a need to develop FAIR-oriented (second level) provenance support for data analysis, which is not available yet.

Current practices: *Not specified*

Types of stakeholders

- Material Scientist: using distributed infrastructure, collect and analyse data
- Climate Scientist: performing scientific data analysis, run and reproduce analytics workflows (i.e. climate modelling groups, impact researchers)

Auxiliary applications: *Not specified*

[1] ECAS: ENES Climate Analytics Service - https://marketplace.eosc-portal.eu/services/enes-climate-analytics-service

[2] Ophidia big data analytics framework - https://ophidia.cmcc.it/

[3] Climate and Forecast convention: http://cfconventions.org/

[4] CMIP6 experiment - https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6

[5] Earth System Grid Federation - https://esgf.llnl.gov/

Objectives of the Use Cases:

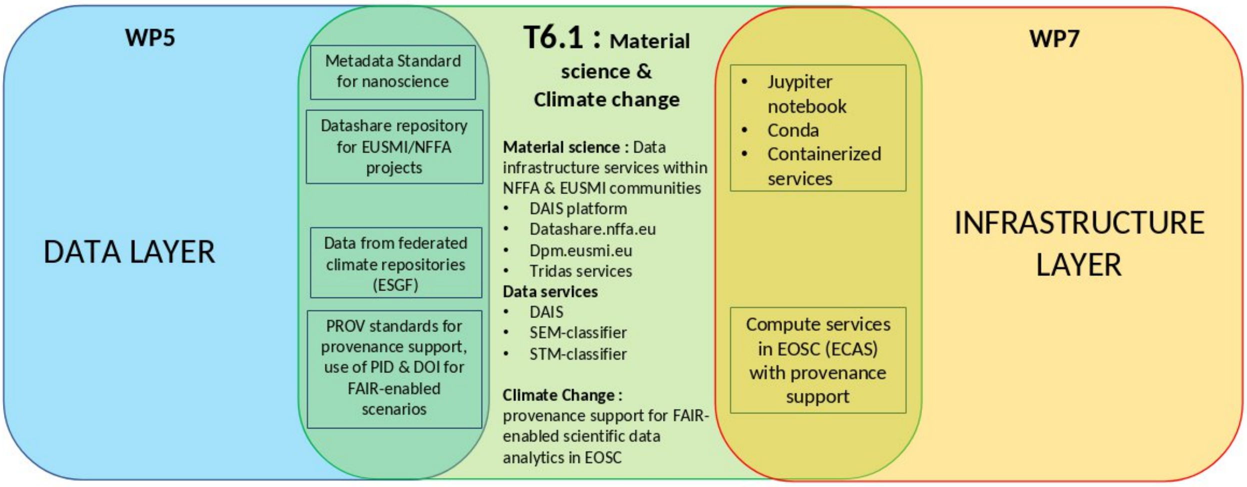
- Elaborate cross-domain, FAIR-oriented procedures and recommendations to enforce data provenance.
- Provide two different workflows (demonstrators) for data provenance on Material Science (CNR/IOM) and Climate Data (CMCC&DKRZ)

Gap Analysis/Missing pieces: *Not specified*

Cross-cutting analysis

- UC 6.2 about data provenance and ISO Standards
- UC 6.9 for data provenance
- UC 6.7 and UC 6.8

Services needed by WP5 and WP7


WP5
DATA LAYER
Metadata Standard for nanoscience
Datashare repository for EUSMI/NFFA projects
Data from federated climate repositories (ESGF)
PROV standards for provenance support, use of PID & DOI for FAIR-enabled scenarios
T6.1 : Material science & Climate change
Material science : Data infrastructure services within NFFA & EUSMI communities
• DAIS platform
• Datashare.nffa.eu
• Dpm.eusmi.eu
• Tridas services
Data services
• DAIS
• SEM-classifier
• STM-classifier
Climate Change : provenance support for FAIR-enabled scientific data analytics in EOSC
• Juypiter notebook
• Conda
• Containerized services
Compute services in EOSC (ECAS) with provenance support
WP7
INFRASTRUCTURE LAYER

3.1.2 T6.2: “Agile FAIR data for environment and earth system communities (ocean, atmosphere, continental surfaces, solid earth)”

Earth Observation involves several domains. In addition, data from in-situ observations and remote sensing data (satellites) are often managed and preserved separately. It makes inter-calibrations, inter-comparisons and integrated uses quite difficult. This **Use Case** aims to be general, to be able to address several categories of users related to **Environment & Geosciences**. The final objective of the EOSC-Pillar Task 6.2 is to set up **Virtual Data Analysis Platform** (VAP) that proposes on-demand data visualization, browsing and processing services to users from environment and earth system communities.

As a consequence, this use case will be divided in three sub-use cases, in order of priorities:

1. **Data Science Notebooks**, that will offer a web based processing environment for scientists and data analysts,
2. **Data services**, that will speed up data access and to facilitate access to data from several domains,
3. **Discovery services**, by the implementation of a cross domain catalog.

However, setting up such a comprehensive data system is really ambitious and the present EOSC-Pillar “Environment & Geo-sciences” UC will focus on some services that require improvements towards:

- A better integration of data from several domains and managed in a distributed way to facilitate cross-domain studies;
- The set-up of on demand data browsing and processing platforms that allows users to develop their own scripts for data analysis.

In order to test the services that will be set up, the following subjects will be specifically addressed as proof of concept:

- Co-localization of satellite data and in-situ data for data intercalibration in the fields of satellite salinity (SMOS satellite) and in-situ salinity at sea surface (Argo Floats and SeaDataNet Sea Cruise observations) and analysis of correlations;
- Co-localization of Ocean Color (Optical imagery from e.g. Sentinel 2) and Chlorophyll in-situ observations (Argo floats and SeaDataNet Sea Cruise observations) and characterization of the extension of blooms.

User scenario (i):

Sub-use case Data Science Notebooks:
This Use Case will be divided in two parts, by order of priority:

- Step 1: **Virtual Research Environments (VRE’s)** will be made available to users, providing data analysis tools, which rely on existing Python scripts. These Python scripts have been developed and will be updated by EOSC-Pillar Task 6.2 participants. Users will run the scripts in a predefined environment for their own purpose, typically to perform analyses on their areas or their time slots of interests. Since these scripts present a fixed user-interface, integration in VRE’s such as D4Science may be considered.

- Step 2: Platforms to develop, upgrade, test on real data sets (that are potentially large), new algorithms will be set up (**Virtual Data Analysis Platform**). These platforms will make available the Pangeo software suite and will have access to the datasets provided by the use case. These platforms will allow collaborative work on the same Python scripts.

The targeted users are different for the two steps: users of preexisting scripts in the first case, software developers and data analysts for the second steps. The script developed for the second step could then be made available for the users of the first step.

Available or developed scripts will be managed using the same repository (Git and/or possibly Software Heritage) with two different access rights: read-only for general users, read/write mode for software developers.

Both of the two parts require user’s authentication and authorization to have access to the computer resources (computing, memory and storage). For the first step, a VRE mechanism such as D4Science may be used for that purpose.

However, the step 2 induces specific IT issues such as security and limitation of required IT facilities (quotas for computing, memory, and storage) since scripts under development can be erroneous or much more consuming than expected. These points have to be carefully considered with people in charge of IT infrastructures and that explains why it has been relegated as a second priority.

Sub-use case Data services:

The input datasets used by the demonstration scripts (cf. sub use case 1) are presently distributed on several repositories: e.g. for script 1, the in-situ observations (Cora, Argo) are not stored in the same repository as satellite SMOS datasets. In addition, all these datasets are stored using NetCDF format which is not fully optimized for parallel computing.

This sub use case will propose to :

1. transfer the data sets (or necessary subset of them) on an e-infrastructure able to run the demo scripts. This transfer could be implemented using micro-services relying on cloud storage technologies (such as iRODS).
2. convert the datasets format to Pangeo compatible format such as Xarray/zarr. That requires to adapt scripts that are available to convert NetCDF files to zarr format and to define the implementation of zarr on the targeted e-infrastructure (size of the chunks...).

In addition, in order to facilitate the adaptation of the demo scripts, the provision of a uniform Python API will be considered, using a common technical catalog such as the Intake catalog derived from the discovery catalog (please refer to sub use-case 6.2.3), which harmonizes access to data for Python developers.

Sub-use case Discovery services:

This sub use case will set up a common cross-domain discovery catalog, relying on a technology such as STAC, and populated by harvesting several pre-existing catalogs, such as ISO 19115 (environmental metadata standard), OAI/PMH or STAC.

Harvesting will be performed using protocols such as Catalog Service for the Web (CSW) or OAI/PMH.

An Intake technical catalog will be derived from this common discovery catalog.

That requires :

- Writing scripts to populate the cross-domain STAC catalog from the existing ISO19115 or OAI-PMH catalogs.
- Writing a script to generate the Intake technical catalog from the STAC catalog.

UC#2 Analysis

Data services for management:

Existing services are:

- SeaDataNet (https://www.seadatanet.org/)
- SMOS satellite (https://www.catds.fr/Products/Available-products-from-CPDC)
- Copernicus/CMEMS (http://marine.copernicus.eu/)
- EuroArgo (http://www.argodatamgt.org/Access-to-data)
- Sentinel2 Imagery (https://sentinel.esa.int/web/sentinel/sentinel-data-access)
- WDCC World data center of climate (https://delos.dkrz.de/WDCC/ui/cerasearch/)

- ECASLab (https://ecaslab.cmcc.it/web/home.html)

Foreseen service : a Virtual Research Environment to access and browse those data.

Datasets type: Domain specific datasets (ocean, earth, atmosphere). Climate datasets are stored in NetCDF format.

Data services for storage:

- Same as the services for data management.
- Foreseen service : “Data Lake” relying on Data cubes (such as Xarray) and No-SQL databases

Need of UC in terms of compute power:

- A few CPUs for running notebooks (~20 in the test phase).
- 4 GPUs for Pangeo machine learning methods.
- Virtual machines for dedicated web services.

Current scale: National/European (for Individual services)

FAIRness of data: FAIR data and services, standardized metadata used

Current practices: Data management and valorisation is domain-dependent, and there is no cross-over.

Types of stakeholders:

- Scientists : Python script developers, data analysts, data managers
- Data managers : manage, process, access data
- Data users : Environmental cross-domain data users or using data from different sources

Auxiliary applications: *Not specified*

Objectives of UC: Provide a **Virtual Data Analysis Platform** (VAP) that proposes on-demand data visualization, browsing and processing services to users from environment and earth system communities, through :

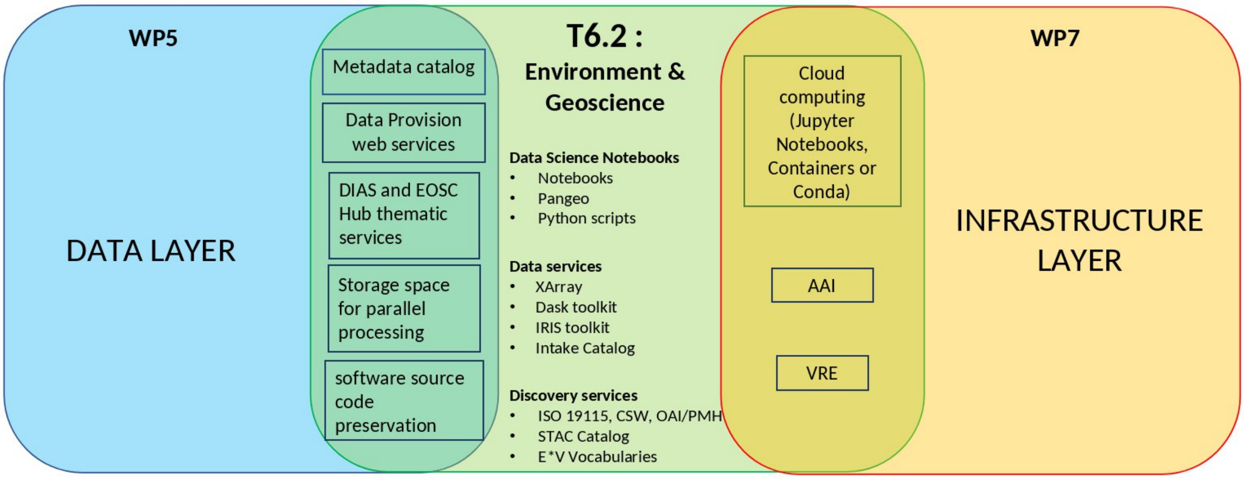
1) **Testing on demand processing platform**, that will offer a web based processing environment for algorithm developers,
2) **Providing data services**, that will speed up data access and to facilitate access to data from several domains,
3) **Providing discovery services**, by the implementation of a cross domain catalog

Gap Analysis/Missing pieces: *Not specified*

Cross-cutting analysis

- UC 6.1 about data provenance and ISO Standards
- UC 6.3 about share requirements on Calalogs and on demand processing
- UC 6.4 about saving source code from VAP developments in software heritage
- UC 6.5 about Spatio-temporal Data
- UC 6.7 about VAP, VRE and Big Data

Services needed by WP5 and WP7



3.1.3 T6.3: “Integration of data repositories into EOSC based on communities’ approaches”

Research organizations have built independently data repositories in order to fit their organizational and technical specificities and needs, using diverse technical solutions, including Dataverse (an increasingly popular open source repository tool). Meanwhile, agri-food systems are getting more and more complex (multiscale, interactions network, dynamics / transient, multidisciplinary).

The aim of this **Use Case** is to leverage EOSC services to create a flexible **federated data ecosystem for the agri-food community** with three main focus areas: **(i) data preservation, (ii) data catalog** and **(iii) cloud computing.**

This UC will be based on several institutional repositories related to the agri-food domain, such as “Data INRAE” (https://data.inrae.fr), the open-source “Phenotyping Hybrid Information System PHIS” (http://www.phis.inra.fr) and “INSERM Data” repositories (i.e. https://www.inserm.fr/en/professional-area/health-research-databases, and repositories that will be set up and designed within UC6).

User scenario (i):

Data providers, data repositories will be able to improve their interoperability and also make their data findable, accessible and usable by a wider community.

Data scientists will be able to find, trust and access a vast amount of agrifood data as well as to process, analyze and visualize the data in-situ with appropriate compute infrastructure, without the need to download them first.

The agrifood community as a whole will access a research environment which fosters collaborations and cross-fertilization.

UC#3 Analysis

Data services for management:

- INRAE Data portal - https://data.inrae.fr
- The open-source Phenotyping Hybrid Information System PHIS
- Inserm data repositories (i.e. https://www.inserm.fr/en/professional-area/health-research-databases)

Datasets type: Research data in relation with food, nutrition, agriculture and environment. It includes experimental, simulation and observation data, omic data, survey and text data. Data INRAE is agnostic in terms of data formats, however, certain file types support additional functionality, which can include downloading in different formats, previews, file-level metadata preservation, file-level data citation and exploration through data visualization and analysis.

Data services for storage: Secured storage for the Phenome-Emphasis project data

Need of UC in terms of compute power: Connection to compute services will be addressed through the Use case; in addition 2 GPU will be provided by France Grilles

Current scale: Currently national and thematic (institutional)

FAIRness of data: FAIR

Current practices: The data can be accessed and reused by anyone depending on the rights defined by the data depositors. Data deposit requires authentication and authorization.

Types of stakeholders: Service providers, Infrastructure provider, Data provider, Technology provider, Super user/supporter, End user

Auxiliary applications *Not specified*

Objectives of UC

Connect a variety of data repositories in the agrifood domain with EOSC in order to enable:

- their long term preservation (focus area 1 : data preservation)
- the findability, access and reusability of their data (focus area 2 : data catalog)

in-situ data computation (focus area 3 : cloud computing) and GPU

Gap Analysis/Missing pieces

Connectors between different services

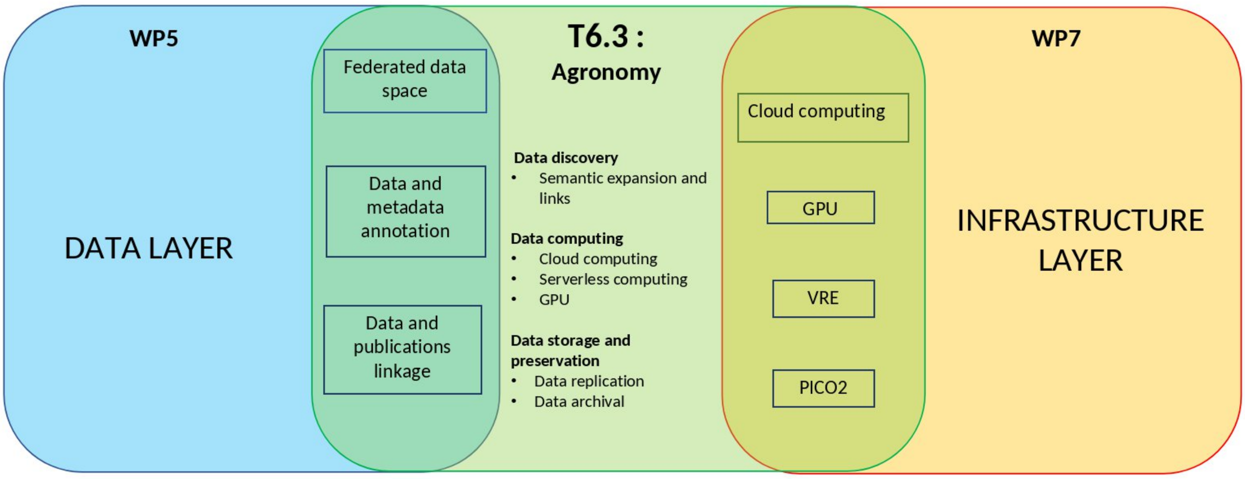
- Link Dataverse and Openstack and make it possible for users to get access from Dataverse to object in Swift/S3 + compute with Sahara and Hadoop; Single authentication to data access and compute
- Data transfer between Dataverse and VITAM
- Data and metadata annotation tools
- Connector between Dataverse and PICO2
- D4Science based VRE including subdataverses in Data INRAE

Cross-cutting analysis

- UC 6.2 about catalogs and on demand processing. Even if application domain and data are different, some objectives of the two use cases are very similar.
- UC 6.6 for data preservation

- A possible link with UC 6.5 through the data catalog so that users can navigate from data to publications and vice versa using semantic linkages.

Services needed by WP5 and WP7



3.1.4 T6.4: "Software source code preservation, reference and access"

Leveraging the experience of **Software Heritage**, the challenge of this **Use Case** will be to **design and pilot a solution for the preservation of massive collections of software source code into EOSC eTDR service** (European Trusted Digital Repository).

The Software Heritage project is building a structured archive of all available source code to preserve human's technological and scientific knowledge. It involves many different aspects, and different ways of gathering said source code. If most of it is actually retrieved using scraping techniques from well-known sources (GitHub, GitLab, etc.), it is also possible to explicitly push source code artifacts to the Archive, via a deposit API.

In this context, Software Heritage will make available a copy of its archive via the EOSC eTDR for long-term preservation and integrate its APIs in the service catalog.

User scenario (i):

1. *Deposit of a software source code in the Software Heritage Archive.*

   A researcher wants to ensure the software source code used to produce valuable results published in a scientific paper is properly identified and cited, with the ability for a reader to retrieve this source code at the time of the paper publication as well as afterwards (with no known time-limit).

   Using an Open Access Repository application provided by EOSC-eTDR, she uploads a compressed archive file (zip or tgz file) containing the software source code as well as the metadata describing the deposited source code, following the deposit best practices as much as possible.

   The Open Access Repository performs a query on the Software Heritage deposit API with the given software source code archive and the related metadata. The response from the deposit API contains a deposit identifier that can be used to later retrieve a unique and persistent identifier (or “SWHID”) that the Open Access Repository application attaches to the new entry created by the end user.

She can use the obtained SWHID in her papers to properly cite the software she used to produce the results presented.

2. *Retrieve an exact copy of the source code used in a published research paper.*

   A scientific researcher wants to reproduce some results from a published paper in which the source code used to produce the presented results is properly identified with a Software Heritage unique ID (or “SWHID”). She downloads the exact copy of the software source code on her workstation from the Software Heritage Archive using the Software Heritage Vault web application.

3. *Look for a software project in the Software Heritage Archive*

   A scientific researcher wants to retrieve the source code used in a published scientific paper, but this is only cited by either its name or a project URL that is no longer valid.She browses the Software Heritage web application and uses the search form to look for a copy of this software source code. Once found, she uses the repository browsing capability of the Software Heritage Archive to read the files and documentation, navigate the source code history, etc. Eventually, she asks for a copy of the browser repository using the “Download” action provided by the Software Heritage Archive web application.

UC#4 Analysis **Data services for management:** The software Heritage stack (www.softwareheritage.org) : • Software Heritage Archive (https://archive.softwareheritage.org) • Software Heritage Web API https://archive.softwareheritage.org/api/ • Software Heritage technical documentation: https://docs.softwareheritage.org/devel/ • Software Heritage Persistent Identifiers (SWHIDs): https://docs.softwareheritage.org/devel/swh-model/persistent-identifiers.html
Datasets type: Text-based files (code)
Data services for storage: Software heritage archive application stack. Data hosted at Inria and on several public cloud-based resources
Need of UC in terms of compute power: A full copy of the Software Heritage Archive requires at least: – 600 TB for hosting the object storage, – 6 TB for hosting the Software Heritage graph in a PostgreSQL database. – Decent machine for running the database (at least 64GB of RAM). – 4 to 8 medium-sized VMs (4 core, 8/16GB) for the consumers of incoming mirror data.
Current scale: Running full stack in house (Inria) and on several public cloud-based resources. Nationally controlled deployment with geographical distribution within Europe
FAIRness of data: *Not specified*
Current practices: Strong existing sharing practices in the software development world
Types of stakeholders: Data providers and consumers are end users (possibly the whole world)

Auxiliary applications: *Not specified*

Objectives of UC

- Develop a deposit API for research software archival.
- Develop an access API for retrieval of archived software.
- Standardize a persistent identifier schema (PID) for referencing billions of archived software artifacts.
- Integrate the above APIs and Services with EOSC eTDR.
- Develop a pilot to fully host the software archive onto existing EOSC infrastructure.

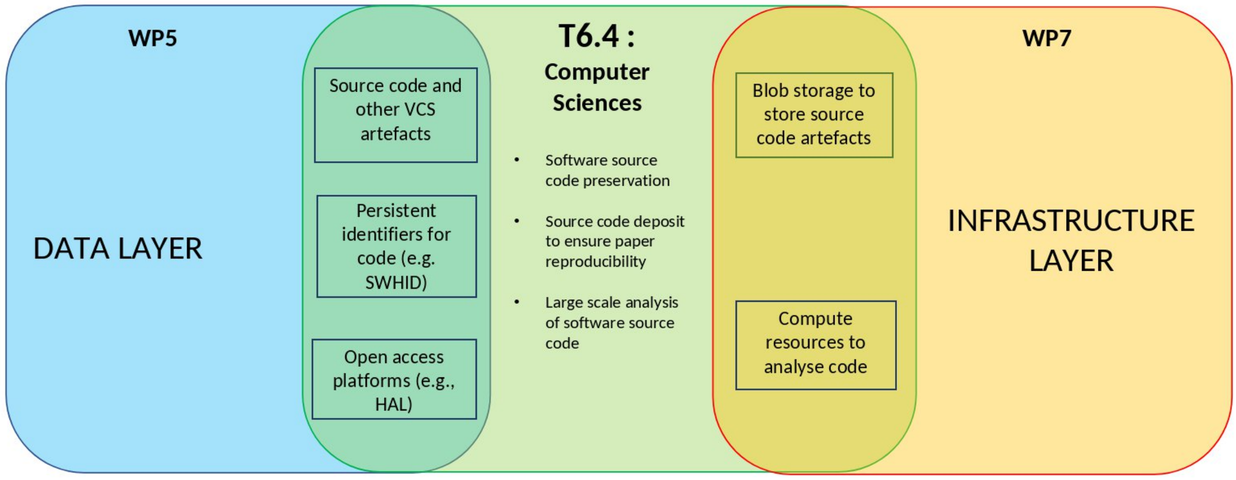
Gap Analysis/Missing pieces

- Backup repository.
- Link to publication systems.

Cross-cutting analysis

- All the UC may be interested in archiving source code
- Privileged relation with UC 6.5 for the link between software heritage and HAL

Services needed by WP5 and WP7


WP5
DATA LAYER
Source code and other VCS artefacts
Persistent identifiers for code (e.g. SWHID)
Open access platforms (e.g., HAL)
T6.4 : Computer Sciences
Software source code preservation
Source code deposit to ensure paper reproducibility
Large scale analysis of software source code
Blob storage to store source code artefacts
Compute resources to analyse code
WP7
INFRASTRUCTURE LAYER

3.1.5 T6.5: “FAIR principles in data life-cycles for humanities”

The aim was to identify specific **Use Cases** based on **SSH** communities’ engagement.

Several Use Cases (7) have been already identified:

1. **Linking Nakala repository to HAL open archive Use Case.** This UC focuses on the link between data and publications by working on a Proof of Concept to link two existing repositories (Nakala and HAL). It will take the form of a new workflow between both platforms. This UC will be a model for linking with other data repositories used in SSH communities (*Analysis matrix below*).
2. **LIDAR data Use Case.** In order to promote open data and collaborative work on huge lidar datasets, the UC aims to develop open access frameworks and infrastructures for aggregating, sharing and collaborating on lidar datasets. The net result of this work will be to

create consistent, comparable datasets of human impacts on the Earth's surface. One of the issues identified is concerning research using lidar data. There are several different researchers across various EU member states working on closely related research questions, usually in isolation across varies disciplines such geomorphometry and forest ecology, and often without access to required resources. In order to set the scale of such a service at a national and transnational level, the EFEO will rely on French national Infrastructures like Huma-Num for hosting or GENCI /IDRIS/In2p3 to power the deep learning processes launched by the Web GIS platform (https://www.efeo.fr/).

3. **Geo-historical semantic data Use Case.** This Use Case is building on the good practices in the constitution of such data and especially on the means of making the database tools interoperable with the tools for disseminating this spatial-temporal data[6]. The main goals here are respectively: (i) to develop bridges between database tool, semantic data tool and spatial portal, (ii) to develop an integrated and automated workflow for all these treatments, and (iii) to promote dissemination of these data as streams into integrated portals.
4. **Open science publication infrastructure with 3D data for humanities Use Case.** The 3D French consortium gather more than 400 3D models that are progressively integrated in the CND3D (Conservatoire National des Données 3D). At this time only one hundred have entered the repository and some metadata still need to be aligned with standard thesaurus like Geonames for localisation, PeriodO for time period and PACTOLS or other standard thesaurus for subject metadata. It is evident that a lot of common features exist with HAL. However, some work needs to be done on the metadata and for the API. For the metadata, it should be aligned with the specific thesaurus of HAL (https://aurehal.archives-ouvertes.fr/) notably for the authors, research structures, scientific domains, ANR and European projects. A mapping between permanent ID (DOI on CND3D and HAL dedicated ones) and url to ease the interconnections should also developed. The HAL recommendation for API and indexing have to be integrated.
5. **3D data open repository for humanities Use Case.** To give the international community a practical tool able to make a complete 3D data deposit in order to preserve 3D data, share and access them in the LOD cloud thanks to the existing 3D Data Repository (*Conservatoire National des Données 3D* / CND3D) is the main purpose of this UC. Actually, due to the fact that it is only supported by national infrastructures, deposit is for the moment limited to results of researches from French teams or in collaboration with French organizations. Due to the volume of considered data, an interconnected infrastructure has to be set-up. The internal infrastructure should be improved to be able to scale up to international interconnections.
6. **OpenArcheo Use Case.** The Semantic Web with its Linked Open Data cloud enables scholars and cultural institutions to publish their data in RDF, using CIDOC CRM ontology as an interlingua that enables a semantically consistent re-interpretation of their data. Nowadays more and more projects have done the task of mapping legacy datasets to CIDOC CRM, and successful Extract-Transform-Load data-integration processes have been performed in this way. A next step is enabling people and applications to actually dynamically explore autonomous datasets using the semantic mediation offered by CIDOC CRM. This is the purpose of OpenArchaeo, a tool for querying archaeological datasets on the LOD cloud.

[6] https://paris-timemachine.huma-num.fr/la-plateforme-de-webmapping-geo/;https://paris-timemachine.huma-num.fr/heurist-une-base-de-donnees-generique-pour-les-sciences-humaines-et-sociales/;https://paris-timemachine.huma-num.fr/oronce-fine-une-plateforme-web-semantique-pour-les-donnees-spatio-temporelles/;https://documentation.huma-num.fr/content/17/130/fr/what-can-huma_num-do-for-you.html;https://documentation.huma-num.fr/content/25/71/fr/l%E2%80%99hebergement-de-sites-web-chez-huma_num.html

7. **ORTOLANG language data Use Case.** ORTOLANG is the leading repository of language data in France (https://www.ortolang.fr/). The goal of this UC is twofold: (i) maintain, secure and improve the national ORTOLANG platform (usability, interoperability...), and (ii) develop a workflow to connect the corpora stored on the platform to the existing corpus tools. Data that have been deposited in formats such as PDF, raw text format, or text processing formats, have to be processed in a more accessible format (e.g. TEI) so that they are ready to be processed by automatic tools such as grammatical analysis. Some of the automatic tools have actually been deposited and saved in ORTOLANG, however the choice of processing tools does not have to be limited only to these tools. The result of the processing has to be exposed in adequate manner so that they can be viewed online, and later downloaded and used for scientific work. In other words, in its current state, the ORTOLANG platform enables researchers to download language data, including corpora, but does not enable them to import the corpora available on ORTOLANG within an existing corpus tool or to explore the corpus data in plain text format (query based on metadata are fully functional). The development of such type of workflow is a recurrent request of the community of linguists, who would be eager to have the possibility to query corpora before downloading them (*e.g.* to verify the presence or absence of certain linguistics forms or topics they are interested in), and to download the existing corpora in a format supported by the leading corpus tools.

In the case of the T6.5, it has been suggested to merge some internal UCs and then some technical sub-groups of UCs have been proposed in order to prioritize actions to be implemented in the framework of EOSC-Pillar. The Task consortium has decided to make the “Linking Nakala repository to HAL open archive Use Case” a priority. In any case, it does not avoid any other developments in relation to the other UCs identified by the SSH communities.

“Linking Nakala repository to HAL open archive Use Case”

User scenario (i):

As a data user I would like to find data used in a publication in order to access these data and exploit them. Result is expected to be a list available on the landing page of the publication. Data could be visualized through the French open archive HAL or directly on the data repository Nakala.

As a SSH researcher I would like to link my publication available in the French open archive and my data in the repository Nakala. Result is expected to be a dissemination of the informations through different ways : ISIDORE search engine in SSH, OAI-PMH, Triplestore.

UC#5 Analysis

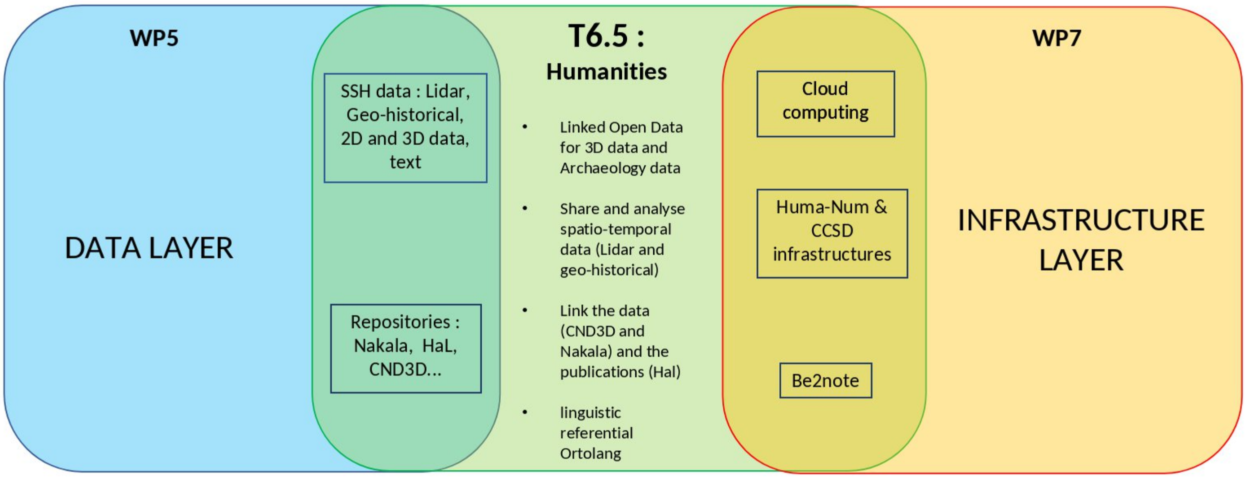
Data services for management:

HAL (https://hal.archives-ouvertes.fr/): French National Open Archive to deposit and access open access research publication. Centralised database providing specific portal and views. It assigns PIDs and stable URLs, curation, time stamping (https://www.ccsd.cnrs.fr/en/open-archives/ https://hal.archives-ouvertes.fr/ https://doc.archives-ouvertes.fr/en/homepage/) and dissemination through OAI-PMH, APIs (https://api.archives-ouvertes.fr/docs) and a triplestore (https://data.archives-ouvertes.fr/).

Nakala (https://www.nakala.fr/): data repository for SSH providing PID, permanent data access and metadata dissemination through a triple store and OAI-PMH (https://documentation.huma-num.fr/content/17/130/fr/what-can-huma_num-do-for-

you.html ; https://www.nakala.fr/ (but the technical documentation through a Web app will evolve in late 2019)
Datasets type: Open Access Research publications (articles, communication, reports, PhD thesis, ...) Text file, audio, video and image.
Data services for storage: HAL and Nakala are two data repositories. HAL is currently archived at CINES
Need of UC in terms of compute power: Not defined.
Current scale: HAL: National archive used by international partners to deposit publications. Should be scalable. What would be necessary to extend the scale of this service in terms of infrastructure. A need for scalability will be more focused on curation (hiring personnel).
FAIRness of data: HAL and Nakala provide FAIR data.
Current practices: National, but involved in European and international networks and possibly useful for other communities and data managers, especially those who already use SWORD (Simple Web-Service Offering Repository Deposit, http://swordapp.org/) protocol.
Types of stakeholders: • Humanum Nakala's team • HAL Technical team • Users : Researchers in SSH
Auxiliary applications Web Annotation Data Model (used with B2NOTE system) to convey information about a resource or associations between resources
Objectives of UC • Linking Data into Nakala to Publications into HAL and developing a new workflow between both platforms. • Define an interoperability model between data repository and publication repository using standards (Web Annotation Data Model). This model can be transposed to other repositories (3D data open repository, DataINRAe, ...)
Gap Analysis/Missing pieces Make NAKALA compatible with SWORD and thus install a server for SWORD, which is an interoperability standard for digital file deposit. Rework the HAL interface so that the user has the least effort to make to access more information/data possible.
Cross-cutting analysis • UC 6.4 to establish the link between Hal and Software Heritage

Services needed by WP5 and WP7


WP5
DATA LAYER
SSH data : Lidar, Geo-historical, 2D and 3D data, text
Repositories : Nakala, HaL, CND3D…
T6.5 : Humanities
• Linked Open Data for 3D data and Archaeology data
• Share and analyse spatio-temporal data (Lidar and geo-historical)
• Link the data (CND3D and Nakala) and the publications (Hal)
• linguistic referential Ortolang
Cloud computing
Huma-Num & CCSD infrastructures
Be2note
WP7
INFRASTRUCTURE LAYER

3.1.6 T6.6: "Exploring reference data through existing computing services for the bioinformatics"

This **Use Case** aims at producing an extensive study of the possible Galaxy deployment scenarios including guidelines and best practices to realise a proof of concept (POC)/prototype of a service allowing access to various reference data sets through Galaxy for the **bioinformatics**.

The purpose of this UC will be to ensure the interoperation between already available Galaxy computing services (developed by Elixir-IT) and data repositories (currently being designed by INSERM), with the final aim of serving the Elixir user community as a whole. It will build on top of existing national services made available in France and Italy by participating partners, and will aim at fulfilling the following objectives:

1. Allow access to reference data from different Galaxy deployments for data analysis.
2. Facilitate the deployment of Galaxy instances close to the data.
3. Provide coherency between different existing Galaxy deployments.
4. Ensure health data security requirements are met throughout the process.

User scenario (i):

As a researcher, I would like to access a subset of genomics data from a given cohort in order to perform analysis/computation on these data without having to transfer them.

As a bioinformatician (Galaxy user) I would like to know how to format my data in order to ensure it can be used properly throughout my workflows

As a data provider (health data) I would like to be sure the data I'm taking care of are both properly secured and yet accessible to various computational tools so that they can be useful to researchers

As a developer, I would like to be able to deploy Galaxy on different Cloud infrastructures across Europe transparently in order to implement and make available new tools and workflows very quickly on the different data spaces available to me.

As a healthcare provider interested in personalized medicine, I would like to be able to use GDPR compliant Galaxy instances deployed over Cloud infrastructures across Europe so I could safely process patients data.

UC#6 Analysis
Data services for management: • INSERM data hub service, currently being set up: to describe, identify and reference data. • Data repositories containing genomics data (https://www.inserm.fr/en/professional-area/health-research-databases), in order to build an integrated and interoperable service for ELIXIR and the wider Life Science user community as a whole. • The repository of reference datasets is shared by all the Galaxy instances, to avoid useless and costly data duplication, through the CERN-VM read-only filesystem.
Datasets type: Mainly Genomic.
Data services for storage: LUKS (Linux Unified Key Setup) is used for provisioning of encrypted volumes to store users' data.
Need of UC in terms of compute power: No, already provided by INDIGO Data Cloud (INFN partner in this project).
Current scale: European.
FAIRness of data: Yes, they are FAIR compliant.
Current practices: • Established and accepted practices in the bioinformatics world. • Sharing constraints when source data is personal/health data.
Types of stakeholders: • Service provider: provide access to the Galaxy instances • Infrastructure provider: access to computing and data • Data providers: access to the reference data • Technology providers: develop solutions and technologies • Super user/supporter: Acts as proxy between service and users • End user: uses the service
Auxiliary applications • PostgreSQL, NGINX, uWSGI, and Proftpd • Indigo components: Proxy server, IAM, SLAM, PaaS Orchestrator, Vault etc.
Objectives of UC • Allow access to reference data from different Galaxy deployments for data analysis;

- Facilitate the deployment of Galaxy instances close to the data;
- Provide coherency between different existing Galaxy deployments; and
- Ensure health data security requirements are met throughout the process.

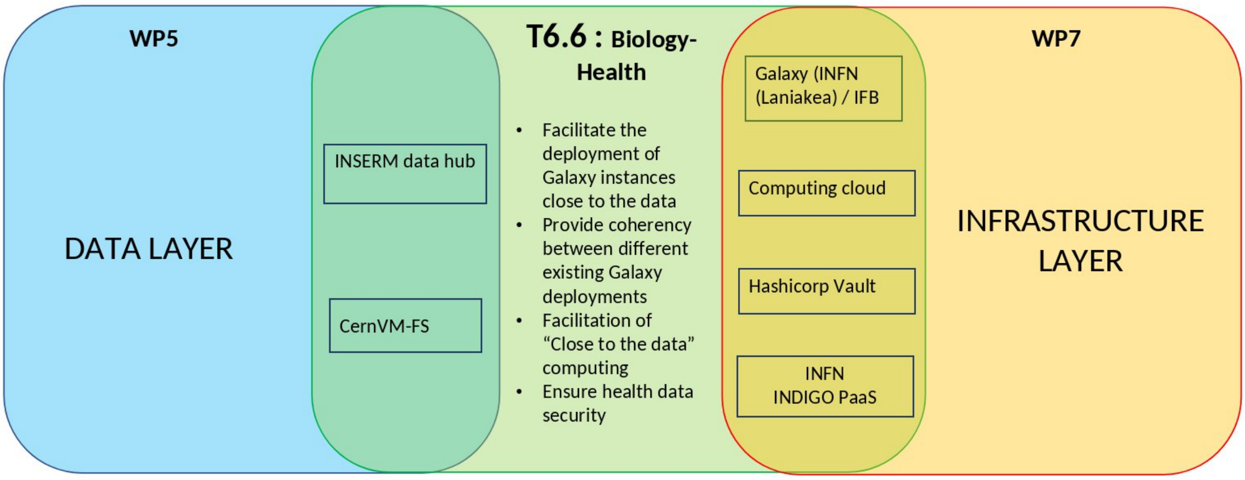
Gap Analysis/Missing pieces

- Strong link between data and compute facilities in order to perform close-to-the-data computing
- Harmonisation of data format and structure in existing/foreseen repositories so that they can be used as input for Galaxy analyses
- Recipes and deployment practices for using Galaxy in a secured (health data compatible) environment

Cross-cutting analysis

- UC 6.2 for computation close to data
- UC 6.3 might be interested on the deployment of Galaxy instances close to data
- UC 6.7 for big data

Services needed by WP5 and WP7


WP5
DATA LAYER
INSERM data hub
CernVM-FS
T6.6 : Biology-Health
Facilitate the deployment of Galaxy instances close to the data
Provide coherency between different existing Galaxy deployments
Facilitation of "Close to the data" computing
Ensure health data security
Galaxy (INFN (Laniakea) / IFB
Computing cloud
Hashicorp Vault
INFN INDIGO PaaS
WP7
INFRASTRUCTURE LAYER

3.1.7 T6.7: "Suitable data formats for seismological big data provisioning via web services"

Seismological web services have been designed some years ago with particular types of users and data in mind, which are not exclusively what we see today. Disruptive techniques, generating data at much a finer resolution, compared to the standard seismic stations, are a challenge for data centres and users. For the data centres, because they have to offer long-term archival of the data, and for the users, because they need to change the way they are used to access data.

This **Use Case** has two main purposes:

1. The **evaluation and adoption of new data formats** for the provision of big seismological datasets, as well as best formats for the archival of the experiments, and the implementation of these recommendations in the current data provisioning services (FDSN Dataselect-WS).

2. The **re-evaluation of the current data provisioning services** specification taking into account the change in formats, volume, usage and computational platforms to be used.

User scenario (i):

As a seismologist processing large volumes of seismic waveforms I would like to be able to request these data and:

1. select a convenient format to access the data in the best way,
2. stage them in an HPC facility or in some cloud environment,
3. attach to the data all needed metadata to process them properly.

As a Data Centre Manager, I would like to:

1. archive experiments generating a large volume of data in the best format considering size and speed of access, as well as allowing their fast processing even in real time,
2. export subsets to be published through the normal (not big-data) services in the standard formats.

UC#7 Analysis **Data services for management:** • GEOFON data centre system implemented FDSN-WS/Data select -ws • FDSN Dataselect http://www.fdsn.org/webservices/fdsnws-dataselect-1.1.pdf • FDSN Station-WS http://www.fdsn.org/webservices/fdsnws-station-1.1.pdf • Possibility to archive data (or experiments generating large volume of data) in different formats
Datasets type: • Earthquake parametric data. • Seismic waveforms. • Datasets sensors.
Data services for storage: stage data in an HPC facility or in some cloud environment.
Need of UC in terms of compute power: computational platforms.
Current scale: internationally, Europe, USA.
FAIRness of data • Yes, but new developments are required to properly and FAIRly serve these data. • Convert on-the-fly big data volumes from proprietary to standard formats.
Current practices: The web service providing seismological data (Dataselect) is an international standard adopted by most of the seismological data centres. Data formats to be evaluated are also known internationally. Extensions to those could be implemented if needed.
Types of stakeholders: • Seismologist, User of services. • Data Centre Manager. • Developer.
Auxiliary applications

- New formats might be used e.g., ASDF
- Provide data quality with a wide spectrum of applications ranging from Tsunami early warning to Infrastructure monitoring

Objectives of UC

- Evaluation and adoption of new data formats for the provision of big seismological datasets, as well as best formats for the archival of the experiments.
- Implementation of these recommendations in the current data provisioning services.
- Re-evaluation of the current data provisioning services specification taking into account the change in formats, volume, usage and computational platforms to be used.
- Propose the results from these activities as recommendations and/or international standards, considering that the members of the group are important players in the organization encompassing all seismological data centres at the global level (FDSN).

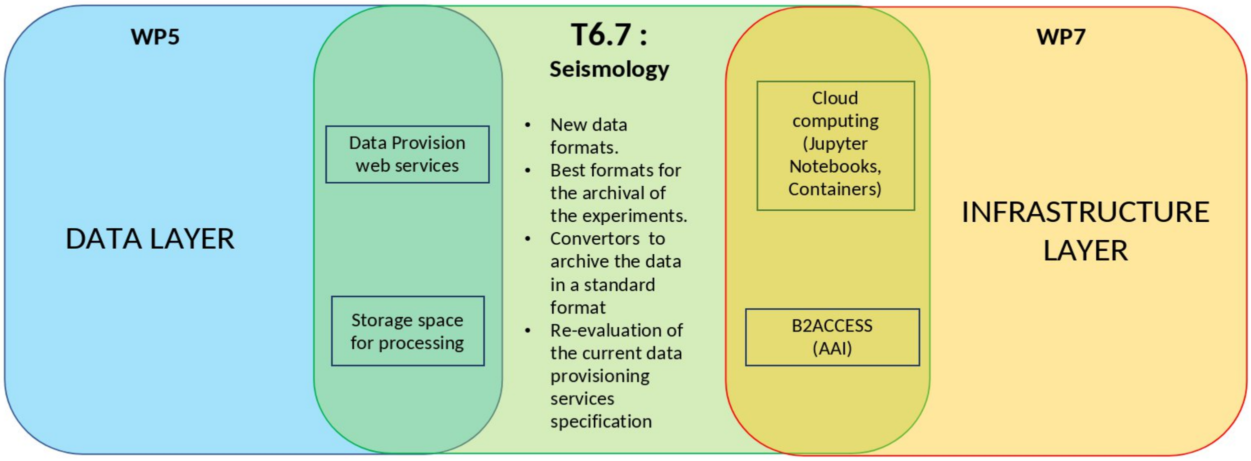
Gap Analysis/Missing pieces

- Web services not suitable to provide big amounts of data, as these are expected to be received synchronously.
- No interactions with other web service providing the needed metadata to properly interpret the data.
- Current used formats are not able to pack data and metadata in the same container.

Cross-cutting analysis

- UC 6.2 for collaboration with RESIF

Services needed by WP5 and WP7


WP5
DATA LAYER
Data Provision web services
Storage space for processing
T6.7 :
Seismology
• New data formats.
• Best formats for the archival of the experiments.
• Convertors to archive the data in a standard format
• Re-evaluation of the current data provisioning services specification
Cloud computing (Jupyter Notebooks, Containers)
B2ACCESS (AAI)
WP7
INFRASTRUCTURE LAYER

3.1.8 T6.8: "Virtual definition of data sets according to RDA recommendations (seismological data)"

Many actors during the data life cycle need at some moment to aggregate data from different sources in order to provide scientists (data consumers) a more comprehensible and high-quality dataset. The capability to aggregate different pieces of information without limitations on data formats and/or structure provides enough flexibility as to publish and merge results from different disciplines and in a variety of situations. For instance, usual activities in data Repositories and Libraries among others. There are already prototypes of this service running internally in different institutions at the global level (e.g. GEOFON, DKRZ, Perseids Project).

This **Use Case** main goal is:

- provide a production-ready *Data Collection system* based on the already available system at GEOFON following the **RDA Recommendations** from the Research Data Collections WG. Include extensions to the specification based on the needs of the involved communities.

User scenario (i):

As a seismologist requesting seismic waveforms to process, I would like to:

1. save the definition of an aggregated dataset with different types of information and formats,
2. which includes not only the data, but also the metadata, information about the earthquake parametric data and technical reports providing a context to the data. The definition of this aggregated dataset should be further editable if needed.

As the operators of a scientific open platform, we want all data we publish to be easily shared and reused by the larger community, at all stages of the publication lifecycle. I would like to persistently identify, version, carry fine-grained provenance metadata and validate it against a profile, schema or other verifiable criteria. To facilitate reuse, I must be able to:

1. describe collection items as machine-actionable data types, independent of their identifier schemes, and the properties of the collection to which they belong.
2. create reusable templates of collection types with standard descriptive properties and capabilities.
3. express relationships between collections, items within a collection, and items across collections using one or more standard ontologies
4. perform simple CRUD/L operations on collections and items in a collection
5. perform more complex discovery operations on collections based upon the properties of individual collection items, such as finding all items across all collections that match or don't match or contain a specific item.

A scientist from Climatology defines a dataset which is being used to obtain a result. On the sake of reproducibility, he/she would like to:

1. save the definition of the dataset,
2. store metadata describing the dataset.

UC#8 Analysis **Data services for management:** • GEOFON data centre system implemented FDSN-WS/Data select -ws • GEOFON offers access to seismic waveform and station metadata using the FDSN web services. https://geofon.gfz-potsdam.de/waveform/builder-dataselect.php
Datasets type: • Pre-assembled datasets at seismological data centres. • Dynamic datasets at seismological data centres.

Data services for storage

- Data Collections System following RDA recommendations: https://www.rd-alliance.org/system/files/rda-collections-recommendation_ref16102017.pdf
- A collection is a 4-tuple of an identifier, capabilities, collection properties, and membership.

Need of UC in terms of compute power: DC System needs basically storage space for a DB and computational power to run the implementation of the API. The volume of data is not too demanding, as only references (i.e. PIDs, links) to the resources are stored and provided.

Current scale:

- International (global), for the implementation at GEOFON.
- 6000 collections and 1.5 million members for datasets pre-defined by the data centre operators.

FAIRness of data: As they follow the recommendations of RDA, it is supposed that datasets are FAIR.

Current practices: *Not specified*

Types of stakeholders:

- Seismologist.
- Data Centre Manager.
- Scientist.
- Librarian.

Auxiliary applications

- Handle server and AAI are already provided
- SeisComP ® 3 Software package (freely available)
- DKRZ is also running a collection-builder instance in the framework of the ENES climate data infrastructure (technically integrated with ESGF, Earth System Grid Federation)

Objectives of UC

- Provide a production-ready implementation of a system following the RDA Recommendations from the Research Data Collections WG.
- The Data Collection System should be running in production mode at GEOFON and available to be used by the community.
- The GEOFON instance of the system will identify created collections with PIDs hosted in their own Handle Server. As a fallback, for other deployments of the system, IDs identifying the collections could be saved locally without including these in the global Handle system (e.g. locally resolvable instead of globally resolvable).

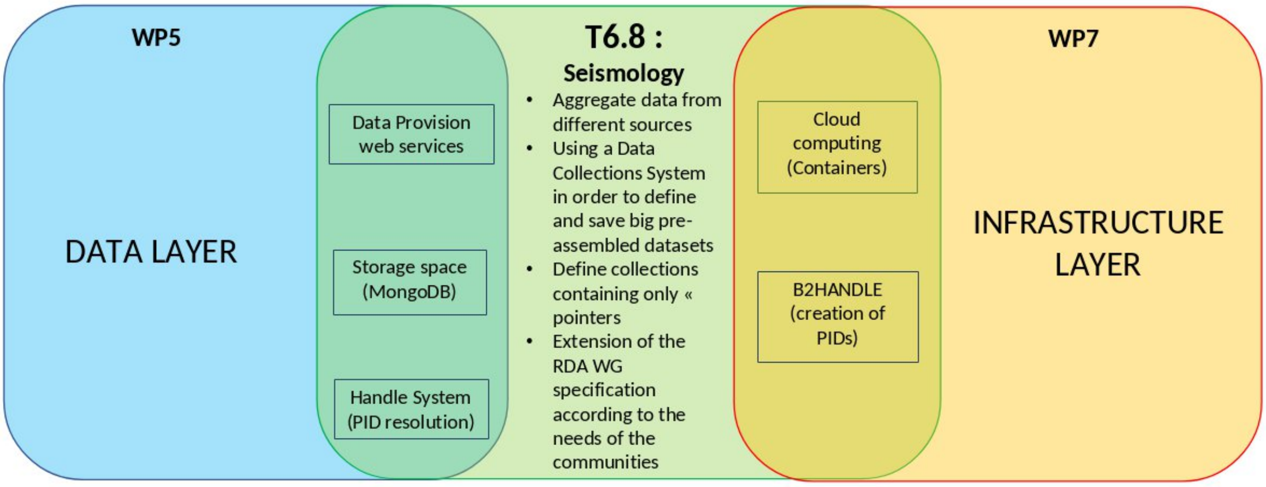
Gap Analysis/Missing pieces

- Develop smart clients to access and download a collection to a local computer, or to stage it to a computational facility.
- In the case that each collection must be identified by a PID, the Handle service should be provided by some partner and the possibility to pay for a prefix should be considered (US$ 50/year).

Cross-cutting analysis

- Use Cases which have to aggregate their data into collections

Services needed by WP5 and WP7


WP5
DATA LAYER
Data Provision web services
Storage space (MongoDB)
Handle System (PID resolution)
T6.8 :
Seismology
• Aggregate data from different sources
• Using a Data Collections System in order to define and save big pre-assembled datasets
• Define collections containing only « pointers
• Extension of the RDA WG specification according to the needs of the communities
Cloud computing (Containers)
B2HANDLE (creation of PIDs)
WP7
INFRASTRUCTURE LAYER

3.1.9 T6.9: "Integrating heterogeneous data on cultural heritage"

The overall purpose of this **Use Case** is to establish connections among data produced by analyses of **heritage** items (e.g. archaeological artefacts, artistic objects, parts of monuments etc.) with generic reports on the same kind of objects. The latter are usually available with poor metadata that the use of TEXTCROWD, a NLP package used for text mining of heritage-related text files, should enrich. This is true especially for legacy data. Moreover, they are usually produced in the language of the country where they are created. For such metadata no standardization is currently in use. Metadata for numeric datasets produced by analyses are instead recorded together with the numeric values according to a standard data organization. NLP is heavily dependent on the language of the text documents and on general domain thesauri, preferably multilingual ones.

This UC aims at porting TEXTCROWD-a version, which has already been tested for archaeological data in the ARIADNE project (for English and Dutch) and in EOSCpilot for Italian, to new languages.

This will require:

1. installing the appropriate linguistic libraries;
2. creating (if not already available) discipline-related multilingual thesauri,
3. porting TEXTCROWD as a container, and
4. testing it for more languages in real use conditions.

User scenario (i):

As a researcher in art history, archaeology, heritage science, or a museum curator, I want to apply TEXTCROWD to a set of multilingual text documents and enrich their metadata, to set up a personal mini-knowledge base.

UC#9 Analysis

Data services for management:

TEXTCROWD-a version: NLP package used for text mining of heritage related text files.

Datasets type: archaeological data, Dublin Core metadata.

Data services for storage: *Not specified*

Need of UC in terms of compute power: NO.

Current scale: European.

FAIRness of data: FAIR principles implementation to improve results accessibility

Current practices: Using Textcrowd to enrich metadata, to set up a mini-knowledge base for the personal research question.

Types of stakeholders:

- Heritage professional / researcher : User in art history, archaeology, heritage science;
- museum curator;
- professional in charge of conservation.

Auxiliary applications: *Not specified*

Objectives of UC

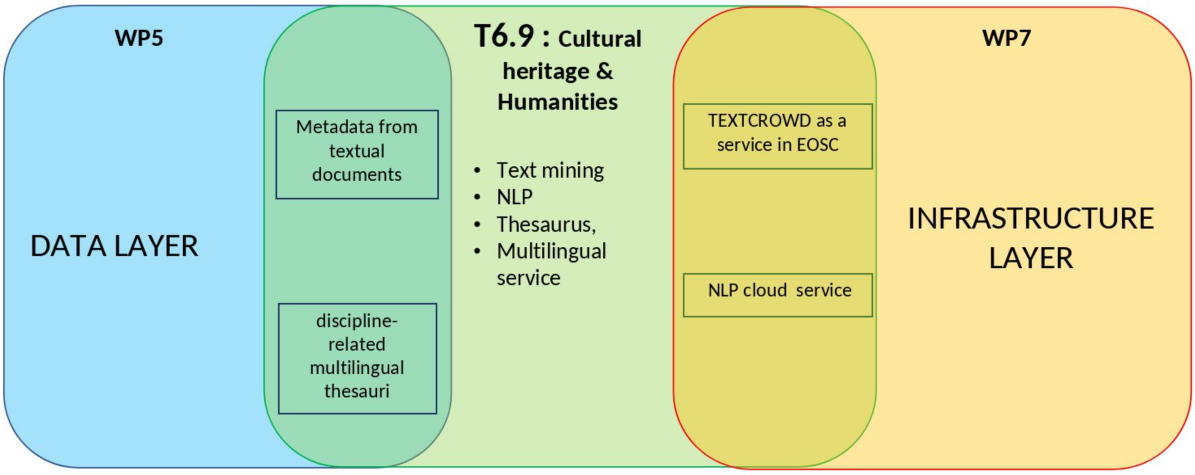
The UC concerns the implementation of a text mining tool to bridge text reports (e.g. from art history, conservation etc.) to datasets deriving from heritage science activities. The tool has already been developed in previous projects and tested in a cloud environment in EOSCpilot, without involving so far the scientific datasets.

Gap Analysis/Missing pieces: *Not specified*

Cross-cutting analysis

- UC 6.1 Data Provenance
- UC 6.5 Ortolang

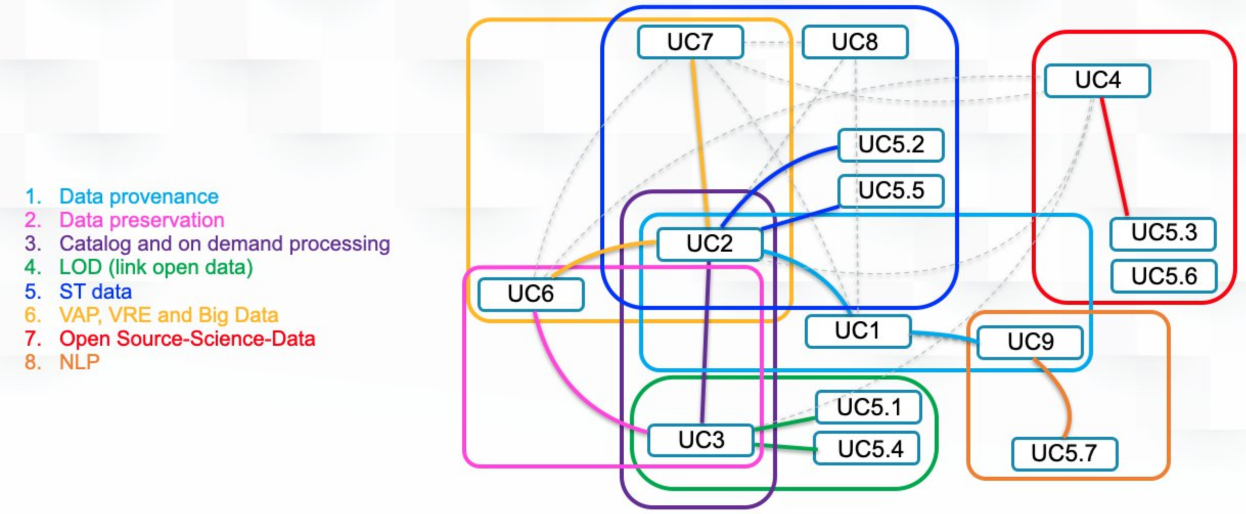
Services needed by WP5 and WP7


WP5
DATA LAYER
Metadata from textual documents
discipline-related multilingual thesauri
T6.9 : Cultural heritage & Humanities
Text mining
NLP
Thesaurus,
Multilingual service
TEXTCROWD as a service in EOSC
NLP cloud service
WP7
INFRASTRUCTURE LAYER

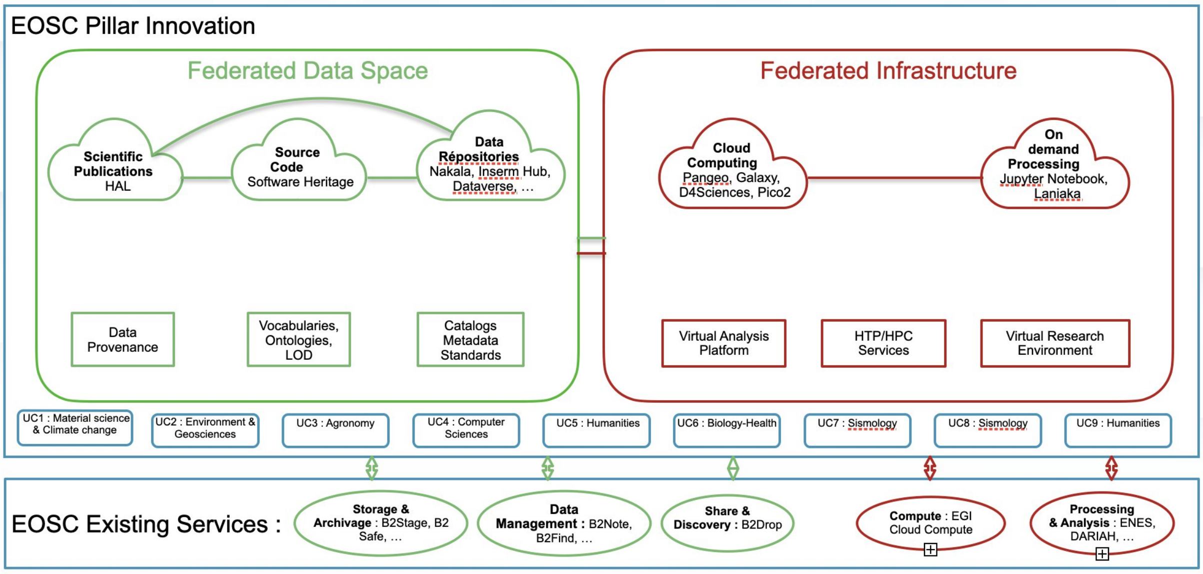
4 Conclusions

The activity that led to the present “state of the art” was particularly complex due to the diversity of the Uses Cases identified by each task within the framework of WP6. Clearly, the Use Cases appear to be quite different from one to another, which has required a more robust cross-cutting analytical approach in order to be able to identify needs and thus define common or clustering services that could be provided within the framework of EOSC-Pillar Consortium.

The figure here-below presents all the Use Cases as a structured network where nodes are the UCs and edges the interactions identified. Eight clusters have been set up.



The eight technical clusters are focused on: **1. Data Provenance, 2. Data Preservation, 3. Catalog and On Demand Processing, 4. LOD (Linked Open Data), 5. Spatio-Temporal Data, 6. VAP (Virtual Analysis Platform), VRE (Virtual Research Environment) and Big Data, 7. Open Source-Science-Data,** and **8. NLP (Natural Language Processing).**


EOSC Pillar Innovation
Federated Data Space
Scientific Publications HAL
Source Code Software Heritage
Data Répositories Nakala, Inserm Hub, Dataverse, …
Data Provenance
Vocabularies, Ontologies, LOD
Catalogs Metadata Standards
Federated Infrastructure
Cloud Computing Pangeo, Galaxy, D4Sciences, Pico2
On demand Processing Jupyter Notebook, Laniaka
Virtual Analysis Platform
HTP/HPC Services
Virtual Research Environment
UC1 : Material science & Climate change
UC2 : Environment & Geosciences
UC3 : Agronomy
UC4 : Computer Sciences
UC5 : Humanities
UC6 : Biology-Health
UC7 : Sismology
UC8 : Sismology
UC9 : Humanities
EOSC Existing Services :
Storage & Archivage : B2Stage, B2 Safe, …
Data Management : B2Note, B2Find, …
Share & Discovery : B2Drop
Compute : EGI Cloud Compute
Processing & Analysis : ENES, DARIAH, …

These potential technical synergies will allow to validate inter-community solutions that could be extended to other communities with minimal changes (innovation system).