Project Acronym:	DataBio
Grant Agreement number:	732064 (H2020-ICT-2016-1 – Innovation Action)
Project Full Title:	Data-Driven Bioeconomy
Project Coordinator:	INTRASOFT International

Funded by the Horizon 2020 Framework Programme of the European Union

This project is part of BDV PPP

DELIVERABLE

D4.4 – Service Documentation

Dissemination level	PU -Public
Type of Document	Report
Contractual date of delivery	M35 – 30/11/2019
Deliverable Leader	EXUS
Status - version, date	Final – v1.0, 30/12/2019
WP / Task responsible	WP4
Keywords:	Platform, component, interface, pilot, pipeline, service, experiment, pilot, trial

Executive Summary

The public deliverable D4.4 describes the software components and processes (here called pipelines as the processes mostly consist of Big Data volumes streaming through successive processing steps) to be utilized by the DataBio Platform and pilots. The pilot services were tested through two phases, Trial 1 and Trial 2 of the project. Most of the components were used in both Trials with some updates in their features for Trial 2. In addition, this deliverable reports which components were deployed in each pilot and the development platform that the pilots tested their Big Data solutions on. The document aggregates information dispersed among various deliverables (namely [REF-01] - [REF-06]). The aim of this deliverable is to create a comprehensive overview of DataBio technical results.

The objective of WP4 "DataBio Platform with Pilot Support" was to configure and adopt Big Data technologies for agriculture, forestry, and fishery. The work package together with WP5 "Earth Observation and Geospatial Data and Services", established a platform for the development of bioeconomy applications. The software and dataset repository DataBio Hub is a central resource of the platform. In doing so, WP4 supported the DataBio pilots in their needs for Big Data technologies.

This deliverable starts with an overview of DataBio building blocks such as platform architecture, software components, datasets, models that offer functionalities primarily for services in the domains of agriculture, forestry, and fishery. Then follows the exploitation for the identification of cross reusable (sub) pipelines ("design patterns") that can be used across the pilots of the project and can be applied to other domains. The pipelines are one of the major exploitable assets of DataBio.

The generic sections of the deliverable are concluded by Chapter 4 that explains the integration of different components into a pipeline and the services that are provided per pilot. The main results for the pilot services and the component updates, from a technological aspect, for both trials 1 and 2 are presented. The concluding chapter outlines the main findings, lessons learned and emerging examples of best practices.

The deliverable comprises contributions from the following tasks:

- T4.1: DataBio Architecture Requirements
- T4.2: Advanced Visualization Services
- T4.3: Predictive Analytics and Machine Learning
- T4.4: Real-time Analytics and Stream Processing
- T4.5: Big Data Variety Management, Storage, Linked Data and Queries
- T4.6: Big Data Acquisition and Curation with Security/Privacy Support
- T5.1: EO Subsystem and Components
- T5.2: EO Data Discovery and Data Management & Acquisition Services
- T5.3: EO Data Processing, Extraction, Conversion and Fusion Services
- T5.5: Meteo Data Management

Deliverable Leader:	EXUS
Contributors:	Maria Plakia (EXUS) Konstantinos Roussopoulos (EXUS) Stefanou Hara (EXUS) Javier Hitado Simarro (ATOS) Miguel Angel Esbri Palomares (ATOS) Caj Södergård (VTT) Pekka Siltanen (VTT) Jarmo Kalaoja (VTT) Ephrem Habyarimana (CREA) Baldur Kubo (CYBER) Ivo Senner (Fraunhofer) Fabiana Fournier (IBM) Arne Berre (SINTEF) Aphrodite Tsalgatidou (SINTEF) Yves Coene (Spacebel) Per Gunnar Auran (SINTEF Fishery) Michal Kepka (UWB) Karel Charvat (LESPRO) Karel Charvat jr (LESPRO) Savvas Rogotis (NP) Stamatis Krommydas (NP)
Reviewers:	Tomas Mildorf (UWB) Yves Coene (Spacebel) Christian Zinke-Wehlmann (infAI) Amit Kirschenbaum (infAI) Iason Kastanis (CSEM)
Approved by:	Athanasios Poulakidas (INTRASOFT)

Document History			
Version	**Date**	**Contributor(s)**	**Description**
0.1	4/10/2019	EXUS	Table of contents (ToC) + pipeline template
0.2	4/11/2019	ATOS	Trial 2 updates and results template
0.3	13/11/2019	Pilot leaders, WP4 and WP5	ToC update and assignments
0.4	22/11/2019	WP4, WP5	Generalized pipeline template + components initial descriptions
0.5	25/11/219	EXUS, ATOS	Trial 1 and 2 initial information

0.6	1/12/2019	Pilot leaders, WP4 and WP5	Pipeline descriptions according to template
0.7	6/12/2019	Pilot leaders, WP4 and WP5	Pipeline descriptions revision
0.8	12/12/2019	Pilot leaders, WP4 and WP5	Pipeline descriptions revision
0.9	13/12/2019	EXUS	Submission for internal review
0.95	18/12/2019	EXUS	Update after internal review
0.96	19/12/2019	SINTEF, EXUS, SPACEBEL	Update after internal review
0.97	20/12/2019	VTT	Update after internal review
0.98	27/12/2019	SINTEF	Update on datasets and confidential data handling
1.0	30/12/2019	INTRASOFT	Final version for submission

Table of Contents

Table of Figures

List of Tables

Definitions, Acronyms and Abbreviations

Acronym	Title
ADES	Application Deployment and Execution Service
API	Application Programming Interface
BDVA	Big Data Value Association
CEOS	Committee on Earth Observation Satellites
CEP	Complex Event Processing
CSV	Comma Separated Values
DIAS	Data and Information Access Services
DOI	Digital Object Identifier
EC	European Commission
EO	Earth Observation
EOEP	EO Exploitation Platform
EOEPCA	EO Exploitation Platform Common Architecture
ESA	European Space Agency
GEO	Group on Earth Observation
GUI	Graphical User Interface
INSPIRE	Infrastructure for Spatial Information in Europe
ICT	Information and Communication Technology
IoT	Internet of Things
ISO	International Organisation for Standardisation
JSON	JavaScript Object Notation
MPC	Multi Party Computation
NASA	National Aeronautics and Space Administration
OGC	Open Geospatial Consortium
OWL	Web Ontology Language
PPP	Public-Private Partnership
PROTON	PROactive Technology ONline
RAD	Rapid Application Development
RDF	Resource Description Framework
SGX	Software Guard Extensions
SME	Small – Medium Enterprise
SPARQL	SPARQL Protocol and RDF Query Language
SVM	Support Vector Machine
TEP	Thematic Exploitation Platform
TRL	Technology Readiness Level
USGC	United States Geological Survey
W3C	World Wide Web Consortium
WCS	Web Coverage Service
WMS	Web Map Service

WP	Work Package
XML	eXtensible Markup Language

Term	Definition
Dataset	Identifiable collection of data. In the EO Community, a dataset is typically called a "collection" or sometimes a "product".
Sentinel-1	The Copernicus Sentinel-1 earth observation mission developed by ESA provides continuity of data from ERS and Envisat missions, with further enhancements in terms of revisit, coverage, timeliness and reliability of service. The SENTINEL-1 mission comprises a constellation of two polar orbiting satellites, operating day and night performing C-band synthetic aperture radar imaging, enabling them to acquire imagery regardless of the weather. The two-satellite constellation offers a 6 days revisit time. A summary of mission objectives is: • Monitoring sea ice zones and the Arctic environment, and • surveillance of marine environment; • Monitoring land surface motion risks; • Mapping of land surfaces: forest, water and soil; • Mapping in support of humanitarian aid in crisis situations; • Spatial Resolution: 5m, 20m, 40m. Source: Wikipedia and Sentinel Online Web site (https://sentinels.copernicus.eu).
Sentinel-2	The Copernicus Sentinel-2 earth observation mission developed by ESA provides continuity to services relying on multi-spectral high-resolution optical observations over global terrestrial surfaces. Sentinel-2 sustains the operational supply of data for services such as forest monitoring, land cover changes detection or natural disasters management. The Sentinel-2 mission offers an unprecedented combination of the following capabilities: • Multi-spectral information with 13 bands in the visible, near infra-red and short wave infra-red part of the spectrum; • Systematic global coverage of land surfaces: from 56°South to 84°North, coastal waters and all Mediterranean Sea; • High revisit: every 5 days at equator under the same viewing conditions; • High spatial resolution: 10m, 20m and 60m; • Wide field of view: 290 km. (https://sentinels.copernicus.eu)
Sentinel-3	The Copernicus Sentinel-3 earth observation mission developed by ESA main objective is to measure sea-surface topography, sea- and land surface temperature and ocean- and land-surface colour. A pair of Sentinel-3 satellites will enable a short revisit time of less than two days for OLCI instrument and

	less than one day for SLSTR at the equator. Mission objectives are: • Measure sea-surface topography, sea-surface height and significant wave height; • Measure ocean and land-surface temperature; • Measure ocean and land-surface colour • Monitor sea and land ice topography; • Sea-water quality and pollution monitoring; • Inland water monitoring, including rivers and lakes; • Aid marine weather forecasting with acquired data; • Climate monitoring and modelling; • Land-use change monitoring; • Forest cover mapping; • Fire detection; • Weather forecasting; • Measuring Earth's thermal radiation for atmospheric applications. The Sentinel-3A mission has now reached the full operational capacity and preparations for Sentinel-3B launch is-going (mission status on 6 December 2017). Sources: Wikipedia and Sentinel Online Web site (https://sentinels.copernicus.eu).
LANDSAT-8	Landsat 8 is an American EO satellite launched on February 11, 2013, being the eighth satellite in the Landsat program; and the seventh to reach orbit successfully. Originally called the LDCM, it is a collaboration between NASA and the USGS. NASA Goddard Space Flight Center in Greenbelt, Maryland, provided development, mission systems engineering, and acquisition of the launch vehicle while the USGS provided for development of the ground systems and will conduct on-going mission operations. Landsat 8 consists of three key mission and science objectives: • Collect and archive medium resolution (30-meter spatial resolution) multispectral image data affording seasonal coverage of the global landmasses for a period of no less than 5 years; • Ensure that Landsat 8 data are sufficiently consistent with data from the earlier Landsat missions in terms of acquisition geometry, calibration, coverage characteristics, spectral characteristics, output product quality, and data availability to permit studies of landcover and land-use change over time; • Distribute Landsat 8 data products to the general public on a non-discriminatory basis at no cost to the user.
Proba-V	PROBA-V is a small satellite, assuring the succession of the Vegetation instruments on board the French SPOT-4 and SPOT-5 Earth observation missions. PROBA-V was initiated by the Space and Aeronautics department of the BELgian Science Policy Office. It is built by QinetiQ Space N.V. and operated

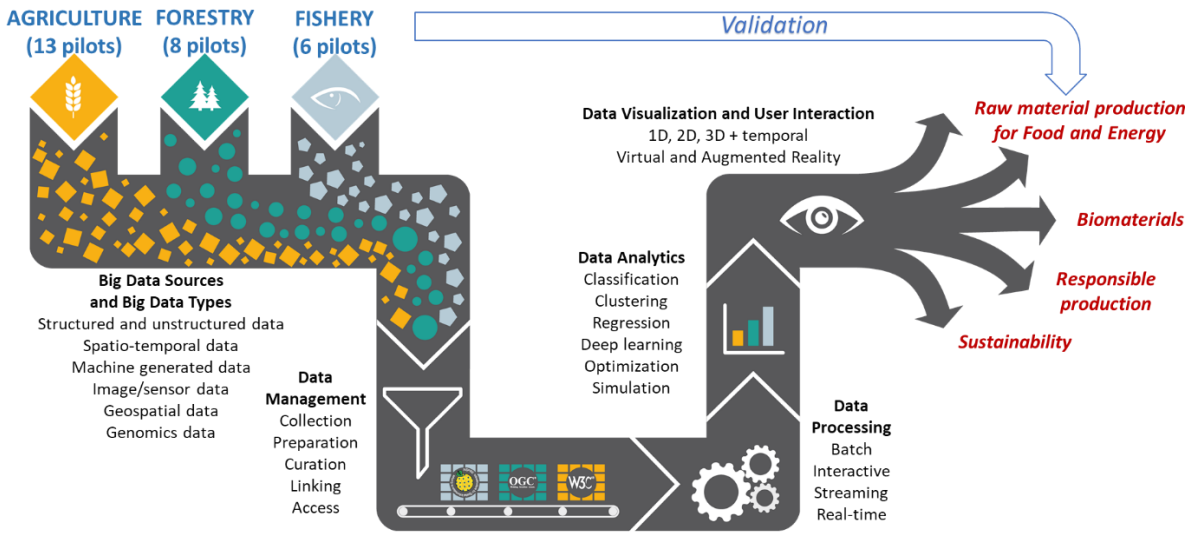
	by ESA and uses a PROBA platform. PROBA-V will support applications such as land use, worldwide vegetation classification, crop monitoring, famine prediction, food security, disaster monitoring and biosphere studies. The mission was originally conceived as a "gap filler" between the SPOT-5 end-of-life (foreseen mid-2014) and the launch of the constellation of the Sentinel-3A and -3B satellites. Due to delays of the Sentinel programme and because some instrument specifications of the Sentinel3 satellites have meanwhile changed, PROBA-V no longer is a gap filler mission but will assure the continuation of the Vegetation programme as such. Source: Wikipedia and Proba-V website (http://proba-v.vgt.vito.be/content/welcome-proba-v-website)
CMEMS SeaLevel	The Copernicus Marine Environment Monitoring Service (CMEMS) is part of the EU's Copernicus Earth observation programme. It is operated by the French centre of global ocean analysis and forecasting, Mercator Océan. CMEMS has been designed to respond to issues emerging in the environmental, business and scientific sectors. Using information from both satellite and in situ observations, it provides daily state-of-the-art analyses and forecasts, which offer an unprecedented capability to observe, understand and anticipate marine environment events. CMEMS is in charge of the processing and distribution of the Sea Level Anomaly (SLA-H) and Absolute Dynamic Topography Heights (ADT-H) in near-real-time product and the Sea Level Anomalies and Absolute Dynamic Topography Heights in delayed-time product (formerly distributed by Aviso+, no change in the scientific content). Source: CMEMS website (http://marine.copernicus.eu)
FOODIE	Farming ontology provides an application vocabulary covering different categories of information dealt by typical farm management tools/apps for their representation in semantic format, and in line with existing standards and best practices (INSPIRE, ISO/OGC standards).
SOSA/SSN	The Semantic Sensor Network (SSN) ontology is an ontology for describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties, as well as actuators. SSN follows a by including called SOSA (Sensor, Observation, Sample, and Actuator) for its elementary classes and properties.
RDF Data Cube Ontology	Data Cube Vocabulary and its SDMX ISO standard extensions are able to publish multi-dimensional data, such as statistics, on the web in such a way that it can be linked to related datasets and concepts. The Data Cube vocabulary is a core foundation which supports extension vocabularies to enable publication of other aspects of statistical data flows or other multi-dimensional datasets.

1 Introduction

1.1 Project Summary

DataBio (Data-driven Bioeconomy) is a H2020 lighthouse project focusing on utilizing Big Data to contribute to the production of the best possible raw materials from agriculture, forestry, and fishery/aquaculture for the bioeconomy industry in order to produce food, energy and biomaterials, also taking into account responsibility and sustainability issues.

DataBio has deployed state-of-the-art Big Data technologies taking advantage of existing partners' infrastructure and solutions. These solutions aggregate Big Data from the three identified sectors (agriculture, forestry, and fishery) and intelligently process, analyse and visualize them. The DataBio software environment allows the three sectors to selectively utilize numerous software components, pipelines and datasets, according to their requirements. The execution has been through continuous cooperation of end-users and technology provider companies, bioeconomy and technology research institutes, and stakeholders from the EU´s Big Data Value PPP programme.



DataBio has been driven by the development, use and evaluation of 27 pilots, where also associated partners and additional stakeholders have been involved. The selected pilot concepts have been transformed into pilot implementations utilizing co-innovative methods and tools. Through intensive matchmaking with the technology partners in DataBio, the pilots have selected and utilized market-ready or near market-ready ICT, Big Data and Earth Observation methods, technologies, tools, datasets and services, mainly provided by the partners within DataBio, in order to offer added-value services in their domain.

Based on the developed technologies and the pilot results, new solutions and new business opportunities are emerging. DataBio has organized a series of stakeholder events, hackathons and trainings to support result take-up and to enable developers outside the consortium to design and develop new tools, services and applications based on the DataBio results.

The DataBio consortium is listed in Table 1. For more information about the project see www.databio.eu.

Table 1: The DataBio consortium partners

Number	Name	Short name	Country
1 (CO)	INTRASOFT INTERNATIONAL SA	**INTRASOFT**	Belgium
2	LESPROJEKT SLUZBY SRO	**LESPRO**	Czech Republic
3	ZAPADOCESKA UNIVERZITA V PLZNI	**UWB**	Czech Republic
4	FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V.	**Fraunhofer**	Germany
5	ATOS SPAIN SA	**ATOS**	Spain
6[1]	STIFTELSEN SINTEF	**SINTEF ICT**	Norway
7	SPACEBEL SA	**SPACEBEL**	Belgium
8	VLAAMSE INSTELLING VOOR TECHNOLOGISCH ONDERZOEK N.V.	**VITO**	Belgium
9	INSTYTUT CHEMII BIOORGANICZNEJ POLSKIEJ AKADEMII NAUK	**PSNC**	Poland
10	CIAOTECH Srl	**CiaoT**	Italy
11	EMPRESA DE TRANSFORMACION AGRARIA SA	**TRAGSA**	Spain
12	INSTITUT FUR ANGEWANDTE INFORMATIK (INFAI) EV	**INFAI**	Germany
13	NEUROPUBLIC AE PLIROFORIKIS & EPIKOINONION	**NP**	Greece
14	Ústav pro hospodářskou úpravu lesů Brandýs nad Labem	**UHUL FMI**	Czech Republic
15	INNOVATION ENGINEERING SRL	**InnoE**	Italy
16	Teknologian tutkimuskeskus VTT Oy	**VTT**	Finland
17	SINTEF FISKERI OG HAVBRUK AS	**SINTEF Fishery**	Norway
18	SUOMEN METSAKESKUS-FINLANDS SKOGSCENTRAL	**METSAK**	Finland
19	IBM ISRAEL - SCIENCE AND TECHNOLOGY LTD	**IBM**	Israel
20	WUUDIS SOLUTIONS OY[2]	**MHGS**	Finland
21	NB ADVIES BV	**NB Advies**	Netherlands
22	CONSIGLIO PER LA RICERCA IN AGRICOLTURA E L'ANALISI DELL'ECONOMIA AGRARIA	**CREA**	Italy
23	FUNDACION AZTI - AZTI FUNDAZIOA	**AZTI**	Spain
24	KINGS BAY AS	**KingsBay**	Norway
25	EROS AS	**Eros**	Norway
26	ERVIK & SAEVIK AS	**ESAS**	Norway
27	LIEGRUPPEN FISKERI AS	**LiegFi**	Norway
28	E-GEOS SPA	**e-geos**	Italy
29	DANMARKS TEKNISKE UNIVERSITET	**DTU**	Denmark

[1] Replaced by partner 49 as of 1/1/2018.
[2] Formerly MHG SYSTEMS OY. Terminated on 27/9/2019.

30	FEDERUNACOMA SRL UNIPERSONALE	**Federu**	Italy
31	CSEM CENTRE SUISSE D'ELECTRONIQUE ET DE MICROTECHNIQUE SA - RECHERCHE ET DEVELOPPEMENT	**CSEM**	Switzerland
32	UNIVERSITAET ST. GALLEN	**UStG**	Switzerland
33	NORGES SILDESALGSLAG SA	**Sildes**	Norway
34	EXUS SOFTWARE LTD	**EXUS**	United Kingdom
35	CYBERNETICA AS	**CYBER**	Estonia
36	GAIA EPICHEIREIN ANONYMI ETAIREIA PSIFIAKON YPIRESION	**GAIA**	Greece
37	SOFTEAM	**Softeam**	France
38	FUNDACION CITOLIVA, CENTRO DE INNOVACION Y TECNOLOGIA DEL OLIVAR Y DEL ACEITE	**CITOLIVA**	Spain
39	TERRASIGNA SRL	**TerraS**	Romania
40	ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS	**CERTH**	Greece
41	METEOROLOGICAL AND ENVIRONMENTAL EARTH OBSERVATION SRL	**MEEO**	Italy
42	ECHEBASTAR FLEET SOCIEDAD LIMITADA	**ECHEBF**	Spain
43	NOVAMONT SPA	**Novam**	Italy
44	SENOP OY	**Senop**	Finland
45	UNIVERSIDAD DEL PAIS VASCO/ EUSKAL HERRIKO UNIBERTSITATEA	**EHU/UPV**	Spain
46	OPEN GEOSPATIAL CONSORTIUM (EUROPE) LIMITED LBG	**OGCE**	United Kingdom
47	ZETOR TRACTORS AS	**ZETOR**	Czech Republic
48	COOPERATIVA AGRICOLA CESENATE SOCIETA COOPERATIVA AGRICOLA	**CAC**	Italy
49	SINTEF AS	**SINTEF**	Norway

1.2 Document Scope

This deliverable describes how technologies (software components, datasets, pipelines) are used in the pilots in agriculture (WP1), forestry (WP2) and fishery (WP3). So-called generalised pipelines used in multiple pilots are laid out. Furthermore, the deliverable describes the components used in each pilot, the provided services, the changes made for Trial 2 of the pilots and lessons learned during Trial 1 and Trial 2.

This deliverable is the final outcome of the two trials for the DataBio pilots (WP1, WP2 and WP3) and the updates of technologies (WP4). Moreover, it concludes the activities and related outcomes of Earth Observation services (WP5).

1.3 Relation with other documents

The software environment developed in DataBio was described in public Deliverables D4.1, D4.2, D4.3 (WP4) and D5.1, D5.2 D5.3 (WP5). All the reports can be found at http://www.databio.eu. Deliverables D4.1-3 defined the Milestone M7 Service ready for Trial 1, whereas Deliverables D5.1-3 defined the Milestone M9 EO Services ready for integration. The platform services and pipelines have been in trials since April 2018 (M16).

Deliverable D4.2 *Services for tests* provides an overview of the component pipelines as identified at month 16 (M16) of the project. It also provides guidelines for the successful implementation and deployment of the pipelines.

Deliverable D4.3 *Data sets, formats and models* were submitted at the end of August 2018. While the two earlier reports deal with software modules, this report focused on the datasets and streams employed in DataBio. Data formats, standards and models were enabling easy findability, access, interoperability, and reusability of data (FAIR principle).

Deliverable D5.1 *EO component specification* includes an analysis of the EO dataset and component related requirements provided by the pilots. It was published at the end of 2017 and contains an overview of best practices of EO access and initial component and dataset requirements based on the DataBio pilot needs.

Deliverable D5.2 *EO component and interfaces* describe, building on D5.1, the Earth Observations component pipelines similarly as D4.2 does for IoT components. It also includes examples of data experimentations with the pipelines.

Deliverable D5.3 EO *services and tools* builds on D5.1 and D5.2 and describes how the technical components from DataBio can be scaled-up to services and tools that are installed as Software as a Service (SaaS) or on-premise. It further provides the information on how and under which conditions these services and tools can be externally accessed.

This public deliverable D4.4 also uses content from the internal deliverables D4.i3 – ” Technology Description for Trial 2” together with the D4.i4 – “Results from Trial 1”, which served as a basis for supporting pilots and components partners to continue with the implementation and deployment of technologies for Trial 2.

1.4 Document Structure

This document is comprised of the following chapters:

Chapter 1 presents an introduction to the project and the document.

Chapter 2 introduces the DataBio technology.

Chapter 3 presents the generalized pipelines in DataBio that can be used across the pilots of the project and can be applicable to other domains.

Chapter 4 provides an overview of the specific pipelines of each pilot in DataBio.

Chapter 5 describes the lessons learned.

The document includes three appendices: **Appendix A** presents a classification of the DataBio components, **Appendix B** presents the DataBio components that are used in each pilot and **Appendix C** presents the benefits from OGC Testbed.

2 DataBio Technology

2.1 DataBio Platform Architecture

As described in D4.1 [REF-01], we understand the concept of a platform in a strictly technical sense as a *software development platform*. With this, we refer to an environment in which a piece of software is developed to be deployed in hardware, virtualized infrastructure, operating system, middleware or a cloud. More specifically, we focus on Big Data platforms that deal with *Big Data* i.e., high volume, high velocity and high variety.

DataBio provides a Big Data *toolset*, which offers functionalities primarily for *services* in the domains of agriculture, forestry and fishery. The functionalities enable new software *components* to be easily and effectively combined with open-source, standards-based Big Data, and proprietary components and infrastructures based on the use of generic and domain-specific components.

The DataBio toolset supports the forming of reusable and deployable *pipelines* of interoperable components (mostly provided by partners), thus extending the impact of DataBio to new bioeconomy projects as well as to other business areas.

DataBio platform consists of a development environment, software components used and developed by DataBio partners and pipelines connecting the components to services. This chapter provides an overview of the update made to the platform since D4.1 was published.

While writing D4.1, we had identified 90 components that could be used in the pilots. Of all the components, 38 of 90 were used by the pilots at that time. In the second trial, 62 of the components offered by the partners are used in one or more of the pilots. As predicted in D4.1, most of the components offered by the DataBio partners are being used in practical pilot applications and at least one component is used from each component provider.

Figure 1 shows a summary of how the components used in the second Trial are classified according to BDVA classification. A detailed list of all the categories and the components that fall into each class is described in Appendix A.

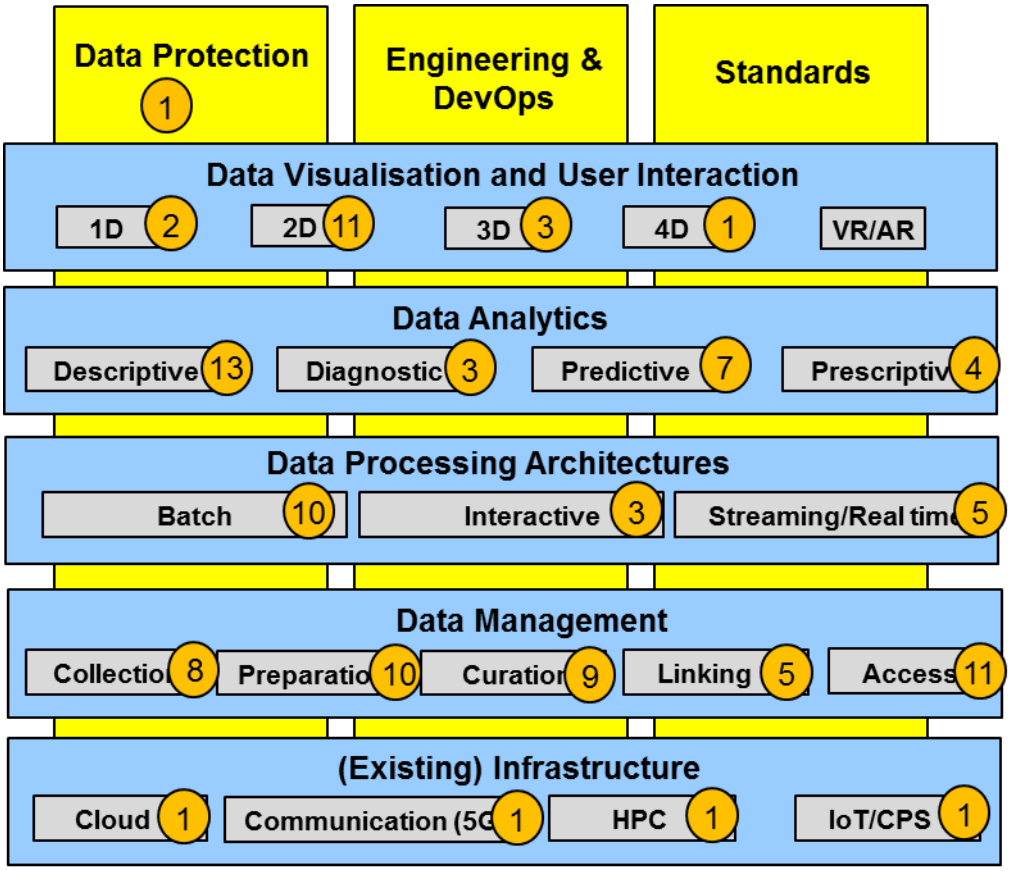


Figure 1: BDVA Reference Architecture: number of DataBio components in each class in Trial 2.

Note that many of the components fall into several categories that are all described in Figure 1. This is because many of the components have several functions: for example, a visualisation software can support both 2D and 3D visualisations.

Each of the pipeline components is used in 1 - 6 pilots, on average in 1.8 pilots. Additionally, the Digital service hub (DataBioHub) and Modelio BA Data modelling tool have been used by all the pilots for information modelling and delivery.

The components have been heavily developed based on pilot demands as shown in the table below. The average rise of the TRL level of the component is expected to be 2.7 until the end of the project.

Table 2: Component developments during Trial 1 and 2

	Trial 1	Trial 2
New User Interfaces	9	5
New APIs	31	28

2.2 DataBio Software Components

DataBio provides a Big Data toolset that offers functionalities primarily for services in the domains of agriculture, forestry and fishery. The functionalities enable new software components to be easily and effectively combined with open-source, standards-based Big Data, and proprietary components and infrastructures based on generic and domain-specific components.

All the software components that DataBio partners provided for the project are described in the deliverables D4.1 Platform and Interfaces [REF-01] and D5.2 EO Components and Interfaces [REF-05]. These deliverables describe the components from the technological point of view and classify the components according to the BDV Reference Model. Figure 2 shows the classification of each of the components according to the BDV Reference Model. A detailed list of all the categories and the components that fall into each category is described in Appendix A.

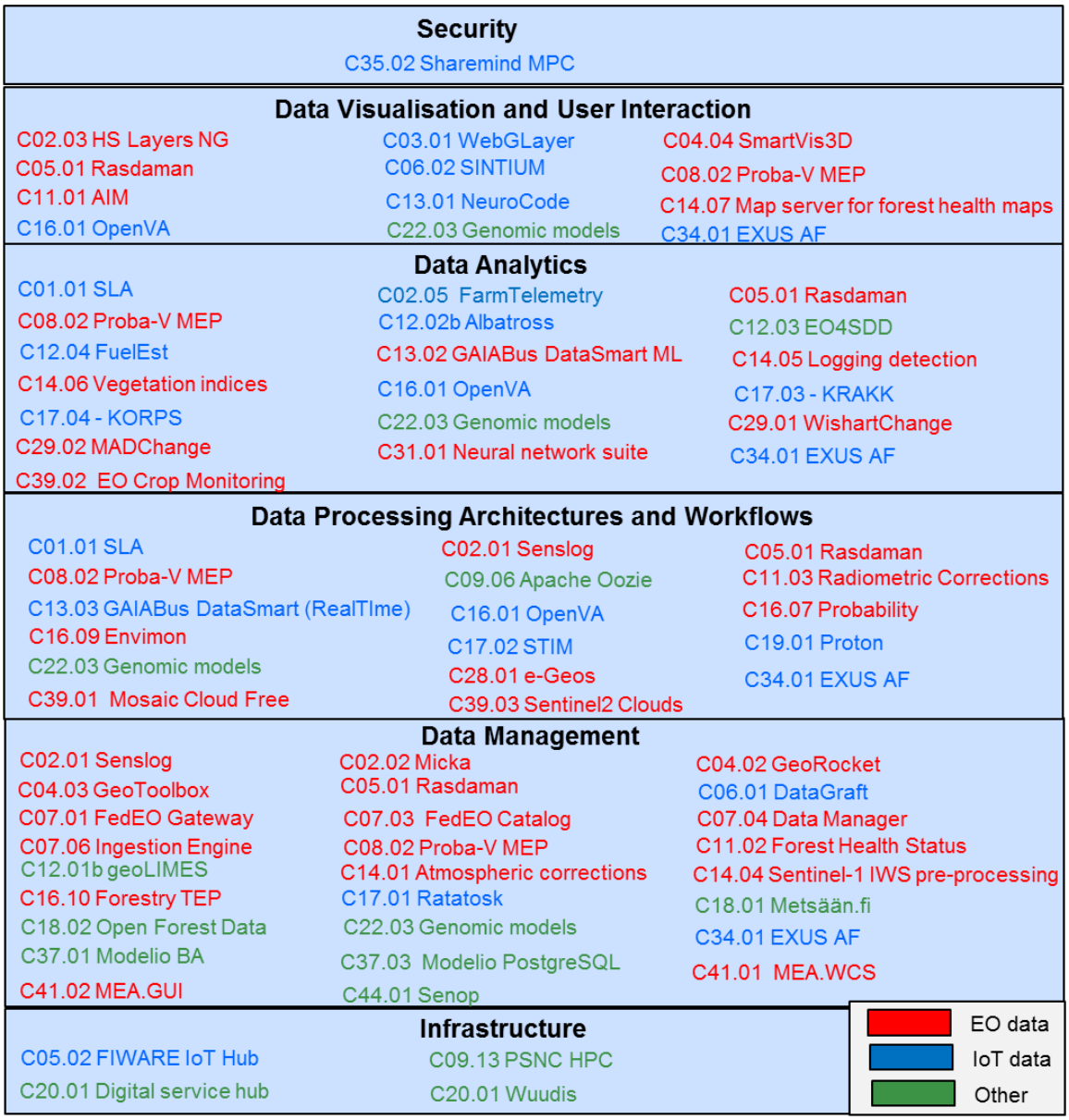


Figure 2: Classification of DataBio components according to the BDVA Reference Model.

In this section an overview of how the components support the different bio-economy pilots are given. Basically, there are two types of components: those that are specific to one pilot domain (agriculture, fishery, forestry) and those that can be used in different domains. Out of 62 components, 19 are domain-specific, while the other 43 can be used in different domains. In practice, the component providers have been concentrating their efforts on supporting pilots in one domain in most cases. However, there are 10 components that are used in at least two domains.

Figure 3 visualizes which components support the DataBio domains. In the middle are the components that are used in several domains. The second layer contains those components that are used in one domain in DataBio but are not domain-specific. The domain-specific components are located at the edges of the image. Most of the components handle mainly

IoT or EO data, which is color-coded in the figure. The components classified as "other" may handle both IoT and EO data, or other data types, such as genomic data.

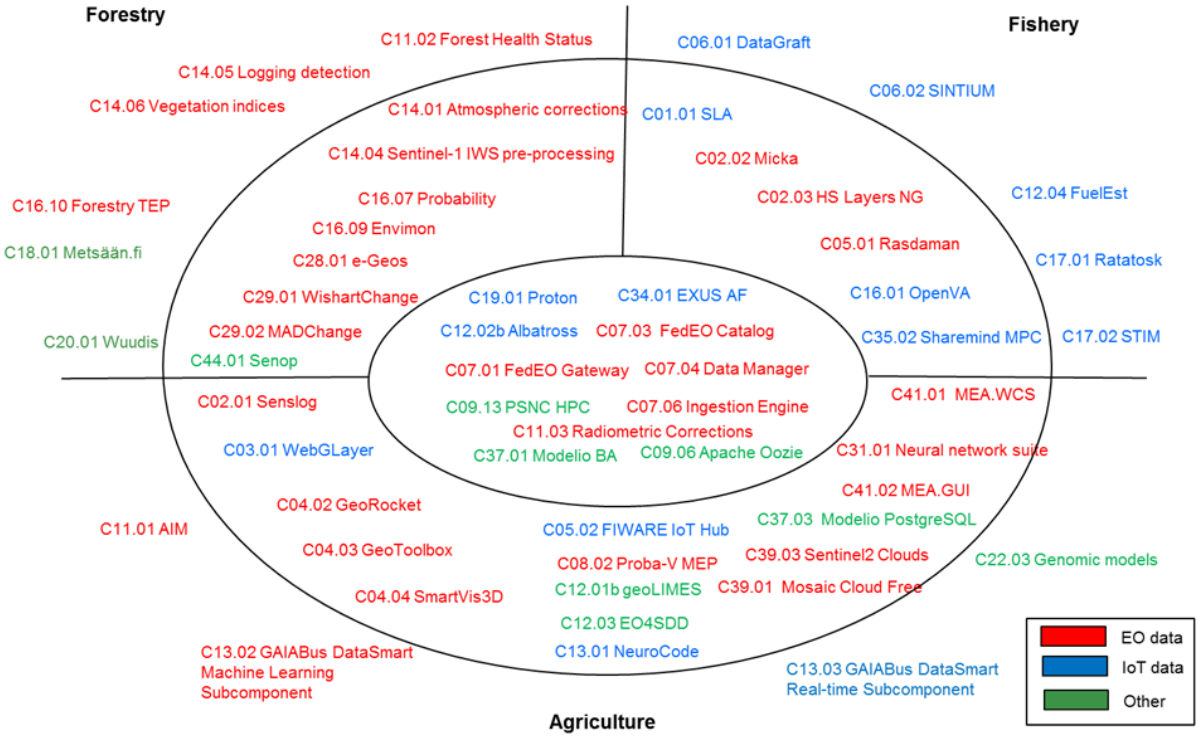


Figure 3: DataBio components used in different bio-economy domains

2.2.1 DataBio Component Descriptions

As stated in DataBio *Description of Action*, the main objective of work packages 4 and 5 was "configuration and adaptation of Big Data technologies in order to enable data-driven innovation with new applications [...] for Agriculture, Forestry and Fishery" and "support to Pilots". In order to do this, the project partners added new features to their software components and configured them based on the pilot needs. All 90 components that the component providers offered for the project have been described in deliverables D4.1 [REF-01] and D5.2 [REF-05]. In order to avoid unnecessary repetition, in this deliverable, we describe only a summary of the new features and configurations that have been done during the project, as well as how the components were used in the pilots.

Table 3: Summary of new features and configurations of the DataBio components used in pilots

Component	New features and usage in the pilots	New interfaces
C01.01 SLA	**Trial 1:** A number of different classification algorithms were deployed and tested to the provided dataset; accuracy and efficiency were measured regarding fish existence identification. The methods tested were Naïve Bayes, k-Nearest Neighbours (kNN) and Support Vector Machine (SVM), both with linear and radial kernels. **Trial 2:** A new bigger dataset was provided by SINTEF and all three methods were applied once again to verify the existence of fish in specific areas. Principal components analysis (PCA) was also examined as a pre-processing method. All classification approaches were tested on MVBS values for different combinations of the five frequencies measured. All methods reach an accuracy of more than 92%. However, the kappa coefficient varies greatly among the different approaches. SVM with radial kernel seems to be more appropriate for the acoustic data, whereas linear kernel totally fails in discriminating the classes of fish presence and fish absence. For all cases, the best results come from using vectors comprising all five frequencies measured.	**Trial 1:** A basic graphical user interface has been implemented, to visualise the main outputs of the classification algorithms and provide information about the presence or absence of fish. **Trial 2:** A user interface for end-users to upload datasets for analyses
C02.01 SensLog	**Trial 1:** Complete redesign and implementation of SensLog core to gain better scalability. Improvement of the data model to reflect scalability features. Dockerization of the solution, testing of Docker deployment. **Trial 2:** Design and implementation of Feeder module and system of Connectors for SensLog. New Feeder reflects the scalability of redesigned SensLog, adds a system of queues to receive data from different sources, mainly in the form of binary protocols over IoT transfer networks. Connectors allow conversion between different APIs and SensLog API. The connector is able to receive or transfer data from endpoints where is not possible to affect local API.	**Trial 1:** Improvement of SensLog API to reflect scalability features. Added filtering mechanism for selecting stored data. **Trial 2:** Implementation of Feeder and Connectors to be able to receive data from other sources.
C02.03 HSLayers	**Trial 1:** Development of a module for visualisation RDF data **Trial 2:** Development of QGIS plugin	**Trial 1:** Interface to RDF data

Component	New features and usage in the pilots	New interfaces
C02.05 Farm Telemetry	**Trial 1:** Complete redesign of the structure as a standalone module with analytics functions. To be separated from SensLog Core functions. **Trial 2:** Design and implementation of the improved version of the FarmTelemetry component. Connecting to SensLog v2 by API, connecting to different telemetry data providers.	**Trial 1:** Simplified API was designed to be able to cooperate with other data sources that contain telemetry data from machinery. **Trial 2:** Interface to communicate with SensLog v2. Integrating data from different telemetry data provides.
C03.01 WebGLayer	**Trial 2:** Features to load the results of live analytics.	**Trial 2**: Loading data from SensLog v2.
C04.02 GeoRocket	**Trial 1:** High-Performance data management for structured geospatial data • Support for multiple storage backends • Fast spatial and property-based queries • Efficient data aggregation queries	**Trial 1:** REST-Interface and Java-API implemented. Simple Query Language for data selection implemented. Simple Aggregation Language for data aggregation implemented.
C04.03 GeoToolBox	**Trial 1:** Improved tools for data preprocessing and optimization for efficient visualization. **Trial 2:** Support for integration of external services	**Trial 1:** REST- and Command Line Interface
C04.04 SmartVis3D	**Trial 1:** Interactive data exploration & visual analytics based on GeoRocket (C04.02) • Interactive data exploration using filters and data-driven colorization • In-depth data inspection of each dataset • Customization visualization of multi-dimensional, data aggregation **Trial 2:** Integrate machine-learning services for improved parcel assessment	**Trial 1:** Web-based graphical user interface **Trial 2:** Interfaces for integrating the machine-learning services

Component	New features and usage in the pilots	New interfaces
C05.01 Rasdaman	**Trial 1:** Component installed and working on PSNC premises. The first set of scripts to insert data have been implemented. **Trials 2:** Latest version of software installed and development of new scripts to insert historical data in netcdf format plus insert vessel data (recently gathered).	
C05.02 FIWARE IoT Hub	**Trial 1:** Necessary software installed in new acquired Raspberry Pi as well as the installation of the necessary software on PSNC cloud infrastructure to facilitate communication with other DataBio components. **Trial 2:** Development of new Python scripts to gather data from sensors.	**Trial 2:** Communication with Proton Component through RESTful NGSI consumers and producers for FIWARE context broker
C06.01 DataGraft	**Trial 1:** DataGraft is a data preparation, curation, transformation, linking and publishing tool. It accepts input in various table formats and CSV, and allows for data cleaning and transformation operations into a new CSV, table form. DataGraft has been extended with Datascraper as a new Python sub-component that fetches, converts and stores data from various web resources into a NoSQL database (CouchDB) or file output	**Trial 2:** Different data source API connectors relevant for the fishery domain have been developed as individual scrapertypes for Datascraper . Used for the Fishery hackathons
C06.02 SINTIUM	**Trial 1**: SINTIUM is a visualization framework for heterogeneous Big Data collections from various sources. The framework consolidates the sources into a single source of data for the creation of coordinated views. The framework has been extended with coordinated views to support different collections of graphs and 2D/3D map interaction in real-time, in particular related to needs by Fishery pilots.	**Trial 2:** SINTIUM is used in web applications visualizing historical and forecasted catch data. Used in particular for the Fishery hackathon demonstrations.
C07.01 FedEO Gateway	**Trial 1:** FedEO Gateway has been extended with an additional connector for the CMEMS web service interface. **Trial 1**, **Trial 2:** Used to retrieve and download products that are utilized as an input of the processing.	

Component	New features and usage in the pilots	New interfaces
C07.03 **FedEO Catalog**	**Trial 1:** Implements an EO catalog server allowing to store EO (satellite) collections (series) and products (datasets) metadata**.** FedEO Catalog has been initially installed in a Fishery pilot. At a later stage, it has been decided to replace it with EO component C05.01 (Rasdaman). **Trial 2:** Extended to comply with OGC Testbed-15 and OGC API Hackaton requirements to host service and application metadata (used as a subcomponent of DataBio Hub for standardized access).	**Trial 2:** Search and metadata interfaces aligned with OGC API - Features and OGC 19-020. Faceted search response added.
C07.04 **Data Manager**	**Trial 1, Trial 2**: Discovery of EO data, download of EO data, preprocessing of downloaded EO data, and registering of the EO data into a cache. Used to retrieve and download products that are utilized as an input of the processing.	
C07.06 **Ingestion Engine**	**Trial 1, Trial 2:** Provides a user interface to access an OpenSearch endpoint implementing a two-step search using OGC 10-032r8 and OGC 13-026r8 specifications. Used to retrieve and download products that are utilized as an input of the processing.	
C08.02 **Proba-V MEP**	**Trial 2:** A first solution to cope with cloudy data inherent to the Sentinel-2 images, VITO has extended the existing services with the possibility to query the number of clouds above an area of interest. This allows pilots to get time-series based information about the percentage of clouds above a field or area of interest. This information can be used in the pilot applications to only show those dates that have relevant and sufficient information. A data fusion algorithm has been developed to cope with cloudy images inherent to the use of Sentinel-2 data. This data fusion uses both Sentinel-1 and Sentinel-2 data to provide cloud-free time series. The backbone infrastructure and services have been optimized to cope with the growing scope of the applications. This enables an increased performance for the pilots using the VITO EO components. A first step towards the creation of a maturity assessment model has been created. The goal of this model is to provide advice to the end-user about the optimal harvesting date.	**Trial 1, Trial 2:** Proba-V MEP offers Virtual Machine interface and a Jupyter Notebook interface, for researchers and developers.

Component	New features and usage in the pilots	New interfaces
C09.13 **PSNC HPC and cloud infrastructure**	**Trial 1, Trial 2:** 1) HPC infrastructure based on GPUs for neural networks modelling for agriculture pilots with Neuropublic: Insurance and CAP support (CSEM); 2) Big Data infrastructure for memory demanding R execution for Forestry scenario with TRAGSA; 3) Cloud infrastructure for deploying RASDAMAN database	
C11.01 **AIM**	**Trial 1, Trial 2:** This service provides information for precision agriculture, mainly based on time series of high resolution (Sentinel-2 type) satellite images, complemented with UAV images and sensor data. The information can be used as input for farm management (operational decisions, tactical decisions). Information layers may include: - Vegetation indices (NDVI, Normalized green red difference index) and derived anomaly maps. This service will offer cost-saving for farmers communities due to better quality management in agricultural zones, especially focused on irrigated crops. Monitoring and managing irrigation policies and agricultural practices will offer meaningful water and energy saving. Besides this, fertilizer control and monitoring can produce, eventually, a prominent economic saving per year and hectare. This better management of hydric and energetic resources is also related to Green-house effect gases reduction, directly linked to better environmental conditions in agriculture. *Data sources* were gathered and processed. *New processing algorithms* were developed.	
C11.02 **Forest Health Status**	**Trial 1, Trial 2:** An EO-based system for monitoring the health of big forest areas was set up (mapping + assessment tools), so authorities will be able to optimise forest management resources.	
C11.03 **Radiometric Corrections**	**Trial 1, Trial 2:** Radiometric improvement of Orthophotos provided by the National Geographic Institute in Spain. This improvement and physical features harmonization (colour, intensity…) allow this data source to be used with similar accuracy than Satellite Images.	

Component	New features and usage in the pilots	New interfaces
C11.014 **IAS Analysis**	The model developed allows estimating AIS invasion risk in Spain, based on a set of factors that greatly influence the geographic pattern and in the degree of invasion. The performance of TRAGSA servers was not enough, as the use of 'R' for producing maps is a very exigent computational process. This has been solved by using an external server thanks to the collaboration with the DataBio partner PSNC.	
C12.01b **geoLIMES (Geo-L)**	**Trial 1:** Speed-up implemented (changing caching-process, download process, mapping-process). Benchmark made (in some cases 5 times faster than earlier version). Error-handling improved (from endpoints) **Trial 2:** Docker image created and tested. More test cases implemented.	**Trial 1:** REST API implemented. **Trial 2:** Full REST API on Docker
C12.02b **Albatross**	**Trial 1:** Flexible data transformation, advanced visualization, flexible data workflows, AI-components (predictions, regression) **Trial 2:** More AI components added. Small EO products included.	**Trial 1:** Visualization user interface **Trial 2:** Visualization user interface expanded.
C12.03 **EO4SDD**	**Trial 1:** Download EO product, train model to detect crop diseases, test model, implement different features to avoid overfitting **Trial 2:** The model expanded; tests expanded.	**Trial 1:** Interface to EO products **Trial 2:** New API.
C12.04 **FuelEsti + ULEI (New subcomponent)**	**Trial 1:** Create feature models for different vessels using a) multivariate regression and b) random forest approaches, predict fuel consumption based on the feature models by given route, including Sentinel 2 data (wind) for prediction. **Trial 2:** Sentinel 2 use enhanced (waves) Documentation and API description. Multi-point routing enabled.	**Trial 1:** API for visualization, REST API **Trial 2:** API expanded.
C13.01 **NeuroCode**	**Trial 1:** Neurocode allows the creation of the main pilot UIs in order to be used by the end-users of the pilot applications.	**Trial 1:** Provided interfaces: F#, Application Interface - Consumed interfaces: XML, PostgreSQL **Trial 2:** Additional pilot UIs explored based on end-user needs

Component	New features and usage in the pilots	New interfaces
C13.02 GAIABus DataSmart Machine Learning Subcomponent	**Trial 1**: Support for EO data preparation and handling functionalities. Support for multi-temporal object-based monitoring and crop type identification. **Trial 2:** New data, features and classification methodologies examined and used taking into account inter-year changes in crop cultivation periods. Integration of agronomic knowledge into the methodological component framework	
C13.03 – GAIABus DataSmart RealTime Subcomponent	**Trial 1:** Real-time data stream monitoring for NP's GAIAtrons Infrastructure installed in all pilot sites. Real-time validation of data and real-time parsing and cross-checking. **Trial 2:** Improved data representation and handling mechanisms, enabling the expansion and/or customized configuration of each GAIAtron station	
C14.01 Atmospheric corrections	**Trial1:** Atmospheric corrections were used in a novel processing chain for automated cloud-free image synthesis based on the analysis of all available Sentinel–2 satellite data for selected sensing period (e.g. the vegetation season from June to August).	
C14.04 Sentinel-1 IWS pre-processing	**Trial 1, Trial 2:** Complex processing chain for satellite data pre-processing and interpretation towards forest health was developed and it was deployed at FMI's infrastructure.	
C14.05 Logging detection	**Trial 1, Trial 2:** Recent sanitary loggings and dead standing wood were identified. Sentinel-2 based maps of forest health were used as one of the inputs to filter all logging activities before 2018.	
C14.06 Time series analysis of Sentinel-2 vegetation indices and biophysical products	**Trial 1, Trial 2:** Selected vegetation indices and image transformations were calculated and their sensitivity against in-situ data from sampled plots was compared. For each dataset, linear regression model between in-situ data and Sentinel–2 indices were calculated and evaluated. For indices yielding best linear fit, a neural network was trained and applied per-pixel to retrieve prediction LAI maps.	
C14.07 Web-based mapserver	**Trial 1, Trial 2:** Maps of retrieved leaf area indices from 2015 to 2018 and their between-year changes are published on FMI's map server.	

Component	New features and usage in the pilots	New interfaces
C16.01 OpenVA	**Trial 1:** New analysis methods that were needed in the pilot were implemented. Better support for Big Data management was implemented because of the pilot requirements. **Trial 2:** New Python analysis backend implemented. VTT OpenVA published as Open Source.	**Trial 2:** Automatic continuous data import, real-time event UI and REST POST interface for receiving messages from Proton
C16.04 Digital Service Hub	**Trial 1:** Migrated to a cloud platform and new website provided for DataBio. User interface revisions for DataBio project. Improved vocabulary for interface description. New template fields for DataBio. **Trial 2:** Vocabularies formalized, partial export for RDF provided.	**Trial 2:** Data exchange interfaces based on new OGC standards.
C16.07 Probability	**Trial 1:** Used for forest parameter estimation, on Forestry TEP (pilot area Hippala, Finland) **Trial 2:** Used for forest parameter estimation, on Forestry TEP (pilot area Galicia, Spain)	
C16.09 Envimon	**Trial 1:** Used for satellite data pre-processing, on Forestry TEP (pilot area Hippala, Finland) **Trial 2:** Used for satellite data pre-processing, on Forestry TEP (pilot area Galicia, Spain)	New approach for producing cloud-free composite data. (The implementation may be a new component.)
C16.10 Forestry TEP	**Trial 1:** Used for implementation of the T2.3.1 'Forest damage remote sensing' processing chain for forest parameter estimates: satellite data acquisition, reference data upload, pre-processing, processing, output storage, delivery. **Trial 2:** Adaptation of the T2.3.1 processing chain for a new type of area (Galicia, Spain) and with new reference data. Use in T2.4.1 in hosting and executing FMI's processing services.	**Trial 1:** Support for the Finnish Forest Information Standard as input data, and provision of a WMS interface for data delivery **Trial 2:** Support for converting raster format data to vector-based (forest stand). Particularly support the Finnish Forest Information Standard (XML) as an output data format, as well as GeoJSON.

Component	New features and usage in the pilots	New interfaces
C17.01 Ratatosk	**Trial 1:** Used for data collection and transfer **Trial 2:** Used for data collection.	
C17.02 STIM	**Trial 1:** Used for data collection and analysis. **Trial 2:** Used for data collection and analysis.	
C17.03 Krakk	**Trial 1:** Used for data collection and explorative analysis. **Trial 2:** Used for data collection, explorative analysis and predictions of fish prices.	
C17.04 Korps	**Trial 2:** Used for enabling stakeholders to perform data analysis in a web browser.	
C18.01 Metsään.fi	**Trial 1:** New version of Wuudis out 25th of Oct 2018 and Case: Wuudis for METSAK forestry advisors. Laatumetsä work quality monitoring app launched in Nov, 2018. Metsään.fi user authentication via Suomi.fi national service architecture portal completed for forestry service providers. **Trial 2:** The data gathered via Wuudis application is utilized in updating METSAK forest resource data. The quality control data is used for updating the Metsään.fi forest resource data. Metsään.fi user authentication via Suomi.fi national service architecture portal for forest owners.	**Trial 1:** The forest resource data system interfaces (standardization, data transfer service, database) **Trial 2:** The forest resource data system interfaces (standardization, data transfer service, database) and processes adjusted accordingly.
C18.02 Open Forest Data	**Trial 1:** Laatumetsä forest damage crowdsourcing app launched in Nov, 2018 Metsään.fi open forest data web service completed. **Trial 2:** The crowdsourced information is received in METSAK via the new data transfer service and in a standardized format. Data for the forest resource data sample plots gathered as field data will be published in Open forest data service.	**Trial 1.** Metsak receives crowdsourced forest damage information via the Laatumetsä mobile application via the spatial interface to METSAK map service. Map service, download service and API's. **Trial 2.** Standardized forest damage message and accordingly updated data transfer interface.

Component	New features and usage in the pilots	New interfaces
		New service for the Open forest data service implemented as well as new dataset for forest resource database.
C19.01 **IBM Proactive Technology Online**	**Trial 1:** Proton CEP engine (generic complex event processing engine used to implement monitoring rules for a number of use cases), Proton adapters used for interfaces with the components (receiving raw events and emitting complex events). Proton dashboard used for displaying situations. Proton CEP application implemented for monitoring events in different pilots: disease and pest status in olives, peaches and grapes; engine parameters for tuna fisheries; primary and auxiliary engine parameters for pelagic fisheries. **Trial 2:** Features in use for the pilots: Proton CEP engine used to implement monitoring rules for a number of use cases. Proton adapters used for interfaces with the components (receiving raw events and emitting complex events) Features Implemented: New dockerized endpoint created. Extended CEP application for tuna fisheries with situation clearing rules. CEP application design for monitoring environment parameters (humidity, temperature)	**Trial 1:** RESTful adapter for pulling sensor reading from RESTful service and RESTful adapter for POSTing complex events to RESTful dashboard File adapter for reading CSV input files and file adapter for writing output files **Trial 2:** RESTful consumer for VTT OpenVA dashboard, data logger interface application sending RESTful messages to Proton RESTful NGSI consumers and producers for FIWARE context broker
C20.01 **Wuudis**	**Trial 1:** New version of Wuudis service was launched on 25th October, 2018. This new version enables easy data sharing and networking particularly between forest owners and forest authority experts. Laatumetsä (developed by Wuudis) work quality monitoring app launched in November, 2018. Laatumetsä (developed by Wuudis) forest damage crowdsourcing app launched in November, 2018. Wuudis tree-wise monitoring minimum viable product (MVP) was launched in June, 2018 and sold to leading forest management association (MHY, Pohjois Karealia, Savotta) in Finland. **Trial 2:** New version of Wuudis service used. Laatumetsä app maintained. Features to analyse the damage data used. Standard WMS and WFS developed for tree-wise monitoring service with VTT (Forestry TEP), Spacebel and FMI. Features	**Trial 1.** Wuudis web service and app was integrated with METSAK database (forest resource data) Laatumetsä app sends work quality control data to METSAK in standardized format via the renewed data transfer service. METSAK receives crowdsourced forest

Component	New features and usage in the pilots	New interfaces
	developed for analyzing different camera images (multispectral, hyperspectral) for boron deficiency recognition	damage information via the Laatumetsä mobile application via the spatial interface to METSAK map service. The results of tree-wise monitoring is delivered as data in Finnish standard format (xml). **Trial 2:** Some interfaces adjusted based on user feedback. Standardized forest damage message and accordingly updated data transfer interface. More interface development as per requirement of METSAK Integrate tree-wise WMS/WFS data into Wuudis. Interface using In Finnish forest standard format (xml) implemented. Integration of different datasets and services with Wuudis
C22.03 Genomic models	**Trial 1, Trial 2:** Used for genomic prediction and selection in CERTH's tomato crops grown in glasshouses. Component detailed, trained and validated in cereals and solanaceae.	
C28.01 eEOPS - e-GEOS EO processing service	**Trial 1, Trial 2:** e-Geos EOPS platform (data fusion service) aggregates the data provided by other components. e-GEOS has implemented a pipeline consisting of several pre-processing steps performed directly on Sentinel-2 and Landsat-8 products, including: • Automated product downloading and archiving • Pre-processing: atmospheric correction and cloud, snow and shadow masking	

Component	New features and usage in the pilots	New interfaces
	• Vegetation index extraction (NDVI, NDMI, EVI, etc.)	
C29.01 WishartChange	**Trial 1, Trial 2:** Change detection for SAR data implemented.	**Trial 1:** Standalone app and command line for Linux and Windows **Trial 2:** UI implementation for Mac
C29.01 MADchange	**Trial 1, Trial 2:** Change detection for optical images implemented.	**Trial 2:** Standalone app and command line for Linux and Windows. **Trial 2:** UI implementation for Mac
C31.01 Neural network suite for image processing	**Trial 1:** Crop analysis from satellite images using ML. Single time point. **Trial 2:** Temporal crop analysis using ML implemented.	**Trial 2:** REST API implemented.
C34.01 EXUS Analytics Framework	**Trial 2:** Weather insurance profile building clustering and feature selection analysis based on satellite and weather measurements, using ML. **Trial 1, 2:** Predict main engine performance and faults in advance applying deep learning.	**Trial 2:** Analytics Dashboard UI and command-line tools for prediction models of the main engine performance and faults in advance.
C35.02 Sharemind MPC	**Trial 2:** A demonstrator, which predicts the best fish catch location and expected catch size on a given day, was developed.	**Trial 2:**. The tool allows input parties to encrypt and import their data. Fisheries can use the tool to train the predictive model and get out of confidential predictions.
C35.03 Sharemind HI	**Trial 2:** The demonstrator, which predicts the best fish catch location and expected catch size on a given day, was developed using trusted execution environments.	**Trial 2:** A web-based interface was developed for the tool.

Component	New features and usage in the pilots	New interfaces
		It allows input parties to encrypt and import their data. Fisheries can use the tool to train the predictive model using their parameters.
C37.01 Modelio BA Data modelling tool	**Trial 1.** Modelio BA Data modeling tool used for all DataBio pilots within Modelio Constellation collaborative environment. **Trial 2:** New Archimate metrics to analyze the created Modelio BA Data modelling tool models developed.	
C37.02 Modelio MongoDB modeller	**Trial 1.** Modelio SQL2NoSQL MongDB Designer used within SensLog pilot. Modelio was used to generate a first Senslog MongoDB from the original PostgreSQL SensLog schema using Modelio MongDB modeller. Test of performances was executed and proved the MongDB to be better suited to SensLog needs. A paper to Data 2018 has been presented on these common achievements.	
C39.01 Mosaic Cloud Free Background - TerraS service	**Trial 1:** The component is deployed on an application server and keeps an up to date collage (mosaic) of Sentinel 1, Sentinel 2 and Landsat 8 images, covering the Romanian territory with the latest, cloud free satellite scenes (in the case of Sentinel 2 and Landsat 8 images). Backgrounds are updated automatically, soon after a new raw scene is available. The whole processing chain is independent and self-content, based on cloud and shadows mask extraction, histogram matching procedures and, finally, a pixel-based analysis. The component has been used in the pre-processing stage of Trial 1 - Pilot C2.1 CAP Support for the AOI situated in Romania. **Trial 2:** Further updates of the background, as soon as new cloud free scenes are available implemented. Integration of the service into the pipeline for the agriculture pilot's AOI prepared.	**Trial 1:** A basic web graphical user interface has been implemented, in order to visualise the main outputs and for the C39.01 component to be supplied as a WMS layer for C39.02 - EO Crop Monitoring. **Trial 2:** As soon as new cloud free scenes are available, the cloud free mosaic backgrounds are uploaded to the developed graphical user interface and also delivered via WMS.
C39.02	**Trial 1:** The service has been tailored to the specific needs of pilot C2.1 - CAP Support. It allows the performing of Big Data analytics to various crop indicators on parcel level, based on	**Trial 1:** A basic web graphical user interface has been

Component	New features and usage in the pilots	New interfaces
EO Crop Monitoring - TerraS service	highly automated algorithms for processing Big Data and relying on multi-temporal series of free and open EO data, with a focus on Copernicus Sentinel-2 data. Trial 1 involved: • the selection of a test area of at least 10.000 km2, situated in southeastern Romania; the selection was performed based on a multi-criteria analysis, taking into account: plots' size, crops' diversity and accessibility; • data collection: EO data – Sentinel 2, Landsat 8, Sentinel 1, for the chosen AOI, in-situ / field data, plots representing farmers' declarations regarding crop types and areas covered; • data preparation (preprocessing the S2 / L8 satellite imagery, reclassifying crop types into crop families, deriving different EO products – e.g.: vegetation indices, band combinations); • data analysis: first test for automatic detection of the potential anomalies in the chosen AOI; performing a pixel-based and a plot-based analysis over the chosen AOI and detecting the anomalies / incongruences between the results and farmers' declarations; • data visualization: designing a way of representing the results of Trial phase 1; assembling of crop family maps, incongruences / potential anomalies / validation maps (pixel-based, plot-based). **Trial 2:** Further development of the Crop Monitoring Service, using the 2018 and 2019 farmers' CAP declarations regarding crop types and areas covered. New farm profile data will be ingested and new Sentinel-2 and Landsat-8 imagery will be processed as soon as they are acquired during the 2019 growing season. New trials, based on a different test area and increasing the number of target crops; testing the algorithms developed using data provided by TRAGSA, for an area of interest situated in Spain. Data collection through on-site visits, in order to evaluate the accuracy of the crop families and crop inadvertencies maps. Visual validation and of Trial 1 results based on Sentinel-2 backgrounds for the test-area. Validation of the results through field visits. Accuracy level computation. Dialog with users / beneficiaries / stakeholders (APIA - the Romanian National Paying Agency). Further optimization of the algorithm based on Trial 1 results. Fine-tuning and further comparisons to the results obtained in Trial 1, in order to prepare the final adjustments for the service. Final optimization of algorithm settings. Final service trials	implemented, in order to visualise the main outputs and for the C39.01 component to be supplied as a WMS layer for C39.02 - EO Crop Monitoring. **Trial 2:** Further updates. Upload of the results of the EO Crop Monitoring Service to the developed graphical user interface and delivery via WMS.

Component	New features and usage in the pilots	New interfaces
C39.03 Sentinel2 Clouds, Shadows and Snow Mask tool	**Trial 1:** The processing chain was developed for Sentinel 2 imagery, based on an in-house formula. The algorithm was intensely tested on many S2 scenes, in all seasons and in various geographical situations. Internal benchmarking shown better performances than other known solution (e.g. fMask or the genuine S2 algorithm). The tool uses unsupervised machine learning techniques for the extraction of the cloud and shadow masks through mutual confirmation. **Trial 2:** Further testing of the algorithm developed by Terrasigna, in various geographical situations. Further comparisons regarding the performances of the developed solution.	**Trial 1, Trial 2:** Stand-alone executable file for Linux environment - command line
C41.01 MEA.WCS	**Trial 1:** The already existing features developed to discover available data collections and to filter them have been exploited to retrieve precipitation data from KNMI meteo data providers. **Trial 2:** The existing component features (data access, data collection, subset...) could be exploited to retrieve further meteo-climate variables, such as temperature.	**Trial1, Trial 2:** The already existing MEA WCS interface has been exploited to access data and to perform basic computing.
C41.02 MEA.GUI	**Trial 1:** The already existing web application Jupyter Notebook implemented through MEA GUI has been exploited to extract precipitation values according to pilot requirements. **Trial 2:** Jupyter Notebook implemented through MEA GUI exploited to perform further data extraction	**Trial 1:** Jupyter Notebook implemented through MEA GUI exploited to perform further data extraction. **Trial 2:** Jupyter Notebook script implementation to retrieve extreme events occurrences and define a classification of such events (risk map) Logically interfaced to EXUS machine learning component
C44.01 Senop	**Trial 2:** SENOP hyperspectral camera was used in Polvijärvi, Finland to identify Boron deficiency in spruce strands.	

2.3 Datasets

This section provides a summary of the datasets in the context of the DataBio platform and pilots. The datasets were identified based on the needs and the requirements of the platform and of the 27 pilots in the domains of agriculture, forestry and fishery. The datatypes of these datasets are quite diverse; they vary from structured data, semi-structured data, or unstructured data to new generation Big Data. The latter include sensor data (IoT data, Drone data, data from hand-held or from mounted optical sensors), machine-generated data (produced by ships, boats and machinery used in agriculture and in forestry), geospatial data (Earth Observation data from various sources and geospatial data from EU, national, local, private and open repositories), and genomics-relevant data categorized as genomic, transcriptomic, phenomic, metabolomic, farm data, in-situ IoT sensors and other environmental datasets, genomic predictions and selection data from plant breeding efforts and biochemical data collected from tomato fruits and sorghum grains. Historical data have also been used in the pilots.

A significant part of the project pilots uses EO (Earth Observation) data, often coming from the Copernicus Sentinel constellation, as input for their specific purposes, in the context of efficient resource use and increasing productivity in agriculture, forestry and fishery, as well as other types of data. In the table below we provide the data types used in four different pilots in order to exemplify their diversity:

Table 4: Data types of pilots A1 and B1.4 in agriculture, B2 in forestry, A2 in fishery

Domain	Name of Pilot	Areas of Interest	Data Used in the Pilot
Agriculture	A1. Precision agriculture in olives, fruits, grapes and vegetables	Greece (Pilot Site A: Chalkidiki - 600 ha, Pilot Site B: Stimagka - 3 000 ha, Pilot Site C: Veria - 10 000 ha)	data directly from the field, collected from a network of telemetric IoT stations called GAIAtrons; remotely with image sensors on in-orbit platforms; and by monitoring the application of inputs and outputs in the farm (e.g. in-situ measurements, farm logs, farm profile)
Agriculture	B1.4 Cereals, biomass crops 4 (Cereal crop monitoring)	8300 ha - Rostenice (Vyskov, Czech Republic); target crops: cereals - winter wheat, spring barley, grain maize	EO data (Landsat 8 -Landsat data repository - (https://espa.cr.usgs.gov), Sentinel 2A/B - (https://scihub.copernicus.eu/), Google Earth Engine platform for fast viewing EO data: (https://earthengine.google.com/), field boundaries from Czech LPIS database as shp or xml (http://eagri.cz/public/app/eagriapp/lpisdata/), orthophotos, topography maps, cadastral maps – as WMS service, farm data - Crop rotation, crop treatments records, yield maps, soil maps

Forestry	B2. Invasive alien species control – plagues – forest management	Spain - the Iberian Peninsula, the Canary Islands and the Balearic Islands	EO data (Sentinel 2, Landsat 8), several alphanumeric Big Data databases - centralized data - WORLDCLIM dataset (provided by the International Journal of Climatology - 19 bioclimatic raster layers with a resolution of 1 km), foreign trade database from Spanish Finance Ministry, Immigration Database by Spanish Statistical Institute, tourism dataset from Ministry of Energy, Tourism and Digital Agenda, GHS - population grid (developed by JRC), Spanish terrestrial transport network (ESRI shp), provided by the National Geographic Institute), NUTS-2, NUTS-3, Municipalities maps from GADM - Global Administrative Areas
Fishery	A2. Small pelagic fisheries immediate operational choices	Small pelagic fishing fleet, covering the North Atlantic Ocean	time-series measurements collected from a variety of sources (power system, navigation system, weather sensors, deck machinery), sonar / hydroacoustic data; EO data evaluated for inclusion in the pilot

The full table of the data types used in the DataBio pilots can be found in the deliverable D4.3 [REF-03].

The datasets have been used in the 27 project pilots and have been classified as follows:

Existing Datasets utilized in DataBio pilots (14 datasets);
Existing Datasets that have been improved in the DataBio project in terms of easier or better findability, accessibility, interoperability or reusability (6 datasets);
New Datasets created by the DataBio project by collecting new data or by combining or processing existing data sources (43 datasets).

In the following we provide a summary of the above datasets in two tables: one table for existing datasets utilized in the project and one table for improved as well as for created datasets.

Table 5: Existing datasets utilized by DataBio Pilots

Name	Responsible or Provider	Reference	Short Description
Open Transport Map	UWB	D03.02	Displays a road network suitable for routing; visualizes average daily Traffic Volumes
Forest resource data	METSAK	D18.01	Forest resource data concerning privately owned Finnish forests;

Name	Responsible or Provider	Reference	Short Description
			consists of basic data of tree stands, growth place data, etc.
Landsat 8 OLI data	NASA & U.S. Geological Survey	https://landsat.gsfc.nasa.gov/landsat-9/	Moderate-resolution measurements of the Earth's terrestrial and polar regions.
Sentinel 3 OLCI data	ESA	https://sentinel.esa.int/web/sentinel/missions/sentinel-3/data-products/olci	Ocean and Land Colour Instrument data
Sentinel 3 SLSTR data	ESA	https://sentinel.esa.int/web/sentinel/missions/sentinel-3/data-products/slstr	Data by the Sea and Land Surface Temperature Radiometer (SLSTR) products
MODIS data	NASA	https://modis.gsfc.nasa.gov/data/	Data concerning the global dynamics of the Earth atmosphere, land, ice and oceans
Proba-V data	Vito	http://proba-v.vgt.vito.be/en	Multispectral images for studying the evolution of the vegetation cover daily and globally.
Global Precipitation Measurement (GPM) mission data	NASA	https://www.nasa.gov/mission_pages/GPM/main/index.html	GPM is an international satellite mission to provide next-generation observations of rain and snow worldwide every three hours
KNMI precipitation data	KNMI Data Centre (KDC)	https://data.knmi.nl/datasets	Weather, climate and seismological datasets of KNMI (Koninklijk Nederlands Meteorologisch Instituut) accessed via the KDC.
CMEMS data	Copernicus	http://marine.copernicus.eu/	Datasets by CMEMS (Copernicus Marine Environment Monitoring Service) about the state of the physical oceans and regional seas.
Sentinel 2A	ESA	D11.01	EO data for multiples geographical areas and various times
Sentinel-2 data	ESA	D14.01, D14.02	EO data
Sentinel 3 SRAL data	ESA	https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-3-altimetry/instrument/sral	Sentinel 3 SRAL (Synthetic Aperture Radar Altimeter) data

Name	Responsible or Provider	Reference	Short Description
Sentinel 3 MWR data	ESA	https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-3-altimetry/instrument/mwr	Sentinel 3 MWR (Microwave Radiometer) data

Table 6: Datasets improved by DataBio and new datasets created during DataBio

Name	Responsible or Data Source - GraphURI	Reference	Description
Canopy height map	FMI	D14.05	Stand age (growth stages) according to canopy height model
gaiasense field	NP	D13.01	Measurements from NP's telemetric IoT agro-climate stations called GAIATrons.
Land use and properties - Greek agriculture pilots	NP	D13.02	Agricultural parcel positions expressed in vectors with several attributes and extracted multi-temporal vegetation indices associated with them.
Customer and forest estate data	METSAK	D18.02	Information on who owns certain forest estates and who has the rights to read and use forest resource data of a certain forest owner.
Orthophotos	IGN	D11.02	Orthophotos provided by Spanish National Geographic Institute
GEOSS Sources	TRAGSA-TRAGSATEC	D11.03	EO data
RPAS data	TRAGSA Group	D11.04	RPAS data and Images
MFE50	TRAGSA-TRAGSATEC	D11.06	Mapa Forestal Español (MFE) - Spanish Forestry Map
Field data - pilot B2	TRAGSA Group	D11.07	Data acquired by IoT Sensors. Scientific data from field samples
Forest damage	FMI	D14.07	In-situ observations of forest damage.
Open Forest Data	METSAK	D18.08	Forest resource data concerning privately owned Finnish forests.

Name	Responsible or Data Source - GraphURI	Reference	Description
Hyperspectral image orthomosaic	Senop	D44.02	Orthorectified hyperspectral mosaic, n-bands, band-matched.
Leaf area index	FMI	D14.06	Leaf area index and canopy closure based on interpretation of digital hemispherical photos
NASA CMR Landsat Datasets via FedEO Gateway	Spacebell	D07.02	Datasets and collections metadata (including Landsat-8 collections) provided by the NASA Common Metadata Repository (CMR).
Gateway Ontology for (Precision) Agriculture	PSNC	D09.01	Representation of data compliant with FOODIE data model in semantic format and their interlinking with established vocabularies and ontologies
Open Land Use	Lespro	D02.01	A composite map intended to create detailed land-use maps of various regions based on certain pan-European datasets enriched by available regional data.
Phenomics, metabolomics, genomics and environmental datasets	CERTH	DS40.01	This dataset includes phenomics, metabolomics, genomics as well as environmental data. Genomic predictions and selection data are also there.
Quality control data	METSAK	D18.04	Datasets consist of estate, sample plot locations, etc.
Sentinels Scientific Hub Datasets via FedEO Gateway	SPACEBEL	D07.01	Sentinel collections and datasets metadata (including product download URL) for the global world.
SigPAC	TRAGSA	D11.05	LPIS - Land parcel identification system
Smart POI dataset	Lespro	D02.01	Seamless and open resource of POIs that is available for all users to download, search or reuse in applications and services
Stand Age Map	FMI	D14.14	Vector layer based on Czech forest management plans and stand age based on detailed forest inventory.

Name	Responsible or Data Source - GraphURI	Reference	Description
Storm and forest damage observations and possible risk areas	METSAK	D18.03a	Consist of location, type of the damage, evaluation of the extent of the damage, tree species and distance from the road.
Forest road condition observations / Roads.ML	METSAK	D18.03b	Consist of location, type of the road based on digiroad map, evaluation of road condition, possible road limitations or obstacles on the road as well as the forest development classes for the road surroundings.
Tree species map	FMI	D14.03	Raster dataset based on classification of Sentinel-2 multi-temporal data and National forest inventory of Czech Republic.
Wuudis data	METSAK	D20.01	different map layers
Fishing ship data	AZTI	https://zenodo.org/record/3563390#.XeicA5MzZhE	One month of IoT data from three fishing ships with measurements from the main and auxiliary engines and from ship sensors.
*Most of the following datasets are **Linked Data** that have been created by transforming and publishing different input datasets from various heterogeneous sources. We would like to note that **the Graph URIs** that appear in some of them, **are not resolvable;** they can be used to refer to the specific dataset in the triplestore.*			
Fishery catch data	Norwegian Directorate of Fisheries (NDF)	https://www.fiskeridir.no/Tall-og-analyse/Aapne-data/Aapne-datasett/Fangstdata-koblet-med-fartoeydata	Data downloaded from NDF and filtered to be used as input for predicting the most likely catch location per day for selected species
Farm Accountancy Data Network (FADN) linked dataset	FADN	https://www.databiohub.eu/registry/#service-view/Farm%20Accountancy%20Data%20Network%20%20(FADN)%20linked%20dataset/0.0.1	Harmonized microeconomic data for agriculture. Input data was in CSV files, and were modeled and aligned by using ontologies like Data Cube Vocabulary and its SDMX ISO extensions.
Polish Land Parcel Information System (LPIS) linked dataset	Graph: http://w3id.org/foodie/open/pl/LPIS/{voivodeship}#	https://www.databiohub.eu/registry/#service-view/Polish%20Land%20Parcel%20Information%20System%20(LPIS)%20linked%20dataset/0.0.1	The LPIS system identifies land use for a country; utilizes aerial photographs and satellite images rendered to extract spatial information. The input data was in the form of geospatial data.

Name	Responsible or Data Source - GraphURI	Reference	Description
Norway fishery stocks linked dataset	Source: http://standardgraphs.ices.dk/StandardGraphsWebServices.asmx/getListStocks?year=0	https://www.databiohub.eu/registry/#service-view/Norway%20fishery%20stocks%20linked%20dataset/0.0.1	Stock data from Norway transformed and published as Linked Data. Data source was the list of all stocks for all years retrieved in XML
FAO water areas classification as linked data	Food and Agriculture Organisation (FAO)	https://www.databiohub.eu/registry/#service-view/FAO%20water%20areas%20classification%20as%20linked%20data/0.0.1	FAO water areas classification, including inland and marine codes, published as linked data. The source of the data is FAO.
Micka (EO metadata)	Lesproject Micka registry	https://www.databiohub.eu/registry/#service-view/Micka%20(EO%20metadata)/0.0.1	Geospatial metadata collected in RDF form and transformed in Linked Data to enable their integration with other datasets.
FedEO gateway as Linked Data	FedEO gateway	https://www.databiohub.eu/registry/#service-view/FedEO%20gateway%20as%20linked%20data/0.0.1	EO metadata accessed through a system called Federated Earth Observation (FedEO) gateway and exposed as Linked Data.
Pilot B2.1 Senslog linked dataset	GraphURI: http://w3id.org/foodie/senslog/	https://www.databiohub.eu/registry/#service-view/Pilot%20B2.1%20Senslog%20linked%20dataset/0.0.1	Data from Pilot 9 [B2.1] Machinery management that are transformed into Linked Data on the fly
FAO ASFIS classification of biological entities	FAO	https://www.databiohub.eu/registry/#service-view/FAO%20ASFIS%20classification%20of%20biological%20entities/0.0.1	Taxonomic classification of biological entities.
Open Transport Map	GraphURI: http://w3id.org/foodie/otm#	https://www.databiohub.eu/registry/#service-view/Open%20Transport%20Map/1.0	A road network for routing which visualizes average daily Traffic Volumes for the whole EU
Pilot B1.4 yield data from Rostenice farm in Czech Republic	LESPRO	https://www.databiohub.eu/registry/#service-view/Pilot%20B1.4%20yield%20data%20from%20Rostenice%20farm%20in%20Czech%20Republic/0.0.1	Linked Data from the farm data which contain information about each field name with the associated cereal crop classifications and arranged by year in the Czech Republic.

Name	Responsible or Data Source - GraphURI	Reference	Description
Pilot B1.4 fields and crop data from Rostenice farm in the Czech Republic	LESPRO	https://www.databiohub.eu/registry/#service-view/Pilot%20B1.4%20fields%20and%20crop%20data%20from%20Rostenice%20farm%20in%20Czech%20Republic/0.0.1	Farm data with the associated cereal crop classifications arranged by year and data about the field boundaries and crop map and yield potential of most of the fields in Rostenice pilot farm.
Corine linked dataset	http://w3id.org/foodie/corine#	https://www.databiohub.eu/registry/#service-view/Corine%20linked%20dataset/0.0.1	Linked data from agriculture-related lands and from main cities in the Czech Republic, Poland and Spain.
Urban atlas linked dataset	http://w3id.org/foodie/atlas#	https://www.databiohub.eu/registry/#service-view/Urban%20atlas%20linked%20dataset/0.0.1	This dataset contains agriculture-related lands and for main cities in the Czech Republic, Poland and Spain
Land parcel dataset (LPIS) from the Czech Republic (CR)	GraphURI: http://w3id.org/foodie/open/cz/pLPIS_180616_WGS#	https://www.databiohub.eu/registry/#service-view/Land%20parcel%20dataset%20(LPIS)%20from%20Czech%20Republic/0.0.1	This dataset contains land parcel and cadastral data collected from the Czech Republic (CR).
Erosion-endangered soil zones linked dataset from the CR	GraphURI: http://w3id.org/foodie/open/cz/Soil_maps_BPEJ_WGSc#	https://www.databiohub.eu/registry/#service-view/Erosion-endangered%20soil%20zones%20linked%20dataset%20from%20Czech%20republic/0.0.1	This data contains the erosion of endangered soil maps from the CR. Source data was in shapefiles which were transformed into Linked Data.
Water buffer linked dataset from CR	GraphURI: http://w3id.org/foodie/open/cz/water_buffer25#	https://www.databiohub.eu/registry/#service-view/Water%20buffer%20linked%20dataset%20from%20Czech%20Republic/0.0.1	Water buffer and water body related data from the CR. Source data was in shapefiles which were transformed into Linked Data.
Norway catch records linked dataset	Norwegian Directorate of Fisheries (NDF)	https://www.databiohub.eu/registry/#service-view/Norway%20catch%20records%20linked%20dataset/0.0.1	Catch records data from Norway (2014-2019), transformed and published as Linked Data.
ISO Country Codes	ISO	https://www.omg.org/spec/LCC/Countries/ISO3166-1-CountryCodes/	ISO Country Codes as Linked Data.

Name	Responsible or Data Source - GraphURI	Reference	Description
ISO Country Subdivision Codes	ISO	https://www.omg.org/spec/LCC/Countries/Regions/ISO3166-2-SubdivisionCodes-NO/	ISO Country Subdivision Codes as Linked Data.
ISO Region Codes	ISO	https://www.omg.org/spec/LCC/Countries/UN-M49-RegionCodes/	ISO Region Codes as Linked Data.

An example of each of the above dataset groups with metadata, as well as the full metadata template that has been used in deliverable D4.3 is shown in the following:

Table 7: Example of existing dataset with metadata utilized by DataBio Pilots: Proba-V data

Field	Value
Internal Name of the Dataset	Proba-V
Name of the Dataset/API Provider	Vito
Short description	The Proba-V mission provides multispectral images to study the evolution of the vegetation cover on a daily and global basis. The 'V' stands for Vegetation. This mission is extending the dataset of the long-established Vegetation instrument, flown as a secondary payload aboard France's SPOT-4 and SPOT-5 satellites launched in 1998 and 2002 respectively. The Proba-V mission has been developed in the frame of the ESA General Support Technology Program (GSTP). The Contributors to the Proba-V mission are Belgium, Luxembourg and Canada.
Extended Description	Proba-V's main applications are related to monitoring plant and forest growth, as well as inland water bodies. The Vegetation instrument can distinguish between different land cover types and plant species, including crops, to reveal their health, as well as to detect water bodies and vegetation burn scars. The VEGETATION instrument is pre-programmed with an indefinitely repeated sequence of acquisitions. This nominal acquisition scenario allows a continuous series of identical products to be generated, aiming to map land cover and vegetation growth across the entire planet every two days.
Geographical Coverage	The mission, developed as part of ESA's Proba Programme, is an ESA EO mission providing global coverage every two days, with latitudes 35-75°N and 35-56°S covered daily, and between 35°N and 35°S every 2 days
Access Mechanism	PROBA-V products can be ordered and downloaded from the PROBA-V Product Distribution Portal (PDP) at http://www.vito-eodata.be/.

	Products are usually available within 24 hours after sensing time (max 48 hours).
URI	https://www.vito-eodata.be/PDF/portal/Application.html#Home

Table 8: Example of dataset with metadata improved by DataBio: RPAS (Remotely Piloted Aircraft Systems) data

Field	**Value**
Internal Name of the Dataset	RPAS
Name of the Dataset/API Provider	TRAGSA
Short description	RPAS data, property of TRAGSA, are provided according to the pilot needs. The images acquired are provided in 6 spectral bands: RGB, Red Edge, NIR, Thermal, as well as point-cloud
Extended Description	The delivery of RPAS imagery started in October 2017 and the areas covered represent small parcels (hectares) within pilot areas in the areas in the Iberian Peninsula - Spain (Extremadura, Andalucia, Castilla y León, Castilla La Mancha, Madrid). RPAS imagery is stored in TRAGSA Premises.
Timespan	From 2017

Table 9: Example of new dataset created during DataBio: Open Forest Data (METSAK - D18.01)

Field	**Value**
Internal Name of the Dataset	D18.01
Name of the Dataset/API Provider	Open Forest Data / METSAK
Short description	The pilot uses METSAK's forest resource data concerning privately owned Finnish forests from METSAK's forest resource data system. The forest resource data consists of basic data of tree stands (development class, dominant tree species, scanned height, scanned intensity, stand measurement date), strata of tree stands (mean age, basal area, number of stems, mean diameter, mean height, total volume, volume of logwood, volume of pulpwood), growth place data (classification, fertility class, soil type, drainage state, ditching year, accessibility, growth place data source, growth place data measurement date), geometry and compartment numbering. The forest resource data is available in a standard format for external use with the consent of a forest owner.
Extended Description	The forest resources are created once in a decade per certain area using remote sensing (airborne laser scanning) and aerial photographs. The new data is analysed and in some parts measured in the field. Other updates on the forest resource data are yearly growth calculations,

	possible notifications of forest use or other forestry operations or so-called Kemera financing operations and possible new aerial photographs to be interpreted.
Version	OGC GeoPackage with 1.2 RTree XML version 1.7
Initial Availability Data	1.3.2018 Download services, Q2/2018 API's
Data Type	Open forest data including forest resource data as well as GIS data
Personal Data	N/A
Rightsholder	METSAK
Dataset/API Owner/Responsible	METSAK Open forest data/ Juha Inkilä
Dataset/API Owner/Responsible Contacts	METSAK Open forest data (Avoin metsätieto)/METSAK / virpi.stenman@metsakeskus.fi
Technology	WMS, WFS and REST
Name of the System	Open forest data (Avoin metsätieto)
Dataset Data Model/API Interface	XML standard/REST, OGC GeoPackage standard / WFS, WMS from Oracle database
Data Model: Standards, Glossaries and metadata standards	https://www.metsatietostandardit.fi/en/
Data Identifier - Standard used	XML, OGC, WFS, WMS, REST
Data Model - Specific Data Model	https://www.bitcomp.fi/metsatietostandardit/
Data Volume	276,8 GB on June 2018
Update Frequency	Daily
Geographical Coverage	Finland
Timespan	From 1.3.2018 onwards
Access Level	Open
URI	https://www.metsaan.fi/rajapinnat

The full metadata template that has been used for describing the metadata in deliverable D4.3 and which can be found in Appendix A of D4.3, is the following:

Field	**Value**
Internal Name of the Dataset	

Name of the Dataset/API Provider	
Short description	
Extended Description	
Version	
Initial Availability Data	
Data Type	
Personal Data	
Rightsholder	
Other Rights Information	
Dataset/API Owner/Responsible	
Dataset/API Owner/Responsible Contacts	
Technology	
Name of the System	
Dataset Data Model/API Interface	
Data Model: Standards, Glossaries and metadata standards	
Data Identifier - Standard used	
Data Model - Specific Data Model	
Data Volume	
Update Frequency	
Data Archiving and preservation	
Geographical Coverage	
Timespan	
Access Level	
Access Mechanism	
URI	

All datasets with metadata have been described in the deliverable D4.3 - Data sets, formats and models [REF-03]. The metadata are very diverse as regards their structure, their encodings, the kinds of resources they describe, their handling as well as publication point of

views. Therefore, the DataBio approach ties data and metadata together (i.e. it ensures updated metadata despite Big Data velocity updates), supports metadata heterogeneity (i.e. enables discovery of static (e.g. datasets) as well as mobile/other resources (e.g., sensors active during agricultural machinery fleet tracking) in a unified platform, uses efficient encodings and integrates metadata in other tools.

Metadata are registered in the DataBioHub (https://www.databiohub.eu/) which is a pivotal tool in the project for data management and data sharing. Specifically, the DataBioHub:

- contains descriptions of DataBio components, pipelines and pilots and their mutual relations;
- makes the DataBio data *findable* as it supports their discovery by i) publishing the metadata according to best practices and standards (geospatial and others), and by ii) applying search keywords (=tags) to the digital objects;
- the actual data are *accessible* through the DataBioHub by consulting the dataset owner;
- contains information about the APIs, the data model and formats as well as about the access methods;
- promotes *interoperability* as the metadata (and many times the data) conform to established standards, e.g. in the Earth Observation (EO) field;
- enables *reusability*, provided that licensing schemes are in place.

Thus, the support of the FAIR principles is materialized in the project through the DataBioHub and according to the Data Management Plan that constitutes Deliverable D6.2 [REF-07] and covers descriptions of the DataBio datasets, data standards, data sharing and long-time preservation of data.

The use of GeoDCAT-AP[3] (and its compacted GeoJSON representation defined in OGC 17-084) is encouraged for the description of datasets, as it is described next in 2.4.2 on Data Sharing, due to the many benefits accruing from this, e.g. increased discoverability of datasets, decentralized publishing, etc.

2.3.1 Public datasets produced and shared by DataBio

A brief description of the public datasets produced and shared by DataBio after the delivery of D4.3 is provided next, while the full list of the created public datasets has been depicted in the respective table at the beginning of this section:

1) Fishing ship data: One month of IoT data from three different fishing ships. Data contains measurements from the main and auxiliary engines from the ships as well as other measurements from the ship sensors. Time-stamped data is measured every ten seconds and it has been anonymized so that the ships cannot be identified.

[3] GeoDCAT is a Geospatial extension to DCAT-AP (DCAT application profile for data portals in Europe).

The dataset has been published in zenodo.org: https://zenodo.org/record/3563390#.XeicA5MzZhE

2) Fishery catch data from the Norwegian Directorate of Fisheries were downloaded and filtered (reduced in size) to be used as input for predicting the most likely catch location per day for selected species demonstrating privacy-preserving computation, please refer to Sections 2.7 and 3.7 for details. The source dataset can be downloaded from the Norwegian Fisheries Directorate through this location: https://www.fiskeridir.no/Tall-og-analyse/Aapne-data/Aapne-datasett/Fangstdata-koblet-med-fartoeydata. For this use-case we focused on datasets for 2016-2018 and species important for whitefish trawling (e.g. north-east arctic cod) to have overlap with the private dataset used for the demo pilot described in Section 3.8.

3) Linked Data datasets: Please refer to Section 3.3.3 for additional information of the full list of linked datasets. Below is given a brief description of the most relevant Linked Datasets published in DataBio.

- **Farm Accountancy Data Network (FADN) linked dataset**: FADN is an instrument for evaluating the income of agricultural holdings and the impacts of the Common Agricultural Policy. FADN consists of an annual survey carried out by the Member States of the European Union. The services responsible in the Union for the operation of the FADN collect every year accountancy data from a sample of the agricultural holdings in the European Union. Derived from national surveys, the FADN is the only source of microeconomic data that is harmonized, i.e. the bookkeeping principles are the same in all countries. Input data for this dataset was in the form of CSV files, which were first modeled and aligned by using ontologies like Data Cube Vocabulary and its SDMX ISO extensions. The source files can be downloaded from https://ec.europa.eu/agriculture/rica/database/consult_std_reports_en.cfm

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/Farm%20Accountancy%20Data%20Network%20%20(FADN)%20linked%20dataset/0.0.1

- **Polish Land Parcel Information System (LPIS) linked dataset:** LPIS is a system to identify land use for a given country. It utilises orthophotos – basically aerial photographs and high precision satellite images that are digitally rendered to extract as much meaningful spatial information as possible. A unique number is given to each land parcel to provide a unique identification in space and time. This information is then updated regularly to monitor the evolution of the land cover and the management of the crops. Polish LPIS data (land parcel and cadastral data), include all the voivodeships. The input data was in the form of geospatial data in shapefile format.

URL in DataBio Hub:

https://www.databiohub.eu/registry/#service-view/Polish%20Land%20Parcel%20Information%20System%20(LPIS)%20linked%20dataset/0.0.1

- **Norway fishery stocks linked dataset:** Stock data from Norway transformed and published as Linked Data. Data source was the list of all stocks for all years retrieved in XML (transformed initially to JSON), collected from service: http://standardgraphs.ices.dk/StandardGraphsWebServices.asmx/getListStocks?year=0.

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/Norway%20fishery%20stocks%20linked%20dataset/0.0.1

- **FAO water areas classification as linked data:** FAO water areas classification, including inland and marine codes, published as linked data. The source of the data is FAO.

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/FAO%20water%20areas%20classification%20as%20linked%20data/0.0.1

- **Micka (EO metadata):** EO metadata collected from the public Lesproject Micka registry (https://micka.lesprojekt.cz/en/), which includes information of over 100K geospatial datasets. Micka is software for spatial data / services metadata management according to ISO, OGC and INSPIRE standards. It allows retrieving the metadata in RDF forms using Geo-DCAT-AP (an extension of DCAT) for the representation of geographic metadata compliant with the DCAT application profile for European data portals. As such metadata cannot be queried as Linked Data, the goal was to make them available in the form of Linked Data to enable their integration with other datasets, e.g., Open Land Use (OLU). The process for publication was straightforward: a dump of all the metadata in RDF format was generated from Micka, which was then loaded into the Virtuoso triplestore. Some exemplary SPARQL queries were then generated to identify connection points for integration, e.g., get OLU entries and their metadata is given a municipal code and type of area (e.g., agricultural lands). The dataset is accessible via: https://www.foodie-cloud.org/sparql.

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/Micka%20(EO%20metadata)/0.0.1

- **FedEO gateway as Linked Data:** This use case for Linked Data for hybrid systems involves the accessing of EO metadata through a system called Federated Earth Observation (FedEO) gateway, which provides interfaces for the access of the EO data. FedEO provides a unique entry point to a growing number of scientific catalogues and services for EO missions. In the case of FedEO, the metadata returned was already using semantic vocabularies in the (Geo)JSON-LD representation, thus it required only to expose the results as Linked Data.

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/FedEO%20gateway%20as%20linked%20data/0.0.1

- **Pilot B2.1 Senslog linked dataset:** Data from Pilot 9 [B2.1] Machinery management in the DataBio project where sensor data from the SensLog service (used by FarmTelemeter service) was transformed into Linked Data on the fly, i.e. data stays at the source and only a virtual semantic layer was created on top of it to access it as Linked Data.

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/Pilot%20B2.1%20Senslog%20linked%20dataset/0.0.1

- **FAO ASFIS classification of biological entities:** Taxonomic classification of biological entities, data is maintained in ASFIS, by FAO. This model implements the ontology design patterns for taxonomic classifications. See http://ontologydesignpatterns.org. This ontology has been made for testing purposes and should not be used as official expressions of any of the classifications systems modelled. Source of the data is from NeOn project, creator: Caterina Caracciolo, KCEW - FA[4]

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/FAO%20ASFIS%20classification%20of%20biological%20entities/0.0.1

- **Open Transport Map:** The Open Transport Map displays a road network which – is suitable for routing –visualizes average daily Traffic Volumes for the whole EU; visualizes time-related Traffic Volumes (in OTN Pilot Cities - Antwerp, Birmingham, Issy-le-Moulineaux, Liberec region). The linked dataset of OTM in PSNC triplestore has the graph: http://w3id.org/foodie/otm#. Presently the OTM linked dataset comprises data from the Czech Republic, Spain and Poland.

[4] http://neon-project.org

RoadLinks only for 'FunctionalRoadClassValue' of type: (‘mainRoad’, ’firstClass’, ’secondClass’, ’thirdClass’, ’fourthClass’)

URL in DataBio Hub:

https://www.databiohub.eu/registry/#service-view/Open%20Transport%20Map/1.0

- **Open Land Use:** Open Land Use Map is a composite map that is intended to create detailed land-use maps of various regions based on certain pan-Europen datasets such as CORINE Landcover, UrbanAtlas enriched by available regional data. The dataset is derived from available open data sources at different levels of detail and coverage. These data sources include: 1) Digital cadastral maps if available 2) Land Parcel Identification System if Available 3) Urban Atlas(European Environmental Agency) 4) CORINE Land Cover 2006 (European Environmental Agency) 5) Open Street Map The order of the data sources is according to the level of detail and, therefore, the priority for data integration. The linked dataset of OLU in PSNC triplestore has graph: http://w3id.org/foodie/olu#. The dataset includes agriculture-related lands (hilucs_code<200) from Czech, Poland, Spain & for few main cities in the Czech Republic (centers of NUTS3 regions), Poland (agglomeration areas from Urban Atlas) and Spain (agglomeration areas from Urban Atlas) Graph for Open Land Use Metadata: http://micka.lesprojekt.cz/catalog/dataset#

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/Open%20Land%20Use/0.0.1

- **Smart POI dataset:** The Smart Points of Interest dataset (SPOI) is the seamless and open resource of POIs that is available for all users to download, search or reuse in applications and services SPOI’s principal target is to provide information as Linked data together with other dataset containing road network. The SPOI dataset is created as a combination of global data (selected points from OpenStreetMap) and local data provided by the SDI4Apps partners or data available on the web. The dataset can be reached by Sparql endpoint (http://data.plan4all.eu/sparql), for detailed information please follow: http://sdi4apps.eu/spoi. The graph in the triple store wherein the linked dataset is stored in PSNC is: http://www.sdi4apps.eu/poi.rdf

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/Smart%20POI%20dataset/0.0.1

- **Pilot B1.4 yield data from Rostenice farm in Czech Republic:** Linked Data from the farm data collected from Czech Pilot 8 [B1.4] Cereals and biomass crops

from the DataBio project, in order to query and access different heterogeneous data sources via an integrated layer. The input datasets used for this experiment include farm data having information about each field name with the associated cereal crop classifications and arranged by year in the Czech Republic.

URL in DataBio Hub:

https://www.databiohub.eu/registry/#service-view/Pilot%20B1.4%20yield%20data%20from%20Rostenice%20farm%20in%20Czech%20Republic/0.0.1

- **Pilot B1.4 fields and crop data from Rostenice farm in the Czech Republic:** This dataset contains farm data having information about each field name with the associated cereal crop classifications and arranged by year and data about the field boundaries and crop map and yield potential of most of the fields in Rostenice pilot farm from the Czech Republic.

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/Pilot%20B1.4%20fields%20and%20crop%20data%20from%20Rostenice%20farm%20in%20Czech%20Republic/0.0.1

- **Corine linked dataset:** Corine Land Cover (Coordination of Information on the Environment Land Cover, CLC) is referring to a European programme establishing a computerised inventory on land cover of the 27 EC member states and other European countries, at an original scale of 1: 100 000, using 44 classes of the 3-level Corine nomenclature. This dataset contains linked data from agriculture-related lands (hilucs_code<200) & for main cities in the Czech Republic (centers of NUTS3 regions), Poland (agglomeration areas from Urban Atlas) and Spain (agglomeration areas from Urban Atlas).

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/Corine%20linked%20dataset/0.0.1

- **Urban atlas linked dataset:** The European Urban Atlas provides reliable, inter-comparable, high-resolution land use maps for over 300 Large Urban Zones and their surroundings (more than 100.000 inhabitants as defined by the Urban Audit) for the 2006 reference year in EU member states and for over 800 Functional Urban Area (FUA) and their surroundings (more than 50.000 inhabitants) for the 2012 reference year in EEA39. This dataset contains agriculture-related lands (hilucs_code<200) & for main cities in the Czech Republic (centers of NUTS3 regions), Poland (agglomeration areas from Urban Atlas) and Spain (agglomeration areas from Urban Atlas)

URL in DataBio Hub:

https://www.databiohub.eu/registry/#service-view/Urban%20atlas%20linked%20dataset/0.0.1

- **Land parcel dataset (LPIS) from the Czech Republic:** LPIS is a system to identify land use for a given country. It utilises orthophotos – basically aerial photographs and high precision satellite images that are digitally rendered to extract as much meaningful spatial information as possible. A unique number is given to each land parcel to provide a unique identification in space and time. This information is then updated regularly to monitor the evolution of the land cover and the management of the crops. This dataset contains land parcel and cadastral data collected from the Czech Republic. The input data source was in shapefiles which were transformed into Linked Data.

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/Land%20parcel%20dataset%20(LPIS)%20from%20Czech%20Republic/0.0.1

- **Erosion-endangered soil zones linked dataset from the Czech Republic:** This data contains the erosion of endangered soil maps from the Czech Republic. The source data was in shapefiles which were transformed into Linked Data.

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/Erosion-endangered%20soil%20zones%20linked%20dataset%20from%20Czech%20republic/0.0.1

- **Water buffer linked dataset from the Czech Republic:** This data contains the water buffer and water body-related data from the Czech Republic. The source data was in shapefiles which were transformed into Linked Data.

  URL in DataBio Hub:

  https://www.databiohub.eu/registry/#service-view/Water%20buffer%20linked%20dataset%20from%20Czech%20Republic/0.0.1

- **Norway catch records linked dataset:** Catch records data from Norway (2014-2019), transformed and published as Linked Data. The source data is in CSV format, which has been processed, transformed and published as Linked Data, enabling their integration with other datasets in the LOD datasets. The catch record data involves the catch amount in tonnes for each year, along with the species caught and vessel and fishermen data. fishing regions and catch area data are also included in the respective years in the dataset. The dataset is

already linked with some of these datasets, including FAO species, FAO fishing areas, and the ISO-3166 standard codes. The source CSV files can be found in https://www.fiskeridir.no/Tall-og-analyse/AApne-data/AApne-datasett/Fangstdata-koblet-med-fartoeydata

URL in DataBio Hub:

https://www.databiohub.eu/registry/#service-view/Norway%20catch%20records%20linked%20dataset/0.0.1

2.4 Data and application sharing

2.4.1 Application sharing

DataBio is pursuing the use of standards to ensure that its multitude of components can be easily combined and interoperate.

The Open Geospatial Consortium, in collaboration with ESA has developed an EO Big Data Architecture allowing deployment and executing applications close to the physical location of EO data[5]. This data includes satellite images, model outputs and in-situ data.

Several OGC Testbed Engineering Reports define the details of this architecture, including the OGC Testbed-14: Application Package Engineering Report [REF-08] and the OGC Testbed-14: ADES & EMS Results and Best Practices Engineering Report [REF-09]. According to this architecture, application components are made available in public or private Docker registries, where they can be retrieved to be deployed as application packages on a cloud infrastructure close to the data. DataBio is following the same approach to make available its architectural building blocks for deployment as part of pipelines from Docker registries (See Section 2.5 below).

The fully distributed landscape of (EO) application components and service endpoints made available by various platforms, including thematic exploitation platforms, mission exploitation platforms, data and processing clouds etc. constitute the "Network of EO Resources". The Data and Information Access Services (DIAS) platforms are an example of foundation blocks of the European EO "Resources Tier Layer". A number of DataBio pilot applications rely on the platform services provided by e.g. the Proba-V MEP and Forestry TEP which are part of the "Platform Services Layer".

[5] https://www.opengeospatial.org/pressroom/pressreleases/2988

Figure 4: Network of EO resources - Layer View (source: ESA)

The ESA "EO Exploitation Platform Common Architecture" initiative aims at defining standard interfaces facilitating the discovery and interoperation of the scattered resources available in this network. These interfaces are provided by the "Platform Services Layer" depicted in Figure 4 and consumed by the Exploitation Layer.

As part of OGC Testbed-15[6], DataBio actively contributed to the EOPAD (Earth Observation Processes and Application Discovery) Thread aiming to define a standards-based approach to discover applications and components either available as service endpoints accessible through OGC interfaces or Docker containers (application packages). The interfaces proposed by OGC to achieve this are defined in the OGC 19-020r1 Engineering Report [REF-10]. DataBio has adopted the proposed GeoJSON/JSON-LD application/metadata encoding (consistent with GeoDCAT-AP [REF-11]) and associated bindings to make the information from the DataBio Hub available. The above Engineering Report contains the details of the DataBio implementation in section 8.3.7. Its public release was approved by the OGC TC in Toulouse in November 2019.

To consolidate the implementation of the Testbed-15 discovery interfaces for the DataBio Hub, and ensure their interoperability tests with other partners, DataBio participated (remotely) to the OGC API Hackaton[7] in June 2019. The results of the performed work related to DataBio Hub are documented in the OGC19-062 Engineering Report [REF-13]. As part of the activity, it was demonstrated that a third-party client (OpenSphere) was able to discover and access metadata information in the DataBio server through its OGC API Common compliant interfaces.

Finally, after the release of OGC API - Features Core [REF-13], the interfaces of the DataBio Hub were aligned with this specification and the STAC specification [REF-14] as part of the

[6] https://www.opengeospatial.org/projects/initiatives/testbed15
[7] https://www.opengeospatial.org/OGCAPI_Hack2019

OGC API - Features and Catalogues Sprint in November 2019 in Arlington U.S.A [REF-15] and are available as [REF-16]. As part of this Sprint, it was demonstrated that external clients were able to access the corresponding DataBio Hub component metadata (in OGC 19-020r1 encoding). Figure 5 below shows how components to be discovered are organised per "organisation" providing the component. The metadata available includes DataBio components and OGC Testbed-15 participants. By adopting the same metadata for DataBio components/applications as defined in the OGC19-020r1 Engineering Report, DataBio components/applications were made available for discovery with the corresponding entry in a Docker registry. We achieved full alignment with the current state-of-the-art practices for application sharing and thus prepared DataBio for being a part, discoverable and reusable in the forthcoming "network of EO resources". Details of the DataBio contribution to the OGC Sprint are documented in [REF-17].

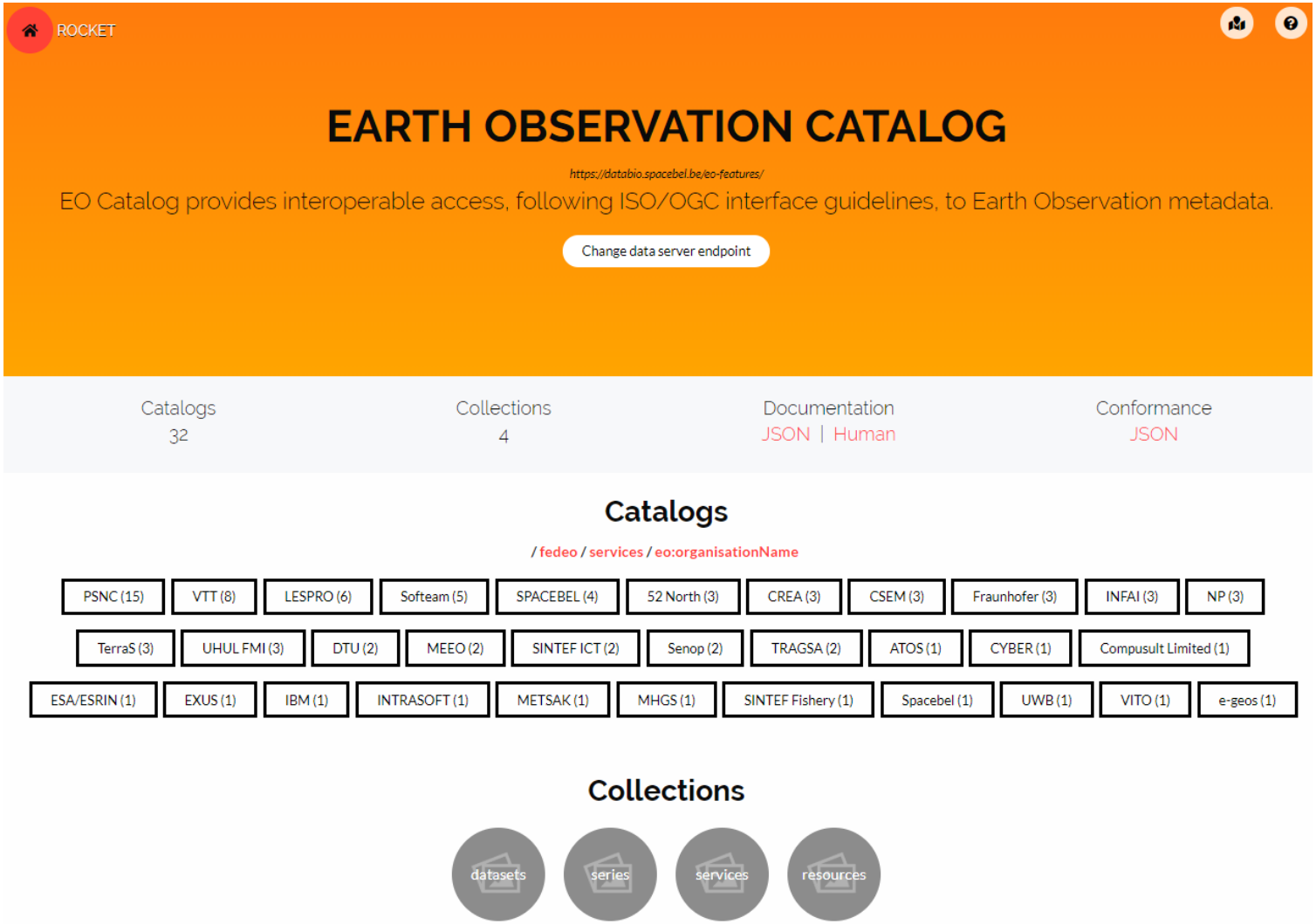


Figure 5: Access to DataBio Hub component/application metadata with third-party https://rocket.snapplanet.io/ application.

The above DataBio contributions and alignment to OGC Testbed-15 recommendations for application (and data) sharing were also covered in presentations at CEOS WGISS#48 (Vietnam, October 2019), the OGC Catalog and Metadata DWG (OGC TC in Toulouse, November 2019) and Accessibility and Discoverability Session of the Data Access and Preservation WG (CERN, Geneva, November 2019).

2.4.2 Data sharing

DataBio encourages the use of GeoDCAT-AP (and its compacted GeoJSON representation defined in OGC 17-084) for describing datasets, including datasets published by the Pilots. DCAT is one of the vocabularies supported by Google Dataset Search[8] in addition to Schema.org, which should allow automatic indexing of such metadata by mass-market search engines. The format can be used in combination with OpenSearch catalogue[9] bindings and as Linked Data. The metadata encoding can contain information about service offerings related to the dataset (e.g. a download link to the actual data, contain binding information to drill down in the dataset, DataCube interfaces available for the dataset etc.), reference the DOI[10] identifier of the dataset, citation information etc.

A typical use case is to publish the new dataset also on Zenodo[11]. The dataset then gets a DOI which can be referenced from the above metadata. In addition, datasets published on Zenodo get automatically indexed by OpenAIRE[12]. The table below provides the information for an example Fishery dataset generated by DataBio.

Table 10: Example Fishery dataset generated by DataBio

Description	***URL***
DataBio Hub GUI	https://www.databiohub.eu/registry/#service-view/Fishing ship sensor data/0.0.1
DataBio API access (OGC 17-084, OGC 19-020r1)	https://databio.spacebel.be/eo-catalog/resources/5de8d0fee4b09c46a6446989
Zenodo	https://zenodo.org/record/3563390#.XejEEihKhaS
DOI	https://doi.org/10.5281/zenodo.3563390
OpenAIRE	https://explore.openaire.eu/search/dataset?datasetId=r37b0ad08687::2000a2b3921e4105261f70d5eafadde6

[8] https://toolbox.google.com/datasetsearch
[9] https://catalogue.nextgeoss.eu/opensearch
[10] https://www.doi.org/
[11] https://zenodo.org
[12] https://explore.openaire.eu/

CEOS agencies, including ESA are spending a lot of effort to make EO datasets (collections and products) discoverable through standard interfaces and describe the datasets with agreed metadata. CEOS IDN[13] is the main entry point, combining EO datasets from ESA FedEO and CWIC. The core metadata specifications describing the EO datasets are DIF-10, and ISO19139-2. There is ongoing work by the ESA supported by the CEOS WGISS community to make EO dataset metadata available in GeoDCAT-AP compliant encoding as defined in OGC 17-084 [REF-18]. As most EO datasets used as input by the DataBio applications are already available in this format (Sentinel, Landsat, Copernicus MEMS etc.) through the ESA FedEO endpoint, DataBio Hub makes available these metadata in this format as well via the catalog contributed to Testbed-15 available at [REF-19].

Table 11: Example EO datasets used by DataBio described with standard metadata

Dataset (i.e. EO collections)	***Metadata via DataBio Hub API***
Sentinel-2	https://databio.spacebel.be/eo-catalog/resources/EOP:ESA:Sentinel-2
Landsat-8	https://databio.spacebel.be/eo-catalog/resources/LANDSAT.ETM.GTC
Proba-V	https://databio.spacebel.be/eo-catalog/resources/?type=collection&platform=Proba-V
CMEMS SeaLevel	https://databio.spacebel.be/eo-catalog/resources/SEALEVEL_GLO_PHY_L4_REP_OBSERVATIONS_008_047

As the dataset metadata is GeoDCAT-AP [REF-11] compliant (via a normative JSON-LD context), it can be accessed also as linked data and combined with other linked data sources as was presented by DataBio (PSNC) at the OGC TC in Toulouse (November 2019) in the "EO for Agriculture Needs" session.

The Linked Data datasets are also discoverable via the DataBio Hub and its API. The table below refers to an example of Linked Data dataset from one of the Fishery pilots. The metadata includes information about the SPARQL endpoint allowing data access.

Table 12: Example Linked Data dataset from one of the Fishery pilots

Dataset (i.e. Linked Data)	***Metadata via DataBio Hub API***
Norway catch records linked	https://www.databiohub.eu/registry/#service-

[13] https://idn.ceos.org/

dataset	view/Norway%20catch%20records%20linked%20dataset/0.0.1 (GUI) https://databio.spacebel.be/eo-catalog/resources/5dea2ee9e4b09c46a644698d (API)

Finally, meteo datasets used by the Pilots are equally made available. For these datasets, the published metadata includes information about the OGC Web Coverage Service (WCS) serving the data. Below an example dataset:

Dataset (i.e. Meteo Data)	***Metadata via DataBio Hub API***
KNMI Precipitation Data	https://www.databiohub.eu/registry/#service-view/KNMI%20(Koninklijk%20Nederlands%20Meteorologisch%20Instituut)%20precipitation%20data/0.0.1 (GUI) https://databio.spacebel.be/eo-catalog/resources/5b86a466e4b042a37ac77ed9 (API)

2.5 Container-Based Deployment

The "Exploitation Platform" concept described in *Appendix B (Benefits from OGC Testbed*) of this document applies to DataBio. Two ESA "Exploitation Platforms" are currently providing the environment to run Forestry and Agriculture pilots, in particular, the Forestry TEP (VTT) and Proba-V Mission Exploitation Platform (VITO). These may benefit from the evolution of this concept in the previous OGC Testbeds and applied in some of such environments.

In the following sections, we briefly introduce some key technologies and how they have been applied to the DataBio context to achieve a DataBio platform fostering reuse and facilitating the rehosting of components and pipelines on heterogeneous environments during and beyond the project.

2.5.1 Docker containerization

2.5.1.1 Technology

Docker provides a set of tools to support software as a service deployment using containerized applications. In the domain of Software Engineering and Deployment, *containerization* refers to the concept of shipping applications along with all dependencies, libraries, configuration files and other resources to setup a well-defined application environment. Consequently, there must be no dependency or other requirements to execute such an application besides a standardized runtime environment provided by Docker. This allows overcoming the need to do application-specific provisioning of the executing host and

complex application setups. Furthermore, this allows for isolating different applications on the same host with ease.

To run software components or complex systems that consist of multiple services, Docker allows to connect different containers using virtual networks. Therefore, services can communicate which each other transparently and without any knowledge about the Docker environment.

Each Docker Container is created from a Docker Container Image, which usually consists of multiple layers. Each layer is the result of a single setup stage such as installing a library, adding a file or modifying the containerized environment. This allows to reuse so-called intermediate layers and hence decrease disk usage and speed up the build process. Images can be delivered using standard transport mechanisms (i.e. TCP) and are usually shared using a centralized repository which is called registry in the context of Docker.

The build-process of an image is defined in a textual file called Dockerfile. Starting with the definition of a base-image it contains a list of instructions to add files, run commands or setup environment variables to setup the application environment. Each of these instructions will result in a separate layer, as mentioned before. Furthermore, the Dockerfile specifies which command should be executed when running the container.

Main advantages:

- Self-contained: single responsibility, easy to reuse and secure
- Independent from the operating system (will run in the same way on all operating systems that can run the Docker engine), even if there is no CUDA support for Machine Learning under Windows Dockers
- Containers are more optimized than a program running on a Virtual Machine
- Easy to share, easy to reuse and built upon public images
- Version control friendly

2.5.1.2 Use in DataBio

To enhance the interchangeability of DataBio components that were introduced in DataBio deliverables D4.1 [REF-01], D5.1 [REF-04] and D5.2 [REF-05], and facilitate their deployment at Pilot sites it was beneficial to remove the dependencies of these components with their underlying infrastructure and package them as Docker containers.

Furthermore, the various (EO) processing components provided by Technology partners (WP4, WP5) or directly by the Pilot partners correspond to “Exploitation Platform Application Packages” in OGC Testbed or EOEPCA terminology. Although run-time deployment through an ADES (Application Deployment and Execution Service) is not required, the packaging proposed by OGC Testbed as Docker container is applicable and enhances the independence of the components and the pipeline processing services and application implementations from the underlying “Pilot” host infrastructure. Figure 6 below depicts the pilot processing services and applications as part of the Platform Tier in the well-known “EO Exploitation

Platform (EOEP) Common Architecture"[14] (EOEPCA) representation. The DataBio Hub API allows the discovery of DataBio applications and datasets, Moreoever, it covers the Resource Management[15] (Application Catalog[16]) functionality exposed to external applications via the "Service API" of the Platform Tier. The DataBio Hub API corresponds to the 'Data Discovery" and "Application Discovery" interfaces accessed by the EOEPCA Client Library[17].

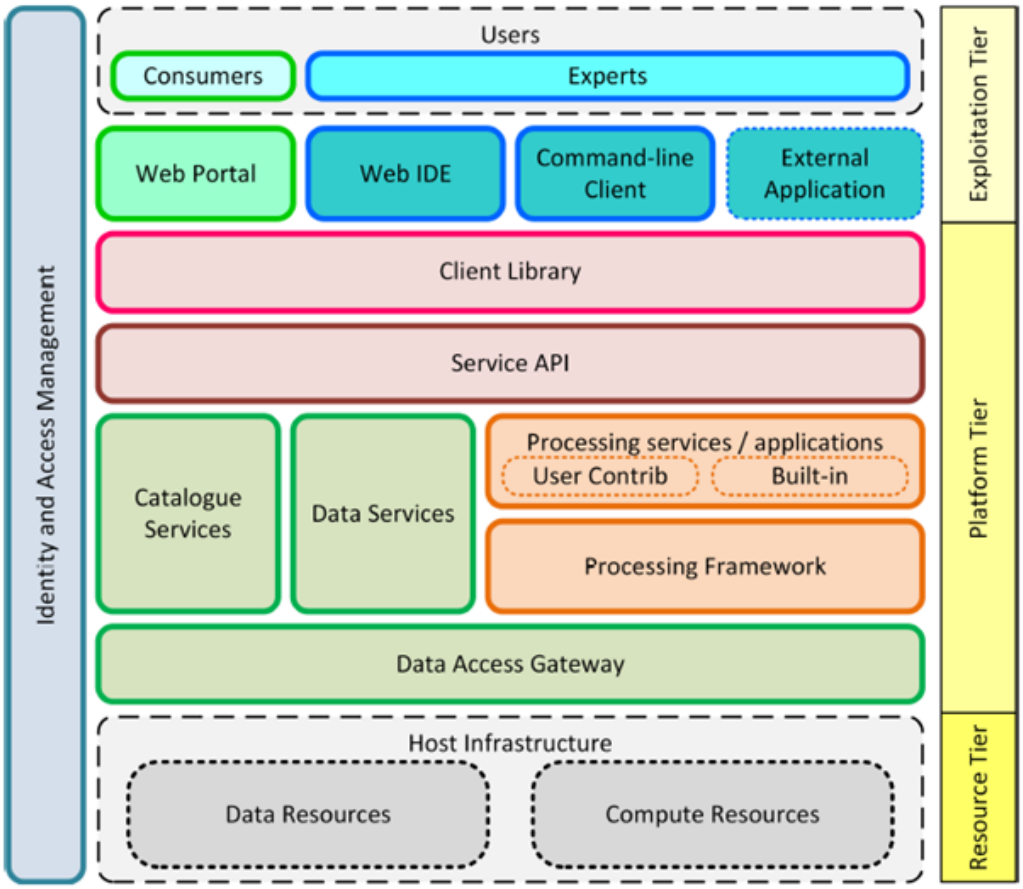


Figure 6: Architecture layers

To be able to support pilot partners, Docker registries have been deployed as part of the DataBio Platform. In this registry, interested partners can find available components ready-to-use as Docker images. These components have not been made available to the general public. The DataBio Docker registries could accommodate components that are open-source, but also components by SME partners which are not necessarily open-source, but proven to fit with other DataBio components to realise useful pipelines as described in D5.2 and D4.1.

[14] https://eoepca.github.io/master-system-design/#img_architectureLayers
[15] https://eoepca.github.io/master-system-design/#_resource_management
[16] https://eoepca.github.io/master-system-design/#mainAppCatalogue
[17] https://eoepca.github.io/master-system-design/#mainClientLibrary

In addition, the DataBio Hub refers interested parties to the corresponding partners and to the DataBio Docker Registry for access to the components and/or "application packages".

There are three possibilities to host Docker images in a Docker registry, a combination of which can be used as well by Partners:

- Host images publicly on Docker hub: this is free, but everybody has access to it.
- Host image on a private (paying) registry on Docker hub: https://hub.docker.com/billing-plans/. "Small" private registries cost less than 15 USD per month.
- Deploy a private registry on premise: There is the official registry from docker with a limited set of functionality (See https://docs.docker.com/registry/deploying/) and third-party solutions, such as VMWare Harbor (see https://goharbor.io/), which provide a more sophisticated range of features. This allows complete control over user access rights. It would also allow Partners to push software from their development machines and Pilots to "pull" components automatically to their Kubernetes cluster (See next section). A private registry for DataBio has been set-up by a project partner PSNC (see Section 2.5.3).

Docker images can be retrieved/pulled without using a registry as well as explained in https://medium.com/@sanketmeghani/docker-transferring-docker-images-without-registry-2ed50726495f.

2.5.2 Container orchestration with Kubernetes

2.5.2.1 Technology

Kubernetes is a production-grade container orchestration system to automate the deployment, scaling, and management of containerized applications across clusters of nodes. Kubernetes provides clustering features such as fault-tolerance, self-healing, dynamic scaling, automated upgrades, etc. Kubernetes is available on many public clouds (Google, Amazon, Azure, Oracle, etc.) and hosted solutions. A Kubernetes cluster can scale up to 5,000 nodes, 100 pods per node, 150,000 pods, and 300,000 containers.

The Kubernetes approach is to declare in a YAML file the desired state of the system in terms of:

- Number of nodes
- What containers
- How many replicas
- What storage
- etc.

The Kubernetes master, based on its knowledge of available resources (nodes and their capacity) will trigger container deployments / destructions in order to reach the desired state.

Kubernetes services are offered by many cloud providers. Below are some examples:

- https://cloud.google.com/kubernetes-engine/
- https://aws.amazon.com/eks/
- https://azure.microsoft.com/en-us/services/container-service/kubernetes/
- https://cloud.oracle.com/containers/kubernetes-engine
- https://www.ibm.com/cloud/container-service

Main advantages:

- Centralized, explicit and declarative infrastructure configuration
- 100% Docker compatible
- Robust infrastructure (Kubernetes is able to restart failed containers)
- Easy to add new nodes
- Optimize resource utilization
- Easy to scale: scaling can be as simple as incrementing the replica number (this will create a new pod, allocate it to a node, create an endpoint for the corresponding service and add it to the load-balancer configuration)
- Opens up a lot of powerful features such as auto-scaling.

2.5.2.2 Use in DataBio

Kubernetes allows horizontal scaling (and load balancing) of components by transparently adding nodes used by the component. In addition, Kubernetes considerably facilitates the deployment of containerised components on clouds as explained in the previous section. As many public cloud providers offer such a solution, it allows DataBio component pipelines (see DataBio D5.2 [REF-05]) to be rehosted easily on different environments/clouds (public, hybrid, private) at different Pilot sites.

As EO data (e.g. Sentinel, Landsat) are and will be available on several of the public clouds (Amazon, Google, DIAS), the co-hosting of DataBio pipelines on the same cloud where the data resides will significantly simplify access to data. Spacebel is using this approach with its "GEP" Exploitation Platform used by EORegions! and hosts currently the entirety of the components and thematic pipelines on the Google Cloud Platform. A number of the same components, scaled over a different number of cluster nodes to achieve scalability, are also running as containers on a Kubernetes cluster on the Interoute cloud.

Deployments on Kubernetes can pull component images directly from the corresponding Docker Registries. The installation (and maintenance) of pipelines consisting of multiple components can thus be greatly automated and simplified.

2.5.3 Infrastructure

The exploitation of Docker technology in DataBio project requires dedicated resources to maintain Docker registry. A private, internal registry is used to support the development process by implementing security and access policies according to project requirements. The

volume of resources provided by the registry can be strictly controlled, to provide efficient and reliable service.

Current Docker registry is located at https://registry.apps.man.poznan.pl . It is maintained by the partner PSNC and hosted on their internal resources. For that purpose, the following resources are allocated: 8GB RAM, 4CPU, 1.5TB storage.

At the time of writing this document, DataBio Docker registry contains three images: OpenVA Realtime, Kartoza/Postgis, Geo-L. The private registry has been successfully used, for example, in a fishery pilot where new software versions of OpenVA Realtime by VTT were updated continuously to the registry and IBM integrated the software with their Proton application. Finally, the Docker containers were deployed to fishing ships with the help of EHU/UPV. The use of Docker greatly helped to overcome usual problems caused by different operating systems.

Publication as Docker containers is not limited to a single centralised registry. Additional DataBio components are available as Docker images on publicly accessible Docker registries (e.g. C05.01 “Rasdaman”, C14.01 “Sen2Cor” etc.) while other partners have used other access restricted registries to publish the Docker images of their closed-source components (e.g. Spacebel).

2.6 DataBio Hub

DataBio Hub provides a registry for sharing information about the project components, datasets, pipelines, and pilots for an easy search of the different project entities. Registered hub users can add new resources and model dependencies between resources. DataBio Hub supports private sandboxes for editing descriptions before publishing them to unregistered users.

DataBio Hub does not offer a repository or operating environment for the services and datasets themselves, as those services will be running on the service providers’ servers or cloud infrastructure (or DataBio -provided cloud). Regardless of the running environment of the services, DataBio Hub offers descriptions and endpoints (= service calls) to all DataBio platform -compatible services and components (and possibly applications) in a single location and with a coherent description.

A special service registry instance has been provided for DataBio project and can be found at http://www.databiohub.eu/. The instance has been deployed as a virtual machine in Microsoft Azure’s cloud computing service. Infrastructure as a Service (IaaS) allows easy server management and increasing computing power and resources if needed. The virtual machine runs on Ubuntu Linux platform and the whole machine is backed up in a geologically redundant recovery service vault.

The Digital service registry has been tailored for DataBio use, which includes the following developments:

- As the service registry was initially developed to register digital services with mainly machine-readable interface descriptions, vocabulary support for new categories of software components such as applications with both human and technical interfaces have been added.
- New interface technologies such as OpenSearch for satellite image services have been added to Service hub interface description vocabularies.
- DataBio Pilot descriptions, data processing pipelines and datasets developed or used in pilots as well as component descriptions are now also included into the registry. Links to textual descriptions in public project deliverables are provided.
- Specific rules for keyword use for DataBio have been enforced to link descriptions to BDVA reference architecture and also help linking component and service descriptions to the overall architecture of the DataBio platform and pilots.
- New fields for human-readable descriptions have been added to improve linking them to pilot development, data models, external distribution platforms such as Docker hubs, and DataBio deliverables with the possibility to include as images the component diagrams exported from the DataBio architecture models.
- Service Hub UI and its website have been tailored for DataBio and linked with other websites of the DataBio project.
- Registration mechanisms for new users outside the DataBio consortium have been restricted during the DataBio platform development.

A new DCAT/JSON-LD based REST API has been implemented for accessing and exporting the information in the hub. During the export, the descriptions can be automatically scanned for improved keyword and theme linking external resources such as DBpedia and ESA thesaurus using spotlight API or FOODIE semantic annotation services.

A machine-to-machine interface for accessing the DataBio Hub metadata via OpenSearch is available at https://databio.spacebel.be/eo-catalog/description?httpAccept=application/opensearchdescription%2Bxml. The service endpoint compatible with OGC Testbed-15 interfaces (OGC 19-020r1) is available at https://databio.spacebel.be/eo-catalog/.

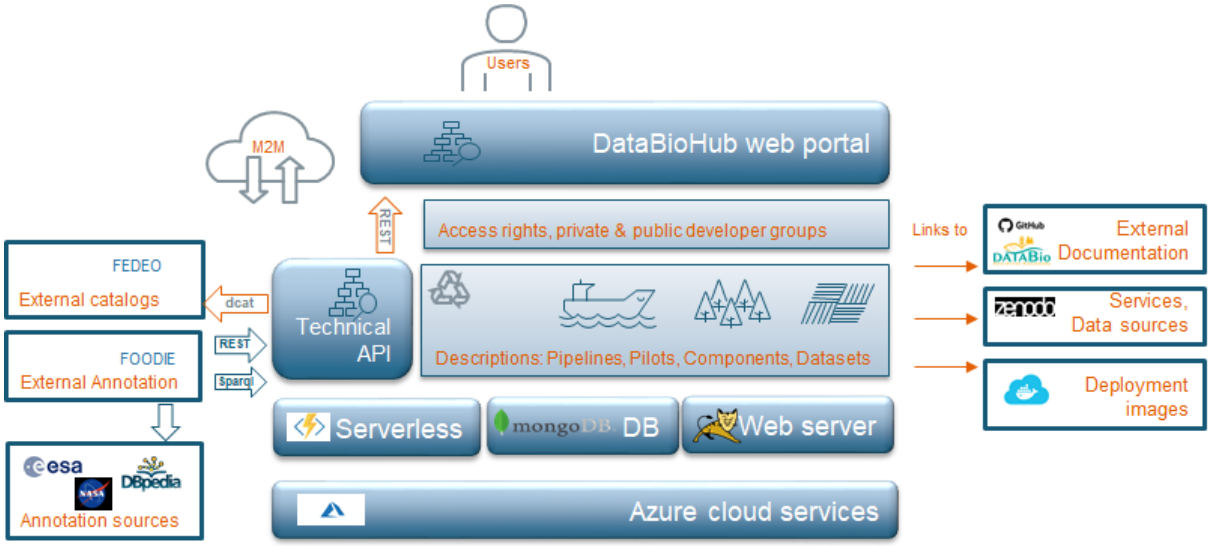


Figure 7: Architecture of DataBioHub

2.7 Confidential data handling and DataBio example

Data analysis and machine learning methods can provide great value in different areas of governance and business. By recognizing patterns in data, visualizing the patterns and developing predictive models we can optimize farming, forestry and fishing operations.

Well-known data analysis and machine learning tools and frameworks can be used when the data originates from public sources such as satellite images or when an agricultural business collects their own data. When the data on the other hand is confidential, current toolsets can protect it only while it is not being used or when data is being transferred. For data analysis common tools need to remove technical protection. Thus, the only protection of the owner of confidential data is contractual or limiting access to data to selected trusted persons.

One of the reasons for combining data from different companies and public sources is to improve the accuracy of machine learning and data analysis methods as data from different entities might capture different patterns or provide increased statistical power due to larger sample size. Learning from combined data can thus provide increased value for the industry. However, companies might be reluctant to share their data to protect the confidentiality of their operations.

Secure computation technologies have been developed which enable processing confidential data without leaking individual values. By using these technologies, we are able to develop data analysis and machine learning software that retains the confidentiality of individual data providers but allows them to collectively gain improved insights from sharing their data.

When using secure computation data is encrypted by the data owner and only then sent to a service processing the data. The host of the service will not have access to the unencrypted data nor the encryption keys. Data protection is not removed even while the data is being processed.

Secure computation technology can be used to develop solutions which are otherwise not possible due to confidentiality restrictions. There are some general types of problems where secure computation technology may be required:

- Outsourcing computations. Secure computation is a solution if one wishes to provide an analysis service to clients without learning the clients' data.
- Analyzing data governed by data protection laws. Secure statistical analysis can be used for decision-making when databases are governed by data protection laws and remain inaccessible for standard statistics software.
- Analyzing data from multiple sources. If data originates from a single provider, then the provider can run analysis using their own infrastructure without giving data access to a third party. If we wish to analyse data from multiple sources without revealing the data to the party running the analysis, we can use secure computation technology.

In this section we will describe two technologies for privacy-preserving data analysis and a demonstrator developed in the DataBio project which uses such technology to predict catch location and expected catch size for fisheries. The business impact of privacy-preserving data analysis and its applicability are also discussed.

2.7.1 Technology

Secure computation approaches can be categorised into software-based cryptographic techniques and hardware-based techniques. We will describe one example in each category.

2.7.1.1 Secure Multi-Party Computation

Secure multi-party computation (MPC) is a cryptographic technique for processing private data while preserving privacy. Sharemind MPC is a technology leveraging MPC which provides a framework for programming secure client-server applications. The roles of different parties involved in a Sharemind MPC process are:

- Input parties who convert their public data into secret data and import it to servers hosted by computation parties.
- Computation parties who perform operations on the secret data without learning the input values or the results.
- Output parties who can retrieve the secret results from computation parties and construct the public result values.

Sharemind MPC uses an approach for MPC called additive secret-sharing where private values are split into random values before being imported into an MPC system. This means that given a private 32-bit value x, two random values x_1, x_2 are generated and x_3 is computed so that $x \equiv x_1 + x_2 + x_3 (mod 2^{32})$.The three values are sent to three independent servers.

The servers can perform arithmetic on secret-shared values. For example, to add two values each server adds their respective shares of the values. After the local additions each server

holds one share of the sum. More complicated operations require network communication between the servers. Figure 8 illustrates how two private values can be added using MPC.

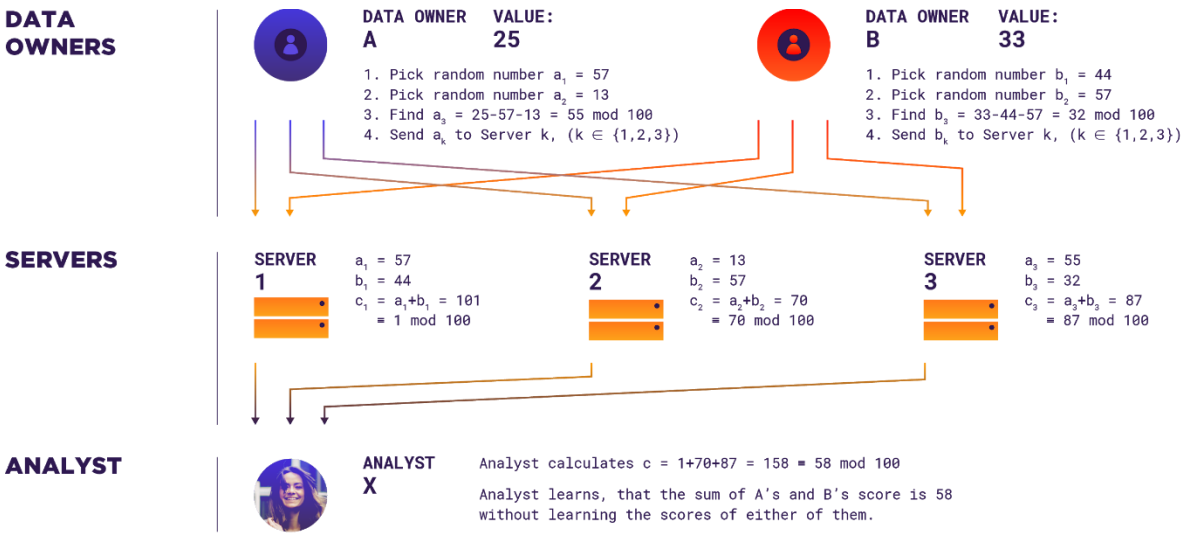


Figure 8: Illustration of adding secret-shared values

As long as at most one of the servers is compromised, privacy remains protected. All three server hosts verify the analysis program before installing it. This ensures that only agreed upon results will be published to output parties. Shared responsibility also means that privacy remains protected if one of the servers is compromised. Sharemind MPC includes an auditing tool to detect tampering.

MPC is a general-purpose programmable technique and has been successfully used to implement a variety of practical applications [REF-22]. The Sharemind MPC technology has been used for tax fraud detection [REF-23], statistical analysis of government databases for a social study [REF-24] and a report on the state of the Estonian IT industry by combining data from companies in the IT sector [REF-25].

The main benefit of MPC is the high security guarantees it provides. A party hosting an MPC server cannot learn anything about the values sent to it. There are no side-channel attacks which sometimes plague cryptographic techniques. Sharemind protects data in transit, in memory, at rest and during computations.

When compared to conventional software MPC increases deployment complexity and decreases performance. Since the three server hosts must be independent, the organisations using MPC must decide on three parties who will be managing the servers. This involves more contracts between parties participating in the process when compared to a single organisation providing an analysis service, but data will be protected technically, not just by the contracts as with usual data analysis tools.

2.7.1.2 Trusted Execution Environments

An alternative to software-based techniques is using a Trusted Execution Environment such as Intel Software Guard Extensions (SGX)[18]. SGX is an extension of the instruction set of Intel processors which enables developing secure applications when even the host operating system is not trusted. SGX relies on three concepts to protect data: enclaves, attestation and data sealing.

SGX is a set of CPU instructions for creating and operating with memory partitions called *enclaves*. When an application creates an enclave, it provides a protected memory area with confidentiality and integrity guarantees. These guarantees hold even if privileged malware is present in the system, meaning that the enclave is protected even from the operating system that is running the enclave. With enclaves, it is possible to significantly reduce the attack surface of an application.

Remote attestation is used to prove to an external party that the expected enclave was created on a remote machine. During remote attestation, the enclave generates a report that can be remotely verified with the help of the Intel Attestation Service. Using remote attestation, an application can verify that a server is running trusted software before private information is uploaded.

Data sealing allows enclaves to store data outside of the enclave without compromising confidentiality and integrity of the data. The sealing is achieved by encrypting the data before it exits the enclave. The encryption key is derived in a way that only the specific enclave on that platform can later decrypt it.

Sharemind Hardware Isolation (HI) is a technology using Intel SGX which provides the ability to process confidential data. Sharemind HI is built as a client-server service similar to Sharemind MPC. The client is an application that calls operations on the server, encrypts data and performs remote attestation on the server. The Sharemind HI server does the bulk of the work and is responsible for the following:

- Checking if a user has the right to access the system.
- Checking if a user has the correct roles to perform an operation.
- Managing the encrypted user data and the encryption keys of the data.
- Managing task descriptions of how a data analysis process is carried out.
- Storing a log of the operations performed in the server.
- Scheduling the tasks to run.

Figure 9 illustrates the security model of Sharemind HI applications. The input data, shown in blue, is encrypted at the client side and sent to the server. The input data encryption keys of the data are securely transferred to the SGX protected enclaves. Likewise, the output data,

[18] Intel® Software Guard Extensions | Intel® Software

shown in green, is encrypted inside of the enclave and stored on the server. When requested, the enclave securely transfers the output data encryption keys to the authorized clients.

At any point during the deployment, a client can request a cryptographic proof of what analysis code is running in the server, shown in blue on the figure. This proof can be compared against a previously generated proof by an auditor who has validated the code to be secure.

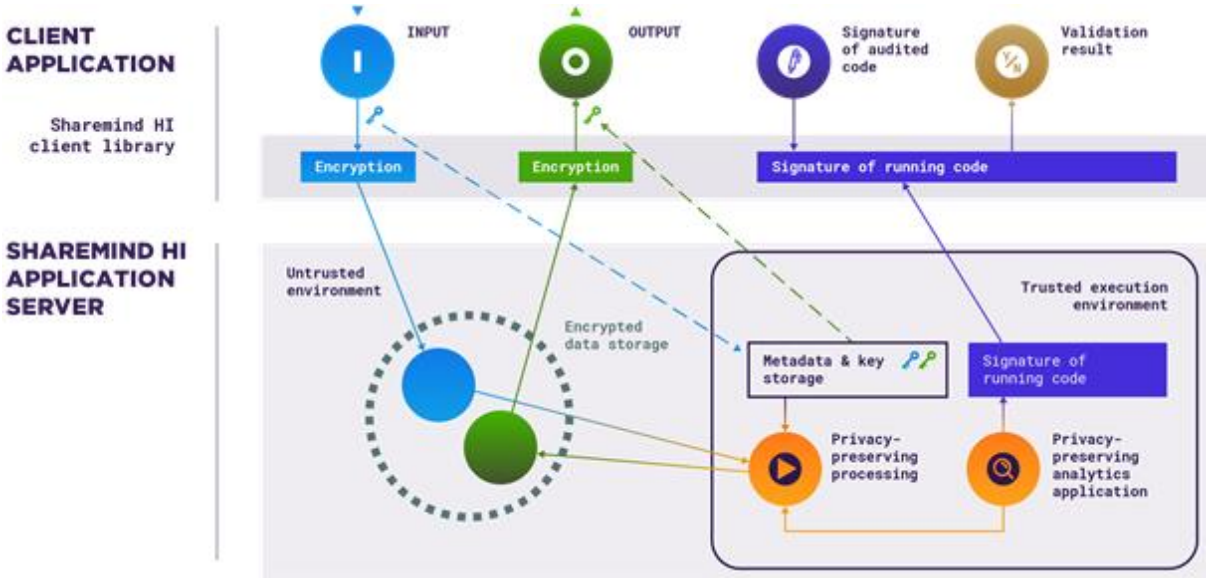


Figure 9: Sharemind HI security model

The main benefits of Sharemind HI over Sharemind MPC are high performance and simpler deployment. There is only one computational party and unlike Sharemind MPC network communication is not required while the enclave is running.

Another benefit of Sharemind HI is that enclaves are programmed in the C++ programming language whereas Sharemind MPC programs are written in a domain-specific language called SecreC which resembles C. This allows Sharemind HI programmers to use libraries and other existing code written in C or C++.

On the other hand, Sharemind HI requires users to trust Intel. Details of how SGX-enabled processors are produced is undisclosed information and Intel cannot prove that SGX is secure. It is also possible that side-channel attacks against SGX will be developed which would require more careful design of the enclave software. Practical applications should consider the security and performance trade-offs between cryptographic and hardware-based techniques.

2.7.1.3 Fully Homomorphic Encryption

Yet another alternative technology for privacy-preserving computation is fully homomorphic encryption (FHE). FHE allows for arbitrary computations on encrypted data. The privacy of the data is ensured by the encryption and is therefore independent of the trustworthiness or security of the server that is executing the computation. FHE has often been called the "swiss army knife of cryptography", since it provides a single tool to be applied to a wide variety of applications.

The idea of computing messages in a secure (i.e. encrypted) form has been around for a long time, but it was not clear if this is actually possible. Partially homomorphic encryption methods (that is, they allow only certain types of operations) have been known since the 1980s. However, fully homomorphic encryption was first developed in 2009 by Gentry et al. Since then, various improved FHE schemes have been proposed to increase the unfeasibly bad performance (with respect to both security and efficiency) of Gentry's first FHE scheme.

Homomorphic means structure-preserving. The mathematical structure of a message is preserved under a homomorphic encryption, so operations can still be performed on the encrypted message in a meaningful way. This property can be limited to certain types of operations (in partially homomorphic encryption) or support arbitrary operations (then the encryption is called *fully homomorphic*). Normally, encryption scrambles messages in such a way that they can no longer be processed without destroying the contained information. So, it is a special property for an encryption scheme to be fully homomorphic. Although FHE schemes may be based on different mathematical foundations, they all function in a similar way on a general level.

Figure 10 visualises the simplest variant with two parties: Alice and Bob. Alice has a message m and some algorithm f that she wants to apply to m to obtain a result $f(m)$. However, she does not wish (or is lacking resources) to perform the necessary computation herself. Instead, she asks Bob to do it for her. To protect her privacy even from Bob, she encrypts m with a homomorphic encryption scheme Enc and sends the encrypted message Enc(m) to Bob, along with instructions how to evaluate f on Enc(m)[1].

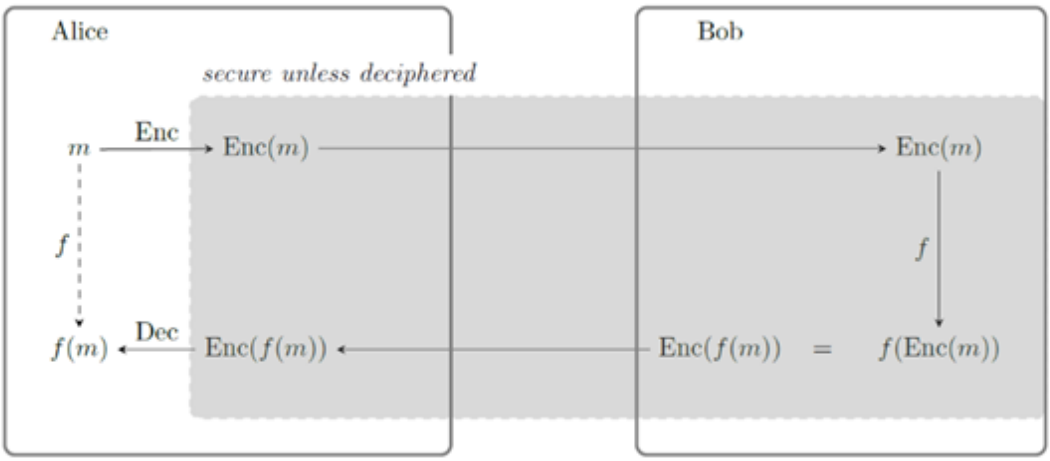


Figure 10: Schematic diagram of a homomorphic encryption scheme (two parties)

Bob then evaluates f on Enc(m) to obtain f(Enc(m)). At this point the homomorphic property of Enc comes into play. It is based on the fact that f(Enc(m)) = Enc(f(m))[2]. Bob can therefore send Enc(f(m)) back to Alice, which she can decrypt. During the whole process, he has only ever seen encrypted messages that he cannot read.

Alice is the only one to have the decryption function Dec and she uses it to obtain $f(m)$. The diagram in Figure 10 *commutes*, i.e. the end result is the same as if she had applied f to m herself (indicated by a dashed arrow). Remember that in a *fully homomorphic* encryption scheme, this can be done for an arbitrary f, while in some other schemes it only works for certain types of functions.

In practise, an FHE scheme needs to meet certain requirements to be feasible. For one, keep in mind that the main motivation for Alice is that she doesn't have the resources to calculate $f(m)$ herself. In order for it to make sense for her to use, it needs to be significantly less costly to perform the Enc and Dec operation instead of f.

Secondly, operations on encrypted messages will typically be more costly than the same operations on unencrypted messages, often by a large factor. Although it can be assumed that Bob has more computing resources than Alice, they are still limited. If the factor of cost-increase due to encryption is too high, the FHE scheme is simply not worth the trouble.

Thirdly, the encryption needs to guarantee a sufficient level of security - both against outside attacks and against a betrayal from Bob's side.

2.7.1.4 On-the-fly MPC by Multi-Key Homomorphic Encryption

One major disadvantage of classical MPC-schemes (such as secret sharing) is that they need to be planned out in advance. The number of participants needs to be known and fixed before the calculation starts. In contrast, there is the concept of *on-the-fly MPC*, which is much more flexible in those regards. The main criteria an on-the-fly MPC scheme should meet are:

1) The cloud can perform arbitrary, dynamically chosen computations
2) It can use data from an arbitrary, non-prefixed set of participants ("on-the-fly")
3) The computations are non-interactive, i.e. they don't require communication with all the participants (like with secret sharing)

On-the-fly MPC can be achieved by using so-called multi-key fully homomorphic encryption (MKFHE). While most FHE-schemes allow only one encryption key to be used, MKFHE schemes allow for multiple (almost arbitrarily many) different keys to be used for one computation.

Figure 11 illustrates how an MKFHE scheme can facilitate on-the-fly MPC. In this case we have 4 different Alices with each their secret message m_1, m_2, m_3, and m_4 (however that number might be higher or lower than 4). Each of them encrypts their message using a different key (k_1, k_2, k_3, and k_4) and sends it to Bob.

Out of these 4 encrypted messages, Bob can choose any subset (say Enc(m_1, k_1), Enc(m_2, k_2), Enc(m_3, k_3)) and any function that he wishes to perform on it (say f). Note that these choices can be made *after* the messages have been encrypted and sent to Bob.

He then calculates f(Enc(m_1, k_1), Enc(m_2, k_2), Enc(m_3, k_3)) and sends the result back to Alice1, Alice2 and Alice3, who agree to approve or disapprove of the calculation. If approved, they

can decrypt the result together and obtain $f(m_1, m_2, m_3)$. The decryption is only possible if the three of them work together.

Note that there is no need for any communication with Alice4, since her message is not involved in the calculation. Also note, that the other three Alices need not communicate until after Bob has finished his calculation. This gives MKFHE a huge advantage over classical MPC in terms of scalability and flexibility. However, like for other FHE schemes, the computation of f is very costly.

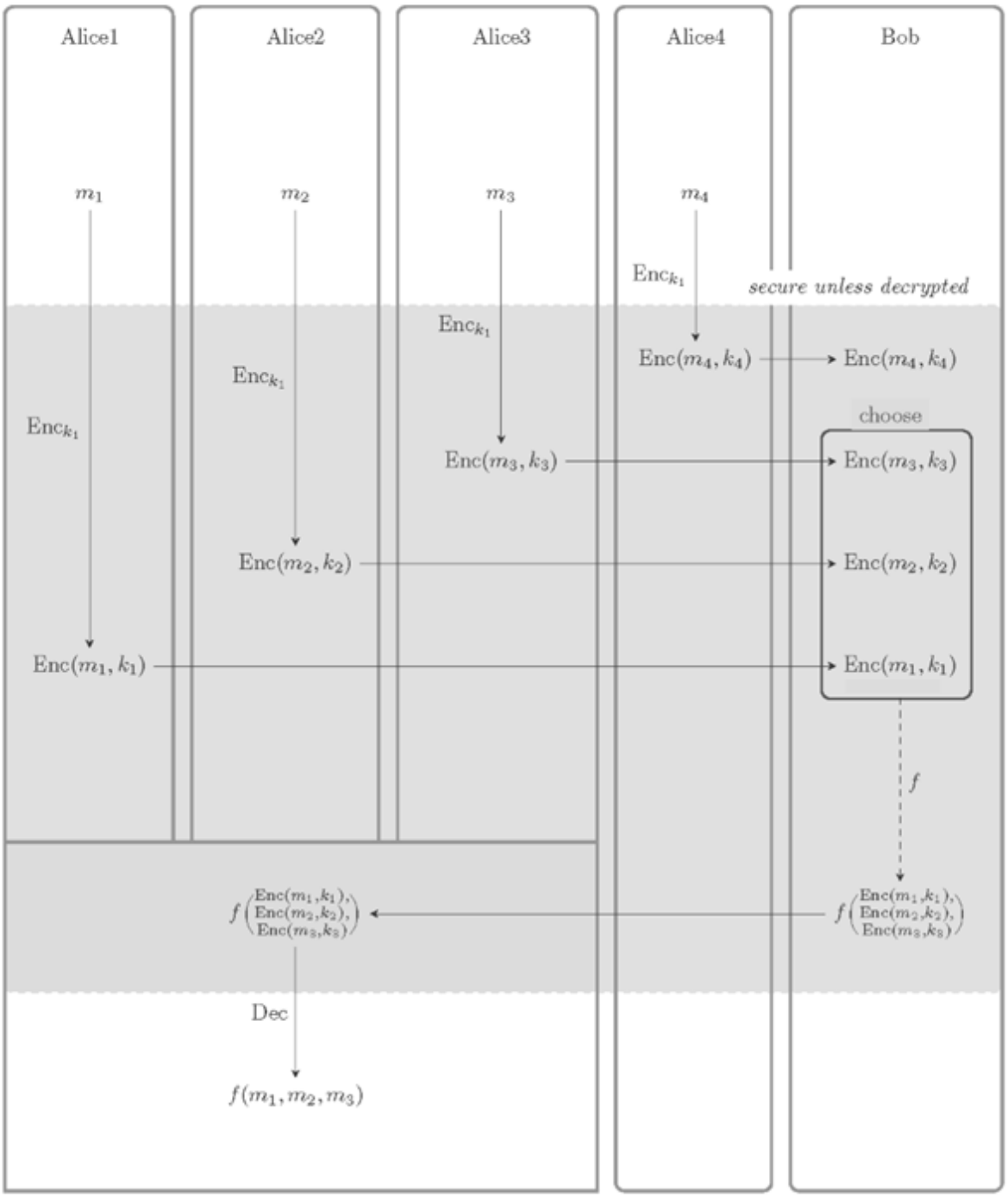


Figure 11: On-the-fly MPC using an MKFHE scheme

2.7.1.5 Comparison of Methods

The methods discussed above have different characteristics. The following describes relevant different properties and characteristics of the methods.

- **MPC by Secret Sharing:** relatively efficient - easy to handle - already mature technology - requires coordinating multiple servers - requires in-advance planning and setup.
- **Trusted Execution Environments:** high efficiency - need for special hardware support.
- **Single-key Homomorphic Encryption:** very flexible - security independent of software and hardware - needs only one server - high computational cost - difficult to understand and use - allows for one key only.
- **Multi-key Homomorphic Encryption:** full flexibility - security independent of software and hardware - on-the-fly execution - high computational cost - difficult to understand and use.

The flexibility that homomorphic encryption schemes offer may be useful for some applications, but in the DataBio proof of concept demonstrator it is of highest importance to demonstrate a solution that is scalable with Big Data and have a feasible computational cost.

We therefore decided that MPC and Trusted Execution Environments would be the most feasible option to pursue and demonstrate further.

2.7.2 Use in DataBio: Secure Machine Learning of best catch locations - Pipeline

To demonstrate how secure computation technologies could be used in agriculture, forestry and fisheries we developed a demonstrator which predicts the best fish catch location and expected catch size on a given day.

Catch data with geographical positions was retrieved from the Norwegian Directorate of Fisheries [REF-26]. Although we used public data for experimentation, our approach demonstrates that secure machine learning models can be trained on data from multiple fisheries and enable combining private data with public data.

We implemented the mathematical model used in the pilot using both Sharemind MPC and Sharemind HI. Due to better performance we chose the Sharemind HI solution as the backend for a web-based tool. The Sharemind MPC version is efficient enough to train models that can be reused for estimation afterwards even if the model is kept private. Since our tool has some fishery-specific parameters the model would need to be trained for each fishery. The Sharemind HI implementation trains the model in the order of a minute instead of a few hours.

Figure 12 illustrates the prediction pipeline using secure machine learning.

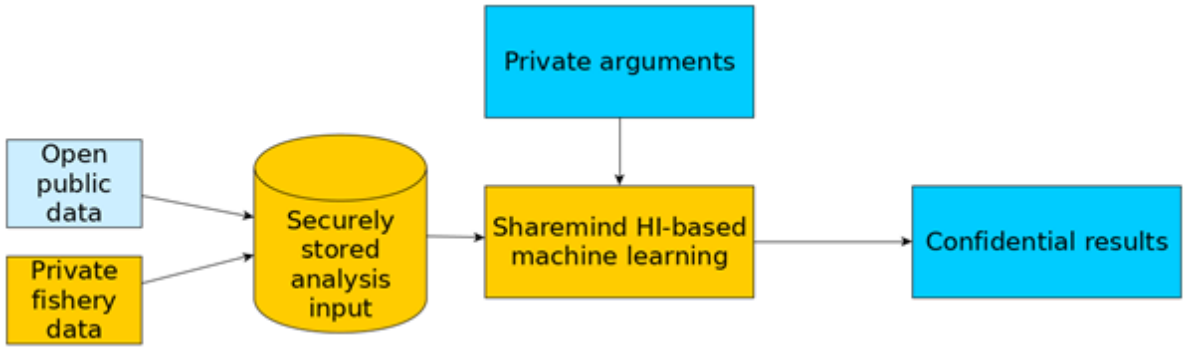


Figure 12: An abstract overview of the proposed Sharemind HI-based solution

The analysis takes into account the following parameters:

- Harbor location
- Distance threshold
- Quantile of best catch
- Size of the ship
- Whether to maximise a single species catch or all species (total biomass output).

2.7.2.1 Model Development

Public catch data was used in the R[19] statistical analysis software to find a method for modelling the data. Since catch size and position varies by season, we could not use linear regression or autoregression for accurate prediction. A local regression method called LOESS was chosen due to its ability to model phenomena not explained by a known function.

The program predicts three variables on a given date: latitude, longitude and catch size by fitting three LOESS regression models. LOESS is a non-linear regression method which was originally developed for smoothing data. It allows one to see trends in scatterplots of noisy data.

LOESS trains a weighted linear regression model for each day by fitting a weighted linear regression model. The point estimated by the trained local model is given as the estimate for that day. A second-degree polynomial was used for local regression.

The user can specify a quantile argument to find the „best" catches to train LOESS models. For example, if the quantile argument is 0.9 then the top 10% data points by catch size are used for training the models. This essentially means that our model will estimate where the best captains are fishing.

The user can also specify their home harbour and a distance threshold to filter out distant locations before fitting the model.

[19] R: The R Project for Statistical Computing

After choosing LOESS, we implemented fitting of LOESS models in both Sharemind MPC and Sharemind HI. We consider experimentation on public or generated data a good practice for finding a suitable model before implementing it using a secure computation technology.

2.7.2.2 User Interface

A web-based interface was developed for the tool. It allows input parties to encrypt and import their data. Fisheries can use the tool to train the predictive model using their parameters.

The user can select the fish species, home harbour, distance threshold, vessel type and top catch quantile. After training the models, the enclave returns three vectors to the client application: latitude curve, longitude curve and catch size curve. The interface will display a map with the estimated position on a given day. The user can change the day with a slider to see how the position changes. The enclave also calculates prediction intervals for the fitted curves which allows the catch area to be displayed as an ellipse.

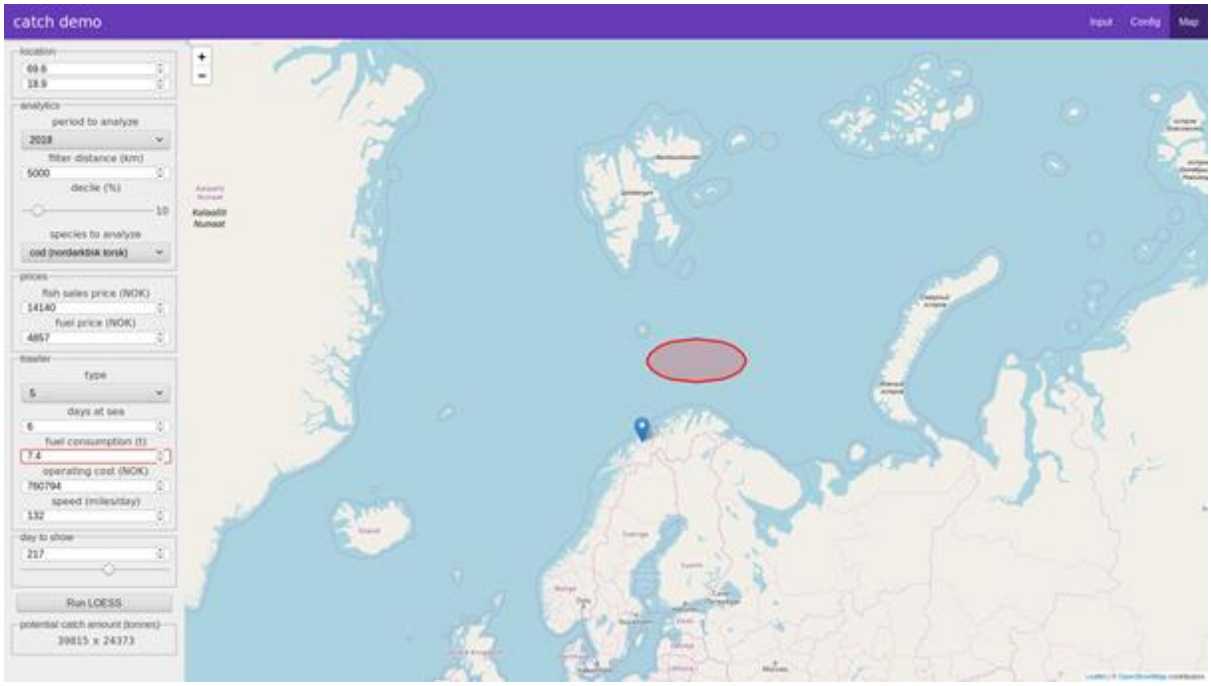


Figure 13: Catch location prediction demonstrator user interface

2.7.2.3 Business Impact

The ability to handle confidential data in privacy-preserving analytics opens up for a number of new applications opportunities, not only in the fishery domain, but also in agriculture and forestry.

There are many situations where sensitive data is not made available because of concerns that the data becomes accessible by competitors or by others that might misuse the data.

The purpose of this demonstrator is to show that it is possible to handle confidential data as part of data analytics, potentially combining open data and confidential data in analytics that both provides business value and preserves data confidentiality.

A wide business impact is foreseen by the example of this demonstrator that shows that this is possible, demonstrating a pipeline that can be reused in future applications where data confidentiality is a concern.

3 DataBio generalized pipelines

3.1 Introduction

Pipelines constitute one of the major exploitable assets of the DataBio project. These were devised with the aim of potential identification of cross reusable (sub) pipelines ("design patterns") that can be used across the pilots of the project and can be applicable to other domains. Approaching the end of the project we would like to test whether this aim was achieved. In other words, we would like to check to what extent we can generalize over the identified pipelines. Generalization can happen at two levels: *technical* and *conceptual*. At the technical level, it means identifying sub-pipelines, e.g.; parts of the technical architecture, that are common to more than one pilot. At the conceptual level, it means the identification of sub-pipelines that are applicable to one or more domains of the projects and beyond.

3.1.1 Top level generic pipeline

The following figure depicts a top-level pipeline, following the DataBio Data Value chain, that is abstract enough, so that it can be specialised in order to describe more specific pipelines, depending on the type of data and the type of processing (e.g. IoT data real-time processing). This top-level pipeline contains the four major steps which are depicted in Figure 14 and analysed next.

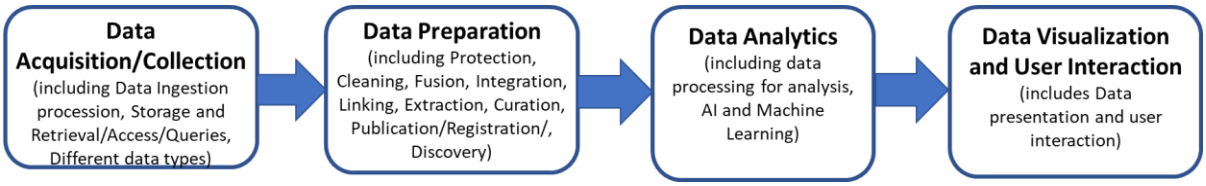


Figure 14: Top Level Generic Pipeline

These steps are in compliance with the activities described in the Reference Architecture for Big Data Application Providers provided in the NIST Big Data Interoperability Framework [REF-40]. A brief description of these steps follows next.

Data Acquisition/Collection

In general, this step handles the interface with the data providers and includes the transportation of data from various sources to a storage medium where it can be accessed, used, and analysed by an organization. Tasks in this step, depending on application implementations, include accepting or performing specific collections of data, pulling data or receiving pushes of data from data providers and storing or buffering data. The cycle Extract, Transform, Load (ETL)/Extract, Load, Transform (ELT) can also be included in this step. At the initial collection stage, sets of data are collected and combined. Initial metadata can also be created to facilitate subsequent aggregation or look-up methods. Security and privacy considerations may also be included in this step, since authentication and authorization activities as well as recording and maintaining data provenance activities are usually

performed during data collection. Last, we would like to note that tasks in this step may vary, depending on the type of the collected data.

Data Preparation

Tasks performed in this step include data validation, like for example checking formats, data cleansing, such as removing outliers or bad fields, extraction of useful information, organization and integration of data collected from various sources, leveraging metadata keys to create an expanded and enhanced dataset, annotation, publication and presentation of the data in order to be available for discovery, reuse and preservation, standardization and reformatting, or encapsulating. Also, in this step, source data are frequently persisted to archive storage and provenance data are verified or associated. The transformation part of the ETL/ELT cycle could also be performed in this step, although advanced transformation is usually included in the next step which is related with data analytics. Optimization of data through manipulations, such as data deduplication and indexing, could also be included here in order to optimize the analytics process.

Data Analytics

In this step, new patterns and relationships, which might be invisible, are discovered so as to provide new insights. The extraction of knowledge from the data is based on the requirements of the vertical application which specify the data processing algorithms. This step can be considered as the most important step as it explores meaningful values, and thus, it is the basis for giving suggestions and making decisions. Hashing, indexing and parallel computing are some of the methods used for Big Data analysis. Machine learning techniques and Artificial Intelligence are also used here, depending on the application requirements.

Data Visualisation and User Interaction

Data can have no value without being interpreted. Visualization assists in the interpretation of data by creating graphical representations of the information conveyed, and thus adding more value to data. This is due to the fact that the human brain processes information much better when it is presented in charts or graphs rather than on spreadsheets or reports. Thus, visualization is an essential step as it assists users to comprehend large amounts of complex data, interact with them, and make decisions according to the results. It is worth to note that effective data visualization needs to keep a balance between the visuals it provides and the way it provides them so that it attracts users' attention and conveys the right messages.

In the following sections, the above steps of the top level pipeline are specialised based on the different data types used in the various project pilots, and are set up differently based on different processing architectures, such as batch, real-time/streaming or interactive. Also, with Machine learning there will be a cycle starting from training data and later using operational data.

The following sections describe the generic/reusable pipelines that have been extracted from the various project pilots, which are:

- IoT data real-time data and decision-making generic pipeline
- Linked Data Integration and Publication generic pipeline
- Earth Observation (raster) and Geospatial/Spatiotemporal (vector) generic pipeline
- Forestry data management/support generic pipeline
- Genomics generic pipeline
- Privacy-aware analytics generic pipeline
- Fisheries decision support in catch planning generic pipeline

Each identified generic pipeline is depicted with an end-to-end data flow diagram the steps of which are mapped to the steps of the top-level generic pipeline described above.

At the end of each section, the project pilots, where each identified generic pipeline has been applied to, are also presented and some conclusions are derived.

3.2 Generic pipeline for IoT data real-time processing and decision-making

3.2.1 General

The “Generic pipeline for IoT data real-time processing and decision making” is an example of a pipeline pattern that fits the two aspects of generalization. It has been applied to three pilots in the project from the agriculture and fishery domain, and it can also be applied to other domains as discussed in the summary section. The main characteristic of this generic pipeline is the collection of real-time data coming from IoT devices to generate insights for operational decision making by applying real-time data analytics on the collected data.

Figure 15 depicts the common data flow among three pilots of the project, two in agriculture (“Prediction and real-time alerts of diseases and pests breakouts in crops” and “Cereals and biomass crop” pilots) and one in fishery (“Monitoring, real-time alerts, and visualization for operation efficiency in tuna fishery vessels” pilot).

Streaming data from IoT sensors are collected in real-time, for example: agricultural sensors, machinery sensors, fishing vessels monitoring equipment. These streaming data (a.k.a. events) can then be pre-processed in order to lower the amount of data to be further analysed. Pre-processing can include filtering of the data (filtering out irrelevant data and filtering in only relevant events), performing simple aggregation of the data, and storing the data (e.g. on cloud or other storage model, or even simply as a computer’s file system) such that conditional notification on data updates to subscribers can be done. After being pre-processed, data enters the complex event processing (CEP) component for further analysis, which generally means finding patterns in time windows (temporal reasoning) over the incoming data to form new more complex events (a.k.a. as situations or alerts/warnings). These complex events are emitted to assist in decision-making process either carried out by humans (“human in the loop”) or automatically by actuators (e.g., sensor that starts irrigation in a greenhouse as a result of a certain alert). The situations can also be displayed using visualization tools to assist humans in the decision-making process. The idea is that the

detected situations can provide useful real-time insights for operational management (e.g.; preventing a possible crop pest or machinery failure).

Figure 15 shows the end-to-end flow. In essence, all components except the data producers (i.e., sensors) and a data consumer (either human or automatic) can be optional. The level of analysis of the data and its level of abstraction is driven by the specific use case. Sometimes, some filtering on the data is enough, while in other cases, the CEP component performs all types of analysis in a central manner. Communication between the software components is performed using standard RESTful APIs while communication between IoT devices and the *Real-time data collection* component is based on standard IoT communication protocols (e.g. MQTT).

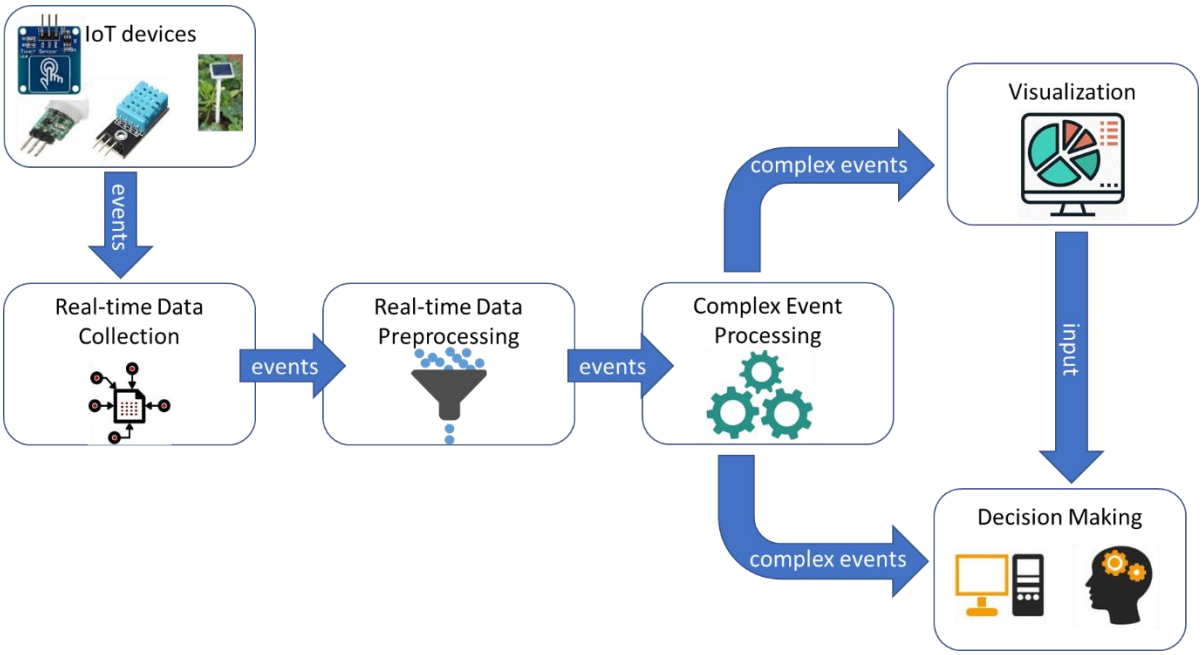


Figure 15: Data flow for real-time IoT data processing and decision-making generic pipeline

Figure 16 depicts the mapping of the steps of the top-level pipeline to the above generic pipeline for real-time IoT data processing and decision making.

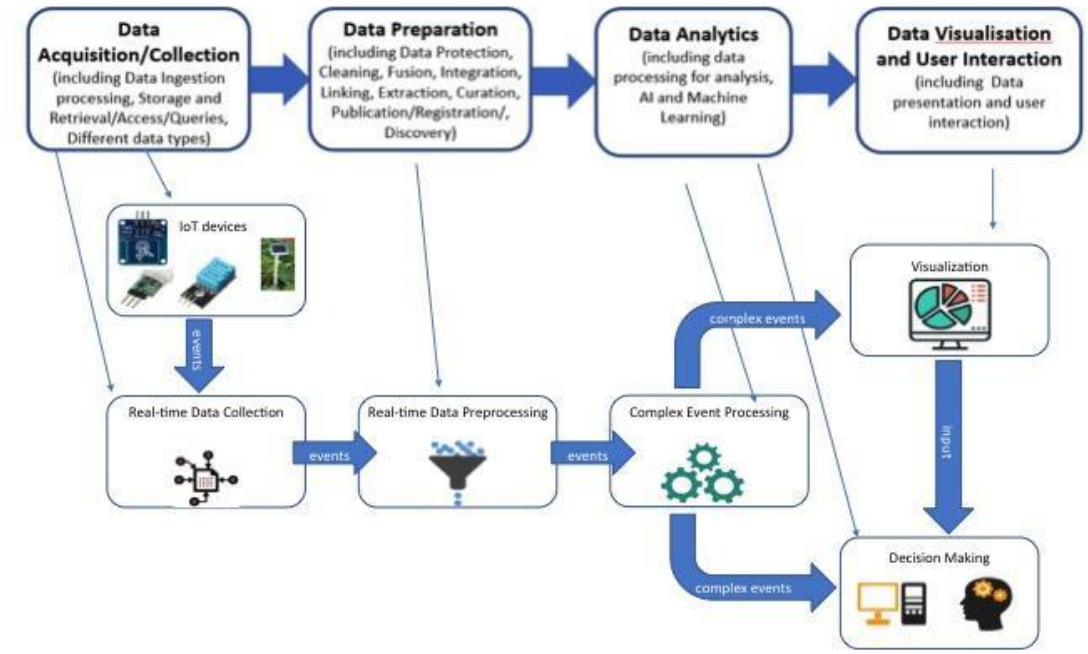


Figure 16: Mapping of the steps of the top-level pipeline (depicted in Fig. 12) to the steps of the generic pipeline for Data flow for real-time IoT data processing and decision-making

3.2.2 Instances of this generic pipeline in DataBio

As aforementioned, the generic pipeline is a generalization of three of the project's pilots. To show this, for each of these three pilot's pipelines we show the mapping of the different generic components into the pilot's pipeline diagram.

Two important notes:

1. The component of the IoT devices is not mapped since the component view diagrams only include software components in the pilot pipeline. We denote for each pipeline the IoT sensors used in the description of the mappings that follows each pilot pipeline.
2. In the DataBio project, we apply the IBM PROactive Technology ONline (PROTON) complex event processing (CEP) tool from partner IBM. Therefore, this is the engine applied in all the pilots that developed an event-based solution for real-time data analytics. However, in the generic case, any other CEP engine can fit as well.

3.2.2.1 A1.1 pilot: Prediction and real-time alerts of diseases and pests breakouts in crops

A1.1 pipeline involves continuous monitoring and prediction of diseases and pests outbreaks on crops. This pipeline is associated with pilot A1.1: Precision agriculture in olives, fruits, grapes (see pilot description in D1.2).

URL in DataBio Hub:

https://www.databiohub.eu/registry/#service-view/WP%201%20Pilot%201%20[A1.1]%20Precision%20agriculture%20in%20olives,%20fruits,%20grapes

This integrated solution collects, validates and stores sensor data through partner NP's GAIAtrons stations. These perform initial processing, monitoring and cross-checking of the data though NP's GAIABus DataSmart RealTime Subcomponent, and push the validated values to PROTON for further analysis in order to enable the detection of trends in data indicating disease or pest outbreaks in real-time. PROTON performs temporal reasoning (processing of data/events in time windows) to derive a more complex event in the form of alerts and warnings. The derived alerts are displayed to the end-user using a visualisation component.

The mapping of the identified generic components into pilot A1.1 component view is shown in Figure 17 and described next.

- **IoT devices:** data from agriculture fields collected by the NP's GAIAtrons stations. The Gaiatron Station is an IoT "Deploy-and-Forget" platform incorporating a wide variety of sensors intended for the continuous surveillance of cultivation environment variables listed in Table 13 below.

Table 13: Data types monitored by Gaiatron Station's.

Variable(s) and (Units) measured
Temperature (C^{o})
Relative Humidity (%)
Barometric Pressure (mBar)
Wind Direction (^{o})
Sun Radiation - Pyranometer (Volts)
Rain Collector (0.2 mm increments- Auto Emptying)
Wind Velocity (km/h)
Leaf Temperature (C^{o})
Leaf Wetness Sensor (Binary -wet/dry)
Leaf Relative Humidity (%) (at two different locations of the field)
Soil Moisture (%) (at up to 7 different depths, depending on the crop needs)
Soil Salinity/Conductivity (dS/m) (at up to 7 different depths depending on the crop needs)

Soil Temperature (C°)

- **Real-time data collection:** is performed by the NP's GAIABus DataSmart RealTime Subcomponent which receives the streams of sensor data from the IoT Gaiatron stations, performs a set of activities including real-time pre-processing, monitoring, validation and cross-checking. The subcomponent is designed for cloud-based operation (constitutes a component of GAIA Cloud), and uses a custom data exchange syntax along with the interconnection of IoT devices and the subcomponent's server-side applications that is designed to optimally address the needs of the offered smart farming applications.
- **Real-time data pre-processing:** is performed by GAIA Cloud which acts as data warehousing, fusion and interfacing component. GAIA Cloud is responsible for: a) storing data at a central point; b) fusing data (risk factor) for the iterative evolution, continuous training and exploitation of tailored pest/disease breakout scientific models, adapted to the specific characteristics of the selected crop types (olive trees, grapes and peaches) and microclimates; and c) managing interfaces, so that this data becomes accessible to the CEP engine for further analysis.
- **CEP:** PROTON's event-driven application searches for patterns to detect potential pests and disease threats to the crops. To this end, more sophisticated analysis is conducted on top of the calculated output of the scientific models (risk factor), exploiting notions of temporal reasoning.
- **Visualization:** The alerts and warnings emitted by the CEP engine are reported back to GAIA Cloud via standard REST API for their visualization. This is effectively managed using a set of Rapid Application Development (RAD) tools for the integrated visualization of spatial and non-spatial data.
- **Decision making:** The alerts and warnings serve as a decision-making tool for the end-user in order to decide whether to take any pro-active measures regarding pest/disease management (e.g. spray).

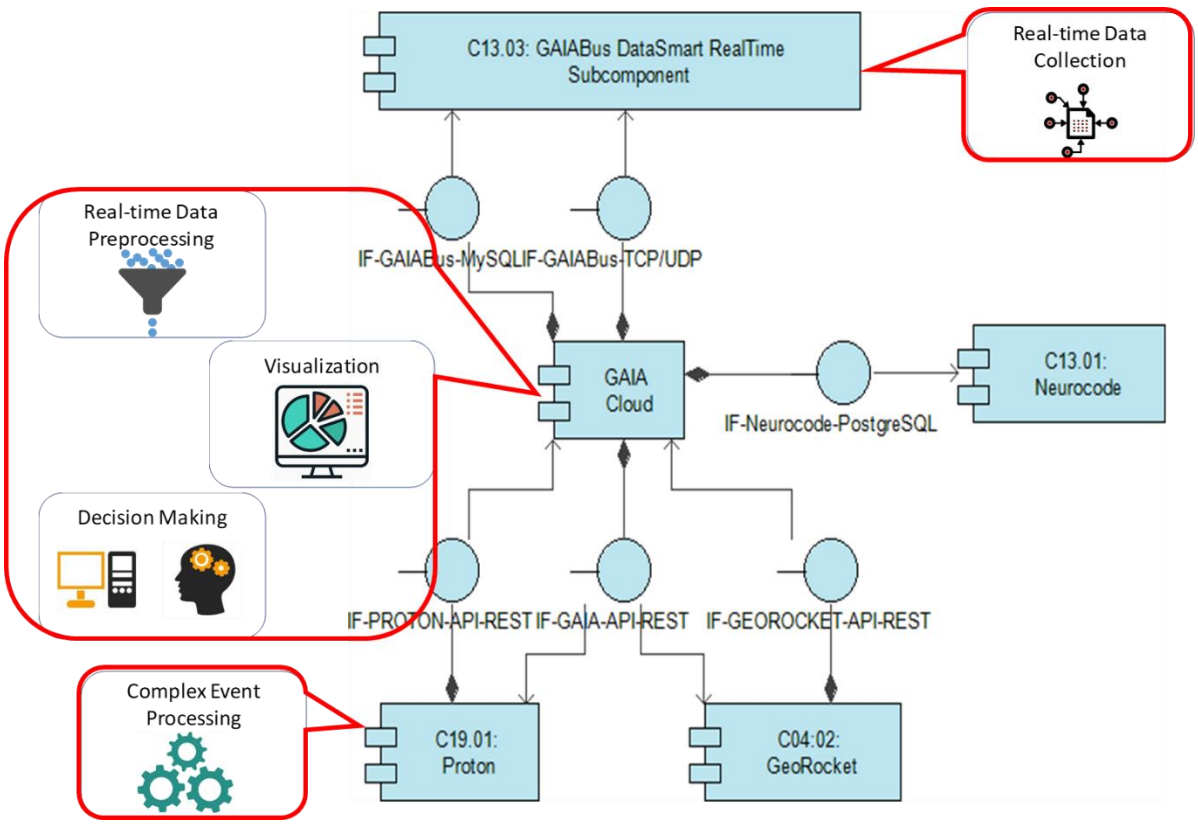


Figure 17: Mapping of generic components into pilot A1.1 component view

3.2.2.2 B1.1 pilot: Cereals and biomass crop

B1.1 pilot focuses on environmental conditions monitoring (such as leaf moisture, light, air pressure) to detect and predict possible problematic conditions for the crops and taking actions to prevent damage. This pipeline is associated with pilot B1.1: Cereals and biomass crop (see pilot description in D1.3 [REF-39]).

URL in DataBio Hub: https://www.databiohub.eu/registry/#service-view/WP%201%20Pilot%205%20[B1.1]%20Cereals%20and%20biomass%20crop/0.0.1

The integrated solution includes IoT power resources like Raspberry PI connected to environmental greenhouse sensors. The data is forwarded and stored in IoT Hub and Orion context broker (from FIWARE EU project), from which the relevant data, based on smart notification settings is forwarded to PROTON CEP engine. PROTON forwards the detected alarms and warnings events back to Orion Context Broker for their storing.

The mapping of the identified generic components into pilot B1.1 component view is shown in Figure 18 and described next.

- **IoT devices:** collected by the analogue environmental sensors (e.g. environmental temperature and humidity, soil moisture, dendrometer and leaf humidity).
- **Real-time data collection**: performed by the Raspberry PI component connected to analogue environmental sensors via digital transformators.

- **Real-time data pre-processing**: performed by FIWARE IoTHub and Orion context broker. The data from Rasberry PI flows to Orion Context broker where it is stored as CB entities. Each change on the entity level is forwarded to subscribers, in this case to the CEP engine.
- **CEP:** PROTON detects patterns indicating potential problems with the plants in the greenhouse (such as not enough water, not enough light) in order to initiate corrective actions (start aggregation/stop aggregation)
- **Visualization:** Dashboards have been created using Grafana's platform to visualise sensors real-time data as well as historical data.
- **Decision making:** The high-level events detected by the CEP engine are reported back to FIWARE Orion context broker component via standard REST API to be stored as alarm entities. From this point on all interested parties subscribed to alarms (such as visualisation components) will receive a notification on the alarms entities to enable the end user to take corrective actions.

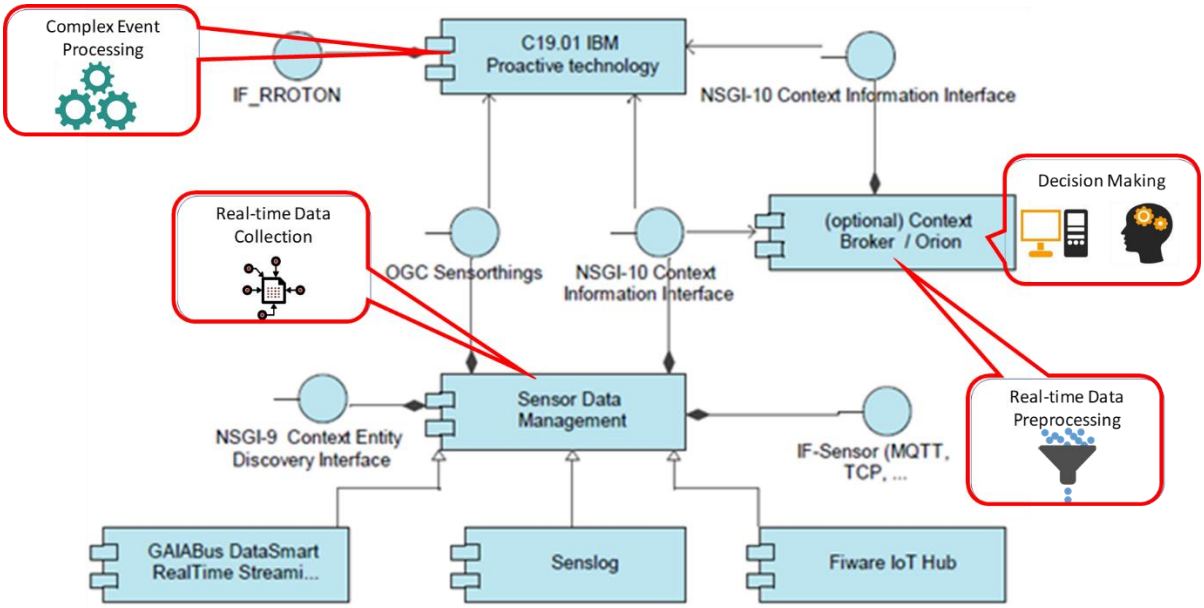


Figure 18: Mapping of generic components into pilot B1.1 component view

3.2.2.3 A1 pilot: Monitoring, real-time alerts, and visualization for operation efficiency in tuna fishery vessels

This pilot aims at improving the efficiency of operations in tuna fishing vessels by implementing a solution which leverages a machinery monitoring system on board as well as offline solution for energy efficiency and machinery condition monitoring. In the operational on-board system measurements such as electric generation/consumption and fuel oil consumption are recorded on computer on board, forwarded using FTP protocol to a file system from which they are consumed by the CEP component. The CEP component installed onboard searches in real-time for possible conditions which might lead to engine faults. In case such conditions are met and alerts or warnings are detected, these are displayed to the

crew on the vessel so that proactive actions can be taken to prevent such situations (see pilot description in D3.3).

URL in DataBio Hub: https://www.databiohub.eu/registry/#service-view/WP3%20%E2%80%93%20Monitoring,%20real-time%20alerts,%20and%20visualization%20for%20operation%20efficiency%20in%20tuna%20fishery%20vessels/0.0.1

In this integrated solution, data collected from engine sensors by the data logger component onboard is forwarded to a computer onboard using FTP, where it is stored in a designated file system using agreed upon format. The data are read by the FileReader component and sent to the CEP engine using provided RESTful API. Once the data are read, they are deleted from the gathering system. PROTON event driven application processes the engine parameters values and searches for potential problematic situations. In case these are detected, the warning and alarms are forwarded via a standard REST API call to the Open VA's visualisation dashboard (from partner VTT) for alerting the ship's operators regarding probable threats and malfunctions. The warnings and alarms are stored in a laptop for further analysis if required.

The mapping of the identified generic components into pilot A1.1 component view is shown in Figure 19 and described next.

- **IoT devices:** Data are collected by the ship's onboard monitoring system.
- **Real-time data collection**: Is performed by the data logger component
- **Real-time data pre-processing**: Data from the data logger is forwarded via FTP to laptop's file system from which it is read by FileReader component. Notifications in the form of RESTful API invocations are sent to CEP component.
- **CEP:** The goal of the event-driven application is the monitoring of engine parameters to alert regarding potential engine problems before these can cause any critical condition to the engine, and therefore to the vessel. The found alarms and warnings are reported to the visualisation dashboard and stored in the computer for historical traceability.
- **Visualization:** The high-level situations detected by the CEP engine are reported to Open VA's visualisation component for alerting the ship's operator of potential engine faults and threats.
- **Decision making:** Based on the dashboard alerts and warnings, operators on board can take proactive actions.

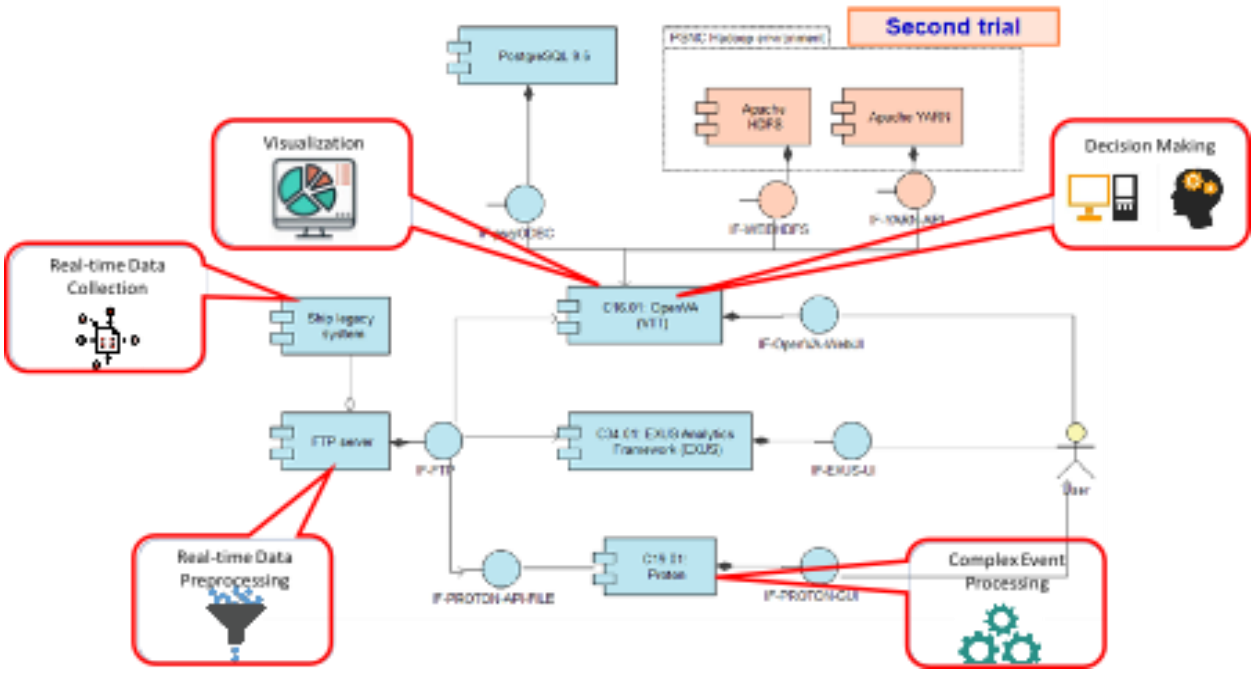


Figure 19: Mapping of generic components into pilot A1 component view (Trial 2)

3.2.3 Summary

The "Generic pipeline for IoT data real-time processing and decision making" is an example of a pipeline pattern that fits the two aspects of generalization. Technically, it has been applied to three pilots in the project from the agriculture and fishery domain and, as such, can be seen as a "pipeline design pattern". Conceptually, it can also be applied to other domains beyond fishery and agriculture. Basically, each use case from any domain in which data is collected from IoT sensors and is analysed in real-time to provide real-time alerts for operational decision making can nicely fit this generic pipeline.

For example: sensor readings from a supply chain scenario in which objects are monitored for track and trace can be collected to further being processed by a CEP engine in order to detect potential delays. The detected situations can be displayed to operators so they can take actions in case such delays are detected (e.g., reschedule trajectory). Another use case can be found in a classical manufacturing process, in which machinery sensors are monitored, aiming at detecting potential machinery failures. The sensor data in the factory is collected and transmitted to a CEP engine that once finds potential failure situations and emits these alerts for decision making (e.g., stop the machine or replace a part in a machine).

3.3 Generic pipeline for linked data integration and publication

3.3.1 General

In DataBio project and some other agri-food projects Linked Data has been extensively used as a federated layer to support large scale harmonization and integration of a large variety of data collected from various heterogeneous sources and to provide an integrated view on

them. The triplestore populated with Linked Data during the course of DataBio project (and few other related projects) resulted in creating a repository of over 1 billion triples, being one of the largest semantic repositories related to agriculture, as recognized by the EC innovation radar naming it the "Arable Farming Data Integrator for Smart Farming". Additionally, projects like DataBio have also helped in deploying different endpoints providing access to the dynamic data sources in their native format as Linked Data by providing a virtual semantic layer on top of them.

This action has been realised in DataBio project through the implementation of the instantiations of a 'Generic Pipeline for the Publication and Integration of Linked Data", which have been applied in different uses cases related to the bioeconomy sectors. The main goal of these pipelines instances is to define and deploy *(semi-) automatic processes* to carry out the necessary steps to transform and publish different input datasets for various heterogeneous sources as Linked Data. Hence, they connect different data processing components to carry out the transformation of data into RDF [REF-27] format or the translation of queries to/from SPARQL [REF-28] and the native data access interface, plus their linking, and including also the mapping specifications to process the input datasets. Each pipeline instance used in DataBio is configured to support specific input dataset types (same format, model and delivery form), and they are created with the following goals:

- Capability of a pipeline to be directly re-executable and re-applicable (e.g. extended/updated datasets)
- Easy reusability of a pipeline
- Easy adaptation of a pipeline for new input datasets
- Automatic execution of a pipeline as far as possible, though the final target is to create fully automated processes
- Pipelines should support both (mostly) static data and dynamic data (e.g. sensor data)

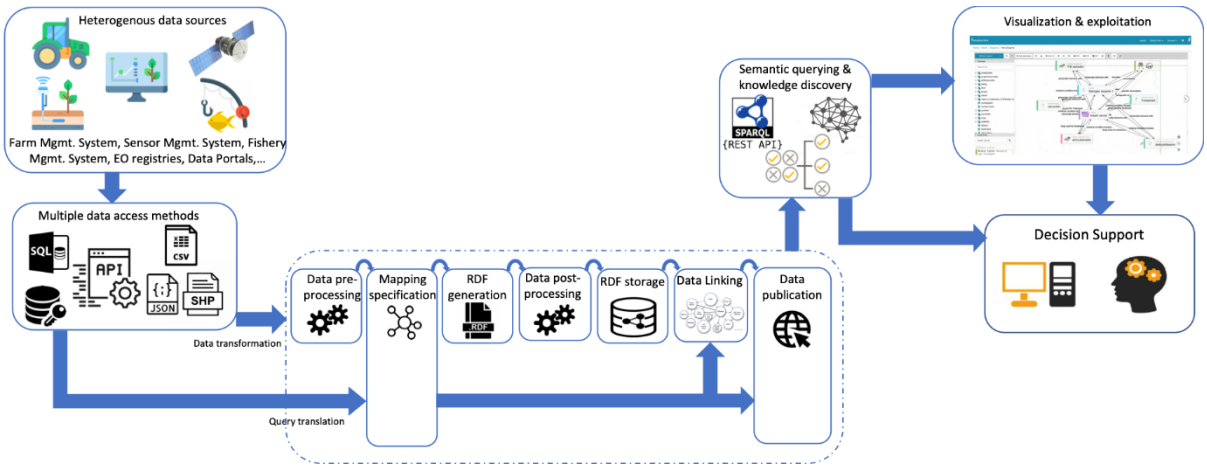


Figure 20: Generic flow for Linked Data Integration and Publication pipeline

A high-level view of the end-to-end flow of the generic pipeline is depicted in Figure 20. In general, following the best practices and guidelines of Linked Data Publication [REF-29],[REF-30], the pipeline:

i. takes as input selected datasets that are collected from heterogeneous sources (shapefiles, GeoJSON, CSV, relational databases, RESTful APIs),
ii. curates and/or pre-process the datasets when needed,
iii. select and/or create/extend the vocabularies (e.g., ontologies) for the representation of data in semantic format,
iv. process and transform the datasets into RDF triples according to underlying ontologies,
v. perform any necessary post-processing operations on the RDF data, vi) identify links with other datasets, and
vi. publish the generated datasets as Linked Data and applying required access control mechanisms.

The transformation process depends on different aspects of the data like the format of the available input data, the purpose (target use case) of the transformation and the volatility of the data (how dynamic is the data). Accordingly, the tools and the methods used to carry out the transformation were determined firstly by the format of the input data.

- Tools like D2RQ[20] were normally used in case of data coming from relational databases,
- tools like GeoTriples[21] was chosen mainly for geospatial data in the form of shapefiles,
- tools like RML Processor[22] for CSV, JSON, XML data formats,
- services like Ephedra[23] (within Metaphactory platform) for Restful APIs.

The list of relevant components identified and used in each pipeline instance will be discussed in the later subsections of this deliverable. Choosing the most suitable tools for the transformation of the source data also depends on the targeted usage of the transformed Linked Data and the goal of accessing the data integrated with other datasets. Finally, based on how often the data is changing (i.e. rate of change), the transformation methods and the related tools are to be further determined. Based on these characteristics, there are two main approaches for making the transformation for a dataset:

- Data upgrade or semantic lifting, which consists of generating RDF data from the source dataset according to mapping descriptions and then storing it in semantic triplestore (e.g., Virtuoso).
- On-the-fly query transformation, which allows evaluating SPARQL [REF-28] queries over a virtual RDF dataset, by re-writing those queries into source query language

[20] http://d2rq.org/
[21] http://geotriples.di.uoa.gr/
[22] https://github.com/IDLabResearch/RMLProcessor
[23] https://www.metaphacts.com/ephedra

according to the mapping descriptions. In this scenario, data physically stay at their source and a new layer is provided to enable access to it over the virtual RDF dataset. This applies mainly to highly dynamic relational datasets (e.g. sensor data) or RESTful APIs.

In every transformation process regardless of the method or tools are chosen, a mapping specification has to be defined to specify the rules to map the source elements (e.g., tables columns, JSON elements, CSV columns, etc.) into target elements (e.g., ontology terms). Generally, this specification is an RDF document itself written in RML[24]/R2RML[25] (and extensions) languages and/or non-standard extensions of SPARQL, e.g., in the case of the Tarql CSV to RDF transformation tool[26].

The resulting datasets can thereafter be exploited through SPARQL queries or through various user interfaces. Some examples of these interfaces include:

- SPARQL endpoint interface to execute queries available in PSNC cloud infrastructure (https://www.foodie-cloud.org/sparql)
- Faceted search interface to navigate the linked datasets available at http://www.foodie-cloud.org/fct/
- Map visualisation via HSLayer demo applications, e.g., http://app.hslayers.org/project-databio/land/
- On Metafactory platform also including hybrid services like Ephedra: http://metaphactory.foodie-cloud.org/resource/Start

3.3.1.1 Generic pipeline

The following diagrams provide i) a simplified top-level representation of the Linked Data Integration and Publication pipeline aligned with the top-level generic pipeline, ii) the pipeline view with specific components of the generic pipeline, respectively.

[24] http://rml.io/spec.html
[25] https://www.w3.org/TR/r2rml/
[26] https://tarql.github.io/

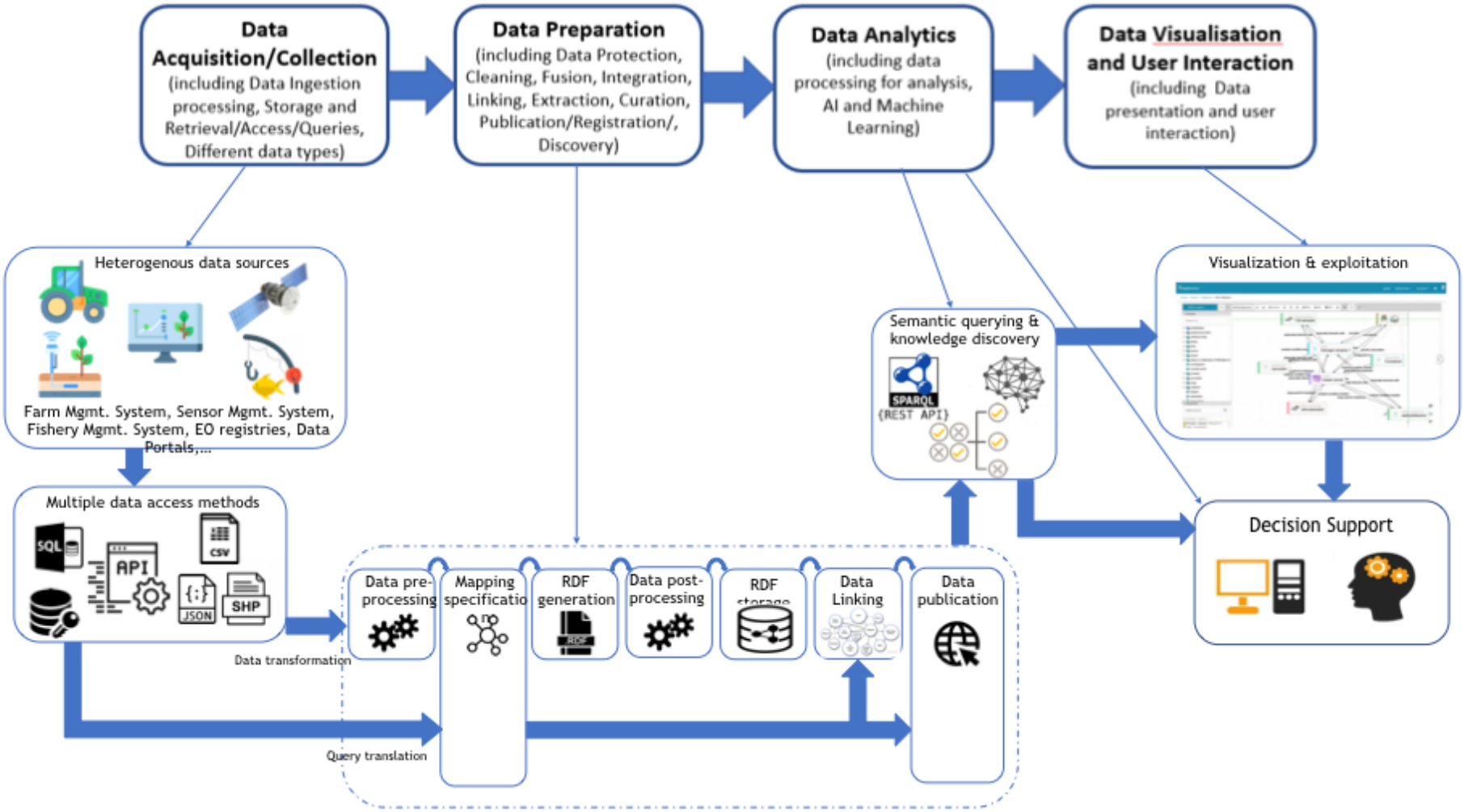


Figure 21: Generic flow for Linked Data Integration and Publication pipeline aligned with top-level generic pipeline

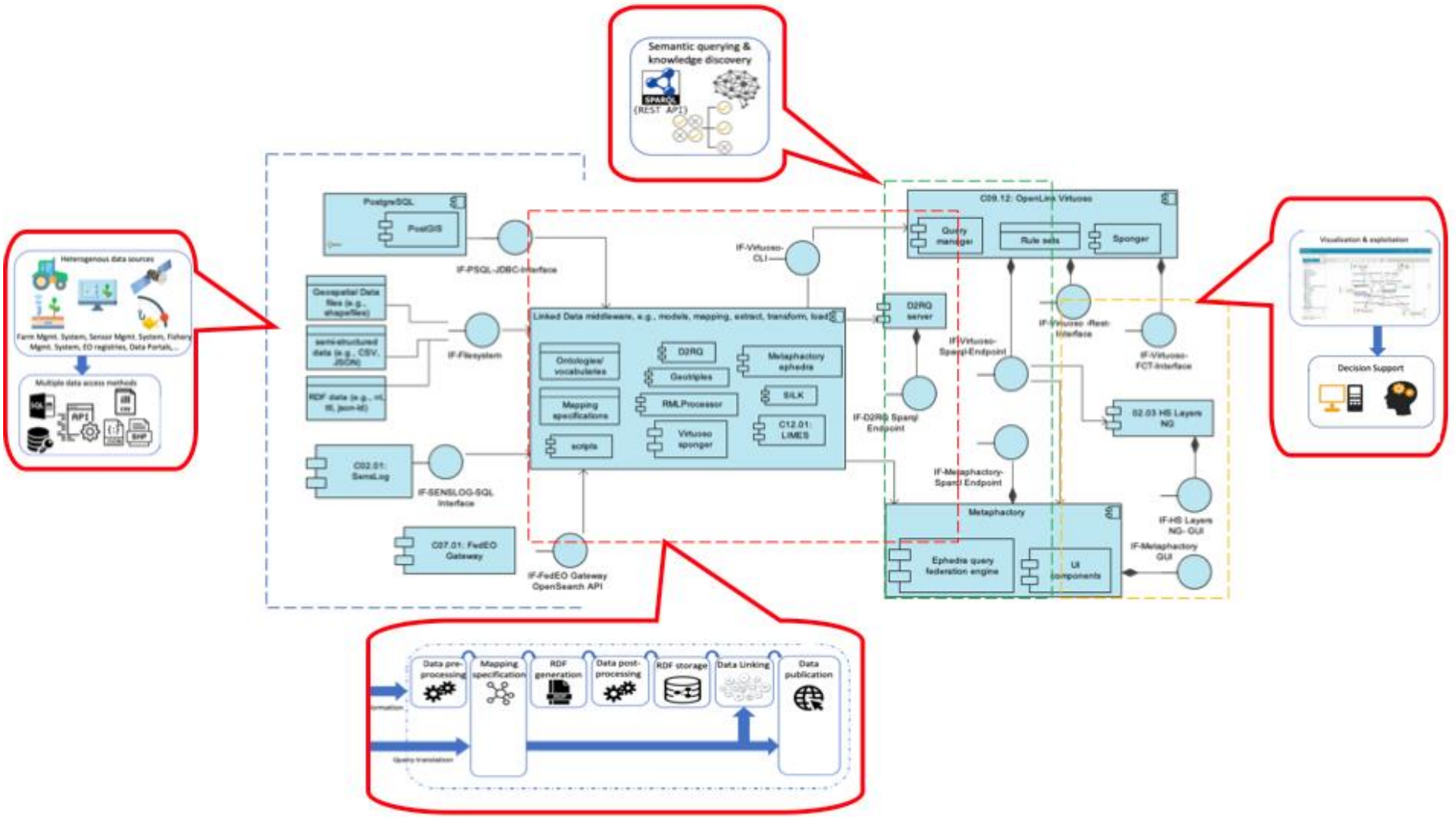


Figure 22: Generic Linked Data publication pipeline component diagram

The URL link of the generic pipeline in the DataBioHub is at:
https://www.databiohub.eu/registry/#service-view/WP1%20-%20Publication%20of%20Linked%20Data%20from%20heterogeneous%20sources%20in%20DataBio%20sectors/0.0.1

Linked Data pipelines implemented and deployed in DataBio project include many different tools and data processing components, as well as models, mappings, and a large number of linked datasets generated in the course of the project. The main tools and applications used during the various stages of the process are summarized below:

- **Virtuoso[27] (incl. sponger, faceted browser):** OpenLink Virtuoso (open source edition of Virtuoso Universal Server)[28] is a middleware and database engine hybrid that combines the functionality of a traditional RDBMS (Relational Database Management System), ORDBMS (Object-Relational Database Management System.), virtual database, RDF, XML, free-text content management & full-text indexing, linked data server, web application server and file server functionality in a single system. Primarily the RDF store is the most important feature of Virtuoso for the provisioning of a semantic database (triplestore) as a service to store the RDF data, as well as for their publication as Linked Data. Along with the RDF triplestore Virtuoso also provides a SPARQL query language support, SPARQL protocol supports inline SPARQL integration within SQL, use of bitmap indices for optimizing storage and management of RDF triples, implementation of the HTTP-based Semantic Bank API that enables client applications to post to its RDF Triple Store, and several RDF insert methods, including http PUT and POST. Virtuoso SPARQL can be used as an inference context for inferring triples (not physically stored) by supporting some RDF schemas and OWL constraints (e.g., owl:sameAs, rdfs:subClassOf etc.). Virtuoso also includes a faceted search interface which allows simple text search and navigation of RDF data through their links. Virtuoso also includes Sponger, which is a Linked Data middleware component. It generates Linked Data from a variety of data sources, and supports a wide variety of data representation and serialization formats (e.g., standard non-rdf like csv, atom, rss, etc., and vendor-specific like facebook, google+, geonames, openstreetmap, etc.). The Sponger is also a full-fledged HTTP proxy service, directly accessible via SOAP or REST interfaces. Similarly, Virtuoso includes RDBMS-to-RDF mapping functionality (also known as Linked Data Views of SQL data).
- **D2RQ[29]**: D2RQ is a system for accessing relational databases as virtual, read-only RDF graphs. It offers RDF-based access to the content of a relational database without having to replicate it into an RDF store. Using D2RQ we can:
    - query a non-RDF database using SPARQL
    - access the content of the database on the fly as Linked Data over the Web.

[27] https://www.databiohub.eu/registry/#service-view/OpenLink%20Virtuoso/0.0.1
[28] http://vos.openlinksw.com/owiki/wiki/VOS
[29] http://d2rq.org/

- create custom dumps of the database in RDF formats for loading into an RDF store
- access information in a non-RDF database using the Apache Jena API

D2R Server uses a customizable D2RQ mapping to map database content into this format and allows the RDF data to be browsed and searched which are the two main access paradigms to the Semantic Web. This on-the-fly translation allows publishing of dynamic RDF data from large live databases and eliminates the need for replicating the data into a dedicated RDF triple store (e.g dynamic sensor data).

- **Geotriples**[30]: Its tool for transforming geospatial data from their original formats (e.g., shapefiles or spatially-enabled relational databases) into RDF. The following input formats are supported: spatially-enabled relational databases (PostGIS and MonetDB), ESRI shapefiles and XML, GML, KML, JSON, GeoJSON and CSV documents. GeoTriples comprises two main components: the mapping generator and the R2RML/RML mapping processor. The mapping generator takes as input a geospatial data source (e.g., a shapefile) and creates automatically an R2RML or RML mapping that can transform the input into an RDF graph which uses the GeoSPARQL vocabulary. The mapping statements can be customised based on any other data models or ontologies in course of its transformation into RDF triples. In the pipeline processes this tool has proved to be a very effective one as it has a versatile usage in many of the pipelines tasks.
- **RML Processor**[31] : This tool is used to process the mainly CSV type data sources in the pipelines, mostly along with the tool Geotriples mentioned before. In the pipeline tasks the mostly the initial RDF mapping statement was generated using the Geotriples. Thereafter the mapping definition is remodelled and customised as per the data needs and then transformed by using the initial source data in CSV format mainly.
- **Silk**[32]**:** Silk Workbench is a web application which guides the user through the process of interlinking between different data sources. Silk Workbench offers the following features:
  - Enabling users to manage different sets of data sources, linking tasks and transformation tasks.
  - Offers a graphical editor which enables the user to easily create and edit linking tasks and transformation tasks.
  - Silk Workbench makes it easy for the user to quickly evaluate the links which are generated by the current link specification.
  - It allows the user to create and edit a set of reference links used to evaluate the current link specification.

[30] http://geotriples.di.uoa.gr/
[31] https://github.com/RMLio/RML-Processor
[32] http://silkframework.org/

- **LIMES[33]:** LIMES (Link Discovery Framework for metric spaces) is an easy and efficient approach for the discovery of the links between various Linked Data sources. It addresses the scalability problem of link discovery by utilizing the triangle inequality in metric spaces to compute estimates of instance similarities. Based on these approximations, LIMES can filter out a large number of instances pairs that do not meet the matching condition set by the user. The real similarities of the remaining instances are then computed, and the matching instances are returned. Large-scale link discovery in Linked Data sources based on the characteristics of metric spaces computes pessimistic approximations of the similarity between instances of different data sources and filters out the instances that do not meet the mapping conditions sufficiently well (non-related data instances).
- **Geo-L[34]:** This is a tool for discovery of geo-spatial links that retrieves specific properties of spatial objects from source and target datasets, through their respective SPARQL endpoints, and finds topological relations between objects in source and target objects according to topological predicates. The specifications of the relevant properties are provided in a configuration file, which allows constraining the number of the object by specifying offset and limit. A dataset can be created through properties that already exist in the graph, and, in addition, Geo-L allows direct construction of ad-hoc values through a SPARQL SELECT statement for a given resource.
- **Hslayers NG[35]:** Hlayers NG is a web mapping library written in JavaScript. It extends OpenLayers 4 functionality and takes basic ideas from the previous HSlayers library, but uses modern JS frameworks instead of ExtSJS 3 at the frontend and provides better adaptability. That's why the NG (Next Generation) is added to its name. HSLayers is open-sourced and is built in a modular way which enables the modules to be freely attached and removed as far as the dependencies for each of them are satisfied. The dependency checking is done automatically. In the case of the pipelines, the tool was mostly used as a visualization tool for exploiting and showcasing the various integration possibilities of the Linked Data from various sources which gave rise to various use cases in many of the above-mentioned projects.
- **Metaphactory[36]:** This platform is one of the most widely used tools in the pipeline mainly for visualization purposes of geospatial data. Metaphactory supports knowledge graph management, rapid application development, and end-user-oriented interaction. Metaphactory runs on top of your on-premise, cloud, or managed graph database and offers capabilities and features to support the entire

[33] https://www.databiohub.eu/registry/#service-view/LIMES/0.0.1
[34] https://github.com/DServSys/geo-L
[35] https://www.databiohub.eu/registry/#service-view/HSLayers%20NG/0.0.1
[36] https://www.metaphacts.com/product

lifecycle of dealing with knowledge graphs. Metaphactory's generic approach based on open standards offers great flexibility in different usage scenarios and across various industries and application areas. Moreover, the Metaphactory platform also provides the hybrid services for RESTFUL APIs e.g. Ephedra.

- **Ephedra**: Type of API wrapper basically is a SPARQL federation engine aimed at processing hybrid queries, which provides a flexible mechanism for including hybrid services into a SPARQL federation. Ephedra is a component of Metaphactory and an end-to-end Knowledge Graph Platform for knowledge graph management, which facilitates rapid application development and end-user-oriented interaction.
- **Scripts**: Additionally, as in many cases, it is necessary to carry out some pre and/or post data processing tasks, e.g., to correct data format or to clean data, a set of shell scripts have been created. The scripts are available to reuse in DataBio. Some of the most generic ones are to remove empty/incomplete triples with no values generated due to the absence of values while transformation; post-processing scripts for removing duplicate triples and selective removal of the transformed data etc.

Furthermore, several ontologies and vocabularies were reused, including:

1. **FOODIE ontology**[37]: FOODIE ontology is based on INSPIRE schema and the ISO 19100 series standards are most suitable in order to represent and model all aspects of the farm and open data from any related input datasets. Its extension includes data elements and relations from the input datasets that were not covered by the main FOODIE ontology [5] but that were critical for the transformation needs. To ensure the maximum degree of data interoperability, the FOODIE data model [6] is based on INSPIRE based generic data models, especially the data models for Agricultural and Aquaculture Facilities (AF), which was extended and specialized in various projects.
2. **SOSA/SSN**[38]: An ontology for describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties. A lightweight but self-contained core ontology called Sensor, Observation, Sample, and Actuator or SOSA.
3. **RDF Data Cube Ontology**[39] : Data Cube Vocabulary and its SDMX ISO standard extensions were effective in aligning multidimensional survey data like in SensLog. The Data Cube includes well-known RDF vocabularies (SKOS, SCOVO, VoiD, FOAF, Dublin Core).
4. **CatchRecord.owl Ontology**[40]**:** A design pattern for populating an ontology of aquatic species catching records. With this vocabulary, a pattern can be modelled for the kind

[37] http://agroportal.lirmm.fr/ontologies/FOODIE
[38] https://www.w3.org/TR/vocab-ssn/
[39] https://www.w3.org/TR/vocab-data-cube/
[40] http://www.ontologydesignpatterns.org/cp/owl/fsdas/catchrecord.owl

of species and the amount of organisms that have been caught, from which areas/countries and at what date and fishing year.

3.3.2 Instances of the generic pipeline in DataBio

The pipeline, as described in the previous section, is a generalization of multiple instantiations, in particular, two specific project's pilots and four additional experiments in DataBio. In order to show how this generic pipeline has been applied in each use case, we present for each pipeline the pipeline view highlighting the specific methods and components used/applied, along with a description of the task performed and results achieved.

3.3.2.1 Linked Data in agriculture related to cereals and biomass crops

This pilot experiment was focused on the publication of INSPIRE-based agricultural Linked Data from the farm data collected from Czech Pilot 8 (pilot B1.4 in [REF-39]) Cereals and biomass crops from the DataBio project, in order to query and access different heterogeneous data sources via an integrated layer. The input datasets used for this experiment include:

- Farm data having information about each field name with the associated cereal crop classifications and arranged by year (from Rostenice area in the Czech Republic)
- Data about the field boundaries and crop map and yield potential of most of the fields in Rostenice pilot farm from the Czech Republic.
- Yield records from two fields (Pivovarka, Predni areas in Czech Republic) within the pilot farm that were harvested in the years 2017 and 2018.

The source data was collected in the form of shapefiles that were transformed into RDF format and then published as Linked Data, using the underlying model of the FOODIE ontology. The resultant Linked Datasets are available for querying and exploitation through the DataBio SPARQL endpoint deployed at PSNC's HPC facilities.

The description of the tasks carried out along with the components used in the whole process is as follows:

- For defining the data model to transform the input datasets into RDF format, FOODIE ontology which is based on INSPIRE schema and the ISO 19100 series standards was used as the base vocabulary and extended as and when needed (with a Czech pilot extension) in order to represent and model all aspects of the farm and open data from the input datasets. The extension included data elements and relations from the input datasets that were not covered by the main FOODIE ontology, but those were critical for the pilot needs.
- Creation of an RDF mapping definition that specifies how to map the contents of a dataset into RDF triples using the FOODIE ontology and its extensions. In this process, firstly a generic R2RML definition of the mapping file was generated from the input data in the format of shapefiles by using applications like GeoTriples and thereafter manually edited as per the FOODIE data model to generate the final mapping

definition. GeoTriples being the main tool for transformation, was also used to generate the RDF dump from the source data contents.

- The final task involved providing an integrated view of the original dataset. As target datasets were particularly large (especially when considering connections with open datasets), and the connections were not of equivalence (i.e., resources are related via some properties, e.g., geometry, but they are not equivalent) it was decided to use queries to access the integrated data as per need rather than using link discovery tools like SILK or LIMES. Hence cross querying within the datasets was done in Virtuoso SPARQL endpoint for some use cases to establish possible links between agricultural and other related open datasets. The public instance of SILK is present in http://silk.foodie-cloud.org/.
- To visualize and explore the Linked Data in a map different application/system prototypes were created using the component called HSLayers NG, as mentioned earlier. (e.g. https://app.hslayers.org/project-databio/land/). One such visualization is shown as:

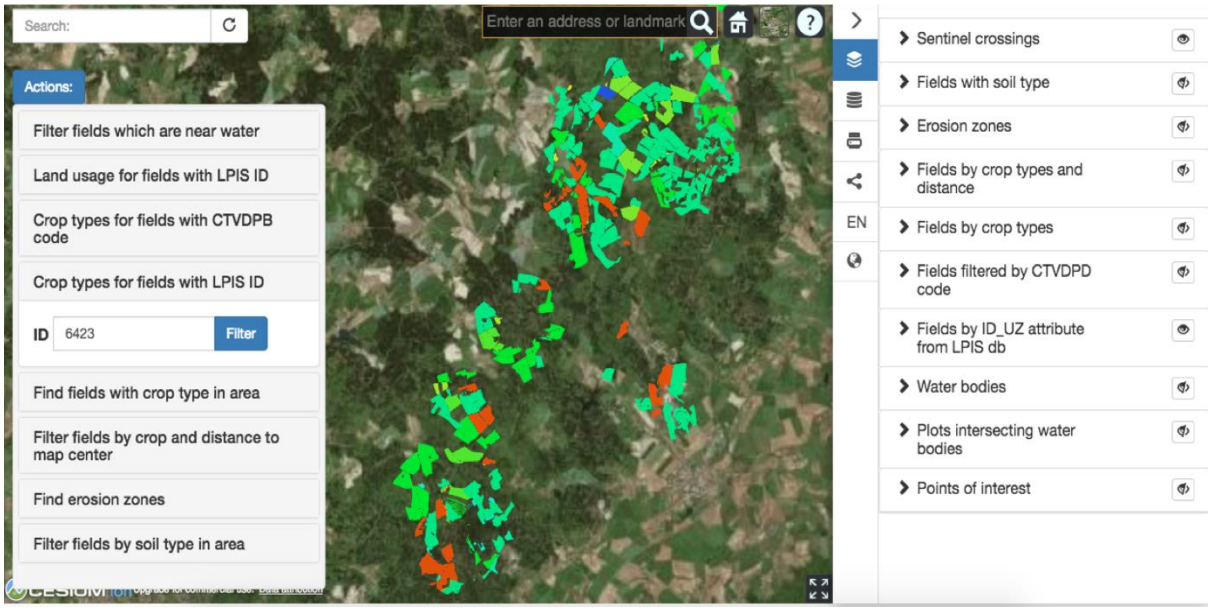


Figure 23: Map visualisation prototype (HSLayer application) - http://app.hslayers.org/project-databio/land/

The resulting linked datasets are accessible via: https://www.foodie-cloud.org/sparql. Figure 24 maps the generic components identified in this pilot. The red highlighted markings indicate the components in use in the pilot.

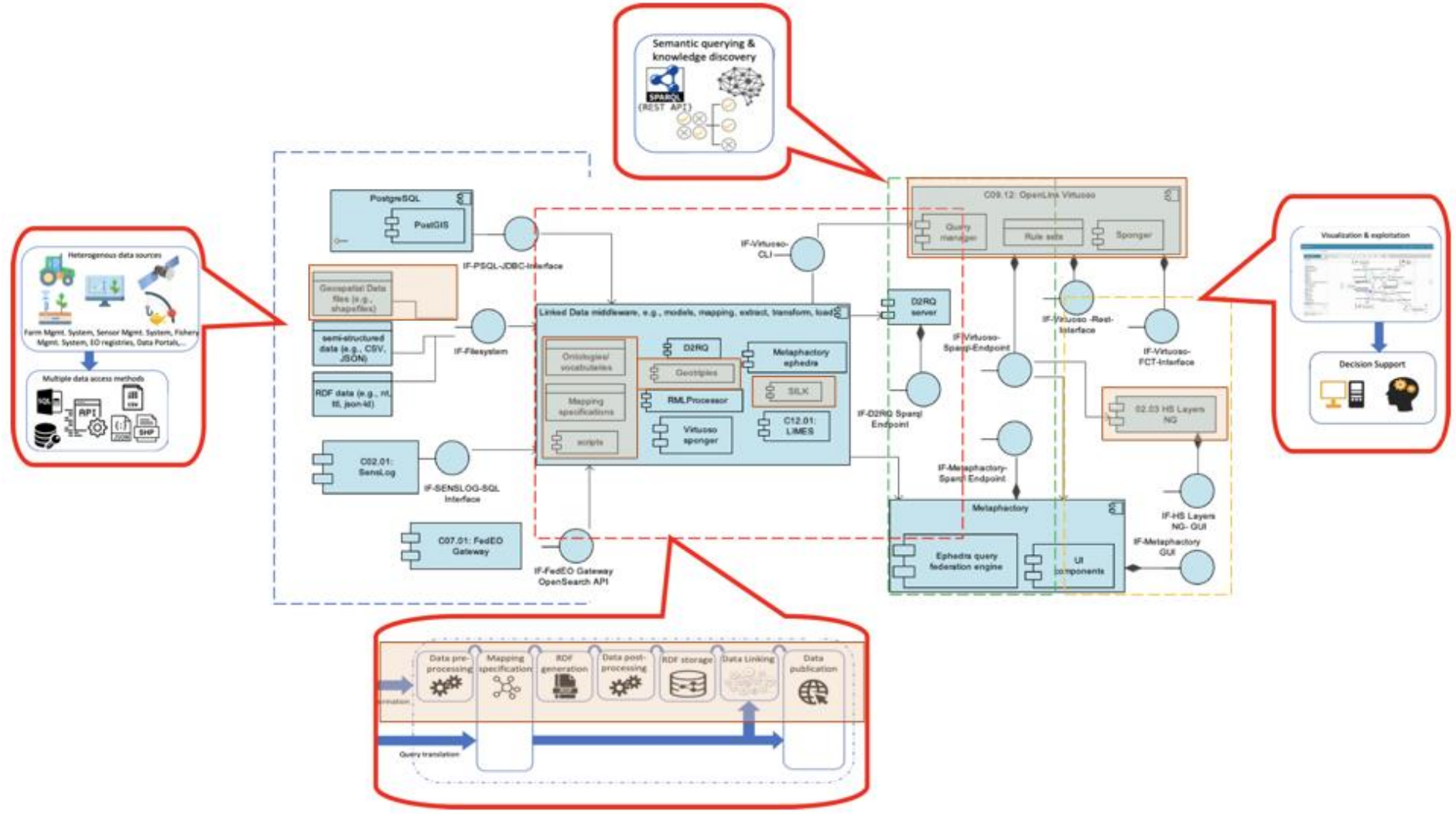


Figure 24: Mapping of the generic components into pilot [B.14] in the pipeline view

3.3.2.2 Linked sensor data from machinery management

This pipeline was performed for Pilot 9 (pilot B2.1 in [REF-39]) Machinery management in the DataBio project where sensor data from the SensLog service (used by FarmTelemeter service) was transformed into Linked Data on the fly, i.e. data stays at the source and only a virtual semantic layer was created on top of it to access it as Linked Data. For modelling the sensor data, the following vocabularies/ontologies were selected:

- Semantic Sensor Network (SSN)[41] ontology for describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties. A lightweight but self-contained core ontology called (Sensor, Observation, Sample, and Actuator) or SOSA was actually used in this specific case to align the SensLog data.
- Data Cube Vocabulary and its SDMX ISO standard extensions were effective in aligning multidimensional survey data like in SensLog. The Data Cube includes well-known RDF vocabularies (SKOS[42], SCOVO[43], VoiD, FOAF[44], Dublin Core[45]).

The SensLog service uses a relational database (PostgreSQL) to store the data. Hence, in the mapping stage, the creation of R2RML/RML definitions required different preprocessing tasks and some on-the-fly assumptions to engineer the alignment between the SensLog database and the ontologies/vocabularies.

Once the mapping file was generated (manually), the RDF Data of the dataset was published using D2RQ server that enables accessing relational database sources as virtual RDF graphs (see Figure 25 and Figure 26). This on-the-fly approach allows publishing of RDF data from large and/or live databases and thus the need for replicating the data into a dedicated RDF triple store is not required. The Linked Data from the sensor data from SensLog (version 1) was published in the PSNC infrastructure in a D2RQ server available at http://senslogrdf.foodie-cloud.org/. The associated SPARQL endpoint to query the data is available at: http://senslogrdf.foodie-cloud.org/sparql.

[41] https://www.w3.org/TR/vocab-ssn/
[42] https://www.w3.org/TR/skos-reference/
[43] http://vocab.deri.ie/scovo
[44] http://www.foaf-project.org/
[45] https://www.dublincore.org/specifications/dublin-core/dces/

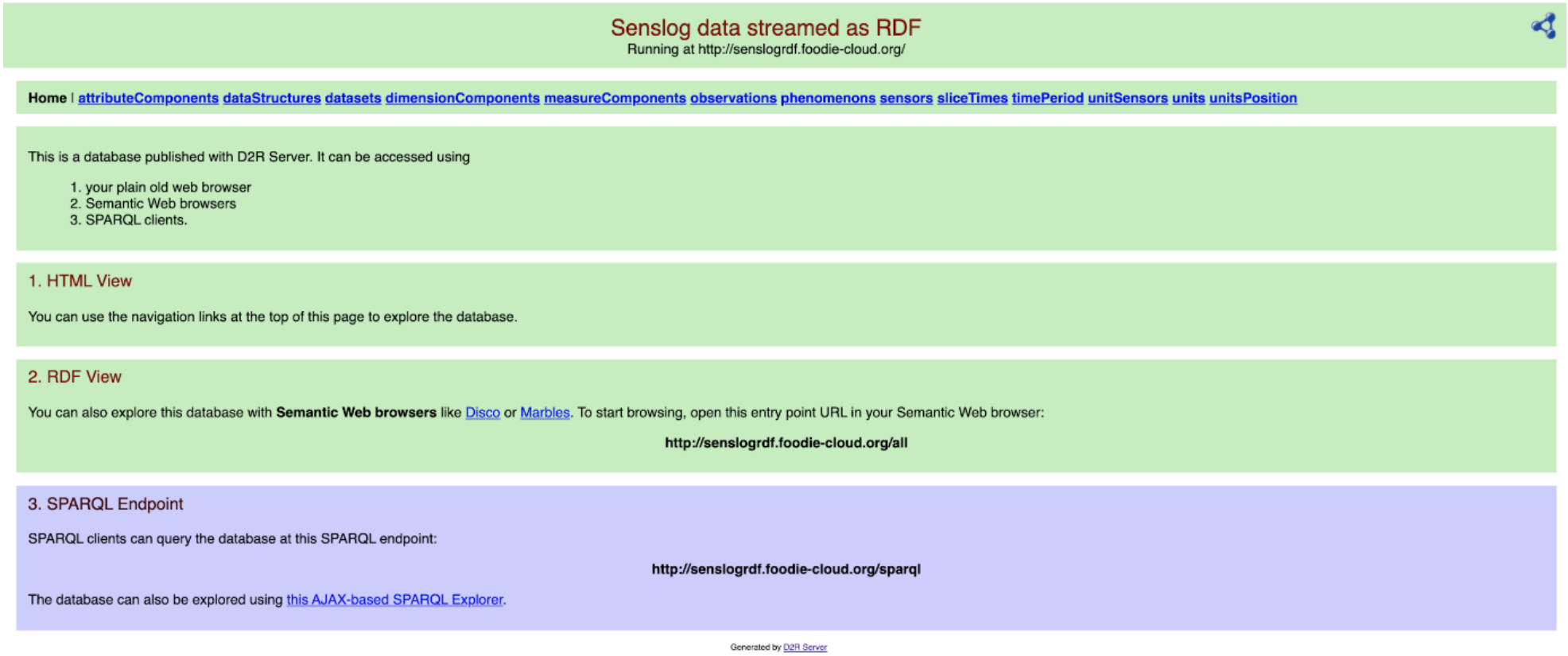


Figure 25: Entry page to the visualization of sensor data as RDF on-the-fly

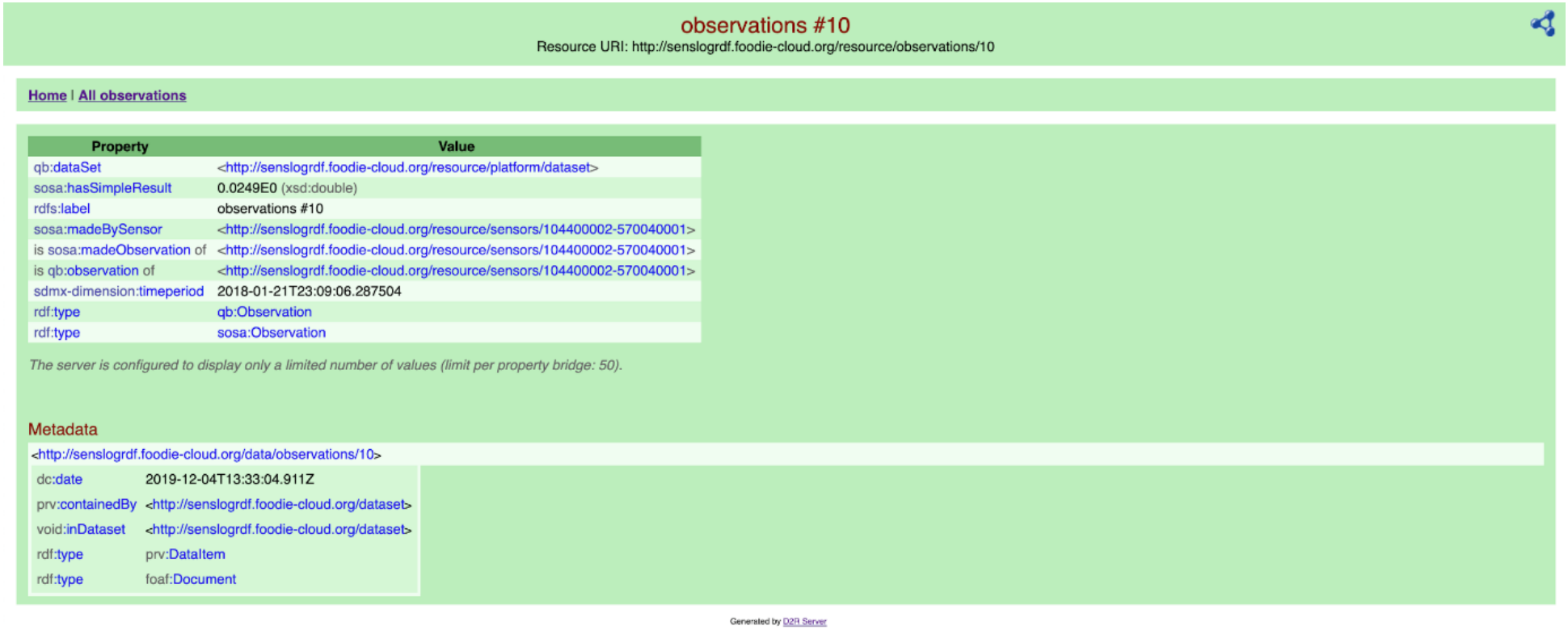


Figure 26: Visualization of an observation details in RDF generated on-the-fly

The figure below highlights the main components used in this pilot from the generic pipeline components.

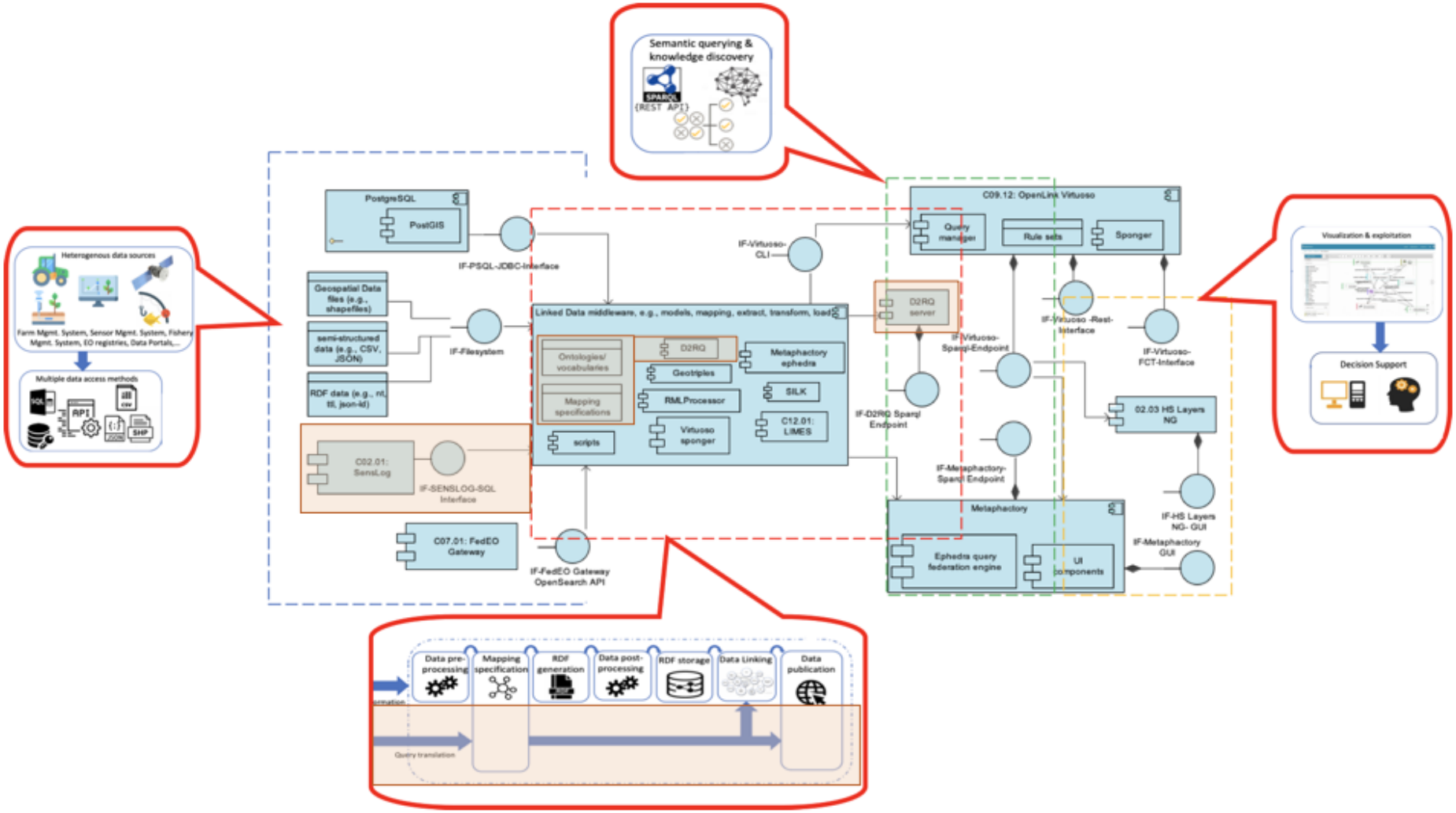


Figure 27: Mapping of the generic components into pilot [B2.1] in the pipeline view

3.3.2.3 Linked Open EU datasets related to agriculture and other bio-sectors

This pipeline includes mainly the EU and national open data from various heterogeneous sources collected from a wide range of applications in the geospatial domain. The purpose was to experiment on the datasets by transforming them into Linked Data and exploiting them on various technology platforms for the purpose of integration and their visualization. The sources for all of these data contents are widely heterogeneous and in various formats (e.g. in shapefiles, CSV format, JSON and in relational databases). The process of transformation required extensive work to identify and determine the most suitable mode for their transformation. This required careful inspection and analysis of the input data contents in order to identify the available ontologies/vocabularies, and any of their extensions if needed, necessary for the representation of such data in RDF format. Additionally, since the source datasets were in different formats, the selection of most suitable tools (mentioned earlier) for their transformation was most vital in order to create the correct (R2RML/RML) mapping definitions. Some of the input datasets, their formats, and the ontologies/vocabularies used for the representation of data in the semantic format are described below along with a brief idea of the whole process:

- Input data regarding land parcel and cadastral data (for Czech Republic & Poland), erosion-endangered soil zones, water buffer and soil type classification, all of which were collected as shapefiles. The ontologies used for the representation of such data included the INSPIRE-based FOODIE ontology as well as different extensions of the ontology created to cover all the necessary information (e.g., erosion zones and restricted area near to water bodies).
- Input data from the Farm Accountancy Data Network (FADN), [46] which was in the form of CSV files, were first modelled and aligned by using ontologies like Data Cube Vocabulary and its SDMX ISO extensions. The preparation of the mapping definitions from the input files required the pipeline shell scripts as mentioned before in order to make it reusable for all FADN data sources in CSV format. Separate CSV files were manually created for each reusable common class type. Once the mapping definitions for each of the CSV were generated, they were integrated to make a reusable common mapping definition covering all the components of the input data.
- The sample data as input from the well-known reviewing site Yelp was also used as a set of JSON files. Different ontologies like review[47], FOAF, schema.org, POI etc. were used to represent the elements from the input data in semantic format during the creation of the mapping definition.
- Other ontologies from previous efforts for the representation of open geospatial datasets like corine, hilucs, olu, otm, urban atlas, were also used. These ontologies are available at https://github.com/FOODIE-cloud/ontology

[46] https://ec.europa.eu/agriculture/rica/
[47] https://vocab.org/review/#

For this process as we can see that the data inputs are from a wide range of heterogeneous sources, thus the generation of the RDF data was carried out using different tools depending on the format of the source data. For example, for shapefiles GeoTriples tool was used, while for the JSON and CSV data the RML Processor tool was used. The resulting RDF datasets were then loaded into Virtuoso triplestore providing SPARQL and faceted search endpoints for further exploitation. Finally, for the provision of an integrated view over the original datasets in case of agricultural and open data, SPARQL queries were generated and possible link discovery was carried out using tools like SILK. For visualization platforms like HS Layers NG and Metaphactory were used.

Figure 1: DataBio Metaphactory (entry page)

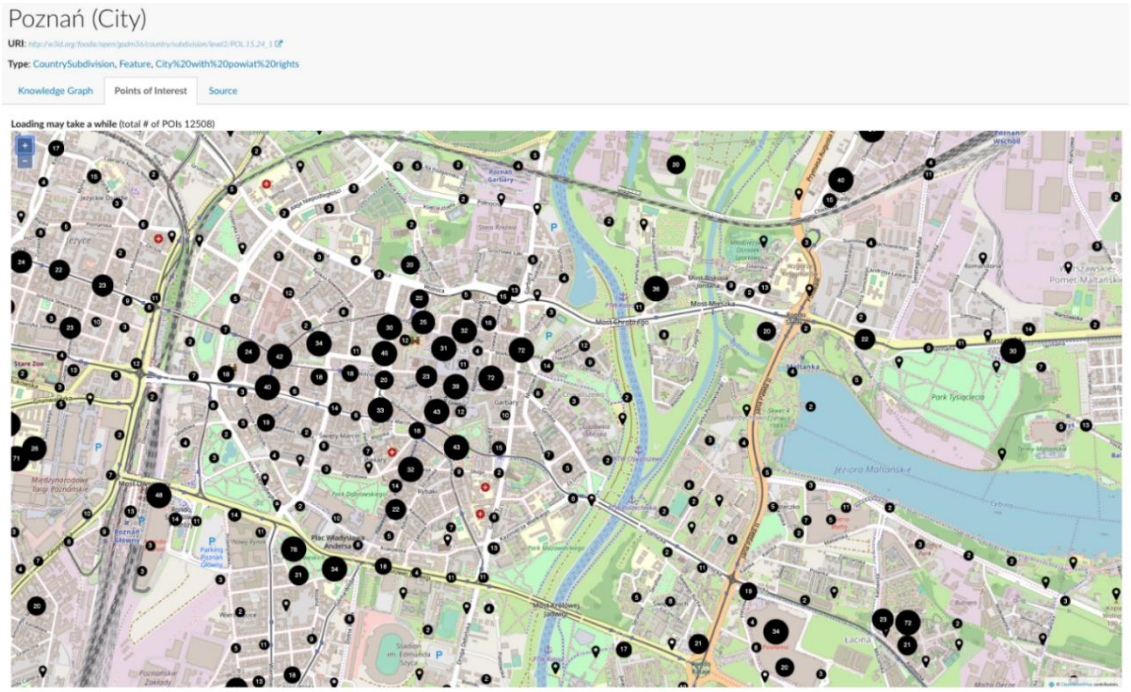


Figure 28: DataBio Metaphactory (map visualisation of points of interest in Poznan city)

The resulting linked datasets are accessible via: https://www.foodie-cloud.org/sparql. Figure 29 highlights the main components used in this pilot from the generic pipeline components.

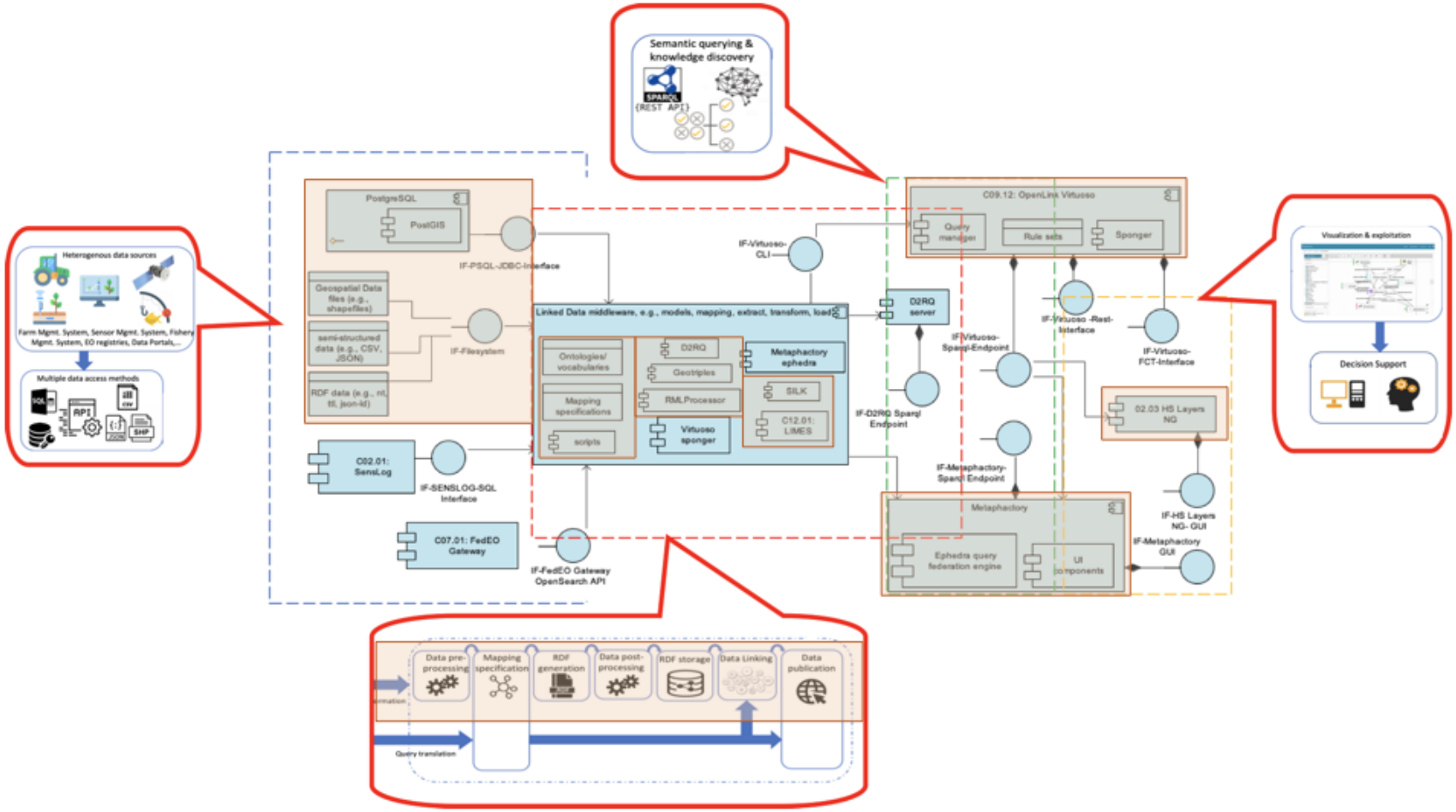


Figure 29: Mapping of the components used in the use case of Linked Open EU-datasets in the pipeline view

3.3.2.4 Linked (meta) data of geospatial datasets

This pipeline developed in DataBio project focuses mainly on the publication of metadata from geospatial data origins as Linked Data. There were two data sources that were transformed.

The first sub-case focused on EO metadata collected from the public Lesproject Micka registry (https://micka.lesprojekt.cz/en/), which includes information of over 100K geospatial datasets. Micka is software for spatial data / services metadata management according to ISO, OGC and INSPIRE standards. It allows retrieving the metadata in RDF forms using Geo-DCAT-AP[48] (an extension of DCAT) for the representation of geographic metadata compliant with the DCAT application profile for European data portals. As such metadata cannot be queried as Linked Data, therefore, the goal was to make it available in the form of Linked Data in order to enable its integration with other datasets, e.g., Open Land Use (OLU). The process for publication was straightforward: a dump of all the metadata in RDF format was generated from Micka, which was then loaded into the Virtuoso triplestore. Some exemplary SPARQL queries were then generated to identify connection points for integration, e.g., get OLU entries and their metadata given a municipal code and type of area (e.g., agricultural lands). The dataset is accessible via: https://www.foodie-cloud.org/sparql.

The second sub-case focused on making Earth Observation (EO) Collections and EO Products metadata available as Linked Data via a SPARQL compliant endpoint which makes a request to non-sparql backends on-the-fly. Hence, it was targeted to enable querying via SPARQL without harvesting all the metadata and storing the data in a triplestore but to access them dynamically via the existing on-line interfaces. This use case for Linked Data for hybrid systems [8] involves the accessing of EO metadata through a system called Federated Earth Observation (FedEO)[49] gateway (by partner Spacebel) which provides interfaces for the access of the EO data. The ESA FedEO gateway provides a unique entry point to a growing number of scientific catalogues and services for EO missions. Regarding the ontologies for EO products metadata, the main idea was to reuse the standard and/or widely used ontologies/vocabularies whenever it is possible and if needed to extend them. In case of FedEO, the metadata returned was already using semantic vocabularies in the (Geo)JSON-LD representation, thus it required only to expose the results as Linked Data. These vocabularies include Dublin Core, DCAT, SKOS, VOID and OM-LITE, as well as terms from the OpenSearch specifications. In this case the component of Ephedra API wrapper was used to access the data by using SPARQL description of a REST Service Signature defining the input and output terms and thereafter configuring a REST Service Repository. Once the RDF data is generated the data can be exposed via a SPARQL endpoint provided in the Metaphactory platform (http://metaphactory.foodie-cloud.org/sparql?repository=ephedra). A demo interface has

[48] https://ec.europa.eu/jrc/en/publication/geodcat-ap-representing-geographic-metadata-using-dcat-application-profile-data-portals-europe
[49] http://ceos.org/ourwork/workinggroups/wgiss/access/fedeo/

also been implemented to visualize the linked data in Metaphactory (entry point: http://metaphactory.foodie-cloud.org/resource/:ESA-datasets).

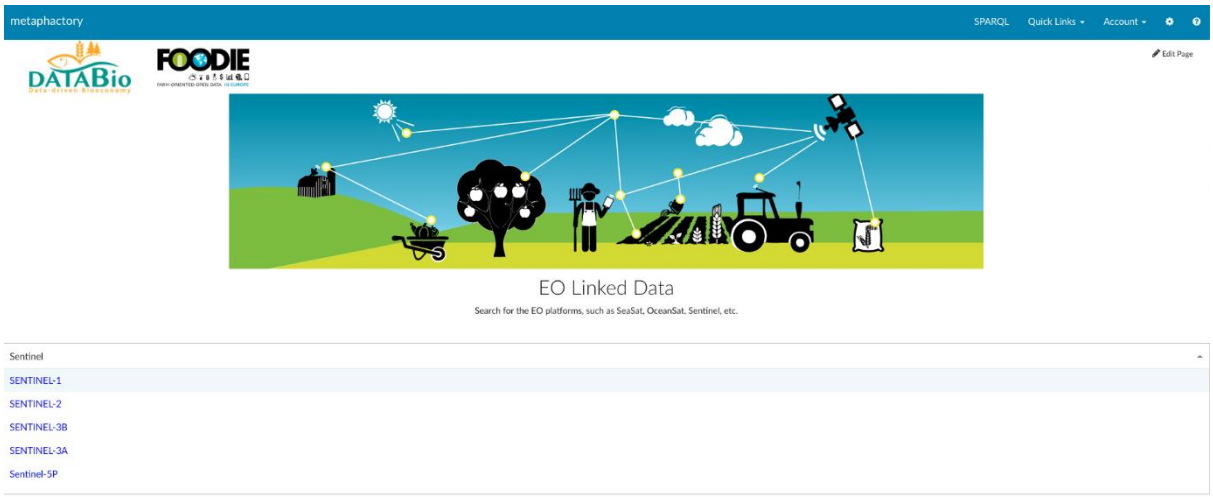


Figure 30: Metaphactory demo application to access FedEO REST API as Linked Data

Figure 31 highlights the main components used in this pilot from the generic pipeline components. In this figure, the components related to the first sub-case (Micka) are highlighted in green, while the components related to the second sub-case (C07.01 FedEO) are highlighted in orange.

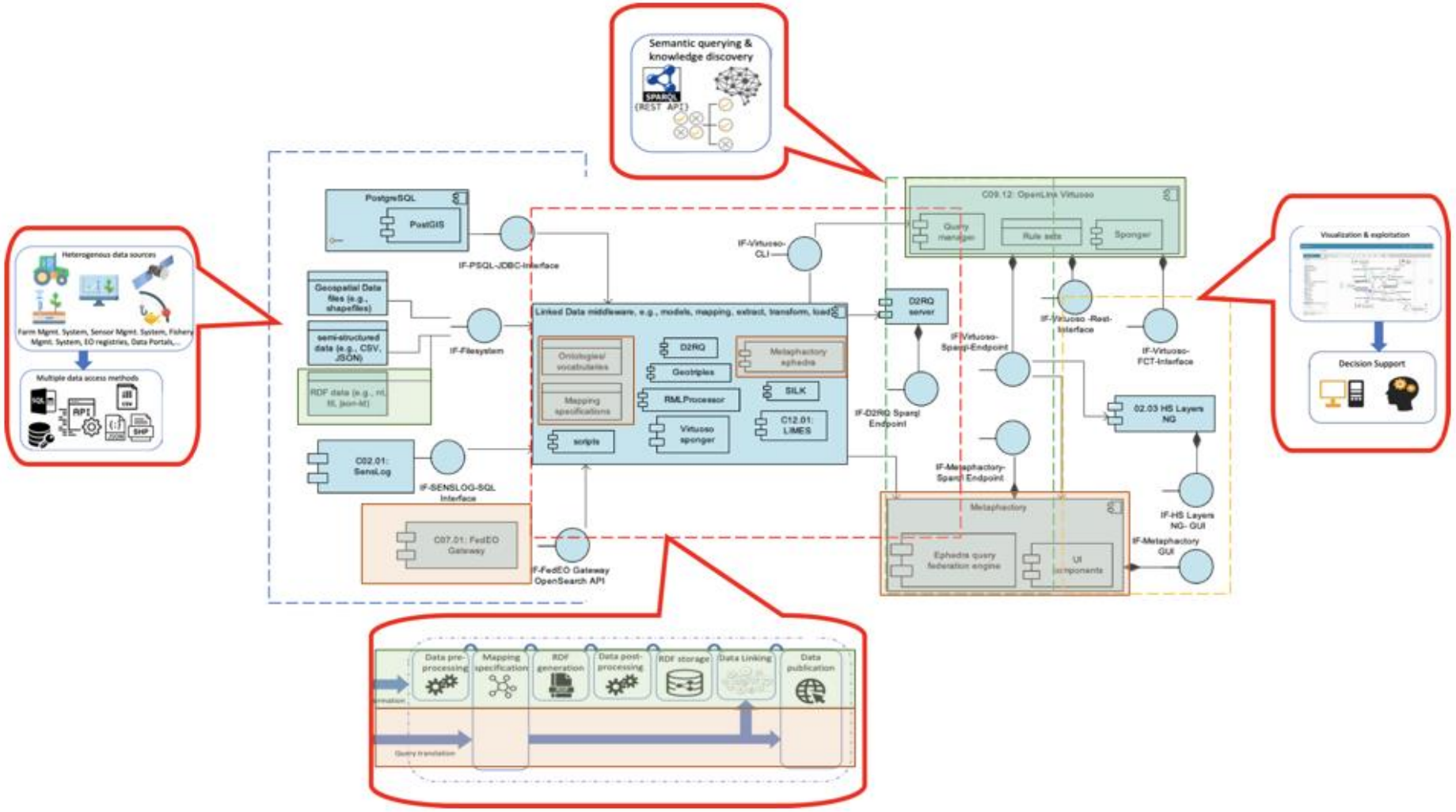


Figure 31: Mapping of the components used in the use case of Linked (meta) data of geospatial datasets in the pipeline view.

3.3.2.5 *Linked data in fishery*

This pipeline for DataBio project focused on the catch record data from the fisheries of Norweign marine regions. The purpose of this pipeline was to publish the catch record data from five years of historical data as Linked Data and perform experimentation operations to exploit and visualize them on various platforms. The input data was in the form of CSV files containing the catch record data of each year. The tasks included in this pipeline are:

- To identify and map which attributes of the data are mostly in line with the transformation procedure and can be mapped with some existing ontology. For the mapping the ontology/vocabulary catch record.owl and mostly an extended version of the vocabulary was used in this case which suits all the relevant data attributes of the main CSV data.
- The CSV files were extensively preprocessed in the process to generate a R2RML/RML mapping definition using the component GeoTriples. The mapping definitions were further analysed and processed to settle with a final mapping definition for the transformation of the CSV data into Linked Data using RML Processor. During the creation of the mapping definitions, the possibility of integration with other Linked Datasets was also considered.
- After the transformation of the Linked Data few post processing were done by using our generic scripts to make the data ready to upload to the Virtuoso Triplestore.
- At present, the catch data from five years were transformed and uploaded to the Virtuoso triplestore providing SPARQL and faceted search endpoints for further exploitation.

To showcase the integration and visualisation of the dataset a web interface using the Metaphactory platform was created, which includes map visualisations and representation of data in the form of charts and graphs. This process is ongoing and more experimentations are ongoing. The interface is presently available at http://metaphactory.foodie-cloud.org/resource/:CatchDataNorway_v2. The resulting linked datasets are accessible via: https://www.foodie-cloud.org/sparql.

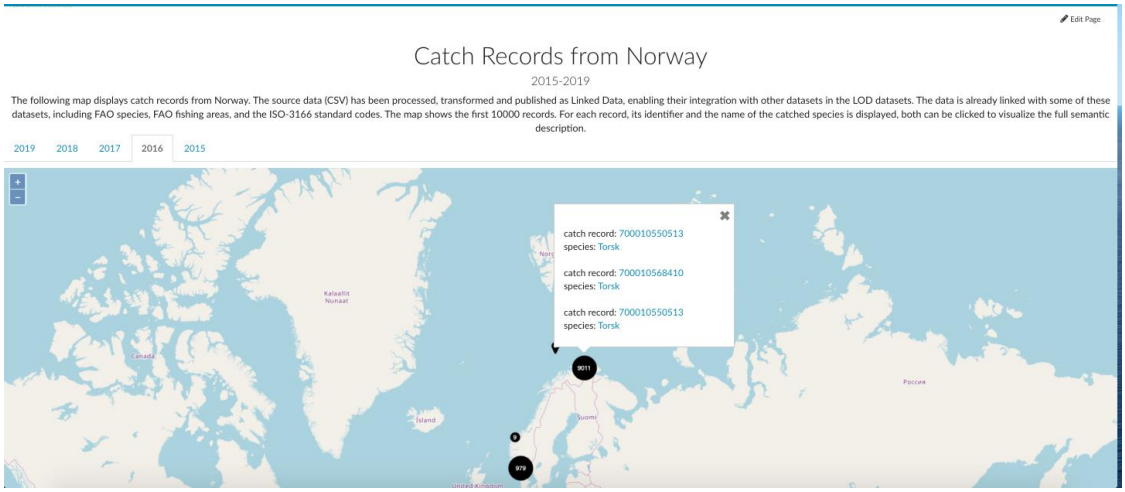


Figure 32: DataBio Metaphactory custom view (map with catch records from Norway)

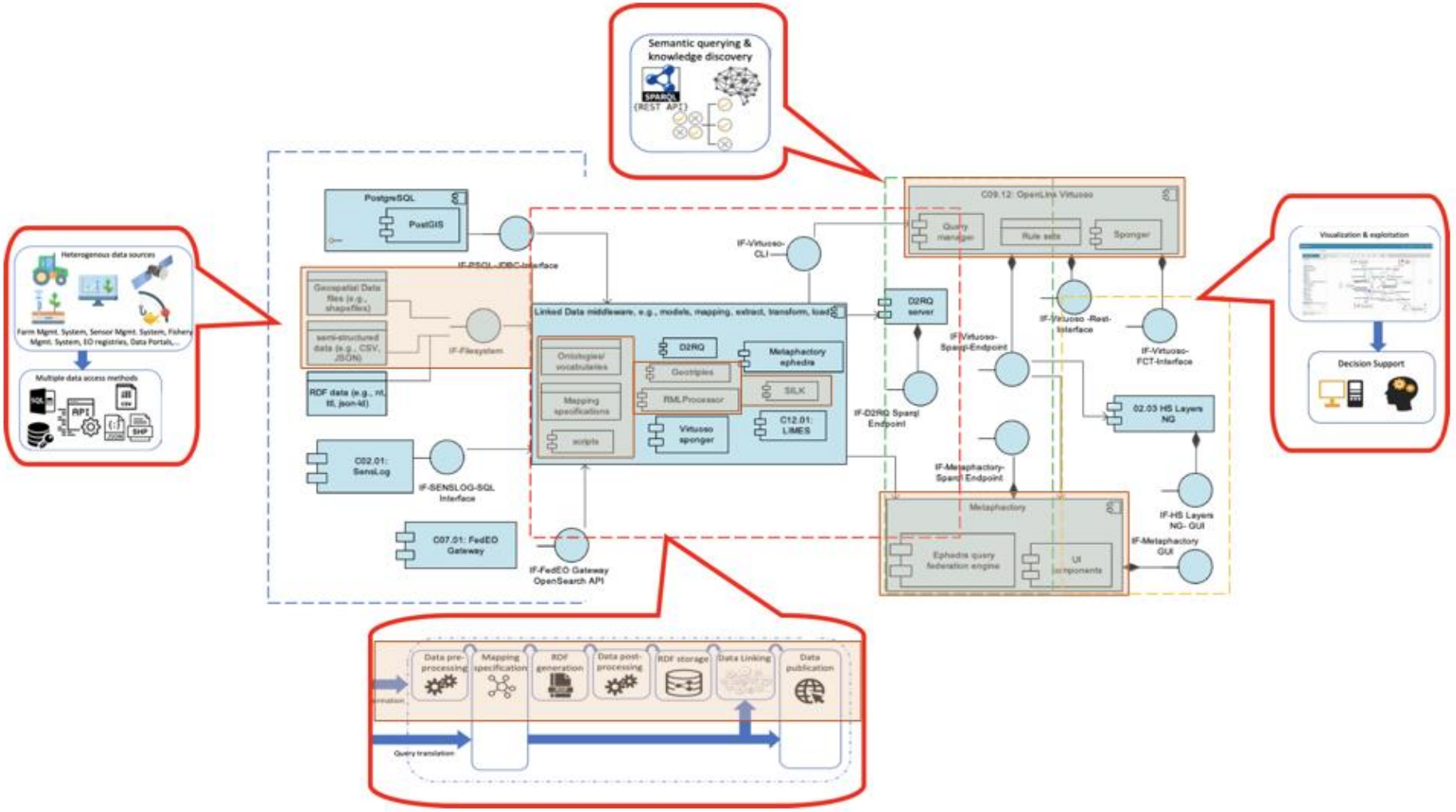


Figure 33: Mapping of the components used in the fishery use case in the pipeline view

3.3.3 Linked datasets

As mentioned above, RDF datasets were deployed in the Virtuoso triple store within PSNC and can be accessed via SPARQL and faceted search endpoints. Currently, the triplestore has over 1 billion triples, being one of the largest semantic repositories related to agriculture, which has been recognised by the EC innovation radar as an agriculture integrator database. The table below shows some of the respective graphs produced by all the pipelines previously described and the number of triples contained in them. The official SPARQL and the faceted search endpoints of the triplestore are: https://www.foodie-cloud.org/sparql and https://www.foodie-cloud.org/fct.

Table 14: RDF graphs produced by pipelines

Graph URI (note: URIs are not resolvable; they can be used to refer to the specific dataset in the triplestore)	**Name of dataset**	**Number of RDF triples**
http://w3id.org/foodie/open/pl/LPIS/{voivodeship}# (where voivodeship in poland = mazowieckie, dolnoslaskie, kujawsko-pomorskie, lodzkie, lubelskie, lubuskie, malopolskie, opolskie, podkarpackie, podlaskie, pomorskie, slaskie, warminsko-mazurskie, wielkopolskie, zachodniopomorskie, swietokrzyskie)	LPIS Poland	727517039
http://w3id.org/foodie/olu# agriculture related lands (hilucs_code<200) in CZ, PL, ES & for main cities in Czech Republic (centers of NUTS3 regions), Poland (agglomeration areas from Urban Atlas) and Spain (agglomeration areas from Urban Atlas)	Open Land Use	127926060
http://w3id.org/foodie/otm# CZ, ES, PL; but RoadLinks only for FunctionalRoadClassValue of type: ('mainRoad', 'firstClass', 'secondClass', 'thirdClass', 'fourthClass') (see http://opentransportmap.info/OSMtoOTM.html)	Open Transport Map	154340785
http://micka.lesprojekt.cz/catalog/dataset#	Open Land Use Metadata	10456676
http://www.sdi4apps.eu/poi.rdf	Smart Points of Interest (SPOI)	407629170
http://w3id.org/foodie/open/cz/pLPIS_180616_WGS#	LPIS Czech Republic	24491282
http://w3id.org/foodie/open/cz/lpis/code/LandUseClassificationValue	LPIS Czech Republic Land Use Classification	83

Graph URI (note: URIs are not resolvable; they can be used to refer to the specific dataset in the triplestore)	**Name of dataset**	**Number of RDF triples**
http://w3id.org/foodie/atlas# agriculture related lands (hilucs_code<200) & for main cities in Czech Republic (centers of NUTS3 regions), Poland (agglomeration areas from Urban Atlas) and Spain (agglomeration areas from Urban Atlas)	Urban atlas	19606088
http://w3id.org/foodie/corine# agriculture related lands (hilucs_code<200) & for main cities in Czech Republic (centers of NUTS3 regions), Poland (agglomeration areas from Urban Atlas) and Spain (agglomeration areas from Urban Atlas)	Corine Land Use	16777595
http://w3id.org/foodie/open/cz/Soil_maps_BPEJ_WGSc#	Czech Soil Maps	8746240
http://w3id.org/foodie/open/cz/water_buffer25#	Czech Water Buffers	3978517
http://w3id.org/foodie/core/cz/Predni_prostredni_vyfiltrovano_UTM#	Yield data in field (CZ Pilot)	1111852
http://w3id.org/foodie/core/cz/Pivovarka_vyfiltrovano#	Yield data in field (CZ Pilot)	437404
http://w3id.org/foodie/core/cz/CZpilot_fields#	CZ pilot fields & crop data	20183
http://ec.europa.eu/agriculture/FADN/{FADN category}# (Where FADN category = year-country, year-country-anc3, year-country-lfa, year-country-organic-tf8, year-country-siz6, year-country-siz6-tf14, year-country-siz6-tf8, year-country-sizc, year-country-tf14, year-country-tf8m, year-country-typology, year-region, year-region-siz6, year-region-siz6-tf8, year-region-sizc, year-region-tf14, year-region-tf8)	FADN	23520756
http://w3id.org/foodie/open/africa/GRIP#	African Roads Network	27586675
http://w3id.org/foodie/open/africa/water_body#	African water bodies	11330
http://w3id.org/foodie/open/gadm36/{level}# where {level} = level0, level1, level2, level3, level4, level5	GADM dataset	7188715
http://w3id.org/foodie/open/kenya/ke_crops_size#	Kenya crop Size	85971

Graph URI (note: URIs are not resolvable; they can be used to refer to the specific dataset in the triplestore)	**Name of dataset**	**Number of RDF triples**
http://w3id.org/foodie/open/kenya/soil_maps#	Kenya Soil Maps	10168
http://www.fao.org/aims/aos/fi/taxonomic#	FAO	318359
http://www.fao.org/aims/aos/fi/water_FAO_areas#	FAO	150
http://www.fao.org/aims/aos/fi/water_FAO_areas/inland#	FAO	15779
http://www.fao.org/aims/aos/fi/water_FAO_areas/marine#	FAO	6768
http://w3id.org/foodie/open/catchrecord/norway/	Catch Record Norway	192867166
http://standardgraphs.ices.dk/stocks#	ICES stocks data	1270280
https://www.omg.org/spec/LCC/Countries/ISO3166-1-CountryCodes/	ISO Country Codes	8629
https://www.omg.org/spec/LCC/Countries/Regions/ISO3166-2-SubdivisionCodes-NO/	ISO Country Subdivision Codes	391
https://www.omg.org/spec/LCC/Countries/UN-M49-RegionCodes/	ISO Region Codes	569

3.3.4 Summary

The “Generic pipeline for Linked Data” is an example of a pipeline pattern that fits different needs, which can be used in different scenarios, and applied with different data types and sources. Technically, it has been applied to two specific pilots from the agriculture domain in DataBio project, but it has also been applied to other use cases, including general fishery data related to the fishery domain, EO metadata from different services in DataBio, and open EU/national datasets relevant to the bioeconomy sectors. Therefore, it can be considered a “pipeline design pattern” that can be easily customized to different needs.

Linked Data is increasingly becoming one of the most popular methods for publishing data on the Web due to several reasons, e.g., improved accessibility, integration, and knowledge discovery. Hence, this pipeline has been created with the aim to collect, transform, and publish data related to DataBio sectors, collected in the form of heterogeneous sources (shapefiles, (Geo)JSON, CSV, relational database, REST APIs) as Linked Data. After the publication of the linked datasets, the pipeline includes different methods for their exploitation and reuse over applications like HSLayersNG, Metaphactory, and other Linked Data consumers tools. Ultimately the Linked Data pipeline aim at providing an integrated view

over different source datasets, which can be used by different analytic and decision support services to provide better and more informed advice to decision-makers.

3.4 Generic pipeline for Earth Observation and Geospatial data Processing

3.4.1 Generic/reusable pipeline for Earth Observation and Geospatial data processing

The following figure depicts a generic pipeline for Earth Observation and Geospatial data processing. Starting from the DataBio Data Value chain, it has been adapted to provide a version focused on Earth Observation and Geospatial data management. To achieve it, we have used the experience gained during the work done in DataBio project, and more specifically the pilots that are using Earth Observation and Geospatial data.

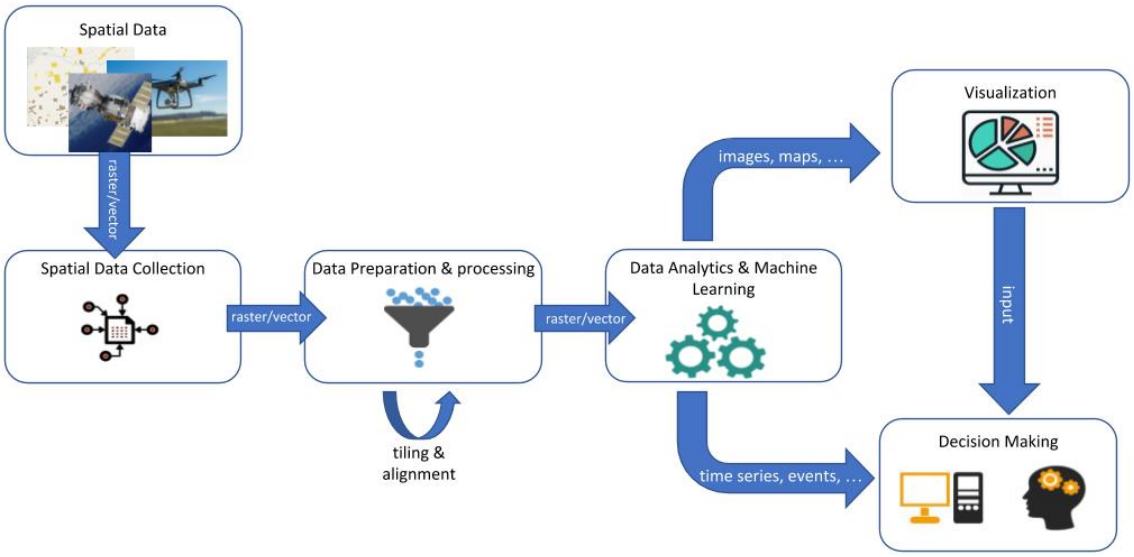


Figure 34: Generic pipeline for Earth Observation and Geospatial data processing

Among the specific characteristics of this pipeline, we would like to highlight the following:

- As initial data input, it has georeferenced data, which might come from a variety of sources such as satellites, drones or even from manual measurements. In general, this will be represented as either in the form of vector or raster data. Vector data usually describes some spatial features in the form of points, lines or polygones. Raster data, on the other hand, is usually generated from imaging-producing sources such as Landsat or Copernicus satellites.
- Information exchanged among the different participants in the pipeline can be either in raster or vector form. Actually, it is possible and even common that the form of the data will change from one step to another. For example, this can result from feature extraction based on image data or pre-rendering of spatial features.
- For visualisation or other types of user interaction options, information can be provided in other forms like: images, maps, spatial features, time series or events.

3.4.2 Instances of this generic pipeline in DataBio

The pipeline depicted in Figure 15 is quite abstract in order to encompass all data types, while the generic pipeline that appears in Figure 34 depicts the common data flow among six project pilots, four of which are from the agricultural domain and two from the fishery domain. Specifically, from the agricultural domain there are two smart farming pilots (A1.1 and B1), one agricultural insurance pilot (C1.1) and a CAP Support pilot (C2.2), while the two pilots from the fishery domain are A1 "Oceanic tuna fisheries immediate operational choice" and B1 "Oceanic tuna fisheries planning".

Therefore, the pipeline in Figure 34 can be considered as a specialization of the top-level pipeline, as it concerns the data processing for Earth Observation and Geospatial data. The mapping of this pipeline to the steps of the pipeline of Figure 15 is depicted in Figure 35.

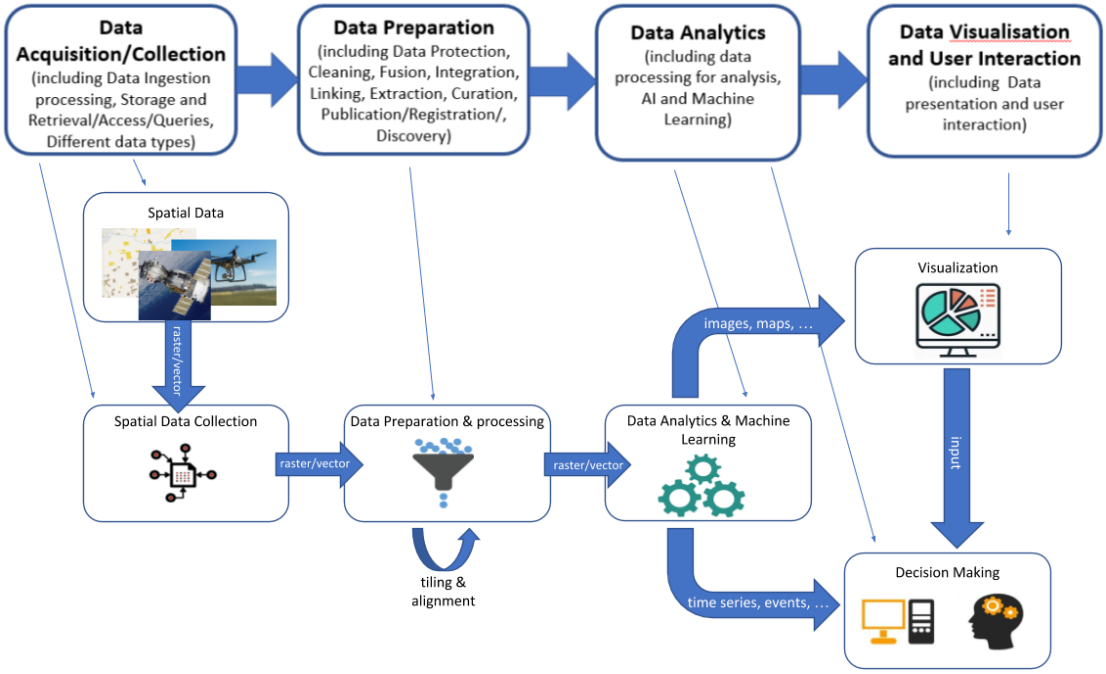


Figure 35: Mapping of the steps of the top-level pipeline (depicted in Fig. 33) to the steps of the generic pipeline Generic pipeline for Earth Observation data processing

3.4.3 Pilot name A1.1, B1.2, C1.1 & C2.2 (Agriculture)

The figure below represents the pipeline shared between the following agricultural pilots: A1.1 "Precision agriculture in olives, fruits, grapes", B1.2 "Cereals, biomass and cotton crops", C1.1 "Insurance (Greece)" & C2.2 "CAP Support (Greece)".

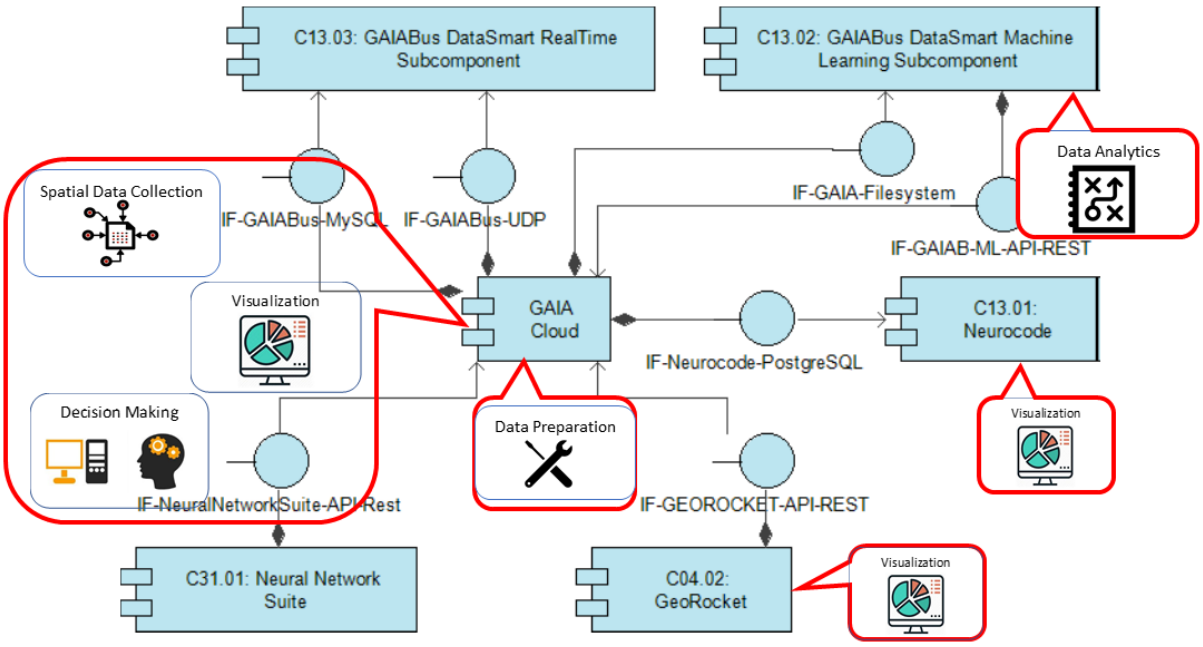


Figure 36: Mapping of the steps of the generic pipeline (depicted in Fig. 33) to the component view shared between the agricultural pilots A1.1, B1.2, C1.1 and C2.2

This pipeline description focuses on the business/pilot context, interfaces and data experimentation from Georocket's point of view, as Georocket is part of the DataBio technology platform. On the other hand, GAIA Cloud is a major horizontal building block of NP's Gaiasense solution and constitutes of multiple cloud computing services that collect, store and combine heterogeneous data to convert them into facts using advanced data analytics techniques.

The pipeline can process a large amount of data in a reasonable time. The components are able to visualise vector data in various clients, effectively exhibiting the interchangeable nature of clients in the pipeline.

With respect to the EO data managed by the pipeline:

- Sentinel-2- generated data: Time series and multiple statistics of EO-based indicators that describe various agri-environmental conditions and are assigned to each agricultural parcel.

The following components are involved in this pipeline:

- NP's private cloud, called GAIA Cloud, along with all its infrastructure and the supporting cloud computing services (e.g. GAIABus DataSmart Machine Learning Subcomponent (C13.025) from NP).
- GeoRocket (C04.026) from Fraunhofer.
- GeoToolbox from Fraunhofer to extend the functionality of GeoRocket - C04.037.

3.4.4 Pilot name A1 & B1 (Fishery)

The figure below represents the pipeline shared between the following fishery pilots: A1 "Oceanic tuna fisheries immediate operational choice" and B1 "Oceanic tuna fisheries planning".

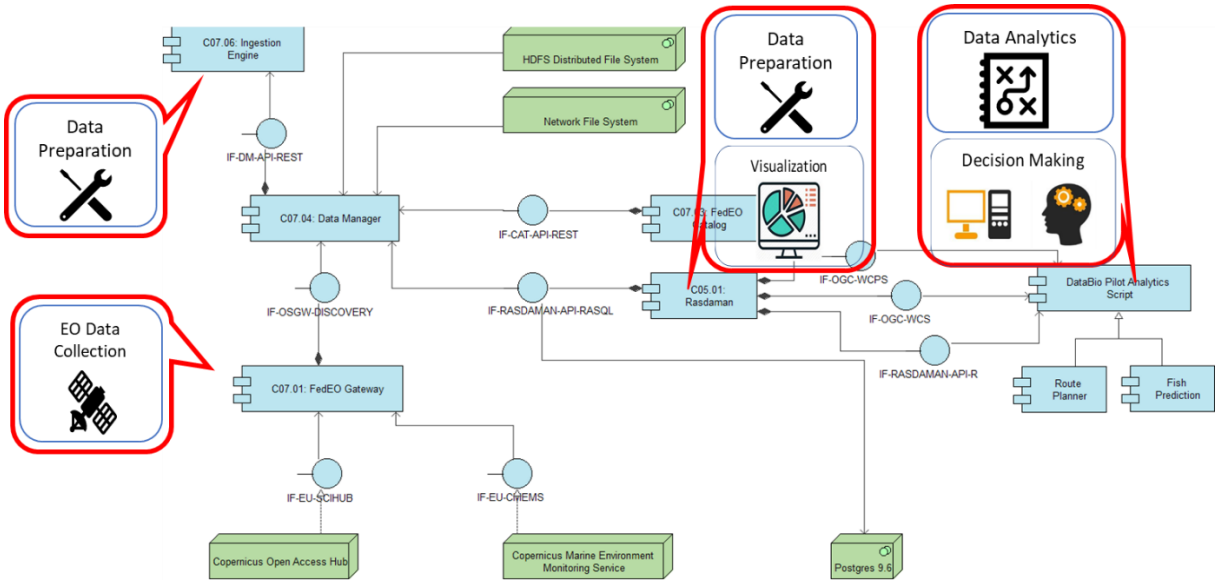


Figure 37: Mapping of the steps of the generic pipeline (depicted in Fig. 33) to the component view shared between the fishery pilots A1 and B.

In the presented pilots, EO data, in particular SENTINEL-3 and Copernicus Marine Environment Monitoring Service data (i.e. sea surface currents, temperature, wind speed, chlorophyll, phytoplankton and other oceanographic parameters) is used in combination with weather conditions information and models as well as in-situ real-time observations from the fleet (i.e., engines, propulsion, route and speed of the vessel, destination) in order to establish a common data management and analysis system.

With respect to the EO data managed by the pipeline:

- Sentinel-3: i) Sentinel-3 SLSTR for Sea Surface Temperature; ii) Sentinel-3 SRAL/MWR for altimetry (Anomalies) and; iii) Sentinel-3 OLCI for Chlorophyll.
- CMEMS Products: Copernicus Marine Environment Monitoring Services.
- Fleet sensor observations: i) Vessel engine sensors (velocity and heading); ii) Position of the vessel and; iii) Fish catches (species, weight).

The following DataBio components are involved in this pipeline:

- Ingestion Engine (C07.06) and Data Manager (C07.04) components use the FedEO Gateway (C07.01) to discover and download new Sentinel-3 and Copernicus Marine products respectively from the Sentinel Scientific Data Hub.

- The Data Manager stores metadata and quicklooks retrieved in a local EO Catalog (C07.03) and makes the downloaded products available in the NFS datastore.
- The Data Manager invokes Rasdaman (C05.01) API and the EO data is published in Rasdaman, providing all relevant data in a common data reference framework (raster-based and same spatial resolution), enabling access from further elements in the pipeline via the OGC WCS and WCPS interfaces as well as the API provided by Rasdaman (enabling thus the integration of the R statistics framework with Rasdaman internal capabilities).

3.4.5 Summary

The "Generic pipeline for Earth Observation and Geospatial data processing" is an example of a pipeline pattern that fits the two aspects of generalization. Technically, it has been applied in four DataBio pilots from the agriculture domain and two DataBio pilots from the fishery domain, that is why it can be considered as a "pipeline design pattern". Below, we include some examples of further usage of this pipeline, mainly in the fishery domain:

- reduce fuel consumption from the interaction between engine data, propulsion data, meteorological data and the vessel design by means of Big Data approaches.
- estimate the expected lifetime of different parts of the engine and propulsion system, or to find out when a part of the engine is close to failure and notify the technical staff in order to have the necessary parts and technicians at the port and reduce the downtime for unexpected failures.
- improve the profitability of oceanic tuna fisheries through savings in fuel costs through fish observation and route optimisation by means of unassisted machine learning. In that regard, all the historical data available i.e., catch logbooks, GPS, Buoys, Observers, etc. will be put in relation to different databases and used to learn the better way to fish with the FAD (Fish Aggregating Devices) method.

Conceptually, this pipeline can also be applied to other domains beyond agriculture and fishery. Basically, it can be applied to any use case from any domain in which Earth Observation and Geospatial data is collected and managed for data storage, analysis and visualisation.

3.5 Generic pipeline for Forestry data management/support

3.5.1 General

The main characteristic of this generic pipeline is the collection of standardized data from multiple data sources to generate a forest data ecosystem where the decision-making can be based on a combination of different data types via the standardized API integrations.

This pipeline has been derived from three pilots in forestry: 2.2.1, 2.2.2 and 2.4.2. In Pilot 2.2.1 "Easy data sharing and networking", the data of the forest data ecosystem is connected with the Wuudis service. In Pilot 2.2.2 "Monitoring and control tools for forest owners", a

new mobile solution called Laatumetsä was produced for enabling the production of the observation data for the forest authorities as a part of the Forest data ecosystem. In Pilot 2.4.2 "Shared multiuser forest data environment", the forest data ecosystem is connected with multiple new data sources as well as standardized open data services with standardized API's were established. Figure 38 illustrates the end to end data flow applied to the three aforementioned pilots of the project.

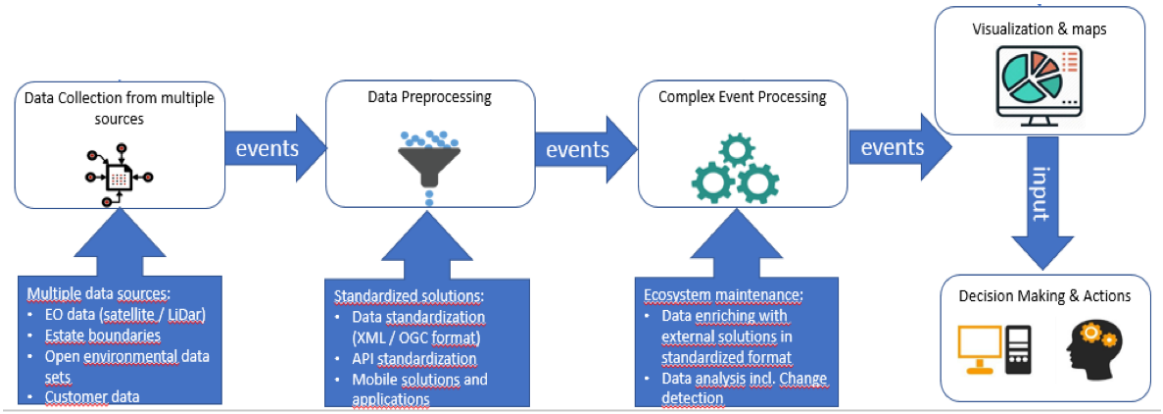


Figure 38: Generic pipeline and Data flow for the Forest data ecosystem data processing and decision-making

In the following figure, we provide a mapping to the steps of the top-level pipeline depicted in Figure 38 to the generic pipeline for the forest data ecosystem data processing and decision making depicted in the previous figure.

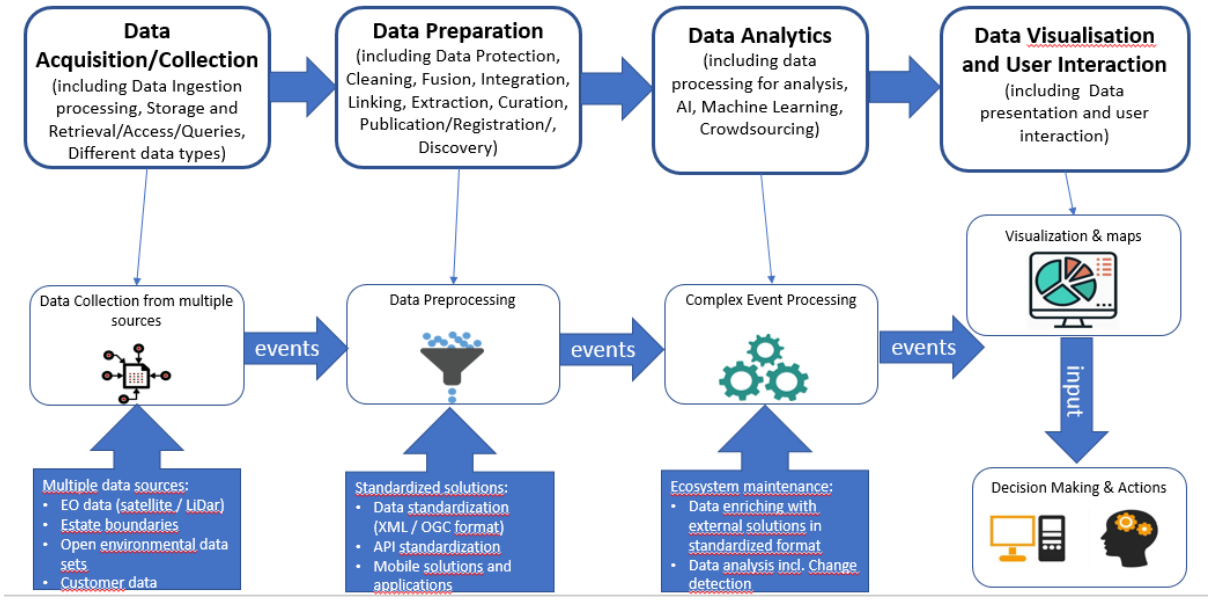


Figure 39: Mapping of the Generic pipeline for the Forest data ecosystem data processing and decision-making to the top-level pipeline depicted in Fig. 37

3.5.2 Instances of this generic pipeline in DataBio

As aforementioned, the generic pipeline is a generalization of three of the project's pilots. To demonstrate this, for each of these three pilot's pipelines, we show the mapping of the different generic components into the pilot's pipeline view diagram.

3.5.2.1 Pilots 2.2.1 Easy data sharing and networking and 2.2.2 Monitoring and control tools for forest owners

The objectives of the Pilot 2.2.1 and 2.2.2 pipeline are as follows:

- Real-time forest management service development based on multiple forest Big Data sources
- Assess and monitor the forest health status
- Handle Big Data in forestry for storing, processing, and transferring of large amounts of forestry data
- Integration between Wuudis and Metsään.fi and extending the Wuudis mobile app to cope with new requirements and integrations.

Wuudis is a commercial service in the market for forest owners, timber buyers, and forestry service companies that enables forestry and forest resource management in one place. The cloud-based platform with mobile interface, and data in Extensible Markup Language (XML) and JSON formats, connects forest owners directly with local contractors and timber buyers. With it, forest owners and other stakeholders can effectively manage their forest resources remotely in real-time. It can be used to obtain real-time information about the forest and its timber resource, track executed silvicultural and harvest activities, plan the needed forest management activities, and bid care works and timber sales online.

URL in DataBio Hub: https://www.databiohub.eu/registry/#service-view/Wuudis/0.0.1

The components involved in the pipeline include:

- Wuudis (MHGS, C20.01)
- Forestry Thematic Exploitation Platform (Forestry TEP, C16.10)
- Metsään.fi eService (C18.01)
- Senop Hyperspectral Camera (C44.01)
- EnsoMOSAIC Fusion software (MosaicMill, C44.02)

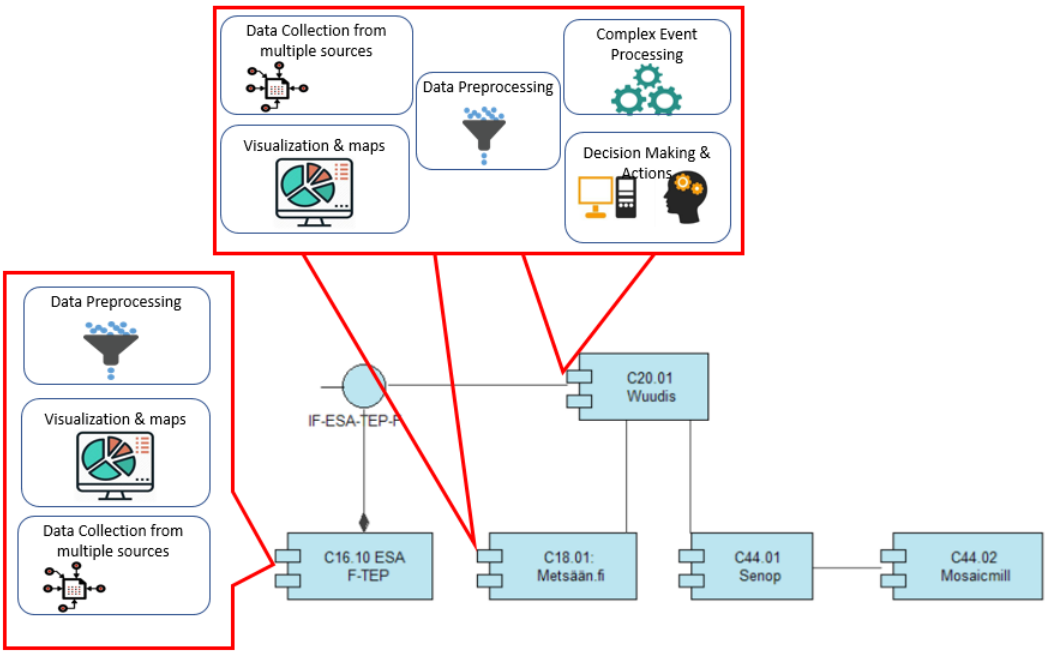


Figure 40: Mapping of generic components into pilot 2.2.1 and 2.2.2 component view

3.5.2.2 Pilot 2.4.2 Shared multiuser forest data environment

The purpose of this pipeline is to provide new maps and functionalities to Metsään.fi and Metsäkeskus concerning crowdsourced data. It is also closely linked with Pilot 2.2.2, which provides prescriptions of quality control for Metsään.fi users. Furthermore, the link with the Pilot 2.2.2 consists of the generalized storm damage data to Metsään.fi via crowdsourcing interface. This pipeline also produces updates on forest data standards to collect quality control data with mobile devices. The pipeline consists of integrations for other data sources for providing user and data security via single-login and easy user role-based authentication and data access permissions.

URL in DataBio Hub: https://www.databiohub.eu/registry/#service-view/WP2%20-%20Shared%20multiuser%20forest%20data%20environment/0.0.1

The components involved in the pipeline include:

- Metsään.fi eService - Metsään.fi is a portal through which people who own forest property in Finland can conduct business related to their forests from the comfort of their own homes. The portal connects owners with related third parties, including providers of forestry services. (C18.01)
- Open forest data service - Open Forest Data is a service providing open data according to Finnish legislation. The data consist of Forest Compartments data, Grid data, Forest Use Declaration Data, Subsidy Applications data and Data from Valuable Habitats of Forest Act. (C18.02)

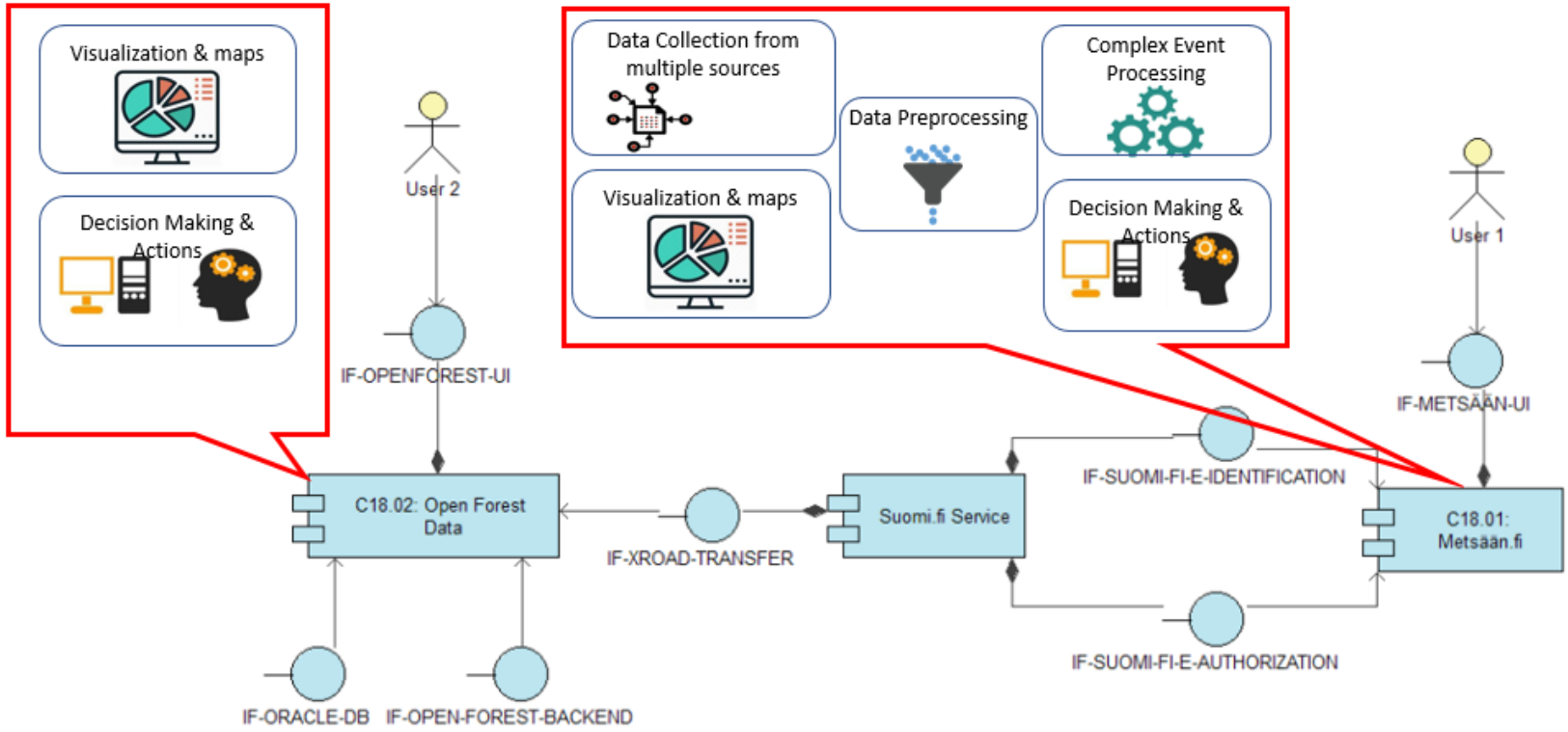


Figure 41: Mapping of generic components into pilot 2.2.4 component view

3.5.3 Summary

The "Generic pipeline for Forest data ecosystem" is an example of a pipeline pattern that fits the two aspects of generalisation. It has been applied in three pilots in the project from the forestry domain and as such, it can be seen as a "pipeline design pattern". Conceptually, the same approach could also be applied to other domains beyond forestry. Basically, all the use cases from any domain in which data is collected in a standardized format from the different data sources and furthermore enriched with additional operations via the standardized API's can nicely fit into this generic pipeline.

For example: the observation data regarding the environmental and physical phenomena collected by the citizens or industry professionals by mobile applications can be further adopted as a part of the forest data ecosystem. This data can be used to analyse the possible magnitude of the forest damages further as well as for preventing the spread of the damage and additional implications such as industry profit losses or ecosystem damages. Another use case is related to the field data such as sample plot data collected by the mobile solution end-users, which can be further combined with other data sources and utilized as a reference data for the satellite data analysis and related data products.

3.6 Genomics

3.6.1 General

The Genomic models (C22.03) can be used across agriculture and husbandry pilots having breeding as core business. This technology was successfully implemented in sorghum [REF-34], wheat [REF-35], potato [REF-36], and several other plant species [REF-34]-[REF-37] of agricultural interest. The main features of this generic pipeline are the collection of whole-genome nucleic acid (DNA) information and phenotypic information in the training population and produce equations that will be used to predict the performance of unphenotyped individuals be they plants or animals. The C22.03 routines are collectively implemented as depicted in Figure 42.

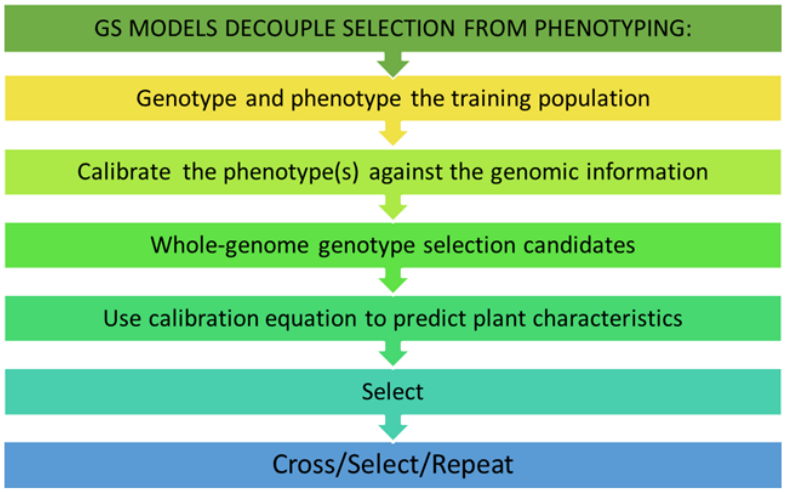


Figure 42: Collective Implementation of the routines of the Genomic Models (C22.03)

Alternative models can be implemented to account for the allelic frequencies that governed the plant characteristics of interest. For instance, some models accommodate the infinitesimal model assumptions based on the action of very many genes, each with a minimal effect, others are best suited to finite loci model with stronger shrinkage of regression coefficients that are close to zero and less shrinkage of those with large absolute values, while others can extend Fisher's infinitesimal model of genetic variation to accommodate non-additive genetic effects. The ultimate outcome is the computation/prediction of the breeding value upon which the superior genotypes are selected. In virtue of the breeder's equation ($\Delta R = ih\sigma_g/t$) and considering that genomic modelling allows accurate prediction of the Mendelian sampling term even at the seed stage before planting, the genomic predictive analytics (C22.03) can be considered as the gold standard technology to expedite crop breeding and allow higher genetic gains per unit time and cost.

The diagram in Figure 43 represents the generic pipeline for data flow genomic selection and prediction: from data collection to data processing and decision-making, and its mapping to the steps of the top-level pipeline of Figure 14.

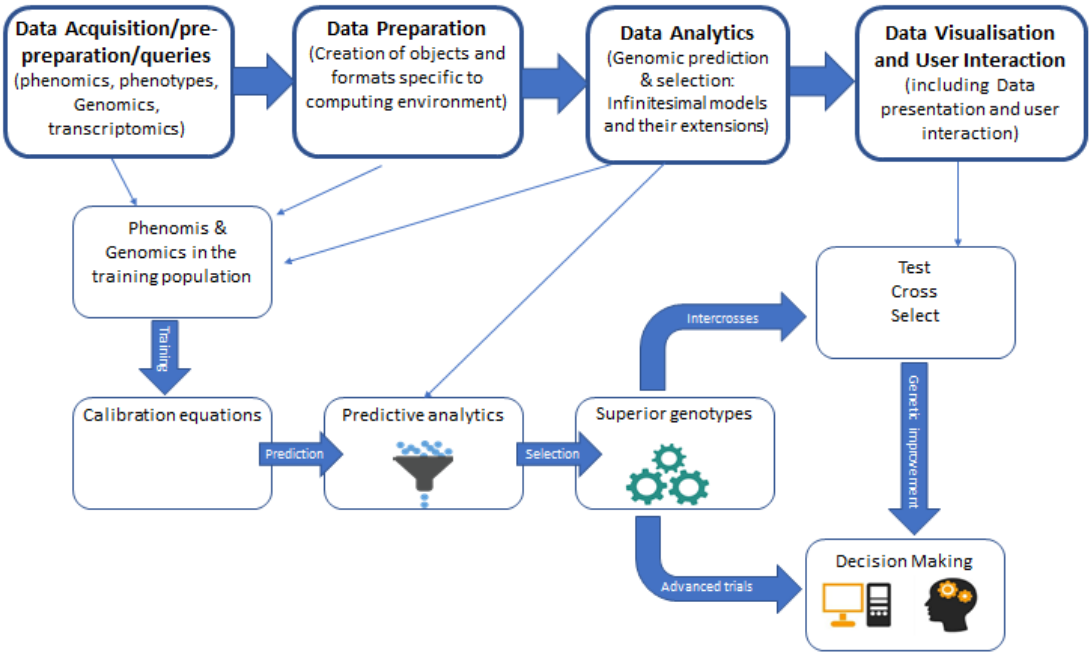


Figure 43: Generic pipeline for data flow genomic selection and prediction and its mapping to the steps of the top-level pipeline.

3.6.2 Instances of this generic pipeline in DataBio

3.6.2.1 A2.1 pilot: Genomic prediction and selection in glasshouse ecosystems

The A2.1 pipeline was designed to model several scenarios including single/multi-traits, single/multi-environments with the aim of simplifying the breeding scheme, cutting costs and time required to develop a superior cultivar. This pipeline is associated with pilot A2.1: Big Data management in greenhouse ecosystems. It was described under the deliverables D1.1

and D1.2 at https://www.databio.eu/publicdeliverables/. The A2.1 pipeline was registered under the DataBio Hub at https://www.databiohub.eu/registry/#services?tag=C22.03. During the project implementation, we observed a slower production of phenotypic data in tomato glasshouses and this prompted us to implement the C22.03 Genomic Models component using similar data produced in biomass sorghum pilot B1.3. The biomass sorghum pilot was described under the deliverables D1.1 and D1.2 at https://www.databio.eu/publicdeliverables/). Phenomics and phenotypic data were used. Phenotypic data included those antioxidants used as raw materials for manufacturing specialty foods to help fight degenerative afflictions [REF-34]. Genomic data included 1,252,091 Single Nucleotide Polymorphism loci [REF-34], [REF-37]. The implementation of the A2.1 pipeline produced encouraging results that were documented in articles published in peer-reviewed international specialized scientific Journals [REF-34], [REF-37].

Figure 44: Phenomics and phenotyping facility in biomass sorghums at CREA, in Italy

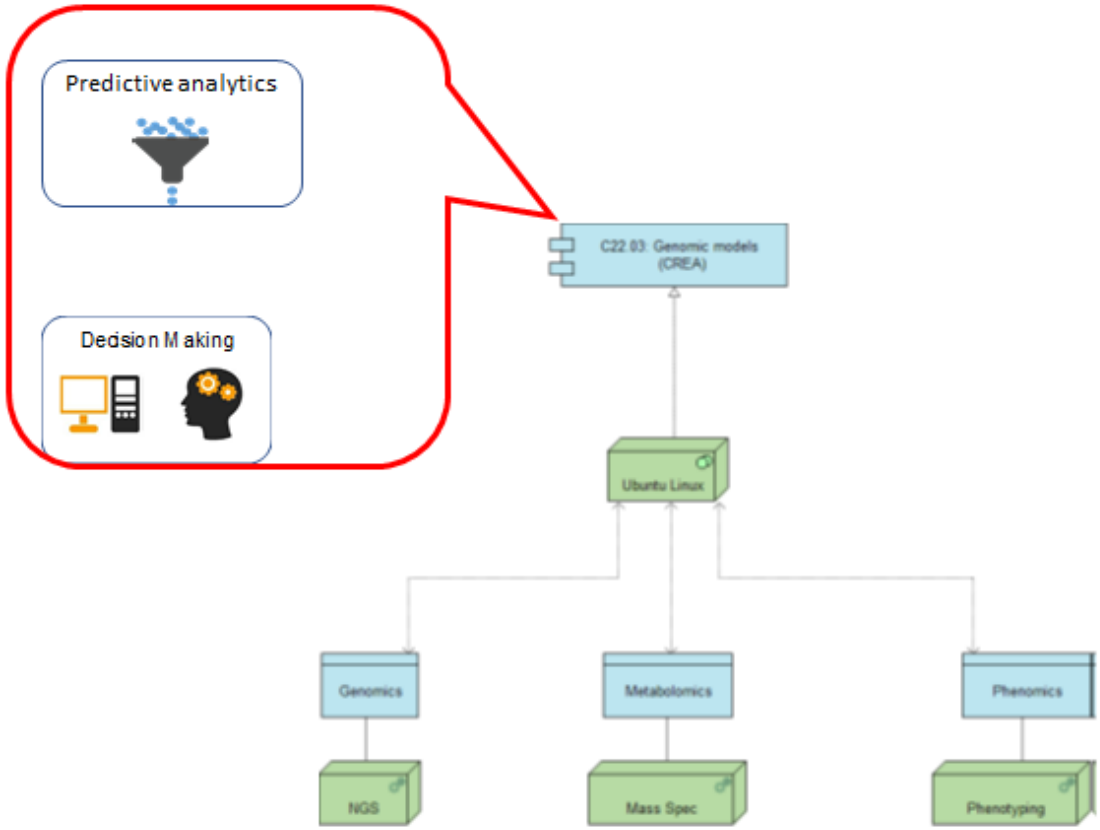


Figure 45: Mapping of generic components into pilot A2.1 component view

3.6.3 Summary

The “Generic pipeline for genomic selection and prediction is an example of a pipeline that can perfectly generalized. Within the framework of DataBio, this pipeline was successfully implemented in biomass sorghum (refer to the agriculture pilot A2.1 in the below sections), and in several other crop species and in animal husbandry. The most basic requirement for the implementation of the C22.03 is any use case from any domain in which models can be calibrated in a training population and the performance of an external genotyped, but unphenotyped sample is predicted. Practical use cases are presented in Chapter 4.

3.7 Generic pipeline for privacy-aware analytics

3.7.1 General

The generic privacy-aware analytics pipeline in Figure 46 is derived from two implementations of secure machine learning in decision support for fisheries, using two different privacy-enhancing technologies. One using secure multi-party computation software, Sharemind MPC, (C35.02) and the other using trusted execution environments based on Intel SGX® technology, Sharemind HI (C35.03). The need for privacy-aware analytics and implementation is described in depth in section 2.7.

One common feature of privacy-aware analytics pipelines is that data owners do not trust their business-critical data to other organizations and want to protect their data during the whole decision-making process. Knowing where to go fishing is a competitive advantage that a fishery is not interested to share with competitors.

The other important aspect of the pipeline is that in a complex decision-making process number of public data sources may be needed for data enrichment. Secure computing technologies are only emerging to the market. The toolsets available to analysts are not yet as comprehensive as those that are developed to handle data openly. Besides the need to implement the required analytical method, secure computing technologies may involve organizational complexities of deployment and computational overheads. Therefore, preprocessing public data sources and private data before data enters the privacy-aware analytics pipeline is highly recommended. Thus, the strengths of common data analysis tools and privacy-aware tools are utilized and complexity is kept optimal.

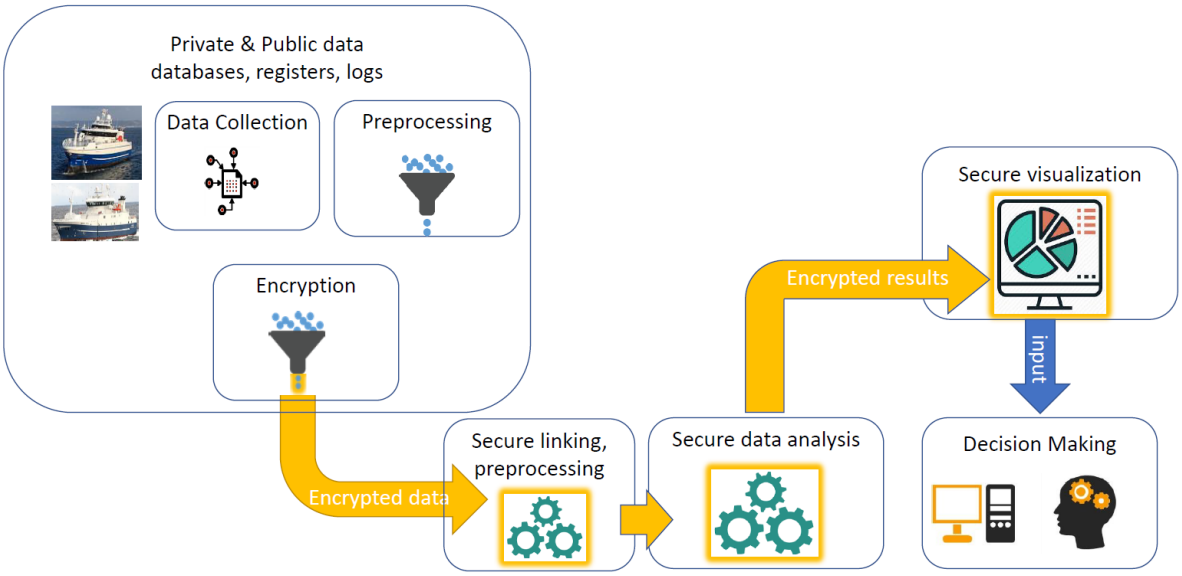


Figure 46: Generic pipeline for privacy-aware analytics

Figure 47 maps the elements of the top-level generic pipeline to the privacy-aware analytics general pipeline. We see that the main difference is maintaining technical protection of confidential or private data during processing when confidential data needs to exit the data owner's organization.

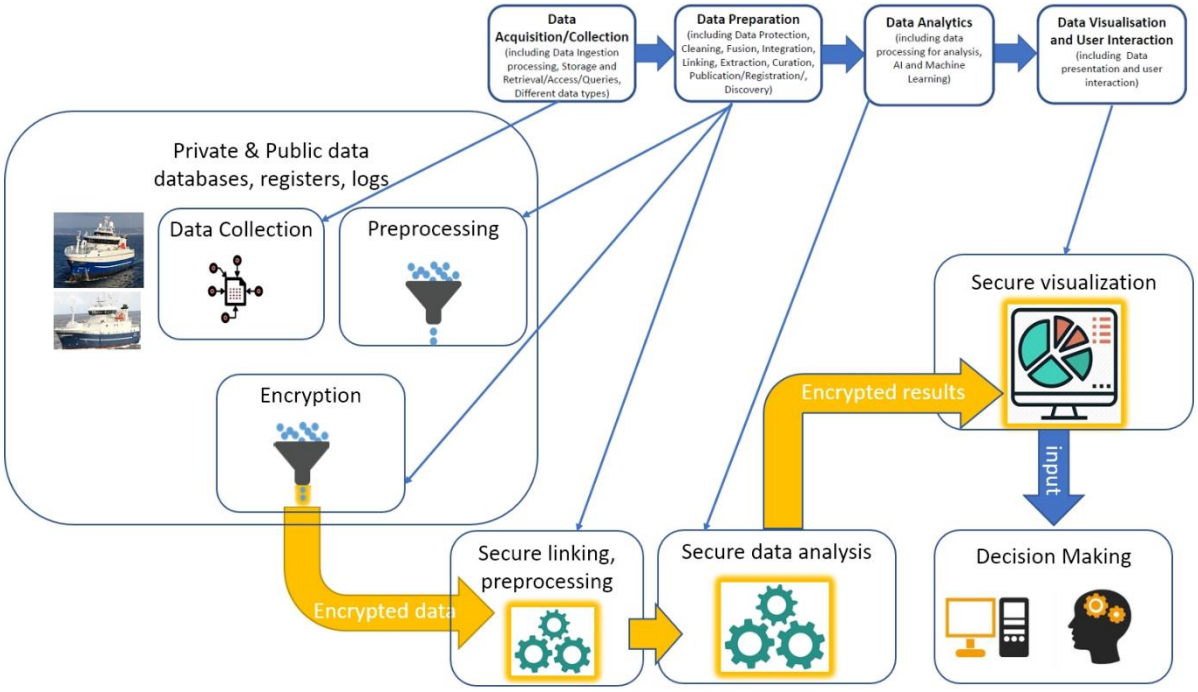


Figure 47: Mapping of the steps of the top-level pipeline to the privacy-aware analytics generic pipeline

3.7.2 Instances of this generic pipeline in DataBio

The two specific pipelines for privacy-aware analytics were implemented in the context of the virtual WP4 demo pilot. See Section 3.7 describing the need to add privacy-aware analytics to the pipelines developed for fishery data and the resulting pipelines.

In this section, we focus on the privacy-aware part and the specifics of the two implementations. Figure 48 describes DataBio's first implementation of a privacy-aware machine learning pipeline to support the integration of private and public data sources. The target was to demonstrate the feasibility of the approach and showcase the results on the EuroGEOSS/Lisbon INSPIRE Fisheries hackathon in June 2019, in a joint effort of SINTEF and Cybernetica. As SINTEF had access to a powerful data analytics and visualization component, SINTIUM (C06.02) and private catch data from one fishery, just the secure machine learning part of the pipeline, was implemented with privacy-aware analytics, component (C35.02) Sharemind MPC. This optimized the joint implementation effort and maximised the speed of delivery of the specific privacy-aware analytics functionality developed for DataBio. Note that in this pipeline, SINTEF acts as a trusted third party for the fishery contributing confidential catch data.

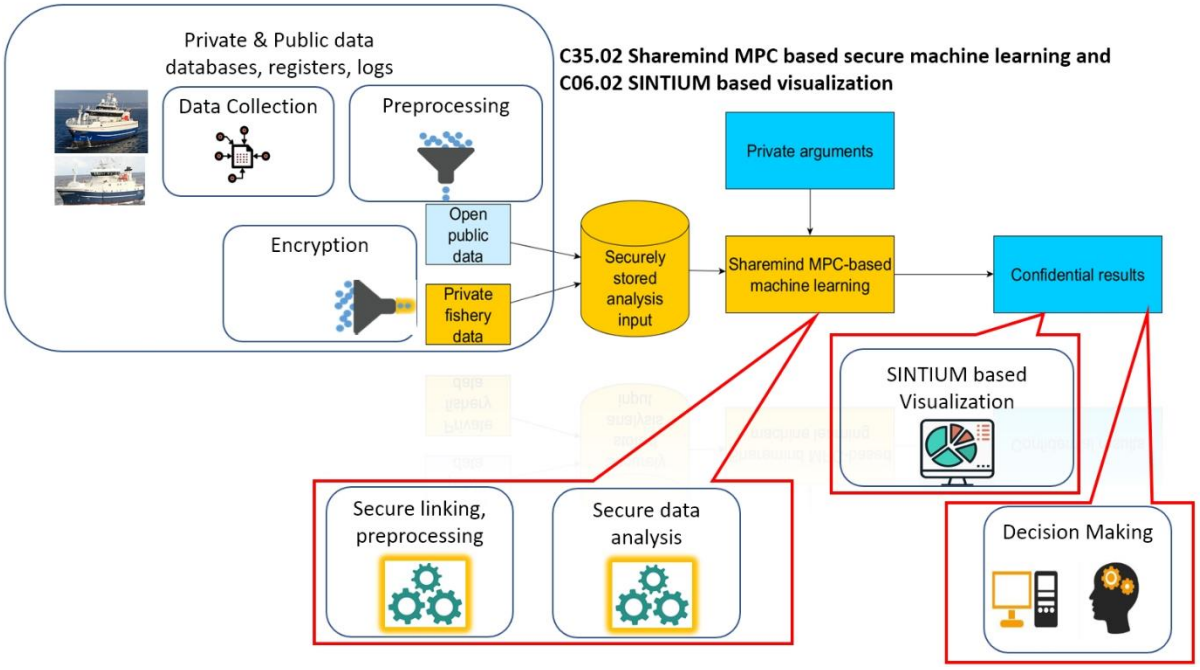


Figure 48: Mapping of the steps of the Privacy-Aware Analytics generic pipeline to the implementation with C35.02 Sharemind MPC and SINTIUM C06.02

Based on the success of the demonstration on the first hackathon, the full privacy-aware analytics pipeline, not needing a trusted third party, was implemented and deployed in IBM public cloud. All the components from data encryption to machine learning and data visualization on the map were implemented using (C35.03) Sharemind HI. Cloud deployment was selected to ease solution use by end customers, while demonstrating feasibility of secure computing in the public cloud in a trusted hardware environment. The solution and pipeline

are described in detail in Section 2.7.2. The target of the demonstrator is to show to fisheries, how they can securely use confidential data to reduce their trips in search for catch locations. This could result in significant economies, increased profitability, reduced waste of fuel and CO_2 emissions. Figure 49 describes the mapping between the specific pipeline implementation and the generic privacy-aware analytics pipeline.

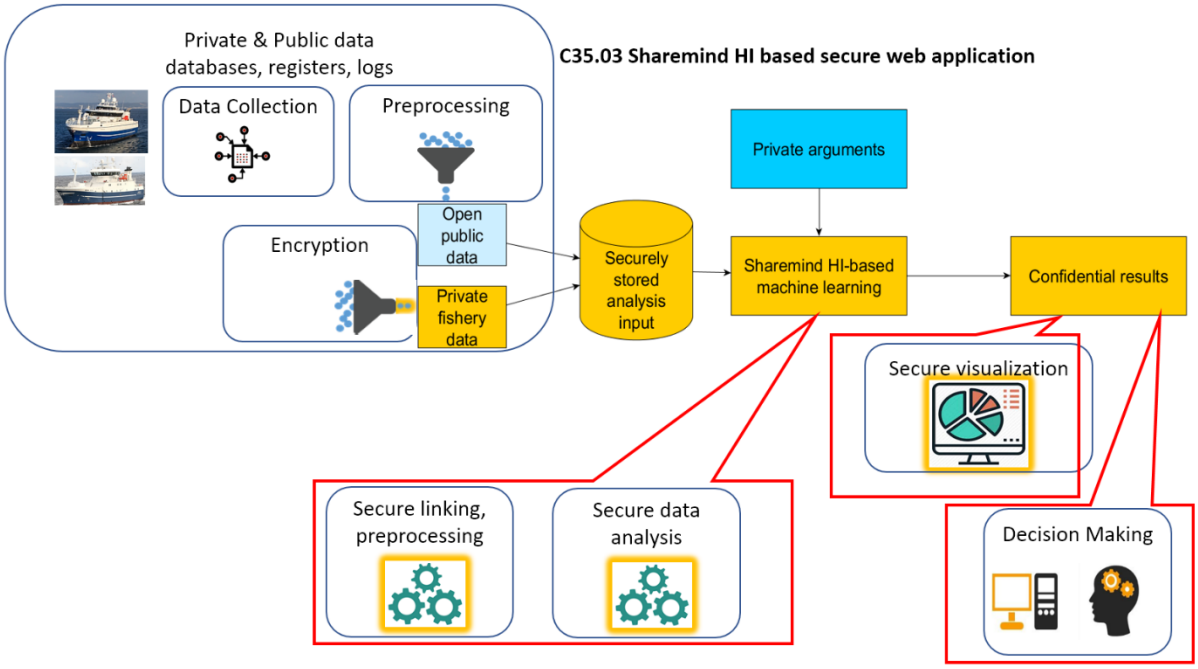


Figure 49: Mapping of the steps of the privacy-aware analytics generic pipeline to the implementation with C35.03 Sharemind HI

The particular pipeline is not limited to the protection of business-critical fishery data. The underlying component, (C35.03) Sharemind HI, has been used for protecting the privacy of people and confidential data of telecom service providers in producing tourism statistics from mobile Big Data in Indonesia[50] and is in the process of development to protect data in healthcare, energy and other domains.

3.7.3 Summary

The "Generic pipeline for Privacy-Aware Analytics" fits two aspects of generalization. Technically it has been implemented using two different methods for protecting confidential data during the whole processing lifecycle. Both use cases: with a natural trusted third party and traditional data analysis components and pure end to end data protection type cases could be covered. The approach is generic and fitting other use cases for private or confidential data processing beyond applications in fisheries.

[50] https://sciencebusiness.net/pursuit-precision-and-privacy

3.8 Generic pipeline for fisheries decision support in catch planning

3.8.1 General

A deciding factor for the efficiency of fishing vessels is the ability to find the best suitable fishing grounds, combined with the best suitable fishing methods and tools. Such decisions are today based on the skills and knowledge of the captain and crew in combining public data (weather, earth observation and public catch data) with private data (catch history, locations and price) to optimize the fisheries. There is a strong need for decision support applications in the planning and execution of open sea fisheries, and the application of Big Data technology in catch and process yield optimization have largely untapped potential in this industry. This challenge goes beyond the need of the agriculture and forestry applications, as it is not normally known where the fish is located when the vessel leaves the port.

The distances travelled involved in finding and harvesting the biomass is of quite another scale than land-based production. The search area is huge, for example, in whitefish trawling a typically three-week fishing trip can circumvent the entire Barents' sea before returning to port in Troms or Vesterålen, a single trip of many thousands of kilometres. With respect to this aspect, fisheries can be said to have more in common with hunting than harvesting. Big Data technology, and in particular large-scale prediction analytics that can help locate the fishing grounds is therefore of key interest to the fishery sector.

A generic pipeline for fisheries data and analytics visualization purposes was devised based on the common reusable (sub) pipeline of the four small pelagic fishery pilots in DataBio. This generalized pipeline represents a valuable and exploitable asset applicable to other use-cases: Here it is demonstrated by its use in an alternative fishery, e.g. whitefish fishery through continuous application development iterated over a series of hackathon events (Open Sea Lab hackathons in 2017 and 2019 and EuroGEOSS/Lisbon Inspire hackathons in 2019.

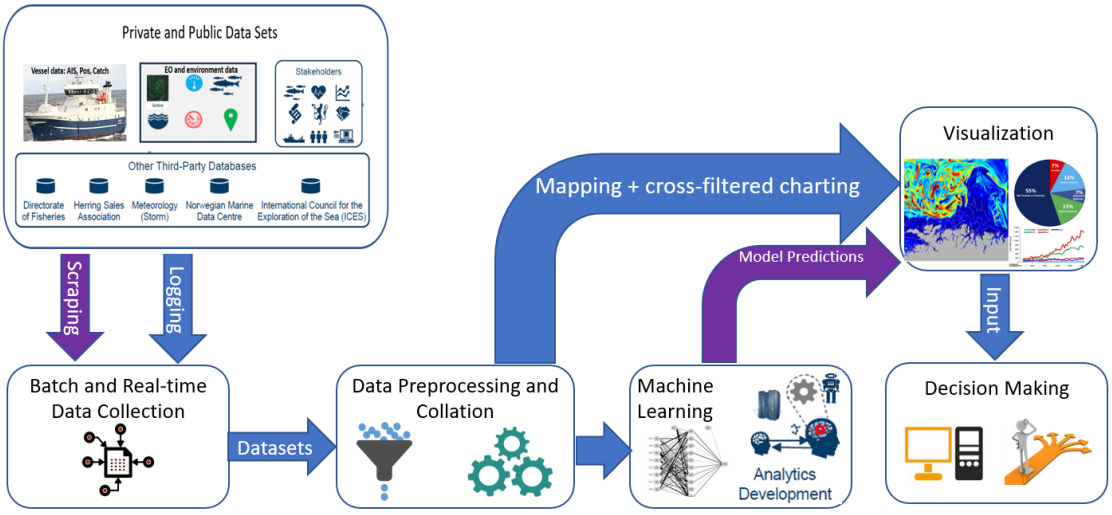


Figure 50: General pipeline for processing heterogeneous datasets for fish catch prediction

Figure 50 shows the end-to-end flow from collecting public and private datasets, with very varying data rates, e.g. from yearly and quarterly statistical datasets to daily, hourly and real-time vessel data like private catch data and sensor data such as vessel location, motion and orientation. These datasets are preprocessed, e.g., filtered, converted and collated, to standardized formats that are mapped, cross-filtered and charted in a graphical user interface and also used to train machine learning models for catch prediction. The output recommendations of the prediction models are displayed in the map GUI together with the collated datasets. Figure 51 shows the same model with its relation to the top-level pipeline abstraction, as can be seen from the figure, there is a logical flow aligned with the generic top-level pipeline.

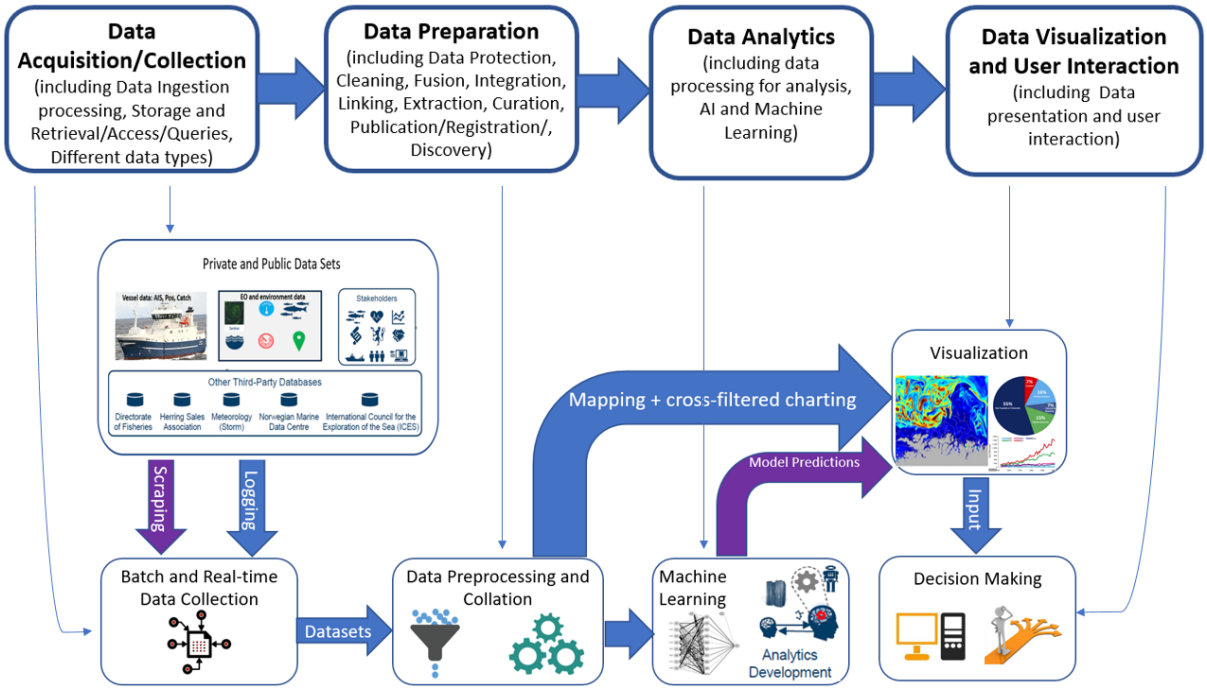


Figure 51: The fisheries pipelines' relation to the top-level generic pipeline abstraction

3.8.2 Instances of this generic pipeline in DataBio

As aforementioned, the third pipeline is a generalization of the initial pipelines designed for the small pelagic fishery pilots in the project, and in particular the B2, C1 and C2 fishery pilots which share a common design (the links are to the descriptions in the DataBioHub):

- Task 3.3.2 Pilot B2: Small pelagic fisheries planning
- Task 3.4.1 Pilot C1: Pelagic fish stock assessment
- Task 3.4.2 Pilot C2: Small pelagic market predictions and traceability

Figure 52 gives the overview of the fisheries and thematic focus of these pilots, and the diagram in Figure 53 illustrates the shared design from data sources to decision support, where each pilot is distinguished by the different focus of the prediction components (as indicated in the red box in the middle right of the figure) . Note that the Pilot A2: Small pelagic

fisheries immediate operational choices also support the design by providing operational data from the vessel into the system.

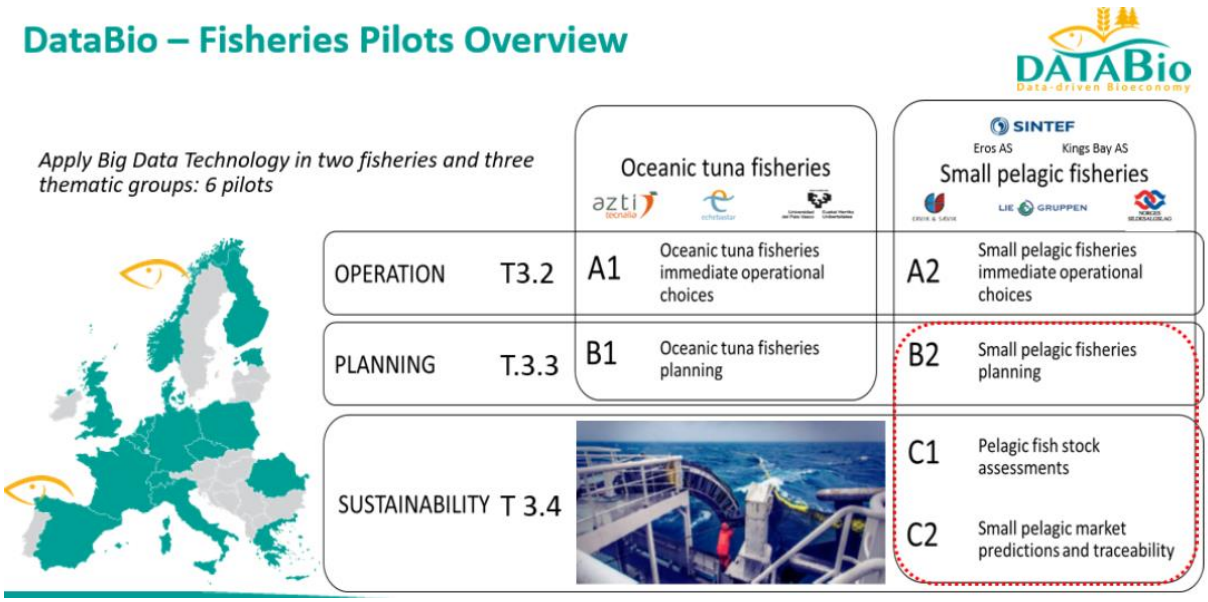


Figure 52: Fisheries pilots overview, indicating the pilots sharing the common data pipeline

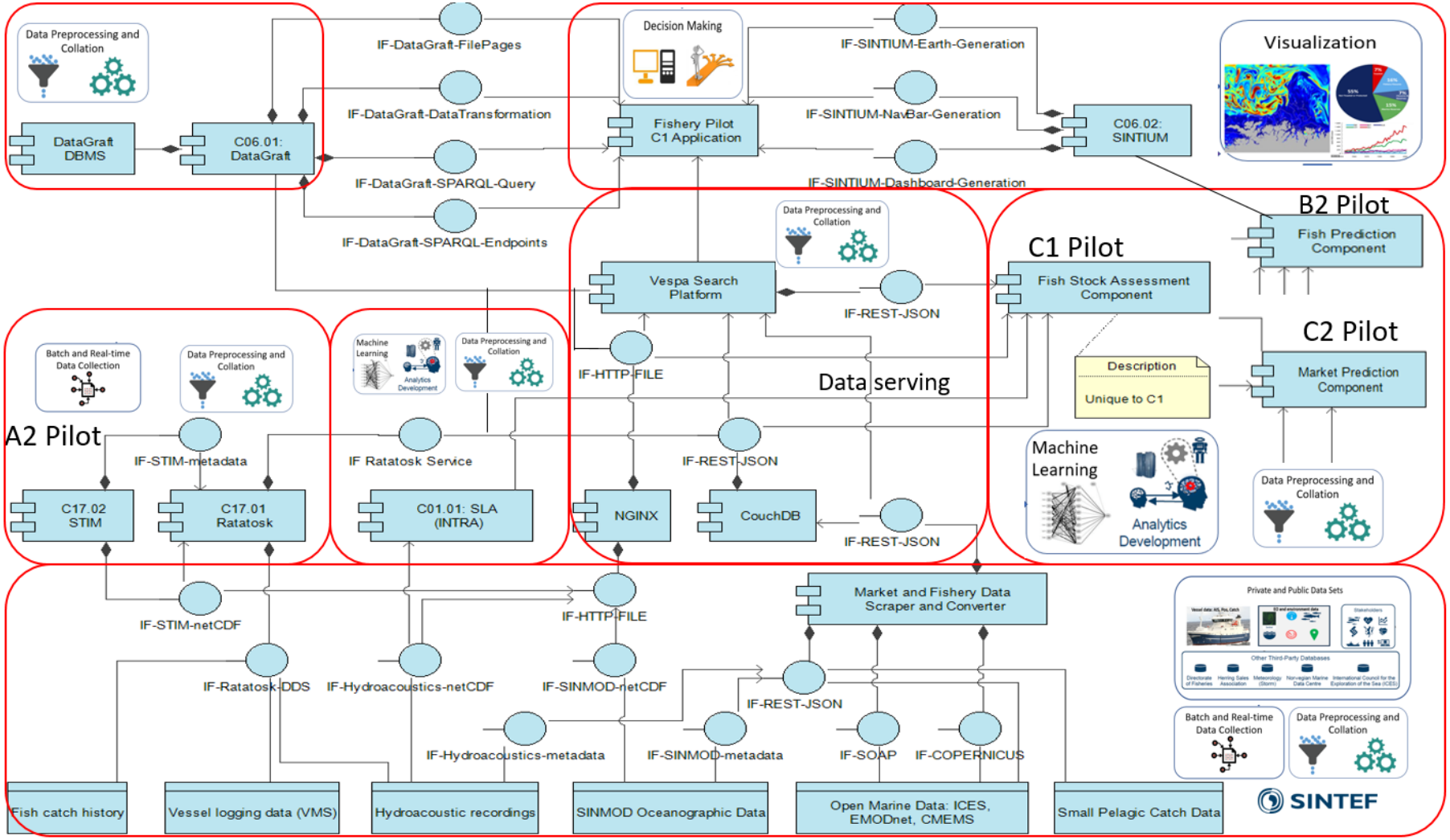


Figure 53: Initial pipeline design for A2, B2, C1, C2 pilots with top level components indicated

The system outlined in Figure 53 is quite complex and represents the proposed design based on identifying the potential components for realizing the small pelagic pilots. Not all of the components were selected or realized by the WP3 pilots, e.g. use of the Vespa, NGINX and DataGraft (C6.01) and SINTIUM (C6.02) components. Therefore, an alternative or virtual pilot was addressed in WP4 together as a joint effort between CYBERNETICA, PSNC and SINTEF with a series of hackathon events. The EuroGEOSS/Lisbon Inspire hackathon 2019, as well as OpenSeaLab 2017 and 2019 hackathons was used as milestone events to accelerate the development and demonstrate the usage of privacy-aware analytics, linked data integration and visualization of historic and real-time fisheries data as well as catch prediction. *As the first two application areas' pipelines are covered by their own sections, e.g. confidential data handling in Section 2.7 and linked data in the fishery in Section 3.3.2.5, we focus on the visualization application of fisheries data for decision support and catch prediction here*. Also, since this virtual pilot demo is not described elsewhere, an overview of the resulting decision support application is included.

3.8.3 Virtual WP4 pilot: Application of the pipeline to whitefish fishery

3.8.3.1 Introduction

The goal of this pilot is to provide the crew and ship owners with information that benefits near-future decisions in fisheries planning. The pilot work has three logical steps:

- Identify, collect, collate a visualize existing fisheries datasets (open and private).
- Train a predictive catch potential model on the combined data.
- Visualize model results together with relevant fishery data.

The pilot combines existing datasets within fisheries activity and catches statistics with meteorological, oceanographic and environmental data. Machine learning has been used to develop a predictive catch model based on combining multiple private and sensitive datasets with open data. This model is used to predict the most likely fishing grounds which is important information providing insight into how the fisheries best can be performed. Hence, very valuable input in the planning and decision process. The benefits of this demo pilot include:

- Saved time and energy while steaming for fishing grounds
- Improved prediction models when joining private data in a secure way
- Showcase DataBio components and services for users external to the project

An alternative fishery, both in terms of species and gear type, was chosen to demonstrate the reusable parts of the fisheries pipeline. The fishery chosen is whitefish fisheries using bottom trawl, mainly due to the availability of detailed private catch data that could be combined with open fisheries datasets. Refer to Figure 54 for the relation to the fisheries planning pilots in DataBio.

During the DataBio project, e.g. in December 2018, the Norwegian Directorate of Fisheries decided to publish and provide access to Norwegian catch data over the last 20 years. Hence,

an interesting new data source became available, with very relevant and interesting fisheries data. The release of this dataset, therefore, became another key driver for demonstrating the combination of private and open catch data in this pilot. This dataset was preprocessed and shared with CYBERNETICA and PSNC for application in the privacy-aware and linked data integration and publication pipelines and future integration with the fishery decision support application.

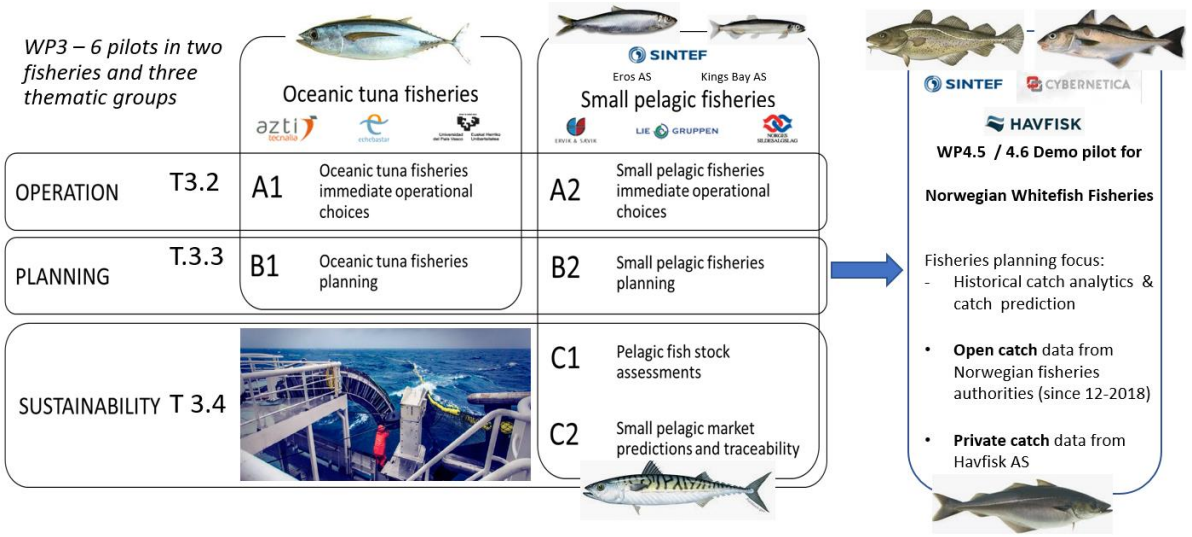


Figure 54: Fisheries pilots overview, showing the relation to the "virtual WP4 demo pilot"

An important observation related to the datasets is that private catch datasets have precise data with respect to location (GPS) and catch time (time of haul) vs. the open catch datasets which are reported when the fish is landed and share a coarse fishing region identificator (covering several square kilometres). Hence, for precise catch prediction of fishing grounds, private data has a substantial higher value, while open data provide interesting context information, e.g. historical catch and volume information.

3.8.3.2 Pipeline demonstrated and components developed further by the pilot

Figure 55 shows an overview of the components of the WP4 demo pilot annotated with the elements of the generic pipeline. The left and middle part of the figure show this demo application is based on adaptation and further development of an existing application from a previous project (ESUSHI), while the right part shows additional extensions from DataBio components included for this pilot. The relationship between this diagram and the pipelines in Figure 53 is that this application is another instance of a fishery decision support application that re-use and expand the functionality of the essential components in the generalized pipeline, e.g. heterogeneous data fetching, collation and data serving for catch analysis and fishing ground predictions.

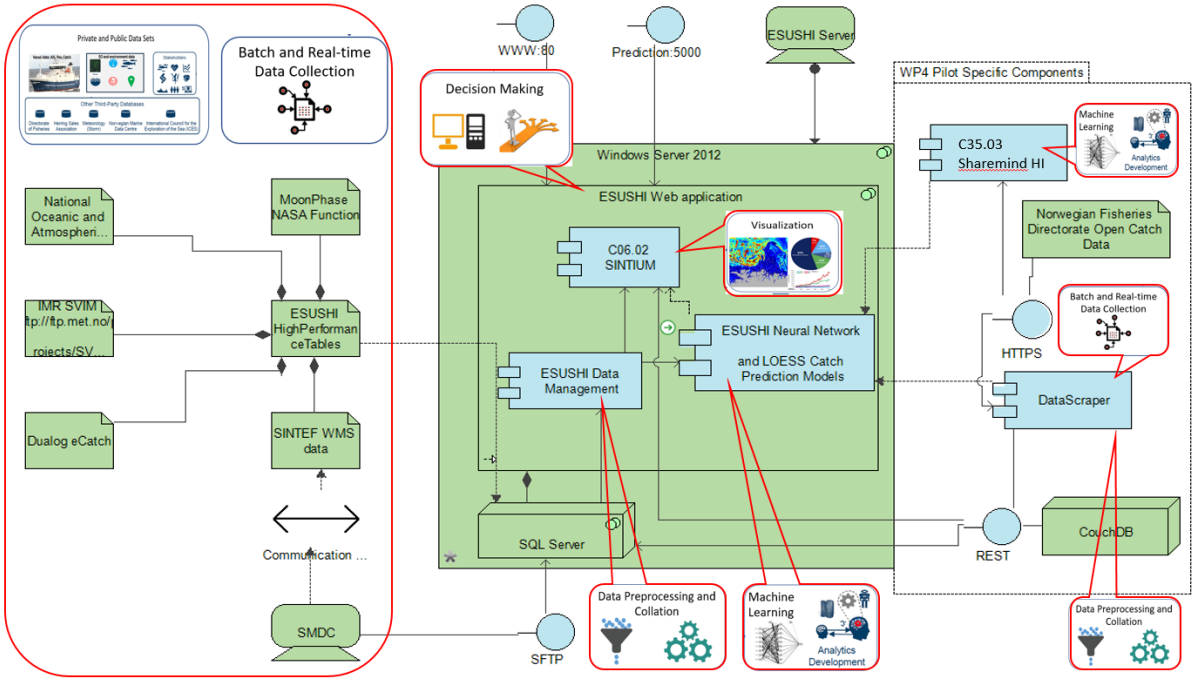


Figure 55: Component diagram showing the "virtual WP4 demo pilot"

To summarize, the work with this pilot has stimulated additional component development beyond the re-use of pipelines:

- The "DataScraper" component is a generalization and further development of the "Market and Fishery Data Scraper and Converter" component in Figure 53, a new component developed by DataBio.
- The SINTIUM component (C06.02) has been developed further by DataBio in multiple iterations to support more datatypes and visualization methods, including
    - Heat-map based visualization of most likely catch areas based on prediction models trained private eCatch data and environment parameters (physical, chemical and biological factors). This was included for the OpenSeaLab 2017 hackathon (ref Team CLP).
    - Integration of the new LOESS catch prediction model developed by CYBERNETICA based on the open catch data from the Norwegian Directorate of Fisheries and applied on the private eCatch data and the display of the prediction output. This was included for the EuroGEOSS/Lisbon INSPIRE Fisheries hackathon in June 2019 (ref. Team Fish slides 7-9 in results section), and developed further afterwards. Please refer to the Privacy-Aware Analytics pipeline section for further details.
    - Visualization of AIS (Automatic Identification System) from BarentsWatch, sea surface temperature, global wind and current forecasts from Copernicus Marine, data integration of bottom sea substrate and vulnerable ecosystems from EMODnet and catch analysis and cross-filtering of the open data from

Norwegian Directorate of Fisheries, and furthermore datascraper API development for ICES datasets (trawl survey and species movements) and adaptation of GlobalFishingWatch machine-learned model for detecting fishing activity based on AIS data to the real-time data-stream from BarentsWatch. This was included in the final hackathon, the 2019 OpenSealab (ref Team CODeFish).

The catch analysis and fishing ground prediction, as well as the added datasets, was integrated into one pilot application that was demonstrated quite successfully in the hackathons.

3.8.3.3 Demo application for fisheries decision support in catch planning

The next section shows a selection of the visualization modes displayed in the demo pilot application realized through this pipeline. Figure 56 shows the global ocean currents from Copernicus Marine (animated), while the diagrams to the right shows catch analytics drill-down of the open catch data, supporting filtering on time, species size and volume.

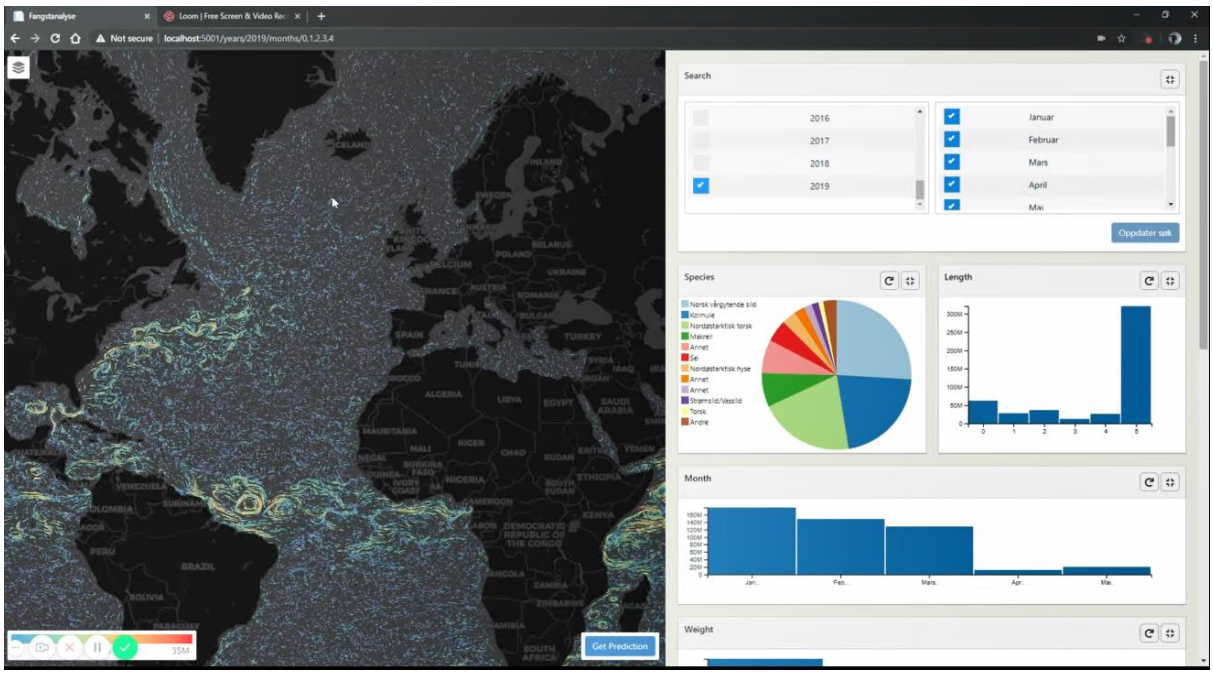


Figure 56: Fisheries decision support web application based on SINTIUM (C06.2)

Figure 57 summarizes other information layers and elements of the application:

- The top left figure shows the real-time location of vessels based on AIS-messages, and the text field right of the figure displays detailed information about one ship that has been selected on the map.
- The bottom left displays global sea-surface temperature from CMEMS.
- In the middle top figure (light colours) is a combined plot of the best fishing locations for North-East Arctic COD as predicted by the Sharemind LOESS model (trained on

2016 top catches, daily predictions) for coarse open data (cyan circles) and private data (darker blue dots). The red dots mark the start and end of the daily predictions for a year. The objective of the figure is to show how much more detailed locations the private data gives using the same prediction model, as the open catch data has quite coarse resolution in location and time.

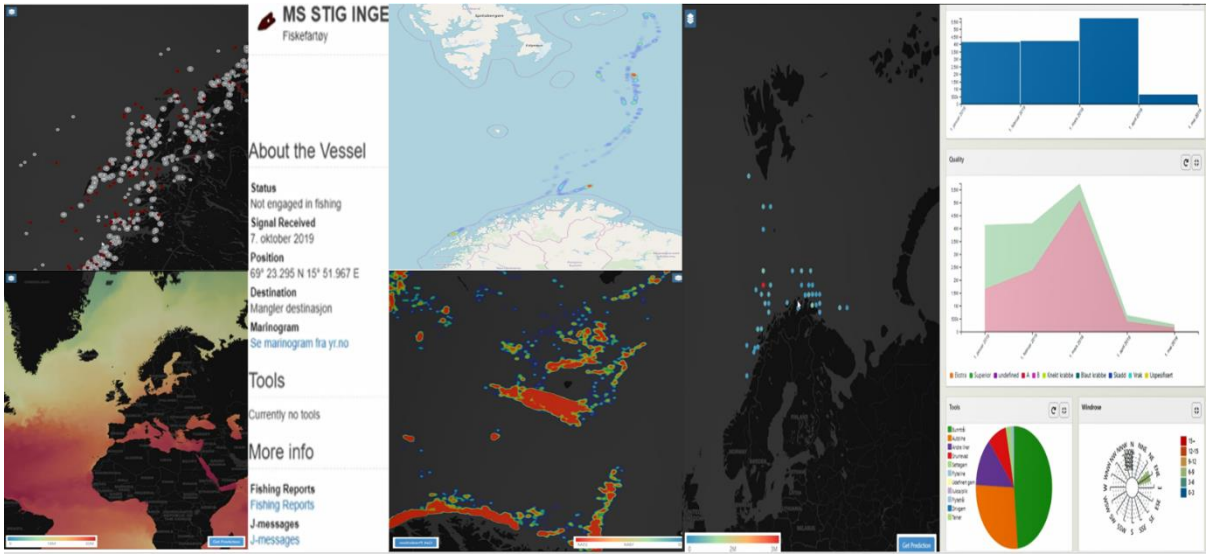


Figure 57: Additional layers/information elements of the decision support application

- This is even more evident in the plot below it. The middle bottom plot displays the output of the best-predicted fishing locations for a selected date as a heatmap based on the results of the trained neural network model. This is based on private catch data. The area is zoomed in north of Norway with the good fishing grounds around Bjørnøya in the middle of the picture. The heatmap details improve as one magnifies the map.
- The larger figure to the right (map and charts) shows the open catch data in both the map and graphs, except the lower right chart that shows the dominant wind direction. In this case, the catch location precision does not improve as one zoom in, since the lat/lon coordinates are the centre positions of the official fishing regions of the Norwegian Directorate of Fisheries.

The development iterations of the WP4 demo pilot application were pitched and show-cased, in the aforementioned series of three hackathons, and received excellent reviews, particularly in its last year of implementation:

- Third price in the first OpenSeaLab hackathon in 2017
- First price in the EuroGEOSS/Lisbon INSPIRE 2019 hackathon
- Two industrial awards at the OpenSeaLab 2019 hackathon: VLIZ and OVH 1st prices

3.8.3.4 Summary

The "Generic pipeline for fisheries decision support" and the pilot demo implementation based on the reusable (sub) pipelines from this design has been tested in an alternative fishery versus what it was designed for. The reason for staying within the fishery domain was motivated by a combination of data availability and synergies with the WP3 pilot pipeline development for pelagic fisheries.

However, as it was noted in Figure 53, the domain-specific elements of the pipeline are mainly the selection of data sources for the domain and the objective of the prediction analytics component and visualization. Recently, both the datascraper, Vespa search (open source) and the SINTIUM (C06.02) component has been applied to another domain: Converting, collating, analysing and visualizing findings from text data harvested from social and online news media and discussion forums in relation to the Norwegian regional election in 2019 with the objective of detecting fake political news and systematic attempts to influence the election through these media. This shows the flexibility of the components involved, they are applicable for alternative applications within the domain as well as across very different domains, i.e. generalized in both the technical and conceptual aspects.

4 DataBio pilot services

1.1 WP1 - Agriculture

4.1.1 Pilot 1 [A1.1] Precision agriculture in olives, fruits, grapes

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

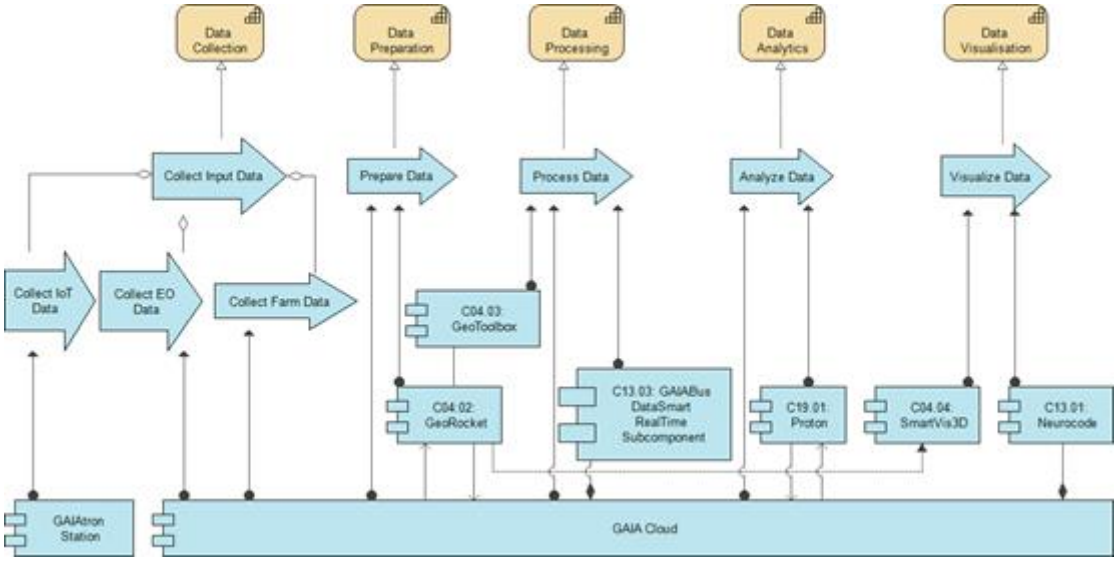


Figure 58: Pilot 1 [A1.1] Precision agriculture in olives, fruits, grapes pipelines

The pipeline shown can process a large amount of data in a reasonable time. Additionally, the components are able to visualise vector data in various clients (effectively exhibiting that interchangeable nature of clients in the pipeline). In particular, data visualization allows data filtering by attribute or attribute-based colour coding on the fly. The pilot uses the following technological components:

- C04.02 GeoRocket (Fraunhofer)
- C04.03 GeoToolbox (Fraunhofer)
- C04.04 SmartVis3D (Fraunhofer)
- C13.01 Neurocode (NP)
- C13.03 GAIABus DataSmart Real-time Subcomponent (NP)
- C19.01 Proton (IBM)

4.1.1.1 Brief summary of results and pending issues from Trial 1

The pilot has been proven successful in Trial 1 as a considerable reduction in the use of agricultural inputs has been witnessed. This has been achieved as farmers and the agricultural advisors showed a collaborative spirit and followed the advice that were generated by DataBio's solutions. As multiple parameters (climate and crop type related) are affecting the

agricultural production it has been proven that a solution "one-fits-all" is not applicable and several factors need to be taken into consideration in translating the trial results (e.g. biennial bearing phenomenon in olive trees, heavy seasonal/regional rains, multi-year fertilization strategies, etc.).

4.1.1.1.1 Earth Observation aspect

The pilot collects **EO data** from Sentinel 2 satellites using the GAIA Cloud component (Pilot component of NP for Sentinel 2 data collection, correction and extraction of indices).

The main - EO related **result** from Trial 1 comprises of the EO data management pipeline that supports the automatic and in regular intervals assignment of NDVI vegetation indices at the monitored agricultural parcels. Moreover, the pipeline optimally handles the EO-related data in order to extract meaningful insights and lead to an enhanced decision-making process when complemented with captions from the crop's phenological stage (agronomic aspect).

4.1.1.1.2 Planned EO related changes for Trial 2

- What is the lesson learned from Trial 1?

The contribution of the EO aspect to the provision of smart farming advisory services should not be neglected. EO data can offer geospatial and temporal distributions of vegetation indices and other EO products, thus, highlighting hidden information, potential abnormalities and/or correlations within the data that would not be visible otherwise.

- Is there a need for a redesign for Trial 2?

No

4.1.1.2 Updates on the work done on Trial 2

Changes for Trial 2:

- C13.01: User Interface integration was performed so that the farm management portal (holding all data of agronomic value and the embedded DSS serving as the endpoint for providing the advisory services) is integrated with the farm electronic calendar (the endpoint where the farmer or the agricultural advisor ingests information to the system regarding the applied cultivation practices, field-level observations, sampling, etc.). Both these tools were developed using the component C13.01. Integration activities were conducted in order to offer seamless user experience and allowing the user to carry out his/her intended operations without going back and forth across different systems,
- C13.03: Improved data representation and handling mechanisms, enabling the expansion and/or customized configuration of each GAIAtron station. As the requirements in terms of sensors deployed for in-the-field usage differ between pilot sites, it became obvious that several adaptations were necessary in respect to C13.03 and the way data was represented for both cloud-based storing and Gaiatron station configuration. More specifically, all relational and EAV (Entity-Attribute-Value) data

representations were adapted to a more flexible and scalable JSON format that performs better in a dynamic IoT measuring environment. The latter is widely acknowledged as JSON has become gradually the standard format for collecting and storing semi-structured datasets that originate from IoT devices. The adaptation to a JSON format for modelling IoT data streams allows the further processing, parsing, integration and sharing of data collections in support of system interoperability through the adaptation on well-established and favoured linked-data approaches (JSON-LD),

- C19.01: PROTON integrated endpoint pulls published events from GAIA Cloud via its provided RESTful endpoint and processes them according to the specified Complex Event Processing application rules. PROTON dashboard is integrated as the current displaying dashboard for the pipeline and is available at http://lnx-blue.sl.cloud9.ibm.com:8081/dashboard/dashboardMain.html.
- Moreover, a new dockerized endpoint is available for supporting the pilot activities and finally, one additional rule (peach disease) was examined by C19.01, pulling data from GAIA Cloud. In total GAIA Cloud provides real-time data for 3 pest and 3 disease models from the pilot sites to C19.01 for more sophisticated temporal processing.
- C04.02 - C04.04: For pilots A1.1, B1.2, C1.1 and C2.2 a new web-based visual analytics agTech platform designed that allows the integration of services on top of it for interactive exploration of heterogeneous data (including satellite imagery), AI products, aggregates and statistics in a user-friendly way

4.1.2 Pilot 2 [A1.2] Precision agriculture in vegetable seed crops

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below in Figure 59.

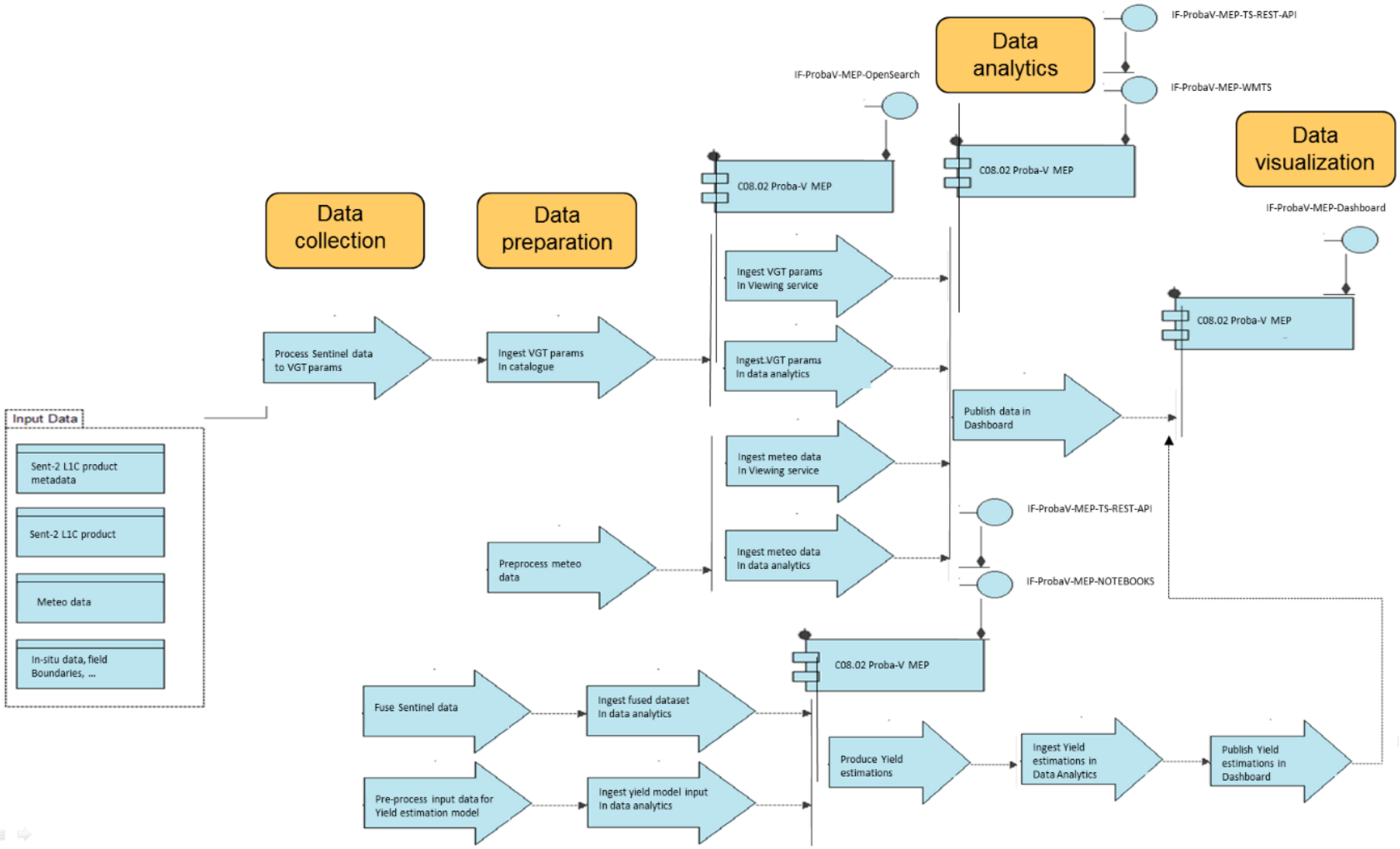


Figure 59: Pilot 2 [A1.2] Precision agriculture in vegetable seed crops pipelines

Preliminary trials in 2017 showed that differences in maturity between sugar beet fields and variability within individual fields can be spotted from satellite greenness index maps. Analysis of the growth curve and discussions with the fieldsmen made CAC seeds and VITO confident that the greenness index (fAPAR) can be used to check when the sugar beet seeds are ready to be harvested. In 2018, during Trial 1, the EO based field monitoring was extended to approximately 90 fields of seed crops. The main part of these fields were sugar beet fields. The good correlation between the "greenness" (fAPAR) index and the maturity of the sugar beet seeds was confirmed in 2018. In addition, a smaller number of fields with soybeans, sunflowers, onions and cabbages were monitored with satellite fAPAR images to investigate if seed maturity could also be assessed for these crops. Good results were obtained for soybeans and sunflowers. Trials for onions and cabbages were less successful. Based on the results of Trial 1 a model was developed to estimate the optimal harvest date for sugar beets based on the greenness index.

The pilot uses the following Earth Observation component:

- C08.02 - Proba-V MEP

4.1.2.1 Brief summary of results and pending issues from Trial 1

- The promising results for sugar beets obtained in preliminary trial were confirmed by the EO and by the in-situ monitoring during the growing season 2018, where a larger number of sugar beet seed production fields have been included in Trial 1.
- Among the new seed crops onions and cabbages showed low reliability in the correlation between index and maturity. Soybean and sunflower looked more promising.

4.1.2.1.1 Earth Observation aspect

As mentioned, the crop monitoring and maturity assessment process use Sentinel-2 **satellite images** as input for fAPAR calculation and generation of "greenness" maps and graphs. The pilot uses **C08.02 Proba-V MEP EO component** for processing, analysing and visualizing the Sentinel-2 fAPAR data.

4.1.2.1.2 Planned EO related changes for Trial 2

- What is the lesson learned from Trial 1?

A larger number of sugar beet seed production fields were included in Trial 1 (season 2018). The promising results for sugar beets obtained in the preliminary trial were confirmed by the EO and in situ monitoring.

Among the new seed crops that were monitored in Trial 1, onions and cabbages showed low reliability in the correlation between the greenness index and seed maturity. The results for soybean and sunflower looked more promising.

Coverage of clouds affected many satellite observations in late spring and early summer

- Is there a need for a redesign for Trial 2?

Cabbage and onions were no longer monitored. Sugar beet was confirmed. Sunflower and soybean will be monitored more closely during Trial 2 to assess the correlation between the greenness index and the technical maturity of the crops.

The current VITO EO components (C08.02 – Proba-V MEP), consisting of a processing platform and services, are being extended to serve the needs of the pilot. During Trial 1 the following requirements that will be implemented/available for Trial 2 have been identified:

- A first solution to cope with cloudy data inherent to the Sentinel-2 images, VITO is extending the existing services with the possibility to query the amount of clouds above an area of interest. This allows pilots to get time-series based information about the percentage of clouds above a field or area of interest. This information can be used in the pilot applications to only show those dates that have relevant and sufficient information.
- As a second step, a data fusion algorithm has been developed to cope with cloudy images inherent to the use of Sentinel-2 data. This data fusion will use both Sentinel-1 and Sentinel-2 data to provide cloud-free time series.
- VITO is also optimizing the backbone infrastructure and services to cope with the growing scope of the applications. This enables an increased performance for the pilots using the VITO EO components.
- A maturity assessment model will be created. The goal of this model is to provide advice to the end-user about the optimal harvesting date.

4.1.2.2 Updates on the work done on Trial 2 and main results

During Trial 2 (2019 season) 77 sugar beet fields and 41 soybean fields were monitored by CAC seeds. Sentinel-2 satellite images provided information on the greenness and health of the crops. From the greenness (fAPAR) curves, the optimal harvest date of the sugar beet and soybean seeds was estimated in near real-time, using the maturity model developed during Trial 1.

EO component C08.02 – Proba-V MEP was updated to cope with cloudy images. Sentinel-1 data were added as input layer and by combining Sentinel-1 and Sentinel-2 images using a data fusion algorithm (CropSAR) cloud-free fAPAR time series could be generated.

From Trial 2 it was found that:

- Sentinel-2 images are beneficial for crop monitoring as they provide information on crop development and variability in crop growth within and between fields.
- It is possible to estimate the optimal harvest date from the fAPAR curve with sufficiently high accuracy (+/- 2 days) when using fAPAR threshold values.
- The accuracy of the harvest date estimation increases when using fused Sentinel-1 and Sentinel-2 fAPAR values are used, especially in cloudy periods.

- The maturity model that is currently used to "forecast" the optimal harvest date in near real-time (simple linear approach to estimate the date that the fAPAR threshold is reached) is not accurate enough. The performance of the model could be further improved by using more advanced modelling techniques such as machine learning techniques to predict fAPAR values and/or by using additional input for modelling such as meteo data (rainfall, temperature).

Given the high interest and the promising results of this pilot, CAC seeds and VITO are currently exploring the possibilities to continue the pilot after the end of the DataBio project.

4.1.3 Pilot 3 [A1.3] Precision agri-culture in vegetables_2 (Potatoes)

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below in Figure 60.

The pipeline presented in Pilot [A1.2] also applies horizontally and similarly to Pilot [A1.3].

About 110-150 ha was monitored in the trial (in the North of the Netherlands). Based on these fields EO data was collected for crop monitoring and establish a correspondence between the yield samples were taken in the field and the parameters derived from the satellite information from Sentinel-2 satellite images and the "greenness" maps of the target-fields were derived throughout the season.

The **Trial 1 results** are visualisation of fAPAR (biomass index) from Sentinel 2 EO data of the area of interest, presenting new imagery every 5-10 days (if cloud coverage permits). The WatchItGrow app was used for the farmers for data entry of parcel information like crop variety, plant date etc.

The pilot uses the following Earth Observation component:

- C08.02 - Proba-V MEP

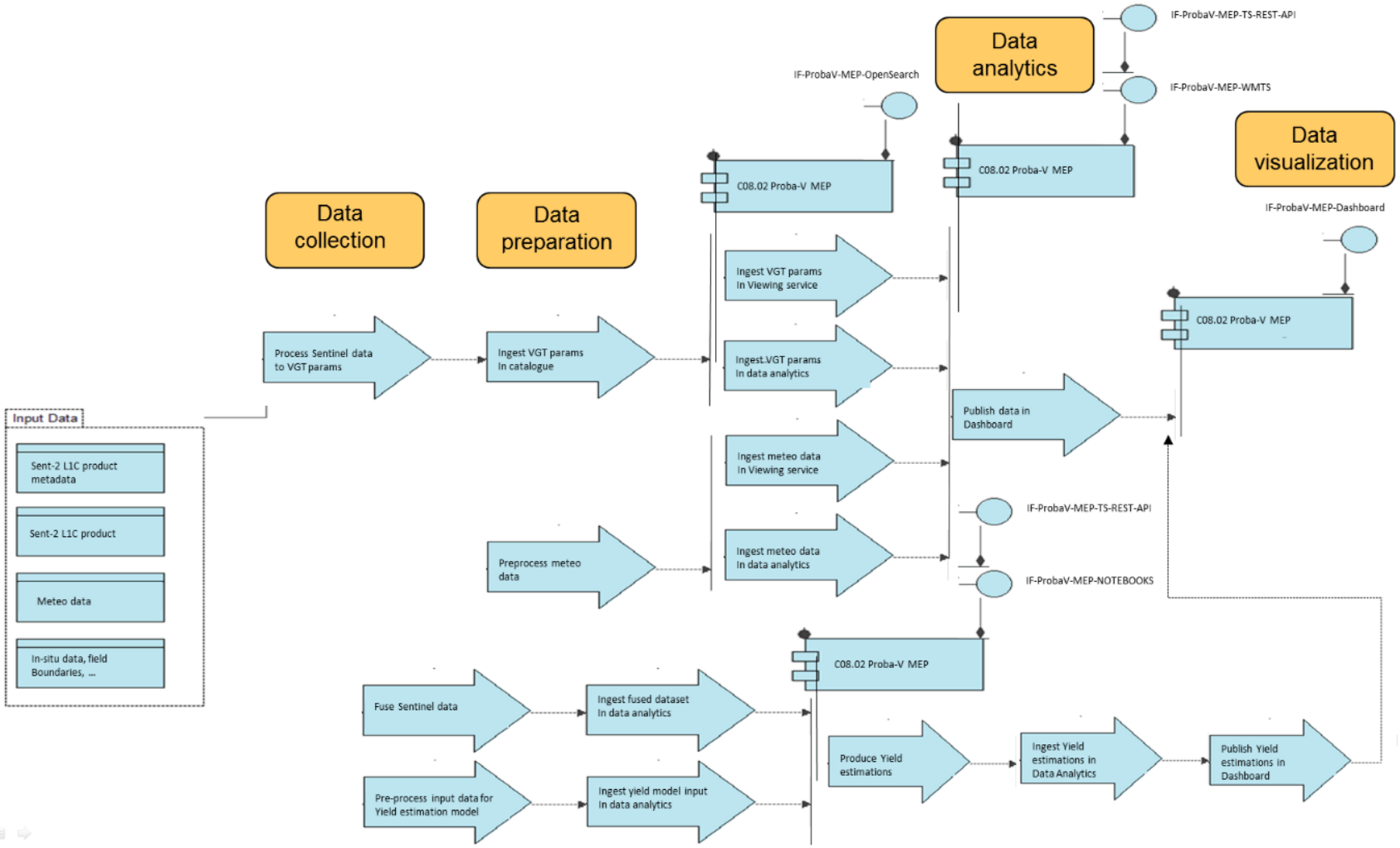


Figure 60: Pilot 3 [A1.3] Precision agriculture in vegetables_2 (Potatoes) pipelines

4.1.3.1 Brief summary of results and pending issues from Trial 1

The farmer gets overloaded with non-actional information by providing him just the Sentinel satellite imagery for 'greenness' in the online platform.

In discussion with farmers, we concluded that farmers want more informative data, so they can act upon the results. We decided to try to improve the benchmark capabilities of the solution by introducing the crop growth model WOFOST as a reference for the actual growth, which will be monitored using EO images.

4.1.3.1.1 Earth Observation aspect

The pilot uses **C08.02 Proba-V MEP EO component** for processing the Sentinel-2 data. The Sentinel-2 **satellite data** were used for "greenness" maps of the target-fields calculation. The "greenness" maps are showing the fraction of Absorbed Photosynthetically Active Radiation (fAPAR), a measure of the crop's primary productivity. fAPAR is often used as an indicator of the state and evolution of crop cover. Low fAPAR values indicate that there is no crop growing on the field (bare soil, fAPAR=0). When the crop emerges, the index will increase until the crop has reached the maximum growing activity (fAPAR=95-100%). Then the fAPAR will decrease again until harvest. From this "crop growth curve" information on phenology and crop development can be retrieved and a model can be designed to predict the yield in an early stage of the growing season (July).

4.1.3.1.2 Planned EO changes for Trial 2

For Trial 2 VITO is working to improve the data curation for cloud coverage. The C08.02 'Proba-V MEP' components offer a Virtual Machine interface and a Jupyter Notebook interface, for researchers and developers. These are used by VITO researchers in WP1 to e.g. work on the fusion of datasets (e.g. Sentinel-1 and Sentinel-2) to provide denser time series than those realised by only using Sentinel-2 because of frequent clouds.

The following Big Data analytics were added during Trial 2: diagnostic, predictive and prescriptive; combining data sources will give input for diagnostics and prediction. Diagnostic analysis needs to tell farmers what brings crops to grow less than expected (yield gap). Predictive analytics needs to give the farmers and the processing industry insight about the yield to be expected in order to facilitate the selling process. Farmers need prescriptive analytics to facilitate them in their decision-making process to improve the plant growth, plant health and yield. To this scope it would be useful to exploit the EO component C08.02 Proba-V MEP which offers powerful data and analytics capabilities.

4.1.3.2 Updates on the work done on Trial 2

For Trial 2 the WOFOST crop growth model was used for predictive analytics on yield. A decision support system was introduced providing an alert-service that will send farmers timely and automated identification of potential yield losses in their fields, saving time and effort to monitor the crop. With feedback information from field visits the system could combine high throughput of field and satellite data with machine learning algorithms.

Eventually, it might be able to autonomously explain the causes of field problems to the farmers.

Technological changes in Trial 2:

- Additional datasets
    - soil data and
    - day-to-day weather data (temperature, precipitation, radiance, evapotranspiration).
    - multispectral images using a UAV (for ground truthing the EO data and crop growth model for different varieties of potatoes)
    - soil moisture sensors
- Additional components
    - Scripts for automated search and transformation of Sentinel-2 data into WOFOST format
    - Scripts for automated retrieval of weather data
    - Calibration of the WOFOST model

4.1.4 Pilot 4 [A2.1] Big Data management in greenhouse eco-system

The pilot uses the following technological component:

- C22.03 - Genomic models

The CREA's Genomic models component was registered under the DataBio platform as component C22.03 (https://www.databiohub.eu/registry/#services?tag=C22.03), while the pilot A2.1 (WP1 Pilot 4 [A2.1] Big Data management in greenhouse eco-system) was registered under DataBio platform at https://www.databiohub.eu/registry/#services?name=a2.1. Alternative modelling solutions were implemented with the main aim of increasing predictive accuracy, reducing predictive bias, and shortening generation intervals. Single trait, multi-trait, and genotype x environment models were implemented. We were able to model optimum index selection and to borrow information from correlated environments. From these breakthroughs, important ramifications are expected: simultaneous improvement of several traits, predicting not only untested genotypes but also untested environments with significant savings.

The following figure depicts a top-level pipeline following the DataBio Data Value chain, adapted to fit C22.03 Genomic models component: The top-level pipeline contains the following four major steps:

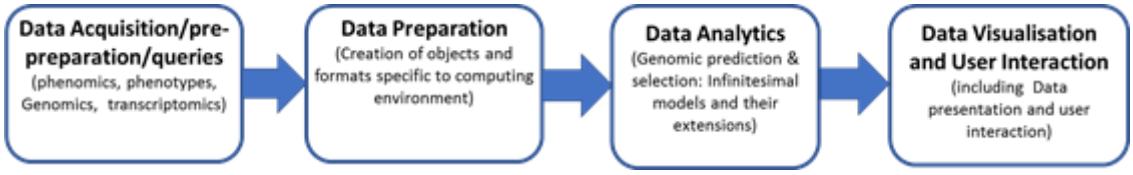


Figure 61: Pilot 4 [A2.1] Big Data management in greenhouse ecosystem top-level pipeline

Below is the simplified deployment view of the component:

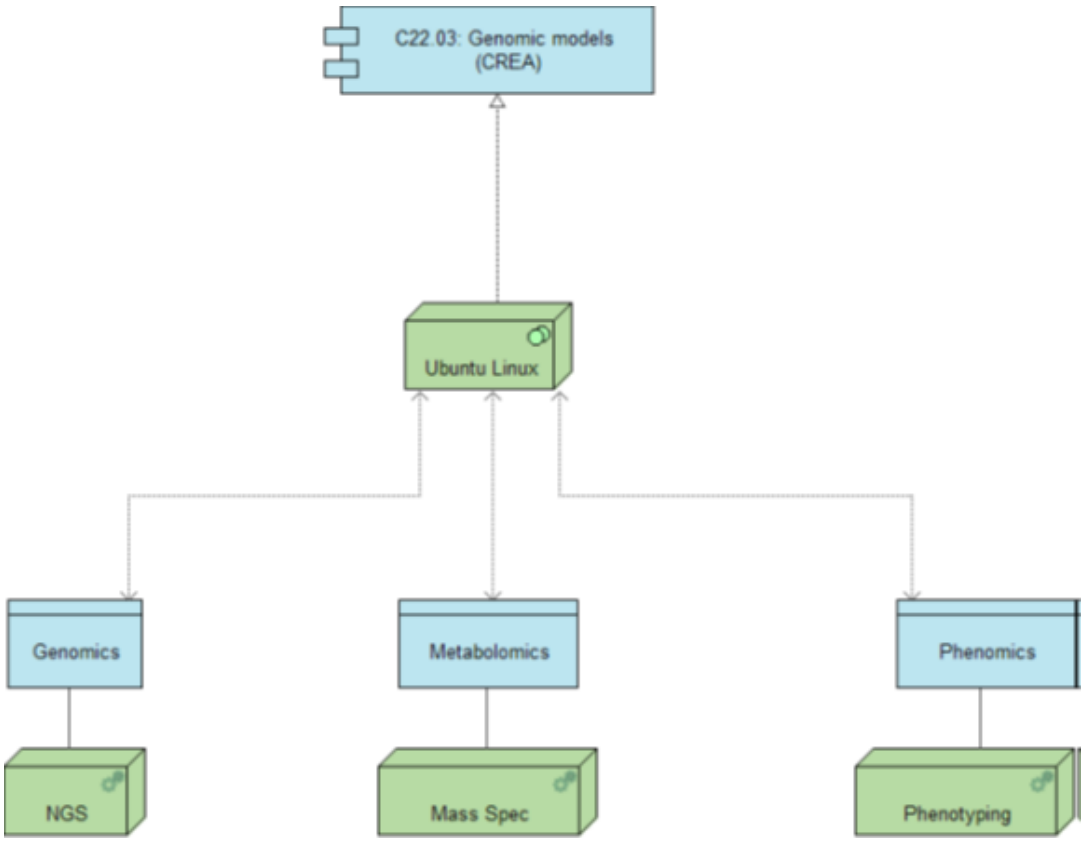


Figure 62: Pilot 4 [A2.1] Big Data management in greenhouse ecosystem pipelines

It is interesting to notice that the C22.03 component is implemented against a crop breeding pipeline schematized at a high level of abstraction as follows:

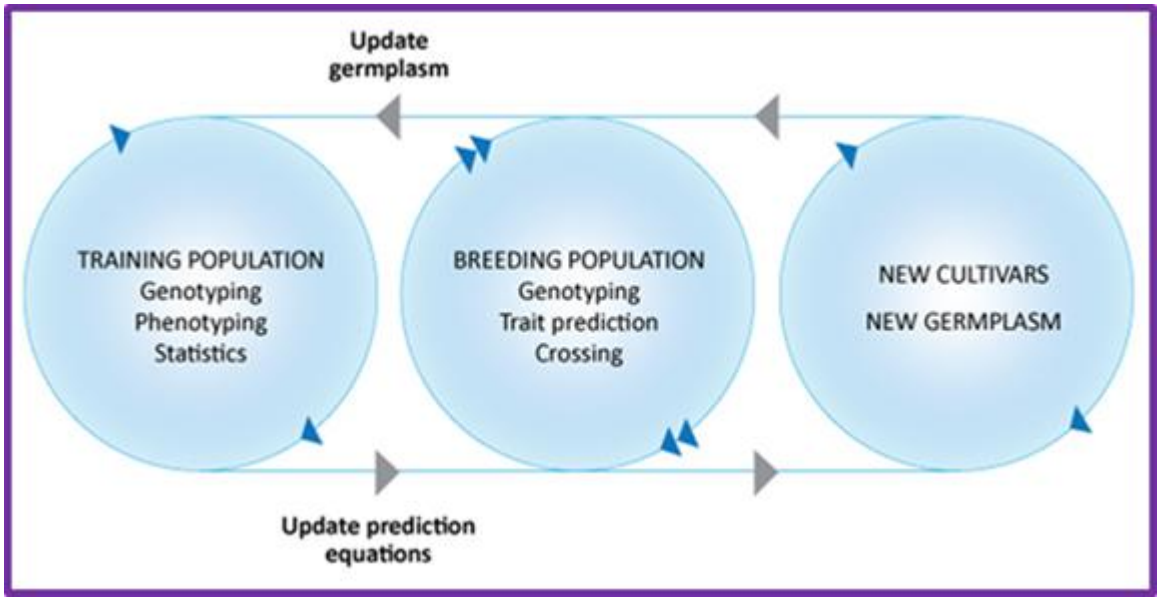


Figure 63: A crop breeding pipeline used in the implementation of C22.03 component

4.1.4.1 Brief summary of results and pending issues from Trial 1

The preliminary and the first-year trials were mostly concerned with setting up the infrastructure and the technological platform. IoT, phenomic, genomic and biochemical data were collected. Genomic models were selected but not usefully implemented under real-world breeding settings due to the current low size of tomato samples phenotyped; this situation was expected to improve in the second year of trials. To achieve the predictive analytics run in 2018, available public datasets were used, and the outcome was encouraging. The analytics anticipated a single and several environments to mimic single or several glasshouses. The DataBio algorithms that were implemented were described in the BDVA reference model diagram, all of which were referred to as C22.03 (genomic models) component designed and deployed by CREA. Our findings in Trial 1 showed that genomic models perform comparably under single environments. Most notably, and as a hallmark of the first trial results, our findings showed that accounting for marker information x environment or implementing the reaction norm model performed comparably and produced superior results. Most notably, a GS Demo was presented in 2018 and CREA and CERTH were selected by the European Commission's Innovation Radar as innovators in the Deep Tech area at exploration level of maturity as documented in the below link: https://www.innoradar.eu/innovation/30863.

4.1.4.2 Updates on the work done on Trial 2

The production of tomato Big Data from the CERTH's Greenhouses was slower than anticipated due to the need for CERTH to produce new genetic data, in order to assess the genetic variability of the crosses and the collection of environmental and phenotypic data. It was not therefore possible to validate the C22.03 on tomato related data. In the last year of the project, the potential of GS algorithms was successfully evaluated in sorghum crops to improve health-promoting compounds (phenolic compounds, flavonoids, tannins, and total

antioxidant capacity) used to manufacture health-promoting and specialty foods. Publicly available datasets including those produced by CREA within the framework of the project DataBio, were ingested and fed to C22.03 models in order to increase the size of training populations and improve model predictive ability. The analytics outcomes were encouraging and were documented in peer-reviewed scientific articles published on specialized international journals with high impact factor (PMID: 31653099 and DOI: 10.3390/genes10110841; Doi: 10.1371/journal.pone.0225979, and Mauscript ID ijms-663847). The diagram below Highlights our findings. Solutions included modelling single trait, multi-traits (optimum index selection), and multivariate GS analytics factoring in genotype x environment interaction and borrowing information from correlated locations to successfully predict the performance of untested environments.

4.1.5 Pilot 5 [B1.1] Cereals and biomass crop

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

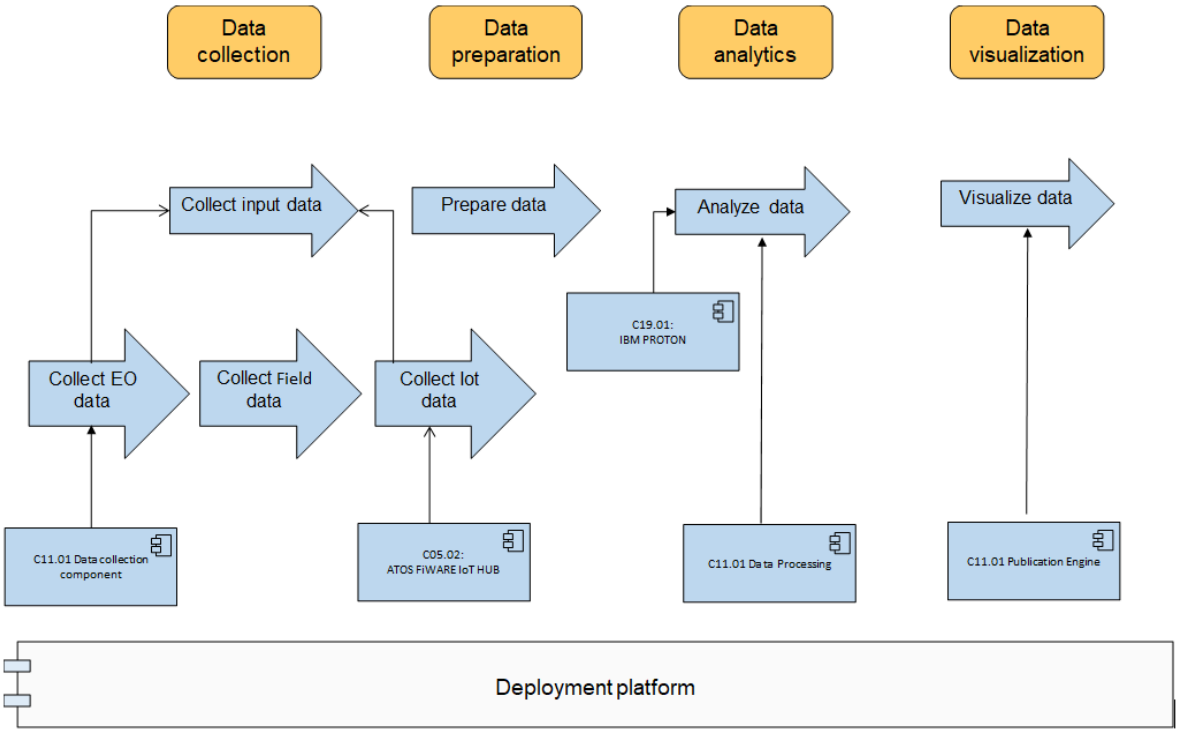


Figure 64: Pilot 5 [B1.1] Cereals and biomass crop pipelines

During Trial 1 agricultural datasets (EO data, sensors data, administrative CAP data, ...) have been gathered and TRAGSA-TRAGSATEC has already started the processing and transformation processes.

One of the main conclusions of this pilot performance for the first period has been that the selection of just one Irrigation Area, despite its interests, highlighted the need for more piloting areas. Due to that, TRAGSA involved new Irrigation Communities in Andalusia.

Because of this, piloting area has increased from 15.000 Ha to 40.000 Ha in the second and final iteration. Therefore, due to the correct development, there has been any redefinition for Trial 2. This pilot is under correct development and it will achieve its defined goals on time.

This pilot uses the following technological components:

- C11.01 - Advanced Irrigation and vigour Monitoring
- C19.01 - IBM Proactive Technology Online
- C05.02 - FIWARE IoT Hub

4.1.5.1 Brief summary of results and pending issues from Trial 1

During Trial 1 agricultural datasets (e.g., EO data, sensors data, administrative CAP data) have been gathered and TRAGSA-TRAGSATEC had already started the processing and transformation processes. Trial 2 were devoted to upgrading the processing tools from few-parcels-input to local and provincial level. The assessment of satellite image and cloud processing service platforms were carried out during Trial 1, and this evaluation has been made through its application to the development of C11.01 component, in order to obtain full functionality in the web application. The vigour maps are based on NDVI – or other indices - built on high frequency, scalable satellite image data at national level. Besides this, an algorithm for the calculation of water needs has also been developed (Applied to the Genil-Cabra pilot zone). The Irrigation Community (Farmers association) has provided this pilot with plot and crop data from 2017. Likewise, Sentinel images, from the same period, have been used in the developed crop-classification process. The final technical result is an accurate map of NDVI (Normalized Difference Vegetation Index) which is the foundation of the water needs calculation algorithm. In Trial 2, TRAGSA and TRAGSATEC will address the integration of the data in the management tool that will provide the specific water needs.

Only text information has been implemented for Trial 1. A GIS viewer was added to the alphanumeric tool. Both management tool, processing algorithm and viewer are the main components of C11.01. For this pilot, it has been asked for support in utilizing Earth Observation data through online platforms.

4.1.5.1.1 Earth Observation aspect

The pilot uses the following **Earth Observation data**: Sentinel-2 (from ESA), Orthophotos (from the National Geographic Institute of Spain) and RPAS: thermal and multispectral observation data from TRAGSA. Parcels are identified by SigPAC (vectors provided by Junta de Castilla y León).

The pilot uses two Earth Observation related components:

- C11.01 Visualization Service on irrigation and vigour status
- C11.03 Radiometric improvements of Orthophotos

The **EO related results** of the pilot are four synthetic maps:

- Large-scale dataset: this set of raster data identifies the major land uses: agricultural, forestry, pasture, unproductive, water and urban.
- Change dataset: The changes observed are grouped into 3 classes: change, doubt and no change.
- Crop dataset and soil cover: this raster dataset is generated on the agricultural land mask of LPIS. It supposes the maximum level of disaggregation of coverages/crops/ land uses to be achieved in each zone, according to the reference data used.
- Dataset of discrepancies between the CAP declarations and the crop dataset obtained by remote sensing.

4.1.5.1.2 Planned EO related changes for Trial 2

Trial 2 has been devoted to upgrading the processing tools from a few-parcels-input to the local and provincial levels. The assessment of satellite image and cloud processing service platforms were carried out during Trial 1, and this evaluation has been made through its application to the development of C11.01 component, in order to obtain full functionality in the web application. The vigour maps are based on NDVI – or other indexes - built on high frequency, scalable satellite image data at national level. Besides this, an algorithm for the calculation of water needs has also been developed (applied to the Genil-Cabra pilot zone). The Irrigation Community (Farmers association) has provided this pilot with plot and crop data from 2017. Likewise, Sentinel images, from the same period, have been used in the developed crop-classification process. The final technical result is an accurate map of NDVI (Normalized Difference Vegetation Index) which is the foundation of the water needs calculation algorithm. In Trial 2, TRAGSA and TRAGSATEC will address the integration of the data in the management tool that will provide the specific water needs.

4.1.5.2 Updates on the work done on Trial 2

Nevertheless, the good performance, there were some changes to be implemented in Trial 2. The goal of the application is to take advantage of the new IoT systems in TRAGSA's pilot located in Maceda, which allows measuring environmental variables as temperature, humidity, air pressure and process this information with CEP to find meaningful patterns for decision making for predictive maintenance activities. The farmer is notified when a specific situation demanding his action is detected. The integration involves FIWARE Orion context broker, which saves raw alert entities (being notified by the field sensors) in a pre-agreed upon format. Alert entities represent different environment variables, such as temperature reading. PROTON subscribes to updates to the alert entities, is notified by FIWARE Context Broker once the state of the entity is changed (temperature going up/down), and processes these changes according to predefined CEP rules to find patterns and fire derived events indicating problematic situations for the farmer. This tool will allow the useful processing of Big Data sources (now stored in a MongoDB No-SQL Database) to be tapped into a useful way both by managers of the irrigation communities and by the farmers themselves. In regards of DataBio tools and components, it is possible to mention:

- C05.02: Latest version of software installed and the development of new scripts to insert historical data in netcdf format plus insert vessel data.
- C05.02: Communication with Proton Component through RESTful NGSI consumers and producers for FIWARE context broker and development of new python scripts to gather data from sensors.
- C19.01: New dockerized endpoint.
- C19.01: Features in use for the trial: Proton CEP engine (generic complex event processing engine used to implement monitoring rules for a number of use cases), Proton adapters used for interfaces with the components (receiving raw events and emitting complex events) Implemented: CEP application design for monitoring environment parameters (humidity, temperature).
- C19.01: RESTful NGSI consumers and producers for FIWARE context broker.

4.1.6 Pilot 6 [B1.2] Cereals and biomass crop_2

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

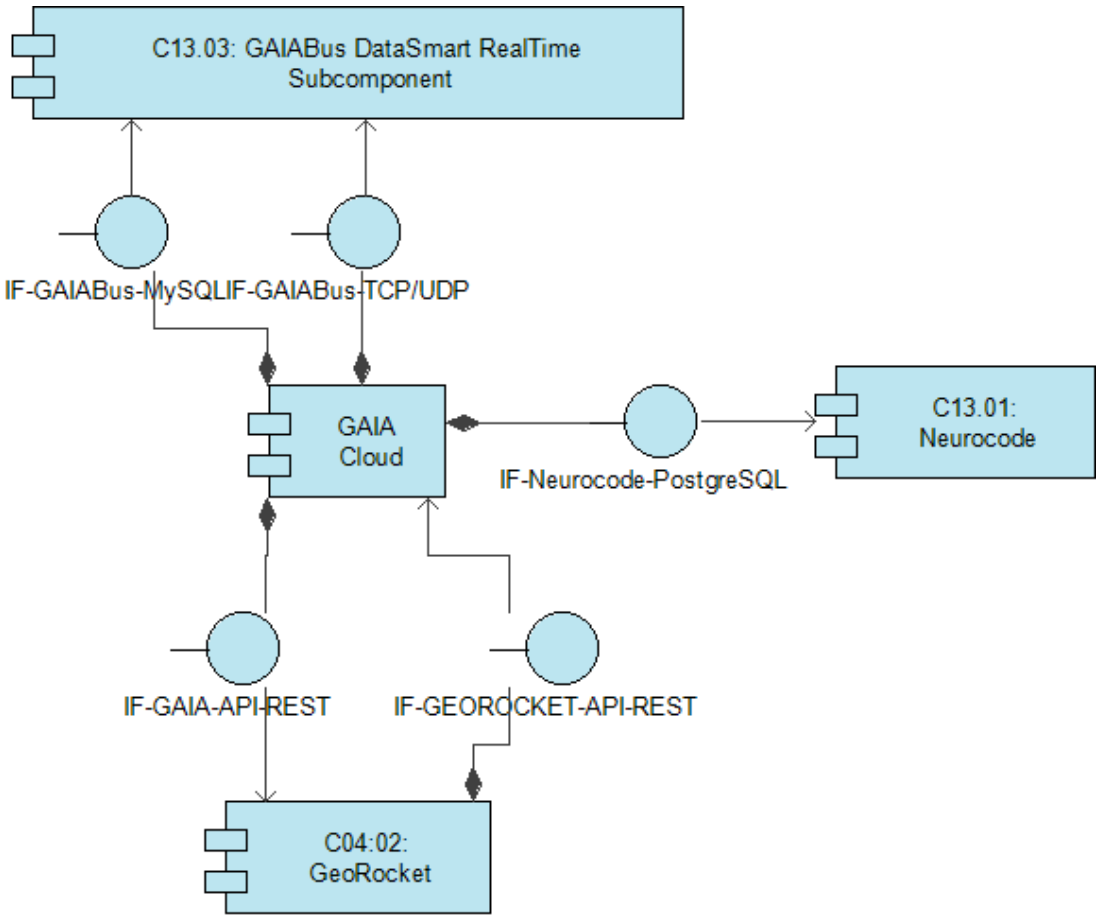


Figure 65: Pilot 6 [B1.2] Cereals and biomass crop_2 pipelines

This pilot uses the following technological components:

- C04.02 – GeoRocket (Fraunhofer)
- C04.03 – GeoToolbox (Fraunhofer)
- C04.04 - SmartVis3D (Fraunhofer)
- C13.01 - NeuroCode (NP)
- C13.03 - GAIABus DataSmart RealTime Subcomponent (NP)

4.1.6.1 Brief summary of results and pending issues from Trial 1

The pilot has been proven successful in Trial 1 as a considerable reduction in the use of agricultural inputs (freshwater) has been witnessed. This has been achieved as farmers and the agricultural advisors showed a collaborative spirit and followed the advice that were generated by DataBio's solutions. As multiple parameters (climate and crop type related) are affecting agricultural production, it has been proven that a solution "one-fits-all" is not applicable and several factors need to be taken into consideration in translating the trial results (e.g. heavy seasonal/regional rains, etc.).

4.1.6.1.1 Earth Observation aspect

The pilot collects **EO data** from Sentinel 2 satellites using the GAIA Cloud component (Pilot component of NP for Sentinel 2 data collection, correction and extraction of indices and evapotranspiration parameters).

The main - EO related **result** from Trial 1 comprises of the EO data management pipeline that supports the automatic and in regular intervals assignment of NDVI vegetation indices and evapotranspiration parameters at the monitored agricultural parcels. Moreover, the pipeline optimally handles the EO-related data in order to extract meaningful insights and lead to an enhanced decision-making process when complemented with captions from the crop's phenological stage (agronomic aspect).

4.1.6.1.2 Planned EO related changes for Trial 2

- What is the lesson learned from Trial 1?

The contribution of the EO aspect to the provision of smart farming advisory services should not be neglected. EO data, within this pilot can contribute to evapotranspiration monitoring. Evapotranspiration offers the means for effectively estimating the soil moisture reduction rate of a given parcel (and thereby the present water content). Combined with soft facts derived from the farmer itself (e.g. date of irrigation, amount of water used, etc.), the system is able to generate irrigation advice without the need for a vast density of hardware infrastructure, thus leading to economies of scale and spectrum. Moreover, the EO aspect can offer geospatial and temporal distributions of vegetation indices and other EO products, thus, highlighting hidden information, potential abnormalities and/or correlations within the data that would not be visible otherwise.

- Is there a need for a redesign for Trial 2?

No

4.1.6.2 Updates on the work done on Trial 2

- C13.01: User Interface integration was performed so that the farm management portal (holding all data of agronomic value and the embedded DSS serving as the endpoint for providing the advisory services) is integrated with the farm electronic calendar (the endpoint where the farmer or the agricultural advisor ingests information to the system regarding the applied cultivation practices, field-level observations, sampling, etc.). Both these tools were developed using the component C13.01. Integration activities were conducted in order to offer seamless user experience and allowing the user to carry out his/her intended operations without going back and forth across different systems,
- C13.03: The improved data representation and handling mechanisms, enabling the expansion and/or customized configuration of each GAIAtron station. As the requirements in terms of sensors deployed for in-the-field usage differ between pilot sites, it became obvious that several adaptations were necessary in respect to C13.03 and the way data was represented for both cloud-based storing and Gaiatron station configuration. More specifically, all relational and EAV (Entity-Attribute-Value) data representations were adapted to a more flexible and scalable JSON format that performs better in a dynamic IoT measuring environment. The latter is widely acknowledged as JSON has become gradually the standard format for collecting and storing semi-structured datasets that originate from IoT devices. The adaptation to a JSON format for modeling IoT data streams allows the further processing, parsing, integration and sharing of data collections in support of system interoperability through the adaptation on well-established and favoured linked-data approaches (JSON-LD),
- C04.02 - C04.04: For pilots A1.1, B1.2, C1.1 and C2.2 a new web-based visual analytics agTech platform was designed that allows the integration of services on top of it for interactive exploration of heterogeneous data (including satellite imagery), AI products, aggregates and statistics in a user-friendly way.

4.1.7 Pilot 7 [B1.3] Cereal and biomass crops_3

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

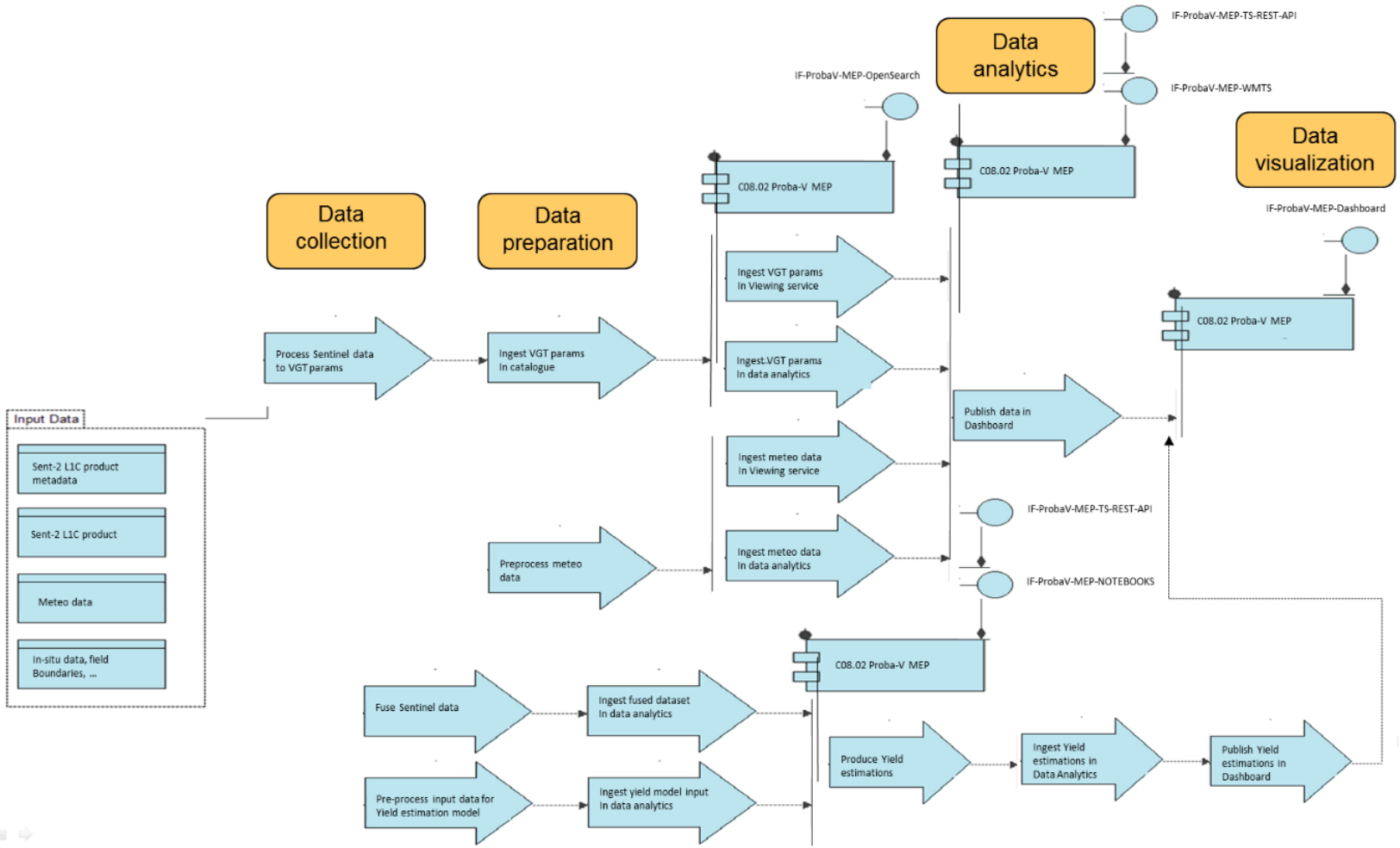


Figure 66: Pilot 7 [B1.3] Cereal and biomass crops_3 pipelines

The pipeline presented in Pilot [A1.2] also applies horizontally and similarly to Pilot [B1.3] and so do its results.

This pilot uses the following technological components:

- C12.03 - EO4SDD

and the following focused on Earth Observation:

- C08.02- Proba-V MEP

4.1.7.1 Brief summary of results and pending issues from Trial 1

The results from the preliminary and the second trials were integrated, and a scientific paper (https://doi.org/10.3390/agronomy9040203) were published in a specialized journal with high Impact Factor. The lesson learned from these trials is that remote sensing technology holds tremendous potential in biomass sorghum business. We were able to predict biomass yields and detect foliar diseases early on (up to six months before harvest) within the cropping season. Overall, Bayesian machine learning technology gave a superior predictive performance.

4.1.7.1.1 Earth Observation aspect

Using **satellite data** and machine learning techniques in sorghum pilots, we were able to predict biomass yields with high precision 6 months ahead of harvesting as shown in the below figure. This pilot uses the C08.02- Proba-V MEP component focused on Earth Observation:

4.1.7.1.2 Planned EO related changes for Trial 2

- As the VITO EO components are shared amongst the different DataBio pilots, they foresee the same changes for pilot B1.3 as mentioned in pilot A1.2. There are no additional requirements identified for pilot B1.3 for Trial 2 as the feedback is the same as the one provided by the other pilots. Below is a short summarization of the features that will be available for pilot B1.3 during Trial 2:
    - A time-series based service to query the amount of clouds above an area of interest
    - Implementation of a data fusion algorithm based on Sentinel-1 and Sentinel-2 data

    Optimizations to VITO infrastructure and services for an increased performance

4.1.7.2 Updates on the work done on Trial 2

- C12.03: In respect to the low amount of data, the model tends to overfit on Trial 1. A lot of efforts have been done to reduce the risks. Within Trial 2 the generated model was expanded with the help of new data.
- C12.03: The tests and use of data augmentation were expanded.
- C12.03: Optional goal was to provide an API to make the model available as a service.

4.1.8 Pilot 8 [B1.4] Cereals and biomass crops_4

During Trial 1 activities, a key result for WP4 and WP5 perspectives was the definition of the pilot pipeline, which is presented below:

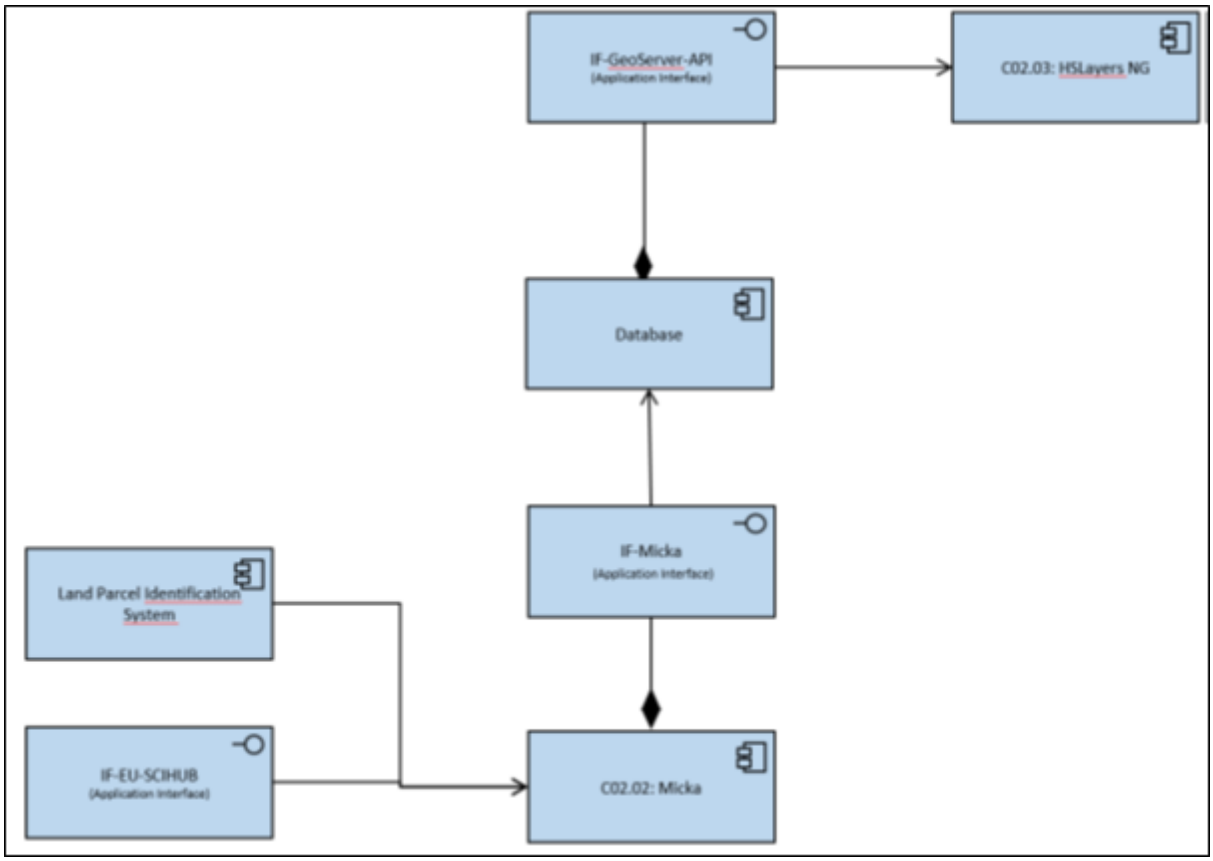


Figure 67: Pilot 8 [B1.4] Cereals and biomass crops_4 pipelines

This pilot uses the following technological components:

- C02.02 - Micka
- C02.03 - HSLayers NG
- C37.01 - Modelio BA Data modelling tool
- C37.03 - Modelio PostgreSQL modeller

4.1.8.1 Brief summary of results and pending issues from Trial 1

Although it is possible to calculate the yield potential for a large area automatically based on freely available input data, the output, in this case, is only a rough estimate. A detailed analysis to be used for planning agricultural operations will always require some manual input of agricultural advisory service and consultation with the agronomist of the farm. The technology line can reduce routine manual operations, but agricultural know-how and local farm knowledge cannot be fully replaced by algorithms at the moment nor in the near future.

4.1.8.1.1 Earth Observation aspect

The Pilot uses the following EO related data:

- Sentinel-2: Scenes covering vegetation period of cereals and meeting cloud cover criteria.
- Landsat 8: Scenes covering vegetation period of cereals and meeting cloud cover criteria.
- LPIS database: Field boundaries provided by the Czech Republic.

4.1.9 Pilot 9 [B2.1] Machinery management

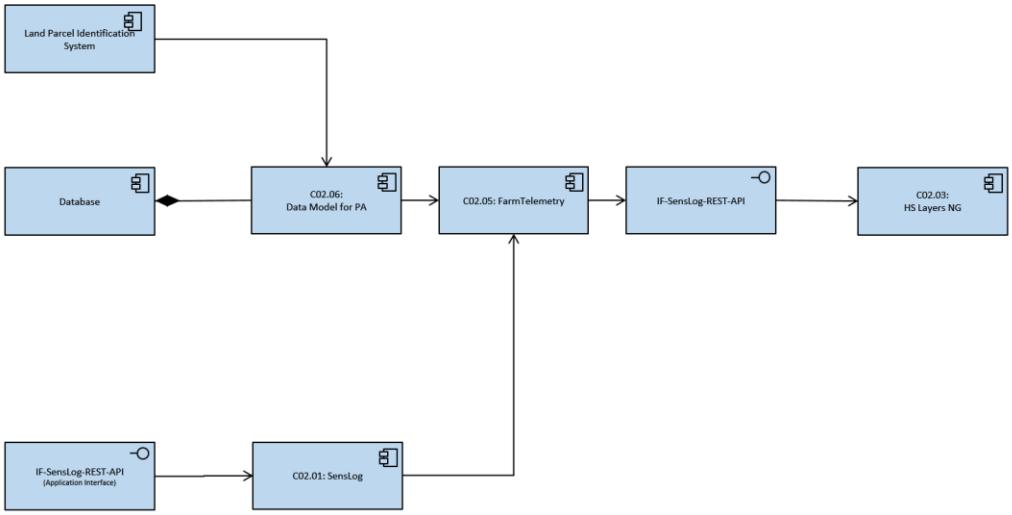


Figure 68: Pilot 9 [B2.1] Machinery management pipelines

4.1.9.1 Brief summary of results and pending issues from Trial 1

This pilot is mainly focused on collecting telemetry data from agriculture machinery from different manufacturers and from different farms. The pilot integrates telemetry data with farm data to one component to provide data on one point for further processing and visualization. The main challenge was that one farm utilizes tractors from different manufacturers using different telematic solutions. The integration of different data sources put the importance of scalability on the receiving component.

4.1.9.2 Updates on the work done on Trial 2

Technological changes for Trial 2:

- C02.01: Design and implementation of Feeder to receive data in binary form reflecting scalability of SensLog version 2 and Connectors to transform data from different sources where any influence on APIs on data source.
- C02.05: Implement a redesigned version and connect by API to SensLog v2 and to another one data source at least. Utilizing of SensLog v2 as data source.

4.1.10 Pilot 10 [C1.1] Insurance (Greece)

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

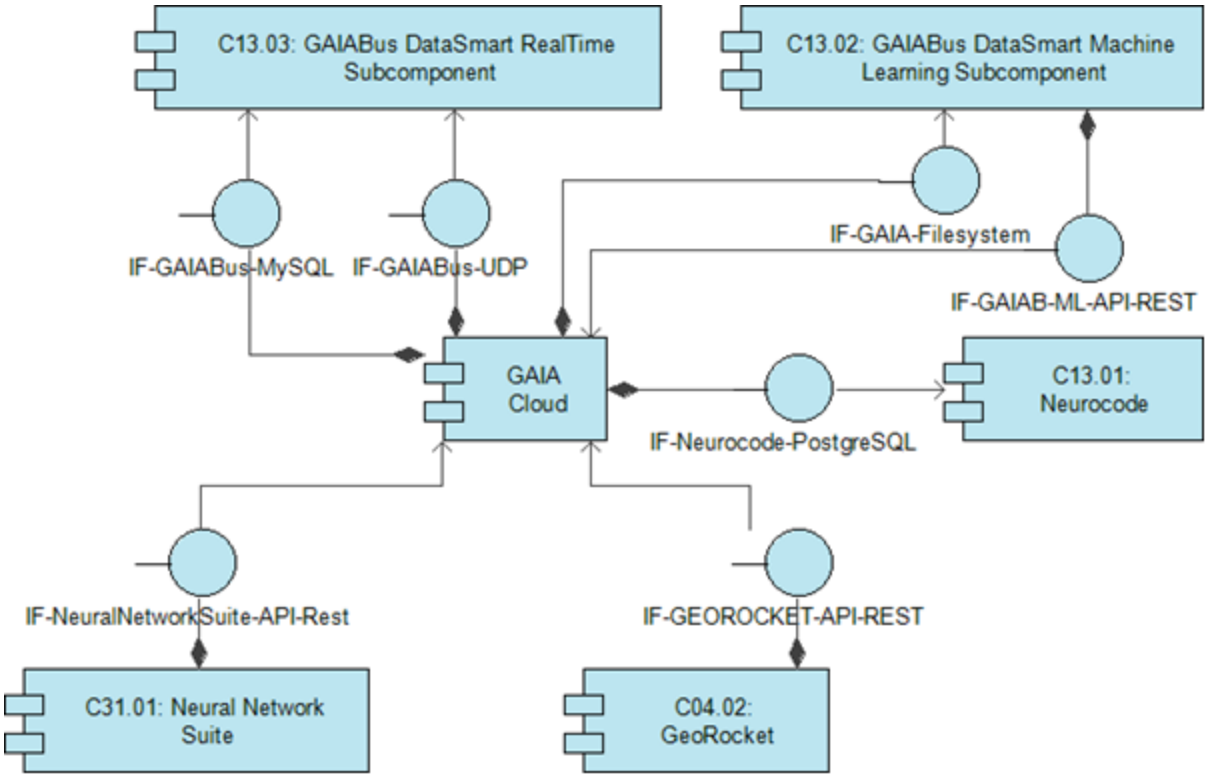


Figure 69: Pilot 10 [C1.1] Insurance (Greece) pipelines

This pilot uses following components:

- C04.02 – GeoRocket (Fraunhofer)
- C04.03 – GeoToolbox (Fraunhofer)
- C04.04 - SmartVis3D (Fraunhofer)
- C13.01- NeuroCode
- C13.02 - GAIABus DataSmart Machine Learning Subcomponent
- C13.03 - GAIABus DataSmart RealTime Subcomponent

4.1.10.1 Brief summary of results and pending issues from Trial 1

The pilot completed the first round of trials during Trial 1 in two regions, namely Evros and Thessaly, Greece. It effectively demonstrated how Big Data-enabled technologies and services dedicated to the agriculture insurance market could eliminate the need for on-the-spot checks for damage assessment and promote rapid payouts. All these assumptions have been validated through a set of pilot KPIs which in their majority met (and in some cases even exceeded) the targeted expectations.

Moreover, valuable insights have been gained from Trial 1. The role of field-level data has been revealed as their collection and monitoring is important in order to determine if critical/disastrous conditions are present (heat waves, heavy rainfalls and high winds). Field-level data can be seen as the “starting point” of the damage assessment methodology, followed within the pilot. Moreover, regional statistics deriving from this data can serve as a

baseline for the agri-climate underwriting processes followed by the insurance companies who design new agricultural insurance products.

In Trial 2, the applied technologies and pipelines will get even more mature and reach their expected TRL by the end of the project. The key pilot stakeholders (i.e. the farmers, which wish to insure their crops against weather-related systemic perils (e.g. floods, high/low temperatures, and drought) and Interamerican, as a major Greek insurance company, with increased interest in agricultural insurance products), will continue (for a second year) to benefit from the EO-based geospatial data analytics, thus, promoting the transition from traditional insurance policies to more flexible index-based methodologies. KPIs were collected along the Trial 2 in order to validate the pilot assumptions.

4.1.10.1.1 Earth Observation aspect

The pilot uses Copernicus EO data and the following components focused on Earth Observation:

- C13.02 - GAIABus DataSmart Machine Learning Subcomponent (NP)
- GAIA Cloud (Pilot component of NP for Sentinel 2 data collection, correction and extraction of indices)
- C31.01 - Neural network suite for image processing (CSEM)

4.1.10.1.2 Planned EO related changes for Trial 2

- What is the lesson learned from Trial 1?

The pilot has been proven successful as Big Data enabling technologies provide the means for improving the procedures for damage assessment (eliminating the need for human expert evaluations and field visits). EO-based components have been deployed within this pilot and attempted to identify the correlations (even hidden within the data) that formulate between different perils and crop types. This task has been proven particularly challenging and requires ground truth and expert knowledge for data validation in order to effectively “translate” the findings.

The following aspects have been taken into consideration in Trial 2:

1. Certain data management capabilities are provided by C04.02 GeoRocket. This is not an EO-specific component but provides geospatial functionalities. For this reason, the status of such component, previously described in the D5.2 – EO component and Interfaces [REF-05], was detailed in the internal deliverable D4.i3 – Description of Platform for Trial 2. For pilot C1.1 further development of the pilot's integration mechanisms is foreseen. Bidirectional information exchange mechanisms between the EO component providers (NEUROPUBLIC, using C13.02 and CSEM, using C31.01) will be explored.
2. For the A1.1, B1.2 and C1.1 pilots, certain aggregation functionalities are part of the component C04.03 GeoToolbox. C04.03 is not an EO-specific component but provides geospatial functionality. For this reason, such component, previously described in the

D5.2 – EO component and Interfaces [REF-05] was detailed in the internal deliverable D4.i3 – Description of Platform for Trial 2.

3. For pilot C1.1 CSEM's data experimentation, for the development of a deep neural network crop classification service (component C31.01), didn't lead to adequate results. The preliminary study in peaches showed that static (in the temporal domain) models are not effective since they don't capture the temporal phenological dependencies that differentiate various crop types. Thereby, the milestone MS1 "Service ready for Pilot 1" wasn't reached in M16 as expected in the DoA for that particular component. CSEM has taken all the necessary corrective measures to speed up the data experimentation process focusing on creating multi-temporal crop models of annual crops (e.g. wheat) that present a more uniform planting continuity (spatially). Preliminary results after the corrective measures have been taken were encouraging. During Trial 2 the temporal models were capable of producing accurate results.
4. C04.03: Support for integration of external services.
5. C04.04: Integrate machine-learning services for improved parcel assessment.
6. C31.01: Temporal crop analysis using ML and implementation of REST API.

4.1.10.2 Updates on the work done on Trial 2

Changes for Trial 2:

- C13.01: Additional pilot UIs were explored based on end-user needs.
- C13.02: Crop type - area tailored models have been created that exploit specific vegetation indices that have proven to be suitable for identifying plant health and yield estimation (NDVI). A dedicated analysis of how different crop types are affected by different weather perils has been explored by NEUROPUBLIC (hail damages in cotton, flooding scenarios, etc.) in close collaboration with INTERAMERICAN. Using statistical tools, the component is able to identify outliers in terms of vegetation health and mark them across multiple time instances so as to determine disaster levels and prioritize the work of the damage evaluators.
- C13.03: Improved data representation and handling mechanisms, enabling the expansion and/or customized configuration of each GAIAtron station. As the requirements in terms of sensors deployed for in-the-field usage differ between pilot sites, it became obvious that several adaptations were necessary in respect to C13.03 and the way data was represented for both cloud-based storing and Gaiatron station configuration. More specifically, all relational and EAV (Entity-Attribute-Value) data representations were adapted to a more flexible and scalable JSON format that performs better in a dynamic IoT measuring environment. JSON has become gradually the standard format for collecting and storing semi-structured datasets that originate from IoT devices. The adaptation to a JSON format for modeling IoT data streams allows the further processing, parsing, integration and sharing of data collections in

support of system interoperability through the adaptation on well-established and favoured linked-data approaches (JSON-LD),

- C04.02 - C04.04: For pilots A1.1, B1.2, C1.1 and C2.2 a new web-based visual analytics agTech platform was designed that allows the integration of services on top of it for interactive exploration of heterogeneous data (including satellite imagery), AI products, aggregates and statistics in a user-friendly way

4.1.11 Pilot 11 [C1.2] Farm Weather Insurance Assessment

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

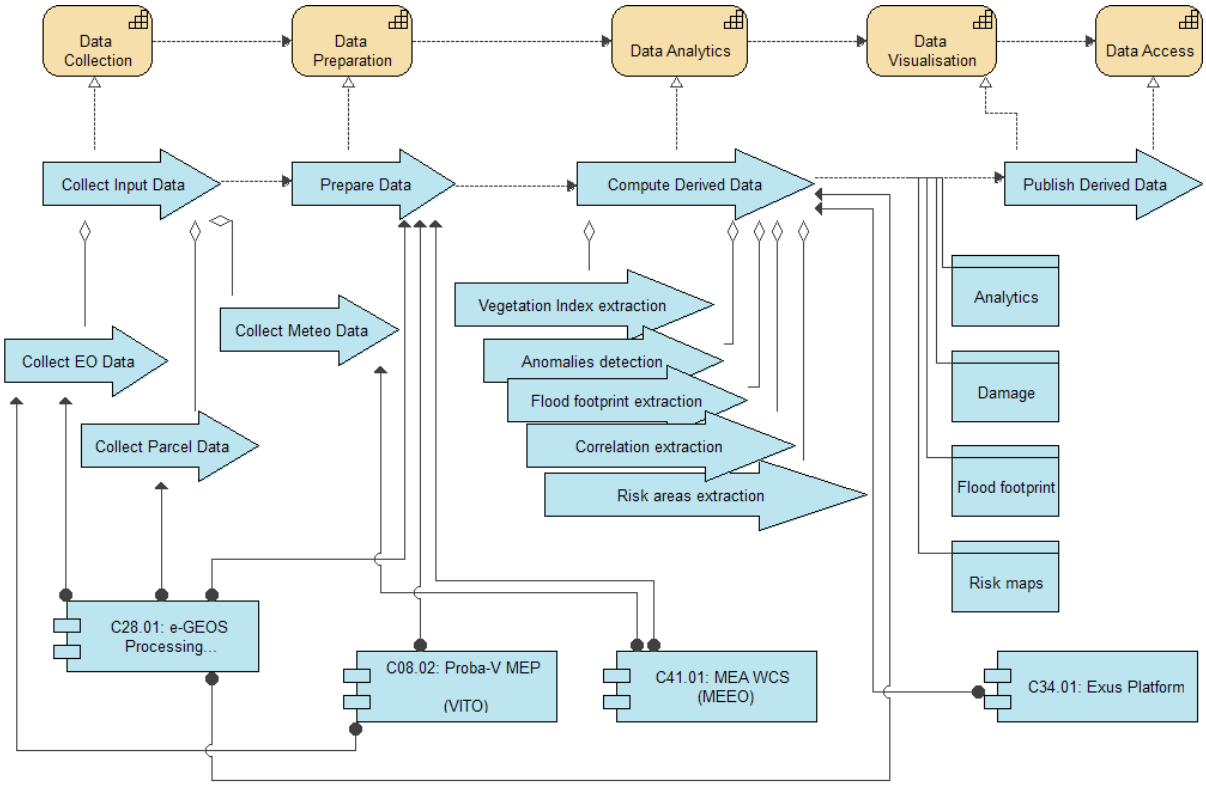


Figure 70: Pilot 11 [C1.2] Farm Weather Insurance Assessment pipelines

The **EO related results** of the pilot are four synthetic maps:

- Historical medium resolution Risk Map
- Field crop growth vs. similar crop (inter-field anomalies)
- Intra-field Anomalies
- Correlation among weather historical data and critical events

This pilot uses following components for Trial 1:

- C28.01: e-GEOS data processing
- C08.02 - Proba-V MEP
- C41.01 - MEA.WCS (meteo data WCS server)

- C41.02 - MEA.GUI (meteo data WCS interface)

This pilot uses following additional components for Trial 2:

- C34.01: EXUS Analytics Framework (EAF)
- C03.01 WebGLayer for web visualization
- C13.01: Additional pilot UIs were explored based on end-user needs

4.1.11.1 Brief summary of results and pending issues from Trial 1

Relevant outputs have been produced during the Trial 1. In particular, the following outputs have been produced based on the analysis of parcel data and past damages occurred in Netherland and related to potatoes fields:

- Historical medium resolution Risk Map: The scope of the service is to provide historical risk maps, based on long-time series of vegetation indices estimated from medium resolution satellite images providing, as output, risk maps per crop (number of critical events for each area). The historical risk map refers to the occurrence of "damage" in the past. The map is based on an index derived from time series of low-medium resolution satellite images. The index is assumed to be correlated with crop yield.
- Field crop growth vs. similar crop (inter-field analysis): Indicator on crop behaviour (average, worst, better) during the current season. Starting from S2 multi-temporal series and from in situ data provided by the Insurance (field boundaries and crop identification for each parcel) vegetation indexes behaviour will be analysed (such as NDVI, EVI e SAVI that could also be related to the specific crop), providing as output :
- Thematic map of parcel exceeding the average index value for that crop
- Graphical Plot of comparison between the single parcel and the average

4.1.11.1.1 Earth Observation aspect

This pilot uses Sentinel-2 **satellite data**, fAPAR, and Meteo climate data and the following **Earth Observation components**:

- C28.01 - e-GEOS data processing
- C08.02 - Proba-V MEP
- C41.01 - MEA.WCS (meteo data WCS server)
- C41.02 - MEA.GUI (meteo data WCS interface)

4.1.11.1.2 EO related changes for Trial 2

- What is the lesson learned from Trial 1?

Important outputs have been produced during Trial 1. The algorithms have been set up and the services described above implemented. One of the main issues for the trial has been data availability. At the beginning of the project it was planned to deliver the aggregated information coming from Meteo sources (MEEO), EO sources (e-GEOS) and from external sources (NBAdvice) to the EXUS analytics platform, which should be in charge of the final analysis using neural networks and presentation of the final information. VITO was

responsible for providing the required datasets to feed and execute the risk model by providing access to the PROBA-V archive. To implement this kind of agriculture dedicated advanced technology (machine learning), it is, however, necessary to provide extremely precise data as input. In the context of this pilot the involved insurance did not provide such data.

- Is there a need for a redesign for Trial 2?

For the above-described reasons, the pilot activities have been redesigned in order to include more datasets and to extract a predictive tool for insurance. In particular, the Trial 2 includes also the usage of Sentinel 1 SAR data. The extremely frequent cloud conditions that affect the pilot region did not allow to have a sufficient set of data to build reliable NDVI trends and to find out valuable correlation between NDVI and weather data by using a machine learning approach.

4.1.11.2 Updates on the work done on Trial 2

In the Trial 2 the pilot was realized introducing a machine learning approach and analysing the correlation among weather data, optical and SAR satellite data and other field parameters in order to support the insurance activities. In particular, the purpose of the Trial 2 was to support:

1. Risk analysis: Create a risk map to determine on a field level the impact of heavy rain on the NDVI value, in three (or more, depending on results of machine learning) categories like heavy risk, medium risk and low risk.
2. Predict on a field basis, during the current season, the impact of a heavy rain event on the crop.

The approach included:

- Clustering the NDVI profiles of each field to classify them according to its risk category. This analysis is performed on all metrics (meteo data, sat indexes) available for the entire potato season (from April to October)
- Predicting the impact / damage during the growing season. This analysis was performed on the metrics (meteo data, sat indexes) available up to the analysis request

4.1.12 Pilot 12 [C2.1] CAP Support

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

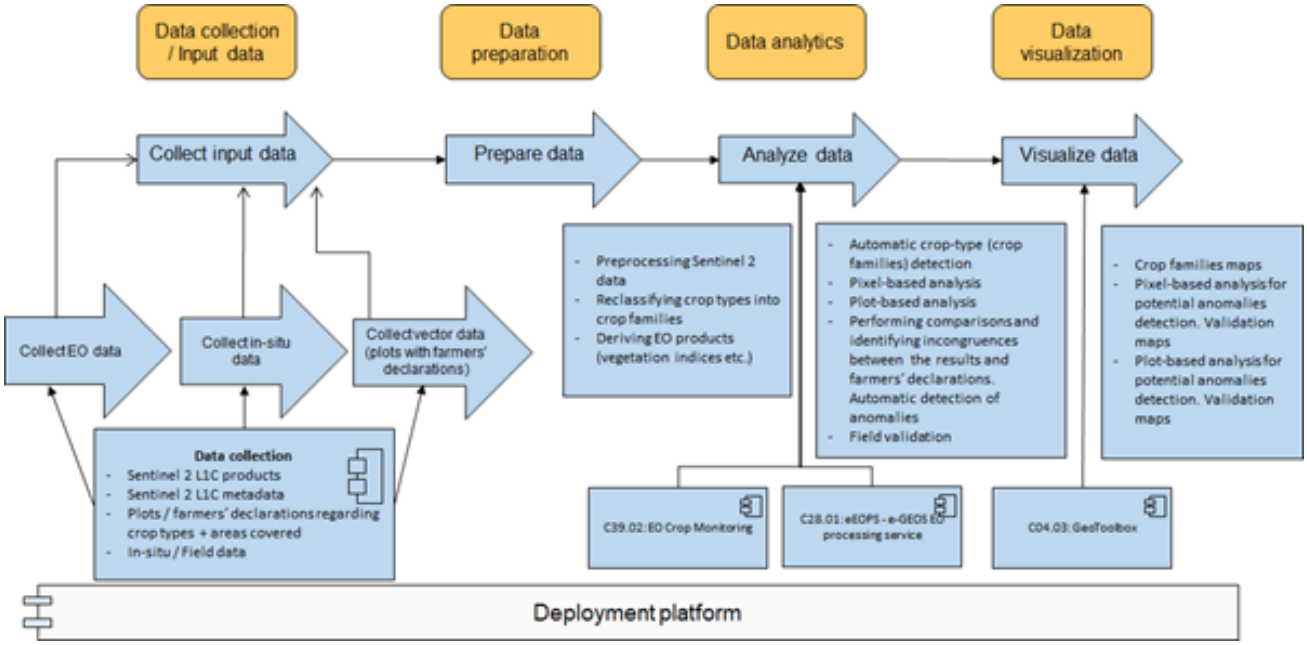


Figure 71: Pilot 12 [C2.1] CAP Support pipelines

During Trial 1, the following actions were performed:

- EO data like Sentinel-2, Landsat 8 and Sentinel-1 have been collected (in the Romanian pilot, access to data has been made available through FedEO Gateway - C07.01 component) and pre-processed (some atmospheric and terrain corrections have been performed) for the chosen AOI;
- crop types have been reclassified into crop families;
- different EO products (e.g. vegetation indices, band combinations) have been derived.
- a pixel-based and plot-based analysis over the chosen AOI has been carried out, followed by the detection of the anomalies / incongruences between the results and farmers' declarations (data linking);
- access for stakeholders has been established via dedicated GeoServer.

The pilot uses the following components:

- C07.01 - FedEO Gateway
- C07.03 - FedEO Catalog
- C07.04 - Data Manager
- C28.01 - e-GEOS Data Processing
- C39.01 - Mosaic Cloud Free Background - TerraS service
- C39.02 - EO Crop Monitoring - TerraS service
- C39.03 - Sentinel2 Clouds, Shadows and Snow Mask tool - TerraS service

4.1.12.1 Brief summary of results and pending issues from Trial 1

The objective of pilot C2.1 – CAP Support was to provide services in support to the National and Local Paying Agencies and the authorized collection offices for more accurate and

complete control of the farmers' declaration related to the obligation introduced by the current Common Agriculture Policy. The provision of products and services was based on specialized highly automated Big Data processing techniques, relying on multi-temporal series of free and open EO data, with a focus on Copernicus Sentinel-2 data.

During Trial 1, developed in 2018, important sets of results have been provided, consisting of crop families' maps and crop inadvertency maps. The results were based on farmers' declarations regarding crop types and areas covered for 2017 and 2018 agricultural seasons and involved five crop families: wheatlike cereals, maizelike cereals, sunflower and related crops, rapeseed and related crops, grassland, pastures and meadows. The results delivered for the 2018 agricultural season have been validated through a series of in-situ data and Sentinel-2 backgrounds for the test-area, resulting in a qualitative assessment, used in order to define Trial 2 actions and expected results.

Trial 2 was planned to overview the key results for Trial 1, identify the emerging needs of the components that were involved in Trial 1 and provide a further development of the Crop Monitoring Service, with products tuned in order to fulfil the requirements of the 2015-20 EU Common Agricultural Policy, based on an adjusted version of the crop detection algorithms, updated according to the previous validation work, aiming to increase accuracy and success rate.

Apart from the optimised version of the algorithms used, Trial 2 objectives included:

- New measurements to be carried out for the analysed area, according to a field tracking plan;
- The delivery of new results for the same area of interest, but based on a higher number of target crops and using crop types instead of crop families;
- Further testing of the algorithms developed.

4.1.12.1.1 Earth Observation aspect

Pilot C2.1 CAP Support uses a **multi-temporal series of Earth Observation data**, consisting of Sentinel-2 and Landsat-8 imagery. Earth Observation (EO) data provide broad and repetitive homogeneous coverage, translated into an unprecedented amount of information generally referred as "Big Data". The technologies benefitting from the data volumes represent a solid solution for continuous monitoring of CAP compliance. The Sentinel-2 satellites, part of the EU Copernicus data stream, hold an enhanced revisiting time, thus delivering regular coverage over large areas and allowing a uniform observation of the agricultural plots. The superior spectral resolution allows the identification of the phenological growth stages and the distinction between various crop types or classes. The free and open availability of Earth Observation data is bringing land monitoring to a completely new level, offering a wide range of opportunities, particularly suited for agricultural purposes, from local to a regional and global scale, in order to enhance the implementation of the Common Agricultural Policy (CAP).

All the **components** used within the pilot (C07.01 - FedEO Gateway, C07.03 - FedEO Catalog, C07.04 - Data Manager, C28.01 - e-GEOS Data Processing, C39.01 - Mosaic Cloud Free Background - TerraS service, C39.02 - EO Crop Monitoring - TerraS service, C39.03 - Sentinel2 Clouds, Shadows and Snow Mask tool - TerraS service) are **EO-related**.

All the **products** developed by Terrasigna under the framework of pilot C2.1 CAP Support are also **EO-related** and consist of:

- Maps of the main types of crops, for an annual agricultural cycle, completed;
- Intermediate maps with the main types of crops, during an annual agricultural cycle (they may serve as early alarms for non-observance of the declared crop type);
- Early discrimination maps between winter and summer crops;
- Layers of additional information, with the degree of confidence for the crop type, maps delivered;
- Maps of the mismatches between the crop type declared by the farmer and the one observed by the application;
- NDVI maps nationwide for a period of time, uncontaminated by clouds and cloud shadows;
- Lists of parcels with problems, in order of the surfaces affected by inconsistencies;
- National maps with RGB aspect mediated for a period of time, uncontaminated by clouds and shadows, obtained through the use of components C39.01 - Mosaic Cloud Free Background Service and C39.03 - S2 Clouds, Shadows and Snow Mask Tool.

4.1.12.1.2 Planned EO related changes for Trial 2

- What is the lesson learned from Trial 1?

The most serious problems that had to be solved and that served as lessons were:

1. the use of data Sentinel-2 and Landsat-8 together - which have a different format and resolution; One possibility, however not used in DataBio, would have been to use the Harmonized Landsat Sentinel product https://hls.gsfc.nasa.gov;
2. correction of the geographical positioning (georeferencing) automatically - which deeply affects the quality of the classification for small or narrow plots;
3. selecting the areas of interest from each image - which are not, as it might seem, the areas uncontaminated by clouds and shadows, but the areas where there is vegetal "activity";
4. the construction of an algorithm that takes into account the matrix of semantic confusion between cultures - which required finding the natural classes of cultures that can be followed simultaneously, without serious mutual confusion.

Geospatial services together with Copernicus data can provide a really powerful tool for monitoring agricultural dynamics. The end-users, the National Paying Agencies, are able to benefit from the modern and effective near real-time service, based on the principles of sustainable agriculture and saving effort both in terms of costs and time. A continuous

agricultural monitoring service based on the processing and analysis of Copernicus satellite imagery time series is not just a CAP compliance tool, but can also offer a great range of supplementary information for both public authorities and citizens.

Regarding the Romanian pilot, the highly-automated fuzzy-based proposed approach developed by Terrasigna allowed the performing of Big Data analytics to various crop indicators, being reliable, cost- and time-saving and allowing a more complete and efficient management of EU subsidies, strongly enhancing their procedure for combating non-compliant behaviours. The developed technique is replicable at any scale level and can be implemented for any other area of interest.

Regarding the Italian pilot, in the Trial 1 the service has been set up by using 2016 Sentinel 2A data. The availability of just one satellite data affected the time series frequency (gap in temporal trends due to cloud conditions) and then negatively impacted the NDVI profile reconstruction. In the Trial 2 the service will be updated including 2017 and 2018 Sentinel-2-time series. In this case the availability of satellite constellation is expected to improve the density and quality of NDVI trends improving as consequence also the service performances.

- *Is there a need for a redesign for Trial 2?*

Trial 2 is considered to be a further development and not necessarily a redesign of the Crop Monitoring Service, using the 2018 and 2019 farmers' CAP declarations regarding crop types and areas covered. During Trial 2 it was foreseen to further evaluate the EO Crop Monitoring (C39.02 component) for its maturity and ability in supporting the needs of the pilot in terms of batch processing functionalities.

Specific EO related changes for Trial 2 were:

- a further development of the pilot's integration mechanisms, as well as the extension of the developed solutions to other geographical areas, both for Romanian and Italian case studies;
- a further evaluation of the EO crop monitoring components (both C39.02 and C28.01) for their maturity and ability in supporting the needs of the pilots in terms of batch processing functionalities;
- implementation of data visualization capabilities through WMS services;
- further development of the crop monitoring services, using the 2018 and 2019 farmers' CAP declarations regarding crop types and areas covered are foreseen (when possible due to GSAA data availability). Planned future update included: new farm profile data ingestion and new Sentinel-2 and Landsat-8 imagery processing as soon as they were acquired during the 2019 growing season; new trials, based on different test areas and increasing the number of target crops;
- further optimization of the algorithms based on Trial 1 results. Fine tuning and further comparisons to the results obtained in Trial 1, in order to prepare the final adjustments for the service. Final optimization of algorithm settings. Final service trials;

- further comparisons regarding the performance of the developed solution.
- preparation for integration of the service into the pipeline for the agriculture pilot's AOI.

4.1.12.2 Updates on the work done on Trial 2

Changes for Trial 2:

For the Italian pilot, starting from the results of Trial 1, Trial 2 was focused on the methodology refinement in terms of crop type macro classification and marker's rules definition, to improve the accuracy of the products.

Trial 2 execution for the Romanian area of interest was entirely based on Terrasigna's toolbox for crop determination, consisting of a set of in-house developed algorithms for calculating CAP support-related products. It started with fine tuning of algorithms. Various parameters have been tested in order to implement a final version of the crop-type detection algorithm. The algorithms have been modified in order to increase the number of analysed crop types and the overall accuracy of the results. Following an automatic learning process, the system became capable of recognizing several types of cultures, of the order of several tens.

For the 2018 agricultural year, Terrasigna extended its service and monitored the CAP declarations for the entire agricultural area of Romania. The total surveyed area exceeded 9 million ha, corresponding to more than 6 million plots of various sizes and shapes. 21% of the total number of plots within the test areas had surfaces below 1 ha. The technology developed by TERRASIGNA was able to recognize a large number of crops families, of the order of tens. For Romania, it addressed the first most cultivated 32 crops families, which together covered more than 97% of the agricultural land. In 2018, the validation of results for a full agricultural season (full phonological cycle) against independent sources revealed promising results, with an accuracy higher than 95% for more than 10 crop types. The performance is quite uniform reported to parcels size and remains high even for parcels smaller than 1 ha.

Finally, during Trial 2, there had also been established a method to access and deliver the results to the stakeholders via WMS services, specifically tailored according to the needs of the end-user for the test area in Romania (the Romanian National Paying Agency).

4.1.13 Pilot 13 [C2.2] CAP support (Greece)

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

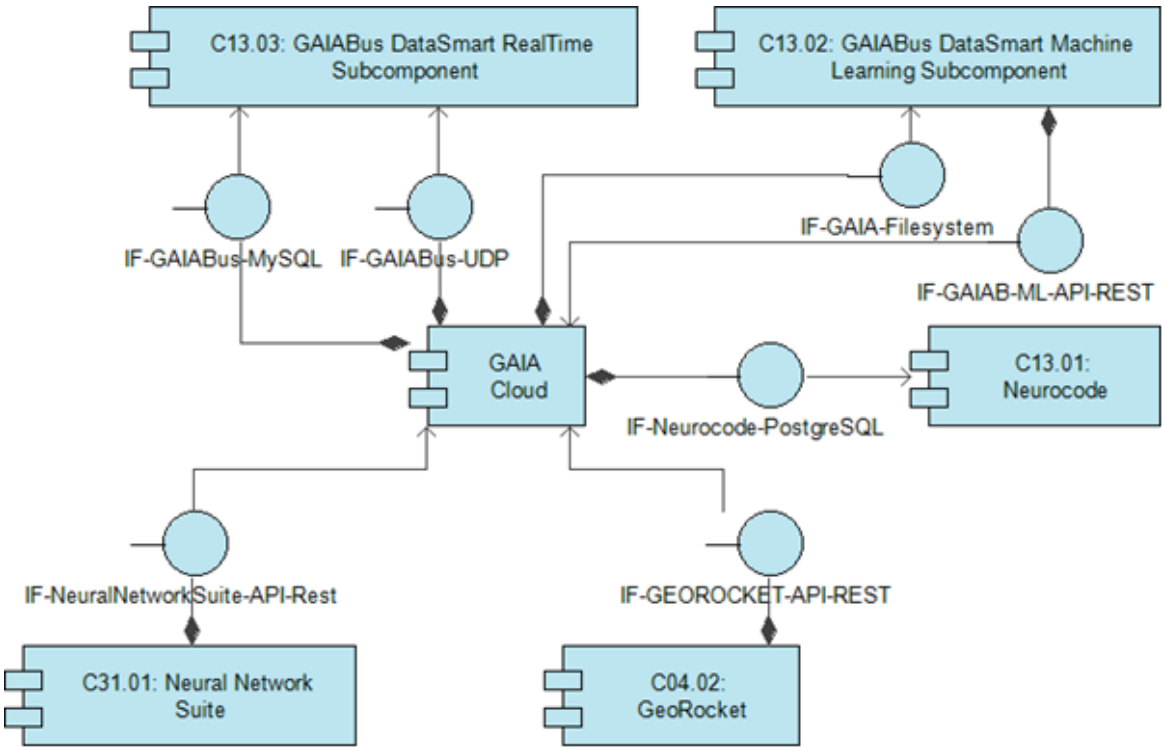


Figure 72: Pilot 13 [C2.2] CAP support (Greece) pipelines

This pilot uses following technological components:

- C04.02 – GeoRocket (Fraunhofer)
- C04.03 – GeoToolbox (Fraunhofer)
- C04.04 - SmartVis3D (Fraunhofer)
- C13.01- NeuroCode (NP)
- C13.02 - GAIABus DataSmart Machine Learning Subcomponent
- C13.03 - GAIABus DataSmart RealTime Subcomponent
- C31.01 - Neural network suite for image processing (CSEM)

4.1.13.1 Brief summary of results and pending issues from Trial 1

The pilot completed the first round of trials during Trial 1 in the greater area of Thessaloniki, Greece. It effectively demonstrated how Big Data enabled technologies and EO-based services can support specific needs of the CAP value chain stakeholders and more specifically the systematic and more automatic assessment of eligibility conditions for "greening" aid declarations. All these assumptions were validated through a set of pilot KPIs which in their majority met (and in some cases even exceeded) the targeted expectations. During Trial 1, NDVI was used as the main source of information to classify parcels in crop types.

In Trial 2, the applied technologies and pipelines got even more mature and reached their expected TRL by the end of the project. The key pilot stakeholders (i.e. the farmers and GAIA that had a supporting role in the crop type declaration process), continued (for a second year) to benefit from the EO-based geospatial data analytics, thus, promoting the simplification and improving the effectiveness of CAP. KPIs were collected along the Trial 2 in order to validate

the pilot assumptions. Lessons-learnt from Trial 1 accommodated the rest of the pilot activities. In Trial 2, alternative EO-based datasets were considered. These datasets provided the means towards evolving the provided services.

4.1.13.1.1 Earth Observation aspect

As mentioned, this pilot uses Copernicus EO data and the following components focused on Earth Observation:

- C13.02 - GAIABus DataSmart Machine Learning Subcomponent (NP)
- GAIA Cloud (Pilot component of NP for Sentinel 2 data collection, correction and extraction of indices)
- C31.01 - Neural network suite for image processing (CSEM)

4.1.13.1.2 Planned EO related changes for Trial 2

- What is the lesson learned from Trial 1?

In 2016 data an average of only ~10% of the parcel declarations were identified as "problematic" and potentially incorrect (for crops like wheat, maize, legumes) based on the followed methodology. This highly encouraging validation performance in terms of accurately classifying crop types at parcel level and detecting outliers, led to the implementation of services aspiring to bring insights in new and unseen data referring to 2018 cultivating period and crop type declarations (acquired with the help of GAIA and the FSC of Thessaloniki that hosts the pilot activities and collected the declarations properly). The application of the crop models in pilot "testing" data of 2018 revealed limitations and a ~70% agreement among the predicted and declared crop type classes. By following a systematic and exhausting data screening parallel activity all this time, we identified that this performance is due to the fact that many crops exhibit inter-year changes in their cultivating period (begin, end, peak, length) originating from climate changes, regulatory and market conditions, regional characteristics etc., thus, making the classification problem in new and unseen data particularly challenging.

- Is there a need for a redesign for Trial 2?

Lessons-learnt from the Trial 1 period were valuable and considered a critical asset for delivering accurate results. They served as a new baseline for the pilot that led the pilot partners to take corrective measures for Trial 2. Thereby, new data, features and classification methodologies were examined taking into account the aforementioned inter-year changes in crop cultivation periods.

Moreover, in Trial 2, it was identified that the pilot would benefit from the exploration of the envisioned agTech platform designed by Fraunhofer using its tools C04:02-04 for pilots A1.1, B1.2 and C1.1. Thereby, these tools were also acknowledged as part of the pilot for Trial 2, as pilot requirements have shaped the respective development activities (especially for coping with handling crop classification and validation of the results).

Several EO aspects were taken into consideration:

1. Certain data management capabilities are provided by C04.02 GeoRocket. This is not an EO-specific component but provides geospatial functionalities. For this reason the status of such component, previously described in the D5.2 – EO component and Interfaces [REF-05], was detailed in the internal deliverable D4.i3 – Description of Platform for Trial 2. For pilot C1.1 further development of the pilot's integration mechanisms is foreseen. Bidirectional information exchange mechanisms between the EO component providers (NEUROPUBLIC, using C13.02 and CSEM, using C31.01) will be explored.
2. For the A1.1, B1.2 and C1.1 pilots, certain aggregation functionalities are part of the component C04.03 GeoToolbox. C04.03 is not an EO-specific component but provides geospatial functionality. For this reason, such component, previously described in the D5.2 – EO component and Interfaces [REF-05] was more detailed in the the internal deliverable D4.i3 – Description of Platform for Trial 2.
3. Data experimentation for developing more accurate crop models using deep neural network (component C31.01) will be extended by CSEM. The new models will be capable of determining if a parcel belongs to a particular crop with a pixel resolution. Further to that, crop health can be assessed for a known crop variety. More features and crop types will be examined for offering better results. Temporal crop analysis using ML and implementation of REST API.
4. C13.02: New data, features and classification methodologies will be examined and used taking into account inter-year changes in crop cultivation periods. Integration of agronomic knowledge into the methodological component framework.

4.1.13.2 Updates on the work done on Trial 2 and main results

Changes for Trial 2:

- C13.01: Additional pilot UIs were explored based on end-user needs.
- C13.02: ML methodologies using EO-products and bands for crop type classification. Specific focus on challenging datasets and performance in new and unseen data of different cultivating periods. Multi-year data have been exploited in order to capture underlying/hidden relations among crop types, phenological stages, inter-year climatological changes and their depiction through different EO products and assigned vegetation indices. A significant improvement in accurately identifying crop types has been witnessed supporting the “greening” aid use case and the implementation of the CAP for farmers in the greater Thessaloniki, Greece area.
- C13.03: Improved data representation and handling mechanisms, enabling the expansion and/or customized configuration of each GAIAtron station. As the requirements in terms of sensors deployed for in-the-field usage differ between pilot sites, it became obvious that several adaptations were necessary in respect to C13.03 and the way data was represented for both cloud-based storing and Gaiatron station configuration. More specifically, all relational and EAV (Entity-Attribute-Value) data

representations were adapted to the more flexible and scalable JSON format that performs better in a dynamic IoT measuring environment. JSON has become gradually the standard format for collecting and storing semi-structured datasets that originate from IoT devices. The adaptation to a JSON format for modeling IoT data streams allows the further processing, parsing, integration and sharing of data collections in support of system interoperability through the adaptation on well-established and favoured linked-data approaches (JSON-LD),

- C04.02 - C04.04: For pilots A1.1, B1.2, C1.1 and C2.2 a new web-based visual analytics agTech platform was designed that allows the integration of services on top of it for interactive exploration of heterogeneous data (including satellite imagery), AI products, aggregates and statistics in a user-friendly way. Support for the integration of external services (including ML) for improved parcel assessment.

4.2 WP2 - Forestry

4.2.1 Pilot 2.2.1: Easy data sharing and networking

This pilot uses the following technological components:

- C18.01 - Metsään.fi
- C20.01 - Wuudis

4.2.1.1 Brief summary of results and pending issues from Trial 1

This pilot has been completed and all the required features on Wuudis are developed.

4.2.1.2 Updates on the work done on Trial 2

Changes for Trial 2:

- Use of Wuudis easy data sharing and networking features by customers of MHGS, METSAK personnel and associated stakeholders. Feedback was gathered from the users of the service.
- C18.01: The data gathered via Wuudis application was utilized in updating METSAK forest resource data.
- C18.01: Metsään.fi user authentication via Suomi.fi national service architecture portal for forest owners.
- C18.01: The forest resource data system (standardization, data transfer service, database) and processes was adjusted accordingly.
- C18.01: The forest resource data system (standardization, data transfer service, database) and processes was adjusted accordingly.
- C18.01: Single-login and User role-based authentication process integrated to Suomi.fi e-authorization process for forest owners, was implemented in Metsään.fi 2.0 version in 2020.
- C20.01: Use of the new version of Wuudis service

- C20.01: Developed standard WMS and WFS for tree-wise monitoring service with VTT (Forestry TEP), Spacebel and FMI
- C20.01: Some interfaces were adjust based on user feedback.
- C20.01: Tree-wise WMS/WFS data was integrated into Wuudis.
- C20.01: In Finnish forest standard format (xml).
- C20.01: Integration of different dataset and service with Wuudis.

4.2.2 Pilot 2.2.2: Monitoring and control tools for forest owners

This pilot uses the following technological components:

- C18.01 - Metsään.fi
- C18.02 - Open Forest Data
- C20.01 - Wuudis

4.2.2.1 Brief summary of results and pending issues from Trial 1

Wuudis launched a work quality monitoring app (Laatumetsä in Finnish) during November 2018 to enhance better work quality monitoring while processing subsidy applications. Currently, this app is under use by METSAK personnel, forestry service providers and forest owners. Feedback from users collected in Q1-Q2/2019.

Forest damage (such as storms, snow, pests and diseases) monitoring through standardized procedures were developed together with METSAK, as well as easy-to-use mobile tools for these damage monitoring needs and non-wood product monitoring needs. Finally, the data will be integrated with METSAK´s Metsaan.fi eService. This allows forest owners and forest specialists willing to monitor and report forest damage information to authorities a direct access to Metsaan.fi's master database.

Basically, the forest damage crowdsourcing app is a feature of Laatumetsä (work quality monitoring) app. Most likely during winter 2018-2019 data from moose damages in young pine stands will be recorded by private forest owners and contractors through the Laatumetsä app. This helps to understand and control better moose populations and thus prevent damages and growth losses.

4.2.2.2 Updates on the work done on Trial 2

Changes for Trial 2:

- C18.01: The data gathered via Wuudis application was utilized in updating METSAK forest resource data.
- C18.01: The quality control data was used for updating the Metsään.fi forest resource data.
- C18.01: Metsään.fi user authentication via Suomi.fi national service architecture portal for forest owners.
- C18.01: The forest resource data system (standardization, data transfer service, database) and processes was adjusted accordingly.

- C18.01: The forest resource data system (standardization, data transfer service, database) and processes was adjusted accordingly.
- C18.01: Single-login and User role-based authentication process integrated to Suomi.fi e-authorization process for forest owners, was implemented in Metsään.fi 2.0 version in 2020.
- C18.02: The crowdsourced information is received in METSAK via the new data transfer service and in standardized format.
- C18.02: Data for the forest resource data sample plots gathered as field data will be published in Open forest data service.
- C18.02: Standardized forest damage message and accordingly updated data transfer interface.
- C18.02: New service for the Open forest data service was implemented as well as new dataset for forest resource database.
- C20.01: Use of the new version of Wuudis service
- C20.01: Maintenance of Laatumetsä app
- C20.01: Analyse the damage data (need to discuss internally with METSAK). Trying (negotiation in place) to integrate and customize the same service for LUKE.
- C20.01: Some interfaces adjusted based on user feedback.
- C20.01: Updates of Laatumetsä app.
- C20.01: Standardized forest damage message and accordingly updated data transfer interface. More interface development as per requirement of METSAK.
- C20.01: Tree-wise WMS/WFS data integrated into Wuudis.
- C20.01: In Finnish forest standard format (xml).
- C20.01: Integration of different dataset and service with Wuudis.

4.2.3 Pilot 2.3.1: Forest Damage Remote Sensing

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

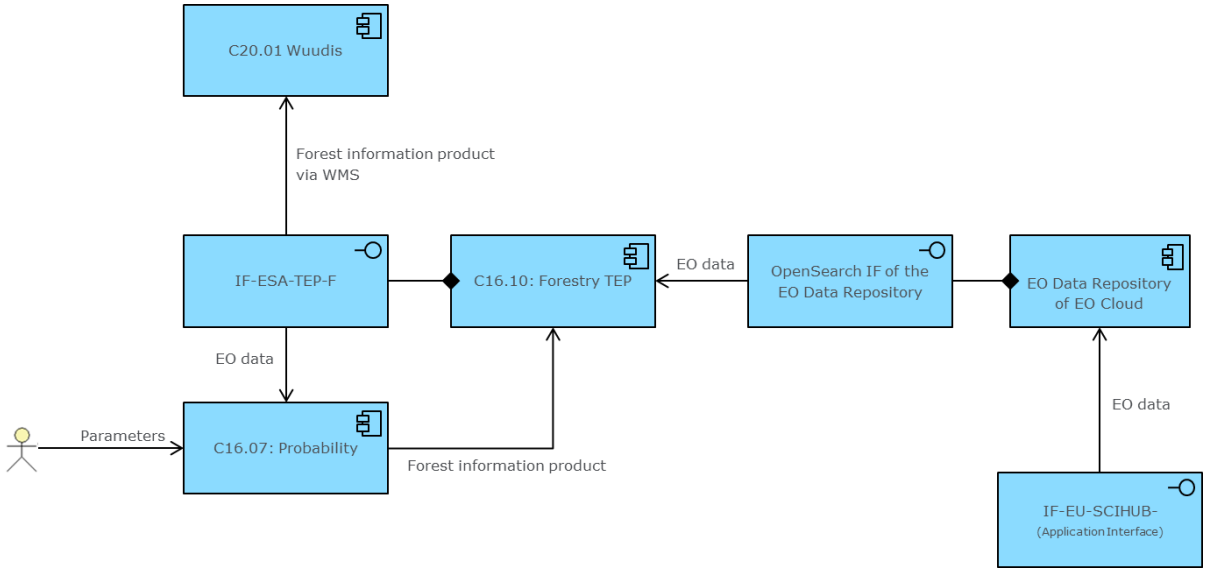


Figure 73: Pilot 2.3.1: Forest Damage Remote Sensing pipelines

This pilot uses the following technological components:

- C16.07 – Probability
- C16.09 – Envimon
- C16.10 - Forestry TEP
- C20.01 – Wuudis
- C44.01 - Senop Hyperspectral Camera

4.2.3.1 Brief summary of results and pending issues from Trial 1

A successful first phase of the pilot has been implemented. With regard to the utilization of Sentinel-2 satellite data, a service chain for forest inventory estimation was established by VTT. The chain utilizes the Forestry TEP platform for data sourcing and processing, and the VTT software Envimon and Probability for processing on the platform. The generated forest variable estimates include: stem number; stem volumes for pine, spruce, broadleaved and total; diameter; basal area; and height. Sample plot data by the Finnish Forest Centre was used as a reference in the estimation model training.

For easy integration of satellite maps and analysed (highlighted) theme maps, standard OGC WMS or WMTS interfaces have been used as a starting point. The Wuudis Service is using OpenLayers 3/4 as the mapping client library. In Trial 1, the forest variable estimates produced by VTT were presented as image raster data (GeoTIFF format) with 10 m pixel resolution, with one image band per variable and each pixel containing the estimated variable value. The output was made available for integration in the Wuudis end-user system via WMS interface from the Forestry TEP.

For the Trial 2 development, Wuudis needs all the data provided in a ready to use format (for the end-users), so no calculations need to be done in the Wuudis platform. Integration activities have been performed with VTT regarding Forestry TEP and with SPACEBEL concerning SF3 and SF4 service and Wuudis. The forest variable estimates can now be

provided from Forestry TEP also in the XML based Finnish Forest Information Standard format, as well as in the GeoJSON format: for each forest stand, a selected measure of central tendency (such as arithmetic average) is produced per variable.

In addition to the aforementioned, the following are operational level components of this pilot:

- Wuudis tree-wise monitoring MVP (minimum viable product) service was launched in June 2018 and sold to leading forest management associations (Mhy Pohjois-Karjala, Savotta and Päijänne.) and forest industries in Finland. Over 4000 hectares were monitored during summer, 2018 by Wuudis network of service providers.
- New pilot (Oct. 2018) started with SENOP hyperspectral camera in Polvijärvi to identify Boron deficiency in Spruce strands. This pilot is sponsored by Yara.

4.2.3.1.1 Earth Observation aspect

Earth Observation (EO) data from multispectral optical aerial, unmanned aerial vehicles (UAV) and satellite sensors present the optimal way to timely collect information on land cover over areas of various sizes. Particularly the availability of the Copernicus Sentinel-2 data and the applicable free data policy presents an excellent opportunity for developing low-cost commercial applications of EO downstream services in the monitoring of the environment. Online platforms, such as the Forestry TEP, enable the creation of services for efficient processing of satellite data to value-added information.

In Trial 1, the forest variable estimates produced by VTT were presented as image raster data (GeoTIFF format) with 10 m pixel resolution, with one image band per variable and each pixel containing the estimated variable value. The output was made available for integration in the Wuudis end-user system via WMS interface from the Forestry TEP.

The pilot uses Earth Observation related components:

- C16.10: Forestry TEP (Forestry Thematic Exploitation Platform)
- C16.07: Probability
- C16.09: Envimon

4.2.3.1.2 Planned EO related changes for Trial 2

- What is the lesson learned from Trial 1?

Big Data methods bring the possibility to both increases the value of forests as well as to decrease the costs within sustainability limits set by natural growth and ecological aspects. The key technology is to gather more and more accurate information about the trees from a range of sensors including a new generation of satellites, UAV images, laser scanning, mobile devices through crowdsourcing and machines operating in the forests. This enables characterization of even single trees, not to mention the accurate monetary value of the forestry.

Specific EO changes for Trial 2:

1. C16.10: Use: Adaptation of the T2.3.1 processing chain for a new type of area (Galicia, Spain) and with new reference data.
2. C16.10: Support for converting raster format data to vector-based format; particularly, support the Finnish Forest Information Standard (XML) as an output data format, as well as GeoJSON. Enables polygon-based output formats for the (originally raster formatted) forest variable estimates, enabling direct mapping to forest stands.

4.2.3.2 Updates on the work done on Trial 2

Changes for Trial 2:

- Commercial scaling of tree-wise MVP service in Finland and Wallonia (together with Spacebel).
- Efforts expanded to Sierra Leone pilot (focus: illegal logging, change in vegetation etc.). Discussion initiated with partners (SPACEBEL and VTT). MHGS has developed a separate pilot description document in place for this purpose. As plan B is to expand to Galicia, Spain if Sierra Leonean initiative fails. MHGS has LoC (Letter of Commitment) in place with Galician forest management association as well as contractor companies.
- C16.07: Use: forest parameter estimation, on Forestry TEP (pilot area Galicia, Spain).
- C16.09: Use: satellite data pre-processing, on Forestry TEP (pilot area Galicia, Spain).
- C16.09: New approach for producing cloud-free composite data. (The implementation may be a new component).
- C16.10: Support for converting raster format data to vector-based format; particularly, support the Finnish Forest Information Standard (XML) as an output data format, as well as GeoJSON [REF-11]. Enables polygon-based output formats for the (originally raster formatted) forest variable estimates, enabling direct mapping to forest stands.
- C20.01: Use of the new version of Wuudis service.
- C20.01: Developed standard WMS and WFS for tree-wise monitoring service with VTT (Forestry TEP), Spacebel and FMI.
- C20.01: Analysing different camera images (multispectral, hyperspectral) for boron deficiency recognition.
- C20.01: New pilot was started in Galicia, Spain during Q1/2019. The focus is on further developing Forestry TEP and other forest health mapping service (utilizing high resolution satellite data) for pilot forest estate managed by forest management association ASEFOGA in Galicia, Spain. The presentation of all data and services done on Wuudis service. Reference data is in place through Wuudis partner network in Spain.
- C20.01: Some interfaces were adjusted based on user feedback.
- C20.01: Standardized forest damage message and accordingly updated data transfer interface. More interface development as per requirement of METSAK.

- C20.01: Tree-wise WMS/WFS data integrated into Wuudis.
- C20.01: In Finnish forest standard format (xml).
- C20.01: Integration of different dataset and service with Wuudis.

4.2.4 Pilot 2.3.2-FH: Monitoring of forest health

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

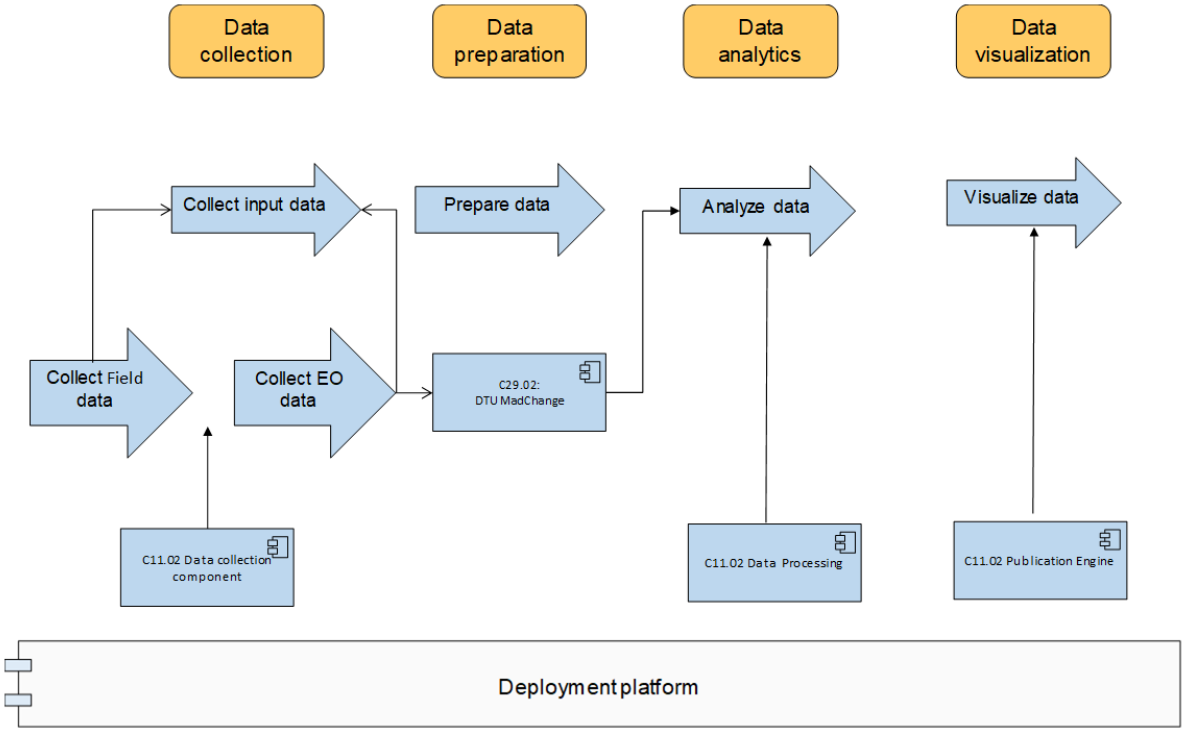


Figure 74: Pilot 2.3.2-FH: Monitoring of forest health pipelines

The goal of the pilot 2.3.2-FH iwas to set up a methodology based on remote sensing images (satellite + aerial + UAV) and field data for the monitoring of the health status of forests in large areas of the Iberian Peninsula. The work focused, particularly in the monitoring of the health of Quercus sp. forests affected by the fungus Phytophthora cinnamomi Rands and of the damage in Eucalyptus plantations affected by the coleoptera Gonipterus scutellatus Gyllenhal.

As seen in the pipeline, using collection components all the data were gathered. Then, all these data were pre-processed to fit the needs of the pilot. e.g. Sentinel L1C data wasprocessed to Level 2 with atmospheric corrections. The prepared data was then analysed using TRAGSA dedicated components. The processed results were delivered to the visualization components, in order to be displayed on a web interface over a base map layer

This pilot used the following technological components:

- C09.13 - PSNC HPC and cloud infrastructure

and the following focused on Earth Observation:

- C11.02 - Forest Health Status - TRAGSA Service
- C11.03 - Radiometric Corrections
- C28.01 - e-Geos Processing Environment
- C29.01 - WishartChange
- C29.02 - MADchange

4.2.4.1 Brief summary of results and pending issues from Trial 1

Despite technical efforts, it has been proved that Sentinel Resolution is not enough for the selected cases in Spain due to Eucalyptus not being not visible in Satellite imageries and Helm Oak density being very low. Consequently, EO techniques have to be complemented with the use of RPAS and UAVs. Therefore, the economical savings could not be so high as defined in the early stages of the project.

Consequently, Sentinel 2 data should be complemented with orthophotos due to the issue previously mentioned. TRAGSA worked in the Radiometric Correction of Aerial Orthophotos (Supported by C11.03 component) to allow aerial images to be used in the same way as Satellite Imageries.

4.2.4.1.1 Earth Observation aspect

For this pilot support, activities focusing on processing Earth Observation data and using it through online platforms were needed. The pilot uses **Earth Observation related components**:

- C11.02 - Forest Health Status - TRAGSA Service
- C11.03 - Radiometric Corrections
- C28.01 - e-Geos Processing Environment
- C29.01 - WishartChange C29.02 - MADchange

Despite technical efforts, it was proved that Sentinel Resolution is not enough for the selected cases in Spain due to the fact that Eucalyptus is not visible in Satellite imageries and Helm Oak density is very low. Consequently, EO techniques had to be complemented with the use of RPAS and UAVs. Therefore, the economical savings could not be so high as defined in the early stages of the project.

4.2.4.1.2 Planned EO related changes for Trial 2

In Trial 2, efforts focused on the development of a more general methodology for monitoring the health status of forest areas, based on spectral indices derived from satellite images (Sentinel 2 / Landsat 8). Auxiliary data about environmental conditions and management procedures were combined with EO-data in order to detect areas under stress and, therefore, more prone to be affected by plagues and diseases. List of technological changes for Trial 2 is provided below:

- Development of a more general methodology for monitoring the health status of forest areas, based on spectral indices derived from satellite images (Sentinel 2 / Landsat 8). Auxiliary data about environmental conditions and management procedures will be combined with EO-data in order to detect areas under stress and, therefore, more prone to be affected by plagues and diseases.
- Development of a more general methodology for monitoring the health status of forest areas, based on spectral indices derived from satellite images (Sentinel 2 / Landsat 8). Auxiliary data about environmental conditions and management procedures will be combined with EO-data in order to detect areas under stress and, therefore, more prone to be affected by plagues and diseases.

4.2.4.2 Updates on the work done on Trial 2 and main results

This is the list of technological changes for Trial 2:

- Quercus forests affected by Phytophthora.
    - Extrapolation of the algorithm obtained from the first campaign to the whole surface of the study site (Haza 'dehesa', 380ha).
    - A second campaign (RPAS flight and field work) took place in spring 2019, so as to evaluate the algorithm obtained from the first campaign.
    - Testing the methodology under different climatic conditions D2.2 – Forestry Pilots Intermediate Report H2020 Contract No. 732064 Final – v1.0, 28/12/2018 Dissemination level: PU -Public Page 41
    - Training the algorithms obtained from RPAS data with satellite-derived spectral variables.
    - The analysis of historic orthophotos (RGB and NIR), so as to analyse the evolution of the affection in the study site.

- Eucalyptus plantations affected by Gonipterus. Results obtained allow to state that it is possible to assess defoliation and assign treatment priorities by using RPAS remote sensing data. Improvements were made in the methodology so as to obtain an objective, operative and affordable service.
    - It was needed to improve the methodology for the automatic extraction of tree crowns and model extension (vertically and horizontally).
    - It was advisable to simplify the model, so as to employ data from a single RPAS-sensor if possible. This way, data acquisition and processing was faster and cheaper.

4.2.5 Pilot 2.3.2-IAS: Invasive Alien Species control and monitoring

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

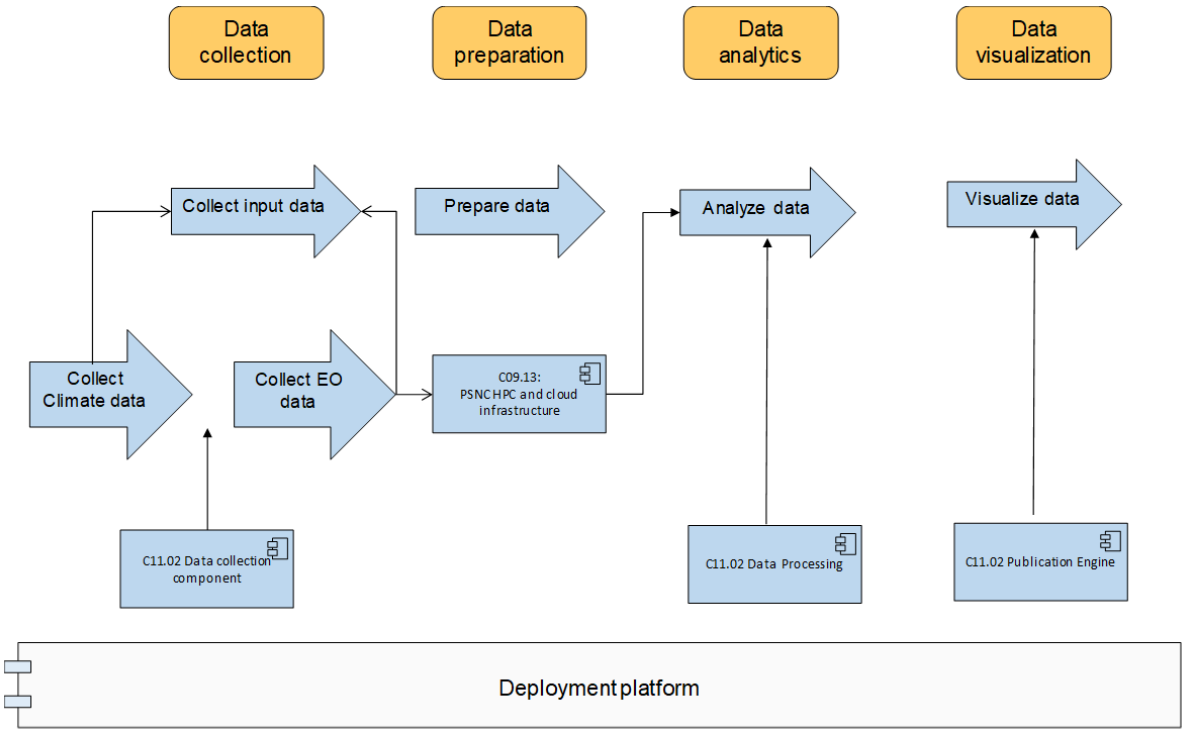


Figure 75: Pilot 2.3.2-IAS: Invasive Alien Species control and monitoring pipelines

As input of the pipeline, several alphanumeric Big Data databases and several TIFF files were used. All these data were pre-processed to fit the needs of the pilot. The processed results were delivered to the visualization components, in order to be displayed through a map viewer.

This pilot used the following technological component:

- C11.014 Forest Health Status - TRAGSA Service (part of C11.02)
- C09.13 - PSNC HPC and cloud infrastructure

4.2.5.1 Brief summary of results and pending issues from Trial 1

The conclusions obtained in Trial 1 were completely aligned with the most advanced results in this research area. Therefore, the algorithms and tools defined could be considered valid. Therefore, this pilot could be considered as especially successful due to the obtained results, external validation of their outcomes through scientific papers and the expressed interest by the Spanish Public Administration. Consequently, despite the development could be considered as completed, TRAGSA continued improving the algorithms and tools.

A first conclusion of the analyses performed indicated that the invasion risk derived from environmental and biogeographic dissimilarity clearly diverges from the current distribution of IAS richness in Spain. This suggested that the environment is not the main driver for the geographic configuration of biological invasions in Spain.

The estimation of IAS invasion risk in the Canary Islands was adjusted, taking into account not only the influence of the world over the islands, but also the invasion risk derived from the rest of Spain.

Models were tested including only climatic and biogeographic dissimilarity (DC+BG), and the first results of propagule pressure and biogeography (PP+BG). The second one (PP+BG) resembles the patterns of IAS richness detected in former works for Spain.

The complete model was developed, which integrates: propagule pressure, climatic dissimilarity, biogeography and ecosystem disturbance. First analyses showed that the complete model resembles the propagule pressure one, highlighting the importance of this factor within the model. This is due to the fact that the propagule pressure shows great differences among the different Spanish provinces, while differences in terms of climatic dissimilarity are much smaller.

First analyses of the complete model showed results in line with the current geographic distribution of richness of exotic vascular plants and birds. Coastal provinces are more threatened than the inland ones (except Madrid). Anthropic factors seem to play a key role in the determination of the geographical pattern of biological invasions

4.2.5.1.1 Earth Observation aspect

The Pilot uses the following EO related data:

- Sentinel-2: A time series of Sentinel-2 L1C images (both A and B satellites) are used to cover 2017-2018-2019 periods and several areas.
- LANDSAT: Despite the resolution being lower than the Sentinel mission, it has been TRAGSA-TRAGSATEC reference data for years. It will be used as reference, contrast and on those dates or areas with no Sentinel coverage.

4.2.6 Pilot 2.4.1: Web-mapping service for government decision making

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

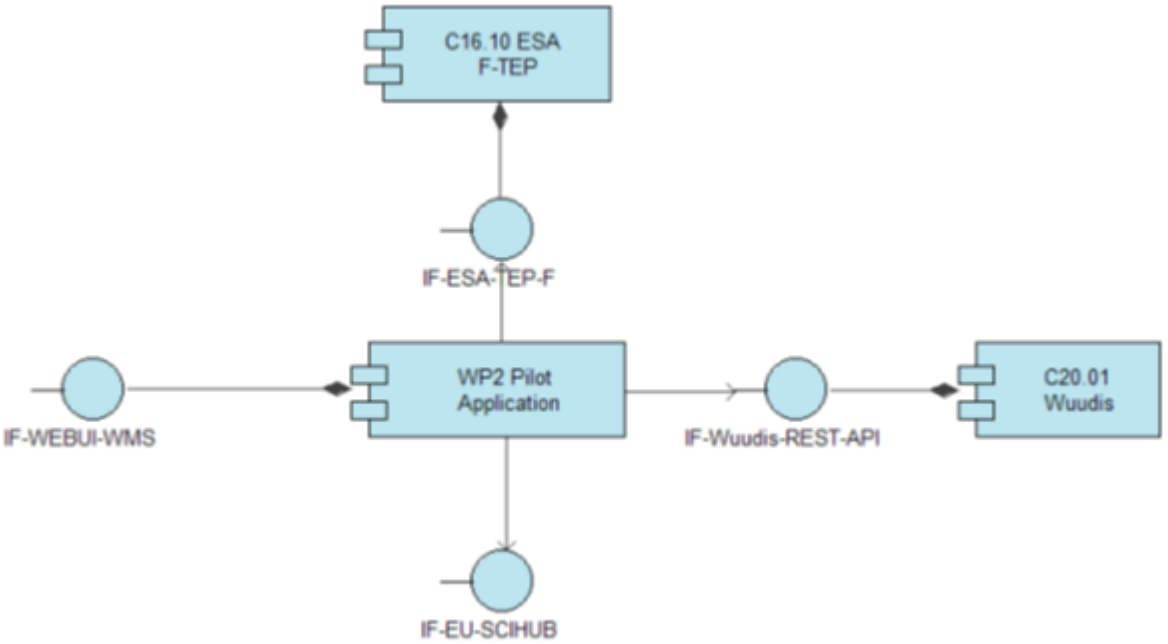


Figure 76: Pilot 2.4.1: Web-mapping service for government decision making pipelines

The goal of the pipeline is to provide a big-data based solution for Sentinel-2 satellite data pre-processing and publication. Raw Sentinel-2 satellite data needs to be systematically corrected for the atmospheric influence (atmospheric corrections) and provided as country-wise cloud-free images.

This pilot uses the following technological components:

- C16.10: Forestry TEP (Forestry Thematic Exploitation Platform)
- C20.01: Wuudis - shows both EO and no-EO features

4.2.6.1 Brief summary of results and pending issues from Trial 1

- Despite the frequent observations of Sentinel-2 satellites, there is relatively high cloud cover in the area of the Czech Republic, especially from October to March which de-facto prevents the use of passive optical remote sensing data for forest monitoring purposes. On the other hand, compared to other Earth observation systems like Landsat, the design of the pair of Sentinel-2 satellites allowed us for the first time to produce high-quality cloud-free images for vegetation growing period (from April to September), which is a basic prerequisite for any further satellite data interpretation.
- We observed a strong linear relationship between the in-situ measurements of leaf area index and the satellite image transformations utilizing shortwave infrared wavelengths (namely the Normalized difference infrared index and the Wetness component of Tasseled Cap transformation). There is therefore a strong potential of those bands for forest health monitoring.
- Forest health cannot be assessed from static remote sensing observations - different forest structures exhibit different reflectance values, which are not per-se related to

their forest health. What makes the remote sensing of forest health possible is the observation of the change in forest reflectances / image transformations sensitive to particular vegetation property (leaf area index in our case). Comparing the observations from the same phenological period is crucial.

4.2.6.1.1 Earth Observation aspect

A complex processing chain for satellite data pre-processing and interpretation towards forest health was developed and started its routine deployment at both the FMI's in-house computing infrastructure and the super computational facilities of IT4Innovations. Forest health trends are assessed from the analysis of time series of Sentinel-2 satellite data which require big-data approach. In principle, the country-wise forest health trends are obtained in two independent steps:

1. Sentinel-2 satellite data pre-processing and cloud-free mask synthesis
2. Retrieval of forest leaf area index (LAI) absolute values and its trends

The key to assessing the health status of forests from remote sensing data is the availability of high quality (i.e. cloud-free) mosaic generated from all-available Sentinel-2 data. This is a basic prerequisite for any remote sensing data interpretation. The developed methodology for forest health assessment under this pilot proposed a novel processing chain for automated cloud-free image synthesis based on the analysis of all available Sentinel–2 satellite data for selected sensing period (e.g. the vegetation season from June to August). The processing chain is implemented in three follow-up processes:

1. batch downloading,
2. atmospheric corrections of raw images (so-called L2 process) and
3. automated synthetic mosaic generation (so-called L3 process, or space-temporal image synthesis

The pilot uses **Earth Observation related components**:

- C16.10: Forestry TEP (Forestry Thematic Exploitation Platform)
- C20.01: Wuudis - shows both EO and no-EO features

4.2.6.1.2 Planned EO related changes for Trial 2

What is the lesson learned from Trial 1?

Despite the frequent observations of Sentinel-2 satellites, there is relatively high cloud cover in the area of Czech Republic, especially from October to March which de-facto prevents the use of passive optical remote sensing data for forest monitoring purposes. On the other hand, compared to other Earth observation systems like Landsat, the design of the pair of Sentinel-2 satellites allowed us for the first time to produce high-quality cloud-free images for vegetation growing period (from April to September), which is a basic prerequisite for any further satellite data interpretation.

The pilot has observed a strong linear relationship between the in-situ measurements of leaf area index and the satellite image transformations utilizing shortwave infrared wavelengths (namely the Normalized difference infrared index and the Wetness component of Tasseled Cap transformation). There is, therefore, strong potential of those bands for forest health monitoring.

4.2.6.2 Updates on the work done on Trial 2

EO changes for Trial 2:

Developed methodology for forest health assessment was implemented in a new pilot focusing on the forests of Galicia, Spain. There are several aspects related to its execution, mainly:

- C16.10: Use in T2.4.1 in hosting and executing FMI's processing services.

4.2.7 Pilot 2.4.2: Shared multiuser forest data environment

This pilot uses the following technological components:

- C18.01 - Metsään.fi
- C18.02 - Open Forest Data

4.2.7.1 Brief summary of results and pending issues from Trial 1

Regarding the Open-data interface to environmental and other public data in Metsään.fi databases the main lessons learned were related to the implementation phase. It's good to reserve enough resources not only for the development activities but also for the maintenance and end-user support as well as for training. The best finding of this pilot was that simple solutions do work!

Regarding the Shared multiuser data environment in Metsään.fi-service topic the main lessons learned were related to technology availability. Due to certain purpose limitation factors the similar authorization processes for all the Metsään.fi end-users could not be applied. As the user role defines the authorization process and this was not yet considered in Suomi.fi services, only part of the authorization process automation could be completed.

Regarding the crowdsourcing solutions and pilotable topics, the main lessons learned was that it is quite easy to produce new crowdsourcing tools. However, the real challenge is to find out how to motivate citizens to produce the information in an unbiased way, which is a difficult question.

4.2.7.2 Updates on the work done on Trial 2

Changes for Trial 2 and beyond:

- For easier database management especially regarding the crowdsourcing solutions including photos, the standardized way of working was implemented by using the XML format. For the first pilots the application programming interface API has been built based on GeoJSON standard, which is not the standardized solution for the data transfer between METSAK and partners.

- At the beginning of 2019 the required XML standard schema version was released and after that, the X-road approach was taken into use also for the crowdsourcing solutions regarding the forest damages reported by Laatumetsä mobile application. This activity was finalized by Q3/2019 and it is more or less technical solution improvement activity and not visible for the end-users.
- C18.01: The data gathered via Wuudis application is utilized in updating METSAK forest resource data.
- C18.01: The quality control data is used for updating the Metsään.fi forest resource data.
- C18.01: Metsään.fi user authentication via Suomi.fi national service architecture portal for forest owners.
- C18.01: The forest resource data system (standardization, data transfer service, database) and processes was adjusted accordingly.
- C18.01: The forest resource data system (standardization, data transfer service, database) and processes was adjusted accordingly.
- C18.01: Single-login and User role-based authentication process integrated to Suomi.fi e-authorization process for forest owners, was implemented in Metsään.fi 2.0 version in 2020.

4.3 WP3 - Fishery

4.3.1 Pilot A1: Oceanic tuna fisheries immediate operational choices

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

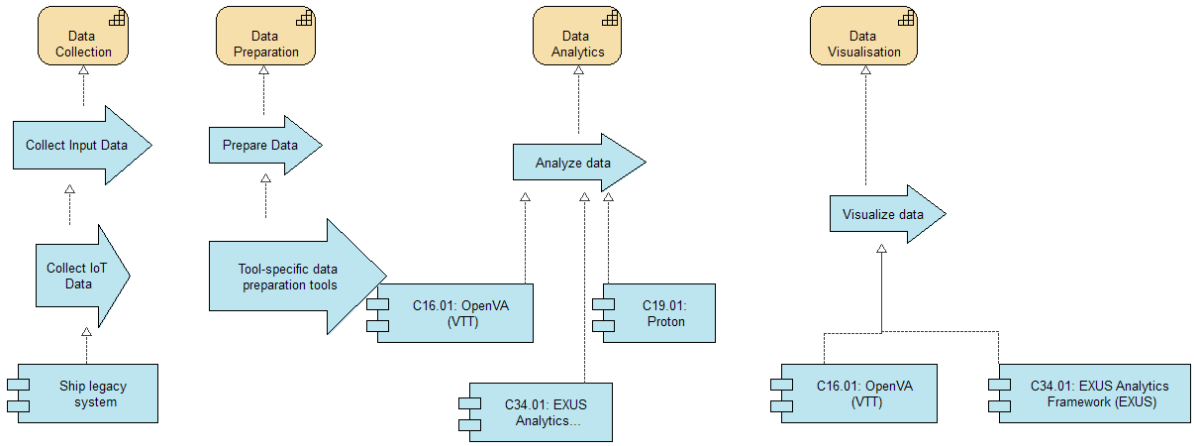


Figure 77: Pilot A1: Oceanic tuna fisheries immediate operational choices pipelines

SENTINEL-3 and Copernicus Marine Environment Monitoring Service data (i.e. sea surface currents, temperature, wind speed, chlorophyll, phytoplankton and other oceanographic parameters) is used in combination with weather conditions information and models as well as in-situ real-time observations from the fleet (i.e., engines, propulsion, route and speed of

the vessel, destination) in order to establish a common data management and analysis system that allows to reduce fuel consumption from the interaction between engine data, propulsion data, meteorological data and the vessels design by means of Big Data approaches and to estimate the expected lifetime of different parts of the engine and propulsion system, or to learn when a part of the engine is close to failure and advise the technical staff in order to have the necessary parts and technicians at port and reduce the downtime for unexpected failures.

This pilot and its trials used the following technological components:

- C02.02 – Micka
- C02.03 - HS Layers NGC12.02b – Albatross
- C12.04 – FuelEst
- C16.01 – OpenVA
- C19.01 - IBM Proactive Technology Online
- C34.01 - EXUS Analytics Framework

4.3.1.1 Brief summary of results and pending issues from Trial 1

Key component is satellite communication with ships. Faulty communication with ships has been a common problem during Trial 1 and this has caused delays and many implementation problems of the solutions. Vessels are close to 30 days in operation in the sea and remain just few days in harbour (usually 3) before departing. This generates operational problems difficult to overcome and impact the project outcome.

Analysing the historical data that is produced by the vessel systems was implemented in Trial 1 using a small subset of the full data. Even though the three vessels were almost identical, the data produced has differences that needed to be preprocessed. However, the first Trial offered good ground for implementing the full-scale data analysis in Trial 2.

4.3.1.2 Updates on the work done on Trial 2 and main results

Real time warnings and alarms onboard: Proton (Complex Event Processing, CEP) and VTT OpenVA (Visualization):

- A RESTful consumer for PROTON situations (output derived events) for integration with VTT dashboard was added. The goal has been to present the alerts to the operators in a manner that suits them. PROTON pushes the derived events to a RESTful endpoint provided by the VTT dashboard.
- A set of event rules with clearing events for CEP situations for the integration of the VTT OpenVA dashboard was implemented, so that visualization of a present alarm/situation is "cleared" once the problematic situation has been taken care of by the operators and PROTON's engine detects (based on raw input events representing the ship's operational parameters) that the problematic situation no longer exists.
- Monitoring rules in for more engine's operational parameters to detect more problematic situations related to engine's operation were added.

- Extended Proton CEP application for tuna fisheries with situation clearing rules, new dockerized endpoint was implemented.
- RESTful interface has been implemented between Proton and OpenVA, data logger interface application sending RESTful messages to Proton. New dockerized endpoint has been implemented for both Proton and VTT OpenVA, to enable easy integration. RESTful NGSI consumers and producers for FIWARE context broker have been implemented in Proton.
- A new user interface has been implemented for VTT OpenVA to visualize the warning and alarm messages produced by Proton.

Results: The final solution has up to 33 different variables monitored with different rules. Basic condition for issuing warnings and alarms is that the engine is running and in steady-state condition. In such conditions, parameters must remain steady. Based on this if there is a deviation of parameters from such steady condition WARNING is issued and if the problem remains or goes worst, ALARM is issued. All of this is executed visually in a clear manner for ease of interpretation by operator (crew). Variables monitored:

- Jacket cooling water temperature.
- Lubricating oil Pressure (in 2 points in the engine).
- Lubricating oil Temperature.
- Cylinder gas outlet temperature, 9 cylinders (deviation from average and high value of temperature).
- Main Bearing Temperature, 11 main bearings (deviation from average).

The warnings and alarm are recorded in files for further analysis. The analysis can cover the condition in which it happened, values of warning/alarm, engine running condition, etc… The values of variables are recorded onboard and in the cloud, so “a posteriori” analysis of the situation can be carried out by crew or by staff in the office onshore. This permits evaluation of the engine condition in more detail than it was available until now, with a better understanding of the condition and improved proactive maintenance of engine by staff on board.

Historical data analysis system (VTT OpenVA):

New analysis reports were implemented: in the final version 56 different performance indicators are calculated and 3 reports are issued grouped in: Main Engine, Auxiliary Engines, Ship.

New visualizations were implemented: in the final version 5 pilot specific visualizations have been implemented, in addition to the 12 standard visualizations offered by VTT OpenVA base configuration.

New Python analysis backend has been implemented to support data import and Python based visualizations. Using this backend, automatic continuous data import has been implemented, enabling daily updates of new data.

New vessel has been included in the fleet analysis (second half of the year a new vessel with similar characteristics to the 3 in operation has been added to the fleet and included in data analysis system).

All these new VTT OpenVA features have been published in open source to enable maintenance of the service after the project with sample data of ships in operation.

Results: A server-based energy performance analysis solution has been developed. The final solution is based on a server in order to be used by shipowner and technical staff in the office, onshore. It is possible to access the solution via browser and an internet connection. The solution provides information about fleet energy performance in a user-friendly manner. The information analysed can be ship-based, i.e., only for one ship, or can be fleet based for comparison and benchmarking, i.e., selecting more than one ship and comparing values between them (reports provide combined information).

The solution provides reports and graphical information with many different visualization options and possibilities also to download data if it is necessary to analyse with other means.

The solution now makes it possible for the technical staff in the offices onshore to analyse the data and interact with the crew on board for more efficient vessel operation, reducing fuel consumption on board. Considering human nature, just the possibility of monitoring crew onboard increases crews' awareness and implication in energy-saving policies during vessel operation.

All these new VTT OpenVA features have been published in open source to enable maintenance of the service after the project with sample data of ships in operation. As well as, giving information of solutions to the technical and scientific community in order to be applied in other marine sectors, and also in other fishing fleets, contributing to the whole fishing and marine industry community.

Fault detection solution (EXUS, Data Analytics Framework):

- Filtering of initial data of healthy engine condition. Implementation of steady-state condition algorithm.
- Obtain models of a healthy engine using ANN with filtered data of healthy engine condition (when new).
- Implement scripts to predict engine running performance parameters (pressure, temperature, etc…) based on certain inputs.
- Design and implement user interface with feedback from EHU and shipowner.
- Implement the offline Application with a friendly user interface in a browser to detect deviations of engine condition from the healthy condition.

Results: It has been possible to develop a digital twin model of ship main engine based on historical data and using machine learning techniques with the aim of reproducing the

engine's running condition virtually, based on engine operational data. In this way, the engine model calculates performance parameters (pressure, temperature) based on some input parameters (FORACK, RPM…).

The application represents graphically:

- Deviations between actual performance parameters of the engine and theoretical values calculated with the digital twin model of healthy engine condition.
- Evolution in time of deviation (deviation between actual and predicted values).
- Deviation relative values and deviation daily average values have been defined to ease analysis.

Although the system improves the work of engine performance analysis and fault detection by a human operator, it is not yet fully automatic and expert human analysis is still needed for assessment of condition. With the solution, some faults have been detected with the engine in operation that would be extremely complicated to find in the "as usual" way of monitoring the engine (i.e., checking hand-written parameters noted by crew).

Anyway, it must be remarked that not all faults are possible to predict, but the solution has meant a big leap in the technical staff capability in the time of evaluating engine condition. Now technical staff onshore is able to analyse and monitor an engine parameter evolution in time with a glance and just making some clicks with the mouse on a computer.

The solution will be used on a regular basis by technical staff in their future usual work to monitor engine conditions.

4.3.2 Pilot B1: Oceanic tuna fisheries planning

During Trial 1 activities, a key result for WP4 and WP5 perspectives is the definition of the pilot pipeline, which is presented below:

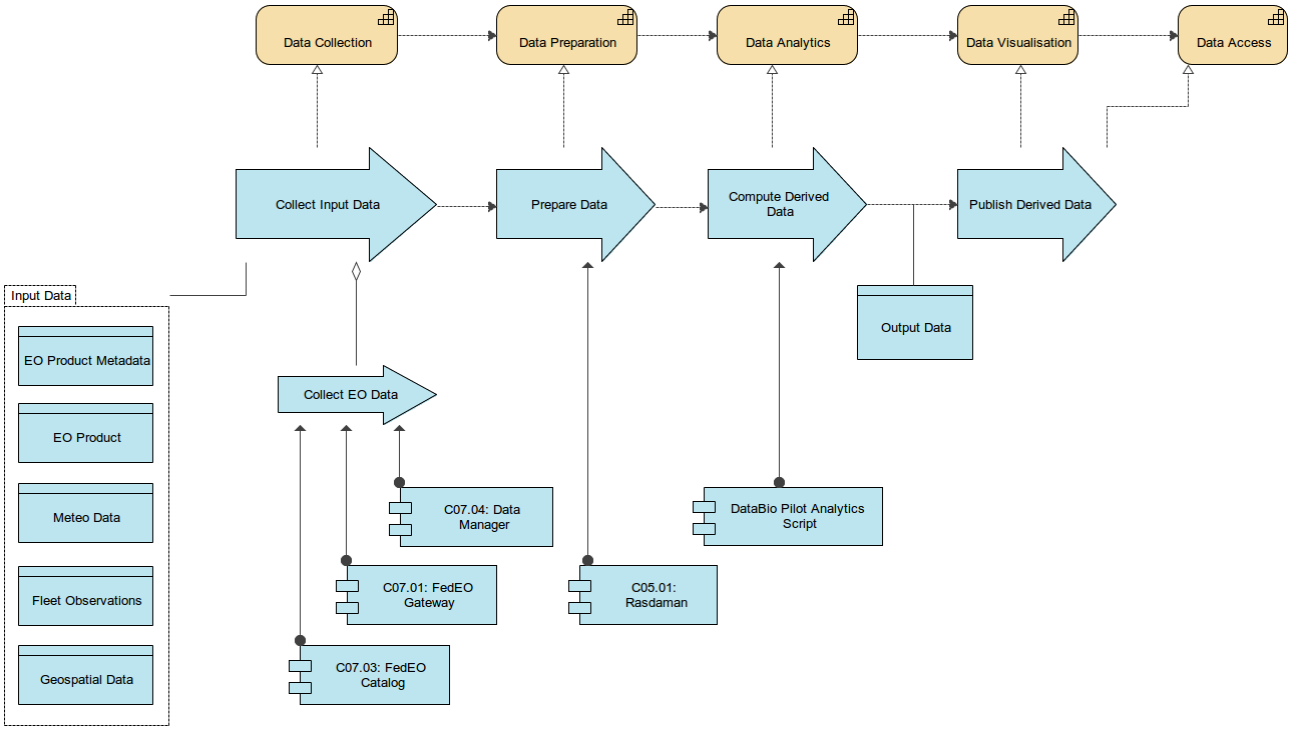


Figure 78: Pilot B1: Oceanic tuna fisheries planning pipelines

SENTINEL-3 and Copernicus Marine Environment Monitoring Service data (i.e. sea surface currents, temperature, wind speed, chlorophyll, phytoplankton and other oceanographic parameters) was used in combination with weather conditions information and models as well as in-situ real-time observations from the fleet (i.e., engines, propulsion, route and speed of the vessel, destination) in order to establish a common data management and analysis system that allows to improve profitability of oceanic tuna fisheries through savings in fuel costs through fish observation and route optimisation by means of unassisted machine learning. In that regard, all the historical data available i.e., catch logbooks, GPS, Buoys, Observers, etc. were put in relation in different databases and used to learn the better way to fish with the FAD (Fish Aggregating Devices) method.

After the Trial 1 - full integrated system with *historical data* was tested by scientists and end-users from fisheries industry.

4.3.2.1 Brief summary of results and pending issues from Trial 1

This pilot had an important part in improving the AZTI hardware capacity in order to accommodate the data acquisition, processing, analysis and visualization components. The solution is composed of three parts. A first part is dedicated to the storage of data and components (software) that configure this platform solution. This first part provides not only the storage but also the computational capacity to run the components. Modelling and analysis are the second part of the solution composed of existing algorithms to be adapted and new algorithms in order to give advice about species distribution and fishing behaviour advice. Finally, the third part deals with the results from the previous parts in order to visualize results and advice not only the fisheries operators and managers in a friendly and

easy to understand way, but also to provide management advice to organizations such as local government, ICES and FAO. This engagement with stakeholders and outcomes users started with Echebastar fleet, presentation at EFARO marine science centers network, involvement in the WKMLEARN ICES working group (machine learning in marine science) and presentation at the FAO Committee on Fisheries (COFI). An informative poster was produced and published (DOI: 10.13140/RG.2.2.22519.32165).

4.3.2.1.1 Earth Observation aspect

The pilot uses Copernicus Marine Environment Monitoring Service (CMEMS) products, namely: Global Ocean 1/12" physics and biochemical analysis and forecast. These **EO data** are then combined with other data, such as Fisheries «point» data from Fisheries monitoring (namely: Catches by Species in Kg and Vessel Monitoring System) to produce a so-called GAMS statistical model [REF-38] for planned species distribution forecasting. The pilot uses six **Earth Observation related components**:

- C07.01 FedEO Portlet (Spacebel)
- C07.01 FedEO Gateway (Spacebel)
- C07.03 FedEO Catalog (Spacebel)
- C07.04 Data Manager (Spacebel)
- C07.06 Ingestion Engine (Spacebel)
- C09.06 Apache Oozie (PSNC)

The **EO related results** of the pilot are the maps of species distribution forecast, whose are calculated by different statistical models using a combination of Earth Observation and survey data as inputs.

- What is the lesson learned from Trial 1?

Not all the data available from the models is necessary for constructing a good forecast model for fish biomass

- Is there a need for a redesign for Trial 2?

The routines for automatic download will be redesigned to download only the data selected in Trial 1, making quicker the downloading of the data needed.

4.3.2.1.2 Planned EO related changes for Trial 2

During Trial 2 only the variables selected in the analysis of Trial 1 will be used and included, all the variables with high intra correlations will be cleared from the system.

4.3.2.2 Updates on the work done on Trial 2

Fuel oil consumed per kilogram of the catch was noticeably reduced. All the five ships of this fleet reduced their fuel consumption with reductions that range from 4 to 30% with a 19% reduction on average. The strategies that have allowed these improvements have been identified, to better plan activities and accomplish further improvements. Further engagement with FAO and other organizations had led to this report (http://www.fao.org/documents/card/en/c/ca7012en) important evaluation of fishing

activity globally through the capacity of AIS technology Big Data and machine learning. The work of WKMLEARN has continued with a long-term strategy to build a network, a planned publication and yearly meetings. We also joined a new ICES working group on shipping impacts (WGSHIP) were DataBio work has been presented.

4.3.3 Pilot A2: Small pelagic fisheries immediate operational choices

This pilot uses the following technological components:

- C17.01 – Ratatosk
- C17.02 – STIM

4.3.3.1 Brief summary of results and pending issues from Trial 1

- Data collection from vessels and integration with a datacentre (SMD) builds a comprehensive dataset of vessel operations that can be utilized.
- Collected data needs augmentation in the form of calculated "synthetic data" in order to form a complete dataset.
- Vessel energy system flexibility and transient operations make use of event processing difficult to implement (Use of C19.01).
- Signal errors and misconfigurations on board the vessels makes the collection and maintaining of decision support data exclusively on-board the vessels difficult, corrections and reconfiguration from shore/datacenter) would be an advantage.

4.3.3.2 Updates on the work done on Trial 2

- C17.01: Continued collection of data on-board vessels.
- C17.02: Creation of a tool to generate decision support database from corrected offline data. The data should be compatible with the on-board decision support system.
- C17.01. Update mechanism for new decision support database shore was made.

4.3.4 Pilot B2: Small pelagic fisheries planning

This pilot uses the following technological components:

- C17.02 – STIM
- C17.03 – KRAKK
- C17.04 – KORPS

4.3.4.1 Brief summary of results and pending issues from Trial 1

- Trial 1 focussed on identifying and providing datasets that are needed to achieve the goals of this pilot. This entails the development of tools and infrastructure for making data available for end-users through visualization technologies.
- Need to get feedback from end-users on what the web application should present.

4.3.4.2 Updates on the work done on Trial 2

- Implemented multiview map with WMTS-T layers from oceanographic simulation model (SINMOD)
- Added playback of WMTS layers with time manipulation together with catch data species and time filtering capabilities
- Implemented access control and security mechanisms for web application
- Improved operationalization of various aspects pertaining to pilot functionality, including source data processing, management, and visualization
  - GeoServer ingestion of selected oceanographic SINMOD variables, with improved performance through configuration of GeoWebCache
  - Careful selection of raster layer palette styles

A **web application** was developed, which allows end-users to scrutinise historical catch data together with biomarine layers, such as temperature and phytoplankton. The application lets the user run playbacks of selected species and time periods to observe trends and correlation between catches and biomarine attributes.

4.3.5 Pilot C1: Pelagic fish stock assessments

This pilot uses the following technological components:

- C01.01 – SLA
- C17.02 – STIM
- C17.03 – KRAKK
- C17.04 – KORPS

4.3.5.1 Brief summary of results and pending issues from Trial 1

- Extending C17.01 Ratatosk (Onboard data acquisition software) with unattended acquisition of hydroacoustics data was found to be too laborious.
- Fish migration simulation may be plausible with ecosystem approach.

4.3.5.2 Updates on the work done on Trial 2

- C01.01: A new bigger dataset was provided by SINTEF Ocean and all three methods were applied once again to verify the existence of fish in specific areas. Principal components analysis (PCA) was also examined as a pre-processing method. All classification approaches were tested on MVBS values for different combinations of the five frequencies measured. All methods reach an accuracy of more than 92%. However, the kappa coefficient varies greatly among the different approaches. Support Vector Machine (SVM) with radial kernel seems to be more appropriate for the acoustic data, whereas linear kernel totally fails in discriminating the classes of fish presence and fish absence. For all cases, the best results come from using vectors comprising all five frequencies measured.
- C01.01: An interface for end-users to upload datasets for analyses.

4.3.6 Pilot C2: Small pelagic market predictions and traceability

This pilot uses the following technological components:

- C17.03 - KRAKK
- C17.04 - KORPS

4.3.6.1 Brief summary of results and pending issues from Trial 1

- Trial 1 has focussed on identifying and providing datasets that can support the overall goals of this pilot. It has developed the necessary tools and infrastructure for making these datasets available for further processing and analysis, and done a preliminary, exploratory analysis of the data to identify the most obvious characteristics.
- It is clear that for Norwegian Mackerel, the price and quantities of the landings are very dependent on time, with strong seasonal dependencies.
- It is also clear that there are other factors which contribute to variations between years, assumed to be related to both market situations, existing storage, present and future quotas and past, present and future production of other fish species and alternative food sources.

4.3.6.2 Updates on the work done on Trial 2

Trial 2 was built upon the technical developments made in Trial 1, as well as the improved understanding about which factors affect the future price variations of Norwegian Mackerel. Trial 2 focussed on two different methods to enable the fishermen to plan better when and where to fish which species:

A **web service** was developed, allowing end-users to analyse previous fish catches and fish transactions all the way back to 2012. This enables the fishermen to understand better how the prices and availability of fish usually depend on factors such as season and moon phase.

Also, a price prediction pipeline for Norwegian Mackerel was developed, where different methods for prediction have been evaluated. As the fish prices are influenced by seemingly chaotic and psychological factors, predicting prices is inherently difficult. In this respect, the fish market is not much different than other markets, such as the stock market. Still, the achieved results are as good as expected, and it might be possible to use this for the benefit of the stakeholders. The next goal will be to develop an operational service which provides regularly updated price predictions. To improve performance, additional datasets should be considered.

5 Lessons learned and best practices

A lot of *opportunities* rise from the DataBio platform environment developed in the project. As stated frequently earlier, the project did not aim at a monolithic operative platform, because such would be extremely difficult to get into use and would rapidly become outdated. Instead, the project has created a *network of resources* with a sandbox to be used in the iterative development of data-driven bioeconomy solutions.

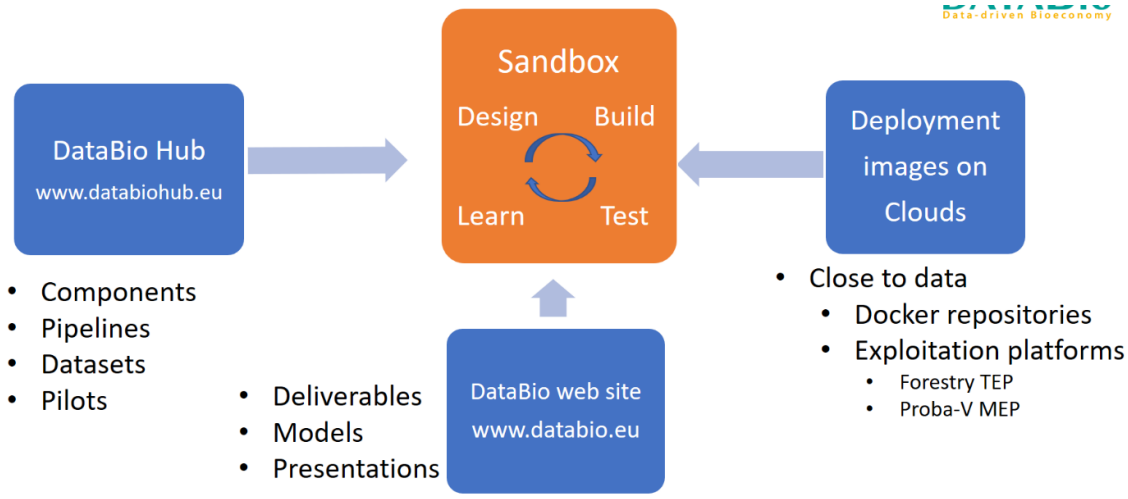


Figure 79: The DataBio platform seen as a development sandbox for data-driven bioeconomy solutions within a network of resources

The sandbox uses as resources mainly the DataBio Hub, but also the project web site and deployed software on public and private clouds. The Hub links to content both on the DataBio website (deliverables, models) and to the Docker repositories on various clouds. This environment has the potential to make it easier and faster to design, build and test digital solutions for the bioeconomy sectors in the future. The development is iterative; after learning from the tests and trials, the designs are modified, and a new circle starts.

Several *limitations* in the applied technologies can be identified. The main ones are

- Cloudy conditions in satellite images can disturb the image analysis, like deciding on the harvesting moment for potato
- Training material from more than one year might be needed in e.g. in reliable crop detection
- Crop detection algorithms are highly dependent on ground data retrieved from farmer´s declarations
- Crowdsourcing of forest conditions requires the active participation of forest owners and others moving in the forests
- Fishery market price prediction is challenging (like the stock market), but still possible to use for stakeholders

When looking at the Climate Crises and a growing global population, IT and AI have to be applied in much larger extent than currently. The DataBio results can here form a

steppingstone for future developments by the DataBio partners themselves, a group of partners and totally external actors. Especially interesting will be to see how much the multi-party pipelines will be used by partners in new business configurations.

6 References

Reference	Name of document
[REF-01]	DataBio deliverable D4.1: Platform and Interfaces, v1.0, 2018-05-30.
[REF-02]	DataBio deliverable D4.2: Services for Tests, v1.0, 2018-05-31.
[REF-03]	DataBio deliverable D4.3: Data sets, formats and models, v1.0, 2018-08-31.
[REF-04]	DataBio deliverable D5.1: EO Component Specification, v1.0, 2017-12-29.
[REF-05]	DataBio deliverable D5.2: EO Component and Interfaces, v1.0, 2018-05-30.
[REF-06]	DataBio deliverable D5.3: EO Services and Tools, v1.0, 2018-06-15.
[REF-07]	DataBio deliverable D6.2: Data Management Plan, v1.0, 2017-06-30.
[REF-08]	OGC 18-049r1, OGC Testbed-14: Application Package Engineering Report, https://docs.opengeospatial.org/per/18-049r1.html
[REF-09]	OGC 18-050r1, OGC Testbed-14: ADES & EMS Results and Best Practices Engineering Report, https://docs.opengeospatial.org/per/18-050r1.html
[REF-10]	OGC 19-020r1, OGC Testbed-15 Catalogue and Discovery Engineering Report, http://docs.opengeospatial.org/per/19-020r1.html
[REF-11]	European Commission, GeoDCAT-AP Version 1.0.1, https://joinup.ec.europa.eu/solution/geodcatapplication-profile-data-portals-europe/distribution/geodcat-ap-101-pdf
[REF-12]	OGC 19-062, OGC API Hackathon 2019 Engineering Report, [http://docs.opengeospatial.org/per/19-062.html]
[REF-13]	OGC 17-069r3, OGC API - Features - Part 1: Core, http://docs.opengeospatial.org/is/17-069r3/17-069r3.html
[REF-14]	SpatioTemporal Asset Catalog specification, https://github.com/radiantearth/stac-spec
[REF-15]	2019 OGC API - Features and Catalogues Sprint, https://www.opengeospatial.org/events/191105apisprint
[REF-16]	https://databio.spacebel.be/eo-features/readme.html
[REF-17]	Radiant Earth Foundation, Community Sprint, Arlington VA, U.S.A, https://github.com/radiantearth/community-sprints/blob/master/11052019-arlignton-va/group-work/progress.md#databio
[REF-18]	OGC 17-084, EO Collection GeoJSON(-LD) Encoding Best Practice, https://portal.opengeospatial.org/files/?artifact_id=83226
[REF-19]	https://databio.spacebel.be/eo-catalog/readme.html
[REF-20]	JSON-LD 1.1, A JSON-based Serialization for Linked Data, W3C Working Draft 10 May 2019, https://www.w3.org/TR/json-ld11/
[REF-21]	Radiant Earth Foundation, Community Sprint, Arlington VA, U.S.A, https://github.com/radiantearth/community-sprints/blob/master/11052019-arlignton-va/group-work/progress.md#databio
[REF-22]	D. W. Archer, D. Bogdanov, B. Pinkas, and P. Pullonen, “Maturity and performance of programmable secure computation.” Cryptology ePrint Archive, Report 2015/1039, 2015, https://eprint.iacr.org/ 2015/1039

[REF-23]	Bogdanov, D., Jõemets, M., Siim, S., & Vaht, M. (2015, January). How the Estonian tax and customs board evaluated a tax fraud detection system based on secure multi-party computation. In *International conference on financial cryptography and data security* (pp. 227-234). Springer, Berlin, Heidelberg.
[REF-24]	Bogdanov, D., Kamm, L., Kubo, B., Rebane, R., Sokk, V., & Talviste, R. (2016). Students and taxes: a privacy-preserving study using secure computation. *Proceedings on Privacy Enhancing Technologies*, 2016(3), 117-135.
[REF-25]	Talviste, R. (2011). Deploying secure multiparty computation for joint data analysis—a case study. Master's thesis.
[REF-26]	Fiskeridirektoratet. "Åpne data: fangstdata koblet med fartøydata." Fiskeridirektoratet, 2 Oct. 2018, https://www.fiskeridir.no/Tall-og-analyse/Aapne-data/Aapne-datasett/Fangstdata-koblet-med-fartoeydata
[REF-27]	Wood, D., Lanthaler, M. & Cyganiak, R. (2014). RDF 1.1 Concepts and Abstract Syntax [W3C Recommendation]. (Technical report, W3C).
[REF-28]	Harris, S. and Seaborne, A. (2013). SPARQL 1.1 Query Language. W3C Recommendation. W3C.
[REF-29]	Hyland, B., Atemezing, G., Villazón-Terrazas, B. (2014). Best Practices for Publishing Linked Data. W3C Working Group Note 09 January 2014. https://www.w3.org/TR/ld-bp/
[REF-30]	Heath, T. and Bizer, C. (2011) Linked Data: Evolving the Web into a Global Data Space (1st edition). *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1:1, 1-136. Morgan & Claypool.
[REF-31]	Palma R., Reznik T., Esbri M., Charvat K., Mazurek C., An INSPIRE-based vocabulary for the publication of Agricultural Linked Data. *Proceedings of the OWLED Workshop: collocated with the ISWC-2015*, Bethlehem PA, USA, October 11-15, 2015
[REF-32]	FOODIE project deliverable D2.2.3: Service Platform Specification. Miguel Ángel Esbrí. 2016
[REF-33]	Nikolov, Andriy et al. "Ephedra: SPARQL Federation over RDF Data and Services." International Semantic Web Conference (2017), https://www.metaphacts.com/images/PDFs/publications/ISWC2017-Ephedra-SPARQL-federation-over-RDF-data-and-services.pdf
[REF-34]	Habyarimana, E.; Lopez-Cruz, M. Genomic Selection for Antioxidant Production in a Panel of Sorghum bicolor and S. bicolor × S. halepense Lines. Genes 2019, 10, 841.
[REF-35]	Habyarimana E, Paris B, Mandolino G. 2017. Genomic prediction for yields, processing and nutritional quality traits in cultivated potato (Solanum tuberosum L.). *Plant Breed*. 136(2): 245-252. DOI:10.1111/pbr.12461
[REF-36]	Habyarimana E. 2016. Genomic prediction for yield improvement and safeguarding genetic diversity in CIMMYT spring wheat (Triticum aestivum L.). *Australian Journal of Crop Science*, 10(1):127-136.
[REF-37]	Habyarimana E, Dall'Agata M, De Franceschi P, Baloch FS (2019) Genome-wide association mapping of total antioxidant capacity, phenols, tannins, and flavonoids in

	a panel of Sorghum bicolor and S. bicolor × S. halepense populations using multi-locus models. PLoS ONE 14(12): e0225979.
[REF-38]	Arrizabalaga, H., Dufour, F., Kell, L., Merino, G., Ibaibarriaga, L., Chust, G., ... & Chifflet, M. (2015). Global habitat preferences of commercially valuable tuna. *Deep Sea Research Part II: Topical Studies in Oceanography*, 113, 102-112
[REF-39]	DataBio deliverable D1.3: Agriculture Pilot Final Report, to be published.
[REF-40]	NIST Big Data Interoperability Framework: Volume 6, Reference Architecture. Available at https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-6.pdf

Appendix A Classification of the components

Data Visualisation and User Interaction	
1D	*2*
Genomic models	*C22.03*
Rasdaman	*C05.01*
2D	*11*
Agriculture Map Generator	*C11.01*
EXUS Analytics Framework	*C34.01*
Genomic models	*C22.03*
HSLayers NG	*C02.03*
HSLayers-NG Cordova mobile application	*C02.04*
Map server for forest health maps	*C14.07*
NeuroCode	*C13.01*
OpenVA	*C16.01*
Rasdaman	*C05.01*
SINTIUM	*C06.02*
WebGLayer	*C03.01*
3D	*3*
Genomic models	*C22.03*
Rasdaman	*C05.01*
SmartVis3D	*C04.04*
4D	*1*
Rasdaman	*C05.01*

VR/AR	*0*
Data Analytics	
Descriptive	*13*
Albatross	*C12.02*
Analysis of vegetation indices and biophysical products	*C14.06*
EO Crop Monitoring	*C39.02*
EXUS Analytics Framework	*C34.01*
Genomic models	*C22.03*
KRAKK	*C17.03*
KORPS	*C17.04*
MADChange	*C29.02*
Neural network suite for image processing	*C31.01*
OpenVA	*C16.01*
Rasdaman	*C05.01*
Supervised learning algorithm	*C01.01*
WishartChange	*C29.01*
Diagnostics	*3Wishart*
Genomic models	*C22.03*
Neural network suite for image processing	*C31.01*
Forest logging detection	*C14.05*
Predictive	*7*
Genomic models	*C22.03*
EO4SDD	*C12.03*

FuelEsti	*C12.04*
EXUS Analytics Framework	*C34.01*
KRAKK	*C17.02*
OpenVA	*C16.01*
SLA	*C01.01*
Prescriptive	*4*
FarmTelemetry	*C02.05*
GAIABus DataSmart Machine Learning Subcomponent	*C13.02*
Genomic models	*C22.03*
OpenVA	*C16.01*
Data Processing Architectures	
Batch	*10*
Apache Oozie	*C09.06*
eEOPS - e-GEOS EO processing service	*C28.01*
Envimon	*C16.09*
EXUS Analytics Framework	*C34.01*
Genomic models	*C22.03*
Mosaic Cloud Free Background	*C39.01*
Probability	*C16.07*
Sentinel2 Clouds, Shadows and Snow Mask	*C39.03*
STIM	*C17.02*
Radiometric Corrections	*C11.03*
Interactive	*3*

Genomic models	*C22.03*
OpenVA	*C16.01*
Rasdaman	*C05.01*
Streaming/Real-time	*5*
GAIABus DataSmart RealTime Subcomponent	*C13.03*
Genomic models	*C22.03*
IBM Proactive Technology Online	*C19.01*
SensLog	*C02.01*
Supervised learning algorithm	*C01.01*
Data Management	
Collection	*8*
Data Manager	*C07.04*
FedEO Catalog	*C07.03*
FedEO Gateway	*C07.01*
Genomic models	*C22.03*
Ingestion Engine	*C07.06*
Proba-V MEP	*C08.02*
Ratatosk	*C17.01*
Senop Hyperspectral Camera	*C44.01*
Preparation	*10*
Forest Health Status	*C11.02*
DataGraft	*C06.01*
Data Manager	*C07.04*

EXUS Analytics Framework	*C34.01*
FedEO Gateway	*C07.01*
Forest Health Status	*C11.02*
Modelio BA Data modelling tool	*C37.01*
Modelio PostgreSQL modeller	*C37.03*
Senop Hyperspectral Camera	*C44.01*
Sentinel-1 IWS pre-processing	*C14.04*
Curation	*9*
Atmospheric corrections of Sentinel-2 data	*C14.01*
DataGraft	*C06.01*
Data Manager	*C07.04*
EXUS Analytics Framework	*C34.01*
FedEO Catalog	*C07.03*
FedEO Gateway	*C07.01*
Forest Health Status	*C11.02*
GeoToolbox	*C04.03*
Sentinel-1 IWS pre-processing	*C14.04*
Linking	*5*
Genomic models	*C22.03*
geoLIMES	*C12.01*
Micka	*C02.02*
MEA WCS	*C41.01*
MEA GUI	*C41.02*

Access	*11*
Data Manager	*C07.04*
GeoRocket	*C04.02*
Genomic models	*C22.03*
FedEO Catalog	*C07.03*
FedEO Gateway	*C07.01*
Forestry TEP -- Forestry Thematic Exploitation Platform	*C16.10*
Metsään.fi	*C18.01*
Open Forest Data	*C18.02*
Rasdaman	*C05.01*
SensLog	*C02.01*
Infrastructure	
Cloud	*1*
Wuudis	*C20.01*
Communication	*1*
Digital service hub	*C16.04*
HPC infrastructure	*1*
PSNC HPC and cloud infrastructure	*C09.13*
Iot/CPS infrastructure	*1*
FIWARE IoT Hub	*C05.02*
Data Protection	
Data Protection and Cybersecurity	*2*
Sharemind MPC	*C35.02*

	Sharemind HI	*C35.03*

Appendix B Components used in pilots

B.1 WP1 - Agriculture

Pilot name	Components used
Pilot 1 [A1.1] Precision agri-culture in olives, fruits, grapes	**C04.02, C04.03, C04.04, C13.01, C13.03, C19.01**
Pilot 2 [A1.2] Precision agri-culture in vegetable seed crops	**C08.02**
Pilot 3 [A1.3] Precision agri-culture in vegetables_2 (Potatoes)	**C08.02**
Pilot 4 [A2.1] Big Data management in greenhouse eco-system	**C22.03**
Pilot 5 [B1.1] Cereals, biomass and cotton crops	**C11.01, C11.03, C19.01, C05.02**
Pilot 6 [B1.2] Cereals, biomass and cotton crop 2	**C04.02, C04.03, C04.04, C13.01, C13.03**
Pilot 7 [B1.3] Cereal, biomass and cotton crops 3	**C08.02, C12.03**
Pilot 8 [B1.4] Cereals, biomass and cotton crops 4	**C02.02, C02.03**
Pilot 9 [B2.1] Machinery management	**C02.01, C02.05, C03.01**
Pilot 10 [C1.1] Insurance (Greece)	**C04.02, C04.03, C04.04, C13.01, C13.02, C13.03, C31.01**
Pilot 11 [C1.2] Farm Weather Insurance Assessment	**C08.02, C41.01, C41.02**
Pilot 12 [C2.1] CAP Support	**C07.01, C07.03, C07.04, C07.06, C09.06, C09.13, C39.01, C39.02, C39.03**
Pilot 13 [C2.2] CAP Support (Greece)	**C04.02, C04.03, C04.04, C13.01, C13.02, C13.03, C31.01**

B.2 WP2 - Forestry

Pilot name	Components used
Pilot 2.2.1: Easy data sharing and networking	**C18.01, C20.01**
Pilot 2.2.2: Monitoring and control tools for forest owners	**C18.01, C18.02, C20.01**
Forestry 2.3.1: Forest Damage Remote Sensing	**C16.07, C16.09, C16.10, C20.01, C44.01**

Pilot 2.3.2-FH: Monitoring of forest health	**C09.13, C11.02,C11.03, C28.01, C29.01, C29.02**
Pilot 2.3.2-IAS: Invasive Alien Species control and monitoring	**C11.014**
Pilot 2.4.1: Web-mapping service for government decision making	**C16.09, C16.10**
Pilot 2.4.2: Shared multiuser forest data environment	**C18.01, C18.02**

B.3 WP3 - Fishery

Pilot name	Components used
Pilot A1: Oceanic tuna fisheries immediate operational choices	**C02.02, C02.03, C07.01, C07.03, C07.04, C07.06, C09.06, C12.02b, C12.04, C16.01, C19.01, C34.01**
Pilot B1: Oceanic tuna fisheries planning	**C02.02, C02.03, C05.01, C07.01, C07.03, C07.04, C07.06, C09.06**
Pilot A2: Small pelagic fisheries immediate operational choices	**C17.01, C17.02, C17.04, C19.01**
Pilot B2: Small pelagic fisheries planning	**C17.02, C17.03, C17.04**
Pilot C1: Pelagic fish stock assessments	**C01.01, C17.02, C17.03, C17.04**
Pilot C2: Small pelagic market predictions and traceability	**C17.03, C17.04**

Appendix C Benefits from OGC Testbed

C.1 Exploitation Platforms

The concept of Exploitation Platforms (EP) was introduced by ESA to promote the exploitation of Earth Observation (EO) data.

An Exploitation Platform is a virtual workspace, providing the user community with access to:

- large volume of data (EO/non-space data),
- algorithm development and integration environment,
- processing software and services (e.g. toolboxes, retrieval baselines, visualization routines),
- computing resources (e.g. hybrid cloud/grid),
- collaboration tools (e.g. forums, wiki, knowledge base, open publications, social networking…), and
- general operation capabilities (e.g. user management, access control, accounting, etc.).

An Exploitation Platform thus provides a complete work environment for its users, enabling them to easily and effectively perform data-intensive research. The platform permits the execution of dedicated processing software close to the data, thereby avoiding moving large volumes of data through the network and spending time on developing tools for sourcing data, basic data manipulation, etc. Moreover, the platform offers a collaboration environment where the scientist can share their algorithm with the community, publish results and perform development.

Exploitation Platforms are usually tailored to a particular scope. For example, a Thematic Exploitation Platform (TEP) is related of a user community and research field (e.g. Hydrology, Forestry, Geohazard), a Regional Exploitation Platform (REP) to a given area of interest (e.g. Europe, Sierra Nevada, Japan), a Mission Exploitation Platform (MEP) to a particular satellite mission (e.g. PROBA-V, SMOS, GOCE).

C.2 OGC Testbeds

OGC Testbeds are part of the OGC Innovation Program initiatives aiming at providing global, hands-on, collaborative prototyping for rapid development and delivery of proven candidate specifications to the OGC Standards Program. In OGC Testbeds, Participants collaborate to examine specific geoprocessing interoperability questions posed by the initiative's Sponsors and propose consensus-based solutions along with their implementation. The results of the Testbeds are documented in Engineering Reports and are presented at a final event. OGC Testbeds are organised on a yearly basis, typically starting in April and ending in December.

The Earth Observation Clouds (EOC) thread of two recent OGC Testbeds sponsored by ESA is of interest to DataBio: Testbed 13 (2017) and Testbed 14 (2018).

C.2.1 EOC thread OGC Testbed 13

C.2.1.1 C.2.1.1 Purpose

The purpose of the EOC thread OGC Testbed 13 was to support the development of ESA TEPs by defining interfaces based on OGC standards to achieve interoperability. The main goal is to define a packaging mechanism for user developed applications and a service to deploy and execute these applications on cloud-based environments.

C.2.1.2 C.2.1.2 Architecture

The EOC thread OGC Testbed 13 architecture is illustrated in the Figure C.1.

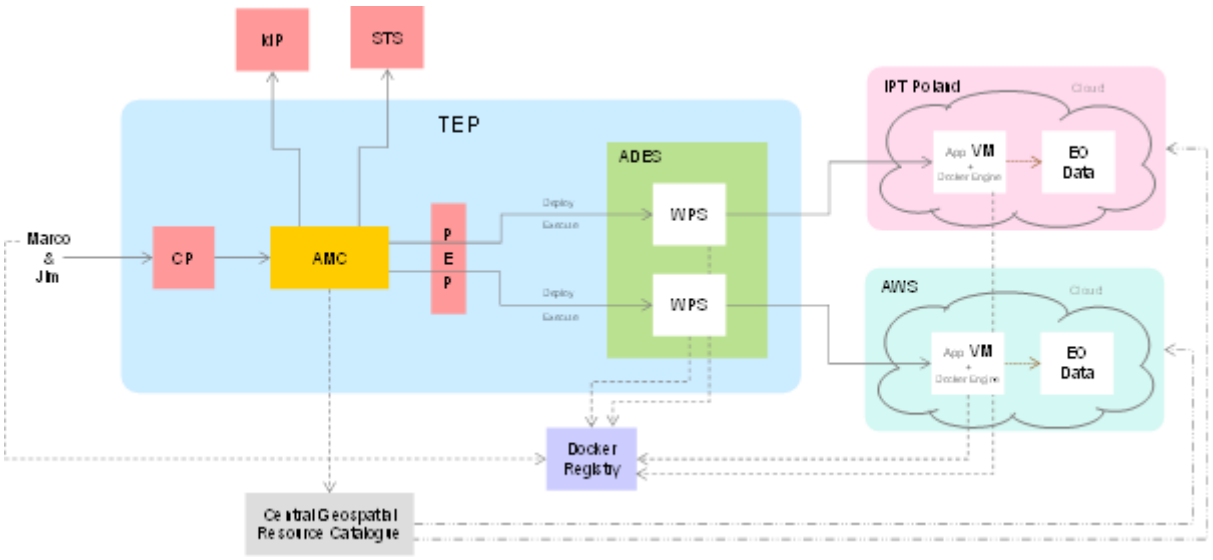


Figure C.1: EOC thread OGC Testbed 13 Deployment

The main elements to be designed and implemented are the following:

- the Exploitation Platform Application Package (in short: EP Application Package), which includes the applications and settings for its execution,
- the Application Management Client (AMC) that provides a visual front end to the user and provides the necessary business logic for deployment and execution, and
- the Application Deployment and Execution Service (ADES), which acts as a front end to cloud environments.

Two cloud environments hosting Sentinel-2 data are used: IPT Poland and Amazon Web Services (AWS).

The discovery of Sentinel-2 data on both cloud environments is performed by using an OpenSearch Gateway [REF-C01] [REF-C02] (called Central Geospatial Resource Catalogue in the figure above) that aggregates the catalogues available in each environment.

C.2.1.3 C.2.1.3 Results

The full results of EOC thread OGC Testbed 13 are documented in Engineering Reports [REF-C03] and [REF-C04]. Highlights are provided below.

Applications are packaged into Docker images. Applications are deployed and executed using an OGC Web Processing Service (WPS) [REF-C05] with the POST/XML binding. This service is also used to discover deployed applications (using the GetCapabilities + DescribeProcess operations).

The EP Application Package is based on the (Atom encoding of the) OGC OWS Context standard [REF-C06]. It includes an Atom entry with an application profile containing the following offerings:

- a Docker offering providing a reference to the Docker image in a Docker Registry, and
- a WPS Process Offering describing the inputs and outputs of the WPS Process used as a façade to the application running in a Docker container.

The EP Application Package includes also an Atom entry with a catalogue profile containing an OpenSearch offering providing endpoint information on the OpenSearch Gateway.

The deployment/undeployment of an application is performed via the WPS in two possible ways:

- either by executing a dedicated DeployProcess/UndeployProcess process, or
- by using the DeployProcess/UndeployProcess operations of a WPS supporting the Transactional extension (WPS-T) [REF-C07].

In the latter case, the EP Application Package is actually translated to a WPS Process Description that includes (using OWS Metadata) the remaining part of the above OWS Context i.e. the Docker offering and the OpenSearch offering.

Due the lack of time, the following aspects were not covered: Bulk Window and Periodic Execution, Authentication and Authorisation, Accounting and Billing.

Note that the interface between the WPS and the application (running inside the container) has not been defined (although most participants used environment variables).

C.2.2 EOC thread OGC Testbed 14

C.2.2.1 Purpose

The purpose of the EOC thread OGC Testbed 14 is to continue the work initiated in EOC thread OGC Testbed 13, to develop clear recommendations on the various aspects left open by EOC thread OGC Testbed 13, and to address the topics that were not covered due to lack of time (see above).

The focus of EOC thread OGC Testbed 14 is on a multi-platform scenario involving a TEP and several MEPs. In this scenario, (simple) workflows are deployed and executed on the TEP and applications (participating in the workflow) are deployed and executed on the MEPs.

C.2.2.2 Architecture

The Testbed 14 architecture is illustrated in the Figure C.2.

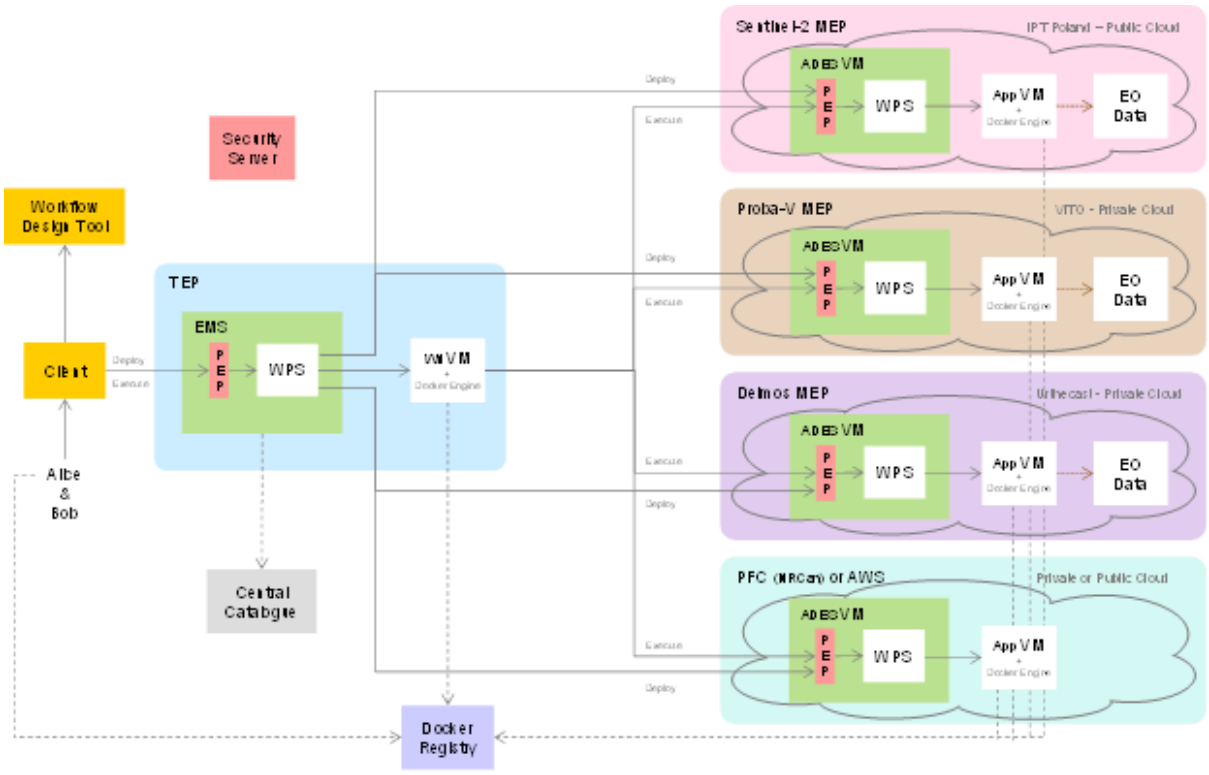


Figure C.2: Testbed 14 Deployment

The main elements to be designed and implemented are the following:

- the Web Client that provides a visual front end to the user and provides the necessary logic for the design of workflows (by Alice) and for the deployment (by Alice) and execution (by Bob) of workflows and applications,
- the Execution Management Service (EMS) which is in charge of dispatching the application execution to the relevant MEP/Execution Platform and which orchestrates the basic chaining of applications, and
- the Application Deployment and Execution Service (ADES), which acts as a front end to cloud environments.

The Figure C.3 shows an example of a (simple) multi-platform workflow where the first step is executed in parallel on several MEPs each hosting EO data from different sensors.

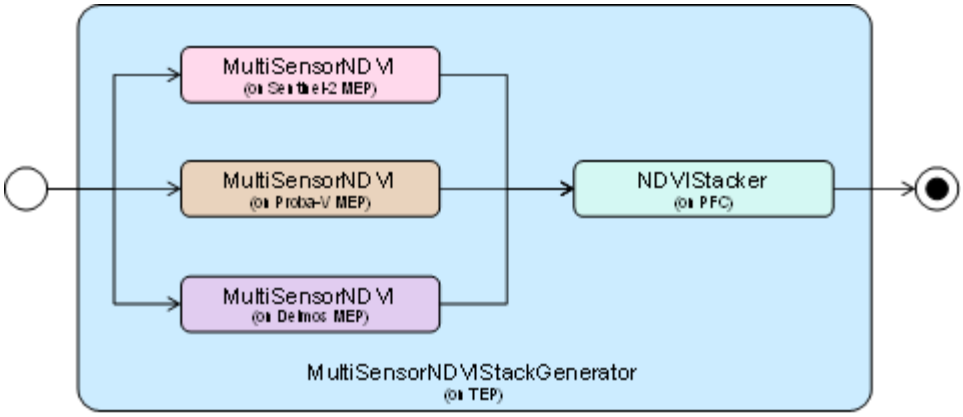


Figure C.3: MultiSensorNDVIStackGenerator Workflow

The TEP provides an environment (possibly Dockerised) where workflows are executed.

Three MEPs are used to host and process the following EO data: Sentinel-2 (IPT Poland), Proba-V (VITO), and Deimos-2 (Urthecast).

As in EOC thread Testbed 13, the discovery of EO data is performed using an OpenSearch Gateway (called Central Catalogue in the figure above) that aggregates the catalogues available in each MEP.

C.2.2.3 Results

The full results of EOC thread Testbed 14 are documented in Engineering Reports [REF-C08], [REF-C09], and [REF-C10]. Highlights are provided below.

The workflows and applications are deployed and executed using WPS services (for both the EMS and ADES) supporting the Transactional extension (WPS-T) with a REST/JSON binding. This interface was defined using an OpenAPI 3.0 specification[51].

The language used to create workflows is the Common Workflow Language (CWL)[52]. The CWL specification actually consists of two parts: a Command Line Tool specification and a Workflow specification.

The recommended way, in Testbed 14, to define (on the ADES) the interface between the WPS and the application (running inside a container) is to use the CWL Command Line Tool specification.

The EP Application Package is no longer based on the (Atom encoding of) an OGC OWS Context document. Instead, the Application Package is a WPS-T deploy process document with an execution unit referring to a Docker image in the case of an application or to a CWL file in the case of a workflow. In the first case (application), metadata is added in the process description to refer to the application CWL file (mandatory on the EMS side) or, possibly, to

[51] https://github.com/OAI/OpenAPI-Specification/blob/master/versions/3.0.0.md
[52] https://www.commonwl.org/v1.0/

provide equivalent information regarding the options/parameters of the application command line (on the ADES side).

The discovery of EO data to be processed by the workflow is done by the EMS instead of by the client (AMC) as in EOC thread Testbed 13. This is transparent to the workflow and the assignment of discovered EO data to a particular MEP is based on the catalogue collection chosen. To achieve this, after deployment on the EMS, the process created on the EMS actually exposes a modified interface with additional inputs to capture the catalogue queryables (collection, area of interest, start date, and end date).

Authentication and authorisation are based on OAuth2 and OpenID Connect rather than by using the Check Point (CP), Identity Provider (IdP), Security Token Service (STS), and Policy Enforcement Point (PEP) components as considered in EOC thread Testbed 13.

The REST/JSON WPS-T OpenAPI interface was enhanced to support:

- Authorisation (security scheme: bearerAuth; operations: getVisibility, changeVisibility),
- Quotation (operations: getQuotationList, getQuotation, executeQuotation), and
- Billing (operations: getBillList, getBill).

Note that the Quotation model used was extremely simple (e.g. price based on size of EO data input).

Note also that the propagation/aggregation of the following aspects to/from the applications in a workflow was not really considered or properly handled: Dismiss (cancel), Quotation, Billing.

C.2.3 OGC Testbed Future Work

Although the aspects left open by Testbed 13 were mainly covered in Testbed 14, some aspects still require additional work, such as:

- Quotation model
- Propagation/aggregation of Dismiss, Quotation, Billing
- Dynamic allocation/deallocation of Virtual Machines (cloud resources)
- Resource control (matching resources required by applications with resources available in the cloud environment)

Note that the last two items could be addressed using container clustering tools such as Kubernetes. This could possibly be tackled in OGC Testbed 15.

C.3 Appendix C References

Reference	Document
[REF-C01]	OGC 10-032r8, OGC OpenSearch Geo and Time Extensions, 2014-04-14

[REF-C02]	OGC 13-026r8, OGC OpenSearch Earth Observation Extension, 2016-07-06
[REF-C03]	OGC 17-023r1, EP Application Package ER, 2017-12-07
[REF-C04]	OGC 17-024r1, Application Deployment and Execution Service ER, 2017-12-07
[REF-C05]	OGC 14-065r2, OGC Web Processing Service Interface Standard, Issue 2.0.2, 2018-02-16
[REF-C06]	OGC 12-084r2, OGC OWS Context Atom Encoding Standard, Issue 1.0, 2014-01-14
[REF-C07]	OGC 13-071r1, OpenGIS WPS 2.0 Transactional Extension, OGC Discussion Paper, 2014-05-28
[REF-C08]	OGC 18-049, Application Package ER
[REF-C09]	OGC 18-050, ADES & EMS Results and Best Practices ER
[REF-C10]	OGC 18-057, Authorisation, Authentication, and Billing ER