D4.2 – Services for Tests (Public version)

<table>
<thead>
<tr>
<th>Dissemination level</th>
<th>PU - Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Document</td>
<td>Report</td>
</tr>
<tr>
<td>Contractual date of delivery</td>
<td>31/5/2018</td>
</tr>
<tr>
<td>Deliverable Leader</td>
<td>IBM</td>
</tr>
<tr>
<td>Status - version, date</td>
<td>Final - v1.0-Public, 12/12/2018</td>
</tr>
<tr>
<td>WP / Task responsible</td>
<td>WP4</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Platform, component, interface, pilot, pipeline, service, experiment, pilot, trial</td>
</tr>
</tbody>
</table>
Executive Summary

Deliverable 4.2 (D4.2) “Services for Tests” provides a comprehensive overview of the DataBio pipelines (hereinafter referred to as pipelines) identified in month 16 (M16, April 2018) of the project. Pipeline is a set of components in the DataBio platform with clear mutual interfaces linking the components together and to the platform environment. A pipeline fulfils pilot functionalities that cannot be supported by a single component. A pipeline can be seen as a white box showing the internal wiring of the components in the pipeline, thus providing technical guidance for configuration and deployment. Platform services can be seen as providers of functionalities to users that typically need to know the usability of the service, but do not need to understand the inner wiring, inner components, nor where the service is deployed.

The main goals of D4.2 are twofold. First, to provide a comprehensive overview of the pipelines’ status in M16 of the project and their level of readiness towards their first trials in M17. Second, to provide specific guidelines for their successful implementation and deployment. To this end, pipelines descriptions are detailed according to a specified template giving information about both the pipeline status and instructions for a successful deployment.

Deliverable 4.2 complements the main report on the platform (Deliverable 4.1 in work package 4) with the pipelines descriptions. Note that the dissemination level of this report is Confidential (CO), that is, only for members of the consortium (including the Commission Services), while D4.1 is a public document. With this in mind, it has been decided that D4.1 should be the main document containing all aspects of the platform, while D4.2 will complement it with the pipelines descriptions.

The pipelines are the heart of the platform as they constitute the cornerstones for the pilot’s experimentations. Next phase includes the identification of commonalities among these pipelines in terms of components and interfaces. The pipelines can be seen as important assets for exploitation at the end of the project. The goal is to have standardized interfaces and reusable pipelines not just in the three bioeconomy domains of the project: agriculture, forestry, and fishery, but in other domains as well.

Relation with Other DataBio Platform Deliverables

The DataBio project includes three piloting work packages (WP1-3) and two related platform work packages (WP4 handling IoT data and WP5 processing Earth Observation and geospatial data) that support the pilots (Figure 1). The DataBio platform provides Big Data capabilities to the pilots by forming software pipelines of components through which data
flows from the sources in agriculture, forestry and fishery through data management, analytics, and visualization stages in the pilots.

**Figure 1: Work packages and their roles in DataBio**

The platform developed in DataBio is described in the Deliverables D4.1, D4.2, D4.3 (WP4) and D5.1, D5.2, D5.3 (WP5) (Figure 1). Deliverables D4.1-3 define the Milestone M7 Service ready for Trial 1, whereas Deliverables D5.1-3 define the Milestone M9 EO Services ready for integration. The platform services and pipelines have been in trials since April 2018 (M16).

More specifically, the public deliverable D4.1 Platforms and interfaces describes the software components to be utilized by the pilots. Most of components are already in use in the first pilot trials. In addition, this deliverable reports the outcome of a matchmaking process, in which the pilots selected which components to deploy in their pilots.

Deliverable D4.2 Services for tests builds on D4.1 and provides an overview of the component pipelines as identified at month 16 (M16) of the project. It also provides guidelines for successful implementation and deployment of the pipelines.

Deliverable D4.3 Data sets, formats and models is due at the end of August 2018. While the two earlier reports deal with software modules, this report will focus on the data sets and streams employed in DataBio. Data formats, standards and models enabling easy findability, access, interoperability, and reusability of data (FAIR principle) will be dealt with. Thus, we will address in this deliverable topics beyond the coverage of single pilots.

Deliverable D5.1 EO component specification includes an analysis of the EO dataset and component related requirements provided by the pilots. It was published in 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 describes, 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 5.1 and 5.2 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 how and under which conditions these services and tools can be externally accessed.
<table>
<thead>
<tr>
<th>Contributors:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caj Södergård – VTT</td>
</tr>
<tr>
<td>Fabiana Fournier – IBM</td>
</tr>
<tr>
<td>Florian Zouhar – Fraunhofer IGD</td>
</tr>
<tr>
<td>Inna Skarbovsky – IBM</td>
</tr>
<tr>
<td>Ivo Senner – Fraunhofer IGD</td>
</tr>
<tr>
<td>Javier Hitado Simarro – ATOS</td>
</tr>
<tr>
<td>Jesus maria Estrada Villegas – TRAGSA</td>
</tr>
<tr>
<td>Joakim Haugen – SINTEF</td>
</tr>
<tr>
<td>Karl Gunnar Aarsæther – SINTEF</td>
</tr>
<tr>
<td>Karel Charvat – Lespro</td>
</tr>
<tr>
<td>Karel Charvat, jr. – Lespro</td>
</tr>
<tr>
<td>Michal Kepka – UWB</td>
</tr>
<tr>
<td>Miguel Angel Esbrí – ATOS</td>
</tr>
<tr>
<td>Ondřej Kaas – UWB</td>
</tr>
<tr>
<td>Panagiotis Ilias – NP</td>
</tr>
<tr>
<td>Pekka Siltanen – VTT</td>
</tr>
<tr>
<td>Per Gunnar Auran – SINTEF</td>
</tr>
<tr>
<td>Savvas Rogotis – NP</td>
</tr>
<tr>
<td>Seppo Huurinainen – MHGS</td>
</tr>
<tr>
<td>Stenman Virpi –METSAK</td>
</tr>
<tr>
<td>Tergujeff Renne –VTT</td>
</tr>
<tr>
<td>Tomas Reznik – Lespro</td>
</tr>
<tr>
<td>Tomas Mildorf – UWB</td>
</tr>
<tr>
<td>Zigor Uriondo Arrue – EHU</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reviewers:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tomas Mildorf – UWB</td>
</tr>
<tr>
<td>Petr Lukes – FMI</td>
</tr>
<tr>
<td>Iñaki Quincoces – AZTI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Approved by:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athanasios Poulakidas – INTRASOFT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Document History</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Version</strong></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>0.2</td>
</tr>
<tr>
<td>0.3</td>
</tr>
<tr>
<td>0.4</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>Version</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>0.6</td>
</tr>
<tr>
<td>1.0</td>
</tr>
<tr>
<td>1.0-Public</td>
</tr>
</tbody>
</table>
Table of Contents

EXECUTIVE SUMMARY .............................................................................................................. 2

RELATION WITH OTHER DATA BIO PLATFORM DELIVERABLES ........................................ 2

TABLE OF CONTENTS .............................................................................................................. 7

TABLE OF FIGURES .................................................................................................................. 9

LIST OF TABLES ........................................................................................................................ 10

DEFINITIONS, ACRONYMS AND ABBREVIATIONS ............................................................... 11

1 INTRODUCTION .................................................................................................................... 13

1.1 PROJECT SUMMARY ........................................................................................................... 13

1.2 DOCUMENT PURPOSE AND SCOPE ............................................................................... 16

1.3 PHASES IN OUR WORK ..................................................................................................... 18

1.4 RELATION WITH OTHER DOCUMENTS ......................................................................... 18

1.5 DOCUMENT STRUCTURE .................................................................................................. 19

2 DATA BIO PIPELINES .......................................................................................................... 20

2.1 WP1 - METADATA, LINKED DATA AND GRAPH DATA .................................................. 21

2.1.1 General information ....................................................................................................... 21

2.1.2 Interfaces ....................................................................................................................... 23

2.1.3 Components involved ................................................................................................... 24

2.1.4 Experimentation ............................................................................................................. 24

2.1.5 Next steps .................................................................................................................... 26

2.2 WP1 – PREDICTION AND REAL-TIME ALERTS OF DISEASES AND PESTS’ BREAKOUTS IN CROPS ............................................................................................................. 27

2.2.1 General information ....................................................................................................... 27

2.2.2 Interfaces ....................................................................................................................... 30

2.2.3 Components involved ................................................................................................... 31

2.2.4 Experimentation ............................................................................................................. 32

2.2.5 Next steps .................................................................................................................... 36

2.3 WP1 - EO DATA ANALYSIS AND MACHINE LEARNING METHODOLOGIES FOR SUPPORTING CAP AND AGRI-INSURANCE USE CASES IN GREECE ....................................................................................................................... 36

2.3.1 General information ....................................................................................................... 36

2.3.2 Interfaces ....................................................................................................................... 40

2.3.3 Components involved ................................................................................................... 40

2.3.4 Experimentation ............................................................................................................. 41

2.3.5 Next steps .................................................................................................................... 45

2.4 WP1 – FIELD DATA ANALYSIS AND REAL-TIME ALERTING FOR DECISION MAKING FOR PRECISE AGRICULTURE .............................................................................................................. 45

2.4.1 General information ....................................................................................................... 45

2.4.2 Interfaces ....................................................................................................................... 47

2.4.3 Components involved ................................................................................................... 48

2.4.4 Experimentation ............................................................................................................. 49

2.4.5 Next steps .................................................................................................................... 51

2.5 WP1 - IOT DATA ACQUISITION, ANALYSES, AND REAL-TIME ALERTS OF POTENTIAL THREAT CONDITIONS FOR CROPS .............................................................................................................. 51

2.5.1 General information ....................................................................................................... 51

2.5.2 Interfaces ....................................................................................................................... 53

2.5.3 Components involved ................................................................................................... 54

2.5.4 Experimentation ............................................................................................................. 55

2.5.5 Next steps .................................................................................................................... 57
2.6 WP1 – PUBLICATION OF LINKED DATA RELATED TO CEREALS AND BIOMASS CROPS ........................................ 58
  2.6.1 General information .................................................................................................................. 58
  2.6.2 Interfaces .................................................................................................................................. 61
  2.6.3 Components involved .............................................................................................................. 62
  2.6.4 Experimentation ....................................................................................................................... 63
  2.6.5 Next steps ............................................................................................................................... 68
2.7 WP1- IOT DATA ACQUISITION, ANALYSES, AND PREDICTION OF POTENTIAL THREAT CONDITIONS FOR CROPS ............ 68
  2.7.1 General information ................................................................................................................ 68
  2.7.2 Components involved ............................................................................................................. 69
  2.7.3 Experimentation ..................................................................................................................... 70
  2.7.4 Next steps ............................................................................................................................. 71
2.8 WP2 - SHARED MULTUSER FOREST DATA ENVIRONMENT ........................................................................ 71
  2.8.1 General information ................................................................................................................ 71
  2.8.2 Interfaces .................................................................................................................................. 73
  2.8.3 Components involved ............................................................................................................. 73
  2.8.4 Experimentation ..................................................................................................................... 73
  2.8.5 Next steps ............................................................................................................................. 74
2.9 WP2 – MULTI-SOURCE DATA ENABLING EFFICIENT FOREST MANAGEMENT ....................................................... 74
  2.9.1 General information ................................................................................................................ 74
  2.9.2 Interfaces .................................................................................................................................. 76
  2.9.3 Components involved ............................................................................................................. 77
  2.9.4 Experimentation ..................................................................................................................... 77
  2.9.5 Next steps ............................................................................................................................. 79
2.10 WP3 – SMART MANAGEMENT OF MACHINERY POWER IN SMALL-PELAGIC FISHERIES VESSEL ................... 79
  2.10.1 General information ................................................................................................................ 79
  2.10.2 Components involved .......................................................................................................... 80
  2.10.3 Experimentation ..................................................................................................................... 81
  2.10.4 Next steps ............................................................................................................................. 83
2.11 WP3 – MONITORING, REAL-TIME ALERTS, AND VISUALIZATION FOR OPERATION EFFICIENCY IN TUNA FISHERY VESSELS 83
  2.11.1 General information ................................................................................................................ 83
  2.11.2 Interfaces .................................................................................................................................. 87
  2.11.3 Components involved .......................................................................................................... 87
  2.11.4 Experimentation ..................................................................................................................... 88
  2.11.5 Next steps ............................................................................................................................. 92
2.12 WP3 - DATA ACQUISITION AND ANALYSES FOR FISHING OPERATIONAL EFFICIENCY ............................................ 92
  2.12.1 General information ................................................................................................................ 92
  2.12.2 Components involved .......................................................................................................... 95
  2.12.3 Experimentation ..................................................................................................................... 97
  2.12.4 Next steps ............................................................................................................................. 98
2.13 WP3 - DATA ANALYTICS FOR PREDICTION OF PELAGIC MARKET SEGMENTS BEHAVIOUR ............................ 99
  2.13.1 General information ................................................................................................................ 99
  2.13.2 Components involved .......................................................................................................... 102
  2.13.3 Experimentation ..................................................................................................................... 104
  2.13.4 Next steps ............................................................................................................................. 105
2.14 PIPELINES SUMMARY .................................................................................................................. 105

3 CONCLUSIONS AND FUTURE WORK ............................................................................................ 108

APPENDIX A PIPELINES DESCRIPTION TEMPLATE ........................................................................ 109
  A.1 PIPELINE NAME .......................................................................................................................... 109
  A.1.1 General information ................................................................................................................. 109
  A.1.2 Interfaces .................................................................................................................................. 109
Table of Figures

Figure 1: WORK PACKAGES AND THEIR ROLES IN DATABio .................................................. 3
Figure 2: DEPLOYMENT ARCHIMATE DIAGRAM FOR THE METADATA, LINKED DATA AND GRAPH DATA PIPELINE ........................................ 22
Figure 3: SCREENSHOT OF THE BETA VERSION OF THE DEVELOPED MAP VIEWER WITH INTEGRATED DISCOVERY MECHANISM FOR SATELLITE IMAGES FROM COPERNICUS SCiHUB API ......................................................................... 26
Figure 4: PILOT 1.2.1.A1.1 - ROLE VIEW .................................................................................. 28
Figure 5: PILOT 1.2.1.A1.1 - PIPELINE VIEW ........................................................................ 29
Figure 6: PILOT 1.2.1.A1.1 - LIFECYCLE VIEW ....................................................................... 29
Figure 7: GAIABus’ USER INTERFACE FOR REAL-TIME SENSORIAL DATA VISUALIZATION ......................................................................................... 30
Figure 8: PROTON’S USER INTERFACE FOR HIGHLY INFORMATIVE INFORMATION VISUALIZATION ................................................................... 31
Figure 9: PILOT 1.2.1.A1.1 - RESULTS FOR THE OLIVE TREES DISEASE ALERTS .................. 35
Figure 10: PILOT 1.2.1.A1.1 - RESULTS FOR THE GRAPE DISEASE ALERTS ......................... 35
Figure 11: PILOT C2.2 - PIPELINE VIEW ................................................................................ 38
Figure 12: PILOT C2.2 - LIFECYCLE VIEW ............................................................................ 38
Figure 13: PILOT C2.2 - PHYSICAL VIEW ............................................................................... 39
Figure 14: WEB-BASED UI OFFERED BY NP, PROVIDING SPATIAL AGGREGATIONS AND ATTRIBUTE-BASED COLOUR CODING .................. 40
Figure 15: PRELIMINARY STUDY AREA COVERING A TOTAL 282.600HA (53.580HA AGRICULTURAL AREA). THE IMAGE HIGHLIGHTS CORRECTLY (GREEN COLOURED PARCELS) AND INCORRECTLY (RED COLOURED PARCELS) IDENTIFIED PARCELS .................................................. 43
Figure 16: PRELIMINARY STUDY RESULTS FOR PARCEL CLASSIFICATION .......................... 44
Figure 17: CSEM’S C31.01 PRELIMINARY STUDY ON PEACHES. THE IMAGE HIGHLIGHTS PEACH FIELDS USED FOR TRAINING (YELLOW COLOURED), PEACH FIELDS USED FOR TESTING (GREEN COLOURED, TRUE POSITIVE SAMPLES) AND WHEAT FIELDS USED FOR TESTING (RED COLOURED, TRUE NEGATIVE SAMPLE) ................................................. 44
Figure 18: PILOT B1.1 - PIPELINE VIEW .............................................................................. 46
Figure 19: PILOT B1.1 - LIFECYCLE VIEW ............................................................................. 46
Figure 20: INITIAL DATA SET FORMAT FOR FIELD DATA ANALYSIS AND REAL-TIME ALERTING FOR PRECISE AGRICULTURE .................. 49
Figure 21: PILOT B1.4 - PIPELINE VIEW .............................................................................. 52
Figure 22: PILOT B1.4 - INTERFACE VIEW .......................................................................... 52
Figure 23: PROTON’S USER INTERFACE FOR INPUT/OUTPUT EVENT VISUALIZATION .......... 54
Figure 24: SENSLog REST API FOR INSERTING ALARMS .......................................................... 55
Figure 25: PILOT 1.3.1.B1.4 - ROLE VIEW ................................................................................ 56
Figure 26: PILOT 1.3.1.B1.4 - PIPELINE VIEW ....................................................................... 56
Figure 27: PILOT 1.3.1.B1.4 - LIFECYCLE VIEW ................................................................. 56
Figure 28: PILOT 1.3.1.B1.4 - SPARQL ENDPOINT FOR LINKED DATA ................................. 66
Figure 29: PILOT 1.3.1.B1.4 - FACETED SEARCH ENDPOINT FOR LINKED DATA ................. 57
Figure 30: PILOT 1.3.1.B1.4 - MAP VISUALIZATION OF LINKED DATA IN HS LAYERS-NG ...... 67
Figure 31: PILOT 1.3.1.B1.4 - MAP VISUALIZATION OF LINKED DATA IN METAPHACTORY ... 67
Figure 32: PILOT 1.3.1.B1.4 - PIPELINE VIEW ....................................................................... 69
Figure 33: ANALYZED DATA AND PREDICTION WITH FUTURE VARIANCE SCREENSHOT ....... 70
Figure 34: PILOT 2.4.2 - PIPELINE VIEW ............................................................................. 72
Figure 35: PILOT 2.4.2 - LIFECYCLE VIEW ............................................................................ 72
Figure 36: PILOTS 2.2.1 AND 2.2.2 - PIPELINE VIEW ........................................................... 75
Figure 37: PILOTS 2.2.1 AND 2.2.2 - LIFECYCLE VIEW ........................................................ 75
Figure 38: WUUDIS - INTERFACE VIEW ................................................................................ 76
Figure 39: PILOT 3.2.2 - PIPELINE VIEW ............................................................................. 79
Figure 40: PILOT 3.2.2 – LOCATION VIEW TRIAL 1 ............................................................. 80
List of Tables

<p>| Table 1: THE DATAIO CONSORTIUM PARTNERS | .......................................................... | 13 |
| Table 2: METADATA, LINKED DATA AND GRAPH DATA PIPELINE - PROVIDED INTERFACES | .......................................................... | 23 |
| Table 3: METADATA, LINKED DATA AND GRAPH DATA PIPELINE - CONSUMED INTERFACES | .......................................................... | 23 |
| Table 4: PILOT 1.2.1.A1.1 – EXPERIMENTATION RESULTS ANALYSIS | .......................................................... | 36 |
| Table 5: EXPERIMENTATION RESULTS STATION 1040 | .......................................................... | 57 |
| Table 6: EXPERIMENTATION RESULTS STATION 1012 | .......................................................... | 57 |
| Table 7: LINKED DATA PUBLICATION PIPELINE - PROVIDED INTERFACES | .......................................................... | 61 |
| Table 8: LINKED DATA PUBLICATION PIPELINE - CONSUMED INTERFACES | .......................................................... | 61 |
| Table 9: EU DATASETS | .......................................................... | 65 |
| Table 10: NATIONAL DATASETS (CZECH) | .......................................................... | 65 |
| Table 11: FARM DATA | .......................................................... | 66 |
| Table 12: WUUDIS INTERFACES | .......................................................... | 76 |
| Table 13: PILOT 3.2.2 - COLLECTED DATA ONBOARD A SINGLE VESSEL | .......................................................... | 82 |
| Table 14: WP4 PIPELINES SUMMARY | .......................................................... | 106 |</p>
<table>
<thead>
<tr>
<th>Acronym/Abbreviation</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>ATOM</td>
<td>ATOM (Syndication Format)</td>
</tr>
<tr>
<td>CAP</td>
<td>Common Agriculture Policy</td>
</tr>
<tr>
<td>CEP</td>
<td>Complex event processing</td>
</tr>
<tr>
<td>CKAN</td>
<td>Comprehensive Kerbal Archive Network</td>
</tr>
<tr>
<td>CSW</td>
<td>Catalogue Service for Web</td>
</tr>
<tr>
<td>DCAT</td>
<td>Data Catalog Vocabulary</td>
</tr>
<tr>
<td>DDS</td>
<td>Data Distribution System</td>
</tr>
<tr>
<td>EO</td>
<td>Earth Observation</td>
</tr>
<tr>
<td>FTP</td>
<td>File Transfer Protocol</td>
</tr>
<tr>
<td>GEMET</td>
<td>GEneral Multilingual Environmental Thesaurus</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>INSPIRE</td>
<td>INfrastructure for SPatial InfoRmation in Europe</td>
</tr>
<tr>
<td>ISO</td>
<td>International Organization for Standardization</td>
</tr>
<tr>
<td>JSON</td>
<td>JavaScript Object Notation</td>
</tr>
<tr>
<td>KML</td>
<td>Keyhole Markup Language</td>
</tr>
<tr>
<td>NG</td>
<td>Next Generation</td>
</tr>
<tr>
<td>NN</td>
<td>Nearest Neighbors</td>
</tr>
<tr>
<td>Paas</td>
<td>Platform as a Service</td>
</tr>
<tr>
<td>ODBC</td>
<td>Open Database Connectivity</td>
</tr>
<tr>
<td>OGC</td>
<td>Open Geospatial Consortium</td>
</tr>
<tr>
<td>OLU</td>
<td>Open Land Use</td>
</tr>
<tr>
<td>OTM</td>
<td>Open Transport Map</td>
</tr>
<tr>
<td>PROTON</td>
<td>IBM PROactive Technology ONline</td>
</tr>
<tr>
<td>RDF</td>
<td>Rich Description Framework</td>
</tr>
<tr>
<td>REST</td>
<td>REpresentational State Transfer</td>
</tr>
<tr>
<td>Saas</td>
<td>Software as a Service</td>
</tr>
<tr>
<td>STIM</td>
<td>Smart Transducer Interface Module (from IEEE standard)</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>TCP</td>
<td>Transmission Control Protocol</td>
</tr>
<tr>
<td>UDP</td>
<td>User Datagram Protocol</td>
</tr>
<tr>
<td>UI</td>
<td>User Interface</td>
</tr>
<tr>
<td>URL</td>
<td>Universal Resource Locator</td>
</tr>
<tr>
<td>WFS</td>
<td>Web Feature Service</td>
</tr>
<tr>
<td>WMS</td>
<td>Web Map Service</td>
</tr>
<tr>
<td>WMTS</td>
<td>Web Map tile service</td>
</tr>
</tbody>
</table>

Definitions, Acronyms and Abbreviations
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTZ</td>
<td>Warning time horizon</td>
</tr>
<tr>
<td>WP</td>
<td>Work Package</td>
</tr>
<tr>
<td>XML</td>
<td>eXtensible Markup Language</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Project summary

The data intensive target sector selected for the DataBio project is the **Data-Driven Bioeconomy**. DataBio focuses 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 output food, energy and biomaterials, also taking into account various responsibility and sustainability issues.

DataBio will deploy state-of-the-art big data technologies and existing partners’ infrastructure and solutions, linked together through the **DataBio Platform**. These will aggregate Big Data from the three identified sectors (**agriculture, forestry and fishery**), intelligently process them and allow the three sectors to selectively utilize numerous platform components, according to their requirements. The execution will be through continuous cooperation of end user and technology provider companies, bioeconomy and technology research institutes, and stakeholders from the big data value PPP programme.

DataBio is driven by the development, use and evaluation of a large number of **pilots** in the 3 identified sectors, where also associated partners and additional stakeholders are involved. The selected pilot concepts will be transformed to pilot implementations utilizing co-innovative methods and tools. The pilots select and utilize the best suitable market ready or almost market ready ICT, Big Data and Earth Observation methods, technologies, tools and services to be integrated to the common DataBio Platform.

Based on the pilot results and the new DataBio Platform, new solutions and new business opportunities are expected to emerge. DataBio will organize a series of trainings and hackathons to support its take-up and to enable developers outside the consortium to design and develop new tools, services and applications based on and for the DataBio Platform.

The DataBio consortium is listed in Table 1. For more information about the project see [www.databio.eu](http://www.databio.eu).

**Table 1: The DataBio consortium partners**

<table>
<thead>
<tr>
<th>Number</th>
<th>Name</th>
<th>Short name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (CO)</td>
<td>INTRASOFT INTERNATIONAL SA</td>
<td>INTRASOFT</td>
<td>Belgium</td>
</tr>
<tr>
<td>2</td>
<td>LESPROJEKT SLUZBY SRO</td>
<td>LESPRO</td>
<td>Czech Republic</td>
</tr>
<tr>
<td>3</td>
<td>ZAPADOCESKA UNIVERZITA V PLZNI</td>
<td>UWB</td>
<td>Czech Republic</td>
</tr>
<tr>
<td></td>
<td>Organization Name</td>
<td>Country</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------------------------------------------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Fraunhofer GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V.</td>
<td>Germany</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ATOS SPAIN SA</td>
<td>Spain</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>STIFTELSEN SINTEF</td>
<td>Norway</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>SPACEBEL SA</td>
<td>Belgium</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>VLAAMSE INSTELLING VOOR TECHNOLOGISCH ONDERZOEK N.V.</td>
<td>Belgium</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>INSTYTUT CHEMII BIOORGANICZNEJ POLSKIEJ AKADEMII NAUK</td>
<td>Poland</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>CIAOTECH Srl</td>
<td>Italy</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>EMPRESA DE TRANSFORMACION AGRARIA SA</td>
<td>Spain</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>INSTITUT FUR ANGEWANDTE INFORMATIK (INFAI) EV</td>
<td>Germany</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>NEUROPUBLIC AE PLIROFORIKIS &amp; EPIKOINNONION</td>
<td>Greece</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Ústav pro hospodářskou úpravu lesů Brandýs nad Labem</td>
<td>Czech Republic</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>INNOVATION ENGINEERING SRL</td>
<td>Italy</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Teknologian tutkimuskeskus VTT Oy</td>
<td>Finland</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>SINTEF FISKERI OG HAVBRUK AS</td>
<td>Norway</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>SUOMEN METSAKESKUS-FINLANDS SKOGSCENTRAL</td>
<td>Finland</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>IBM ISRAEL - SCIENCE AND TECHNOLOGY LTD</td>
<td>Israel</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>MHG SYSTEMS OY - MHGS</td>
<td>Finland</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>NB ADVIES BV</td>
<td>Netherlands</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>CONSIGLIO PER LA RICERCA IN AGRICOLTURA E L'ANALISI DELL'ECONOMIAAGRARIA</td>
<td>Italy</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>FUNDACION AZTI - AZTI FUNDAZIOA</td>
<td>Spain</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>KINGS BAY AS</td>
<td>Norway</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>EROS AS</td>
<td>Norway</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>ERVIK &amp; SAEVIK AS</td>
<td>Norway</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>LIEGRUPPEN FISKERI AS</td>
<td>Norway</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>E-GEOS SPA</td>
<td>Italy</td>
<td></td>
</tr>
<tr>
<td>No.</td>
<td>Organization Name</td>
<td>Contribution</td>
<td>Country</td>
</tr>
<tr>
<td>-----</td>
<td>-----------------------------------------------------------------------------------</td>
<td>--------------</td>
<td>----------</td>
</tr>
<tr>
<td>29</td>
<td>DANMARKS TEKNISKE UNIVERSITET</td>
<td>DTU</td>
<td>Denmark</td>
</tr>
<tr>
<td>30</td>
<td>FEDERUNACOMA SRL UNIPERSONALE</td>
<td>Federu</td>
<td>Italy</td>
</tr>
<tr>
<td>31</td>
<td>CSEM CENTRE SUISSE D'ELECTRONIQUE ET DE MICROTECHNIQUE SA - RECHERCHE ET DEVELOPPEMENT</td>
<td>CSEM</td>
<td>Switzerland</td>
</tr>
<tr>
<td>32</td>
<td>UNIVERSITAET ST. GALLEN</td>
<td>UStG</td>
<td>Switzerland</td>
</tr>
<tr>
<td>33</td>
<td>NORGES SILDESALGSLAG SA</td>
<td>Sildes</td>
<td>Norway</td>
</tr>
<tr>
<td>34</td>
<td>EXUS SOFTWARE LTD</td>
<td>EXUS</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>35</td>
<td>CYBERNETICA AS</td>
<td>CYBER</td>
<td>Estonia</td>
</tr>
<tr>
<td>36</td>
<td>GAIA EPICHEIREIN ANONYMI ETAIREIA PSIFIAKON YPIRESION</td>
<td>GAIA</td>
<td>Greece</td>
</tr>
<tr>
<td>37</td>
<td>SOFTEAM</td>
<td>Softeam</td>
<td>France</td>
</tr>
<tr>
<td>38</td>
<td>FUNDACION CITOLIVA, CENTRO DE INNOVACION Y TECNOLOGIA DEL OLIVAR Y DEL ACEITE</td>
<td>CITOLIVA</td>
<td>Spain</td>
</tr>
<tr>
<td>39</td>
<td>TERRASIGNA SRL</td>
<td>TerraS</td>
<td>Romania</td>
</tr>
<tr>
<td>40</td>
<td>ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS</td>
<td>CERTH</td>
<td>Greece</td>
</tr>
<tr>
<td>41</td>
<td>METEOROLOGICAL AND ENVIRONMENTAL EARTH OBSERVATION SRL</td>
<td>MEEO</td>
<td>Italy</td>
</tr>
<tr>
<td>42</td>
<td>ECHEBASTAR FLEET SOCIEDAD LIMITADA</td>
<td>ECHEBF</td>
<td>Spain</td>
</tr>
<tr>
<td>43</td>
<td>NOVAMONT SPA</td>
<td>Novam</td>
<td>Italy</td>
</tr>
<tr>
<td>44</td>
<td>SENOP OY</td>
<td>Senop</td>
<td>Finland</td>
</tr>
<tr>
<td>45</td>
<td>UNIVERSIDAD DEL PAIS VASCO/ EUSKAL HERRIKO UNIBERTSITATEA</td>
<td>EHU/UPV</td>
<td>Spain</td>
</tr>
<tr>
<td>46</td>
<td>OPEN GEOSPATIAL CONSORTIUM (EUROPE) LIMITED LBG</td>
<td>OGCE</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>47</td>
<td>ZETOR TRACTORS AS</td>
<td>ZETOR</td>
<td>Czech Republic</td>
</tr>
<tr>
<td>48</td>
<td>COOPERATIVA AGRICOLA CESENATE SOCIETA COOPERATIVA AGRICOLA</td>
<td>CAC</td>
<td>Italy</td>
</tr>
</tbody>
</table>
1.2 Document purpose and scope

DataBio proposes a big data platform, the DataBio platform, for the bioeconomy industry in the domains of agriculture, forestry, and fishery. Towards this end, in month 17 (M17) of the project the first two technical deliverables of work package (WP) 4 “DataBio Platform with Pilot Support” in which the DataBio platform is setup were published – Deliverable 4.1 (D4.1) “Platform and interfaces” and D4.2 “Services for tests”. Specifically, D4.2 aims at providing a comprehensive overview of the different services and pipelines (see definitions below) that have been identified in the DataBio platform to support specific requirements of the pilots in the project.

In the scope of the project, we understand the concept of a platform in a strict technical sense as a software development platform referring to an environment in which software is developed to be deployed in hardware, operating system, middleware (e.g. a virtual machine), or a cloud. More specifically, we focus on Big Data platforms, i.e., that deal with Big Data (high volume, high velocity and high variety). Within this scope, the DataBio platform (hereinafter referred to as platform) provides a big data toolset and framework, which offers functionalities primarily for the domains of agriculture, forestry, and fishery. The framework supports the forming of reusable and deployable pipelines of interoperable components thus extending the impact of DataBio to new bioeconomy projects as well as to other business areas.

At the heart of the platform, we identify pipelines as a set of components with clear mutual interfaces linking them together and to the platform environment. A pipeline fulfils pilot functionalities that cannot be supported by a single component. A pipeline can be seen as a white box showing internal wiring of the components in the pipeline, thus providing technical guidance for configuration and deployment. Platform services (hereinafter referred to as services), can be seen as providers of functionalities to users that typically need to know the usability of the service, but do not need to understand the inner wiring, inner components, nor where the service is deployed. These services are typically accessed via standardized interfaces like application programming interfaces, e.g. web services or library interfaces, interactive user interfaces, standard data transfer, or remote call protocols. Services refer (often) to end points that are “black box” activated remotely and executed in the cloud. Note that components (like PROTON) can be both an end-point service and a part of a pipeline. In terms of deployment, containers, virtual machines, and web services that use the pipelines make the deployment in alternative clouds straightforward. When pipelines are deployed they can become services.

Without the loss of generality, we extend the definition of pipelines to cover services, thus referring to both pipelines and services in this report as pipelines, as we mainly focus on...
providing relevant information regarding the inner components wiring and interfaces, as well as instructions for their deployment.

As aforementioned, pipelines are intended to meet key (business) pilot requirements or target a common pilot goal that cannot be fulfilled by a single component. Usually, one component does not provide all required functionalities by a pilot and, therefore, there is a need to orchestrate some components that together fulfil specific functionalities. The pipelines are the heart of the platform as they are the cornerstones of the pilot experimentations. A higher-level goal of the pipelines is the identification of commonalities among different pipelines in terms of components and interfaces that could lead to the definition of “design patterns” or “exploitable assets” that could be applied to multiple pilots and even to other business domains.

D4.2 focuses on the pipelines that have been identified in the project so far. D4.2 complements the main report on the platform (D4.1) with the pipelines descriptions. The goal is to provide technical information and guidelines for the pilots’ experimentation and readiness for trial 1. Note that the dissemination level of this report is Confidential (CO), that is, only for members of the consortium (including the Commission Services), while D4.1 is a public document. With this in mind, it has been decided that D4.1 should be the main document containing all aspects of the platform, while D4.2 will complement it with the pipelines descriptions.

Naturally, the level of granularity and maturity of each pipeline varies among the different pilots as they depend on several factors, including:

- Maturity of the pilot(s) involved
- Maturity of the components/existing technologies comprising the pipeline
- Complexity of the pipeline (number of components, data sets, interfaces)
- Development required by the different components to support the pipeline
- Effort required to deploy the pipeline to be tested with real data
- Environmental requirements, such as weather and seasonal conditions (spring, summer, autumn, winter) can also impact the pipeline’s status, as they might influence the experimentation schedule in the relevant domains of agriculture, forestry, and fishery.

This report provides the pipelines overall picture in M16 of the project. Eventually, the pipelines will evolve during the course of the project to provide the full capabilities required for their full deployment.

The goal of this report is to bring the necessary technical details to enable the experimentation of the pipeline, that is, from the pipeline design to its deployment and experimentation. For details regarding a comprehensive description of the platform please refer to D4.1. For information regarding background or motivation of the different pilots refer to WP1, WP2, and WP3 deliverables (see section 1.4 for specific documents references).
1.3 Phases in our work

In essence, as pipelines are targeted to meet some (business) requirements, they are the outcome of collaborations between technical people (component owners in the platform) and business-oriented people (pilot owners).

As the number of components in the platform is large, the work towards the identification of the pipelines was progressively continuing. The process included as the first step the identification of components that can fulfil some of the pilot requirements. We called this process a matchmaking process. Matchmaking was a two-way process: component providers have expressed their interest for supporting pilots and pilots have expressed their interest for the different components according to specific features/capabilities. As a later step, the orchestration of some components into pipelines took place.

As the technical deliverables were due only in M17, to mitigate any potential risk, it has been decided that two interim documents (dubbed D4i.1 and D4i.2) would be internally submitted to assure the technical deliverables of the project are submitted in time and as high-quality outputs. Di.1 “Component and data set descriptions” internally submitted in October 2017 provided a general description of the platform with emphasis on pilot requirements and on the matchmaking process; while Di.2 “Full platform description for trial 1” internally submitted in February 2018 relied on Di.1 and basically constituted the very first versions of D4.1 and D4.2. As part of this effort, definitions of some central concepts of the platform were defined by a dedicated task force comprising of technical leading members in the project. The aim was to create a common understanding of core terms and concepts in the project, such as DataBio platform, pipelines, and services.

In an attempt to harmonize, as much as possible, the work for the pipelines definition, a template was articulated and distributed among the pipelines collaborators. Chapter 2 gathers all the pipeline descriptions. The template itself is presented in Appendix A.

As per today, we have a total of 13 pipelines to be tested for the first phase of experimentation. We expect this number to increase for the second phase of experimentation.

1.4 Relation with other documents

D4.2 complements the platform and interfaces report (D4.1) and therefore there is a tight connection between these two documents.

As pipelines are comprised of different components of the platform, there may be some overlapping in the pipelines with components belonging to WP5 - Earth Observation and GeoSpatial Data and Services. Therefore, this report is also related to the two deliverables to be submitted as well as M17 in WP5: D5.2 “EO Component and interfaces” and D5.3 EO Services and tools”. See the Executive Summary for an overview of the different platform-related deliverables in the project and their interrelationships.
Finally, for information regarding the pilots refer to following deliverables submitted at M6: D1.1 “Agriculture pilot definition”; D2.1 “Forestry pilot definition”; and D3.1 “Fishery pilot definition”.

1.5 Document structure

This document is comprised of the following chapters:

Chapter 1 presents an overall introduction to the document, including the purpose and scope of the document, phases in our work, and relation with other documents.

Chapter 2 describes the pipelines following the structure of a predefined template. The descriptions are listed by bioeconomy sectors and work packages, starting from agriculture (WP1), Forestry (WP2), and Fishery (WP3).

Chapter 3 summarizes the work that was carried out and presents the future work to be carried out.

Appendix A presents the template for pipeline description.

---

2 DataBio pipelines

This section is the core part of D4.2 as it introduces the DataBio pipelines identified and developed so far in the project. As aforementioned in the introductory chapter, the pipelines vary in terms of maturity and progress, as the pilots themselves are inherent different each from the other and are influenced by many parameters, including different functional requirements, complexity of scenarios, maturity of technical components and level of development needed to fulfil the requested functionality, and environmental factors. Notwithstanding, an attempt to harmonize among the different descriptions has been made by the circulation of a template to be filled out by the pipelines collaborators.

Appendix A presents the template that was circulated along with instructions for each section. In addition, a complete populated template for one of the pipelines was distributed to serve as reference. The goal is to give enough information for trial 1 phase in the project and the ongoing development of the pipelines.

Template structure:

- Pipeline name – A name fitting the pipeline objectives. Each name is pre-fixed with the WP number (i.e., WP1, WP2, and WP3) to ease the attribution of the pipeline to the specific bioeconomy domain.

- General information – This section includes sub-sections that describe the objectives of the pipelines (what specific needs does the pipeline address?); diagrams and views from Modelio and DataBio Hub; the pilot(s) the pipelines is related to; and the level of reusability of the pipeline. Note that Modelio has been selected for the modelling tool in the project. DataBio Hub is a common communication tool that stores all platform related entities for easy search and tracking (Note that you must be logged in as a member of DataBio Pipelines group to view the pipelines). For more details on both tools please refer to D4.1. As the pipelines usually describe the data flow and the process in the specific scenario, common views in Modelio are pipeline, role, and lifecycle (for explanation on the different Modelio views refer to D4.1). Reusability is one important aspect of the pipeline as it has a direct implication on potential exploitation of the pipeline setup. Some of the pipelines are tailored to specific pilot needs and rely on domain specific tools (e.g., Shared multiuser forest data environment pipeline), while some of the pipelines can be used in different scenarios and even in other domains besides the ones in the project (e.g., Field data analysis and real-time alerting for decision making for precise agriculture pipeline, that relies on two generic tools: IoT Hub and PROTON). The goal is to identify potential reusable sets of components among the pilots as some form of “design patterns” that can be exploited by the different partners. For example, in the “Data analytics for prediction of pelagic market segments behaviour” pipeline, the core of this pipeline is intended to be common to all the four pilots in small pelagic fishery, e.g. Pilots 3.2.2, 3.3.2, 3.4.1 and 3.4.2, that all share common infrastructure (logging system, data centre) and have similar needs for data management, processing, analysis and visualization. The data
acquisition, management and serving will share a common pipeline, while each pilot has unique requirements for certain datasets, analysis techniques and display methods. However, the framework for analysis and visualization is shared, although some datasets, algorithms and display methods may differ.

- Interfaces to the external world
- Components involved – names and links to DataBio Hub of the components that constitute the pipeline, as well as the interfaces defined between the components in the pipeline.
- Experimentation – Testing carried out, including deployment of the components, data sets used, and results from experimenting with the pipeline. Again, there are some pilots that have already tested real data in their experiments (e.g., Prediction and real-time alerts of diseases and pest breakouts in crops pipeline) while others are in earlier phases of data acquisition (e.g., Data acquisition and analyses for fishing operational efficiency).
- Next steps – the experimentation phase has just started, and it is crucial that each pipeline has a work plan with regards to its next steps to bring the pipeline into operation.

2.1 WP1 - Metadata, linked data and graph data

2.1.1 General information

One of the DataBio project challenges lies in (Big) metadata integration. Within this scope, there are various metadata standards being handled within various communities; such as sensor networks, semantic services, Earth Observation data and other geospatial data, open data, etc. Another perspective brings the granularity of resources: different metadata appear on a dataset level than on a level of a (Web) service or image tile.

All kinds of metadata, linked and graph data should be handled homogenously and in a user-friendly way for all users.

2.1.1.1 Objectives

We may identify several goals of the Metadata, Linked Data and Graph Data Pipeline. These are, at least:

- to integrate metadata from various standards, i.e. defined through various schemas;
- to offer simple to use graphical user interface that allows
  - to search for any resource according to several criteria such as theme, location, time etc.,
  - to handle similarly various kinds of resources, no matter whether it is a dataset or a mobile sensor,
  - to perform analyses on the top of metadata to support more efficient (Big) data management,
  - to (re)use and design features of Semantic Web.
2.1.1.2 Diagnostics and views

The pipeline is, so far due to the ongoing development, not a physical part of the DataBio Hub. The ArchiMate diagram in Figure 2 presents the deployment viewpoint.

![Diagram](image)

**Figure 2: Deployment ArchiMate diagram for the metadata, linked data and graph data pipeline**

2.1.1.3 Associated pilot

The associated WP1 Pilot [B2.1] Machinery management is focused mainly on collecting telemetry data from machinery and analysing them in relation with other farm data. The main challenge is an access to data and data integration, when farmer uses tractors and equipment from various manufacturers with different telematics solutions and different data ownership/sharing policy.

URL to the DataBio hub:

https://www.databiohub.eu/registry/#service-view/WP%201%20Pilot%209%20[B2.1]%20Machinery%20management

2.1.1.4 Reusability

The pipeline has been developed according to the standards in the domain of geographic information (systems and science). As such, it is scalable, transferable, and re-usable. The only condition for the deployment is to comply with the same standards as the Metadata, Linked Data and Graph Data Pipeline follows. In more specific terms, it means that the pipeline deployment is feasible when an interface for at least one of the following standards is supported:

- OGC Catalogue Service for Web 2.0.2, including the International Organization for Standardization (ISO) Application profile 1.0,
- OpenSearch,
- Comprehensive Kerbal Archive Network (CKAN), (Geo) Data Catalog Vocabulary (DCAT) respectively,
• ATOM (Syndication Format) interface in the INfrastucture for SPatial InfoRmation in Europe (INSPIRE) modification,
• Open Geospatial Consortium (OGC) Web Map Service (WMS) from version 1.0.0 to version 1.3.0, including the INSPIRE modification,
• OGC Web Feature Service from version 1.0.0 to version 2.0.0, including the INSPIRE modification and Gazetteer enhancement,

2.1.2 Interfaces
The Metadata, Linked Data and Graph Data Pipeline provides the following interfaces (see Table 2):

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF-Micka-GUI</td>
<td>A graphical user interface (GUI) for (Open)MicKA.</td>
</tr>
<tr>
<td>IF-Metadata-Editor-GUI</td>
<td>A user interface for Metadata Editor as a subcomponent of (Open)MicKA.</td>
</tr>
<tr>
<td>IF-OGC-CSW</td>
<td>An interface for a communication to other Catalogue Service for Web (CSW) servers. The interface follows the full implementation of the OGC CSW 2.0.2 implementation specification including International Organization for Standardization (ISO) Application Profile 1.0.</td>
</tr>
<tr>
<td>IF-OpenSearch</td>
<td>An interface designed and developed to publish of search results in a format suitable for syndication and aggregation.</td>
</tr>
</tbody>
</table>

The Metadata, Linked Data and Graph Data Pipeline consumes the following interfaces (see Table 3):

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF-INSPIRE-ATOM</td>
<td>An object-oriented interface for a communication according to the ATOM Syndication format, in a version as defined in the INSPIRE Technical Guidelines for Download services.</td>
</tr>
<tr>
<td>IF-OGC-WMS</td>
<td>An interface defined according to the OGC WMS 1.0.0 – 1.3.0 implementation specifications (including INSPIRE modifications) in order to automatically create metadata from a WMS instance.</td>
</tr>
<tr>
<td>IF-GEMET-Rich Description Framework (RDF)</td>
<td>An interface designed and developed for a communication to the GEneral Multilingual Environmental Thesaurus (GEMET). The implementation follows guidelines for GEMET API as described under the URL: <a href="http://www.eionet.europa.eu/gemet/en/webservices/">http://www.eionet.europa.eu/gemet/en/webservices/</a>.</td>
</tr>
<tr>
<td>IF-OGC-WFS</td>
<td>An interface defined according to the OGC WMS 1.0.0 – 2.0.0 implementation specifications (including INSPIRE modifications) to</td>
</tr>
<tr>
<td>ID</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>automatically create metadata from a Web Feature Service (WFS) instance.</td>
</tr>
<tr>
<td>IF-OGC-CSW</td>
<td>An interface for a communication to other CSW servers. The interface follows the full implementation of the OGC CSW 2.0.2 implementation specification including ISO Application Profile 1.0.</td>
</tr>
</tbody>
</table>

### 2.1.3 Components involved

#### 2.1.3.1 Names

Three components participate in the Metadata, Linked Data and Graph Data Pipeline:

- C02.02 Micka (together with its subcomponent Metadata Editor), URL: [https://www.databiohub.eu/registry/#components?name=Micka](https://www.databiohub.eu/registry/#components?name=Micka),
- C02.03 HSLayers Next Generation (NG), URL: [https://www.databiohub.eu/registry/#components?name=HSLayers%20NG](https://www.databiohub.eu/registry/#components?name=HSLayers%20NG),

#### 2.1.3.2 Interfaces between components

Metadata publication and visualization between the C02.02 Micka component on the one hand and the C02.03 HSLayers NG as well as the C02.04 HSLayers Mobile components on the other hand, is being realized through the following interfaces:

- OGC Catalogue Service for Web 2.0.2, including the ISO Application profile 1.0,
- OGC Web Map Service from version 1.0.0 to version 1.3.0, including the INSPIRE modification,
- REpresentational State Transfer (REST) API.

Communication between the C02.02 Micka component and the Metadata editor subcomponent is being realized through ODBC mechanisms.

### 2.1.4 Experimentation

The progress of the team of developers from LESPRO and UWB has been presented, discussed and awarded at several hackathons within the last year, e.g. the Prague hack ([https://inspire.ec.europa.eu/events/prague-inspire-hack-2018](https://inspire.ec.europa.eu/events/prague-inspire-hack-2018)) or the Open Geospatial Consortium (OGC)/INSPIRE Orléans hack ([http://www.plan4all.eu/orleans-ogc-inspire-hackathon-2018/](http://www.plan4all.eu/orleans-ogc-inspire-hackathon-2018/)).

The goal was to develop an easy-to-use satellite images discovery integrated directly in a map viewer. Such map viewer with discovery mechanisms was developed for desktop as well as mobile device environments. Relevant resources to a currently zoomed area appear directly through links to satellite programmes’ APIs, such as Copernicus’ Open Access Hub. The developed application will be based on OpenLayers 4-based framework called HSLayers NG (component C02.03 for desktop, C02.04 for mobile devices).
Within the scope of the DataBio project, the current status allows to store, discover and obtain all the relevant resources within the WP1 Pilot [B2.1] Machinery management. It means, that the following kinds of information are available through the Metadata, Linked Data and Graph Data Pipeline:

- Long-term geographic information
  - topographic as well as thematic maps,
  - satellite images,
- Dynamic geographic information
  - trajectories of farm machinery fleet,
  - sensor readings.

2.1.4.1 Deployment
The Metadata, Linked Data and Graph Data Pipeline is deployed at two physically as well as logically separate locations. The first location is the testing server at the LESPRO infrastructure (with publicly available URL http://geoportal.gov.cz/php/micka/csw/index.php), while the second location is within the PSNC infrastructure to verify the pipeline’s functionality in the cloud environment (with publicly available URL http://portal.foodie-cloud.org/php/metadata/csw/). The invocation of a service should follow the principles described in the OGC CSW implementation specification.

2.1.4.2 Data sets
No datasets were used during the development of the Metadata, Linked Data and Graph Data Pipeline as this pipeline is focused on metadata, i.e. derived information also from the dataset level. Inputs are in several types, ranging from local storage, databases to consumed APIs and/or Web services. The outputs comprise of metadata in an eXtensible Markup Language (XML), Rich Description Framework (RDF), JavaScript Object Notation (JSON), ATOM, or Keyhole Markup Language (KML) formats. Metadata on long-term geographic information counts in dozens of megabytes. Volume of dynamic geographic information depends on the number of monitored vehicles as well as farm size. For instance, one 1’000 hectares farm with 10 vehicles generates one megabyte of metadata from sensors and machinery measurements daily.

2.1.4.3 Results
The Metadata, Linked Data and Graph Data Pipeline proved as a valid approach to unify description of resources (data sets, services, sensor and machinery measurements) at one place. This allows a farmer to discover which particular data, services and/or functionality is available for his/her farm.

Moreover, a novel approach for data and metadata integration has been developed and validated. All forms of metadata were integrated together with data in one application, a map viewer. Such a map viewer also acts as a catalogue. A user doesn’t need to work in two
separate applications (map viewer and catalogue) but uses single integrated solution. A beta version of the developed application is depicted in Figure 3.

![Figure 3: Screenshot of the beta version of the developed map viewer with integrated discovery mechanism for satellite images from Copernicus SciHub API](image)

The beta version shown on Figure 3 accesses directly the Copernicus SciHub API, parses relevant metainformation from the obtained data, stores metadata in the GeoJSON format and visualizes metadata in a human friendly way. It means that a user can discover e.g. which satellite images are available, satellite data type (radar, multispectral etc.), see the percentage of cloudiness, and/or directly download the data.

The current results also demonstrate the capability to fulfil all user needs on metadata, linked and graph data within the WP1 Pilot [B2.1] Machinery management.

### 2.1.5 Next steps

The ongoing development aims at two functionality enhancements:

- Filtering capabilities enabling to a user to filter relevant resources during the discovery mechanism. In other words, a user will be able to interactively search for and download e.g. multispectral satellite images from the last month that have only up to 30% cloudiness.
• Notification capabilities to inform a user that e.g. a new satellite image is available. A user then may download the new satellite image at just one click or to make the notification dismissed.

2.2 WP1 – Prediction and real-time alerts of diseases and pests’ breakouts in crops

2.2.1 General information
The following pipeline description presents a data/information exchange pipeline that supports the activities of WP1’s pilot A1.1. It focuses on the advancements of two (2) components that are part of the DataBio technology platform, namely C13.02 GAIABus DataSmart RealTime Subcomponent (from partner NP) and C13.03 Proton (from partner IBM) and their interoperability with GAIA Cloud. As the main focus of this report is the technical aspect of the pipeline, it is worth mentioning that all development activities related to the expansion and technological evolution of NP’s GAIA Cloud will be described in respective WP1 deliverables. GAIA Cloud is a major horizontal building block of NP’s Gaiasense4 solution and constitutes of multiple cloud computing services, which collect, store and combine heterogeneous data to convert them into facts using advanced data analytics techniques.

2.2.1.1 Objectives
Smart farming is an approach of agricultural production management which allows (and is based on) the use of data for improved decision making regarding cultivation practices and the application of inputs, aiming at the optimisation of production both quantitatively and qualitatively. It is the outcome of the combination of technological tools with various data types, scientific research, knowledge and practical experience.

Smart farming involves the continuous monitoring and prediction of diseases and pests’ breakouts and alerting in case these are found. Farmers can be facilitated in the effective application of fertilizers by having an overview of the possibility of infection or infestation of their crop by a specific disease or pest respectively, based on the environmental conditions that apply in the specific area. Hence, they can schedule the application of sprayings at the most appropriate time, maximising their effectiveness and sometimes reducing the number of required applications.

This pipeline addresses the monitoring and prediction of specific disease and pest infection in olives, peaches, and grapes and the real-time triggering of alerts once these are detected.

The goal is to have an integrated solution in which we collect, validate and store data through NP’s GAIAtrons stations, perform initial processing, monitoring and cross-checking on the raw data though NP’s GAIABus DataSmart RealTime Subcomponent, and push the validated values to IBM PROactive Technology ONline (PROTON) for further analysis (temporal reasoning) to enable the triggering of early warnings/alerts in real-time.

4 http://www.gaiasense.gr.
PROTON is a complex event processing (CEP) tool thus performs real-time processing on events as they arrive to the system and alerts in case a situation is found (according to a pre-defined set of event rules). In this case, it enables receiving the already processed raw data by GAIABus DataSmart RealTime Subcomponent, apply further processing on the data in time windows (temporal reasoning) to derive a more (complex) situation or event emitted by the CEP system.

2.2.1.2 Diagrams and views

![Diagram of the system structure](image)

*Figure 4: Pilot 1.2.1.A1.1 - Role view*
Figure 5: Pilot 1.2.1.A1.1 - Pipeline view

Figure 6: Pilot 1.2.1.A1.1 - Lifecycle view
2.2.1.3 Associated pilot
Pilot 1.2.1.A1.1: Precision agriculture in olives, fruits, grapes.

URL in DataBio Hub:

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

2.2.1.4 Reusability
While PROTON is a generic tool, GAIABus DataSmart RealTime Subcomponent is designed specifically to handle Internet of Things (IoT) data streams from NP’s telemetry agri-sensors, called GAIAtrons, and perform a set of activities including real-time stream processing, monitoring, validation and cross-checking. Having said this, it is possible to reuse the same pipeline in a different agriculture scenario, assuming the rules, thresholds and scientific models for the calculations are known. This will eventually trigger new event rules in the existing event processing implementation or a complete new implementation.

2.2.2 Interfaces

2.2.2.1 GAIABus DataSmart RealTime Subcomponent
The subcomponent is able to communicate with GAIAtrons through a mixed TCP/UDP (Transmission Control Protocol/User Datagram Protocol) data communication scheme and JSON (JavaScript Object Notation)-enabled format. The subcomponent is designed for cloud-based operation and uses a custom data exchange syntax along 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. A web-based tool has been designed for data visualization, providing flexible graphs and UIs.

Figure 7: GAIABus’ user interface for real-time sensorial data visualization
2.2.2.2 **PROTON**

The outputs of the application are recorded into a CSV file for future testing and analysis. In addition, to have a graphical user interface for the outputs, we apply PROTON’s dashboard to show the input events as well as the situations or output events emitted by the engine in real-time. Figure 8 below shows a screenshot of the dashboard showing `DegreeDaysFirstGenerationWarning` and alarms for pest monitoring. Alarm events are denoted with a red exclamation mark, while warnings are denoted with a yellow exclamation mark. Figure 8 shows an example of a warning and an alarm shown in the dashboard and the reasons these have been detected.

![PROTON Dashboard Screenshot](image)

*Figure 8: PROTON’s user interface for highly informative information visualization*

2.2.3 **Components involved**

2.2.3.1 **Names**

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 from NP,
- GAIABus DataSmart RealTime Subcomponent (C13.03 - [https://www.databiohub.eu/registry/#service-view/GAIABus%20DataSmart%20RealTime%20Subcomponent/0.0.1](https://www.databiohub.eu/registry/#service-view/GAIABus%20DataSmart%20RealTime%20Subcomponent/0.0.1)) from NP,
- PROTON (C19.01 - [https://www.databiohub.eu/registry/#services?name=IBM](https://www.databiohub.eu/registry/#services?name=IBM)) from IBM

2.2.3.2 **Interfaces between components**

Within the pipeline, REST endpoints will be designed for allowing end-to-end communications and data/information exchange among all components.

- Events from GAIABus and GAIA Cloud to PROTON (see Figure 5):
2.2.4 Experimentation

GAIATron stations allow the recording of various atmospheric and soil parameters, such as temperature, humidity, solar irradiance, wind velocity and direction, rainfall etc., through their integrated sensors. Data collected from the GAIATron telemetric stations are transmitted to the Gaiasense network (GAIA Cloud), which is responsible not only for the collection, but for the processing of data and their storage at a central point, so that these data become accessible to all involved stakeholders. GAIABus DataSmart RealTime subcomponent is responsible for:

- Real-time monitoring of the incoming data streams (raw data)
- Real-time validation based on pre-defined configuration attributes (for example, if the type of the request is a valid JSON string, if it contains a specific Authentication Token, if it has specific type of fields)
- Real-time parsing and cross-checking the continuously incoming sensor measurements (checks if each value is within predefined ranges)
- Event driven triggering (when the server encounters an invalid value).

Then, the collected sensor data is pushed for further processing and analysis in GAIA Cloud’s SmartFarm services for the iterative evolution, continuous training and exploitation of tailored scientific models, adapted to the specific characteristics of the selected crop types (olive trees, grapes and peaches) and microclimates. As already mentioned, more details
about GAIA Cloud and its associated SmartFarm services are provided in WP1 deliverables, as this description focuses on components which constitute part of the DataBio technology platform.

The extracted numerical output of GAIA Cloud’s SmartFarm data processing (risk factor associated with diseases and pests breakouts) is then passed to PROTON for more sophisticated analysis on it (applying temporal reasoning) and real time alerting of pests and diseases in the analysed crops.

In our current application, PROTON applies the following monitoring rules, exploiting crop-related scientific models offered by NP:

- **For olive trees**
  - Monitor and predict disease (*Spilocaea oleaginea*) breakout among the crop
    - Report when infection percentage is over a critical threshold
    - Report when infection percentage is continuously increasing, with at least 5 such readings, within 24 hours and on the way to reach critical threshold
  - Monitor and predict pest (*Bactocera olea*) infection among the crop
    - Report when the degree days sum crosses specified thresholds (alarms)
    - Report when the degree days sum is steadily increasing and coming close to alarm thresholds (warnings)

- **For grapes**
  - Monitor and predict disease (downy mildew) breakout among the crop
    - Report when infection percentage is over a critical threshold
    - Report when infection percentage is continuously and significantly increasing (with 1.3 ratio of increase between readings), with at least 3 such readings, within 72 hours and on the way to reach critical threshold
  - Monitor and predict pest (*Lobesia botrana*) infection among the crop
    - Report when the degree days sum crosses specified thresholds (alarms)
    - Report when the degree days sum is steadily increasing and coming close to alarm thresholds (warnings)

- **For peaches**
  - Monitor and predict pest (*Grapholita molesta*) infection among the crop
    - Report when the degree days sum crosses specified thresholds (alarms)
Report when the degree days sum is steadily increasing and coming close to alarm thresholds (warnings)

2.2.4.1 Deployment
NP has built and operates its own data centre offering cloud services at platform level (PaaS) and software level (SaaS). GAIA Cloud supports NP’s pilot activities at various levels and serves as the backbone of the pilots’ pipelines, in terms of infrastructure and applications/services.

NP’s GAIABus RealTime subcomponent is part of NP’s GAIA Cloud. In order to support these functionalities, for its deployment the GAIABus DataSmart RealTime Subcomponent exploits a set of technologies comprising of Javascript, Sockets, NodeJS, MySQL, RESTful endpoints.

2.2.4.2 Data sets
The input data consisted of 2 CSV files for each of the 3 crops:

- Hourly processed data (pre-calculated infection percentage) for disease monitoring and prediction
- Daily processed data (pre-calculation of degree days summation) for pest infection monitoring

For hourly sensor data, around 2 weeks of data for each crop, for daily sensor data around 1 month of data.

2.2.4.3 Results
In general, for all the crops, PROTON was able to emit preliminary alerts that precede “real” (based on the exploited scientific models) alerts, thus enabling a timely reaction to the situation. For example, Figure 9 depicts the results for the olive trees crop. All situations/alarms emitted by PROTON (annotated in blue on the graph) happen before the “real” alerts occur (in red on the graph). Figure 10 depicts the results for the grapes disease alerts. Again, “warnings” emitted by PROTON precede the “alerts” thus giving sometime for corrective actions.
Table 4 depicts the results from the described experimentation. It’s important to note that **Alarms** are emitted when the monitored condition passes critical threshold value, while **Warnings** are emitted when there is an increasing trend over time in the monitored condition which grows closer and closer to critical threshold (but still there is “sometime” for corrective actions before an alarm is triggered). In order to assess the potential time saving from applying these warnings we calculate a **Warning time horizon** (WTZ).

WTZs are calculated as the difference between the first warning event and the alarm event. In other words, this is the elapsed time between a detection of a potential problem until the latter is actually detected (alarm), therefore allowing for proactive actions to take place before the monitored condition reaches critical values.

The warning time horizon is adjustable as they depend on the temporal intervals and thresholds used in the monitoring application (in short time windows the potential gain cannot be large). Those can be adjusted to give more precise results/longer warning time.
horizon. In our specific implementation the WTZs are shown in Table 4, ranging from an hour to 2 days.

Table 4: Pilot 1.2.1.A1.1 – Experimentation results analysis

<table>
<thead>
<tr>
<th>Crop</th>
<th>Disease</th>
<th>Pest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olive trees</td>
<td>05/05/2017-17/05/2017 289 entries (hourly)</td>
<td>Dataset 01/01/2017-27/06/20:7</td>
</tr>
<tr>
<td></td>
<td>Results: 3 alarms, 5 warnings</td>
<td>178 entries (daily)</td>
</tr>
<tr>
<td></td>
<td><strong>Warning time horizon</strong>: on average 4 hours before alarm</td>
<td>1 alarm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 warnings</td>
</tr>
<tr>
<td>Grapes</td>
<td>02/05/2017-26/05/2017 337 entries (hourly)</td>
<td>Dataset 01/01/2017-27/06/20:7</td>
</tr>
<tr>
<td></td>
<td>Results: 3 alarms, 2 warnings</td>
<td>178 entries (daily)</td>
</tr>
<tr>
<td></td>
<td><strong>Warning time horizon</strong>: on average 1 hour before alarm</td>
<td>3 alarms (first, second and third generation)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 warnings</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Warning time horizon</strong>: on average 1.5 days before alarm</td>
</tr>
<tr>
<td>Peaches</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.2.5 Next steps

We are currently testing the end-to-end scenario in which real-time data is provided by GAIA Cloud to PROTON for further analysis via a RESTful interface. As a next step we will consider replacing PROTON’s dashboard as consumer by GAIATrons dashboard. In addition, the experimentation will be extended to support the continuously monitoring and alerting for more pest and disease breakouts for the specific crop types taking part in pilot A1.1.

2.3 WP1 - EO data analysis and machine learning methodologies for supporting CAP and Agri-Insurance use cases in Greece

2.3.1 General information

The following pipeline description presents a data/information exchange pipeline that supports two of DataBio’s pilot activities in WP1. It focuses on the advancements of two components that are part of the DataBio technology platform, namely C13.02 GAIABus DataSmart Machine Learning Subcomponent (NP) and C31.01 Neural Network Suite (CSEM) and their interoperability with GAIA Cloud. As the main target here is to focus on DataBio technology, it is worth mentioning that all development activities related to the expansion and technological evolution of NP’s GAIA Cloud will be described in respective WP1.
deliverables. GAIA Cloud is a major horizontal building block of NP’s Gaiasense\(^5\) solution and constitutes of multiple cloud computing services, which collect, store and combine heterogeneous data in order to convert them into facts using advanced data analytics techniques.

At this point, it should be noted that this pipeline constitutes also part of D5.2, as it primarily exploits EO data for creating data-driven crop models. Therefore, it is shown in the respective pipeline description in both documents.

### 2.3.1.1 Objectives

This pipeline copes with the need to analyse and model data originating from various sources including (most importantly) remote sensing, in situ measurements and land use data. The main target is to address challenges imposed by the increasing volume, variety and velocity of the data, flowing mainly within NP’s GAIA Cloud, and exploit big data components, like C13.02 GAIABus DataSmart Machine Learning Subcomponent and C31.01 Neural Network Suite in order to transform them into valuable insights for Common Agriculture Policy (CAP) Support and Insurance pilot cases in Greece.

The objectives of the pipeline are summarized below:

- To develop a set of cloud EO (Earth Observation)-based services for the agricultural sector
- To emphasize on the decision-making process, providing innovative solutions & applications
- To integrate earth observation data with image processing, machine learning and spatial modelling.
- To offer supporting technology solutions and services for NP’s CAP Support and Insurance pilots in Greece. More specifically to address specific problems and business needs like crop type identification and damage assessment respectively
- To experiment with alternative methodologies (C13.02 – C31.03) and identify their strengths and weaknesses within the specific business case scenarios (supporting DataBio’s vision for interchangeable and interoperable components).

### 2.3.1.2 Diagrams and views

The following C2.2 agriculture pilot views are indicative of the pipeline’s importance in the specific pilot cases.

\(^5\) [www.gaiasense.gr](http://www.gaiasense.gr)
Figure 11: Pilot C2.2 - Pipeline View

Figure 12: Pilot C2.2 - Lifecycle view
2.3.1.3 Associated pilot
Pilot 1.4.2C2.2: CAP Support (Greece)

URL in DataBio hub:
https://www.databiohub.eu/registry/#service-view/WP%201%20Pilot%2013%20[C2.2]%20CAP%20Support%20(Greece)

Pilot 1.4.1C1.1: Insurance (Greece)

URL in DataBio hub:
https://www.databiohub.eu/registry/#service-view/WP%201%20Pilot%2010%20[C1.1]%20Insurance%20(Greece)/0.0.1

2.3.1.4 Reusability
Since the pipeline is exploiting the functionalities offered by its machine learning/modelling components, it can be easily transferred to other application scenarios through specific adaptations (in terms of model types, data sets, operational characteristics, etc.). This is already verified within DataBio as it constitutes an important building block of two agriculture pilots (C1.1 and C2.2 agriculture pilots).
2.3.2 Interfaces

NP’s GAIA Cloud infrastructure is the core of this pipeline and as such manages most of the links and interfaces. The pipeline leverages providers’ distribution services to access EO data as soon as they are available. This mainly refer to connection links with entrusted EO data providers, namely the Copernicus Open Access Hub and its contributing Sentinel missions.

Further, the pipeline’s findings are accessible via Web Interfaces which were developed exploiting different technologies. Several implementations are being explored in order to address the needs of each individual business case. Indicatively, the following implementations are being examined:

- Report generation that uses the state-of-the art NeuroCode software tool (C13.01), designed for Rapid Application Development by partner NP.
- GAIA Cloud’s client (Agrimonitor UI (User Interface)). A new web-based interface has been designed by NP as an interface out of this pipeline. The new client (UI) has been developed to support the activities of the pilots and of the collaboration among the components. The UI uses browser-based libraries (e.g. leaflet) and is able to present multi-temporal object-based (parcel) data and information along with spatially aggregated statistics in a user-friendly way (Figure 14).

![Figure 14: Web-based UI offered by NP, providing spatial aggregations and attribute-based colour coding](image)

2.3.3 Components involved

2.3.3.1 Names

The components participating in this pipeline are:

- NP’s private cloud called, GAIA Cloud, along with all its infrastructure and the supporting cloud computing services from NP,
- C13.01 Neurocode from NP (interface out of this pipeline),

URL in DataBio hub:
https://www.databiohub.eu/registry/#service-view/NeuroCode

- C13.02 GAIABus Datasmart Machine Learning Subcomponents from NP,
  URL in DataBio hub:
  https://www.databiohub.eu/registry/#service-view/GAIABus%20DataSmart%20Machine%20Learning%20Subcomponent

- C31.01 Neural Network Suite from CSEM,
  URL in DataBio hub:
  https://www.databiohub.eu/registry/#service-view/Neural%20network%20suite%20for%20image%20processing/0.0.1

2.3.3.2 Interfaces between components
The pipeline plans to exploit well-specified Representational State Transfer (REST) endpoints along the way for bidirectional information and data sharing, as graphically depicted at the pilots’ pipeline views.

2.3.4 Experimentation
NP through its GAIA Cloud infrastructure performs all the CETL activities (connect, extract, transform and load) of the EO data. It has implemented a pipeline consisting of several pre-processing steps performed directly on Sentinel-2 products, including:

- Automated product acquisition and indexing
- Transformation to higher-level products. Quality enhancement and noise removal.
- Cloud annotation using the Fmask algorithm\(^6\), which is intended for masking different kinds of clouds and snow according to a cirrus cloud probability. This step alleviates the problem caused by the presence of clouds and cloud shadows within the optical imagery.
- Building on top of the generated higher-order (Level-2) Sentinel-2 products, several vegetation indices are extracted.

Most importantly, NP has implemented a land cover classification framework using indices that targets towards estimating the parcel’s crop type from EO data. The methodological framework uses a set of automatically generated and updated markers (parcel-related features which refer to a temporal snapshot of land’s use) and signals (time-series of the features) using statistical operations at parcel level (mean, median, max and standard deviation) for the creation of data-driven crop type models. The parcel signals are “data aggregates” that can be related to the physical properties of the land cover (and thus its crop type). The basic assumption to derive meaningful information from the EO and EO-generated

---

data is that the declared parcel contains a significant number of pixels. Further, C13.02 also supports the automatic outlier detection to remove noisy captures (e.g., snow, cloud, shadows) from the stack of images. The detection of such noise can be greatly facilitated by including metadata and meteorological data from the GAIAtrons in the processing. Finally, various machine learning methodologies (SVM, NNs) are exploited, where appropriate, for the development of crop classification models. Although this workflow is mostly concentrated on the CAP Support use case, several processes are being used in the context of the Insurance use case as well (e.g. CETL activities, time series extraction of vegetation indices at parcel level).

On the other hand, CSEM received EO data and parcel data from NP covering one Sentinel tile. The first step was to process the data before handing it over to the machine learning algorithm. Using simple statistics, outliers were removed from the training data set. The focus of the trial study was on peach trees. Using the C31.01 Neural Network Suite for Image Processing component a deep neural network was trained for a time point on this data set and its performance was measured using a previously unseen data set in various conditions, same crop and subtype, same crop and different subtype and completely different crop.

2.3.4.1 Deployment
NP has built and operates its own data centre offering cloud services at platform level (PaaS) and software level (SaaS). GAIA Cloud supports NP’s pilot activities at various levels and serves as the backbone of the pilots’ pipelines, in terms of infrastructure and applications/services. NP’s C13.02 GAIABus Machine Learning subcomponent is deployed for cloud-based operation within GAIA Cloud’s infrastructure. In order to support the aforementioned functionalities, GAIABus DataSmart Machine Learning Subcomponent exploits a set of technologies comprising among others of python-powered scripts, relational databases and REST endpoints.

The C31.01 Neural Network Suite is currently installed locally in CSEM and it is planned to be installed in a location accessible from other project partners.

2.3.4.2 Data sets
In order to support its activities, NP already collects and stores heterogeneous data in its private cloud infrastructure. The sources of data that are used for the purposes of data experimentation are the following:

- **EO data** (data used to feed the algorithms for modelling/classification purposes). Remote sensing data from the new Sentinel 2 optical products (13 spectral bands) are being extracted and stored since the beginning of 2016. The latter comprise both raw and generated data (corrected products, vegetation indices like NDVI and NDWI). In terms of this data experimentation and defining crop-specific models, EO and EO-generated data from several tiles from Northern Greece have been used that correspond to the 2016 and 2017 cultivating periods (several TBs/year).
• **IoT (Internet of Things, In-Situ) data** from GAIATron Atmo stations (ancillary data to support the decisions from the satellite imagery). GAIATrons offer configurable data collection and transmission rates. GAIATron stations collect and transmit several atmospheric (Atmo version) and soil (Soil version)-related measurements in NP’s cloud infrastructure contributing in several dimensions of big data (velocity, variety, value). More specifically, they continuously capture data related to soil temperature, soil humidity (multi-depth), ambient temperature, ambient humidity, barometric pressure, solar radiation, leaf wetness, rainfall volume, wind speed and direction.

• **Farm data.** Agricultural parcel information including position, crop type for 2016 and 2017, size, etc. (tailored for the individual pilot business case).

### 2.3.4.3 Results

**C13.02 GAIABus DataSmart Machine Learning Subcomponent (NP)**

During the preparatory phase of the project, data experimentation initially targeted towards modelling wheat, stone fruits, legumes, maize, forest trees, fallow and pasture exhibiting interesting findings (Figure 15, Figure 16).

![Image](image_url)

*Figure 15: Preliminary study area covering a total 282.600ha (53.580ha agricultural area). The image highlights correctly (green coloured parcels) and incorrectly (red coloured parcels) identified parcels.*
Figure 16: Preliminary study results for parcel classification

However, in order to meet business specific needs relative to the C2.2 CAP Support pilot case, more focus has been given on annual crops with an important footprint in the Greek agri-food sector, and indicatively: durum wheat, maize, cotton, etc.

For the training phase: Training is performed with parcels that their crop type is known for the growing seasons of 2016 and 2017. Training parcels are labelled with unique crop type categories and the training is performed at this stage separately for the two growing seasons. False labels can exist and are taken into consideration.

For the testing phase: Each agricultural parcel is a discrete testing sample for classification. The classifier should return a unique label accompanied with a confidence level for the result. At this stage, the data that need to be classified are from 2018-01-01 – now (updated every month to support continuous monitoring).

C31.03 Neural Network Suite (CSEM)
Preliminary results on peaches has shown that it is possible to classify them. Further optimization is required to increase the accuracy of the method.

Figure 17: CSEM’s C31.01 preliminary study on peaches. The image highlights peach fields used for training (yellow coloured), peach fields used for testing (green coloured, true positive samples) and wheat fields used for testing (red coloured, true negative sample).
2.3.5 Next steps

Data experimentation will keep progressing following the specific requirements of each pilot use case (CAP Support and Insurance in Greece) in terms of crop type model support and classification accuracy. From another perspective, the component will directly associate temporal variations in parcels’ extracted indices, with specific Insurance business case inputs (e.g. crop damages). All these activities will benefit from further optimization in terms of data cleaning and outlier removal. The described steps will ensure the wider uptake of the offered pipeline which target towards a highly ambitious set of pilot cases. Hence, KPIs have been defined in WP1 in order to guide the development process of the DataBio components participating in the present pipeline. Moreover, the components will be offered as services, exposing REST endpoints for data/information sharing. This will allow their integration with GAIA Cloud and the supporting -out-of-the-pipeline- interfaces in a seamless manner.

2.4 WP1 – Field data analysis and real-time alerting for decision making for precise agriculture

2.4.1 General information

2.4.1.1 Objectives

Pilot B1.1 aims to develop "irrigation maps" and "vigor maps" (combining EO data and IoT sensors data) which allows mapping different areas in Spain and set up an informative and management system for early warning of inhomogeneities. These services provide analytical and accurate finding of heterogeneities in crops related to irregular irrigation, mechanical problems affecting irrigation systems, incorrect distribution of fertilizers or any other sources of inhomogeneity that could explain crops growing differences. This service is a powerful preventive tool for general farmers and land owners to avoid production losses.

EO tasks are performed internally by TRAGSA Group (TRAGSA and TRAGSATEC), therefore the interfaces and pipelines are internals.

The goal of this pipeline is to take advantage of the use of new IoT systems in TRAGSA’s pilot located in Maceda (Spain), that will allow to measure environmental entities like temperature, humidity, air pressure, etc. To achieve this objective, low power resources like Raspberry Pi mounted with different sensors will be used to collect the data. Having available all this information and after processing it, will help to achieve a precision agriculture goals, where farmers will be supported by information for decision making as well as for predictive maintenance activities.

B1.1 Pilot will obtain data from C05.02: IoT Hub. This data will be then leveraged by C19.01: PROTON to develop an event driven warning system. Within the scope of this pilot, we will be able to validate FIWARE technologies in an agriculture environment, as PROTON is an important component in the FIWARE ecosystem and the IoT hub was born using different Generic Enablers from FIWARE. Being both components parts of the FIWARE program facilitates the integration between them.
2.4.1.2 Diagrams and views

**Figure 18: Pilot B1.1 - Pipeline view**

**Figure 19: Pilot B1.1 - Lifecycle view**

2.4.1.3 Associated pilot
Pilot B1.1: Cereals and Biomass Crops.

URL in DataBio Hub:


2.4.1.4 Reusability
All the EO processes on Sentinel Data could be exportable and reusables in any other forest or agriculture pilot.
Both PROTON and IoT Hub are generic tools, which means that the same pipeline could be reuse in other agriculture scenarios, and even in other domains. Obviously, new event rules and configurations will have to be defined.

2.4.2 Interfaces

Visualization Service on irrigation and vigour status aims to provide 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 indexes (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. In addition, fertilizers 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.

The outputs of the PROTON application are recorded into a CSV file for future testing and analysis. In addition, in order to have a graphical user interface for the outputs, we apply PROTON’s dashboard to show the input events as well as the situations or output events emitted by the engine in real-time.

IoT Hub is a middleware component to support continuous data collection from IoT based resources. In a typical IT infrastructure, it is usually located at gateway level, between data producers (resources) and data consumers. One of its main features is that it can be deployed in low level devices, like Raspberry PI. Other general characteristics of the component are: Interoperability, by supporting well-known and most common protocols for IoT related data sources. Scalability, due the amount of data generated by IoT devices, the module’s performance behaviour should remain stable during its use, being able to react to different data loads situations. Easy-to-use and install, simplifying the communication between uses cases and platform. Take advantage of the use of Open Sources technologies. The IoT Hub is composed by a set of four sub-components that tackle three different issues: communication, management of the devices and handling of data. Communication: grating interoperability and adaptation between different protocols and the matching between consumers and providers of data. Device Management functionalities to enclose generic information about devices and also to address their security and connectivity. Supporting data handling functionalities, IoT Hub is also granted with a sub component that ensures that the data obtained in the IoT world is pre-filtered before being passed to a data consumer or platform, reducing the flow or the quantity of inaccurate data.

IoT Hub offers two different interfaces:

- A Restful API for configuration that can be divided as well in two parts:
2.4.3 Components involved

2.4.3.1 Names

Three components participate in this pipeline:

- PROTON from partner IBM (C19.01 URL: https://www.databiohub.eu/registry/#services?name=IBM).
- FIWARE IoT Hub from ATOS:
  
  https://www.databiohub.eu/registry/#service-view/FIWARE%20IoT%20Hub/0.0.1
- Advanced Irrigation and Vigour Monitoring – Data visualization (C11.01):
  
  https://www.databiohub.eu/registry/#service-view/Advanced%20Irrigation%20and%20Vigour%20Monitoring%20Data%20Visualization/0.0.1

2.4.3.2 Interfaces between components

For now, the pipeline has not been completed in the sense that the integration between IoT Hub and PROTON has not been achieved yet. As both components can follow the FIWARE NGSI communication protocol, there exists two different possibilities to integrate the components:

1. Setting up an instance of the Orion Context Broker from FIWARE in the middle.
2. Use the Restful API offered by PROTON.

Option 2 has been selected as it fulfils the requirements so far as can be seen from Figure 19. The body of the event hasn’t been defined yet.

At this point, it is also relevant to highlight that the communication between the sensors and the IoT Hub is done using the MQTT protocol. IoT Hub includes a MQTT broker and when a device is registered, the necessary MQTT topics are created. Using those specific topics, data is published by the devices and gathered at the IoT Hub.

Example of the topics created:

```json
{
  "topic_single_attr": "/DATABIO/tempSensor/attributes/*",
  "topic_mult_attr": "/DATABIO/tempSensor/attributes",
  "topic_command_req": "/DATABIO/command/request",
  "topic_command_resp": "/DATABIO/tempSensor/command/response"
}
```

Example of data publishing:

```
mosquitto_pub -h localhost -t /DATABIO/tempSensor/attributes -m '{"temperature":"25", "humidity":"95"}'
```
2.4.4 Experimentation
At the time of writing this report, we are still working on the definition of the event rules and data types and therefore we don’t have any results to report yet. Experimentation has been possible only with the IoT Hub component (see the results section).

2.4.4.1 Deployment
IoT Hub will be deployed in the local premises of TRAGSA, close to the data sources. The service endpoint for PROTON is: https://hrldevops.eu-central.containers.mybluemix.net/ProtonOnWebServer/rest/events.

2.4.4.2 Data sets
Final data sets are still in the process to be defined. Because of the use of proprietary software in TRAGSA’s premises, it has been difficult to access data from such systems. It is expected that the data sets will consist on real-time data containing the data gathered by the final sensors to be used. An example of a possible data set to be used is shown in the next figure:

![Initial data set format for field data analysis and real-time alerting for precise agriculture](image)

Figure 20: Initial data set format for field data analysis and real-time alerting for precise agriculture

2.4.4.3 Results
So far, experimentation included only the IoT Hub component. This is since only recently the option of including the IBM component PROTON has been brought to the table, and we are still in the process of defining the data sets.
Up to now, IoT Hub has been able to gather data from a temperature and a humidity sensor, to process it and to send it to a control panel where the data is shown. See next figures:

Sensor used: DHT11, capable to gather temperature and humidity:
Raspberry PI 2 Model B, where the IoT Hub is running:

Sensor and Raspberry are connected as follows:

By using the API offered by the IoT Hub, it is possible to define the sensor as well as the statements to be applied on the data to generate new events, i.e.:

"statements": ["INSERT INTO Alert SELECT temperature, humidity FROM Room Where temperature > 15"]
• Data from sensors comes into the IoT Hub using the MQTT protocol.
• Below an example how the data coming from sensors is visualised:

![Example of data visualisation](image)

2.4.5 Next steps
Main next steps for this pipeline are:

• Definition of the final data sets to be used.
• Integration between IoT Hub and PROTON.

A first operative version of the whole pilot B1.1 will be available at the end of 2018.

2.5 WP1 - IoT data acquisition, analyses, and real-time alerts of potential threat conditions for crops

2.5.1 General information

2.5.1.1 Objectives
Precision agriculture involves the monitoring and prediction of conditions which might indicate a threat to the crops. Specifically, this pipeline addresses the monitoring of temperature and air pressure sensor measurements to detect and alert about possible unfavourable weather conditions, specifically a morning freeze. This condition is indicated by a low temperature together with high air pressure in the early morning hours which lasts for several hours. Additionally, rapid drop in the air pressure along with low temperatures might indicate a sudden weather change towards this condition. This has occurred frequently during 2017 and was the reason for large number of crops die-outs.

The goal is to have an integrated solution in which we collect data via Senslog component and push the data to IBM PROactive Technology ONline (PROTON) for analysis of the data (temporal reasoning) along with the triggering of early warnings/alerts in real-time.
2.5.1.2 Diagrams and views
Figure 21 describes the general view of the components collaboration in the pipeline – from receiving data from field sensors, storing the data to a DB, and publishing it to HSLayers visualizations. In parallel, the values of the selected phenomena are provided to PROTON for real time analysis and pattern detection, and the detected events are posted through Senslog REST APIs back to Senslog for visualization, storage, and further analyses.

![Diagram of pipeline view](image1)

Figure 21: Pilot B1.4 - Pipeline view

Figure 22 describes the interfaces between the SensLog and PROTON components:

![Diagram of interface view](image2)

Figure 22: Pilot B1.4 - Interface view

The interface view in Figure 22 includes the following elements:
C02.01: SensLog - SensLog is web-based sensor data management system. SensLog is a solution that is suitable for static in-situ monitoring devices as well as for mobile devices with live tracking ability. It is able to provide analytical functions of incoming data and to provide variety of data publishing ways. More info at: senslog.org.

C19.01: PROTON - IBM Proactive Technology Online (PROTON) is an open source complex event processing engine developed at IBM Research - Haifa. PROTON was developed as a research project and extended in the scope several EU projects. It provides language primitives for defining, submitting, and executing event processing networks. The goal of the system is to respond to raw events and identify meaningful events within contexts. The system comes with a set of built-in operators (such as sequence, all, etc.) for determining CEP patterns. It also has extendable APIs for adding additional custom operators. The system comes with existing source/sink adapters, allowing it to extract raw events from files or pull them from RESTful services. It also provides extendable APIs for adding more adapter types.

IF-PROTON-API-REST - PROTON's RESTful full API description is provided here: http://forge.fiware.org/plugins/mediawiki/wiki/fiware/index.php/Complex_Event_Processing_Open_RESTful_API_Specification. In the case of GAIA interface, the API is used for pushing GAIA sensor events to be input events for PROTON CEP application monitoring crops.

IF-SENSLOG-REST-API - SensLog has implemented own RESTful API. It uses JSON encoding of published data. API is covering full functionality of the data model of SensLog. It is straightforward and self-described. More info at: http://www.senslog.org/api/.

2.5.1.3 Associated pilot
Pilot 1.3.1.B1.4: Cereals and biomass crops

URL in DataBio Hub:
https://www.databiohub.eu/registry/#service-view/WP1%20B1.4%20Cereals%20and%20biomass%20crops_4%20Pipeline/0.0.1

2.5.1.4 Reusability
While PROTON is a generic tool, SensLog is designed to acquire IoT data from sensors and perform a set of activities including data monitoring, storage, pre-processing and analyses. Having said this, it is possible to reuse the same pipeline in a different agriculture scenario assuming the rules, thresholds and models for the calculations are known. Basically, SensLog and PROTON together provide a flexible solution for IoT sensor data acquisition, storage, and analysis.

2.5.2 Interfaces

2.5.2.1 SensLog
SensLog has implemented its own RESTful APIs. It uses JSON encoding of published data. API is covering full functionality of the data model of SensLog. It is straightforward and self-described. The following link provides the full description of the APIs: http://www.senslog.org/api/#InsertAlertEv.
2.5.2.2 PROTON

Currently the output of the CEP application is recorded both in output file for future analysis but also in a dashboard graphical application. It presents the input events as well as the situations or derived events emitted by the engine in real-time. Figure 8 below shows a screenshot of the dashboard showing the sensor reading input events and LowTemperatureFor2Hours alert and the reasoning behind the alert derivation. The alarm level situations are denoted with a red exclamation mark.

![LowTemperatureFor2Hours alert](image)

For very low temperature sensor reading (temperature is < 0 degrees) calculate if within two hours from its arrival the temperature did not go above 0 degrees, if so report an alarm.

**Figure 23: PROTON’s user interface for input/output event visualization**

2.5.3 Components involved

2.5.3.1 Names

Two components participating in this pipeline are: SensLog from partner UWB (02.01 URL: https://www.databiohub.eu/registry/#services?name=Senslog) and PROTON from partner IBM (C19.01 URL: https://www.databiohub.eu/registry/#services?name=IBM).

2.5.3.2 Interfaces between components

PROTON provides its own set of interfaces for streaming in input data, called adapters. Some of these adapters include file adapters, JMS adapters, and RESTful adapters. It also provides an extensible architecture for implementing and adding custom adapters for data input. In the current application, PROTON provides REST API for streaming in sensor data. Additional information on PROTON’s RESTful API can be found here: http://forge.fiware.org/plugins/mediawiki/wiki/fiware/index.php/Complex_Event_Processing_Open_RESTful_API_Specification.

- Events from SensLog to PROTON (see Figure 22: Pilot B1.4 - Interface view):
  - Are submitted to PROTON through PROTON’s RESTful API.
PROTON reports back the derived events to SensLog via RESTful API SensLog provides. The API to receive the alarm data:

http://www.senslog.org/api/#InsertAlertEvent

Figure 24: SensLog REST API for inserting alarms

- Events from PROTON to RESTful consumer (either dashboard or SensLog). See Figure 21.
  - The body of the event for different alarms:
    - {Name:DroppingTemperatureAlert,airPressureValue:962.900024,lowestTemperatureValue=2.4}
    - {Name:LowTemperatureFor2HoursAlert}

2.5.4 Experimentation
Sensor readings are collected by SensLog and pushed to PROTON for further analyses and pattern detection.

Specifically, PROTON applied the following monitoring rules:

- **Low temperature** – Alert when temperature drops below zero and stays below zero for 2 hours
- **Dropping temperature and high air pressure** – Alert when there are five consecutive observations of decreasing and low temperature while at the same time air pressure is high

2.5.4.1 Deployment
The service endpoint for PROTON is:
2.5.4.2 Data sets
The input data consisted of:

- 2 CSV files from 2 meteo stations (1012, 1040) from different parts of Czechia
- Sample of 6 days observation series for both stations
  - For meteo station 1040 sensor readings every 10 mins, 900 entries
  - For meteo station 1012 sensor readings every 15 mins, 500 entries

2.5.4.3 Results
PROTON was able to detect and alert on the trends for freezing mornings.

Specifically:

- For the meteo station 1040 reported LowTemperature (in the mornings) and DroppingTemperature (in the evening hours)
- For the meteo station 1012 reported LowTemperature alarms

Table 5 and Table 6 depict the results from the described experimentation. Please note that the data was supplied for areas which experienced a series of freezing mornings in the provided dates, which was captured in the results.
Table 5: Experimentation results station 1040

<table>
<thead>
<tr>
<th>Date/time</th>
<th>Situation reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>18/04 17:20</td>
<td>DroppingTemperatures</td>
</tr>
<tr>
<td>19/04 07:00-10:40</td>
<td>LowTemperatureFor2Hours</td>
</tr>
<tr>
<td>20/04 07:00-09:50</td>
<td>LowTemperatureFor2Hours</td>
</tr>
<tr>
<td>20/04 18:00-19:00,20:00-20:20</td>
<td>DroppingTemperatures</td>
</tr>
<tr>
<td>21/04 00:50</td>
<td>DroppingTemperatures</td>
</tr>
<tr>
<td>21/04 01:30-06:30</td>
<td>LowTemperatureFor2Hours</td>
</tr>
<tr>
<td>21/04 18:50-19:10,20:10-20:20</td>
<td>DroppingTemperatures</td>
</tr>
<tr>
<td>22/04 19:40-21:20</td>
<td>DroppingTemperatures</td>
</tr>
<tr>
<td>23/04 16:50-19:10,20:30-20:50</td>
<td>DroppingTemperatures</td>
</tr>
<tr>
<td>24/04 01:40-05:40</td>
<td>LowTemperatureFor2Hours</td>
</tr>
<tr>
<td>24/04 02:40</td>
<td>DroppingTemperatures</td>
</tr>
</tbody>
</table>

Table 6: Experimentation results station 1012

<table>
<thead>
<tr>
<th>Date/time</th>
<th>Situation reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>19/04 07:30</td>
<td>LowTemperatureFor2Hours</td>
</tr>
<tr>
<td>20/04 22:45-06:30</td>
<td>LowTemperatureFor2Hours</td>
</tr>
</tbody>
</table>

2.5.5 Next steps

We are currently testing the end to end scenario in which real-time data is provided to PROTON for further analysis via a RESTful interface. As a next step we will consider replacing PROTON's dashboard as consumer by other applications, such as feeding back alarms to SensLog using their provided RESTful APIs (see Figure 24).
2.6 WP1 - Publication of Linked Data related to cereals and biomass crops

2.6.1 General information

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. This pipeline aims to collect, transform, and publish farm data and relevant open datasets as Linked Data, as well as their exploitation and reuse over applications like HSLayers NG and other Linked Data visualization tools. Initially, this pipeline has been tested with data from pilot B1.4 - cereal and biomass crops.

The process so far involved identification and collection of farm data and open datasets, pre-processing and transformation of the datasets into RDF triples according to existing ontologies, post-processing operations on the RDF data, and publishing the datasets as Linked Data in the Virtuoso available in the PSNC cloud infrastructure. Additional possible links for specific components with other external linked datasets have been analyzed using special Linked Data applications.

2.6.1.1 Objectives

Linked data defines simple principles for publishing and interlinking structured data that is accessible by both humans and machines. Some of the key objectives for publishing data as linked data include:

- Improved data accessibility by lowering the barriers on finding and reusing this data
- Harmonization and integration of different datasets related to the agriculture sector using linked data as a federated layer
- Knowledge discovery through the connections with other datasets and potentially application of inferencing
- Exploitation of data through advanced visualization interfaces providing integrated view on the data
- Query and access to data through semantic queries
2.6.1.2 Diagrams and views

Figure 25: Pilot 1.3.1.B1.4 - Role view

Figure 26: Pilot 1.3.1.B1.4 - Pipeline view
2.6.1.3 Associated pilot

WP 1 Pilot B1.4 - Cereals and biomass crops_4

URL in DataBio Hub:

https://www.databiohub.eu/registry/#service-view/WP%20Pilot%20[8%20B1.4]%20Cereals%20and%20biomass%20crops_4/0.0.1

2.6.1.4 Reusability

This pipeline is based on best practices for publishing linked data\(^7\) and has been realised following standards and well-known vocabularies and components. Thus, the pipeline can be reused for publishing linked data in different application domains, although it is focused on agriculture-related data with geo-spatial dimension. The pipeline has been applied and tested in one of the pilots related to agriculture in DataBio and could potentially be applied to other agri-pilots. The main effort for transferring and applying this pipeline to other scenarios is in creating the mapping specification.

The pipeline supports data in different formats and from different sources (e.g., relational databases, CSV, JSON, XML, shapefiles), although the experiments were carried out with datasets in relational format (in PostgreSQL) and with shapefiles.

The main ontology used for the generation of RDF data, called FOODIE 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). The model (reused and extended from FOODIE project) has received positive feedback from experts in various institutions. When applying the pipeline to other pilots, it may be necessary to create small extensions to the ontology to cover some pilot specific data necessary for the mapping specification.

\(^7\) https://www.w3.org/TR/ld-bp/
On the component level, Virtuoso is the main platform used. Being a generic RDF triple store and Linked Data Publication framework, it can be (re-)used to publish any dataset as linked data. Virtuoso is an optimized triple store than can manage millions of triples with exceptional performance, making it a very scalable solution. Virtuoso provides a SPARQL endpoint and a faceted search interface for querying and browsing the Linked Data. Other components used include data semantization tools D2RQ and geotriples, link discovery tools silk and limes, and linked data visualization tools HSLayers and metaphactory.

2.6.2 Interfaces

The Linked Data publication pipeline provides and consumes the following interfaces (Table 7 and Table 8):

**Table 7: Linked Data publication pipeline - provided interfaces**

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF-Virtuoso-SPARQL-Service-Endpoint</td>
<td>This interface allows the interaction between the Virtuoso Universal Server and the open standard compliant services and application via SPARQL service. In this pipeline this interface allows the connection between Virtuoso and HS Layers NG.</td>
</tr>
<tr>
<td>IF-Virtuoso-Rest-Interface</td>
<td>This interface allows the interaction between the Virtuoso Universal Server and the open standard compliant services and application via REST service.</td>
</tr>
<tr>
<td>IF-Virtuoso-Fct-Interface</td>
<td>This interface allows the interaction between the Virtuoso Universal Server and the open standard compliant services and application via a faceted search user interface.</td>
</tr>
<tr>
<td>IF-HS Layers NG-GUI</td>
<td>This is the interface originating from the component HS Layers NG providing the user interface for the map and Linked Data visualization</td>
</tr>
</tbody>
</table>

**Table 8: Linked Data publication pipeline - consumed interfaces**

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF-JDBC-Interface</td>
<td>This interface forms the connection between the databases (PostgreSQL + PostGIS) through the Data Semantization Application (geotriples, D2RQ etc.)</td>
</tr>
</tbody>
</table>
to the main component Virtuoso where the RDF dumps generated after semantization are loaded and queried.

<table>
<thead>
<tr>
<th>Interface</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF-Filesystem</td>
<td>In a more specific case this interface is used while dealing with geospatial datasets in different file formats, such as CSV, JSON, and shapefiles, which are used as input in this pipeline. This interface is application-specific for the application Geotriples and allows to load the files as input for data semantization. After semantization the RDF data passes though IF-Virtuoso-CLI interface leading to the main component of the pipeline called Openlink Virtuoso for storage.</td>
</tr>
<tr>
<td>IF-Virtuoso-CLI</td>
<td>This interface allows to load the rdf dumps generated by the semantization application to the Virtuoso Universal Server.</td>
</tr>
<tr>
<td>IF-SENSLOG-REST-API</td>
<td>The connection to this interface is yet to be realized and will be available in the future versions where the application component SensLog will connect to the main components via REST API.</td>
</tr>
</tbody>
</table>

2.6.3 Components involved

2.6.3.1 Names

The following components are involved in this pipeline along with the respective link of DataBio Hub:

- C09.12: Openlink Virtuoso.
  - [https://www.databiohub.eu/registry/#service-view/OpenLink%20Virtuoso/0.0.1](https://www.databiohub.eu/registry/#service-view/OpenLink%20Virtuoso/0.0.1)
- C02.03: HS Layers NG.
  - [https://www.databiohub.eu/registry/#service-view/HSLayers%20NG/0.0.1](https://www.databiohub.eu/registry/#service-view/HSLayers%20NG/0.0.1)
- C12.01: LIMES
  - [https://www.databiohub.eu/registry/#services?name=limes](https://www.databiohub.eu/registry/#services?name=limes)
- C02.01: SensLog
  - [https://www.databiohub.eu/registry/#services?name=Sens](https://www.databiohub.eu/registry/#services?name=Sens)

2.6.3.2 Interfaces between components

This pipeline has the following main interfaces between the respective components:
- IF-JDBC-Interface forms the interface between the databases (PostgreSQL + PostGIS) and the main component Virtuoso through the Data Semantization Application (geotriples, D2RQ, etc.) that generates the RDF dumps that can be loaded and published in Virtuoso via the IF-Virtuoso-CLI interface and queried by consumers through the different interfaces provided by the pipeline.

- The IF-Filesystem interface is used by the semantization application like Geotriples to load files as input for data semantization. These input files are geospatial datasets that can be in different formats, such as CSV, JSON and shapefiles. After semantization, the RDF data is loaded into Virtuoso using the IF-Virtuoso-CLI interface for storage and further exploitation over different platforms.

- The interface IF-Virtuoso-CLI allows to load the RDF dumps generated by the semantization application to the Virtuoso Universal Server.

- IF-Virtuoso-SPARQL-Service-Endpoint interface allows the interaction between the Virtuoso and the open standard compliant services and application via SPARQL service. In this pipeline this interface allows the connection between the components Virtuoso and HS Layers NG where the Linked Data can be exploited and visualized.

- The connection to the interface IF-SENSLOG-REST-API is yet to be realized and will be available in the future versions where the application component SensLog will connect to the main components via REST API.

2.6.4 Experimentation

Multiple experiments were carried out and some of the results have been showcased at different events (e.g., INSPIRE hackathons, Link Open Data Workshop, and code camp).

In the first experiment we published some EU datasets as linked data, including Open Land Use (OLU) and Open Transport Map (OTM). Because of OLU dataset depends on objects from other datasets (Corine, Urban Atlas, Cadastral Parcels) it also practically meant to convert all those objects to RDF form as well. Before the transformation, ontologies for each of these datasets had to be created (see https://github.com/FOODIE-cloud/ontology). The next step was to carry out the transformation itself. As these datasets were in a relational database (PostgreSQL+PostGIS), we used the D2RQ Platform (http://d2rq.org/), which uses mapping files (in RDF) to execute the transformation. As we were dealing with extremely large pan-European datasets, not all the data was converted, but a large portion was, mainly focused in three countries: ES, PL, CZ. After the data was transformed, it was loaded to the triple store in Virtuoso (see endpoint link in next section). Additionally, we loaded other open datasets, including the Smart Points of Interest (SPOI), NUTS, Eurovoc and Emergel. Regarding the links, the datasets were generated already with some links. For instance, OLU has links to Urban, Corine, and Czech cadaster. SPOI has links to different datasets including dbpedia and others. Hence the next step was to show these connections by creating some interesting semantic queries (in SPARQL). The data returned by each query can be visualised as a layer in map using the HSLayers application developed by LESPRO (see links in next section).
Another experiment, using the same components was carried out but with farm data, although not directly from a DataBio pilot. In this case, the farm data was also available in a PostgreSQL (+PostGIS) database, and included information about the plots, crops, some treatments, and productions. Hence, the underlying model for the generation of the RDF data was FOODIE ontology (https://github.com/FOODIE-cloud/ontology). The published dataset, connected with the EU datasets, can be also visualized via the HSLayers application (see link below).

Then, in a more relevant experiment, sample Farm and Open datasets from the DataBio pilot B1.4 were collected in the form of shape files which were then analyzed and pre-processed to transform them into the same coordinate system (EPG 4326). Then the shapefiles were transformed to RDF using GeoTriples tool (http://geotriples.di.uoa.gr/), based on the mapping specification created in RDF. Various attributes in the newly created mapping files were analyzed and extensions were created according to the FOODIE ontology to match and cover those terms used in the shapefiles which could be associated to those used in the ontology. In this process various other ontologies were also considered to match the attributes.

Using GeoTriples, RDF dumps were generated from the processed shapefiles. After few post-processing steps (e.g. removing duplicates, and coordinate system strings) these dumps were imported into Virtuoso. In the post-processing action, the possibility of linking the resultant datasets with other previously published datasets (dbpedia, NUTS etc.) using the application called SILK was identified. The datasets can be accessed via the Virtuoso SPARQL endpoint and be used in various application platforms for visualization (e.g. Hs Layers-NG, Metaphactory, and Sextant).

2.6.4.1 Deployment
Openlink Virtuoso is deployed in the cloud infrastructure of PSNC which provides SPARQL endpoint and a faceted search for the datasets as Linked Data. Virtuoso has 16GB of RAM allocated in a VM with 24 GB of RAM. The size of the current database is about 180GB.

The SPARQL endpoint can be accessed from https://www.foodie-cloud.org/sparql

The faceted search interface can be accessed from http://www.foodie-cloud.org/fct/

One of the existing linked data application interfaces consuming our datasets in virtuoso is available at: https://foodie.graphs.com/resource/Start

Some visualisation example using HSLayers are available at:

- https://goo.gl/HQsXwr
- http://ng.hslayers.org/examples/foodie-zones/

2.6.4.2 Data sets
The input data for the EU datasets experiment consisted of a large database in PostgreSQL (+PostGIS), while the first farm dataset was available in another smaller PostgreSQL database.
The input data for the DataBio pilot experiment consisted of three shapefiles from Farm Data and three shapefiles from Open Data from Czech Republic.

- **Farm Data:** Data for farms and field names with the name of crops and the coordinates and code identifier, details for specific crops in two designated fields with the specific parameters like moisture and yield masses collected periodically.

- **Open Data:** LPIS data from the Czech Republic, collected data of the Erosion zones, open data of the water bodies.

### 2.6.4.3 Results

The result of the process is the publication of Linked Data related to the agriculture domain, in compliance with the FOODIE agriculture ontology (that in turns is based on the INSPIRE and ISO standards). These data are available in a Virtuoso triple store installation in PSNC infrastructure. Currently Virtuoso stores over 700 million triples. Some statistics below.

#### Table 9: EU datasets

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Graph in FOODIE endpoint</th>
<th>Source</th>
<th>Triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLU</td>
<td><a href="http://w3id.org/foodie/olu#">http://w3id.org/foodie/olu#</a></td>
<td>Transformed from PostgreSQL</td>
<td>127,925,971</td>
</tr>
<tr>
<td>SPOI</td>
<td><a href="http://www.sdi4apps.eu/poi.rdf">http://www.sdi4apps.eu/poi.rdf</a></td>
<td>Provided by WRLS + Improvements</td>
<td>381,393,555</td>
</tr>
<tr>
<td>NUTS</td>
<td><a href="http://nuts.geovocab.org/">http://nuts.geovocab.org/</a></td>
<td>Open Source</td>
<td>316,238</td>
</tr>
<tr>
<td>OTM</td>
<td><a href="http://w3id.org/foodie/otm#">http://w3id.org/foodie/otm#</a></td>
<td>Transformed from PostgreSQL</td>
<td>154,340,611</td>
</tr>
<tr>
<td>Hiluc classification</td>
<td><a href="http://w3id.org/foodie/hiluc#">http://w3id.org/foodie/hiluc#</a></td>
<td>Transformed from PostgreSQL</td>
<td>397</td>
</tr>
<tr>
<td>Urban Atlas</td>
<td><a href="http://w3id.org/foodie/atlas#">http://w3id.org/foodie/atlas#</a></td>
<td>Transformed from PostgreSQL</td>
<td>19,606,025</td>
</tr>
<tr>
<td>Corine</td>
<td><a href="http://w3id.org/foodie/corine#">http://w3id.org/foodie/corine#</a></td>
<td>Transformed from PostgreSQL</td>
<td>16,777,533</td>
</tr>
<tr>
<td>Eurovoc</td>
<td><a href="http://foodie-cloud.org/eurovoc">http://foodie-cloud.org/eurovoc</a></td>
<td>Open Source</td>
<td>425,067</td>
</tr>
<tr>
<td>Emergel</td>
<td><a href="http://foodie-cloud.org/emergel">http://foodie-cloud.org/emergel</a></td>
<td>CTIC</td>
<td>256,239</td>
</tr>
</tbody>
</table>

#### Table 10: National datasets (Czech)

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Graph in FOODIE endpoint</th>
<th>Source</th>
<th>Triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPIS data (Czech)</td>
<td><a href="http://w3id.org/foodie/open/cz/180308_pLPIS_WGS#">http://w3id.org/foodie/open/cz/180308_pLPIS_WGS#</a></td>
<td>Transformed from Shapefiles</td>
<td>18,188,795</td>
</tr>
<tr>
<td>Water Bodies (Czech)</td>
<td><a href="http://w3id.org/foodie/open/cz/Water_bodies_buffer25m_WGS#">http://w3id.org/foodie/open/cz/Water_bodies_buffer25m_WGS#</a></td>
<td>Transformed from Shapefiles</td>
<td>3,869,783</td>
</tr>
<tr>
<td>Erosion Zones (Czech)</td>
<td><a href="http://w3id.org/foodie/open/cz/Erosion_zones_WGS#">http://w3id.org/foodie/open/cz/Erosion_zones_WGS#</a></td>
<td>Transformed from Shapefiles</td>
<td>81</td>
</tr>
</tbody>
</table>
Table 11: Farm data

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Graph in FOODIE endpoint</th>
<th>Source</th>
<th>Triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farm data (Spain)</td>
<td><a href="http://w3d.org/foodie/core/es/">http://w3d.org/foodie/core/es/</a>*</td>
<td>Transformed from PostgreSQL</td>
<td>57,160</td>
</tr>
<tr>
<td>Farm data (Czech)</td>
<td><a href="http://w3d.org/foodie/core/cz/">http://w3d.org/foodie/core/cz/</a>*</td>
<td>Transformed from Shapefiles</td>
<td>1,555,422</td>
</tr>
</tbody>
</table>

The Linked Data from the input datasets can be queried and browsed via the SPARQL endpoint and faceted search browser provided by Virtuoso, and across various visualization platforms for graph data. Some of these applications where the Linked Data can be visualized are Metaphactory, HsLayers NG, and Sextant. Some screenshots of these applications showcasing the published data are depicted in the figures below.

Figure 28: Pilot 1.3.1.B1.4 - SPARQL endpoint for Linked Data

Figure 29: Pilot 1.3.1.B1.4 - Faceted search endpoint for Linked Data
Figure 30: Pilot 1.3.1.B1.4 - Map visualization of Linked Data in HS Layers-NG

Figure 31: Pilot 1.3.1.B1.4 - Map visualization of Linked Data in Metaphactory
2.6.5 Next steps

The further steps include:

1. Addressing the requirements of a protected environment for the published Linked Data,

2. realising the connection with SensLog to enable the publication of streamed sensor data as linked data, and

3. extending existing GUIs and test other Linked data visualization tools like Sextant.

4. We also aim at exploring other security issues regarding organizational, legal and commercial aspects of the Linked Data sharing.

At this moment we are addressing the first point, i.e., securing Virtuoso endpoint. We decided to implement two mechanisms:

- Secure the SPARQL Endpoint Guide via SQL Accounts. The authentication will be available via the /sparql-auth interface, while the main public interface /sparql will be available for non-authenticated users (to access only public data)

- Graph-level security. Virtuoso supports graph-level security for "physical" RDF storage. This is somewhat similar to table-level security in SQL, and works as follows: When a SPARQL query should check whether a given user has permission to access a given graph, the order of checks is:
  
  - permissions of the user on the specific graph;
  - default permissions of the user on all graphs;
  - public permissions on the specific graph;
  - public permissions on all graphs.
  - If no permission is set, then the access is "read/write/sponge/list".

2.7 WP1- IoT data acquisition, analyses, and prediction of potential threat conditions for crops

2.7.1 General information

This pipeline can be interpreted as an alternative pipeline to 2.5.

2.7.1.1 Objectives

The usage of IoT is crucial for precision agriculture and require the monitoring and prediction of conditions which might indicate a threat to the crops. This pipeline addresses the analysis and prediction of specific conditions like temperature and air pressure to enhance pipeline 2.5.
The goal is to have a pipeline which enables the data collection with the help of the SensLog component. Via the SensLog API, the component Albatross analyzes (time-series-analysis) the data and predict near future conditions – to enable a warning or alert system.

2.7.1.2 Diagrams and views

Figure 32 shows the connection from SensLog to Albatross. Albatross can consume the SensLog API\(^8\) and analyzes the resulting JSON.

```
Figure 32: Pilot 1.3.1.B1.4 - Pipeline view
```

2.7.1.3 Associated pilot

Pilot 1.3.1.B1.4: Cereals and biomass crops

2.7.1.4 Reusability

As stated out, this pipeline is an alternative pipeline for IoT data management, analysis, and prediction in various agricultural settings.

2.7.2 Components involved

2.7.2.1 Names

SensLog

SensLog (C02.01) has implemented its own RESTful APIs. It uses JSON encoding of published data. API is covering full functionality of the data model of SensLog. It is straightforward and self-described. The following link provides the full description of the APIs: [http://www.senslog.org/api](http://www.senslog.org/api).

\(^8\) http://www.senslog.org/api/
Albatross

Albatross (C12.02a) is a follow-up software solution of ESTA-LD (C12.02). The software is Open Source and helps to analyze data. In addition, it supports simple prediction models.

2.7.2.2 Interfaces between components
IF-SENSLOG-REST-API is the REST API of SensLog producing a JSON file. The interface is available at: http://ckan.ccss.cz/SensLog/.

2.7.3 Experimentation
The Sensor data had been read, structured, and analyzed. A simple prediction model has been used to predict the conditions. Figure 33 shows the analyzed data and the prediction with the future variance.

Figure 33: Analyzed data and prediction with future variance screenshot

2.7.3.1 Deployment
Up to now, the test has been deployed on an internal server.

2.7.3.2 Datasets
The data consist of five stations (including 4 sensors) from a Czech field with more than 200 samples.

2.7.3.3 Results
With the application of a trained prediction, it is possible to detect bad conditions before they happen and enable pro-active actions if needed.
2.7.4 Next steps
Next steps include the enhancement of tests with by real-time data and complex multivariate prediction models. Furthermore, there is no function to feed the SensLog alert API, which must be implemented.

2.8 WP2 - Shared multiuser forest data environment

2.8.1 General information
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 (see Figure 34 with User 1). Metsän.fi provides also an XML service via the forest owners and forestry service providers can download the standardized forest resource data. This data can be utilized for instance for updating the forest stand information via mobile applications. This data is also linked to personal data and therefore it is available only for registered users.

Open forest data service is a portal where users (User 2 in Figure 34) can download, utilize API’s, or review the data classified as environmental data. This data can be utilized as a basis, for instance, in the crowdsourcing applications. Open forest data does not have any information regarding the property owners and therefore it is classified as public information.

In addition to these services, METSAK is activating the X-road approach provided by Suomi.fi platform (national service architecture). This approach enables Metsakeskus to receive the standardized data for instance provided by the crowdsourcing applications.

All the services and components are located in Finland.

2.8.1.1 Objectives
Objectives of the pipeline are:

- Provide new maps and functionalities to Metsään.fi and Metsäkeskus concerning crowdsourced data
- Provide prescriptions of quality control for Metsään.fi users
- Bring generalized storm damage data to Metsään.fi via crowdsourcing interface
- Produce updates on forest data standard to collect quality control data with mobile devices
- Provide user and data security, via single-login and easy user role based authentication and data access permissions

2.8.1.2 Diagrams and views
Figure 34 and Figure 35 depict pipeline and lifecycle views of the pipeline respectively.
2.8.1.3 Associated pilot

Pilot 2.4.2: Shared multiuser forest data environment (https://www.databiohub.eu/registry/#service-view/WP2)

2.8.1.4 Reusability

The pipeline is specifically tailored for this pilot, however the Suomi.fi based data transfer service enables the data transfer in standardized way between the METSAK and other partners. Also standardized forest data can be utilized for other purposes and on different
scenarios. Suomi.fi service is also applied for the user identification and authentication by Metsään.fi service and many other public organizations in Finland.

2.8.2 Interfaces

- Database interfaces to data storage.
- Interfaces to data transfer services.
- Interfaces to backend systems.
- Interfaces to map services.
- Interface to end user tracking system.

2.8.3 Components involved

2.8.3.1 Names

- C18.01 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.
- C18.02 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.
- C20.01 Wuudis (https://www.databiohub.eu/registry/#service-view/Wuudis/0.0.1) - 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, data in XML and JSON formats, connects forest owners directly with local contractors and timber buyers.

2.8.3.2 Interfaces between components

Metsään.fi eService and Open forest data service will be integrated to Wuudis component via the national service architecture Suomi.fi based data transfer service and API’s.

2.8.4 Experimentation

Metsään.fi data has been integrated with Wuudis mobile application. The interface has been built between the Metsään.fi service and Wuudis mobile application to enable the standardized forest resource data transfer between these two components. Wuudis application has been tested and taken into use by the forest advisors working in Finnish Forest Centre (METSAK) and the feedback has been positive. Further development is needed so that the updated forest resource data during the field work activities can be returned back to Finnish Forest Centre (METSAK) backend systems. This solution will be technically available after the Suomi.fi X-Road implementation for the data transfer services.

Another experimentation has been completed by implementing the national service architecture-based user identification and authorisation. This new technical setup has replaced the old customized setup for the user access management.
2.8.4.1 Deployment
All the components are installed in Finland.

2.8.4.2 Data sets
The total amount of Metsään.fi data is currently around 330 GB, while it was approximately 200 GB at the beginning of the project. The Metsään.fi data covers currently around 11.5 Mha of the private forests.

Open forest data service was implemented on 1st of March 2018. Currently the total amount of the data available in this service is 276.8 GB. After two months around 1486.6 GB of data has been downloaded via this service.

Due to the delay in data transfer service implementation the crowdsourced data is not yet available. However, the technical readiness should be available by the end of 2018, when the pilot experiments can be kicked off.

2.8.4.3 Results
Metsään.fi service user identification and authentication has been implemented according to Suomi.fi national service architecture. This means that the standardized way of Metsään.fi user login has been taken into use. This decreases the customized solution support needs and streamlines the user access management.

2.8.5 Next steps
The next step is to build the technical readiness for collecting the crowdsourced information. This means development of the forest data standards for the forest damages, building the API interface for the Open Forest Data Service, enabling the X-Road approach for the data transfer service and developing the map service for the data visualization.

2.9 WP2 – Multi-source data enabling efficient forest management

2.9.1 General information

2.9.1.1 Objectives
The objectives of this pipeline are as follows:

- Real-time forest management service development based on multiple forest big data sources
- Assess and monitor forest health status
- Handle big data in forestry for storing, processing, and transferring of large amounts of forestry data

Towards this end, currently we focus on 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.

2.9.1.2 Diagrams and views
Figure 36 presents the pipeline view and Figure 37 the lifecycle view corresponding to the pilots.

**Figure 36: Pilots 2.2.1 and 2.2.2 - Pipeline view**

**Figure 37: Pilots 2.2.1 and 2.2.2 - Lifecycle view**

2.9.1.3 Associated pilot
The pipeline described here enables implementation of forestry pilots 2.2.1, 2.2.2 and 2.3.1.

Pilot 2.2.1: Easy data sharing and networking
Pilot 2.2.2: Monitoring and control tools for forest owners

With regards to pilot 2.3.1, a pipeline description in D5.2 focuses on exploitation of Earth Observation data using the Forestry TEP.

2.9.1.4 Reusability

Wuudis is a commercial service and it is integrable with any third-party solution. With this feature in consideration, it is possible to reuse the same pipeline in a different forestry pilot scenario following the standardized approach developed in this pilot.

For example, the same pipeline procedure can be followed in scaling the solution to other EU countries (e.g.: discussion with Galician forest association and authority is undergoing). Depending on the forest authority system existing IT system in other cases, this might trigger new event rules in the existing event processing implementation or a complete new implementation.

2.9.2 Interfaces

External interfaces of the pipeline are via Wuudis, which provides the following interfaces (see Figure 38 and Table 12: Wuudis interfaces):

![Diagram of Wuudis interfaces](image)

**Figure 38: Wuudis - Interface view**

**Table 12: Wuudis interfaces**

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF-Wuudis-Mobile-App-UI</td>
<td>Mobile user interface for a forest owner/timber buyer/service provider/authority expert.</td>
</tr>
<tr>
<td>IF-Wuudis-Forest-XML-Export</td>
<td>REST API of Wuudis uses the Finnish forest information XML standard as basic data export format.</td>
</tr>
<tr>
<td>IF-WUUDIS-Forest-XML-Import</td>
<td>REST API of Wuudis uses the Finnish forest information XML standard as basic data import format.</td>
</tr>
<tr>
<td>IF-Wuudis-REST-API</td>
<td>REST API of Wuudis that uses plain JSON as import/export format.</td>
</tr>
</tbody>
</table>
2.9.3 Components involved

2.9.3.1 Names

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)

2.9.3.2 Interfaces between components

In this pipeline, flow of information is towards Wuudis, in order to collect and visualize information from multiple sources to the end user. Forestry TEP supports Wuudis with a WMS based interface, via which a forest information layer based on analysis of Sentinel-2 satellite images is offered. Metsään.fi by METSAK offers its standard interface, which is directly exploitable in Wuudis. Senop provides hyperspectral airborne imaging data, which is processed by MosaicMill before integration to Wuudis.

2.9.4 Experimentation

Quality monitoring

- Pilots are covering monitoring: work quality and damages
- An app is developed for METSAK called Laatumetsä (Quality Forest in English and Wuudis Qforest is the internal technical name for this). It will be published under Finnish Forest Centre brand and “Powered by Wuudis” will be present to market our own brand. IPR is shared.
- The integration with METSAK will start as soon they have interfaces ready. It is to be integrated with Suomi.fi Integration Bus (Palveluväylä). Currently, we are waiting for METSAK interface readiness.

Networking

- This feature means that Wuudis is integrated to the METSAK forest data interface, so their experts can use their internal forest data with Wuudis mobile app. Daily use of the Wuudis app has been ongoing since February 2018 within METSAK’s experts country wide.
2.9.4.1 Deployment
Wuudis is deployed in DataCenter Finland’s Service Cluster. Forestry TEP is currently deployed on the EO Cloud infrastructure of CloudFerro, in Poland. Metsään.fi is hosted by the Finnish Forest Centre in Finland. The systems of Senop and MosaicMill are similarly based in Finland.

2.9.4.2 Data sets
Data sets applicable to this pipeline are described in the following paragraphs.

Forest information: Wuudis service data model is based on the Finnish forest information standard. All development activities during the DataBio project that will affect to the Wuudis data model are based on Finnish forest information standard. Forest information standard includes a set of different standardized schemas (such as timber sales and logistics). Some of these schemas can be used in the DataBio and some new specifications are developed during project.

Map layers: One important dataset for Wuudis is different map layers. Wuudis uses global map services like Google and Microsoft (Bing) to provide world-wide satellite map layers to the end users. Wuudis also provides map layers from National Land Survey of Finland’s WMS/WMTS service.

Earth Observation data: Forest information layer, with various estimated parameters is produced by VTT, based on analysis of images from the optical Sentinel-2 satellites. The Forestry TEP provides also the satellite data - originally sourced from ESA - as well as a platform for processing and delivery of the information products.

Hyperspectral data: Imagery from Senop’s snapshot hyperspectral camera in VIS-VNIR spectral range.

2.9.4.3 Results
Results from development and testing the pipeline so far include:

- Integration between Wuudis and metsään.fi service, which allows the automated transfer of each forest stands on Wuudis platform.

- Acquisition of hyperspectral data by Senop, and processing by MosaicMill

Additionally, related:

- Wuudis app taken to daily use by METSAK’s experts since February 2018 for networking with forest owners while consultancy in their forests.

- Wuudis work quality monitoring app ready. Integration with suomi.fi service bus and interfacing with METSAK is only pending.

- Definition of Wuudis crowd sourcing app i.e., creating pipeline between Wuudis and METSAK.
2.9.5 Next steps
As a next step, we will consider integrating with suomi.fi service bus for one step login authentication. In addition, integrating with METSAK website for the visualization of crowdsourced data.

2.10 WP3 – Smart management of machinery power in small-pelagic fisheries vessel

2.10.1 General information

2.10.1.1 Objectives
This pilot aims to improve the operation of relatively complex machinery arrangements on-board small-pelagic fisheries vessel through presentation of measurement of current state and historic performance. The energy needs of the vessel for propulsion power, deck machinery, fish processing and general consumption are met by the same power generation system which on newer vessel can be configured to produce and distribute power in a variety of ways. The vessels machinery systems may meet the requirements of the crew in a variety of ways, but do not contain a feedback on efficiency or suggest actions to re-configure power production and distribution.

2.10.1.2 Diagrams and views
Dependencies and interfaces for this pilot are found the DataBio hub page: https://www.databiohub.eu/registry/#service-view/WP 3 Fishery Pilot A2. The pipeline view of this pilot is shown in Figure 39.

Figure 39: Pilot 3.2.2 - Pipeline view

The pilot is intended to operate on-board a vessel at sea with limited, to no, connectivity. The location view for this pilot is therefore shown as two diagrams in Figure 40 and Figure 41.
Figure 40: Pilot 3.2.2 – Location view trial 1

Figure 41: Pilot 3.2.2 - Intended location view in operation

2.10.1.3 Associated pilot
Pilot T3.3.2 - small pelagic fisheries immediate operational choices.

URL in DataBio Hub: https://www.databiohub.eu/registry/#service-view/WP%203%20Fishery%20Pilot%2023%20%5BA2%5D%20Small%20pelagic%20fisheries%20immediate%20operational%20choices/0.0.1

2.10.1.4 Reusability
The instrumentation and data collection facilities used by this pilot is shared among all the "small pelagic" fisheries pilots. The small pelagic fisheries pilots share the data collection pipeline, which is extensively used in this pilot. The other small pelagic fisheries pilots use the same pipeline as parts of their overall pipeline.

2.10.2 Components involved
The components involved are dealing with the collection and analysis of vessel monitoring system data and event processing. VTTs OpenVA will additionally be tried out as an expert tool in order to understand the relationships between the variables in the collected vessel monitoring system data.

2.10.2.1 1.3.1 Names
C17.01 Ratatosk – on-board data collection system (https://www.databiohub.eu/registry/#service-view/Ratatosk/2.1.1)

C17.02 STIM – Time series manipulation and processing (https://www.databiohub.eu/registry/#service-view/STIM/2.0.3)

C19.01 PROTON – Event processing (https://www.databiohub.eu/registry/#services?name=IBM)
2.10.2.2 Interfaces between components

Data Distribution System (DDS) is used between Ratatosk components and between Ratatosk (vessel monitoring system) and the A2 application.

netCDF is used for storage of vessel monitoring system data and storage of analyses performed by STIM. netCDF is then read by the A2 application.

2.10.3 Experimentation

The data collection pipeline is deployed on three of four vessels. The fourth vessel will be deployed in spring 2018 (actual date depending on power system supplier deliveries and port-calls for the vessel). The pilot will also experiment with VTTs OpenVA for understanding of the data generated on-board the vessels. Formulation of rules for PROTON is dependent on a proper understanding of the data in order to code event triggers. Initial experimentation of PROTON suffered of event rules which were not mutually exclusive. This is due to the number of possible variations under the hybrid power production plants on-board modern small pelagic vessels. Therefore, for next implementation of rules, special attention will be put in the articulation of mutually exclusive event rules.

2.10.3.1 Deployment

The data collection pipeline has been deployed on three of four vessels. The fourth deployment of the data collection pipeline is under final preparation. An initial A2 application is deployed on the three vessels, but without access to additional DataBio components (PROTON).

2.10.3.2 Data sets

Example data sets (collected data) on-board a single vessel is shown in Table 13. The signals shown in the table are sampled at 1 Hz and distributed by the Ratatosk framework to on-board storage (netCDF) and to the A2 application (current status). The stored netCDF files are processed by STIM (Smart Transducer Interface Module, from IEEE standard).
### Table 13: Pilot 3.2.2 - Collected data onboard a single vessel

<table>
<thead>
<tr>
<th>#NMEA</th>
<th>Navigation</th>
<th>MACHINERY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nav_Pos</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nav_Gyro</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nav_Wind</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nav_Depth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nav_Log</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Modbus ACON</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speed_Log</td>
<td>NAV NULL</td>
<td></td>
</tr>
<tr>
<td>Rudder_Angle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Draught_AP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Draught_FP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ME1_RPM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ME1_Consumption</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ME1_Load</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUX1_Consumption</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUX1_Load</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUX1_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUX1_KW_From_Gen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUX2_Consumption</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUX2_Load</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUX2_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUX2_KW_From_Gen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shaftline_Torque</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shaftline_Power</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propeller_RPM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propeller_Pitch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propeller_KW_Input</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bow_Thruster_KW_Input</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bow_Thruster_RPM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bow_Thruster_Pitch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freezing_Plant_KW_Input</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fish Factory_KW_Input</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engine_Room_Ventilation_Power_Input</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accomodation_Power_Input</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bow_Thruster_Consumed_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S40_Supply_Fan_Engine_Room_Kw</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S41_Supply_Fan_Engine_Room_Kw</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S52_Supply_Fan_Engine_Room_Kw</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydraulic_Pumps_Factory1_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydraulic_Pumps_Factory2_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MBB1_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MBB2_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFE_To_SG_PTI_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SG1_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROW_Supply1_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROW_Supply2_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Window_Heating_Bridge_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control_Cabinet_Freezing_Plant_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freezing_Compressor1_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freezing_Compressor2_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSW_Compressor_KW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydraulic_Pumps_Factory3_KW</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.10.3.3 Results
Current results are collected data, preparation for installation on-board the fourth vessel as well as evaluation of other components in the DataBio platform.

2.10.4 Next steps
The next steps will focus on:

- The deployment of the data collection on-board the last vessel.
- Improved rules for detection of undesired operation states for PROTON
- Testing connection to PROTON from Ratatosk enabled application (A2 application)
- Deploying initial A2 application on-board vessels

2.11 WP3 – Monitoring, real-time alerts, and visualization for operation efficiency in tuna fishery vessels

2.11.1 General information
The vessels involved in the Tuna Fisheries are usually operating in remote areas with complex machinery. The vessels involved in the project have installed a monitoring system on board that monitors different pieces of machinery: main engine, propeller and propulsion system, electric generation/consumption, fuel oil consumption, and fish freezing holds. In addition, benchmarking skippers, sister vessels with similar characteristics in same ocean are in operation in the company, with more than one skipper per ship. Vessels have evolved; more technical skippers may use different fishing strategies that may be evaluated not only with fishing catches; but with operational efficiency.
In the project’s vessels, all the measurements are recorded in a computer on board and uploaded every day to owner’s File Transfer Protocol (FTP) server.

The historical data of vessel operation is being used to design the data processing tools that are being prepared.

VTT OpenVA component will be used for historical and actual data visualization and analysis of all monitored parameters on board for a clear and simple evaluation of energy consumption and machinery performance on board. Due to big processing requirements, specific server and Hadoop environment will be necessary.

EXUS, through their Analytics Framework, will analyze main engine propulsion data to develop a Condition Based Maintenance tool that by analyzing the data uploaded to the server by the ship in a daily manner, detects anomalies and predicts possible faults in advance, prior to the occurrence. In this way crew and technical staff will be able to act in advance prior to the fault and avoiding possible damages to the machinery.

IBM will implement the PROTON system on board on line to work in real time to detect events and inform crew to act in advance and avoid critical machinery faults prior to the occurrence.

All above developments will be user centric and will be evaluated and validated by users.

2.11.1.1 Objectives

Objectives of the pipeline are:

- Improve vessel energy efficiency; reduce costs from fuel consumption on board while keeping the amount of catches.
- Reduce maintenance costs and ship downtime predicting machinery condition and possible events prior to occurrence employing machine learning techniques and machinery sensors readings. In addition, increase safety of ships and crew on board due to reduced unexpected failures in remote areas.
- Optimize loading of vessel in the different possible configurations to reduce the hull resistance and reduce fuel consumption. How the vessel loading affects vessel movements and ship resistance.
- Reduce fuel consumption by taking weather conditions into account to calculate fuel consumption in different possible fishing routes, from one point to other. Reduction of the fuel consumption by means of considering sea surface currents and meteorological conditions to make an intelligent sailing and include fuel consumption calculation as input to decision making by skippers.

2.11.1.2 Diagrams and views

URL to DataBio Hub: https://www.databiohub.eu/registry/#service-view/WP%203%20Fishery%20Pilot%20A1%20(EHU,%20UPV)/0.0.1
Figure 42: Pipeline view

Figure 43: Pilot 3.2.1.A1 - Lifecycle view
2.11.1.3 Associated pilot
Pilot 3.2.1.A1 – Oceanic tuna fisheries immediate operational choices (A1)
https://www.databiohub.eu/registry/#service-view/WP%203%20Fishery%20Pilot%2021%20[A1]:%20Oceanic%20tuna%20fisheries%20immediate%20operational%20choices/0.0.1

2.11.1.4 Reusability
Data input is tailored for this pilot; however, analytic components can be reused for different purposes. I.e., machine learning techniques used for condition-based maintenance could be later reused with different data inputs in different conditions and different ships.
2.11.2 Interfaces

- FTP interface to ship legacy system.
- Database interfaces to data storage.
- WEBHDFS interface to Hadoop file system storage (will be used in the second trial iteration).
- Apache YARN interface for running analysis in Hadoop system (will be used in the second trial iteration).
- User Interfaces of Open VA, IBM PROTON and EXUS described in the Experimentation paragraph.

PROTON interface

Currently the output of the CEP application is recorded both in output file for future analyses but also in a dashboard graphical application. It presents the situations or derived events emitted by the engine in real-time. Figure 45 below shows a screenshot of the dashboard showing the sensor reading input events and **LubricatingOilEngineInletPressureWarning** and **LubricatingOilEngineInletPressureAlarm** alerts and the reasoning behind the alert derivation. The alarm level situations are denoted with a red exclamation mark, while warning level situations are denoted with a yellow exclamation mark.

![Figure 45: PROTON interface-dashboard for displaying events and alarms](image)

2.11.3 Components involved

2.11.3.1 Names

C16.01 OpenVA (https://www.databiohub.eu/registry/#service-view/OpenVA/0.0.1)

C34.01 Exus analytics framework (https://www.databiohub.eu/registry/#service-view/EXUS%20Analytics%20Framework/0.0.1)

19.01 PROTON : https://www.databiohub.eu/registry/#services?name=IBM
2.11.3.2 Interfaces between components

**PROTON** - receives sensor readings related to main engine monitoring parameters from the ship’s monitoring and logging system. To this end, the DATALOGGER component tracks and logs main engine parameters (117 parameters each 10 seconds). The information is stored in the file system via FTP, from which it is read by Proton’s file adapter and streamed into Proton engine for processing.

2.11.4 Experimentation

In the experimentation 117 measurements from the engine and propulsion system are used. Measurements are recorded every 10 seconds and a file with records is uploaded to ship owner server every 24 hours. A new file is generated every day. Access to real time data from the vessel will be considered in a second phase. At the moment, it is not possible to predict machine failures in real time, since the data are not available in shore premises in real time.

Vessel route data is not yet available, because it is confidential. It may be possible to get history data of the routes to build optimization models in the next phase. Also, oceanographic and meteorological parameters are planned to be used in the next phase.

The data used in the experiments for Open VA consists of values of about two months of measurements delivered as CSV files. The data used for EXUS consist of values of up to 3 years of operation (depending on vessel) with measurements delivered as CSV files and ship crews’ log of maintenance and faults on board for cross validation of developed predictive models.

Henceforth, we describe the role of each component and its status in the current experimentation.

**OpenVA**

OpenVA imports the files from file system, purifies them and transforms the data to OpenVA database structure using Open Database Connectivity (ODBC) interface. The data is visualized in a web browser user interface.

In the next phase using Apache Hadoop file system as data backend, storage will be tested. This pipeline would enable integration to full set of Apache big data management tools provided by PSNC in the DataBio platform.

In first phase of the experimentation static images generated in the server are shown in the user interface. The user interface allows selecting an object to be analysed (in the experiment tuna ships) and measurements to be analysed. Based on the selections, OpenVA offers user suitable analysis methods. By using them, the user may run the selected analysis by given required parameters (e.g. time frame of the measurements). Example of the visualizations is shown in Figure 46.
EXUS Analytics Framework

EXUS Analytics Framework will be used in the framework of pilot 3.2.1.A1 to help predict boat engine failures using machine learning. EHU has provided EXUS with a description of the monitoring system, a description of the collected parameters, sample data files for initial analysis and summary of faults and events on board from crews’ log book. The dataset includes:

1. Data files with information of main engine and propulsion system in healthy state.
2. Definition of condition in which engine is in steady state to be used as reference for evaluation.
3. Faults occurred on board detected and corrected by crew and recorded in ship maintenance log book.

EXUS has begun the analysis of the provided dataset and has defined the better techniques for the data analysis. At the moment, EXUS has preliminary results that have to be refined for better performance. Specifically, EXUS is investigating the most appropriate time-window to be used in the fault detection module, which compares the estimated and actual engine parameters, such that true positives (actual failures) are predicted while keeping false positives low.

PROTON

PROTON is used to implement event-driven analysis that deals with engine condition monitoring, that is, proactive corrective operations in engine based on machinery sensors measurements. More specifically, the goal 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 data for the experimentation includes a CSV file corresponding to the record of one day of data from a vessel (8500 entries). Each row includes 117 different parameters/variables.
received from the different sensors located in the vessel. The variables are divided in some subgroups according to the characteristics of the data and the hardware used for recording. In this initial implementation we have focused on main engine parameters monitoring. This data is injected into the event processing engine as in real time.

PROTON possesses an authoring tool for the definitions of the event-driven application. The definitions include the specification of the entire event processing network (EPN) comprising of the input and output events, event processing agents (EPAs), consumers, and producers. Figure 47 shows a screenshot for the definition of the DecreasingTrendOilEnginePressure event processing agent for this application.

![Screenshot of event driven application definition](image)

**Figure 47: Event driven application definition screenshot**

The outputs of the application are recorded into a CSV file for future testing and analysis. In addition, to have a graphical user interface for the outputs, we apply PROTON’s dashboard to show the input events as well as the situations or output events emitted by the engine in real-time. Figure 48 below shows the input events (denoted by a blue i icon), and the situations (warnings are denoted with the yellow exclamation icon while alarms are denoted with the red exclamation icon).
Figure 48: Event driven application dashboard screenshot

Figure 49 depicts an overall view of the offered solution. As already expressed, we use PROTON’s authoring tool for the definition of the event rules. As the producer of the input events, we read the events from a CSV file that stores historical records of 117 measurements from the engine and propulsion system (one recorded input event every 10 seconds). As event consumers, we apply PROTON’s dashboard and a CSV file that records the events and the situations emitted by PROTON.

Figure 49: Event driven application overall view
2.11.4.1 Deployment

![Diagram showing components and locations for deployment]

Figure 50: Pilot 3.2.1.A1 - Location view

2.11.4.2 Data sets

Data sets contain 117 measurements from the engines and propulsion systems from three vessels. Measurements are mostly recorded every 10 seconds. Test data used in first trial in Open VA are collected during a three months period.

In EXUS Analytic Framework data used is data from 3 years of operation of one of the vessels of the ship owner fleet. There are three sister vessels with similar monitoring system and similar data sets.

Data has been exported from vessel legacy system as CSV files.

2.11.4.3 Results

The development of the entire pipeline is still under development and therefore there are still no results to report.

2.11.5 Next steps

Closer integration of the components, using PSNC Hadoop environment to run analysis.

IBM PROTON will be implemented on board in a dedicated PC for monitoring and event processing of propulsion data in real time.

EXUS will define best analytic solution and will later be tested periodically with new vessel data sets collected nowadays in actual operating conditions.

2.12 WP3 - Data acquisition and analyses for fishing operational efficiency

2.12.1 General information

2.12.1.1 Objectives

The objective of this pipeline is to demonstrate that the combination of information from various assets can be used to produce better population dynamics estimates for pelagic species and provide decision support to fishermen and chiefs of fishing operation. It is anticipated that a crowdsourced data collection effort from fishing vessels combined with
public/private data assets and data analytics can increase both the accuracy and precision of stock assessments and predictions on future efficacy of operation and catch.

### 2.12.1.2 Diagrams and views

The pipeline, lifecycle, and role views for the pilots are shown in Figure 51, Figure 52, and Figure 53 below. The associated URL on DataBio Hub is [DataBio Hub Pipeline Fishery B2/C1](#) (You must be logged in and member of DataBio Pipelines group to view the pipeline. Notice also that it inherits from a common pipeline, which is a common foundation across pilots B2, C1, C2.

![Diagram](image_url)

*Figure 51: Pilot 3.3.2 and 3.4.2 - Pipeline View*
Figure 52: Pilot 3.3.2 and 3.4.2 - Lifecycle view

Figure 53: Pilot 3.3.2 and 3.4.2 - Role view
2.12.1.3 Associated pilots

- Pilot 3.3.2: Small pelagic fisheries planning [DataBio Hub]
- Pilot 3.4.2: Pelagic fish stock assessments [DataBio Hub]

2.12.1.4 Reusability

The core of this pipeline is intended to be common to all the four pilots in small pelagic fishery, e.g. Pilots 3.2.2, 3.3.2, 3.4.1 and 3.4.2 (this one), that all share common infrastructure (logging system, data centre) and have similar needs for data management, processing, analysis and visualization. The data acquisition, management and serving will share a common pipeline, while each pilot has unique requirements for certain datasets, analysis techniques and display methods. However, the framework for analysis and visualization is shared, although some datasets, algorithms and display methods may differ.

2.12.2 Components involved

2.12.2.1 Names

Open source components for data management and integration:

- CouchDB as local database for metadata and thin data
- NGINX to serve large data, i.e. referred through metadata
  - o [https://www.databiohub.eu/registry/#service-view/NGINX/0.0.1](https://www.databiohub.eu/registry/#service-view/NGINX/0.0.1)
- Vespa to serve datasets searchable and filter on geographic location and time
  - o [https://www.databiohub.eu/registry/#service-view/VESPA/0.0.1](https://www.databiohub.eu/registry/#service-view/VESPA/0.0.1)

DataBio partner components:

- Ratatosk (C17.01) for logging vessel data and hydroacoustic data
- STIM (C17.02) for initial analysis of vessel data
- DataGraft (C06.01) for data linking and semantic relationships
- SINTIUM (C06.02) for data visualization
- Supervised Learning Algorithm (C01.01) for some hydroacoustics classification

Other components:

- Market and Fishery data scraper and conversion (tool developed as part of pilot implementation)
- SINMOD oceanographic model (Background software from SINTEF)
  - o [https://www.databiohub.eu/registry/#service-view/SINMOD%20Oceanographic%20model/0.0.1](https://www.databiohub.eu/registry/#service-view/SINMOD%20Oceanographic%20model/0.0.1)

Fish Stock Assessment Component and Fish Prediction Component: To be decided, either an open source software or a suitable DataBio partner component [DataBio Hub]

2.12.2.2 Interfaces between components

Dataset formats:
- JSON is used for thin data like market data sets and metadata for large data, e.g. source, location and timestamp of meta data. Public statistical data, e.g. World Bank, Eurostat, Statistics Norway have REST interfaces serving JSON and both VESPA and couchDB support JSON as the document format. Moreover, statistical organizations use JSON-stat.
- netCDF is used for Earth Observation (EO), weather, marine, oceanographic, and hydroacoustic data.
- Many sensors have proprietary data formats, for example hydroacoustic fish finding sensors and implementation is ongoing to adapt to a common data format used for other data sources on board the vessels (see Figure 54).

Figure 54: Class diagram for hydroacoustics Ratatoskr component module.

- The Data Distribution Service (DDS) standard from OpenGroup is used for real-time data logging from vessels.
  - This publish-subscribe pattern is used as the adapter to various data communication formats found on board vessels, including NMEA, Modbus, and proprietary protocols.

Data management and serving:
• JSON for metadata, market and fishery data and statistics.
• netCDF where possible for marine data, EO, and logged time series data
• NGINX will be used to serve historical data linked from searchable meta data, including netCDF and raw data where needed.

2.12.3 Experimentation
The pipeline is new and not based on previous work or project results, so the focus is on establishing the data feeds for hydroacoustic data and earth observation services. This involves linking data sources and do preparation for explorative predictive analytics. As such, the pilot is not yet ready to hand over the data needed to evaluate analytics components from WP4/5. Trial 1 is intended to explore and identify components and algorithms from which a preferred implementation will be carried out in Trial 2.

2.12.3.1 Deployment
The Location view in Figure 55 indicates where the components are installed. Since the pipeline has not been deployed yet, the plan is subject to changes. The initial pilot development has been prioritized towards data acquisition for the unique data sources needed for these pilots, and the current situation is that programmatic access to the data services have been added to the data scraper and work on data acquisition from hydroacoustics is in progress. The pipeline is to be deployed in the SINTEF Marine Datacenter, and additional deployment of VESPA and other components are being discussed with the PSNC data centre partner.
Figure 55: Pilot 3.3.2 and 3.4.2 - Location view

2.12.3.2 Data sets
No data sets to report. The plan is to ensure that hydroacoustic netCDF follows the CF-1.7 (Climate and Forecast) convention for discrete sampling geometries\(^9\).

2.12.3.3 Results
No results yet.

2.12.4 Next steps
The next steps will focus on:

- The deployment of the data scraper to get regularly update data feeds.
- Meta data and standard formats to make datasets searchable (latitude, longitude, time overlap) and served by Vespa.
- Map-based interface to interactively display such datasets using SINTIUM
- Finish and deploy to vessels the hydroacoustic logging component module

\(^9\) http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.html#discrete-sampling-geometries
2.13 WP3 - Data analytics for prediction of pelagic market segments behaviour

2.13.1 General information

2.13.1.1 Objectives

The goal of this pilot is to provide information for predicting the development of various pelagic market segments, so that the fisheries may be targeted against the species that will give the highest yield given the predicted economic outlook. The basis for the market predictions will be to combine as many data sources as possible relevant for price development, for example: predicted volume of species, international and national financial data, state of competing fish stocks and predicted fished volume of these and prices of alternative protein food. Machine learning and predictive analytics will be employed to model consumption habits as well as the relation between market development and other factors. These models will then be used to provide predictions for how various market segments will develop in the future. For further details on background and motivation for this pilot, consult chapter 8 in DataBio public deliverable D3.1 – Fishery Pilot Definition.

2.13.1.2 Diagrams and views

The role, pipeline and lifecycle views for the 3.4.2 pilot are shown in Figure 56, Figure 57, and Figure 58, respectively.

![Figure 56: Pilot 3.4.2- Role view](image-url)
Figure 57: Pilot 3.4.2 - Pipeline view

Note that the focus on trial 1 (2018) is on data acquisition, management and serving and, hence, the Market Prediction Component has yet not been decided as analysis, prediction and visualization are focused in trial 2 (2019).
2.13.1.3 Associated pilot

Main pipeline association:

- Pilot 3.4.2 (Fishery C2 / 26): Small pelagic market predictions and traceability:
  - URL in DataBioHub: [https://www.databiohub.eu/registry/#service-view/WP%203%20Fishery%20Pilot%2026%20[C2]:%20Small%20pelagic%20market%20predictions%20and%20traceability/0.0.1](https://www.databiohub.eu/registry/#service-view/WP%203%20Fishery%20Pilot%2026%20[C2]:%20Small%20pelagic%20market%20predictions%20and%20traceability/0.0.1)

Related Fishery pilots sharing core pipeline components with pilot 3.4.2 (by relevance):

- Pilot 3.3.2 (Fishery B2 / 24):
  - URL in DataBioHub: [https://www.databiohub.eu/registry/#service-view/WP%203%20Fishery%20Pilot%2024%20[B2]:%20Small%20pelagic%20fisheries%20planning/0.0.1](https://www.databiohub.eu/registry/#service-view/WP%203%20Fishery%20Pilot%2024%20[B2]:%20Small%20pelagic%20fisheries%20planning/0.0.1)

- Pilot 3.4.1 (Fishery C1 / 25): Small pelagic fish stock assessment
  - URL in DataBioHub: [https://www.databiohub.eu/registry/#service-view/WP%203%20Fishery%20Pilot%2025%20[C1]:%20Pelagic%20fish%20stock%20assessments/0.0.1](https://www.databiohub.eu/registry/#service-view/WP%203%20Fishery%20Pilot%2025%20[C1]:%20Pelagic%20fish%20stock%20assessments/0.0.1)

- Pilot 3.2.2 (Fishery A2 / 23): Small pelagic fisheries immediate operational choices
  - URL in DataBioHub: [https://www.databiohub.eu/registry/#service-view/WP%203%20Fishery%20Pilot%2023%20[A2]:%20Small%20pelagic%20fisheries%20immediate%20operational%20choices/0.0.1](https://www.databiohub.eu/registry/#service-view/WP%203%20Fishery%20Pilot%2023%20[A2]:%20Small%20pelagic%20fisheries%20immediate%20operational%20choices/0.0.1)
A common pipeline (set of components) have been made in DataBio Hub for the small pelagic pilots to identify common components for these pilots vs. the unique ones per pilot:

- URL in DataBio Hub: https://www.databiohub.eu/registry/#service-view/WP%203%20Small%20Pelagic%20Fisheries%20Pipeline/0.0.1

2.13.1.4 Reusability
The core of this pipeline is intended to be common to all the four pilots in small pelagic fishery, e.g. Pilots 3.2.2, 3.3.2, 3.4.1 and 3.4.2 (this one), who all share common infrastructure (logging system, data centre) and have similar needs for data management, processing, analysis and visualization. The data acquisition, management and serving will share a common pipeline, while each pilot has unique requirements for certain datasets, analysis techniques and display methods. However, the framework for analysis and visualization is shared, although some datasets, algorithms, and display methods may differ.

2.13.2 Components involved

2.13.2.1 Names
Open source components for data management and integration:

- CouchDB as local database for metadata and thin data
- NGINX to serve large data, i.e. referred through metadata
  - [https://www.databiohub.eu/registry/#service-view/NGINX/0.0.1](https://www.databiohub.eu/registry/#service-view/NGINX/0.0.1)
- Vespa to serve datasets searchable and filter on geographic location and time
  - [https://www.databiohub.eu/registry/#service-view/VESPA/0.0.1](https://www.databiohub.eu/registry/#service-view/VESPA/0.0.1)

DataBio partner components:

- Ratatosk (C17.01) for logging vessel data ([https://www.databiohub.eu/registry/#service-view/Ratatosk/2.1.1](https://www.databiohub.eu/registry/#service-view/Ratatosk/2.1.1))
- STIM (C17.02) for initial analysis of vessel data ([https://www.databiohub.eu/registry/#service-view/STIM/2.0.3](https://www.databiohub.eu/registry/#service-view/STIM/2.0.3))
- DataGraft (C06.01) for data linking and semantic relationships ([https://www.databiohub.eu/registry/#service-view/DataGraft/0.0.1](https://www.databiohub.eu/registry/#service-view/DataGraft/0.0.1))
- SINTIUM (C06.02) for data visualization ([https://www.databiohub.eu/registry/#service-view/SINTIUM/0.0.1](https://www.databiohub.eu/registry/#service-view/SINTIUM/0.0.1))

Other components:

- Market and Fishery data scraper (tool developed as part of pilot implementation)
- SINMOD oceanographic model (Background software from SINTEF)
  - [https://www.databiohub.eu/registry/#service-view/SINMOD%20Oceanographic%20model/0.0.1](https://www.databiohub.eu/registry/#service-view/SINMOD%20Oceanographic%20model/0.0.1)
- Market (price) prediction component: Specific component not yet decided, either an open source software or a suitable DataBio partner component?
2.13.2.2 Interfaces between components

Dataset formats:

- JSON is used for thin data like market data sets and metadata for large data, e.g. source, location and timestamp of meta data. Public statistical data, e.g. World Bank, Eurostat, Statistics Norway have REST interfaces serving JSON and both VESPA and couchDB support JSON as the document format. Moreover, statistical organization
- netCDF is used for EO, weather, marine and oceanographic data.
- Many sensors have proprietary data formats, for example hydroacoustic fish finding sensors.
- The DDS data standard from OpenGroup is used for real time data logging from vessels.

Data management and serving:

- JSON for metadata, market and fishery data and statistics.
- netCDF where possible for marine data, EO, and logged time series data
- NGINX will be used to serve historical data linked from searchable meta data, including netCDF and raw data where needed.

Example data:

- World Bank: economy and growth (topic 3), financial sector (topic 7) and trade (topic 21) indicators:
  - Method: POST
  - The MIME type: application/json
  - Example data from data scraper (only GDP variants for topic 3 shown):
    
    ```json
    { "summary": { "NY.GDP.PCAP.PP.KD": "GDP per capita, PPP (constant 2011 international $)",
    "NY.GDP.PCAP.PP.PC": "GDP per capita, PPP (current international $)",
    "NY.GDP.PCAP.KN": "GDP per capita (constant LCU)",
    "NY.GDP.PCAP.KD.ZG": "GDP per capita growth (annual %)",
    "NY.GDP.PCAP.KD": "GDP per capita (constant 2010 US$)",
    "NY.GDP.PCAP.CN": "GDP per capita (current LCU)",
    "NY.GDP.PCAP.CD": "GDP per capita (current US$)",
    "NY.GDP.MKTP.PP.KD": "GDP, PPP (constant 2011 international $)",
    "NY.GDP.MKTP.PP.CD": "GDP, PPP (current international $)",
    "NY.GDP.MKTP.KN": "GDP (constant LCU)",
    "NY.GDP.MKTP.KD.ZG": "GDP growth (annual %)",
    "NY.GDP.MKTP.KD": "GDP (constant 2010 US$)",
    "NY.GDP.MKTP.CN.AO": "GDP: linked series (current LCU)",
    "NY.GDP.MKTP.CN": "GDP (current LCU)",
    "NY.GDP.MKTP.CD": "GDP (current US$)",
    "NY.GDP.FCST.KN": "Gross value added at factor cost (constant LCU)",
    "NY.GDP.FCST.KD": "Gross value added at factor cost (constant 2010 US$)",
    "NY.GDP.FCST.CN": "Gross value added at factor cost (current LCU)",
    ```
"NY.GDP.FCST.CD": "Gross value added at factor cost (current US$)",
"NY.GDP.DISC.KN": "Discrepancy in expenditure estimate of GDP (constant LCU)",
"NY.GDP.DISC.CN": "Discrepancy in expenditure estimate of GDP (current LCU)",
"NY.GDP.DEFL.ZS": "GDP deflator (base year varies by country)",
"NY.GDP.DEFL.KD.ZG": "Inflation, GDP deflator (annual %)",
...} }

- The data scraper can fetch a series of indicators for a specified set of countries and years which then is stored as vectors in couchDB.

- **ICES: Fishery species example**
  - Method: SOAP
  - URL: [http://ecosystemdata.ices.dk/WebServices/EcoSystemWebServices.asmx](http://ecosystemdata.ices.dk/WebServices/EcoSystemWebServices.asmx)
  - The MIME type: application/XML
  - Example JSON data produced from data scraper accessing the ICES SOAP service (only a couple species shown):

```json
{"species": [{"speciesName": "Melanogrammus aeglefinus", "key": 669, "numMeasurements": 637273, "numYears": 51, "dateFirstMeasurement": "30/03/1965", "dateLastMeasurement": "24/01/2015"}, {"speciesName": "Pecten maximus", "key": 711, "numMeasurements": 150, "numYears": 5, "dateFirstMeasurement": "01/02/2006", "dateLastMeasurement": "09/02/2012"}, ...
]}
```

### 2.13.3 Experimentation

The 3.4.2 pilot is initializing up its development work time during pilot trial 1, i.e. not based on previous work or project results, so the main focus is on establishing the data feeds for catch and market data and do explorative predictive analytics work. As such, the pilot is not yet ready to hand over the data needed to evaluate analytics components from WP4/5. Trial 1 is intended to explore and identify components and algorithms from which a preferred implementation will be carried out in trial 2.

The experimentation for this pilot has so far focused on data acquisition and collaboration tool development for the relevant economic and market data, as well as the initial pipeline design. The development has been focused on the Market and fishery data scraper which now access the following data sources: Slides private catch data, World Bank, Comtrade, Eurostat and Statistics Norway economic data, ICES and Eumofa fishery data.

#### 2.13.3.1 Deployment

The pipeline has not yet been deployed, it is an architecture outline/design overview of what is the proposed pilot architecture. The initial pilot development has been prioritized towards data acquisition for the unique data sources needed for this pilot, and the current situation is that programmatic access to the afore-mentioned data services has been added to the data scraper and that the results are stored in CouchDB as JSON documents.

The pipeline is to be deployed in the SINTEF Marine Datacenter, and additional deployment of VESPA and other components are being discussed with the PSNC data centre partner.
2.13.3.2 Data sets
Data sets are standardized on JSON format, including the more compact JSON-stat format used by some of the statistical services. Update frequency varies from yearly, quarterly, monthly, weekly and daily for these data sets. The market data and catch reports are not big in volume compared to sensor data, e.g. in particular hydroacoustic data, but the main challenge here is to collate, manage and serve a diverse set of data sources.

2.13.3.3 Results
The results so far are limited to data download, conversion and storage using CouchDB from the previous mentioned sources. Data rates and regular update schedules and mechanisms are not yet in place for the pilot but are among the next priorities.

2.13.4 Next steps
The next steps will focus on:

- The deployment of the data scraper to get regularly updated data feeds.
- Meta data and standard formats to make datasets searchable (latitude, longitude, time overlap) and served by Vespa.
- Map based interface to interactively display such datasets using SINTIUM.

2.14 Pipelines summary
At M16 of the project we have a total of 10 pipelines in WP4 at different levels of maturity, ranging from data acquisition to real data experimentation and results.

Table 14 shows the pipelines name along with associated pilot(s) and components.

Each pipeline has a clear work plan with regards to next steps to be done, so the pipeline can be deployed and be tested with real data.

From the project point of view, next phase involves two important interrelated aspects:

- The identification of dedicated (sub) pipelines applicable to one or more pilots and project domains. These are tailored to meet specific pilot requirements, usually because the components in the pipelines are dedicated to a specific domain and requirements. In parallel, the identification of cross reusable (sub) pipelines (“design patterns”) that can be used across the pilots of the project and beyond, that is, to be applicable to other domains. These two activities complement each other and have a direct impact on the exploitable assets of the project.
- The identification of (standardized) interfaces that can be used across these sub-sets of components. For further information on the platform interfaces and standards refer to D4.1.
Table 14: WP4 pipelines summary

<table>
<thead>
<tr>
<th>#</th>
<th>WP#</th>
<th>Pipeline name</th>
<th>Associated pilot</th>
<th>components involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Metadata, Linked data and graph data</td>
<td>B2.1: Machinery management</td>
<td>C02.02: Micka</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C02.03: HSLayers Next Generation (NG)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C02.04: HSLayers Mobile</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Prediction and real-time alerts of diseases and pests breakouts in crops</td>
<td>1.2.1.A1.1: Precision agriculture in olives, fruits, grapes</td>
<td>C13.03: GAIABus Real Time Subcomponent GAIA cloud</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C19.01: PROTON</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Big data analysis for decision making support for CAP and Insurance</td>
<td>1.4.2C2.2: CAP Support (Greece) Pilot 1.4.1C1.1: Insurance (Greece)</td>
<td>C13.03: Neurcode</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Note: This pipeline shared also by WP5 (D5.2)</td>
<td></td>
<td>C31.01: Neural Network Suite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C13.02: GAIABus Datasmart Machine Learning Subcomponents GAIA Cloud</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>Field data analysis and real-time alerting for decision making for precise agriculture</td>
<td>1.3.1.B1.1: Cereals and biomass crop</td>
<td>C05.02: IoT hub</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C19.01: PROTON</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>IoT data acquisition, analyses, and real-time alerts of potential threat conditions for crops</td>
<td>1.3.1.B1.4: Cereals and biomass crops</td>
<td>C02.01: SensLog</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C19.01: PROTON</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>Publication of Linked Data related to cereals and biomass crops</td>
<td>Pilot B1.4 - Cereals and biomass crops_4</td>
<td>C09.12: Openlink Virtuoso.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C02.03: HS Layers NG.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C12.01: LIMES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C02.01: SensLog</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>IoT data acquisition, analyses, and prediction of potential threat conditions for crops</td>
<td>1.3.1.B1.4: Cereals and biomass crops</td>
<td>C02.01: SensLog</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C12.02 a: Albatross</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>Shared multiuser forest data environment</td>
<td>2.4.2: Shared multiuser forest data environment</td>
<td>C18.01: Metsään.fi eService</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C18.02: Open forest data service</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C20.01: Wuudis</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>Multi-source data enabling efficient forest management</td>
<td>Pilot 2.2.1: Easy data sharing and networking Pilot 2.2.2: Monitoring and control tools for forest owners</td>
<td>C20.1: Wuudis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C16.10: Forestry Thematic Exploitation Platform</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C18.01: Metsään.fi eService</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C44.01: Senop Hyperspectral Camera</td>
</tr>
<tr>
<td>Page</td>
<td>Task</td>
<td>Description</td>
<td>Related Tasks</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>-------------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>Smart management of machinery power in small-pelagic fisheries vessel</td>
<td>T3.3.2: Small pelagic fisheries immediate operational choices</td>
<td></td>
</tr>
</tbody>
</table>
|      |      | C44.02: EnsoMOSAIC Fusion software | C17.01: Ratatosk  
|      |      | C17.02: STIM  
|      |      | C19.01: PROTON |
| 11   | 3    | Monitoring, real-time alerts, and visualization for operation efficiency in tuna fishery vessels | Pilot 3.2.1.A1 – Oceanic tuna fisheries immediate operational choices (A1) |
|      |      | C16.01: OpenVA  
|      |      | C09.13: PSNC computing platform and Apache Hadoop stack  
|      |      | C19.01: PROTON  
|      |      | C34.01: Exus analytics framework |
| 12   | 3    | Data acquisition and analyses for fishing operational efficiency | 3.3.2: Small pelagic fisheries planning  
|      |      | 3.4.2: Pelagic fish stock assessments |
|      |      | C17.01: Ratatosk  
|      |      | C17.02: STIM  
|      |      | C06.01: DataGraft  
|      |      | C06.02: SINTIUM  
|      |      | C01.01: Supervised Learning Algorithm  
|      |      | Data management and integration: COUCHDB, NGINX, Vespa  
|      |      | Market and Fishery data scraper and conversion SINMOD oceanographic model |
| 13   | 3    | Data analytics for prediction of pelagic market segments behaviour | 3.4.2 (Fishery C2 / 26): Small pelagic market predictions and traceability  
|      |      | Related Fishery pilots sharing core pipeline components: 3.3.2, 3.4.1, and 3.2.2 |
|      |      | C17.01: Ratatosk  
|      |      | C17.02: STIM  
|      |      | C06.01: DataGraft  
|      |      | C06.02: SINTIUM  
|      |      | C01.01: Supervised Learning Algorithm  
|      |      | Data management and integration: COUCHDB, NGINX, Vespa  
|      |      | Market and Fishery data scraper and conversion SINMOD oceanographic model  
|      |      | Market (price) prediction component |
3 Conclusions and future work

At the core of the DataBio platform is the ability to support all project pilots’ requirements. From the early stages of the project it was clear that many times one component does not provide all required functionalities by a pilot and therefore there is need to orchestrate some components that together fulfil some specific functionalities. These sets of components that provide well defined interfaces and means of deployment to meet some business goals are dubbed pipelines.

The main goal of D4.2 is twofold. First, to provide a comprehensive overview of the pipelines’ status in M16 of the project and their level of readiness towards their first trial in M17. Second, to provide specific guidelines for their successful implementation and deployment. To this end, pipelines descriptions are detailed according to a specified template and aim to give enough information about both the pipeline status as well as guidelines for a successful set up of the pipeline.

The work on the pipelines has been progressive and is the outcome of an ongoing collaboration between pilot and component owners. Currently we have a total of 13 pipelines. It is expected that the work will continue, and refinements will be introduced to the existing pipelines due to experimentation.

Each pipeline has a clear work plan with regards to next steps to be carried out, so the pipeline can be deployed and tested in a realistic scenario with real data.

From project perspective, next phase includes the identification of commonalities among these pipelines in terms of components and interfaces. The pipelines can be seen as important assets for exploitation at the end of the project. The goal is to have standardized APIs and reusable “assets” or services not just in the three bioeconomy domains of the project (agriculture, forestry, and fishery), but in other domains as well.

This document relates to the pipelines in the scope of WP4. Analogously, the deliverable D5.2 relates to the pipelines related to the WP5 components. Deliverable D4.3 “Data sets, formats and models” to be submitted in M20 will include updates and refinements to the pipelines. The technical deliverables in the project aim at providing all technical support to successful experimentations phases of the pilots.
Appendix A  Pipelines description template

A.1  Pipeline name

A.1.1  General information
This section will include general information with regards the pipeline

A.1.1.1  Objectives
What is the purpose of this pipeline? Why it has been created?

A.1.1.2  Diagrams and views
Insert here the URL of the pipeline in DataBio Hub if exists otherwise the diagram (s) in Modelio

A.1.1.3  Associated pilot
What is the pilot name and identification this pipeline belongs to? + URL to the pilot design in DataBio Hub/Modelio diagram

A.1.1.4  Reusability
State whether this pipeline is specifically tailored to the pilot or can be reused in different settings

A.1.2  Interfaces
The interfaces into the pipeline and out of the pipeline (interfaces between the components in the flow are described later in a separate section).

A.1.3  Components involved

A.1.3.1  Names
What are the components that are part of this pipeline? List of names + identifier (e.g., C019) of the components in the pipeline + URLs to the DataBio Hub

A.1.3.2  Interfaces between components
Interface from one component to the other

A.1.4  Experimentation
We would like to include anything we have with regards to tests made with this pipeline up-to-now

A.1.4.1  Deployment
Where the components are installed? Any relevant information re the invocation of the pipeline service should go here.

A.1.4.2  Data sets
What data served for the experiments? Describe the data structure (input and output) along with the data set characteristics, e.g., volume and rate.
A.1.4.3 Results
Results so far from testing

A.1.5 Next steps
What are the plans for next experimentation with this pipeline.