Big Data to Enable Global Disruption of the Grapevine-powered Industries

D1.3 - Annual Public Report

DELIVERABLE NUMBER	D1.3
DELIVERABLE TITLE	Annual Public Report
RESPONSIBLE AUTHOR	Pythagoras Karampiperis (Agroknow)

Co-funded by the Horizon 2020
Framework Programme of the European Union

GRANT AGREEMENT N.	780751
PROJECT ACRONYM	BigDataGrapes
PROJECT FULL NAME	Big Data to Enable Global Disruption of the Grapevine-powered industries
STARTING DATE (DUR.)	01/01/2018 (36 months)
ENDING DATE	31/12/2020
PROJECT WEBSITE	http://www.bigdatagrapes.eu/
COORDINATOR	Pythagoras Karampiperis
ADDRESS	110 Pentelis Str., Marousi, GR15126, Greece
REPLY TO	pythk@agroknow.com
PHONE	+30 210 6897 905
EU PROJECT OFFICER	Mr. Riku Leppanen
WORKPACKAGE N. | TITLE	WP1 | Project Management
WORKPACKAGE LEADER	Agroknow
DELIVERABLE N. | TITLE	D1.3 | Annual Public Report
RESPONSIBLE AUTHOR	Pythagoras Karampiperis (Agroknow)
REPLY TO	pythk@agroknow.com
DOCUMENT URL	http://www.bigdatagrapes.eu/
DATE OF DELIVERY (CONTRACTUAL)	31 December 2018 (M12)
DATE OF DELIVERY (SUBMITTED)	20 December 2018 (M12)
VERSION | STATUS	1.0 | Final
NATURE	Report (R)
DISSEMINATION LEVEL	Public (PU)
AUTHORS (PARTNER)	Aikaterini Kasimati (AUA), Maritina Stavrakaki (AUA), Coraline Damasio (INRA), Florian Schlenz (GEOCLEDIAN), Simone Parisi (Abaco), Eleni Foufa (APIGEA), Pythagoras Karampiperis (Agroknow), Panagiotis Zervas (Agroknow), Milena Yankova (ONTOTEXT), Nikola Rusinov (ONTOTEXT) Raffaele Perego (CNR), Nicola Tonellotto (CNR), Franco-Maria Nardini (CNR), Nyi-Nyi Htun (KULeuven)
CONTRIBUTORS	-
REVIEWER	All partners

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Table of Contents	03/12/2018	Pythagoras Karampiperis (Agroknow), Panagiotis Zervas (Agroknow)
0.4	Section 1, 2, 3, 5	10/12/2018	Aikaterini Kasimati (AUA), Maritina Stavrakaki (AUA), Coraline Damasio (INRA), Florian Schlenz (GEOCLEDIAN), Simone Parisi (Abaco), Eleni Foufa (APIGEA), Pythagoras Karampiperis (Agroknow
0.6	Section 4	14/12/2018	Pythagoras Karampiperis (Agroknow), Panagiotis Zervas (Agroknow), Milena Yankova (ONTOTEXT), Nikola Rusinov (ONTOTEXT) Raffaele Perego (CNR), Nicola Tonellotto (CNR), Franco-Maria Nardini (CNR), Nyi-Nyi Htun (KULeuven)
1.0	Final version	20/12/2018	Pythagoras Karampiperis (Agroknow), Panagiotis Zervas (Agroknow)

PARTICIPANTS		CONTACT
Agroknow IKE (Agroknow, Greece)	Agroknow	Pythagoras Karampiperis Email: pythk@agroknow.com
Ontotext AD (ONTOTEXT, Bulgaria)		Todor Primov Email: todor.primov@ontotext.com
Consiglio Nazionale DelleRicherche (CNR, Italy)		Raffaele Perego Email: raffaele.perego@isti.cnr.it
Katholieke Universiteit Leuven (KULeuven, Belgium)	KATHOLIEKE UNIVERSITEIT LEUVEN	Katrien Verbert Email: katrien.verbert@cs.kuleuven.be
Geocledian GmbH (GEOCLEDIAN Germany)	geo|cledian	Stefan Scherer Email: stefan.scherer@geocledian.com
Institut National de la Recherché Agronomique (INRA, France)	INRA SCIENCE & IMPACT	Pascal Neveu Email: pascal.neveu@inra.fr
Agricultural University of Athens (AUA, Greece)		Katerina Biniari Email: kbiniari@aua.gr
Abaco SpA (ABACO, Italy)		Simone Parisi Email: s.parisi@abacogroup.eu
APIGAIA (APIGEA, Greece)		Eleni Foufa Email: Foufa-e@apigea.com

ACRONYMS LIST

APIs	Application Programming Interfaces
AUA	Agricultural University of Athens
BDG	BigDataGrapes
UEPR	Unit of Pech Rouge
LDBC	Linked Data Benchmark Council
DSS	decision support systems
PA	Precision Agriculture

EXECUTIVE SUMMARY

This report presents the BigDataGrapes project vision and ambition and summarises the project's advancements achieved during the first year of its execution. Its target audience comprises representatives of external interested communities as well as the general public.

The document summarizes the technical and implementation details of the BigDataGrapes infrastructure, describes the technical choices and the rationale behind them, and discusses the research and scientific particularities faced by each of the BigDataGrapes pilots.

More specifically, the document establishes the main objectives of the BigDataGrapes project and summarizes the core outcomes of the four pilot communities represented in the project. Based on these two axes, the technical advancements achieved during the reporting period are contextualised and discussed at a high level. Finally, the report concludes with the presentation of the foreseen next steps and the progress to be achieved during the upcoming second year of the project.

The document is structured as follows. Chapter 1 provides an introduction to the main issues tackled by the BigDataGrapes project, whereas Chapter 2 provides an overview of the use cases with details on the methodology followed in order to define them and associate them with BigDataGrapes pilots. Chapter 3 identifies the annual progress of each of the four selected pilots, while chapter 4 presents the technical advancements of the project. Finally, the last Chapter, discusses the next steps that will be implemented in order to improve pilots' yield during the upcoming second year of the project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

BigDataGrapes is a 36-month Research and Innovation action, supported by the European Commission through the H2020 Research and Innovation programme, under grant agreement no. 780751.

BigDataGrapes aspires to help European companies in the wine and natural cosmetics industries become more competitive in the international markets. Specifically, it tries to help companies across the grapevine-powered value chain ride the big data wave, supporting business decisions with real time and cross-stream analysis of very large, diverse and multimodal data sources.

In particular, BigDataGrapes aims to improve the competitive positioning of companies in the European IT sector that are serving companies and organizations with software applications:

- Software companies developing farm management and precision agriculture systems for companies in the agriculture sector.
- Software companies developing food risk assessment monitoring and prediction systems for companies in the food sector.
- Software companies developing quality control and compliance software for companies in the beauty and cosmetics sector.

To this end, the project develops, extends and provides the necessary specifications, mechanisms, fault-tolerant tools and components for allowing the rapid and intuitive development of variegating data analysis workflows, where the functionalities for data collection and storage, dataset creation, results visualization and deployment are provided by specialized services utilizing European large-scale, cloud-based infrastructures.

Thus, the vision of BigDataGrapes project is to manage technology challenges of the grapevine-powered data economy as its business problems and decisions requires processing, analysis and visualisation of data with rapidly increasing volume, velocity and variety: satellite and weather data, environmental and geological data, phenotypic and genetic plant data, food supply chain data, economic and financial data and more. It therefore makes a perfectly suitable cross-sector and cross-country combination of industries that are of high European significance and value.

The main objectives of BigDataGrapes is to build upon the rich historical, cultural and artisan heritage of Europe, aiming to support all European companies active in two key industries powered by grapevines: the grape and wine industry and the natural cosmetics one. It will help them respond to the significant opportunity that big data is creating in their relevant markets, by pursuing two ambitious goals:

- To develop and demonstrate powerful data processing technologies that will increase the efficiency of companies that need to take important business decisions dependent on access to vast and complex amounts of data.
- To catalyse the creation of a data ecosystem and economy that will increase the competitive advantage of companies that serve with IT solutions these sectors.

2 PILOTS OVERVIEW

BigDataGrapes collects and monitors sensor data derived from all test sites owned or accessible by consortium members, bringing an expansive and diverse collection of datasets. These streams of data and datasets serve as the basis for carrying out research and technical work and are used as the testbed for enabling the implemented technical components to efficiently handle the volume and intricacies of these data, clearly acquired from realistic in- field conditions. The data analysis phase is part of the definition of the BigDataGrapes use cases and the BigDataGrapes pilots. Thus, four (4) overarching use cases have been identified and were then further divided in different scenarios.

Table 1: Use Cases and Scenarios

Use Cases (Generic)	**Use Case Scenarios**
A. Data Anomaly Detection & Classification	A. Earth Observation Data Anomaly Detection & Classification
B. Prediction	B1. Yield Prediction B2. Predicting Biological Efficacy B3. Crop Quality Prediction • for Optimizing Post Harvest Treatments of Table Grapes (B3-1) • for Optimizing Winemaking (B3-2)
C. Farm Management	C1. Optimization of Farm Practices in the Vineyard C2. Management Zones Delineation for Vineyards
D. Risk Assessment	D1. Grape and Wine Quality Risk Assessment (safety) D2. Environmental Impact D3. Long-term Risk Assessment (Insurance Scenario)

Moving from testing in laboratory conditions to testing in real-world settings, BigDataGrapes has designed and is executing human-centred assessment activities, the application pilots, pertaining to the defined use cases. The pilots defined, namely the Table and Wine Grapes pilot, the Wine Making pilot, the Farm Management pilot, and the Natural Cosmetics pilot constitute instantiations of these use cases. They are fully defined grapevine-powered industry use cases' demonstrators, developed in order to allow the evaluation of the BigDataGrapes components within real-world settings, fulfilling industry-centred and specific end-user requirements.

3 BIGDATAGRAPES PILOTS ADVANCEMENT

3.1 TABLE AND WINE GRAPES PILOT

3.1.1 Pilot description

Table and Wine Grapes Pilot aims to denote correlations between precision agriculture information and phenological data and grape and wine chemical analysis. Another goal is to associate the aforementioned data with earth observation data in order to examine the effectiveness of applying machine learning techniques and eventually train the relevant machine learning components.

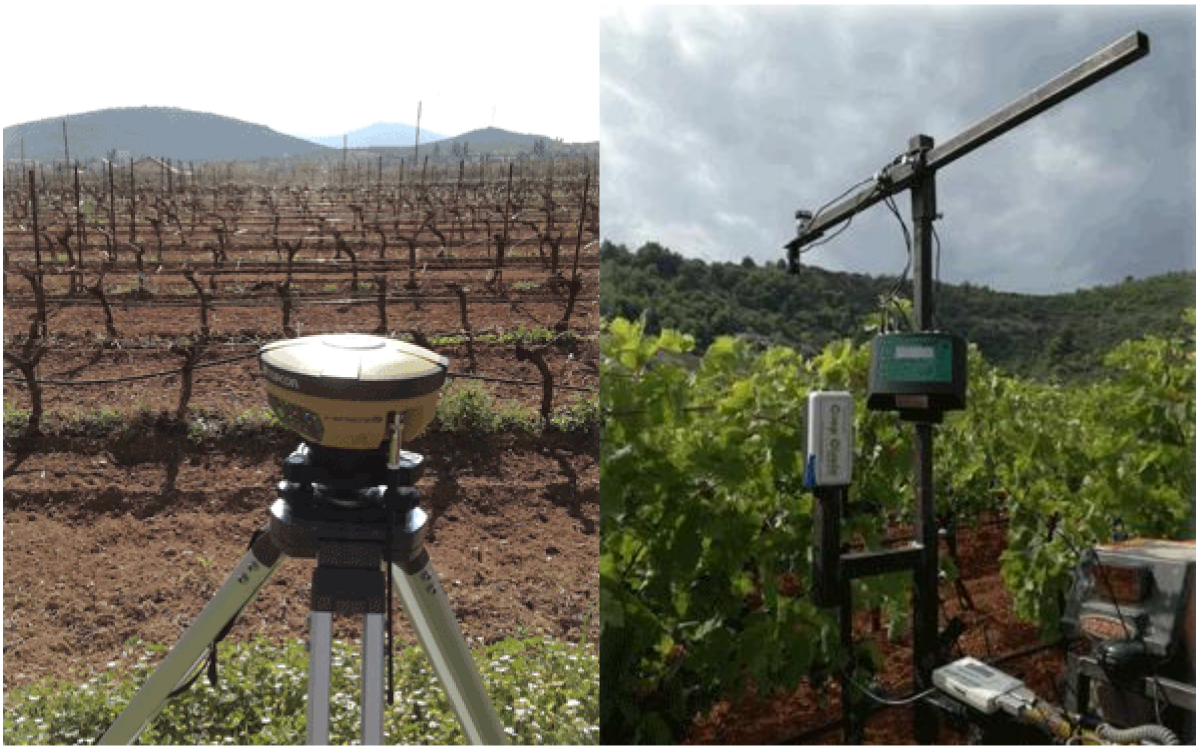


Figure 1: BigDataGrapes device installations from AUA

The responsible partner of this pilot, Agricultural University of Athens (AUA), collects and monitors sensor, farming and phenological data derived from all test sites located in Greece. Soil properties, climate conditions and cultivation techniques constitute significant variables, which affect the quality of the final product. In particular, soil data affect both crop quality data and crop quantity data. Deriving meaningful knowledge from many relevant, yet heterogeneous data sources is important, acting as the basis for future decision-making processes.

3.1.2 Specific Goals

Some of the goals to be achieved through sensor and farming data collection is to denote correlations between precision agriculture information and phenological data and grape and wine chemical analysis. Finally, the ultimate goal is to correlate the aforementioned data with earth observation data in order to examine the effectiveness of applying machine learning techniques and eventually train the relevant machine learning components.

3.1.3 Site Description

Three test sites have been chosen for data collection. These are situated in the regional unit of Corinthia, in the north-eastern part of Peloponnese, Greece. In particular,

- Palivou Estate and Kontogiannis Estate for winemaking production
- Fasoulis Estate for table grapes production

3.1.4 Expected Timeline

Measurements related to the Table and Wine Grapes pilot are taking place during the whole duration of the project. Emphasis is given during the summer months, May through September, a very important period of the grapevine's vegetative cycle, from the maturation of grapes until harvest time.

3.1.5 Envisaged Outcomes

The expansive and diverse collection of datasets for BigDataGrapes serves as the basis for carrying out research and technical work. The data streams are used as the testbed for enabling the implemented technical components to efficiently handle the volume and intricacies of these data clearly acquired from realistic in-field conditions. As the project progresses, the data pool will be continuously enriched in volume and range, in accordance with the needs and requirements of the covered use cases.

3.1.6 Advancement during the first year

During the first year, AUA engaged to collect all necessary data from the three test sites that have been chosen for data collection, namely Palivou Estate and Kontogiannis Estate for winemaking production and Fasoulis Estate for table grapes production. In particular, the requested data are: positioning data, soil electrical conductivity data, weather data, data related to the quality of the grapes and vegetative data. For the realization of the data collection, special equipment is used. For example, Geonics EM38-MK2 Ground Conductivity Meter was used for soil electrical conductivity data collection, Crop Circle, Rapid Scan and SpectroSense2+GPS were used to retrieve classic spectral vegetative index data such as NDVI, NDRE and LAI and HiPer V RTK GPS was used to record topographical data such as field boundary points, and elevation data. Moreover, two weather stations were installed at Palivou and Kontogiannis Estates respectively, in order to measure the wind speed and direction, air temperature sensor, air humidity sensor and barometer in order to monitor the atmospheric pressure. Last but not least, collected samples have been transferred to the Laboratory of Viticulture for further qualitative analysis (pH, Sugar Content, Titratable Acidity etc.).

3.2 Wine making pilot

3.2.1 Pilot description

The Wine Making Pilot is dedicated to research in the fields of viticulture and oenology with an integrated point of view that allows a transversal approach from the vineyard to the packaged final product.

The INRA's experimental unit of Pech Rouge (UEPR) conducts research and technological experiments on:

- Viticulture and the ecophysiology of the vine, with as a main issue a better knowledge and better control of grape quality.
- Enology with, as major research axes, the expression of quality potential existing in the grapes and wines and the on-line monitoring and control of the alcoholic fermentation.
- Technological processes with the aim to propose and study innovative technologies applicable to various steps of winemaking.
- The valuation of coproducts, extraction of molecules and environmental impacts.

3.2.2 Site Description

The INRA Pech Rouge Experimental Unit is located in the Occitanie (ancient name Languedoc-Roussillon) region of France.

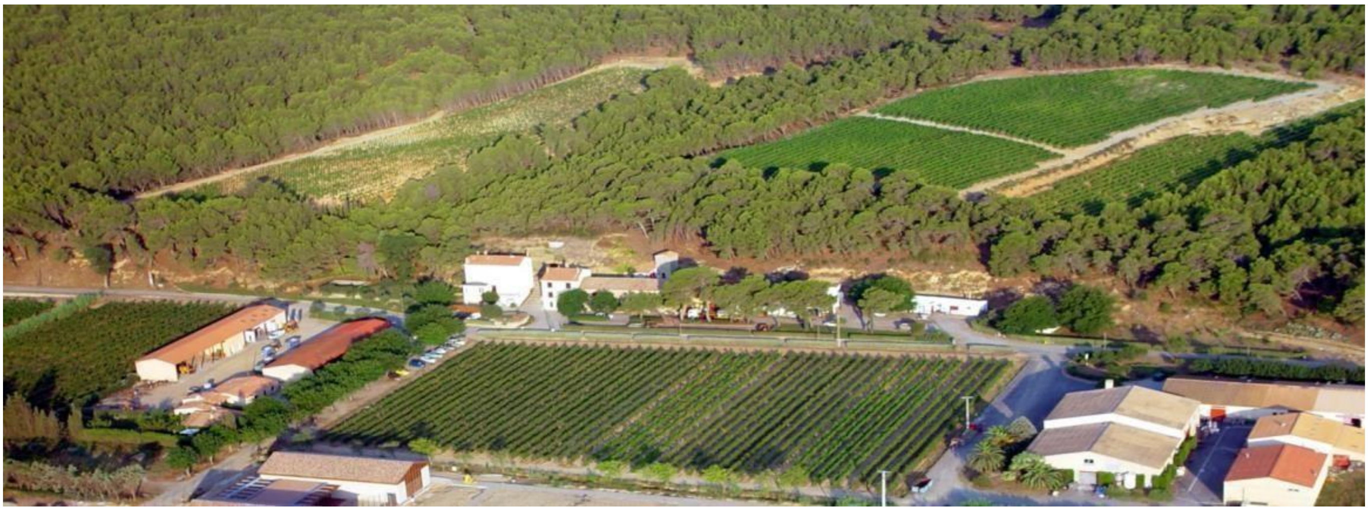


Figure 2: The experimental vineyard of INRA UEPR

3.2.3 Expected Timeline

Measurements related to the winemaking pilot are also taking place during the whole duration of the project. In order to adapt crop varieties, crop management practices and modes of canopy dressing to the requirements of research, the vineyard of Pech Rouge is continuously evolved, with the focus on the following activities, so as to better serve the needs for the BigDataGrapes pilot:

- Collecting information of fields, terrain, product quality
- Measuring and monitoring field activities and winemaking activities

3.2.4 Envisaged Outcomes

Historical and ongoing experimentation on Pech Rouge experimental Unit will provide a large-scale datasets about winemaking and the vine-grape-wine continuum. Those data and datasets will be benefit for:

- The application and test of the BigDataGrapes solution
- The validation of the BigDataGrapes components in real-life conditions and with complex dataset.

The pilot also aims to:

- design and test methods for improving data quality (correction)
- have a better understanding of the relation of observed data (e.g. on-the-field weather measurements, vine water status etc.) with the quality of the end product
- discover hidden patterns and knowledge in order to design new viticulture/ vinification systems

3.2.5 Advancement during the first year

INRA, the responsible partner of the Wine Making Pilot, was engaged to create a dataset with water potential, radial growth and visual observations data. INRA's experimental unit of Pech Rouge has conducted a lot of research experiments for private wine companies and INRA in the frame of its engagement, made a special effort to contact with the wine companies, owners of those experimental data in order its large-scale dataset to be enriched.

3.3 Farm management pilot

3.3.1 Pilot description

The Farm Management Pilot aims to develop a unique system that satisfies the following needs:

- A Farm Management system with all the functionalities to support the farmer in his day by day activities and in gathering data from the field
- Hosting data from different sources with proper tools and functionalities for comparisons and easy data management
- Data exchange. A "day by day" data producer, to feed the generated data into the other BDG components, and make use of the incoming information from the other BDG components.
- Data visualization. The data related to the farmer should be displayed in a way that provides an added value and new insights to the farmer for his activities.

Two wine makers were identified as actors in this pilot. They will be involved in the pilot in two ways:

- Their work will be supported by making the developed products and systems available to them. In addition to the farm management system itself, this includes sensors and measurements that will provide data as basis for decision support.
- On the other hand, these actors can help in designing the new system by providing input and knowhow about their needs and activities. They can also give insights on how to disseminate results, approach and ideas of the BigDataGrapes Project.

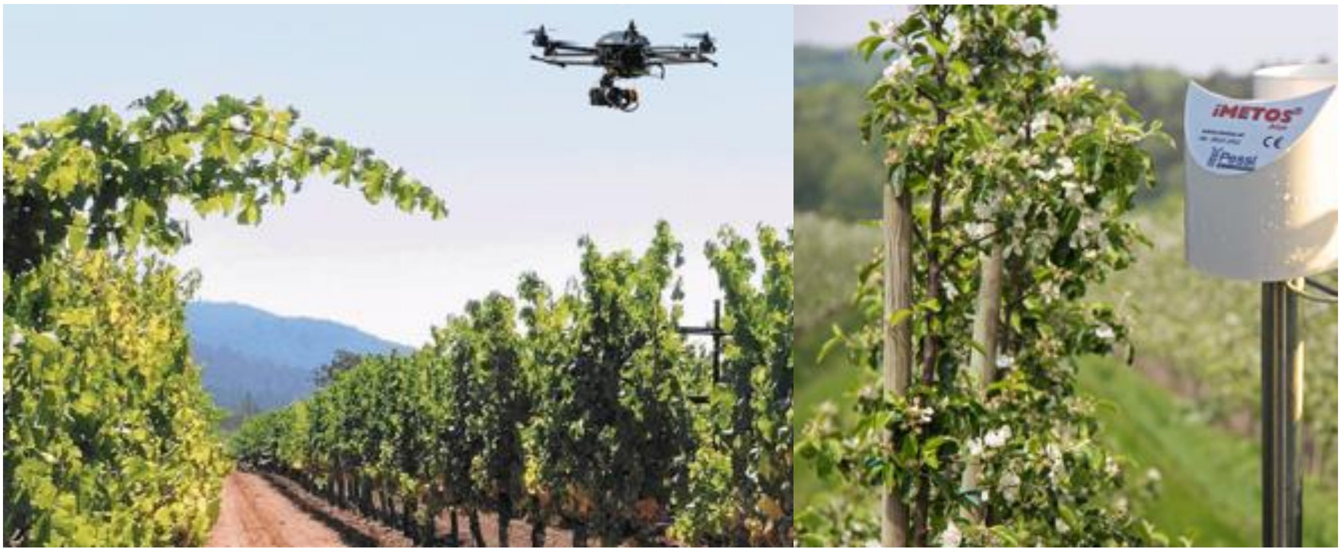
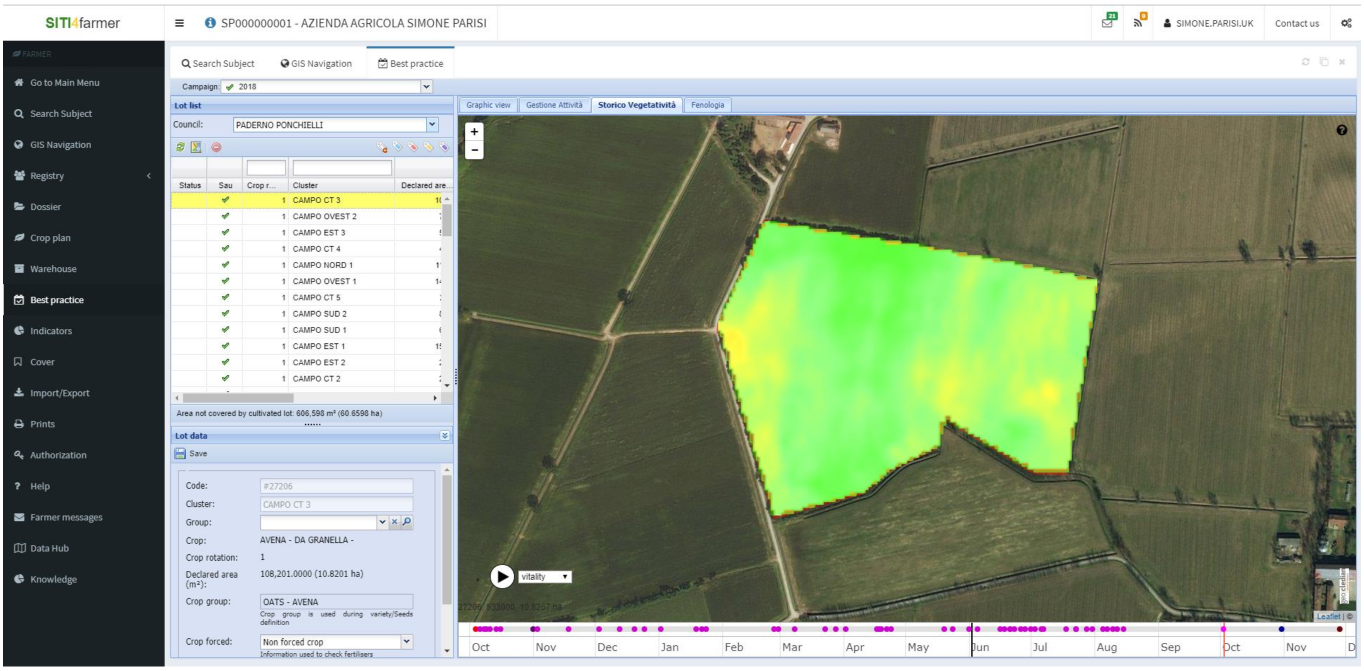


Figure 3: Drones and sensors operating in BigDataGrapes pilot sites in Tuscany

In the following figures, the SITI4Farmer platform, the platform that is going to be used by the two Tuscany pilots winegrowers is demonstrated. It can be used to support various activities: e.g. to load best practice data, to manage variable rate fertilizer maps, to manage different information layers on soil, meteorology, satellite data and so on.



SP000000001 - AZIENDA AGRICOLA SIMONE PARISI

Campaign: 2018

Lot list

Council: PADERNO PONCHIELLI

Status	Sau	Crop r...	Cluster	Declared are...
	✓	1	CAMPO CT 3	10
	✓	1	CAMPO OVEST 2	7
	✓	1	CAMPO EST 3	5
	✓	1	CAMPO CT 4	4
	✓	1	CAMPO NORD 1	11
	✓	1	CAMPO OVEST 1	14
	✓	1	CAMPO CT 5	[illegible]
	✓	1	CAMPO SUD 2	8
	✓	1	CAMPO SUD 1	6
	✓	1	CAMPO EST 1	15
	✓	1	CAMPO EST 2	2
	✓	1	CAMPO CT 2	2

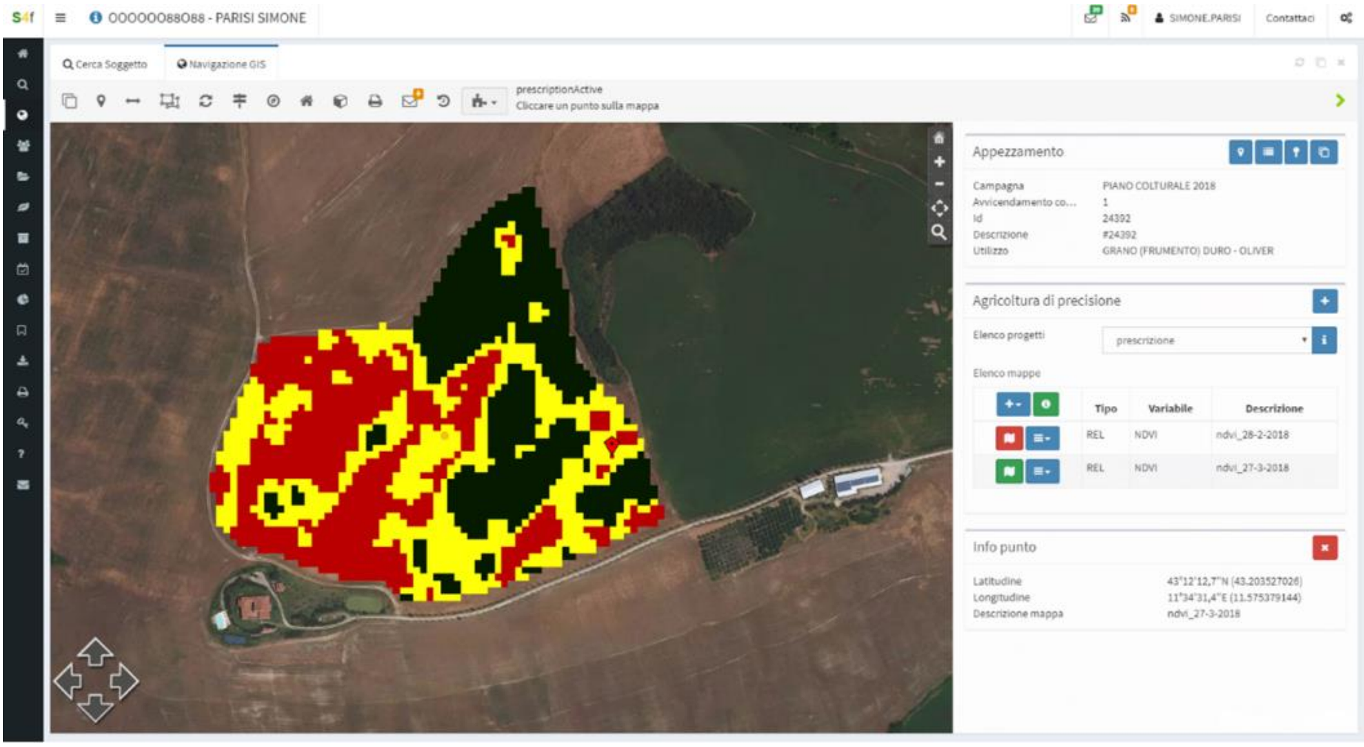
Area not covered by cultivated lot: 606,598 m² (60.6598 ha)

Lot data

Save

Code: #27206
Cluster: CAMPO CT 3
Group:
Crop: AVENA - DA GRANELLA -
Crop rotation: 1
Declared area (m²): 108,201.0000 (10.8201 ha)
Crop group: OATS - AVENA
Crop group is used during variety/Seeds definition
Crop forced: Non forced crop
Information used to check fertilisers

Figure 4: A satellite NDVI image (from Geocledian APIs) for a parcel in SITI4Farmer.



00000088088 - PARISI SIMONE

prescriptionActive
Cliccare un punto sulla mappa

Appezzamento

Campagna: PIANO COLTURALE 2018
Avvicendamento co...: 1
Id: 24392
Descrizione: #24392
Utilizzo: GRANO (FRUMENTO) DURO - OLIVER

Agricoltura di precisione

Elenco progetti: prescrizione

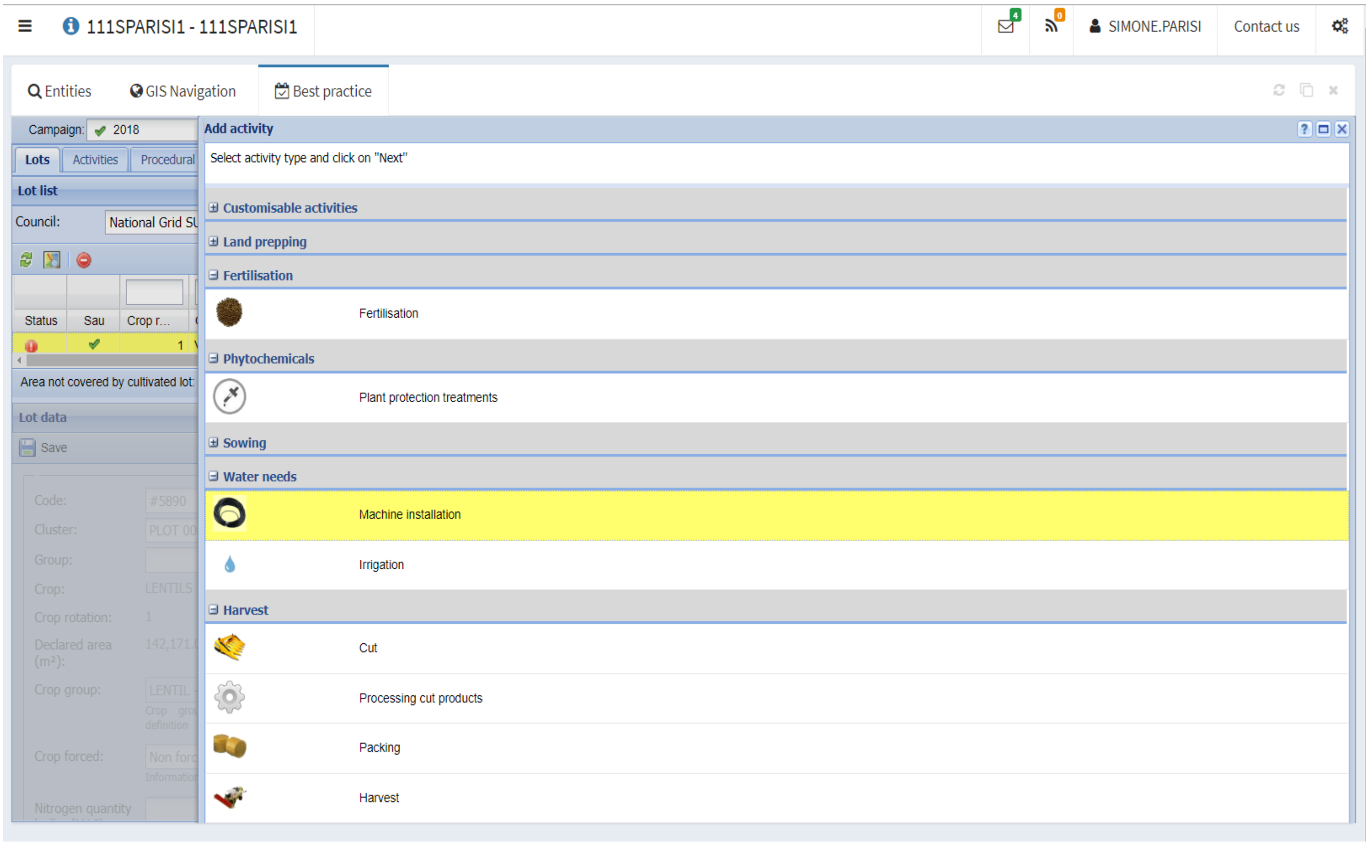
Elenco mappe

	Tipo	Variabile	Descrizione
	REL	NDVI	ndvi_28-2-2018
	REL	NDVI	ndvi_27-3-2018

Info punto

Latitudine: 43°12'12,7"N (43.203527026)
Longitudine: 11°34'31,4"E (11.575379144)
Descrizione mappa: ndvi_27-3-2018

Figure 5: A prescription map created on the basis of NDVI and Soil maps applying a cluster analysis in order to identify different management zones.



Campaign: 2018

Council: National Grid SU

Status	Sau	Crop r...
●	✓	1

Area not covered by cultivated lot:

Lot data

Code: #5890

Cluster: PLOT 00

Group:

Crop: LENTILS

Crop rotation: 1

Declared area (m²): 142,171.

Crop group: LENTIL

Crop forced: Non forc

Nitrogen quantity

Select activity type and click on "Next"

⊞ Customisable activities

⊞ Land prepping

⊟ Fertilisation

Fertilisation

⊟ Phytochemicals

Plant protection treatments

⊞ Sowing

⊟ Water needs

Machine installation

Irrigation

⊟ Harvest

Cut

Processing cut products

Packing

Harvest

Figure 6: The selection mask for different kinds of best practices.

3.3.2 Site Description

The pilot involves 2 wineries, making them an active part of the project, collecting data from the field, in automatic and manual manners, and therefore contribute to the results. Both wineries are located in Tuscany, Italy and their names are:

- Casato Prime Donne: 12 HA of wineyards of Brunello of Montalcino
- Cantina Il Palazzo: 35 HA of Wineyards of Chianti D.O.C.

In this 2 farms-winery has already been installed 2 professional weather stations provided also with soils sensors and plant sensors for the continuous monitoring of the temperature and moisture of plant, soil, atmosphere system.

3.3.3 Expected Timeline

After the engagement of the pilot wineries and the installation of sensors by ABACO, the field data collection will start and run throughout the pilot duration. In parallel GEOCLEDIAN has started to acquire, process and provide satellite data for all pilot test sites throughout the pilot run time and will develop advanced information products based on those. When all data sets are available, they will be analyzed and combined to support the development of meaningful products for the users of the system. In an iterative process these will be refined with the user’s feedback to enhance the Farm Management System with winery-specific products.

3.3.4 Envisaged Outcomes

In the frame of the pilot, GEOCLEDIAN will further develop their current data processing platform into a real Big Data Processing Platform that will allow the scalable production, provision & analysis of large-scale data sets.

The combination of satellite remote sensing data with in situ field & weather data will enable the following developments:

- Data anomaly models to automatically detect interesting field events
- Combined analysis tools for farm management relevant applications
- Preparation of Management Zones Maps and therefore Prescription Maps for the variable rate best practices
- Integration of VHR data and additional data sources together with data quality monitoring tools
- New, grape-specific higher-level information products
- User-specific Visualization of big data analytics that are relevant for the farmer

ABACO is going to make use of the output from GEOCLEDIAN, from sensors, and from the users of the system, to create knowledge maps and data systems to enable the connection of the culture quality with all the other variables. GEOCLEDIAN & ABACO will assess the added value of the developed products for the farmers and improve them with the farmers' input.

3.3.5 Advancement during the first year

During the first year the two responsible partners selected the two wine makers pilot partners. GEOCLEDIAN implemented developments for the improvement of pilot's data products, data quality and processing system monitoring tools. In particular, extensive tests have been performed to analyse the performance and scalability of pilot's system to identify bottlenecks that need to be improved and prepared for the BDG activities. Moreover, the newly set up server and processing chain have been used to study atmospheric corrections and improved cloud masking methods to prepare the delivery of improved data products in the frame of BDG. For the collection of weather and soil data a weather station and soil sensors have been installed. The acquisition, processing and data delivery of satellite data for all pilot sites was started. ABACO implemented connectors to improve communication with weather station and in general sensors from external providers. Therefore, ABACO is continuously upgrading interface and functionalities related to Best Practices, Precision Farming issues and remote sensing image import and visualization according to GEOCLEDIAN.

3.4 Natural cosmetics pilot

3.4.1 Pilot description

The Natural Cosmetics pilot intends to gather samples of vineyard by-products across the Greek territory. There is a need to extract the most out of pharmaceutical plants for both economic and environmental reasons. A real challenge is to add high value to by-products. Wine making produces a lot of by-products that may have a significant biological value if there are adequate data concerning farm management. These data can lead to decisions concerning the processing of by-products in order to produce high added value active ingredients for cosmetics and food supplements. Bioactive compounds from winery by-products have disclosed interesting health promoting activities both in vitro and in vivo. If properly recovered, they show a wide range of potential and remunerative applications in many industrial sectors, including cosmetics, pharmaceuticals, biomaterials and food. In fact, winemaking by-products are outstanding sources of oil, phenolic compounds and dietary fibre and possess numerous health benefits and multifunctional characteristics, such as antioxidant, colouring, antimicrobial and texturizing properties.

3.4.2 Specific goal

The scenario presumes that precision farming and control of parameters linked to the quality of wine (soil characteristics, GIS data etc) may provide by-products of superior quality. In particular, the pilot intends to gather samples of vineyard by-products across the Greek territory and more specifically vine leaves of two different grape varieties (Agiorgitiko and Mandilaria) and test their phytochemical profile and biological value after extraction.

3.4.3 Site Description

Two important indigenous grape varieties (Agiorgitiko and Mandilaria) were chosen for the needs of the Natural Cosmetics pilot. Samples of these two varieties were selected from sixteen vineyards from four regions of the Greek territory, namely Peloponnese, Northern Greece, Aegean and Crete.

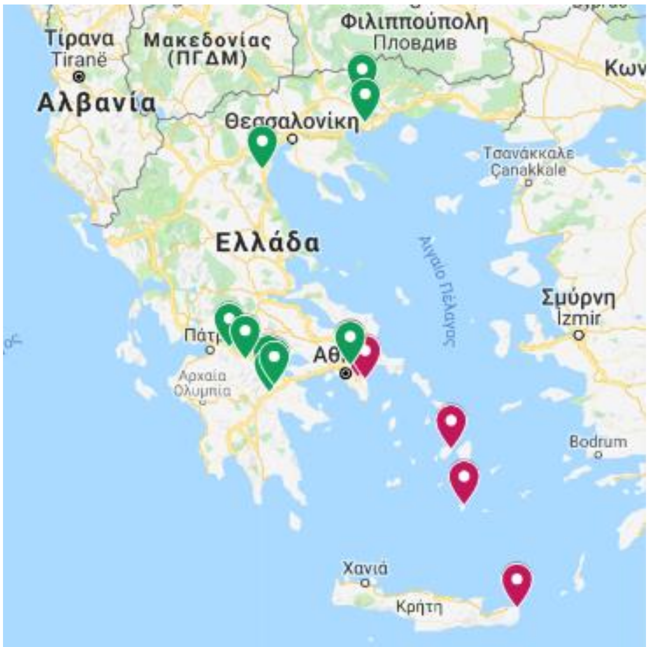


Figure 7: Geographical distribution of selected vineyards for the Natural Cosmetics Pilot

3.4.4 Expected Timeline

Measurements related to the Natural Cosmetics pilot will take place during the whole duration of the project. The collection of samples from the chosen vineyards will be repeated every year, following extraction using two different methods and measurements of biological efficacy of developed extracts.

3.4.5 Envisaged Outcomes

Bioactive compounds found in wine-making by-products such as vine leaves possess multifunctional characteristics and show a wide range of potential and remunerative applications, concerning health promoting activities. Nevertheless, the quality of these by-products and more specifically their biological efficacy can vary depending on multiple parameters, such as the origin of the sample, the recovery process and more.

The collected data from the natural cosmetics pilot will provide the necessary information for the evaluation of the quality of each sample, linked with the special characteristics of the vineyard of origin. The goal is to face the challenge: "how data from the field can be linked to the biological efficacy of final products - an application on wine making by-products".

3.4.6 Advancement during the first year

From the very first months of the project, APIGEA, as pilot's responsible partner trained producers in the correct collection of the leaves and stems and especially the drying process that is very important in order to avoid contamination and be able to process further the products for the pilot's deliverable. Two important indigenous grape varieties (Agiorgitiko and Mandilaria) were chosen for the needs of the Natural Cosmetics pilot and sixteen over twenty-four vineyards that APIGEA approached have agreed to gather samples of dried leaves of the two different varieties from their vineyards for the first year. On the second phase of sample gathering a bigger amount is planned. The first data on the results from the gathered samples on the biological efficacy (pH, refractive index, Total microbial count, Yeasts & Moulds, TPC, TFC, Antioxidant activity) have been analysed by APIGEA.

4 BIGDATAGRAPES TECHNICAL SOLUTION

4.1 OVERALL APPROACH

The described data problems and their respective application scenarios demand the provision of a complete computational solution that serves all aspects of the Big Data management value chain. To this end, BigDataGrapes aims to build and deploy a set of components that carry out the various processes in order to (a) solve the data problems of the grapevine-powered industry and (b) ensure transferability and extensibility to other data-driven businesses by adopting a coherent architectural approach.

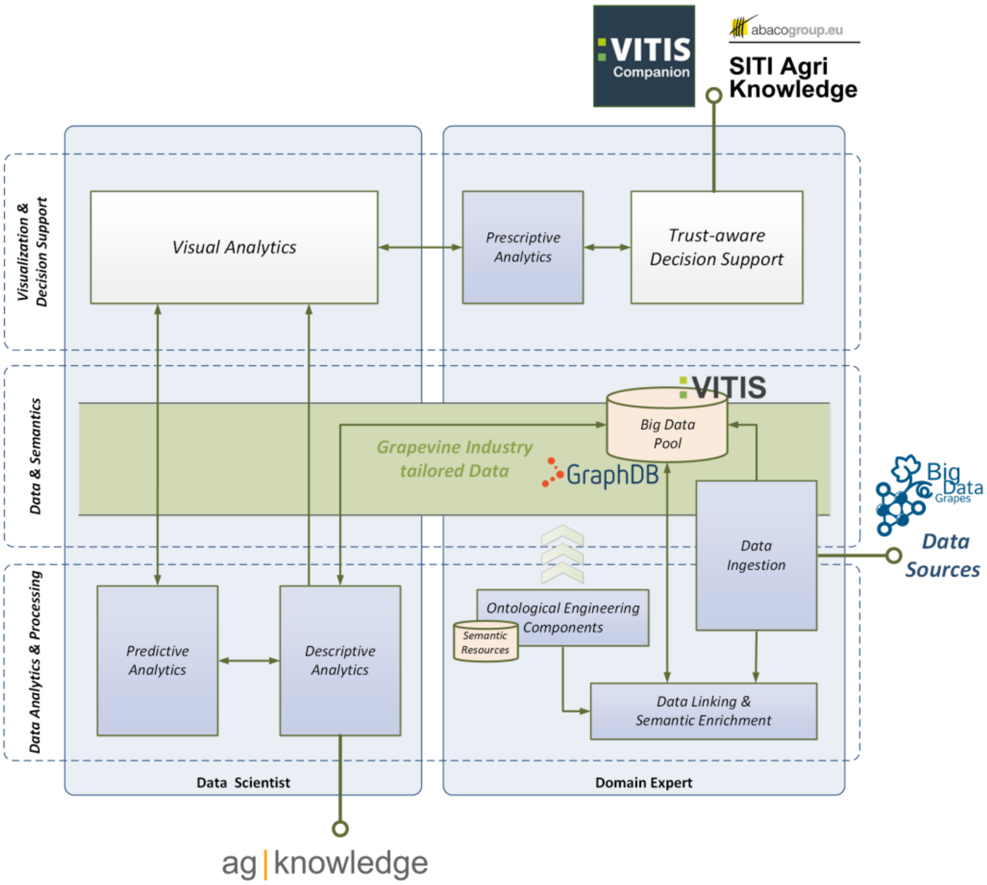


Figure 8: BigDataGrapes top-level architecture

A critical goal for the overall system is to ensure that the relevant data sources are semantically annotated and integrated as parts of a common data pool comprising disparate yet interconnected data assets. To this end, BigDataGrapes designs and develops methods and components for the semantic linking and enrichment of the available data and furthermore, makes available to the domain experts tools for annotating and describing their data in order to be incorporated in the BigDataGrapes pool.

At the processing stage, BigDataGrapes employs the necessary components for carrying out typical analytics processes, making sure that the execution environment and methodology retain scalability and efficient use of computational resources. Additionally, the project designs and implements inference and machine learning methods to produce advanced predictive analytics over the entirety of the available data pool, tackling open issues like the movement to distributed architectures for these tasks.

Finally, BigDataGrapes leverages the value elicited from these data, by translating them to intuitive and actionable knowledge and using them as the foundation for decision support in complex environments.

4.2 Advancement on Data & Semantics

During the first year of the project, the initial and most important steps for identifying and making available the tools and components pertaining to BigDataGrapes semantic layer have been carried out. More specifically, the WP3 Data & Semantics Layer has presented:

- The sort of data represented in a semantic way
- Defining methodology for Semantic Data Integration
- Relevant AgroBio ontologies and problems that we have found in them
- Specific project data
- Specific data processing and ingestion requirements
- Specific data access requirements and relevant tools
- Analysis of the specific data requirements received by the use case partners (data sources, entities, relations, etc)

Regarding the components, the WP3 includes the first version of the software components developed and discusses the preliminary results obtained. In particular, a novel compression technique for inverted indexes based on Variable-Byte, a well-known and widely adopted method for coding integer sequences by saving memory space and enabling fast search operations is presented and preliminary results obtained by applying the novel method on a large inverted index for RDF data are provided. Work is in progress to refine the initial approach and to propose a novel compressed index for RDF big data providing fast access to triples for answering complex SPARQL queries.

Furthermore, in the WP3 the main tools available to index and manage graph and time series are identified. These tools are very popular and have been successfully used in a plenty of applications. They allow for implementing advanced indexing and inference techniques to reflect the data shape and volume of the use-cases of the project.

Finally, an approach for applying linguistic pipelines and semantic enrichment for indexing scientific literature with regards to active compounds has been outlines. It features natural language processing with embedded graph data to enrich the original content and transform its metadata into semantic search indexes. Initial information needs related to project's use-cases have been identified to build functional requirements for the semantic enrichment module.

4.3 Advancement on Data Analytics & Processing

During the reporting period, the BigDataGrapes platform has been consolidated and extended to support grapevine-powered industries to take important business decisions. The platform managed to provide mechanisms, fault-tolerant tools and components to carry out rigorous analytics processes on complex and heterogeneous data helping companies and organizations in the sector to evolve methods, standards and processes based on insights extracted from their data. Core technologies and frameworks have been identified, deployed and used for efficient processing of large datasets, such as Apache Hadoop and Apache Spark, making sure that the execution environment and methodology retain scalability and efficient use of the available computational resources.

On top of BDG software stack, the BDG platform managed to enable distributed predictive big data analytics by effectively exploiting scalable Machine Learning algorithms using efficiently the computational resources of the underlying infrastructure. The software components enabling BDG predictive data analytics have been designed and deployed using Docker containers. They thus include everything needed to run the supported predictive data analytics tools on any system that can run a Docker engine.

In WP4 (Analytics & Processing Layer), the effectiveness of different predictive data analytics tools in terms of a standard and popular metrics such as Accuracy was assessed while in the future tools and methods for a deep analysis of the performance of different machine learning libraries is expected to be provided.

During the reporting period we also investigated a methodology for assessing the performance of a big data system. The methodology takes into account the characteristics of the system and also the heterogeneity and distributed nature of big data. We first analysed the state-of-the-art for big data benchmarking considering different challenges ranging from preserving the 4V properties of big data, streaming and scalability issues, to the main limitations of the current benchmarking in the context of relational databases and semantic repositories. We then presented the steps that can be followed for providing a rigorous testing of the BDG system. We provided some valid solutions for our project, for example, the benchmarks proposed by the Linked Data Benchmark Council (LDBC) which ensure linearity, reliability, and repeatability. We also provided a first proposal on the metrics to use for assessing the performance of the BDG system. Such metrics are chosen based on the datasets employed in the use cases of the BDG project. Finally, we contributed with some guidelines that can be helpful in the process of rigorous testing a big data system. We believe that a good approach would be to follow a standardized benchmarking methodology which is divided into different stages going from the selection of the application domain to the execution of the tests. Since the BDG system is not finalized yet, these guidelines and metrics are general and will be refined and concretized once the system will be developed.

4.4 Advancement on Visualisation & Decision Support

Driven by the requirements posed by the BigDataGrapes supported research communities, the visualisation technologies that were exploited within the project were identified and assessed in the context of a specific visualisation architectural approach.

The benefits of visualising complex data arise from being able to better interact and understand data by aggregating, filtering, searching or scaling down relevant information. Thus, a large volume of data with potentially complex information becomes more easily consumable. Due to such benefits, visualisation tools have been widely used in various domains to assist with tasks that might otherwise require significant cognitive effort. For instance, analysis of past volcanic activities using a time series graph can uncover trends and can aid in predicting future eruption of a volcano.

As a starting point for the Visualisation Layer, a systematic review was conducted where a total of 140 research papers were thoroughly analysed. The review discovered various decision support systems (DSS) and visualisation tools that have been proposed in the domain of agriculture. Based on the findings of the review, a total of six interactive visualisation components have been developed. To ensure that the components can be easily reused by partners, the Polymer framework was used. It provides a JavaScript library for building web applications with custom Web Components. Sample codes for each of the components highlighting their usability were also provided. All of the codes have been published at the github repository of BigDataGrapes (https://github.com/BigDataGrapes-EU/visualisationcomponents) and a manual with steps for obtaining the codes and serving the components has been provided.

Next, a trust-aware decision support system for stakeholders in the grapevine-powered industry has been developed too. After extensive research, it was discovered that the majority of DSS in agriculture use at least one interactive visualisation technique which includes: heatmap (overlaid over geographical maps), time-series, histograms, bar chart, pie chart, radar chart, clustering, temporal pattern and cross section representations of a farm. However, to date, the models employed for agricultural decision support remain opaque to users and hidden behind the software. This often leads to trust issues, notably when suggestions coming from a DSS fail to provide meaningful explanations. Research in various domains have shown that explaining a model's predictions can mitigate this issue (i.e. by earning users' trust). Previous work in Precision Agriculture (PA) has also demonstrated usefulness of visualisations in clearly communicating uncertainty emerging from both data and prediction models. To demonstrate this, a trust-aware DSS was developed for BigDataGrapes, using a wine quality prediction scenario as an exemplar. The system includes two visualisation components, namely: parallel coordinates and waterfall plot, which explain a model's predictions, and the influence of each variable on prediction outcomes. A demo of this system is available at: https://bigdatagrapes-eu.github.io/deliverable5.3/. The codes have been published at the github repository of BigDataGrapes (https://github.com/BigDataGrapes-EU/deliverable5.3), together with a manual of steps for obtaining the codes and serving the components.

4.5 THE BIGDATAGRAPES SOFTWARE STACK

The individual technical components produced in the context of BigDataGrapes, are integrated in a configurable and deployable software stack. The software stack will be continuously updated throughout the duration of the project, and will be released annualy, insync with the implementation milestones of the project. The BigDataGrapes software stack is logically organised into architectural layers as presented in Figure 9.

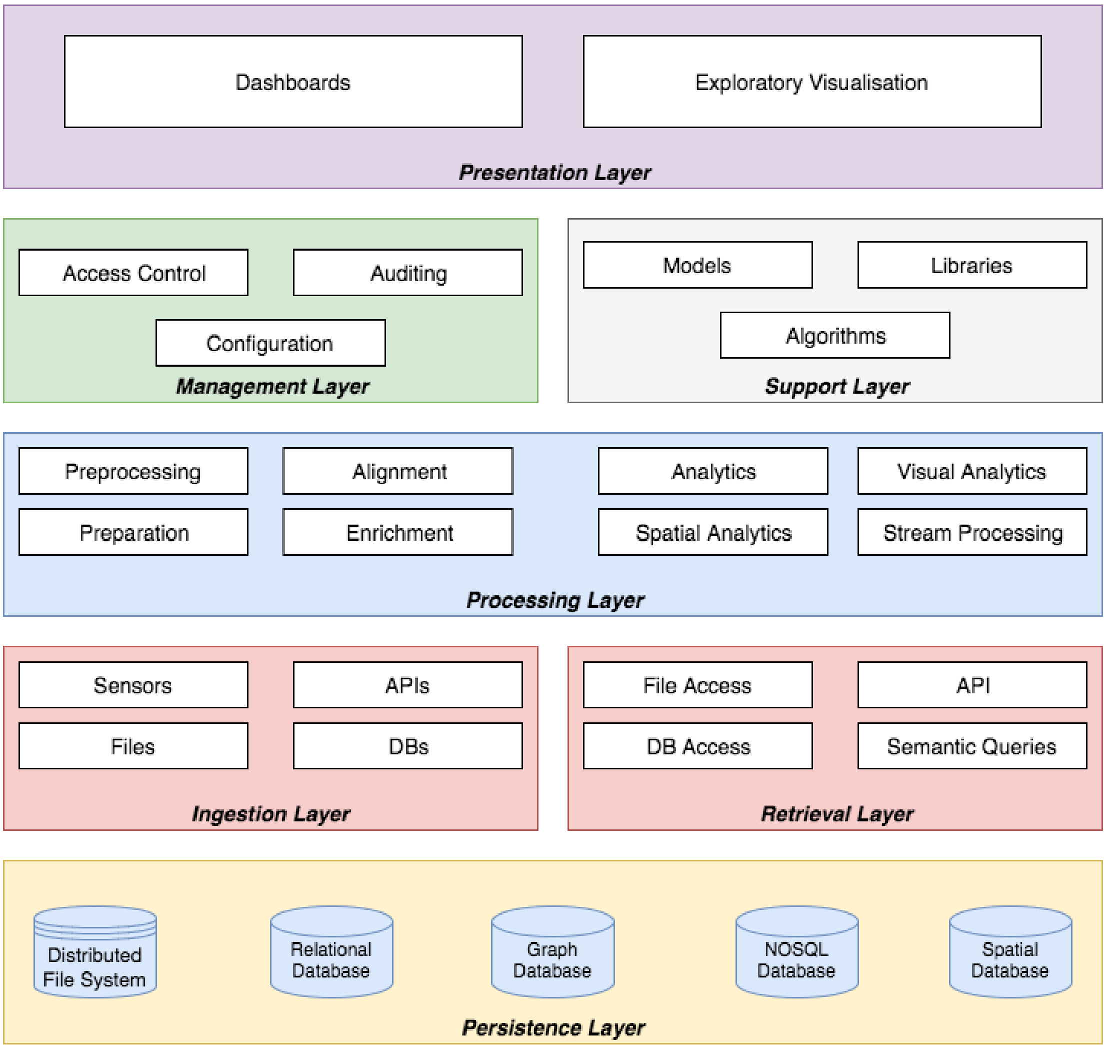


Figure 9: BigDataGrapes Software Stack Layers

The purpose and scope of each of the seven core layers are summarised in the following subsections.

4.5.1 Persistence Layer

The layer deals with the long-term storage and management of data handled by the platform. Its purpose is to consistently and reliably make the data available to the processing layer. The layer incorporates persistence technologies in accordance with the data sources summarised in section 3 of the document, organised in appropriate federations in order to ensure performance, redundancy and uptime.

4.5.2 Data Ingestion Layer

The Ingestion Layer comprises the components responsible for moving external data assets to the persistence layer of the platform. The ingestion layer will incorporate all the connectors required for accessing the sources to be used in the context of the BigDataGrapes use cases, as well as, the necessary logic to transform the input data into the formats and schemas agreed upon for the relevant component of the persistence layer where a given resource will be stored.

4.5.3 Data Retrieval Layer

In similar fashion, components of the Data Retrieval layer are responsible for exposing the stored data to the Processing and Presentation layers. Depending on the needs of each processing component, data can be retrieved via:

- Access to files lying on a distributed file system;
- Direct access to relational or NOSQL databases, via the execution of custom or pre-defined queries (depending on the use case and the degree of control that the end-users should have);
- Direct access to Graph databases and triple stores, via the execution of custom or pre-defined semantic queries (depending on the use case and the degree of control that the end-users should have);
- Calls on Application Programming Interfaces (APIs) that expose the underlying stored data in a controlled fashion.

4.5.4 Processing Layer

The Processing Layer implements the core processes for data management and analysis towards serving the analytics and decision support requirements of the BigDataGrapes use cases. These operations are classified under the following main categories:

- ***Pre-processing***: processes designed to validate and pre-process incoming datasets. Different data sources evidently will require different pre-processing mechanisms. Exemplary operations carried out by pre-processing components are generation of provenance metadata, validation and anomaly detection, feature extraction, etc.
- ***Alignment***: the alignment components are responsible for discovering and proposing links between semantic resources imported into the platform, either at the schema (ontologies, taxonomies, vocabularies) or at the instance level (identity or similarity between datasets or data items).
- ***Enrichment***: the enrichment components carry out the automatic annotation of the available data with semantic information, using the semantic resources available to the platform.
- ***Preparation***: The structure in which data are stored after the integration may not be suitable to perform the target analysis. The preparation modules adapt the data to match the format expected by the analytics components. Since each analytical model may expect data in a different format, data preparation is specific to each analytic model and the preparation components will be enriched and extended as necessary.

- ***Stream Processing***: The stream processing components are responsible for carrying out analysis over live data streams, as opposed to persistent data collections.
- ***Data Analytics***: The components receive as input the prepared data from the preparation components and apply statistical and machine learning methods to extract knowledge and make predictions. At this level, descriptive analysis is done providing some statistical insights on the characteristics and behaviour of the variables under study. In turn, the predictive techniques are based on machine learning models used to explain, classify and predict the targeted variables.
- ***Visual Analytics***: In similar fashion, the visual analytics components apply analytical algorithms and methods over the appropriately prepared data to generate the augmented visualisations to be presented to the end-user/ decision maker via the relevant components in the presentation layer.
- ***Spatial Analytics***: The spatial analytics components entail the functionality for geospatial analysis, making use of the relevant geospatial information, as well as, invoking and using directly the other analytics components of the platform.

4.5.5 Management Layer

The layer incorporates the tools for managing and configuring the operation of the BigDataGrapes platform itself. It targets administrators and managers of a deployment and provides the functionalities for user and role definition and access credentials, log monitoring and auditing, component configuration etc.

4.5.6 Support Layer

The support layer incorporates the modules and functions to be used by the processing components of the platform. These tentatively include implementations of machine learning algorithms, analytical models, geospatial operators, transformation libraries, etc.

4.5.7 Presentation Layer

The presentation layer entails all the user-facing components and environments of the platform. These include platform interfaces in the different modalities supported by BigDataGrapes (web, mobile, on-site equipment), content browsing and management dashboards, administration platforms, etc. The layer also includes the components for presenting analytics results as derived from the operations of the processing layer, and the appropriate environments for directly executing data retrieval queries.

The BigDataGrapes integrated software stack is a container of dockerized versions of the individual components, and can be accessed at:

https://hub.docker.com/u/bigdatagrapes

Full documentation of the BigDataGrapes integrated software stack, describing the installation steps and providing the necessary configurations, can be accessed at:

https://github.com/BigDataGrapes-EU/deliverable-D6.1

5 Conclusion

The first year of the project set the operational framework of BigDataGrapes project, by (a) defining and analysing a pool of selected Use Cases and Use Case Scenarios that are linked and related to the BigDataGrapes pilots as well as pilots' advancement and (b) identifying and assessing the enabling technologies, and their advancement, in order to serve the BigDataGrapes pilots.

During the project's second year we expect to put the established pilots in full use, by further refining and extending the available enabling technologies and incorporating the required amount of research resources for a realistic full-scale assessment of the BigDataGrapes pilots:

Table and wine grapes pilot: will continue to collect and monitor sensor, farming and phenological data from all test sites located in Greece. The first round of controlled pilot trials will implement a first version of the pilot, using the first versions of newly developed BigDataGrapes components. These will be restricted piloting trials in terms of scale and complexity in order to provide the needed data for the assessment of early BigDataGrapes components and the refinement of the pilot. Basic reflectance information from plant canopies and soil as well as classic spectral vegetative index data (NDVI, NDRE etc.) and photosynthesis measurements will be recorded multiple times per season, in order to cover in the most precise way the phenological growth stage of the grapevines. Weather information will be continuously recorded throughout the growing season. Quantitative and qualitative analysis (pH, Sugar Content, Titratable Acidity) data will be collected during the harvesting period. On top of the data collected during the first year, the Table and Wine Grapes Pilot will add to the list of datasets data from multispectral and thermal infrared sensors mount on drones. Additionally, this pilot is going to assist with the analysis and the semantic annotation and enrichment of the examined datasets, in order to elicit additional knowledge and insights from the initial data.

Wine making pilot: During this second year, our main objective is to have all data accessible for the project, for all partners. This task can be achieved using web semantic and the help of the consortium. Collecting data at every stage of vine production (including genetic data), from field to wine will be continued as well as collecting climatic data. Indeed, the behaviour of vineyards which directly impact wine quality in response to a changing environment is a current issue very interesting for Big Data Grapes. To connect data at each step of production until the final product, and the link with climate change is our guideline.

INRA, the wine making pilot, has a real expertise on the issue «crop quality prediction for optimizing winemaking». We will keep on working on that topic by challenging data and try with technical partners to give answers to this issue.

Farm management pilot: In the second year the focus of the pilot will lie on data collection, data analysis and the development of new and improved products. The satellite and in field data collection & provision to partners will be ongoing. We plan to start the analysis of collected data of the first year to find interesting connections between weather, soil, vegetation and satellite data as well as farm management actions, especially related to Fertilization, Farm Management Zones and data anomaly detection with a very important focus on Variable Rate Best Practices. The analysis and improvement of the performance and scalability of the pilot's system will be continued.

Natural cosmetics pilot: In the course of next year, the Natural Cosmetics Pilot first of all intends to work on diffusing data collected during last year amongst all partners. Also, we will start analysing the data collected in order to reach conclusions regarding our hypothesis and be able to refine it accordingly. At the same time, during late spring and summer, collection of the second batch of samples (dried vine leaves) is planned, if possible, with the collaboration of even more wine makers, in order to increase the sample diffusion and total number. The samples gathered will then be measured on their biological efficacy. As a final goal, the Natural Cosmetics Pilot intends to work as a link between data from the field and biological efficacy of final products, along with the valuable help of the consortium which will tackle the technological challenges for said correlation.