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

D2.2 - Data Management Plan & Support Pack

DELIVERABLE NUMBER	D2.2
DELIVERABLE TITLE	Data Management Plan & Support Pack
RESPONSIBLE AUTHOR	Stoitsis Giannis (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	Nikos Manouselis
ADDRESS	110 Pentelis Str., Marousi, GR15126, Greece
REPLY TO	nikosm@agroknow.com
PHONE	+30 210 6897 905
EU PROJECT OFFICER	Ms. Annamária Nagy
WORKPACKAGE N. | TITLE	WP2 | Grapevine-powered Industries Big Data Challenges
WORKPACKAGE LEADER	Agroknow
DELIVERABLE N. | TITLE	D2.2 | Data Management Plan & Support Pack
RESPONSIBLE AUTHOR	Stoitsis Giannis (Agroknow)
REPLY TO	nikosm@agroknow.com
DOCUMENT URL	http://www.bigdatagrapes.eu/
DATE OF DELIVERY (CONTRACTUAL)	31 March 2018 (M3), 30 June 2019 (M18, Updated Version), 30 October (M21, Revised Version), September (M33, Final Version)
DATE OF DELIVERY (SUBMITTED)	30 March 2018 (M3), 28 June 2019 (M18, Updated Version), 30 October (M22, Revised Version), September (M33, Final Version)
VERSION | STATUS	2.1 | Final
NATURE	R (Report)
DISSEMINATION LEVEL	PU (Public)
AUTHORS (PARTNER)	Panagiotis Zervas (Agroknow), Pythagoras Karampiperis (Agroknow), Ioanna Polichronou (Agroknow), Margarita Gourgourini (Agroknow), Katrachoura Athina (Agroknow), Thanopoulos Charalampos (Agroknow)
CONTRIBUTORS	Evangellos Anastasiou (AUA), Katerina Kassimati (AUA), Maritina Stavrakaki (AUA), Florian Schlenz (GEOCLEDIAN), Stefan Scherer (GEOCLEDIAN), Simone Speringo (ABACO), Sabine Karen Yemadje Lammoglia (INRA), Coraline Damasio (INRA), Constantina Litsa (Symbeeosis), Pantelis Natskoulis (Symbeeosis), Eva Bozou (Agroknow)
REVIEWER	All partners

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Initial ToC and document structure	06/03/2018	Pythagoras Karampiperis (Agroknow)
0.2	Input from pilot partners to data management aspects of the identified use case and scenarios	19/03/2018	Evangellos Anastasiou (AUA), Katerina Kassimati (AUA), Maritina Stavrakaki (AUA), Florian Schlenz (Geocledian), Stefan Scherer (Geocledian), Simone Speringo (ABACO), Sabine Karen Yemadje Lammoglia (INRA), Constantina Litsa (Symbeeosis)
0.3	Chapters 1, 2, 3, 4, 5	23/03/2018	Pythagoras Karampiperis (Agroknow)
0.5	Peer-review comments	29/03/2018	Raffaele Perego (CNR)
1.0	Final version	30/03/2018	Pythagoras Karampiperis (Agroknow)
1.3	Pre-final draft for 2nd version	6/6/2019	Katerina Kassimati (AUA), Maritina Stavrakaki (AUA), Florian Schlenz (Geocledian), Simone Speringo (ABACO), Coraline Damasio (INRA), Pantelis Natskoulis (Symbeeosis)
1.5	Comments from internal review	12/6/2019	Katerina Kassimati (AUA), Maritina Stavrakaki (AUA), Florian Schlenz (Geocledian), Simone Speringo (ABACO), Coraline Damasio (INRA), Pantelis Natskoulis (Symbeeosis)
1.8	Pre-final draft	14/6/2019	Panagiotis Zervas (Agroknow),
2.0	Updated version	28/6/2019	Panagiotis Zervas (Agroknow), Eva Bozou (Agroknow)
2.1	Revised version based on the instructions of the reviews	30/10/2019	Ioanna Polychronou (Agroknow)

2.3	Add 5th pilot	8/8/2020	Ioanna Polychronou (Agroknow), Giannis Stoitsis (Agroknow)
2.4	Final version	17/9/2020	Giannis Stoitsis (Agroknow)

PARTICIPANTS		CONTACT
Agroknow IKE (Agroknow, Greece)		Nikos Manouselis Email: nikosm@agroknow.com
Ontotext AD (ONTOTEXT, Bulgaria)		Todor Primov Email: todor.primov@ontotext.com
Consiglio Nazionale DelleRicerche (CNR, Italy)		Raffaele Perego Email: raffaele.perego@isti.cnr.it
Katholieke Universiteit Leuven (KULeuven, Belgium)		Katrien Verbert Email: katrien.verbert@cs.kuleuven.be
Geocledian GmbH (GEOCLEDIAN Germany)		Stefan Scherer Email: stefan.scherer@geocledian.com
Institut National de la Recherché Agronomique (INRA, France)		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
SYMBEEOSIS LONG LIVE LIFE S.A. (Symbeeosis, Greece)		Konstantinos Rodopoulos Email: rodopoulos-k@symbeeosis.com

ACRONYMS LIST

DMP	Data Management Plan
EC	European Commission
H2020	Horizon 2020 EU Framework Programme for Research and Innovation
ICASA	International Consortium for Agricultural Systems Applications
IPR	Intellectual Property Rights
GI	Geographic Information
IPRs	Intellectual Property Rights

EXECUTIVE SUMMARY

This deliverable outlines the strategy for data management to be followed throughout the course of the project and presents the associated support pack including: (a) the data management guidelines and (b) a template that will be instantiated for all datasets that will be used and/or produced by the project pilots as part of the identified use cases and relevant scenarios.

Based on the progress so far of "T2.1 Use Cases & Requirements", three (3) use cases have been identified which were then further divided in different scenarios. The pilot that will be later defined will constitute instantiations of these use cases. For these use case and scenarios, all of the supporting datasets have already been described with the help of a data management plan template, which has been produced following the Guidelines on FAIR Data Management in Horizon 2020. These templates per different scenarios will be periodically updated to take account of additional decisions or best practices adopted during the project lifetime.

Until the end of the project, they will include detailed individual Data Management Plans (DMPs) for the datasets (or groups of related datasets) that are used per use case. More specifically, these plans address a number of questions related to hosting the data (persistence), appropriately describing the data (data provenance, relevant audience for re-use, discoverability), access and sharing (rights, privacy, limitations) and information about the human and physical resources expected to carry out the plans.

TABLE OF CONTENTS

1 INTRODUCTION 9

2 METHODOLOGY 10

3 DATA MANAGEMENT PLAN GUIDELINES 12

3.1 DATASET CONTENT AND PROVENANCE 12

3.2 STANDARDS AND METADATA 12

3.3 DATA ACCESS AND SHARING 14

3.4 DATA ARCHIVING, MAINTENANCE AND PRESERVATION 15

4 DATASET-SPECIFIC PLANS 17

4.1 OVERVIEW 17

4.2 DMP TEMPLATE 17

4.2.1 Dataset content and Provenance 17

4.2.2 Standards and Metadata 17

4.2.3 Data Access and Sharing 18

4.2.4 Archiving, Maintenance and Preservation 18

5 SUMMARY 19

6 APPENDIX – FILLED DMP TEMPLATES 20

6.1 SCENARIO A: EARTH OBSERVATION DATA ANOMALY DETECTION & CLASSIFICATION 20

6.2 SCENARIO B1: YIELD PREDICTION 21

6.3 SCENARIO B2: PREDICTING BIOLOGICAL EFFICACY 23

6.4 SCENARIO B3-1: CROP QUALITY PREDICTION FOR OPTIMIZING POST HARVEST TREATMENTS OF TABLE GRAPES 25

6.5 SCENARIO B3-2: CROP QUALITY PREDICTION FOR OPTIMIZING WINEMAKING 26

6.6 SCENARIO C1: OPTIMIZATION OF FARM PRACTICES IN THE VINEYARD 27

6.7 SCENARIO C2: CROP QUALITY PREDICTION FOR OPTIMIZING WINEMAKING 29

6.8 SCENARIO D: PREDICTIVE ANALYTICS FOR FOOD SAFETY SECTOR 31

LIST OF TABLES

Table 1: Use Cases and Scenarios 10
Table 2: DMP Template Elements - Content and Provenance 17
Table 3: DMP Template Elements - Standards and Metadata 17
Table 4: DMP Template Elements - Data Access and Sharing 18
Table 5: DMP Template Elements - Archiving, Maintenance and Preservation 18

1 INTRODUCTION

Free and open access to scientific publications and research data is nowadays critically important for researchers, in order to base their work on them and make the next step in their research fields, instead of having to duplicate existing experiments and research work. However, scientific publications are usually accessible only through commercial publishers and accompanied by an access fee, which needs to be paid either by the researcher's institutional library (as an annual subscription fee or on a request basis) or by the researcher himself (in case the institutional library does not have an agreement with the specific publisher). At the same time, research data are not always accessible or at least easily discoverable, as data publishing is not a common practice yet even for institutional repositories. As a result, such data remain stored in offline locations, such as the hard disks and other storage solutions used by the researchers. This issue is not only due to the fact that researchers are not aware of common practices or specific solutions available for the storage and preservation of research data, but also due to the (usually) huge volume of research data which renders commercial data sharing solutions often inappropriate for the specific purpose.

This situation was noticed by the European Commission (EC)[1] and it was decided that actions should be taken for ensuring that at least research publications and relevant datasets that have been funded through programmes of the EC have to be publicly available to all stakeholders. The first steps were taken in the context of the Open Access Pilot of the FP7 funding programme[2], where the design and implementation of an Open Access Plan by projects funded through the FP7 programme was optional, followed by the Horizon 2020 programme in which the Open Access and Data Management Plan is a mandatory part of the proposals.

In the context of the Horizon 2020 programme, the European Commission published a document titled "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020"[3]. The document clearly describes the need that led to the mandate for open access to scientific publications, research data and their associated metadata that have been produced under the Horizon 2020 programme. At the same time, the document states the European Commission's view on the important aspect of data re-use: "*information already paid for by the public purse should not be paid for again each time it is accessed or used, and that it should benefit European companies and citizens to the full*".

In this context, this document provides the plan for the management of research outcomes (and more specifically, the research publications and datasets) that will be produced during the BigDataGrapes project lifetime, as well as those that will be collected from the BigDataGrapes partners (i.e. ABACO, Symbeeosis, AUA, INRA and GEOCLEDIAN) for the respective use cases. It aims to ensure that the research activities of the project are compliant with the H2020 Open Access policy and the recommendations of the Open Research Data pilot. In this context, the project's Data Management Plan (DMP) described in this document outlines how research data and metadata will be collected, processed or generated within the project; what methodology and standards will be adopted; whether and how this data will be shared and/or made open; and how this data will be curated and preserved during and after the project.

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[2,3] http://ec.europa.eu/research/science-society/document_library/pdf_06/open-access-pilot_en.pdf

2 METHODOLOGY

The first step towards the implementation of the BigDataGrapes DMP is the identification and analysis of the characteristics of the data that will be collected and generated within the project.

In this document we define as dataset any set of data (no matter how many files it contains) that is meaningful to be considered as a unit from a data management perspective.

Examples of possible datasets are the following:

- Any data used and produced in the context of the use cases.;
- The set of posts and articles produced by the BigDataGrapes partners;
- Publications and reports produced in the context of the project;

The data analysis phase is part of the definition of the BigDataGrapes use cases (WP2) and the BigDataGrapes pilots (WP8) focusing on the data types and formats, metadata standards, as well as the existing licensing options used. The latter is particularly important to allow the DMP to meet any specific requirement originating from the usage license applied on data.

Based on the progress so far of "T2.1-Use Cases & Requirements", three overarching (3) use cases have been identified which were then further divided in seven (7) different scenarios. The pilots that will be later defined will constitute instantiations of these scenarios. For these use case and scenarios, an adequate number of the supporting datasets have already been described with the help of the Data Management Plan template presented in Section 4. More specifically, to initiate this procedure for each use case, a questionnaire in the form of a spreadsheet has been created and circulated to the project partners regarding the datasets relevant to the use cases and scenarios (see Table 1).

Table 1: Use Cases and Scenarios

Use cases	Scenarios
A. Data Anomaly Detection & Classification (link)	A. Earth Observation Data Anomaly Detection & Classification
B. Prediction (link)	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 (link)	C1. Optimization of Farm Practices in the Vineyard C2. Management Zones Delineation for Vineyards
D. Food Protection (link)	D. Supply Chain Risk Prediction Dashboard, Price Prediction Dashboard, Price & Fraud Correlation Dashboard

In that way, the challenges that the partners face when accessing published data will be defined. Their needs in terms of support for publishing data collected from their communities will be also captured according to the Open Access mandate of the European Commission through Horizon 2020. Through this process we aim to map the landscape of data in the specific context of the BigDataGrapes project and to obtain a better understanding on the context in which the DMP would function. On top of that, the latest version of the Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 (published in March 2017)[4] was used for ensuring that the project's DMP will correspond to the Guidelines, meeting all latest and updated requirements.

The next and final version of the questionnaire spreadsheets per use case that will be delivered at M33, will follow the final definition of the use cases, as reported in D2.1-Use Cases & Technical Requirements Specification. Both these documents will provide additional details on the following aspects:

- The final list of datasets that will be used for each use case/ scenario.
- The data analysis phase, which aims to the identification, extraction, organization and analysis of all related information from the use cases' partners (ABACO, Symbeeosis, AUA, INRA), i.e. how the data will be collected, when the data will be collected, and generally all dimensions addressed by the questionnaires.

[4]http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

3 DATA MANAGEMENT PLAN GUIDELINES

Since the project includes different use cases, this section outlines how data and datasets related to these use cases were managed. Every use case contains a group of datasets sharing similar characteristics, e.g. created or collected under similar circumstances, owning the same sharing access plans and/ or Intellectual Property Rights (IPRs). The guidelines provided in this section outline what information will be provided by the DMP. It identifies 12 specific questions, categorized in four groups (presented in each subsection below). The four DMP template sub-parts in Section 4 correspond to these groups and will answer the same questions.

3.1 DATASET CONTENT AND PROVENANCE

1. What type of data has been collected or created?

In the BigDataGrapes pilots (WP8), data can be derived from one or more datasets that relate to each use case. The DMP will explain the background (retaining provenance) of the described dataset. Imported data can then be combined, processed and analyzed, generating additional data. A description of the operations leading to these newly generated datasets should also be included. As an example, the various use cases will possibly include geographic information (GI) and time series, often combined, leading to spatio-temporal datasets.

3.2 STANDARDS AND METADATA

2. Which data standards will the data conform to?

The consortium will strive to comply or reuse existing standards whenever possible. Although original data sources may conform to different formats and standards, data processed by the BigDataGrapes data layer will likely be transformed into formats complying with a set of well-known standards for the agri-food sector. As an example, relevant standards could be the following:

- AgroVoc [5]: a controlled vocabulary for describing food, nutrition, agricultural, marine, forestry, environmental information. It is also part of the GACS initiative[6], which aims to map the core concepts of three major thesauri, namely AgroVoc, CAB[7] and NAL[8].
- ICASA[9]: data format for documenting experiments and modelling crop growth and development, facilitating exchange of information and software.
- QUDT[10] a set of 80 ontologies covering Units of Measure, Quantity Kinds, Dimensions and Types.
- SOSA[11] an ontology used to describe sensors, observations, samples and actuators.
- ISO 19115 [12] provides information about the identification, the extent, the quality, the spatial and temporal aspects, the content, the spatial reference, the portrayal, distribution, and other properties of digital geographic data and services.

[5] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus
[6] http://www.agrisemantics.org/gacs/
[7] http://www.cabi.org/cabthesaurus/
[8] https://agclass.nal.usda.gov/
[9] http://dssat.net/data/exchange/
[10] http://www.qudt.org/pages/QUDToverviewPage.html
[11] https://arxiv.org/abs/1805.09979
[12] https://www.iso.org/standard/53798.html

- GeoDCAT-AP[13] was designed to enable the cross-sector and cross-platform sharing and re-use of INSPIRE and, more in general, metadata following the ISO 19115/19119 standards and the corresponding XML-based implementation (ISO 19139).
- INSPIRE[14] defines the minimum set of metadata elements necessary to comply with the INSPIRE Directive. In essence it is a profile of ISO 19115 for discovery purposes. It allows a variety of possible implementations.
- UK GEMINI[15] (GEo-spatial Metadata INteroperability iNItiative) is a specification for a set of metadata elements for describing geospatial data resources.
- VoID[16](Vocabulary of Interlinked Datasets) is an RDF Schema vocabulary for expressing metadata about RDF datasets. It is intended as a bridge between the publishers and users of RDF data, with applications ranging from data discovery to cataloging and archiving of datasets

As a general principle, the consortium is going to reuse conceptualizations and adopt broader standards where possible (dcterms, foaf, etc.). As the project will support a Linked Data approach, when applicable, the vast majority of resulting datasets are expected to comply with semantic standards (RDF/S), and additional standardization activities done by the World Wide Web consortium (W3C), such as OAI-ORE's JSON-LD implementation. All of the created RDF data is expected to follow the SKOS-XL[17] data model for better description and linking of lexical entities.

3. What documentation and metadata will accompany the data?

In addition to the data collection activities, BigDataGrapes will also generate its own valuable data assets in terms of metadata that will improve the description, interlinking, normalization, unification, and quality assessment of the collected datasets. The use of W3C standards such as PROV-O[18] for provenance, and DCAT[19] for data catalogue description will be encouraged. Upon decision of the data authors, several datasets will be published as data papers in order to increase their discoverability and usability through the long-established dissemination channels of the journal publishing industry. The consortium will also investigate the possibility of publishing datasets and will consider using existing solutions provided by CNR on implementing a resource catalogue based on CKAN[20] technology. Alternatively, for similar purposes, DCAT-AP can be used. DCAT-AP is a European application of DCAT[21]and driven by DG Connect it is the EC recommendation for open dataset descriptions throughout the continent. A DCAT-AP entry's metadata description will itself cover most of the individual questions posed by the DMP. To cover the geospatial aspects of the identified data assets, it seems appropriate to further build upon similar standards, like GeoDCAT. GeoDCAT-AP is an extension of DCAT-AP for describing geospatial datasets, dataset series, and services. It provides an RDF syntax binding for the union of metadata elements defined in the core profile of ISO 19115:2003 and those defined in the framework of the INSPIRE Directive. Its basic use case is to make spatial datasets, data series, and services searchable on general data portals, thereby making geospatial information better searchable across borders and sectors. This can be achieved by the exchange of descriptions of data sets among data portals. Its purpose is giving owners of geospatial metadata, the possibility to achieve more by providing an additional RDF syntax binding. Further details regarding the metadata layer of the BigDataGrapes platform will be provided in WP3.

13 https://inspire.ec.europa.eu/good-practice/geodcat-ap

14 https://inspire.ec.europa.eu/documents/Metadata/MD_IR_and_ISO_20131029.pdf

15 https://www.agi.org.uk/agi-groups/standards-committee/uk-gemini

16 http://vocab.deri.ie/void

17 https://www.w3.org/TR/skos-reference/skos-xl.html

18 http://www.w3.org/TR/prov-o/

19 http://www.w3.org/TR/vocab-dcat/

20 https://ckan.org/about/

21 https://joinup.ec.europa.eu/asset/dcat_application_profile/description

3.3 DATA ACCESS AND SHARING

4. Which data is open, re-usable and what licenses are applicable?

It is envisaged that most of the datasets resulting out of project activities, will be of an open nature, i.e., data which is freely accessible and protected by minimally restrictive or unrestricted licenses. However, some data could also be obtained via private access. In both cases, the consortium will ensure that any imported data conforms to existing or indicated licenses. In particular, the attachment of the Open Data Commons Open Database License [22](ODbL) to open datasets could be adopted, promoting the three core requirements of: attribution, share-alike and the retention of its open nature. Additional usage and sharing restrictions on the dataset will be defined through additional licenses or modifications of existing alternatives. Justifications for restrictions to dataset access or re-use will also be included in the updated questionnaire spreadsheets per use case at M33. As an alternative, the various Creative Commons licenses could be used as a licensing schema of the BigDataGrapes processed data and also for datasets, publications, research papers and outcomes. Data authors may select the license that fits best to their needs from the following open data licensing options:

- Open Data Commons Attribution License[23]
- Creative Commons CC-Zero Waiver[24]
- Open Data Commons Public Domain Dedication and License[25]

Especially for the public deliverables and publications, it is suggested to apply a CC-BY-4.0 Creative Commons license[26]. Data and software tools produced by BigDataGrapes will not only be available in open access, but also published as data papers and software description papers in appropriated journals. Moreover, it is important for the software components to apply an open software license such as Mozilla Public License[27] or GNU General Public License[28]:

The DMP includes also data protection components, copyright and Intellectual Property Rights issues where necessary or relevant.

5. How will open data be accessible and how will such access be maintained?

Data access will vary depending on storage location (see question 8). Starting with the use case data, measures will be taken to enable third parties to freely access, re-use, analyze, exploit and disseminate the data (bound by the license specifications). To ease the interpretation of a dataset and associated third party agreements, even in a machine-readable manner, the consortium strongly considers publishing a DCAT-AP representation for each dataset on the project's portal. Different access procedures will be implemented, enabling the export of an entire dataset as well as the provision of a querying interface for the retrieval of relevant subsets. Access mechanisms will also be supported as much as possible by metadata enabling search engines and other automated processes to access the data using standard Web mechanisms.

6. Which privacy protocols are implemented?

Typically, BigDataGrapes does not make use of any sensitive data. In the case that a dataset contains sensitive corporate or personal data, privacy protocols need to be established and followed throughout the aggregation, processing and publishing stages. The anonymization of personal information should precede the processing

[22] http://opendatacommons.org/licenses/odbl/
[23]http://www.opendatacommons.org/licenses/by/1.0/
[24]http://creativecommons.org/publicdomain/zero/1.0/
[25]http://www.opendatacommons.org/licenses/pddl/1-0/
[26]https://creativecommons.org/licenses/by/4.0/
[27]https://www.mozilla.org/en-US/MPL/
[28]https://www.gnu.org/licenses/gpl-3.0.en.html

stage. If additional data pre-processing measures need to be taken to safeguard individuals or groups, they will be specified in the updated DMP. If the data processing results still produce sensitive data, access controls will be enforced and described (refer to Question 4).

3.4 DATA ARCHIVING, MAINTENANCE AND PRESERVATION

7. Where will each dataset be physically stored?

Data resulting from each pilot will initially be stored in a repository hosted by a partner participating in the consortium (as it has been defined in the questionnaire spreadsheets per use case). Depending on the nature of the data, a dataset might eventually be moved to an external repository, e.g. the European Open Data Portal[29] or Zenodo[30]. Data generated via other means can have additional hosting arrangements. All of the datasets will be stored and processed in the designed BDG stack and depending on the license will be uploaded to other open data portals or remain stored in the BDG platform. The software produced in the project will be publicly available for accessing and downloading. An open repository such as Github[31], will be used to store the source code of all the software components produced by the project and the related documentation.

8. Where will the data be processed?

In the pilot use cases, data will also be processed strictly within the BigDataGrapes processing layer which is set up for that purpose (as it has been defined in the questionnaire spreadsheets per use case). Any deviations from this understanding should be specified and motivated.

9. What physical resources are required to carry out the plan?

During the pilot project phase, hosting, persistence and access will be managed by the project partners' infrastructure. Partners with the most suitable hosting and processing capabilities have been identified early in the project lifetime. Information about the physical resources required for long term maintenance of the data, e.g. hosting capabilities, big data processing clusters, virtual machines, cloud services, etc., will be provided during the course of the project. This information should also include an approximation of the costs involved.

10. What are the physical security protection features?

During the pilot project phase, different security measures will be setup to restrict data and processing (e.g. the use of SSH public keys). Once a dataset is published and its access enabled, state-of-the-art security solutions will be exploited to ensure that the data cannot be tampered with and its veracity can be guaranteed.

11. How will each dataset be preserved to ensure long-term value?

Since the majority of data integrated and generated within the BigDataGrapes infrastructure will abide by the Linked Open Data (LOD) principles, the consortium will follow the best practices for supporting the life cycle of LOD. This includes its curation, repair and evolution, thus also increasing the likelihood that machine-readable structured datasets (and associated metadata) resulting out of project efforts can also be of long-term use for third parties.

12. Who is responsible to deliver the plan?

[29] http://open-data.europa.eu/en/data/
[30] https://www.zenodo.org
[31] https://github.com/BigDataGrapes

Different consortium members will be tasked with carrying out different aspects of the DMP. The coordinator is in charge of the overall management of the DMP and the partners' responsibilities. If the responsibilities are split, the DMP should outline them.

4 DATASET-SPECIFIC PLANS

4.1 OVERVIEW

The scope of this section is to present in detail an appropriate template that will be used to establish DMPs for each dataset aggregated or produced during the project per pilot as part of the identified (until the time of writing this report) use cases and relevant scenarios. Examples of the DMP template filled with values for different datasets to be used within two different scenarios are presented in the Appendix.

4.2 DMP TEMPLATE

4.2.1 Dataset content and Provenance

Table 2: DMP Template Elements - Content and Provenance

Dataset name/ title	*The title of the dataset/ data package*
Responsible(s)	*Responsible for the dataset/ collection*
Description	*A general description of the dataset, indicating whether it has been:* 1. *aggregated from existing source(s)* 2. *created from scratch* 3. *transformed from existing data in other formats* 4. *generated via (a series of) other operations on existing dataset* *The description will also include the reasons leading to the dataset, information about its nature and size and links to scientific reports or publications which refer to the dataset (if any).*
Original sources (Provenance)	*Links and credits to original data sources*
Operations performed	*If the dataset is a result of transformation or other operations (including queries, inference, etc.) over existing datasets, this information will be retained.*

Note: *when completing this section, also refer to question (and answer) 1 in Section 3.1.*

4.2.2 Standards and Metadata

Table 3: DMP Template Elements - Standards and Metadata

Format	*Identification of the format used and underlying standards. In case the DMP refers to a collection of related datasets, indicate all.*
Metadata	*Specify what metadata has been provided to also enable machine-readable descriptions of the dataset. Include a link if a DCAT-AP or EML representation for the dataset has been published.*

Note: *when completing this section, also refer to questions (and answers) 2-3 in Section 3.2.*

4.2.3 Data Access and Sharing

Table 4: DMP Template Elements - Data Access and Sharing

Data Access and Sharing Policy	*To specify extent of access:* • *Widely open* • *Restricted to specific groups* • *Closed* *When access is closed, justifications will be cited (ethical, personal data, intellectual property, commercial, privacy-related and security-related).*
Copyright and IPR	*Where relevant, specific information regarding copyrights and intellectual property should also be provided.*
Access Procedures	*To specify how and in which manner can the data be accessed, retrieved, queried, visualized, etc.*
Dissemination and reuseProcedures	*To outline technical mechanisms for dissemination and re-use, including special software, services, APIs or other tools.*

Note: *when completing this section, also refer to questions (and answers) 4-6 in Section 3.3.*

4.2.4 Archiving, Maintenance and Preservation

Table 5: DMP Template Elements - Archiving, Maintenance and Preservation

Storage	*Physical repository where data will be stored and made available for access (if relevant) and indication of type:* • *Owned by BigDataGrapes partner* • *BigDataGrapes Triple Store* • *Key domain-specific repository* • *Open repository* • *Other*
Preservation	*Procedures for guaranteed long-term data preservation and backup. Targeted length of preservation.*
Physical Resources	*Resources and infrastructures required to carry out the plan, especially regarding long-term access and persistence. Information about access mechanism including physical security features*
Expected costs	*Approximate hosting, access, maintenance costs for the expected end volume, and a strategy to cover them*
Responsible(s)	*Individual and/or entities responsible for ensuring that the DMP is adhered to the data resources.*

Note: *when completing this section, also refer to questions (and answers) 7-12 in Section 3.4.*

5 SUMMARY

This deliverable outlined the BigDataGrapes project strategy for data management plan. It includes the methodology and guidelines that was followed as well as the template that was used for all datasets corresponding to the use cases and relevant scenarios. The DMP presented was updated during the duration of the project through the questionnaire spreadsheets following the detailed definition of the use cases the relevant scenarios, as well as their instantiation via specific pilots.

Before the start of the pilots, all aspects related to the datasets that were used/produced as part of the project pilots were clarified and resolved. These aspects included questions related to hosting the data (persistence), appropriately describing the data (data provenance, relevant audience for re-use, discoverability), access and sharing (rights, privacy, limitations) and information about the human and physical resources that will carry out the data management plans per dataset.

6 APPENDIX – FILLED DMP TEMPLATES

6.1 SCENARIO A: EARTH OBSERVATION DATA ANOMALY DETECTION & CLASSIFICATION

Use Case	*A. EO Data Anomaly detection & classification*
Scenario	*A. Earth Observation Data Anomaly detection & classification*
Real life problem	*In order to make efficient use of Earth Observation (EO) data for Farm Management applications it is crucial to be able to differentiate between data issues and anomalies. This is not a trivial thing. This is a prerequisite to be able to provide warnings to farmers about Management practices. Anomalies detection is possible through the detection of deviations between Expectation and Observation. Inputs that can support this are: Static Heterogeneity of the field (Management Zones) & Typical patterns of expected crop development for the observed environmental conditions; Classification of anomalies should be able to differentiate between Data errors (clouds, shadows, atmospheric disturbances) & Farm Management related issues (Pests, diseases, vegetation stress through missing water or fertilizer or weather related damage).*
Scenario Hypothesis	***HYP1:*** *We are able to detect anomalies in EO data with the support of Management Zones & Typical patterns of expected crop development* • ***GOALS1:*** ○ *Find out if and how we can detect data anomalies in EO data. (what kind of data anomalies (e.g. clouds, cloud shadows, atmospheric disturbances (e.g. fire smoke), vegetation vitality decrease due to water stress, nutrient deficit, pest, disease, damage)?* ○ *Which spatial resolution is necessary for which data anomaly?* ○ *Which benefit do Farm Management Zones bring for this?* ○ *Which benefit do we gain from the comparison of fields with expected patterns of crop development (e.g. by comparing fields with other (reference) fields in the area)* ***HYP2:*** *We are able to classify detected anomalies into data issues and farm management related issues* • ***GOALS2:*** ○ *Find out if and how we can classify anomalies into data issues & farm management related issues. (Understand what kind of farm management related issues are observable in EO data;* ○ *Develop methods to differentiate between data anomalies and farm management related issues)* • ***Potential methods:*** ○ *Time series analysis (e.g. outlier detection after curve smoothing); Data issues like clouds or shadows exhibit characteristic features (e.g. NDVI drops) in specific vegetation indexes and the multispectral images themselves while a decreased vegetation vitality shows other features* ○ *Single Image analysis (e.g. analysis of spectral information or indices)* ○ *Compare spatial patterns in images with expected patterns (Management Zones)* ○ *Compare time series of fields with reference time series (e.g. by generating a reference curve through aggregating all similar fields in an area and searching for deviations)*

A sample dataset of this scenario is presented below:

Dataset name/ title	*Sentinel-2*
Responsible(s)	*Geocledian*
Description	*Sentinel-2A/B MSI visible & NIR bands*
Original sources (Provenance)	*Copernicus EO Programme, ESA*
Operations performed	*Preprocessing, Atmospheric Corrections, Normalized Difference Vegetation Index (NDVI), other Vegetation Indices*

Format	*JSON, GEOTIFF, PNG*
Metadata	*JSON*

Data Access and Sharing Policy	*Widely open*
Copyright and IPR	*N/A*
Access Procedures	*Registration / Subscriptions*
Dissemination and reuse Procedures	*ToU*
Storage	*Geocledian Server (File system, DB)* *In terms of volume the dataset has ~2K scenes with a total size of ~1.5TB of data*
Preservation	*Data will be ingested into the BigDataGrapes Software Stack Persistence layer, which is* **cloud-hosted** with **daily back-ups.** The specific data asset will be updated **every 2 days**
Physical Resources	*Registration / Subscriptions*
Expected costs	*N/A*
Responsible(s)	*Geocledian*

All the rest datasets of this scenario are catalogued in this spreadsheet: https://bit.ly/2K86igC

6.2 SCENARIO B1: YIELD PREDICTION

Use Case	*B. Prediction*
Scenario	*B1.* Yield prediction
Real life problem	*Ever since the first adoptions of precision viticulture, the need for accurate yield predictions has become obvious. Yield estimations can help the growers optimize variable rate applications, the timing of harvest operations, as well as storage and shipping of their*

	production. However, the difficulty of sampling and the lack of efficient methodologies become obstacles that greatly limit the growth and development of the sector.
Scenario Hypothesis	*Ever since the first adoptions of precision viticulture, the need of accurate yield predictions has become obvious. Yield estimations can help the growers optimize variable rate applications, the timing of harvest operations, as well as storage and shipping of their production. However, the difficulty of sampling and the lack of efficient methodologies become obstacles that greatly limit the growth and development of the sector. The traditional method used by table and wine grapes growers for predicting yield is based on weight measurement of bunches, an inefficient and time-consuming operation, which also fails to provide accurate estimations. More sophisticated yield prediction models have been developed based on data of temporal soil and weather patterns; However, the accuracy is still not adequate. Proximal and satellite data that can be converted into vegetation indices demonstrate high correlation to yield, but the number of factors that can affect the indices' values is far too great to be considered a reliable stand-alone data source. The knowledge of spatial patterns within a field is critical to select homogenous zones with site-specific input to better understand and predict the impact of weather, soil and landscape characteristics on spatial and temporal patterns of crop yields to enhance resource use efficiency at field level.*

A sample dataset of this scenario is presented below:

Dataset name/ title	*Yield data*
Responsible(s)	AUA PA Lab
Description	HIstorical yield data
Original sources (Provenance)	Field measurement
Operations performed	N/A

Format	xls
Metadata	N/A

Data Access and Sharing Policy	Confidential
Copyright and IPR	*N/A*
Access Procedures	N/A
Dissemination and reuse Procedures	N/A

Storage	Disc drive storage In terms of volume the provided dataset has >100 rows
Preservation	Data will be ingested into the BigDataGrapes Software Stack Persistence layer, which is **Cloud-hosted** with **daily back-ups**. The specific data asset will be updated approximately **every 6 months**

Physical Resources	Registration / Subscriptions
Expected costs	100 euro/year (https://www.dropbox.com/buy)
Responsible(s)	AUA PA Lab

All the rest datasets of this scenario are catalogued in this spreadsheet: https://bit.ly/2Xt5Jkq

6.3 SCENARIO B2: PREDICTING BIOLOGICAL EFFICACY

Use Case	*B. Prediction*
Scenario	*B2.* Predicting Biological efficacy
Real life problem	*There is a need in extracting 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.* *The real-life challenge applies to both farmers and companies.* *Farmers during the wine making process produce valuable high-quality by-products that may be used in other industries. Nevertheless, a farmer using the state of the art doesn't exploit all the significant parameters and data that play an important role to the final quality and value of its products. The challenge is to be able to exploit data from diverse sources in order to predict some key quality parameters of the products and by-products that will eventually find an application in the industry.* *Potential buyers or companies on the other hand, perform market research and evaluations in order to choose suppliers of raw materials. Nevertheless, using the state of the art there is no efficient and economical way of knowing which one best suits a specific need, except trial and error by sampling and performing lab measurements on every raw material. Linking data such as the location of a domain, the weather conditions in the area or the cultivation methods can lead to conclusions regarding the most suitable supplier and raw material for a specific product with a specific biological function.* *The goal of the pilot is to prove the correlation between data from the field and the quality of extracts developed from vine materials.*
Scenario Hypothesis	*The main purpose is to find how we can link crop location and weather conditions to the biological quality of the products. A company can then choose a list of suppliers for a specific need, just by evaluating crop location and weather conditions and thereby reaching conclusions regarding biological activity of by-products. A farmer on the other hand, can perform decision making by evaluating location and weather conditions on his field and thereby reaching conclusions regarding biological activity of its products. The farmer will then be able to make decisions on the commercialization of the by-products.* *This scenario hypothesis is aiming to create a predictive model that will correlate parameters concerning weather conditions and parameters linked with biological efficacy. The appropriate algorithms will be created that will use existing datasets and explore the relationship between them. Datasets concerning weather conditions will work as independent variables, while the datasets concerning biological efficacy will work as the dependent variables. A number of potential correlations will be generated (regression models?) between them. Once the correlations are generated, the selection process of the ideal correlation will focus on minimum complexity and error.*

	This scenario hypothesis has the potential for increased scalability using additional weather and spatial data by choosing larger territories as points of interest. *Our goal is to develop a decision support system (DSS) that nurtures users' trust. To achieve this goal, the system must be transparent, meaning it must be able to clearly communicate the prediction model with users and show differing effects of input variables on the model's output. Research had suggested that visual tools are the most efficient for these tasks. Start from quality parameters and see which data (inputs) impact them.*

The dataset will be used in conjunction with data provided by Geocledian.
A sample dataset of this scenario is presented below:

Dataset name/ title	Vine leaf extract var. 1
Responsible(s)	Symbeeosis,
Description	Extraction using water soluble solvents
Original sources (Provenance)	Laboratory analyses files
Operations performed	Ultrasounds, Maceration

Format	csv, xls
Metadata	N/A

Data Access and Sharing Policy	widely open
Copyright and IPR	N/A
Access Procedures	Registration / Subscriptions
Dissemination and reuse Procedures	ToU

Storage	Disc drive
Preservation	Cloud storage Data will be ingested into the BigDataGrapes Software Stack Persistence layer, which is **cloud-hosted** with **daily back-ups**. The specific data asset will be updated **every 2 days**
Physical Resources	Registration / Subscriptions
Expected costs	200 euro/year + PMs
Responsible(s)	Symbeeosis

All the rest datasets of this scenario are catalogued in this spreadsheet: https://bit.ly/2I134t1

6.4 SCENARIO B3-1: CROP QUALITY PREDICTION FOR OPTIMIZING POST HARVEST TREATMENTS OF TABLE GRAPES

Use Case	*B. Prediction*
Scenario	*B3-1.* Crop Quality Prediction for Optimizing Post Harvest Treatments of Table Grapes
Real life problem	*TVineyards demonstrate high levels of variability of both yield and quality of the production. Selective harvesting is an example of targeted management, in which split picking of fruit at harvest is performed according to different yield and quality parameters of management zones within the field. Strategies such as selective picking may be highly profitable for grape growers; however, predictions of high accuracy are a requirement for this approach to be effective. A tool that can provide highly accurate information on crop quality can help the growers to optimize harvest, storage and processing of table grapes.*
Scenario Hypothesis	*The table grapes growers are in need of a system that will help them optimize the timing of table grapes harvest and storage based on data from multiple sources. They will also receive production data and assure them whether the production covers the specific standards that are set by the supermarkets. A powerful system that allows growers to efficiently plan the harvest and storage of their table grapes will greatly benefit them and increase the overall production quality of the sector.*

A sample dataset of this scenario is presented below:

Dataset name/ title	*Canopy sensing*
Responsible(s)	AUA PA Lab
Description	Canopy sensing and vegetation indices data
Original sources (Provenance)	Proximal remote sensors
Operations performed	Data filtering for outliers

Format	csv, xls
Metadata	N/A

Data Access and Sharing Policy	Confidential
Copyright and IPR	*N/A*
Access Procedures	*N/A*
Dissemination and reuse Procedures	*N/A*

Storage	SD Card, disc drive storage In terms of volume this dataset has ~100K rows

Preservation	Data will be ingested into the BigDataGrapes Software Stack Persistence layer, which is **cloud-hosted** with **daily back-ups**. The specific data asset will be updated approximately **every 6 months**
Physical Resources	Registration / Subscriptions
Expected costs	100 euro/year (https://www.dropbox.com/buy)
Responsible(s)	AUA PA Lab

All the rest datasets of this scenario are catalogued in this spreadsheet: https://bit.ly/2QXxnUb

6.5 SCENARIO B3-2: CROP QUALITY PREDICTION FOR OPTIMIZING WINEMAKING

Use Case	*B. Prediction*
Scenario	*B3-2. Crop Quality Prediction for Optimizing Winemaking*
Real life problem	*Wine making needs knowledge on grape quality at harvest. Different sugar content in wine grapes can produce wine with different characteristics. Moreover, some quality parameters like the concentration of nitrogen in wine grapes can affect the vinification process. As a result, wine makers have no idea on the quality of the wine grapes that they buy from the wine grape growers and adapt their vinification process accordingly while growers cannot obtain higher selling prices for their products due to better quality.* *Moreover, during wine making process, the effects of each step on the final wine quality are well known. The quality of wine is mainly due to human actions after harvest.* *The main current purpose is to understand how to improve grapes quality to make a good wine and which parameters drive grapes quality.*
Scenario Hypothesis	*The main purpose is to know the variables that have an effect on quality at harvest and then on the final product (wine).* ***Goals:*** • *To identify the relevant leverage (vintage, yield) on a quality measurement (polyphenol content, aromas for example) to optimise wine making.* • *To link quality parameters with environmental data* ***Hypothesis:*** • *Which parameters (environment, genetic, …) have the main impact on grapes quality to optimise winemaking? to optimise alcoholic fermentation (fermentation strategy)?* • *How to obtain a satisfying grapes quality at harvest to make a good wine and have a satisfying alcoholic fermentation/ aroma composition?* ***Solution:*** • *Understand what is going to influence alcoholic fermentation/ aroma production for a variety of vine (year of production, environment) to be able to have an optimised fermentation and winemaking strategy.* • *Maybe we need to work separately in function of the type of wine (red, white, rosé)* • *Start from quality parameters and see which data (inputs) impact them.*

A sample dataset of this scenario is presented below:

Dataset name/ title	*Genetic Data*
Responsible(s)	INRA
Description	Genetic profile, Morphological description, origin, etc.
Original sources (Provenance)	Reseau Français des Conservatoires de Vigne
Operations performed	Query on the database

Format	csv
Metadata	N/A

Data Access and Sharing Policy	Restricted to specific groups
Copyright and IPR	*N/A*
Access Procedures	Registration - access given only by the database owner / Subscriptions
Dissemination and reuse Procedures	The RDF triples produced during BDG project

Storage	csv produced, graphDB In terms of volume the provided dataset has ~2.6K rows
Preservation	Data will be ingested into the BigDataGrapes Software Stack Persistence layer, which is **dataverse INRA, zenodo** with **daily back-ups**. The specific data asset will be updated **every year**
Physical Resources	*N/A*
Expected costs	Free for INRA agent
Responsible(s)	INRA

All the rest datasets of this scenario are catalogued in this spreadsheet: https://bit.ly/2F09s1Q

6.6 SCENARIO C1: OPTIMIZATION OF FARM PRACTICES IN THE VINEYARD

Use Case	*C. Farm Management*
Scenario	*C1.* Optimization of Farm Practices in the Vineyard
Real life problem	Management Practices as irrigation, fertilization and phytochemicals are regularly over or underestimated with respect to the real plant needs. Especially in case of overestimation in quantities there is a negative countereffect towards the plants.

Scenario Hypothesis	*For the last campaign 2018, in the two Italian vineyards pilots we will perform a data analysis crossing mid and high-resolution multispectral satellite data and best practices data from the farmer.* *The hypothesis and relative goals are the following:* • ***HYP 1:*** *The fertilization or phytochems spraying actions can be supported by satellite data.* ○ ***GOAL 1:*** *Before and after the management actions are there detectable differences in satellite data? Is an increasing of the plant health detectable afterwards? (e.g. increases in certain vegetation indexes or changes in patterns on the field (homogenization?))* ○ ***Goal 2:*** *Are the actions justified by a real evident problem in the vineyard? (is an anomaly (see above) detectable before the event?)* ○ ***Goal 3:*** *Do the satellite data provide a benefit to the farmer when he has to decide on fertilization? Additional for the next campaign (2019)* • ***HYP 2:*** *There are interaction effects between weather, management practices and vegetation/fruit qualities that we are able to detect (what kind of effects? Methods: Machine Learning?)* ○ ***Goal 4:*** *Develop a model to predict where and when a fertilization is required* ○ ***Goal 5:*** *Provide a real Decision Support System to the farmer to improve grape quality and quantity.*

A sample dataset of this scenario is presented below:

Dataset name/ title	*SENTEK DRILL & DROP TRISCAN*
Responsible(s)	Abaco
Description	Soil moisture and temperature profiler from 10 cm to 60 cm deep
Original sources (Provenance)	sensor data stored in Fieldclimate cloud
Operations performed	Data comparison with water balance output

Format	Json, csv, xls
Metadata	N/A

Data Access and Sharing Policy	data access reserved to registered users
Copyright and IPR	*N/A*
Access Procedures	Registration / Subscriptions
Dissemination and reuse Procedures	SITI4Farmer cloud platform

Storage	Abaco server / Pessl Fieldclimate cloud In terms of volume this dataset has ~230K data points

Preservation	*N/A*
Physical Resources	*N/A*
Expected costs	*N/A*
Responsible(s)	Abaco

All the rest datasets of this scenario are catalogued in this spreadsheet: https://bit.ly/2WtrUWg

6.7 SCENARIO C2: CROP QUALITY PREDICTION FOR OPTIMIZING WINEMAKING

Use Case	*C. Farm Management*
Scenario	*C2. Management Zones Delineation for Vineyards*
Real life problem	*Delineation of management zones has provided many advantages to the table and wine grape growers by decreasing their input costs and increasing their production value due to selective harvesting. Management zones represent subfield regions within a field with homogeneous characteristics and allow for site-specific management. Nevertheless, the determination of subfield areas is difficult due to the complex factors that can affect crop yield.* *The optimum number of zones to use when dividing a field may vary from year to year and are mainly functions of the following multivariate spatial and temporal attributes: yield maps, soil and topographic properties, electrical conductivity data, remote sensing, vegetation indices and weather data.* *Current methodologies for the delineation of management zones includes the following:* *1) the use of principal component analysis (PCA) to summarize and aggregate the selected datasets,* *2) the Management Zone Analyst (MZA) software using a fuzzy c-means unsupervised clustering algorithm that assigns field information into like classes or potential management zones, and* *3) using multivariate geostatistical analysis, a method based on the coefficient of variation (CV) of each data.* *While these approaches address the complexity of delineating management zones, they do not imply linear, cost-effective and practical usage when more than two variables are introduced to the models. Spatiotemporal modelling, compared to static input, is a challenging task since it includes input dynamics as part of the problem. Additionally, there is insufficient information regarding efficient algorithms that combine further data layers over different spatial scales, in order to define the management zones based on more variables.*
Scenario Hypothesis	*This scenario hypothesis is aiming to create a dynamic tool to delineate management zones. Algorithms will be developed (WP4) to address upcoming important developments of big data, smart farming and open source satellite data. They will aim at extending the existing methodologies using a dynamic multivariate approach and adopting algorithm designs for solving problems that address aspects of the spatiotemporal domain. The proposed method for the delineation of management zones is the use of algorithms that will convert a high dimensional input signal into a simpler low dimensional discrete signal, such that the distance and proximity relationships and topology are preserved. The*

	algorithms will generate a number of options for delineating management zones, taking into consideration two or more variables. Once the options are generated using specific data spectrums, the selection process of the ideal option will focus primarily on the minimum dimensionality, lowest cost and minimum complexity criteria. Those options that are being characterized, as multidimensional, costly and/or complex will be eliminated from the final selection process of the optimal management zones delineation for a specific field and data spectrum. The selection process will depend on the following rules: (1) the number of management zones in which the field is divided should be feasible and dependent on the field size (2) use variable weighting factors by ranking variables according to their importance 3) supports resource optimisation (e.g. for irrigation, fertilization). In addition to the above stated rules, factors such as the existing topography, as well as the minimum actuation from current status, including the complexity of equipment reinstallations and the distance from current state, should be taken into consideration for the final selection of the ideal management zones delineation. This scenario hypothesis has the potential for increased scalability using additional spatial and temporal data.

A sample dataset of this scenario is presented below:

Dataset name/ title	*Eca sensing*
Responsible(s)	AUA PA Lab
Description	Georeferenced soil electrical conductivity data
Original sources (Provenance)	Proximal sensors
Operations performed	Data filtering for outliers

Format	csv, xls
Metadata	N/A

Data Access and Sharing Policy	Confidential
Copyright and IPR	*N/A*
Access Procedures	*N/A*
Dissemination and reuse Procedures	*N/A*

Storage	SD Card, disc drive storage In terms of volume the 17 provided datasets on RapidScan have a total of ~10.5K rows
Preservation	Data will be ingested into the BigDataGrapes Software Stack Persistence layer, which is **cloud-hosted** with **daily back-ups**. The specific data asset will be updated approximately **every 6 months**
Physical Resources	Registration / Subscriptions

Expected costs	100 euro/year (https://www.dropbox.com/buy)
Responsible(s)	AUA PA Lab

All the rest datasets of this scenario are catalogued in this spreadsheet: https://bit.ly/2Wysffs

6.8 SCENARIO D: PREDICTIVE ANALYTICS FOR FOOD SAFETY SECTOR

Use Case	*D. Food Protection*
Scenario	*D. Supply Chain Risk Prediction Dashboard, Price Prediction Dashboard, Price & Fraud Correlation Dashboard*
Real life problem	*It is a common belief that Risk assessment is a very critical part of a food safety system in order to prevent food safety incidents in the supply chain. Today, Quality Assurance and Safety Experts that are working in food companies are using risk estimation approaches that are based on static data such as literature and guidelines published by National Authorities. Such risk estimation approaches are not taking into account the emerging and increasing risks of the global supply chain and cannot predict the risk. This results in several serious food safety incidents that may impact public health, can cause large financial loss for farmers and industries and can damage their "brand" and lose customers. For instance, during the last 15 years, very important fraud issues like the "2013 horse meat scandal" and the "2008 Chinese milk scandal" has greatly affected the food industry and the public health. According to Allianz Product Recall Report, the average cost of a food or beverage recall to a business is more than 9.5$ million. There were 440 recalls of FDA and USDA- regulated foods last year. More than a third of them posed potentially serious health risks. Recent food recalls have included:* • *Contamination from glass, plastic or metal parts* • *Potential contamination from germs* • *Allergens or labels missing allergy ingredient* *Food protection, including safety and fraud, is one of the most critical parameters in food production highly affecting the food companies from the financial and brand point of view. Agroknow is providing a digital solution for the food industry that delivers trends and risk estimation for raw materials, ingredients and finished products, FOODAKAI. This solution is helping the Quality Assurance (QA) and Food Safety (FS) experts working in the food industry to identify risk in their supply chain. The current solution is limited to alarms, statistics, simple trends and search mechanisms. In particular, FOODAKAI analyzes food incidents and recalls globally, delivering insights to the food industry about potential hazards in raw materials, ingredients and products. Moreover, it offers an intelligent online system that minimizes food safety risks by strategically gathering, processing and delivering live food safety data and analytics in an easy, fast and cost-effective way.* *During the last years of FOODAKAI existence, Agroknow has performed a series of focused group and consultation meetings with several companies of the food industry, such as Gallo Winery, Conagra, Campbell, Pepsico, Hershey and Lamb Weston.*
Scenario Hypothesis	*Thus, the main objective of this pilot is to enhance the current digital solution with new modules that will address further needs of the grape and wine supply chain. The enhancement will mainly focus on the further development of Agroknow's Big Data platform with new software modules that will enable advanced data analysis and risk prediction using machine learning and deep learning methods.* *The specific goals of the food protection pilot are:*

	• *To develop a software module that will be able to predict emerging and increasing risks for chemical hazards in the grapes and wines supply chain;* • *To develop a price prediction dashboard that will include algorithms able to predict the prices of agricultural products, including grapes;* • *To develop a food fraud dashboard that will help experts working in the food industry to perform an effective vulnerability assessment for products;*

A sample dataset of this scenario is presented below:

Dataset name/ title	*Laboratory Test*
Responsible(s)	Agroknow
Description	This dataset contains the laboratory tests performed on food ingredients
Original sources (Provenance)	Laboratory results
Operations performed	Data transformation and curation

Format	csv, xls file following a different schema for each authority
Metadata	*Internal data model that is based on FRBR*

Data Access and Sharing Policy	Available to the people that have purchased access to FOODAKAI and/or Data API of the data platform
Copyright and IPR	*Lab test data are collected from the National Authorities and then processed and enriched by Agroknow. Agroknow is the creator and owner of the Food Safety Dataset, an original database. The first component of a database is the structure and organization of the database, which means the types of data that the creator chose to include and how he chose to organize it. This is the "selection or arrangement". The second component is the specific data contained in the database.*
Access Procedures	Registration - access given only by the database owner / Subscriptions
Dissemination and reuse Procedures	*A Data API is provided*

Storage	Agroknow Big Data platform. In terms of volume the provided dataset has ~91.185.000 rows
Preservation	Data will be ingested into the BigDataGrapes Software Stack Persistence layer, which is **cloud-hosted** with **daily back-ups**. The specific data asset will be updated approximately **every 6 months**
Physical Resources	Registration / Subscriptions
Expected costs	*A Virtual Machine on a commercial cloud is used to host the laboratory test data. The cost for such a VM is 50 euros per month. This cost will be covered by the income from premium subscriptions.*
Responsible(s)	Agroknow

Dataset name/ title	*Food Price*
Responsible(s)	Agroknow
Description	This dataset includes food price data for more than 800 products
Original sources (Provenance)	Open data published by the governments (Hellenic Food Market, European Commission and FAO)
Operations performed	Data transformation, translation, and curation

Format	csv, xls file following a different schema for each authority
Metadata	*Internal data model that is based on FRBR*

Data Access and Sharing Policy	Restricted to specific groups.
Copyright and IPR	*Price data are collected from the European Commission and FAO, then processed and enriched by Agroknow. Agroknow is the creator and owner of the Food Safety Dataset, an original database. The first component of a database is the structure and organization of the database, which means the types of data that the creator chose to include and how he chose to organize it. This is the "selection or arrangement". The second component is the specific data contained in the database.*
Access Procedures	Registration - access given only by the database owner / Subscriptions
Dissemination and reuse Procedures	*A Data API is provided*

Storage	Agroknow Big Data platform. In terms of volume the provided dataset has ~388.000 rows
Preservation	Data will be ingested into the BigDataGrapes Software Stack Persistence layer, which is **cloud-hosted** with **daily back-ups**. The specific data asset will be updated approximately **every day**
Physical Resources	Registration / Subscriptions
Expected costs	*A Virtual Machine on a commercial cloud is used to host the pricing data. The cost for such a VM is 20 euros per month. This cost will be covered by the income from premium subscriptions.*
Responsible(s)	Agroknow

Dataset name/ title	*Food Incidents*
Responsible(s)	Agroknow
Description	This dataset contains the food ingredients (recalls and border rejections)

Original sources (Provenance)	The provided dataset is Announced on a yearly basis by each Food safety Authority worldwide
Operations performed	Data transformation, translation, and curation

Format	csv, xls file following a different schema for each authority, html, pdf
Metadata	*Internal data model that is based on FRBR*

Data Access and Sharing Policy	Available to the people that have purchased access to FOODAKAI and/or Data API of the data platform.
Copyright and IPR	*Incidents data are collected from the National Authorities and then processed and enriched by Agroknow. Agroknow is the creator and owner of the Food Safety Dataset, an original database. The first component of a database is the structure and organization of the database, which means the types of data that the creator chose to include and how he chose to organize it. This is the "selection or arrangement". The second component is the specific data contained in the database.*
Access Procedures	Registration - access given only by the database owner / Subscriptions
Dissemination and reuse Procedures	*A Data API is provided*

Storage	Agroknow Big Data platform. In terms of volume the provided dataset has ~91.185.000 rows
Preservation	Data will be ingested into the BigDataGrapes Software Stack Persistence layer, which is **cloud-hosted** with **daily back-ups**. The specific data asset will be updated approximately **every day**
Physical Resources	Registration / Subscriptions
Expected costs	*A Virtual Machine on a commercial cloud is used to host the food incidents data. The cost for such a VM is 50 euros per month. This cost will be covered by the income from premium subscriptions.*
Responsible(s)	Agroknow