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	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	WP2 | Grapevine-powered Industries Big Data Challenges
WORKPACKAGE LEADER	Agroknow
DELIVERABLE N. | TITLE	D2.2 | Data Management Plan & Support Pack
RESPONSIBLE AUTHOR	Panagiotis Zervas (Agroknow)
REPLY TO	pzervas@agroknow.com
DOCUMENT URL	http://www.bigdatagrapes.eu/
DATE OF DELIVERY (CONTRACTUAL)	31/03/2018
DATE OF DELIVERY (SUBMITTED)	30/03/2018
VERSION | STATUS	1.0 | Final
NATURE	REPORT
DISSEMINATION LEVEL	PUBLIC
AUTHORS (PARTNER)	Pythagoras Karampiperis (Agroknow)
CONTRIBUTORS	Evangellos Anastasiou, Katerina Kassimati, Maritina Stavrakaki (AUA), Florian Schlenz, Stefan Scherer (Geocledian), Simone Speringo (ABACO), Sabine Karen Yemadje Lammoglia (INRA), Constantina Litsa (APIGEA)
REVIEWER	Raffaele Perego (CNR)

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, Katerina Kassimati, Maritina Stavrakaki (AUA), Florian Schlenz, Stefan Scherer (Geocledian), Simone Speringo (ABACO), Sabine Karen Yemadje Lammoglia (INRA), Constantina Litsa (APIGEA)
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)

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 DelleRicerche (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 Speringo Email: s.speringo@abacogroup.eu
APIGAIA (APIGEA, Greece)		Constantina Litsa Email: litsa-c@apigea.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

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”, four (4) 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, some of the supporting data have already been described with the help of the data management plan template. 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 ensuing datasets (or groups of related datasets). 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

LIST OF TABLES

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 isa 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, APIGEA, AUA and INRA) 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. 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", four overarching (4) use cases have been identified which were then further divided in ten (10) 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. Risk Assessment (link)	D1. Grape and Wine Quality Risk Assessment (safety) D2. Environmental Impact D3. Long-term Risk Assessment (Insurance Scenario)

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.

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

The next version of the questionnaire spreadsheets per use case that will delivered at M9 (according to Part B, Section 3.1.3 – Versioning of Deliverables)will follow the final definition of the use cases (D2.1 Use Cases & Technical Requirements Specification). Both these documents will provide additional details on the following aspects:

- The data 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, APIGEA, AUA, INRA), i.e. how the data will be collected, when the data will be collected, and generally all dimensions addressed by the questionnaires.

3 DATA MANAGEMENT PLAN GUIDELINES

Since the project includes different use cases, this section briefly outlines how data and datasets related to these use cases could be 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 is expected to 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 modeling crop growth and development, facilitating exchange of information and software.

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.

[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/

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[10] for provenance, and DCAT[11] 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[12] technology. Alternatively, for similar purposes, DCAT-AP can be used. DCAT-AP is a European application of DCAT[13]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. Further details regarding the metadata layer of the BigDataGrapes platform will be provided in WP3.

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 [14](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 M9. 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[15]
- Creative Commons CC-Zero Waiver[16]
- Open Data Commons Public Domain Dedication and License[17]

Especially for the public deliverables and publications, it is suggested to apply a CC-BY-4.0 Creative Commons license[18]. 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

[10] http://www.w3.org/TR/prov-o/
[11] http://www.w3.org/TR/vocab-dcat/
[12] https://ckan.org/about/
[13] https://joinup.ec.europa.eu/asset/dcat_application_profile/description
[14] http://opendatacommons.org/licenses/odbl/
[15] http://www.opendatacommons.org/licenses/by/1.0/
[16] http://creativecommons.org/publicdomain/zero/1.0/
[17] http://www.opendatacommons.org/licenses/pddl/1-0/
[18] https://creativecommons.org/licenses/by/4.0/

important for the software components to apply an open software license such as Mozilla Public License[19] or GNU General Public License[20]:

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 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 will be defined in the updated questionnaire spreadsheets per use case at M9). Depending on the nature of the data, a dataset might eventually be moved to an external repository, e.g. the European Open Data Portal[21] or Zenodo[22]. Data generated via other means can have additional hosting arrangements. The software produced in the project will be publicly available for accessing and downloading. An open repository such as Github[23], 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 will be defined in the updated questionnaire spreadsheets per use case at M9). Any deviations from this understanding should be specified and motivated.

[19] https://www.mozilla.org/en-US/MPL/
[20] https://www.gnu.org/licenses/gpl-3.0.en.html
[21] http://open-data.europa.eu/en/data/
[22] https://www.zenodo.org
[23] https://github.com/BigDataGrapes

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?

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 details 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	*To specify extent of access:* • *Widely open*

Sharing Policy	• *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 AND NEXT STEPS

This deliverable outlined the BigDataGrapes project strategy for data management plan. It includes the basic methodology and guidelines that will be followed as well as the template that can be instantiated for all datasets corresponding to the identified (until the time of writing this report) use cases and relevant scenarios. The DMP presented here will be constantly updated through the questionnaire spreadsheets following the detailed definition of the use cases the relevant scenarios, as well as their instantiation via specific pilots.

It is expected that before the start of the pilots, all aspects related to the datasets that will be used/produced as part of the project pilots will have been clarified and resolved. These aspects include 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 data management plans per dataset.

6 APPENDIX – FILLED DMP TEMPLATES

6.1 EARTH OBSERVATION DATA ANOMALY DETECTION & CLASSIFICATION

Use Case	*A. 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 deviation of Expectation and Observation. Needed knowledge: Static Heterogeneity of the field (Management Zones) & Typical pattern of crop development for current environmental conditions; Classification of anomalies is possible into Data errors (clouds, shadows, atmospheric disturbances) & Farm Management related issues (Pests, diseases, missing water or fertilizer, weather related damage…). This is also interesting for insurances that have to detect damage events.*
Scenario Hypothesis	*We should be able to develop models that allow us to differentiate between EO data issues and anomalies. This would allow triggering warnings to farmers and insurances concerning Farm Management practices or damage events.*

Dataset name/ title	*Sentinel-2*
Responsible(s)	*ESA*
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)*
Preservation	*N/A*
Physical Resources	*N/A*
Expected costs	*N/A*
Responsible(s)	*Geocledian*

6.2 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 the grapes 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 the last cannot achieve higher selling prices for their products due to better quality.*
Scenario Hypothesis	*Giannis has a big winery in North Greece while he cultivates 40 ha of wine grapes of different varieties. Also, collects grapes from local producers. Giannis wants to know the wine grape quality before harvest both from his vineyards and the vineyards of the producers that he collaborates with for optimizing the vinification in his winery. He believes that a holistic approach that includes imagery from UAVs and satellites, weather data from infield weather stations and open source weather along with in field non destructive measurements will fulfill his needs and help him to provide better prices to the wine grape growers according to the produced crop yield quality.*

Dataset name/ title	*Land-based Weather Data*
Responsible(s)	*National Centers for Environmental Information (NCEI)*
Description	*Land-based Weather Data*
Original sources (Provenance)	*National Centers for Environmental Information (NCEI)* *https://www.ncdc.noaa.gov/data-access*
Operations performed	*Soil Temperature, Soil Moisture/Water Content, Humidity*

Format	*Binary*
Metadata	*N/A*

Data Access and Sharing Policy	*Open*
Copyright and IPR	*N/A*
Access Procedures	*Access to the data processed by the pilot is provided as REST API*
Dissemination and reuse Procedures	*The RDF triples produced by processing this dataset are stored into VITIS Open Triple Store. A description of the schema used for the data is provided in the pilot.*

Storage	*Triple Store*
Preservation	*Data are replicated in two cluster nodes, and preserved at least two years after the project end.*
Physical Resources	*Resources: Cluster of at least 4 nodes Access: Registered (free) users of VITIS.*
Expected costs	*Costs: N/A (Infrastructure from GRNET has been provided) Cost Cover Plan: N/A*
Responsible(s)	*Agroknow*