Funded by the European Union

DIVINE - Grant Agreement 101060884 HORIZON-CL6-2021-GOVERNANCE-01-20

DIVINE - DemonstratIng the Value of agri data sharIng for boostiNg data Economy in agriculture

Deliverable D1.1

Title: Data Management Plan

Dissemination Level:	**PU**
Nature of the Deliverable:	**R**
Date:	**31/03/2023**
Distribution:	**WP1**
Editor:	**Soumya Kanti Datta (DIGI)**
Reviewers:	**John McLaughlin (SETU), Kieran Sullivan (SETU), Delia Milazzo (ENG), Stavros Xynogalas (ICCS), Ioanna Roussaki (ICCS)**
Contributors:	**Soumya Kanti Datta (DIGI), Debolina Paul (DIGI)**

Disclaimer

Revision History

Date	Rev.	Description	Partner
12/01/2023	0.1	Document template	DIGI
22/03/2023	0.2	Compilation of all inputs	DIGI
23/03/2023	1.0	Submitted for internal review	DIGI
30/03/2023	1.5	Internal reviews	SETU, ENG
30/03/2023	2.0	Updated version and released for coordinator's approval	DIGI
31/03/2023	3.0	Final review and approval	ICCS

Table of Contents

Index of Tables

Glossary - Acronyms

CERN	European Organization for Nuclear Research
DMP	Data Management Plan
DOI	Digital Object Identifier
EC	European Commission
EEA	European Economic Area
EU	European Union
FAIR	Findable, Accessible, Interoperable, and Reusable
GDPR	General Data Protection Regulations
HE	Horizon Europe
HSTS	HTTP Strict Transport Security
IoT	Internet of Things
IPR	Intellectual property rights
SAML	Security Assertion Markup Language
WP	Work package

Executive Summary

This deliverable provides an initial data management plan of the EU Horizon Europe DIVINE project. The main purpose of this deliverable is to report how research data is generated, collected, and shared throughout the project duration and beyond. In addition to that, standards adopted for common data models, semantic interoperability, data curation, and long-term preservation are summarised. The project activities as well as the DMP follows FAIR principles.

1. Introduction

Digitalisation of agri-food value chain will capture data from farmers, agri-food processors, transport providers, retailers, and other stakeholders. DIVINE will generate new agri-data at its four pilots[1]. But the main focus of the project is the sharing of existing public and private data. DIVINE is capable of dealing with data in the TB range. Most of the data will be collected from established sources e.g., Eurostat, national agricultural ministries, private organisations operating farm management systems, and the infrastructure operated by the pilot owners of DIVINE. Thus, a project-wide data management policy and strategy is required.

DIVINE dedicates Task 1.5 of WP1 for data management plan and this report (D1.1) corresponds to the initial Data Management Plan (DMP). It is a living document with the initial version being delivered at M06 and will be constantly updated, while the next official versions will be prepared at M18 (version to be integrated in D1.3, project report year 2 and released at M24) and at M36 (version to be integrated in D1.4, project report year 3).

1.1 Purpose of the Data Management Plan of DIVINE project

The DIVINE project participates in the Open Research Data pilot, a part of the open access to scientific publications and research data. All project data and outputs, except those deemed to hold exploitable IP will be released with open access to researchers, industries, policy officers, and other agri-data stakeholders. This will foster reuse of data paving a way for higher impact of public investment in research.

The purpose of the first version of the DIVINE DMP is to present a consolidated view for the DIVINE partners regarding research data management. DIVINE policy on data (including agri-data) generation, collection, and sharing throughout the project lifetime and data retention beyond are reported. Routes for reuse of the data within and outside the consortium are also governed by this report. Standards for semantic interoperability, data curation, preservation along with the legal framework are summarised. The DMP ensures that the DIVINE consortium adheres to the FAIR principles [1]. The European Commission's guidelines for research data quality, trusted sharing, secure storage, and privacy aware data collection are followed.

1.2 Structure

This deliverable is structured as follows. Section 1 provides a general reference to the data sharing approach of DIVINE, purpose of the data management plan, and an initial consideration of the ethical and legal framework relevant to the DIVINE activities and this report.

[1] A detailed description of the pilots will be provided in deliverable D5.1 - DIVINE pilot design, management, and deployment - Release 1.

Section 2 provides a summary on the purpose of data generation and collection (at the pilot sites and through stakeholder survey), reuse of existing data, nature & origin of data, its utility, storage, curation, and data retention.

Section 3 dives into the FAIR data management discussing DIVINE's approach for making data findable, accessible, interoperable, and reuseable.

Section 4 and 5 summarise allocation of resources for data storage, curation, retention, and preservation along with data security considerations.

1.3 Ethical and Legal Frameworks

The ethical and legal framework for data collection, storage, processing, and sharing is broadly based on the GDPR [2] which is applicable to EU27 and EEA countries. GDPR regulates the personal data related to individuals within the EU (e.g., farmers in the context of DIVINE) by a legal and/or private person.

The stakeholder data collection (e.g., applicable in Tasks 2.4, 5.1, and 5.4) will be performed in accordance with the GDPR and ethical standards in the respective countries of the data collection (including any additional application regulations for data privacy and protection). Only some parts of the collected data are personal data. In case of extensive personal data collection requirement arises during the project, DIVINE partners will request personal data processing rights from the data controllers leading to formation of personal data processing agreements among the data controllers and DIVINE partners.

In addition to that the framework put forward by the EU Code of conduct on agricultural data sharing by contractual agreement [3] has also been considered. It recognises the data originator's right within the agricultural sector to profit from the usage of the data obtained from their activity. It mentions the importance to grant the data owner a leading decision on the usage and control of their business data to everyone who might use it. The agreement should define the compensation from the exchanged value, monetary or non-monetary. Finally, the contract should use simple language to be easily understood by an average farmer or person in the agricultural sector. The contract should specify terms, definitions, data objectives, rights, obligations, process rules for sharing, security, legal, used application for storage and usage, verification mechanism, and potential future applications.

Further details on the ethical management and legal frameworks will be provided in DIVINE deliverables D1.3 and D1.4 which are due at M24 and M36, respectively.

2. Data Summary

2.1 Purpose of data generation and collection

The main motivation behind agri-data generation and collection in DIVINE is to build an ecosystem for data sharing and demonstrate that with four real-world pilots. In addition to that, DIVINE will carry out a holistic impact assessment of agri-data sharing to promote the analysis to farmers, farmer associations, and other agri-data stakeholders. However, it is to be noted that agri-data sharing is a complex subject due to sensitivity on security, data protection, transparency, trust, and interoperability [4]. All datasets of the DIVINE project may therefore not be shared (any dataset with sensitive data will provide a summary of the data using anonymisation, aggregation, and/or summarisation). Such data sharing is reflected in the following objectives and expected outcomes (EO) of the project (mentioned in Table 1 and Table 2).

Table 1 – Objectives relevant for this deliverable

Objective No.	Objective description
Objective 1	Establish an ecosystem to aggregate private and public agri data that interconnects existing agri data spaces, thereby enabling the sharing of data among stakeholders.
Objective 2	Analyse and adapt information models in the agriculture sector to enable data interoperability across existing agriculture data spaces, Farm Management Information Systems (FMIS), and Agricultural Knowledge Information Systems (AKIS).
Objective 3	Enhance the ecosystem with facilities ensuring increased transparency in data sharing, data trust & sovereignty, data traceability & usage monitoring.
Objective 5	Analyse and adapt agri data-sharing governance models and policies to enable their use by public/ government services; mechanisms to monitor the impact of these models will be built into the ecosystem.

Table 2 – Expected outcomes relevant to this deliverable

EO No.	EO description
EO2	Increase in transparency in data sharing in the agricultural value chain.
EO3	Increased sharing of agricultural data, and the effective and efficient use of private and public data for private and public purpose, particularly through the demonstration of the costs, benefits, risks, and added value as well as the economic and societal potential of agricultural data sharing taking an EU perspective.

EO4	Contribute to an increased uptake of digital and data technologies in the agricultural sector and indirectly contribute an increase in environmental and economic performance of the agricultural sector through increased and enhanced used of digital technologies and data.
EO5	Strengthen policy-making and-monitoring capacities in agriculture and data technologies.

2.2 Reuse of existing data

Each pilot of the DIVINE project reuses existing data that have been generated from public and private data sources. It has been tabulated in Table 3.

Table 3 – Reuse of existing data in DIVINE

Pilot	Nature of data	What data are reused
Slovenia	Public data	Slovenian Open Data portal SiStat - statistical data relevant for farming area; Slovenian Environment Agency for meteorological data; Ministry of Agriculture, Forestry and Food – data collected as part of mandatory reporting from farmers.
	Private data	Previous years data from farmers' registers - Register of cattle (pigs) on the farm - Volos, FADN (farm accounting), Govedo (milk control), register of agricultural holdings, data from application for direct payments.
Ireland	Public data	• Met Éireann (Irish meteorological data) • Department of Agriculture, Food, and the Marine (Variety testing data) • Earth Observation data products (e.g. bio-physical indices such as NDVI, FAPAR, NDRE, CIred-edge, Leaf area content index), • Agricultural statistics (Central Statistics Office, Teagasc National • Irish Soil Information System (for soil type)
	Private data	Previously collected data from – • Recordings in Digital Farm calendars (observations of growth-Stages, fertilisers applications, pesticides applications, irrigations, scouting for pest infestations, seeding/harvesting events, quantity/quality characteristics of harvested yield), • Sensor recordings (Temperature (atmospheric and canopy),

		humidity (atmospheric and canopy), wind (direction, speed), precipitation, soil moisture/salinity (in different depths), leaf wetness, leaf area index, machinery data logs, tissue analysis), • Additional data types (Soil quality, carbon footprint, financial inputs/outputs), • Other data owner providers (Farmers, Irish Farmers Association, Irish Tillage and Land Use Society).
Greece	Public data	• Sentinel II Earth Observation data. • Public weather data. • Agricultural statistics from the national bodies.
	Private data	• Previously collected data from the Gaiasense infrastructure of NP. • CAP indicators.
Spain	Public data	Historical information from the Southern Spain Meteorological Network with over 120 weather stations
	Private data	• Data registered within the farm of ADSC. • Other private data sources on productivity, water performance, irrigations, biofertilizer inputs, mechanical labour and pest control.

2.3 Data types, formats, and collection methods

DIVINE partners are collecting data primarily through the pilots and as feedback from stakeholders. The data collection methods can be divided into two categories –

- The Slovenian pilot utilises **manual data collection**. At present, KGZS, the Farmer Advisory Service of the Slovenian pilot, operates with a benchmarking tool whereby interested farmers manually enter their data. The advisers can then process and analyse this data leading to advise for the farmers.
- The Irish pilot utilises a mix of **manual and automated data collection**. The automated data includes weather data, soil temperature, soil moisture, leaf wetness, machinery and field operation data (e.g., chemical applications). Plant counts & density, growth stage, and disease presence/severity are manually collected.
- The Greek and Spanish pilots employ digital platforms which **automate data collection** through the use of secure web services.
    - The Greek pilot utilises NP's Gaiasense data collection [5] web services collecting IoT data from sensors (e.g., temperature, humidity etc.) and farmers' calendar.
    - The Spanish pilot utilises its own IT infrastructure to collect IoT data from weather stations and other onsite sensors.

In addition to the pilots, WFO has carried out a survey about technical requirements from its member organisations. In this case, manual data collection was performed.

An example of the consent form and an information sheet containing how collected survey answers will be processed will be provided as an annex of deliverable D5.1 (DIVINE pilot design, management, and deployment – Release 1).

At the pilot level, WP5 activities will generate and collect alpha-numeric data (e.g., related to livestock production, crop yields, IoT sensors on agriculture fields). The Irish pilot foresee collecting images of crops as well.

At the project level, a dataset can include various types and formats including stakeholder validation report, interview video files, notes, images of pilot sites. DIVINE will cater for common formats of data generation for the following –

- Project deliverables, internal documents, technical reports, and publications – pdf, txt, doc, docx, rtf, odt
- Spreadsheets – xls, xlsx, ods
- Datasets – csv
- Audio – mp3, wav
- Video – mp4, avi, mov
- Images – png, jpg, jpeg

2.4 Origin of data

DIVINE will collect data not only at the four pilot locations, but also through the national network of CREA, and European and international network of WFO. There are several primary sources of data collection associated with DIVINE:

- Stakeholder interview through an online questionnaire – performed at each pilot site, using the network of WFO.
- Feedback from stakeholder workshops organised for DIVINE data space validation at all pilot sites.
- Market survey for the agri-data marketplace and DIVINE business plan.
- Re-use of existing data (e.g., public data sources from Eurostat).
- Automated collection of IoT sensor data (especially at Greek and Spanish pilots).

2.5 Data storage repository and curation

Each pilot will utilise its own existing database and infrastructure for data collection and storage. The processed data will be stored at a repository provisioned by SETU via its own in-house cloud infrastructure. Open Archival Information System (OAIS), defined in ISO 14721:20120 [6], will be utilised

as a reference model for long term curation and preservation of processed data in this project. The post-project preservation strategy will be finalised by the partners at a consortium-wide discussion around M34. The strategy will also consider any commercial potential and IPR of the partners for future use and access of the curated data.

Through the curation process, the data will be organised, described, cleaned, enhanced, and preserved for later use. The consortium has taken a collective decision to archive open results (including reports, deliverables, publications, and datasets) via a 'community' on Zenodo, which is an open research data repository of CERN.

2.6 Data retention

Although data retention will be triggered post project, DIVINE has considered an initial data retention policy covering all open access reports, deliverables, publications, and datasets (collected or generated at the pilots) will be archived at Zenodo repository of CERN, for at-least 10 years beyond the project conclusion[2].

2.7 Data utility

DIVINE partners have undertaken an initial assessment of external stakeholders who will find the collected and generated data and data sharing approach highly beneficial:

- Individual farmers
- Farm advisory organisations, crop advisors, agronomists, farm associations
- Ministries of agriculture, chambers of agriculture, agricultural institutes
- National and European policy officers working in the agriculture domain
- Agriculture field owners and cooperative associations

2.8 Data management in DIVINE

Data management is performed centrally in DIVINE at Task 1.5: Data and ethical management, planning and assessment, under the task leadership of DIGI. As a part of overall project management and coordination of WP1, data management is performed at the project level as well as the work package level. In this scope, DIGI is responsible for:

- Developing and implementing the data management plan including ethical, legal, and regulatory frameworks relevant to DIVINE.

[2] According to the policies of Zenodo, items will be retained for lifetime, https://about.zenodo.org/policies/

- It also covers writing this report on the project's initial data management plan. It will be updated throughout the project, with the second version prepared at M18 and integrated in D1.3 (released at M24) and the final version to be released at M36 as a part of deliverable D1.4.

- Coordinating data collection through stakeholder surveys (performed in Task 5.1) by WFO, ITC, KGZS, UCD, NP, and ADSC.

3. FAIR Data Management

FAIR data management is the main guiding principle for the DMP of DIVINE to maximise the access to and reuse of research data collected and generated in this project. Also, datasets containing sensitive data will not be completely released as open access. The following sections detail about the FAIR practices adopted in DIVINE.

3.1 Making data findable, including metadata provisions

All data produced and used in the project will be made discoverable with metadata, identifiable, and locatable by means of a standard, persistent, unique identification mechanism called Digital Object Identifier.

3.1.1 DIVINE community in Zenodo

The consortium is currently setting up the community which acts as the central repository, complete with metadata, of all project work in a popular database and is aligned with OpenAIRE. A dedicated glossary will be created to label all data sets and results, and to provide search functionality for stakeholders seeking particular data. Zenodo supports adding search keywords or tags, that optimise the possibilities for datasets finding and their reuse.

3.1.2 Provisions for metadata

Each dataset published in Zenodo and DIVINE data repository will be accompanied with a metadata as follows. A dataset naming convention will be prepared by DIVINE. At present, the naming convention to be used covers the following:

- Dataset name
- DOI
- Dataset description
- Data utility
- Type
- Nature
- Keywords
- Scale
- Origin
- License
- Version number
- Personal data/anonymized data

- Standards and metadata
- Data sharing policies
- Archiving and preservation
- Project name
- Grant agreement number

3.2 Making data accessible

As stated previously, data will be released openly with exceptions such as -

- Data coming from a private source or a third party with a license restricting open sharing or open access.
- Datasets that contain personal information which must be treated in compliance with GDPR.
- Data that have commercial potential and which are therefore treated with restricted IPR licenses.

The project envisages several of its results and data to be necessary for data owners and other stakeholders to reproduce our results, including: (a) raw agri-data from pilots and other private/public sources, with appropriate licenses if required; (b) validation data, generated during the project's work and made available via DIVINE community on Zenodo; (c) metadata and support documentation to enable comparable architecture and information models to be established; (d) underlying data from publications. In addition to the metadata and support documentation that will accompany project data/results, tutorials and training materials will also be provided. Background context and insight will be provided on potentially sensitive data (e.g., farmers' data sets contain 'personally identifiable' information) to ensure as broad as possible an understanding among stakeholders of how to replicate the work themselves.

It is therefore a central interest of DIVINE to make the data accessible. Open results (consisting of data, technical and non-technical publications, software tools) will be centrally stored in Zenodo. This will facilitate reproducibility of project results where access to data for validation of the result is a must.

For the scientific publications, project researchers will be initially directed to Sherpa-Romeo [7] to check the open access policy of any targeted Journals. This will help inform which Journal is eventually chosen. This will create a central repository, complete with metadata, of all project work in a popular database that is aligned with OpenAIRE[3]. DIVINE will also create open access video-abstracts for project papers and public deliverables. Green open access (self-archiving) and/or Gold open access (open access publishing) will be done for the scientific publications and public deliverables.

[3] OpenAIRE – Science. Set Free.: https://www.openaire.eu

3.3 Making data interoperable

One of the intrinsic goals of DIVINE is to assure data produced and collected in the project is interoperable. This facilitates easier data exchange and re-use among researchers, industries, end users, and external stakeholders. To assure data interoperability, DIVINE focuses on the creation of common data models as well as the development of tools and services (in WP2 and WP3) to supply the modules derived from the project's pilots with processed and adapted multi-source data. Adoption of common data models and semantic interoperability mechanisms enable sharing heterogeneous datasets in a standardised way. The characteristics of these common data models are an extension of existing ontologies and vocabularies which are directly or indirectly related to the agricultural domain.

Table 4 – Ontologies for common data models and semantic interoperability

Domain	Related ontology references
Agrifood domain	SAREF4AGRI, DEMETER Agriculture Information Model, AGROVOC, AgroXML/AgroRDF, FoodOn, eCrop, ADAPT, AFarCloud, FOODIE
Agri-machinery	ISOBUS
Weather data	SWEET
IoT	OGC SensorThings API, SSN / SOSA Ontology, sensorML, FIWARE
Statistical and multidimensional data	RDF data cube vocabulary, QB4ST and SDMX
Supply chain	EPCIS from GS1, TOVE
E-commerce retail stores	Good Relations and Schema.org
User generated data	RDF review vocabulary, MUTO ontology
Land and transport data	OLU and OTM ontologies
Data catalogues	DCAT/GeoDCAT-IP
Plant data	EPPO Global Database

The partners will exploit vocabularies that are standard in agriculture domain for all data types present in collected dataset paving way for interoperability. Should it be necessary to use uncommon ontologies/vocabularies, or generate project specific ones, the partners will provide mappings to more commonly used ontologies mentioned in Table 4.

3.4 Increase data reuse

This section outlines the steps that will be taken by the DIVINE partners throughout the project to increase the project data reuse. To increase data reuse, it needs to be appropriately licensed. According to the Grant Agreement, data (including scientific publications and software tools) are owned by the beneficiary or beneficiaries that generate(s) them. Owner(s) of such results will be encouraged to release them with Creative Commons Attribution 4.0 (CC-BY-4.0)[4]. The data retention policy of DIVINE (presented in Section 2.6) will also facilitate increased data reuse for at-least 10 years beyond the project conclusion.

[4] http://opendefinition.org/licenses/cc-by/

4. Allocation of resources

The DIVINE partners do not foresee an immediate cost for short and long term storage, curation, and preservation of the data generated in this project. The Zenodo repository also does not require allocation of any cost.

In case any unforeseen cost arises for providing open access to project data, that will be covered through the responsible partners budget in this project.

5. Data security

Data security is important to the DIVINE reference architecture that is being formulated in Task 2.1 of WP2. In addition to that, WP3 (Agriculture Data Spaces Ecosystem) dedicates Task 3.3 (Data Transparency, Protection, Trust, Sovereignty, Traceability & Usage Monitoring) to ensure data security, transparency, protection, and trust on data. To further illustrate the methods, DIVINE will use claim-based access control using AuthZ/AuthN sharing data schemes, such as SAML, to ensure separation of authentication & authorisation concerns. All data transfers utilise secure connections (e.g., HSTS).

For data curation and preservation, data security benefits from the infrastructure security policies of Zenodo[5]. All the project data uploaded to this certified repository are stored in CERN's EOS service[6]. It is composed of an eighteen petabytes disk cluster. Each file copy of the deposited data has two replicas located on different disk servers for data recovery. Data integrity is guaranteed with two independent MD5 checksums.

[5] https://about.zenodo.org/infrastructure/
[6] https://eos-web.web.cern.ch/eos-web/

5. Conclusion

This deliverable is a report on the initial data management plan of the DIVINE project following the FAIR guidelines. The developed DMP relies on standards like DOI, OpenAIRE initiative, and Zenodo open access data repository for sound and seamless implementation of the data management methodology. It will assure FAIR principles observed to the open results (e.g., data, publications, and software tools) of the project. This document will be updated throughout the project duration and a final DMP will be released at M36 together with D1.4 (Project Report Year 3).

6. References

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

[2] REGULATION (EU) 2016/679, accessible at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[3] Report on EU Code of conduct on agricultural data sharing by contractual agreement, accessible at https://cema-agri.org/images/publications/brochures/EU_Code_of_conduct_on_agricultural_data_sharing_by_contractual_agreement_update_2019.pdf

[4] I. Roussaki, K. Doolin, A. Skarmeta, G. Routis, J. A. López-Morales, E. Claffey et al., "Building an interoperable space for smart agriculture", Digital Communications and Networks, 2022, ISSN 2352-8648. https://doi.org/10.1016/j.dcan.2022.02.004

[5] The gaiasense smart farming system is an innovative tool for developing digital advisory for agriculture, accessible at https://www.neuropublic.gr/en/smart-farming-gaiasense/

[6] ISO 14721:2012 - Space data and information transfer systems — Open archival information system (OAIS) — Reference model, accessible at - https://www.iso.org/standard/57284.html

[7] Sherpa Romeo is an online resource that aggregates and analyses publisher open access policies from around the world and provides summaries of publisher copyright and open access archiving policies on a journal-by-journal basis, accessible at - https://v2.sherpa.ac.uk/romeo/

[8] Open Scholarly Communication Infrastructure, accessible at - https://www.openaire.eu/