Data Management Plan

Deliverable D33

28 April 2021

Alexandra Korcheva, Anna Sapundzhieva

Pensoft Publishers

Ignasi Bartomeus

CSIC

SHOWCASE

SHOWCASing synergies between agriculture, biodiversity and Ecosystem services to help farmers capitalising on native biodiversity

This project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 862480.

Prepared under contract from the European Commission

Grant agreement No. 862480
EU Horizon 2020 Research and Innovation action

Project acronym:	**SHOWCASE**
Project full title:	**SHOWCASing synergies between agriculture, biodiversity and Ecosystem services to help farmers capitalising on native biodiversity**
Start of the project:	November 2020
Duration:	60 months
Project coordinator:	Prof. David Kleijn Wageningen University
Deliverable title:	Data Management Plan
Deliverable n°:	D33
Nature of the deliverable:	Report
Dissemination level:	Public
WP responsible:	WP4
Lead beneficiary:	PENSOFT
Citation:	Korcheva A, Sapundzhieva A, Bartomeus I (2021). *Data Management Plan*. Deliverable D33 EU Horizon 2020 SHOWCASE Project, Grant agreement No 862480.
Due date of deliverable:	Month 6
Actual submission date:	Month 8

Deliverable status:

Version	Status	Date	Author(s)
1.0	Draft	9 April 2021	Pensoft
2.0	Draft	15 April 2021	CSIC
3.0	Draft	28 June 2021	WU
4.0	Final	28 June 2021	Pensoft

Table of contents

Preface

The SHOWCASE Data Management Plan (DMP) outlines the scope of exploitation of research data within the project. It provides the basis of the project consortium's data management life cycle for the data collected, generated and processed by the participants in the project.

As SHOWCASE participates in the Open Research Data Pilot (ORDP), we aim to improve and maximise access to and reuse of research data generated by Horizon 2020 projects. Therefore, SHOWCASE will develop a DMP within the first six months of the project and will keep it up-to-date until the end of its duration. The main aim of the DMP is to adhere to the FAIR (findable, accessible, interoperable, reusable) data management criteria of Horizon 2020.

Based on the *Milestone 33 'Data requirements collected among all partners'*, the SHOWCASE DMP has been developed in close collaboration with all project partners who provided information on their planned generation, collection, and processing of data for the purposes of building a resilient data management strategy of the project which meets all criteria for open research.

Summary

The SHOWCASE DMP is structured into five sections, which aim to establish the scope and terms of use of research data within the project in accordance with the Horizon 2020 requirements of data management.

The first section provides an introduction to the plan, which outlines the main data management practices that SHOWCASE would implement throughout the five-year project duration, as well as aspects of sustainable management of results and data after the conclusion of the project period.

The second section of the document provides an overview of the commitments that SHOWCASE has made in relation to handling data in a controlled and transparent way, and ensuring an open access to research data and results in line with the EU's Open Research Data Pilot and FAIR data management.

The third section describes the details of data management within the project, focusing on different aspects of the process - from data collection, through data processing, to storage and access provision. The section features information on personal data protection in accordance with the General Data Protection Regulation (GDPR), as well as a break-down of the research data usage into project work packages. Recommendations for relevant data management practices are described in the section.

The fourth section includes an overview of the specific data management details for the project work packages. The specific data formats and data management requirements of work packages are described.

The fifth section of the DMP features concluding remarks on the data management strategy adopted by the project, and it outlines future updates and additions to the plan, which are going to be presented at a later stage of the project's development.

List of abbreviations

EU	European Union
EBA	experimental biodiversity area
DMP	Data Management Plan
GDPR	General Data Protection Regulation
FAIR	Findable, Accessible, Interoperable, Reusable
ORDP	Open Research Data Pilot
EML	Ecological Metadata Language

1 Introduction

As part of WP4 Communication and Dissemination, SHOWCASE has the responsibility to create a detailed plan for sustainable data management of research data used within the project, including ethical aspects such as the collection and usage of personal data. The current document is based on the information about data usage provided by SHOWCASE partners in *Milestone 33 'Data requirements collected among all partners'*, and it specifies the foundations of the data management life cycle for data collected, generated and processed by project partners. The SHOWCASE DMP addresses all key aspects of data management, including the handling of research data during and after the end of the project, data collection and data processing by the project, as well as data storage and preservation.

The plan for the management of research data aims to regulate SHOWCASE data throughout the whole project, and beyond the project lifetime. The data management focuses on both the generation of data within the research network of SHOWCASE (e.g. data from experimental biodiversity areas (EBAs)), and the acquisition and processing of primary and secondary data from outside the SHOWCASE internal research network (e.g. use of publicly available datasets).

SHOWCASE shall collect, generate and process data to achieve the objectives of the project for achieving a better understanding of biodiversity-friendly farming incentives and synergies. The project offers an interdisciplinary research agenda which would encompass a number of scientific disciplines, including ecology and agricultural economics. The ambitious objectives and the implicit interdisciplinarity of the project research call for the use of a variety of research methods. This implies the need for various types of research data, and makes the importance of a well-developed data management strategy from the beginning of the project even more important.

The SHOWCASE DMP features a thorough framework for data management, which covers the handling of research data on every level - from data generation and acquisition to data storage and accessibility. The current document also features a framework for the management of personal data within the project, taking into account the necessary personal data protection measures according to the relevant legislation and EU law.

A separate overview for data management according to work packages is presented, in that way describing the specific data management practices that are necessary to be implemented within the separate research activities of the project. This step is of importance, as SHOWCASE is an interdisciplinary project which would require different types of research data, and therefore different specific data management practices to be added to the general data management framework.

The concluding section of the current document lays out the planned updates of the initial DMP and points out the directions in which it would be enriched as project research proceeds.

2 Open access statement and data sharing

In the context of the Horizon 2020 programme, the European Commission launched the Open Research Data Pilot (ORDP) which aims to ensure that all research data that is generated and used within the Horizon 2020 research projects is findable, accessible, interoperable, and reusable (FAIR). The primary purpose of the FAIR data prescription is to maximise the utility of research data by ensuring its open access to different scientists. The Horizon 2020 Open Research Data Pilot is the most comprehensive EU mandate for open access to scientific publications, research data, and metadata that have been produced under funding from the Horizon 2020 programme.

The DMP aims to ensure that the research activities of the project are compliant with the Horizon 2020 Open Access policy and the recommendations of the Open Research Data Pilot, in all cases in which these are applicable. In this context, the project's Data Management Plan described in this document outlines how research data will be collected, processed or generated within the project and how these data will be curated and preserved during and after the project.

FAIR data is findable, accessible, interoperable and reusable, and applying these principles to the SHOWCASE project, we aim to ensure it is robustly managed. The SHOWCASE DMP thus features key initial provisions on ensuring that the FAIR data guidelines are followed and the project's research data is managed in accordance with Horizon 2020 prescriptions for ensuring the open access and reuse possibilities of data generated and processed within the project.

3 Overview of data

The SHOWCASE project will use a wide variety of datasets from different fields of research and obtained through different data gathering methods. The project partners will generate and process both primary and secondary research data. Data is going to be collected from various sources, including publicly available datasets, the SHOWCASE EBAs, citizen science initiatives, and business organisations.

Based on a survey that was conducted among partners (Milestone 33), the main purposes of data collection were outlined. These are:

- experimental design and literature review
- networking and communication
- inform Experimental Biodiversity Areas
- perform empirical validation
- modelling
- facilitating communication within the project and with relevant stakeholders, communication channels performance analysis, email marketing
- assessment of farmers' motives, perceptions and opinions on biodiversity management; willingness to accept of public and private incentive schemes
- assessment of remote sensing measures of diversity
- assessment of citizen science observations

Both research and personal data are administered within the SHOWCASE project. For that reason, the data management strategy consists of two interdependent streams - namely management of research data, and management of personal data. The measures for assuring sustainable management of both types of data are outlined in the following subchapters of the current document.

3.1 Data summary

The project is going to use different types of data in order to satisfy the necessities of research activities within the separate work packages. The research work of SHOWCASE will use a significant amount of self-generated data which is going to be collected through field work and interview surveys. The other major source of data is to be collected through literature reviews and publicly available datasets access.
N.B. The results of the questionnaire are presented in percentage of the total number of partners, who filled out the survey.

Types of data

The main data formats that the project partners are going to use are as follows:

- .csv - 55%
- .xls (.odt) - 28%
- .tab - 22%
- .xml - 17%

Other data formats to be used within SHOWCASE include .tiff, .geotiff, .json, .txt, raster files, and shape files.

Recommendation: We recommend that partners choose data formats for final data deposition which are lightweight, non-proprietary and widely used (e.g. .csv) when possible.

Origin of data

The origin of the research data that is going to be used within the SHOWCASE project varies. Most partners have declared that they gather data from two or more sources for their research within the SHOWCASE project. The following categories of data origin stand out as more frequently used by partners (in % of respondents using each data type):

- publicly available (free access) - 50%
- institutional or private (permit required) - 28%
- self-generated data - 94%
- other - 22%

The research data acquired by project partners is registered under different data licences, including:

- Open Data Commons Public Domain Dedication and License
- Creative Commons CC-Zero Waiver
- Open Data Commons Attribution

The data platforms used by project partners to obtain data are as follows:

- Zenodo
- OpenAIRE
- Copernicus Hub

Recommendation: The SHOWCASE project recommends that the partners focus on data that originates from acknowledged platforms (e.g. Zenodo, OpenAIRE) that are compliant to the EU data regulation policies and comply with existing data licences.

Data collection

Data is going to be collected for the purposes of the research tasks that are going to be performed within the project's lifetime. Most partners are going to implement a combination of two or more data collection procedures. The main data collection methods that are going to be used, according to the preliminary survey among project partner, are as follows:

- Field work - 61%
- Literature review - 67%
- Interview survey - 61%

Another data collection technique that is to be used by some of SHOWCASE's research teams will include the use of citizen science application entries by farmers and volunteers.

Acquisition of reuse permission

In terms of acquiring reuse permission for data that is property of an external research institution, business organisation, or another type of legal entity, project partners rely on personally contacting data owners as a method to secure reuse permission. Permissions are gathered by partners and are stored by them, along with the relevant research data. Information about data reuse may be featured in respective scientific publications that present results which have been obtained through reused data in accordance with the policies of the scientific publisher (journal, conference proceedings, book, etc.).

Use of metadata

Both descriptive and structural metadata is going to be used within the project's research actions. In generating metadata, the project would aim to follow a unified metadata description standard, which would aim to support the findable, accessible, interoperable and reusable nature of data.

Recommendation: The metadata generated within the SHOWCASE project should be standardised for all project partners. Applying a metadata standard such as the Ecological Metadata Language (EML); Document, Discover and Interoperate (DDI) is recommended. Widely supported descriptive metadata standards have the clear advantage of being highly interoperable, as well as open source or subscription-based software for metadata generation and management. Another option for metadata categorisation applicable to some of the SHOWCASE research data is the geographical metadata standard ISO 19115-1:2014 which is applicable to geo-spatial analysis data that would be generated within the respective research tasks of the project.

We recommend that project datasets follow a unified name convention, which will be agreed upon prior to data publishing. An example of a name convention could be [SHOWCASE_datatset.name_version_creation.date].

An exemplary structure of the minimum characteristics of metadata is proposed below:

- Author(s)
- Year
- Dataset Title
- Data Repository or Archive
- Global Persistent Identifier
- Version, or Subset, and/or Access Date
- Language
- Metadata language
- License of use
- Date of metadata creation
- Hierarchy level
- Character encoding
- Format version

Personal data

A significant part of SHOWCASE partners (more than 78% of the survey respondents) are going to use personal data. The necessary personal data protection measures are therefore of importance to the data management strategy of SHOWCASE.

Use of personal data in research datasets requires a necessary amount of anonymisation. Therefore, data preparation in the project necessarily includes technical steps for achieving a required level of anonymisation of entries, mentioned in 3.4 Data archiving and data protection.

Personal information for non-research purposes is regulated in accordance with the EU requirements of the GDPR, with each institutional partner that would collect and store personal data providing the necessary protection of the data, and performing administrator roles that ensure personal data is not only secure, but also that it is not used for purposes that are not related to the project.

Each institutional partner has an appointed data protection officer (DPO), and has presented their contact information to the consortium members and the European Commission in Deliverable 47. In case partners are not required to appoint a DPO, partners declare that they will treat personal data in line with European and national privacy regulations and follow the project guidelines for the use of personal data. They will keep a detailed data protection policy on file, which will be submitted to the Agency on request.

Recommendation: Use of personal data is a strictly regulated topic, and expressed consent from the parties that provide their data voluntarily is recommended for the purposes of data collection within the project. We recommend the provision of consent forms translated into relevant languages upon requesting and/or receiving research-relevant data from individuals and/or organisations.

3.2 Data access and data sharing

Access to data is governed in accordance with the FAIR objectives of Horizon 2020 and the Horizon 2020 ORDP. For that reason, research data is going to be made available for reuse. Governance of data access and sharing revolves around two major reasons. The first is the provision of data in accordance with the ORDP access principles, and the second is related to data protection.

Project partners have stated that the main way in which they are going to make their data available is through publishing it separately from the research results deriving from the data. This is very much in accordance with the global tendencies to perceive data as a separate research good, which not only serves to make the research results subject to scrutiny and methodological examination, but also is a source of potential new research as well.

The stated preferences for data publishing among project partners are as follows:

- publishing of data in an official permanent data repository (e.g. Zenodo)
- publishing the data in an open access data paper in an academic journal

More specific research data is going to be published as sets of R packages, and on a public website.

Management of data access and data sharing is also governed by the protection of personal data. In order to protect personal data, partners are going to anonymise the datasets on the data preparation level, and might impose some non-monetary restrictions on data access on

the complete datasets. Half of the project partners that participated in the survey stated that they are going to restrict their data under certain conditions. The restrictions of datasets are going to be due to personal data administration, or due to a necessity of protecting intellectual property rights.

Recommendation: SHOWCASE partners have the freedom to choose the means of data publishing, as long as they are compliant with the FAIR principles.

The recommended features of data publishing would assure the following characteristics of FAIR data:

- obtain a digital object identifier (DOI) for published research data to make it findable
- register data under an open data licence (e.g. Open Data Commons Attribution Licence) to make it accessible
- ensuring metadata to be machine readable in order to secure its interoperability
- providing permanent open access to research data for promoting its usability with clear licensing.

Two means of data publishing are outlined as most appropriate - publishing of datasets in respective open access data platforms, and publishing datasets as open access data papers in academic journals.

3.2.1 Data publishing in data platforms

Data publishing in data platforms is a popular means of making data accessible. The main advantage of specialised data platforms such as Zenodo and OpenAIRE is that they make data easily accessible to a large audience, and the individual search costs are minimised. The two platforms that would be recommended within this data management plan are Zenodo and OpenAIRE.

The OpenAIRE platform is part of the EU open research infrastructure, and is designed in accordance with the FAIR principles. The platform offers an easy-to-use opportunity to make research public, and reach out to a large audience of research-oriented stakeholders. OpenAIRE provides a powerful search engine, and having your research presented there is definitely a plus. However, the platform can be used on different stages of the research publishing process, with providing the possibility to upload research data or metadata. The platform can link existing research, no matter whether it is open access or closed access. The platform links existing handles (or DOIs, if available) (OpenAIRE 2021).

Zenodo is a research data repository launched by OpenAIRE and CERN to provide a place for researchers to deposit datasets of up to 50 GB in any subject area. Zenodo is an open source platform. Zenodo has a partial integration with GitHub, where work-in-progress, open issues, and roadmaps are shared. All metadata is openly available under Creative Commons CC0 Waiver. Open access content is freely obtainable. Zenodo assigns a digital object identifier (DOI) to each stored research object, or uses the original DOI of the articles or research objects, if available. The repository allows non-open-access materials to be uploaded but not displayed in public, except for their metadata which remains freely available (Zenodo 2021).

3.2.2 Data publishing in academic papers - description of data papers and recommendation for data types and respective journals

A data paper is a publication in a scholarly journal that serves the purpose to describe a dataset or a group of datasets, rather than report the results of a particular research investigation. As such, it contains facts about data, not hypotheses and scientific analysis (Penev et al. 2017).

The main purpose of the data paper is to provide a citable journal publication that brings credit to data publishers; to describe the data in a structured and easily readable form; to bring the existence of the datasets to the attention of the relevant research community (Penev et al. 2017). Producing a data paper is an option for gaining another citable publication irrelevant of the fact whether there is a research result that comes from the data or not.

Different academic journals have developed the option to publish data papers. Some such journals include Biodiversity Data Journal, ZooKeys, Scientific Data, Data in Brief, etc.

3.3 Data storage

Efficient data storage practices are of crucial importance for SHOWCASE partners, as it allows researchers to keep digital information in a convenient way to be retrieved in the future, with a sufficient level of security. As the size of research datasets is expected to vary greatly, and the amount of data stored to be of significant volume, it is necessary that the data storage practices of the project be designed at the beginning of the project lifetime.

After analysing the partner survey results, the most appropriate data storage practices to be applied within SHOWCASE were outlined to be the use of institutional servers to store research data. Focusing on such an approach is cost-efficient, as no additional costs for data storage are imposed.

Another preferred option for data storage is pointed out to be web hosting, however only a couple of partners have stated a preference towards this data storage option.

The security protocols and control of access to data are therefore the institutional procedures applied by the institutional partners of SHOWCASE, where project-relevant data is going to be stored. The institutional servers that are going to store SHOWCASE data are situated in the European Union (EU), with the exception of one server which is situated in the partner institution within the European Economic Area (EEA). One server is located in Great Britain (GB).

Recommendation: The decentralised data storage practices adopted by the project relies on the assumption that each partner performs regular back up and keeps track of the correct dataset version being stored and made available.

As an alternative to institutional server storage, we recommend the application of large data storage platforms (e.g. Git-based online repositories), which can ensure back up, data versioning, as well as integration with online applications for data sharing (example GitHub, GitLab, OpenAIRE).

3.4 Data archiving and data protection

Data archiving is going to be performed in accordance with the rules and procedures of each partner institution which is going to store SHOWCASE data. Each institution has established backup procedures and has ensured access protection.

Access to data is controlled according to the procedures of the institutions where SHOWCASE data is stored. According to the survey taken among partners, the most popular means of access restriction are:

- password protection
- two-factor authentication
- local network access

Backup procedures are in place for all data that is related to the project. Each institution that stores SHOWCASE data has established backup procedures, with the backup frequency most

often applied being once a day (however, some institutions perform backup more frequently, and others - less frequently).

Securing data protection is also related to personal data protection and storage. Each project partner has ensured that personal data that is gathered and stored within their research activities, is properly protected. Each partner has an appointed data protection officer (DPO) that is responsible for protection of SHOWCASE data stored by partners.

4 Data usage per work package

For each work package using data, a detailed initial framework of data management is provided in the following structure:

- Types and formats of data to be generated/collected; the purpose of data;
- Collection/generation and its relation to the objectives for the project;
- Reuse of existing data;
- Origin of the data;
- Expected size of the data.

Work package 1: Implementing Showcase's network of Experimental Biodiversity Areas (EBA) using a multi-actor approach

- The partners that are performing research within this work package are going to collect data in the following formats: .csv, .tab, .xls, .tiff. The purpose of data collection is related to research performed within the SHOWCASE network of experimental biodiversity areas (EBAs).
- The collection of research data is related to the relevant tasks in the work package, as appointed in the SHOWCASE Grant Agreement.
- Although this work package will mostly rely on self-generated data, reuse of existing data is also going to be performed by some of the partners. Reuse permissions are going to be acquired through contacting the respective institution/owner of the data.
- All partners are going to use self-generate data. Some of the partners are going to support their research with institutional data (permit required), and publicly available datasets.
- Total expected size of the data: unknown. One partner, who focuses on geo-spatial analysis, estimates that the size of their final sets is going to be 1TB.

Work package 2: Identifying incentives to promote biodiversity and ecosystem services in agricultural landscapes

- The partners that are performing research within this work package are going to collect data in the following formats: .xml, .csv, .tab, .xls (.odt), .tiff. The purpose of data collection is related to research performed within SHOWCASE to identify relevant incentives to promote biodiversity in agriculture and farming.
- The collection of research data is related to the relevant tasks in the work package, as appointed in the SHOWCASE Grant Agreement.
- Although this work package will mostly rely on self-generated data, reuse of existing data is also going to be performed by most partners. Reuse permissions are going to be acquired through contacting the respective institution/owner of the data.

- All partners are going to use self-generated data. However, a significant part of the partners participating in the work package are going to support their research with publicly available datasets (64%) and institutional data (permit required) (54%).
- Expected size of the data: unknown.

Work package 3: Increasing the evidence base for synergies between agriculture and biodiversity

- The partners that are performing research within this work package are going to collect data in the following formats: .csv, .tab, .xls, .xml, .json. The purpose of data collection is related to research performed within SHOWCASE in order to increase the evidence base for synergies between agriculture and biodiversity.
- The collection of research data is related to the relevant tasks in the work package, as appointed in the SHOWCASE Grant Agreement.
- Although this work package will mostly rely on self-generated data, reuse of existing data is also going to be performed by some of the partners. Reuse permissions are going to be acquired through contacting the respective institution/owner of the data.
- All partners are going to use self-generated data. Some of the partners are going to support their research with institutional data (permit required) (18%), and publicly available datasets (36%).
- Expected size of the data: 100-500MB per partner.

Work package 4: Communicating the benefits of agrobiodiversity through multi-stakeholder knowledge exchange

- The partners that are performing research within this work package are going to collect data in the following formats: .csv, .xml. The purpose of data collection is related to communication and dissemination purposes, as well as social media network analysis of narratives.
- The collection of data is related to the relevant tasks in the work package, as appointed in the SHOWCASE Grant Agreement.
- Data within WP4 is mostly personal data which is used for the facilitation of communication and dissemination. Partners use self-generated datasets and databases, as well as publicly available social media information for analysis of social networks. All data within the work package is managed in accordance with the EU personal data protection regulation.
- Expected size of the data: unknown.

Work package 5: Management and Coordination

- No research is being done in this work package and no research or personal data is being collected.

5 Conclusion

The SHOWCASE DMP aims to provide a framework for the governance of data within the project. In order to create the framework, project partners coordinated their data management practices, the results of this coordination being made available in the *Milestone 33 'Data requirements collected among all partners'*. The SHOWCASE DMP, thus, established the basic rules for data collection, processing, and protection within the project.

As the data management plan is created at an early stage of the project development, it is scheduled that an update will take place once additions and/or alterations are identified. All changes in the DMP would be applied in the current document.

The recommendations made in the current data management plan are going to be exported, synthesised and designed into *SHOWCASE Data Management Guidelines*, which will be circulated among project partners. The purpose of these guidelines will be to raise awareness on the importance of proper data management, as well as to provide concise and useful information on data management.

Sources

Penev L, Mietchen D, Chavan V, Hagedorn G, Smith V, Shotton D, Ó Tuama É, Senderov V, Georgiev T, Stoev P, Groom Q, Remsen D, Edmunds S (2017). Strategies and guidelines for scholarly publishing of biodiversity data. Research Ideas and Outcomes 3: e12431. https://doi.org/10.3897/rio.3.e12431

OpenAIRE (2021). About OpenAIRE. Retrieved at about (openaire.eu). Accessed on 28 April 2021.

Zenodo (2021). About Zenodo. Retrieved at https://about.zenodo.org. Accessed on 28 April 2021.