

BIOservicES

Linking soil biodiversity and ecosystem functions and services in different land uses: from the identification of drivers, pressures, and climate change resilience to their economic valuation



Title Data Management Plan

Deliverable D7.4

Authors: Antonio Peñalver, Eva Lloret and Raúl Zornoza

Summary

This Data Management Plan (DMP) will contribute to the management of research data collected or generated in the course of BIOServicES work and how these data is going to be stored, published, cited and made FAIR (findable, accessible, interoperable and reusable) beyond the project life. Our goal is to meet the requirements of excellent scientific practice and to allow for accessibility, interoperability, reproducibility of research results. This DMP aims to contribute to the development of the European Soil Observatory (EUSO). To ensure that relevant knowledge and project outputs (including data) are fed to the EUSO, datasets of relevance to the EUSO have been identified.

Report

Document summary	
Document Title	Data Management Plan
Author	Antonio Peñalver, Eva Lloret and Raúl Zornoza
E-mail of principal author	raul.zornoza@upct.es
Lead beneficiary	UPCT
Deliverable No.	D7.4
Work Package	WP7: Project coordination and management
Dissemination type	DMP – Data Management Plan
Dissemination Level	PU - Public
Estimated release date	M6 (29/02/2024)
Actual release date	M6 (29/02/2024)
Date of update	15/07/2025
Data Managers	Eva Lloret and Raúl Zornoza
Single contact point with EUSO	Raúl Zornoza (raul.zornoza@upct.es)
Copyright	© 2024 BIOServicES Project and Consortium



Table of Contents

1. Data summary	4
2. FAIR data	6
2.1 Making data findable, including provisions for metadata.....	6
2.2 Making data accessible.....	7
2.3 Making data interoperable	9
2.4 Increase data re-use	9
3. Other research outputs.....	9
4. Allocation of resources	10
5. Data security	11
6. Ethics	11



1. Data summary

Data collection is an essential part of the BIOServicES project which aims to understand the **interconnection between soil organisms (virus, bacteria, archaea, fungi, protists, nematodes, microarthropods, earthworms, isopods, millipedes, insects and spiders) and the delivery of multiple soil ecosystem functions and services at different scales** (field vs landscape), identifying the pressures and drivers resulting from different land uses and climate change, and performing an economic valuation of the contribution of soil organisms to ecosystem services. For this, BIOServicES is using experimental sites across 8 land uses and 5 biogeographic regions from Europe, as central hubs for co-creation and co-design (multi-actor approach, responsible research and innovation (RRI) and open science).

On the one hand, data corresponding to **history of the experimental sites** will be collected, such as classification of soil, climate, property, crop history, land management and experimental design of the experimental sites (geolocation, plot characteristics, replications, treatment applied, management practices, sampling and analyses) (WP2). These data will be integrated in **Excel spreadsheets (.xlsx)**, including all metadata required for suitable understanding and reuse of the data.

Data will also derive from collection of **existing databases, outcomes from previous and ongoing European and national projects** (e.g., EXCALIBUR, SoildiverAgro, InBestSoil, Diverfarming, SOILGUARD, EJP SOIL, etc), literature review (sources: Zenodo, PubMed, SCOPUS, Web of Science, ScienceDirect, etc.), as well as reports and technical documents (re-used data) to i) unravel the relationship between soil biota and ecosystem functions and services, as well as to predict the potential effect of climate change and/or land use changes on soil biodiversity and the delivery of ecosystem functions and services (WP3); ii) develop a common integrated framework to assess and value, in economic terms, of soil ecosystem services associated to soil organisms in different biogeographic regions and land uses (WP4); and iii) horizontal mapping review of existing European, national and regional policy incentives and nudges contributing to preserving and restoring soil biodiversity (WP5). These data will be integrated in **Excel spreadsheets (.xlsx)**, including all metadata required for suitable understanding and reuse of the data.

We will collect data related to **stakeholders' information** and contacts, mostly in WP1, WP4 and WP5, gathered in **Excel datasets (.xlsx)**. Owing to the **sensitive information** related to persons, this information will never become public nor accessible, being used only by project partners for the objective of



the project, deleting the file at the end of the project. It will be shared on BLOserviceES Teams platform for internal use only.

New data will derive from soil analyses from samples collected in all experimental sites in WP2, which will provide the needed data (characterization of soil organisms abundance, richness, diversity, function and structure, soil structure, provision of food, fibre, raw materials or water, climate regulation, water/air purification, soil pollution reduction, nutrient cycling, flood regulation and water storage, etc). These data will be integrated in **Excel spreadsheets (.xlsx)**. for data analysis, with proper metadata for understanding and proper reuse. WP2 will also produce new DNA sequencing data of soil biological communities (taxonomic), that will be stored in **.bam and .fastq files**, which can be operated with the open access R software. After bioinformatic analysis, taxonomic and functional data on soil organisms will be shared as **Excel spreadsheets**. WP2 will also produce **simulation data** from the optimization of IR spectroscopy for the analysis of soil samples (models between soil spectra and data). These data will be integrated in **Excel spreadsheets (.xlsx)**.

WP3 will exploit the data generated within WP2 to shed light on the yet unanswered issue of the **relationship** between soil biota, the biogeographic region, the type and intensity of land use, and climate change. WP3 will also **upscale** WP2 results, providing tools that can support the management decisions. Thus, the biostatistical analysis of soil biota and identification of links between its composition and environmental variables and the soil ecosystem functions and services will be integrated in **Excel spreadsheets (.xlsx), R, and Python files**. The identification of **niches of microbial taxa and development of hybrid Species Distribution Models** will be integrated in **Excel spreadsheets (.xlsx), R, and Python files**. The identification of a **set of indicators** linking ecosystem services to biodiversity and the landscape upscaling data will be integrated in **Excel spreadsheets (.xlsx), R files, Python files, geotiff files, and .img, .asc and .shp extensions**.

Data generated from the **economic valuation** will be gathered in **Excel spreadsheets (.xlsx)** to identify the real valuation of the contribution of soil biodiversity to marketed and nonmarketed ecosystem services (WP4).

Carbon farming certification scheme will be analysed (WP5) with the aim to evaluate the possibility to complemented it with indicators that ensure the accountability of **soil biodiversity** results beyond carbon, introducing the supplementary indicators provided by WP2. Financial and economic incentives will be also analysed as a part of value proposition. These data will be stored as **.xlxs files**.



Overall, the datasets generated by BIOserviceES to fulfil its objectives will be diverse but mostly numeric and, to a lesser extent, images for mapping. Datasets will mainly consist of observational data (geolocation data, experimental design data, surveys/questionnaires, economic and financial assessment,), experimental data captured by field and laboratory equipment (soil properties) and simulation data (models applied to unravel the relationship between soil biota and ecosystem functions and services, drivers and upscaling in WP3).

Regarding the size of the generated data, relatively many datasets will be produced, but none will be overly big, which will allow them to be collected on spreadsheets for easy accessibility. Just raw data derived from DNA sequencing for identification of taxonomical and functional biological diversity (metabarcoding and metagenomics) will need high storage space for bioinformatics and biostatistics, provided by partner Wageningen University and FiBL.

Data from **soil analyses and models in the different experimental sites** (WP2 and WP3) will be of interest to **feed EUSO**, with indicators representing physical, chemical and biological characteristics and processes of soils from different land uses and regions. BIOserviceES generated data will be useful for researchers in the field of soil health and soil biodiversity, land managers, public administration and professionals/business sectors in the field of soil management and soil recovery.

Concerning scientific publications, all publications will be published Open Access, as Gold publication in open access journals or by publication on Zenodo of the article pre-prints.

2. FAIR data

2.1 Making data findable, including provisions for metadata

Due to the fact that the collected datasets will be diverse and belong to different fields, the **metadata** standard used to describe the dataset will be the Dublin Core Schema, as it is a flexible and commonly used standard and is also the one adopted by the European OpenAIRE repository, which will be used. We have already created the **Community “BIOserviceES” within Zenodo** to increase the efficiency when finding the data of this project. The depositors will create the metadata manually when uploading the datasets to Zenodo. **Keywords** will also be manually added to each dataset uploaded to Zenodo to optimize the possibility for discovery and then potential reuse. These keywords will be provided based on the expertise of the depositors, who are familiar with the most commonly used keywords in their field. All datasets will include some



tabs or pages with metadata and description of data factors and variable names and units.

2.2 Making data accessible

All datasets supporting publications will be made openly and publicly available on the OpenAire repository **Zenodo** upon acceptance of the manuscript for publication. For this, the Community “BIOserviceES” has been created. Moreover, all datasets generated along the project work plan will be made **public three years after completion of the BIOserviceES project**, regardless of manuscript publication. The repository Zenodo assigns a **DOI** for persistent identification and citability of the dataset. The Project dataset identification follows the naming: BIOserviceES_<WPno>_<serial number of dataset>_<dataset-title>_versionnumber(date). Example:

BIOserviceES_WP2_1_soil-indicators_v1(2025-02-10). All datasets will be uploaded to the BIOserviceES Community of Zenodo for easy and rapid finding.

Metagenomic data will be also made openly and publicly available on the free repository GitLab. This will include the whole metagenome data processing pipeline, including self-written scripts (Python, R). **Sequencing data (.bam and .fastq files) will be also uploaded to the repository NCBI** once it has been obtained with an embargo period. This embargo will be the manuscript acceptance or three years after the end of BIOserviceES project. NCBI can generate a temporary link for manuscript reviewers to have access to the data. Once the corresponding manuscript is accepted, the embargo is released. Only data gathered by partners outside the project work plan and protected by IPR, or inside the work plan but containing confidential information (e.g. related to personal interviews, stakeholder’s contact information), will be kept closed for privacy reasons.

Data will be shared among members of work packages continuously as they are being generated, by using the BIOserviceES Teams library associated to the O365 platform of tools for collaborative work (online platform belonging to the coordinator UPCT). Furthermore, **when a dataset turns openly accessible, it will be announced on the web page of the BIOserviceES project www.bioservices.co**, where a link to the dataset on Zenodo, GitLab and NCBI will be available.

Most of the project outputs will be made available as spreadsheets (.xlsx format), with other formats (.bam or .fastq, R files, Python files, geotiff files, and .img, .asc and .shp) for domain-specific datasets (Table 1). Information about the software needed to open these common files will be provided within each dataset or file.



Table 1. Summary of data generated during BIOserviceES project life. The WP in which they will be generated, the format, date of publication and repository used, the metadata or treatment, the license and the security of the data are also indicated

Data Asset	Format	Publication date*	Repository used	Metadata/treatment	Licence	Security
Soil indicators (WP2)	Excel files (.xlsx)	31/08/2031	Zenodo	Dublin Core Schema	CC-BY or CCBY-SA	CERN
New metabarcoding and metagenomics data of soil biological communities (WP2)	BAM and fastq files	31/12/2026	Zenodo, GitLab, NCBI			
Outputs of models between IR spectra and analytical data (WP2)	Excel files	28/02/2027	Zenodo			
Outputs of models to establish the connection between soil organisms and EFS (WP3)	Excel files (.xlsx), R, Python, geotiff, .img, .asc and .shp	31/12/2027	Zenodo			
Soil biological targets (WP3)	Excel files (.xlsx)	31/12/2027	Zenodo			
Monetary values of soil ecosystem services (WP4)	Excel files (.xlsx)	30/06/2028	Zenodo			
New data from surveys, polls and questionnaires and the decision-making process (WP1, WP4, WP5)	Excel files (.xlsx)	31/03/2028	Zenodo			



2.3 Making data interoperable

The depositors will strive to use commonly and internationally used **metadata** in their field of research, since no widely accepted specific metadata vocabularies exist. Data variable names and units, commonly accepted by the **international scientific community and International System of Units**, will be used. The vast majority of the datasets will be made available as spreadsheets, .xlsx format, and to a lesser extent as .bam or .fastq, R files, Python files, geotiff files, and .img, .asc and .shp extensions, so that they can be fully used by potential users.

2.4 Increase data re-use

The project will make use, in general, of the CC-BY license. The particular license applicable to each dataset, will be decided on an individual basis, though the preference will be either **CC-BY or CC-BY-SA**.

The data will remain re-usable after the end of the project by anyone interested in it, with no access or time restrictions, since the Zenodo, GitLab and NCBI repositories will be used. Each archived dataset will have its own permanent repository ID and will be easily accessible. We expect most of the data generated to be made available without restrictions and only datasets subject to IPR and confidentiality issues will be restricted. Where this is going to be the case, agreements will be made based on the individual datasets. Requests for the use of the data by externals will be approved by the project consortium.

Regarding data quality, since all datasets will support publication in peer-reviewed open access journals, the standards of quality (validation of the sample, replication and comparison with results of similar studies) will be met.

New data will be announced on the BIOServicES website and the different profiles in the social media, with links to the page with the dataset on Zenodo, GitLab or NCBI. **Internally, all partners and EUSO will be informed by email about the availability of new datasets**, including the link to Zenodo, GitLab or NCBI.

3. Other research outputs

In addition to the management of data, BIOServicES beneficiaries will also manage other research outputs generated within the workplan of the project, which will be openly accessible, findable and usable (Table 2).



Table 2. Key exploitable results from BIOServicES, with the exploitation strategy to be followed and the protection mechanism.

Partner	Key Exploitable Result (KER)	WP	Exploitation strategy / Exploitation path	Protection mechanism
LGI	Collaborative platform for stakeholders and international initiatives	WP1	Enable stakeholders to share results and exchange information beyond the project	/
FiBL / UPCT	Model to estimate soil biological indicators	WP2	Utility model for end-users	Utility model
CREA	Toolkit of soil biological targets linked to key ecosystem functions and services	WP3	Open licence for end-users to apply the knowledge to monitor soil organisms that are key for the delivery of ES by soil	CC-BY licence
UPCT	Web-based application to analyse the biota of soil samples against reference	WP3	Demonstration of the beta version of the application in scientific conferences and workshops. Provide the release version in Open Access to stakeholders	Open code application
UPCT	Web-based calculator for the economic value of soil ecosystem services	WP4	Utility model for end-users to apply the process to give a monetary value to their soils	Utility model
CMCC	Policy conditions and catalogues of existing soil biodiversity related economic and policy incentives	WP5	Science-policy events (e.g. EU Green Week and/or Conference on modelling for policy support)	CC-BY licence
UNITUS	Toolbox of incentive schemes for soil health	WP5	Distributed via repositories (e.g. Zenodo, OPPLA, Knowledge4Policy)	CC-BY licence
CMCC	Report on Policy recommendations for soil health and biodiversity protection.	WP5	Distributed standalone and as a part of European assessment reports such as by EFDRR, EEA.	CC-BY licence

4. Allocation of resources

There are **no costs** associated to the described mechanisms to make the database FAIR and long term preserved since we are using the Zenodo, GitLab and NCBI repositories. Data files quality will be checked by the data managers who will be also responsible of submit them to the repositories. The project



coordinator has the ultimate responsibility for the data management in the project.

5. Data security

On one hand, the Zenodo, GitLab and NCBI repositories take care of security and backup of the deposited data. On the other hand, regarding the data-sharing process involved between partners during research, Teams Online libraries associated to the O365 platform will be used, and data protection services are provided by Microsoft to prevent the loss of data, within the institutional account of UPCT.

6. Ethics

The ethical aspects related to the personal data collected in BIOserviceES datasets are addressed in project deliverable D7.1. Ethics principles and guidelines for protection of personal data.

