



BIOremediation systems exploiting SYnergies for improved removal of Mixed pOllutants

Deliverable D8.3

Intermediate Data Management Plan

Deliverable information

Responsible Partner:	IDENER R&D
Work Package:	WP8
Author(s):	IDENER R&D
Contributing Partner(s):	All partners
Dissemination Level:	PU – Public
Type:	DMP – Data Management Plan
Due Date:	28/02/2025
Submission Date:	28/02/2025
Version:	2.0



Funded by the European Union under the GA no **101060211**.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or REA. Neither the European Union nor the granting authority can be held responsible for them.

Project Profile

Programme	Horizon Europe
Call	HORIZON-CL6-2021-ZEROPOLLUTION-01
Topic	HORIZON-CL6-2021-ZEROPOLLUTION-01-10
Number	101060211
Acronym	BIOSYSMO
Title	BIOremediation systems exploiting SYnergieS for improved removal of Mixed pOllutants
Start Date	1 September 2022
Duration	48 months

Document History

Version	Date	Author	Remarks
1.1	11/02/2025	IDENER R&D Team	Draft version
1.2	13/02/2025	IDENER R&D Team	Draft version including reviewed version of Annex II with partners' inputs
1.3	14/02/2025	IDENER R&D Team	Revision by EXELISIS; additional information for Annex II
2.0	26/02/2025	IDENER R&D Team	Final review; version to submit

Disclaimer

All the contributors to this deliverable declare that they:

- Are aware that plagiarism and/or literal utilization (copy) of materials and texts from other projects, works and deliverables must be avoided and may be subject to disciplinary actions against the related partners and/or the Project consortium by the EU.
- Confirm that all their individual contributions to this deliverable are genuine and their own work or the work of their teams working in the Project, except where is explicitly indicated otherwise.
- Have followed the required conventions in referencing the thoughts, ideas and texts made outside the Project.

Executive Summary

Deliverable D8.3 provides the intermediate version of the Data Management Plan (DMP) for the BIOSYSMO project, aligning with the Horizon Europe Programme's open science priorities. This document defines the project's research data management strategy, ensuring that data remains findable, accessible, interoperable, and reusable (FAIR). This update includes newly identified relevant data categories, an updated data summary, enhancements to the data management strategy, and details on the initial open data generated and shared within the BIOSYSMO collection in Zenodo and other repositories. The final version of this deliverable will be submitted in month 46 as part of Deliverable 8.4.

Disclaimer: During the preparation of this deliverable the Université Marie et Louis Pasteur (UMLP) was officially established, merging UBFC (Partner 5) and UFC (AE 5.1). However, at the time of publication, most repository links associated with the university had not yet been updated, as the entity still retains its previous domain. Consequently, some links referenced in this document may become non-functional over time. These repository links will be revised and updated in the final version of the DMP once the transition is fully completed.

Table of Contents

Executive Summary	3
Table of Contents.....	4
List of Figures	6
List of Tables.....	7
Table of Abbreviations.....	8
1 Introduction	9
2 Data summary.....	12
3 FAIR data.....	13
3.1 Making data findable, including provisions for metadata.....	13
3.2 Making data accessible	17
3.3 Making data interoperable	18
3.4 Increasing data re-use	20
4 Other research outputs.....	21
5 Allocation of resources	21
6 Data security	22
7 Ethics	22
8 Other issues.....	23
9 Available resources on the BIOSYSMO collection.....	23
9.1 Open data.....	23
9.2 Restricted-access datasets.....	24
9.3 Future focus	25
10 References.....	25
Annex I: Data management questionnaire.....	26
Annex II: Data summary.....	30
WP1. Field sampling and characterization of polluted and treated samples	30
WP2. Computational tools for biosystems design and implementation	35
WP3. Development and optimization of consortia and plant-microbe biosystems.....	39
WP4. Application of improved biosystems and optimization of the bioremediation strategy.....	44
WP5. Field studies, validation and roadmap for wider implementation	47
WP6. Environmental, socio-economic, risk and regulatory assessment	49
WP7. Communication, dissemination and exploitation	51

Annex III: Metadata required by any Zenodo deposit	54
Annex IV: Protocol for uploading datasets to Zenodo	60
Annex V: Dataset index.....	63
Annex VI: Overview of data available in the BIOSYSMO collection for Zenodo	67

List of Figures

Figure 1. Schematic representation of the typical research data management (RDM) lifecycle. Image sourced from Bongaerts (2022) ²	9
--	---

List of Tables

Table 1. Repositories that may be used by the BIOSYSMO consortium.....	14
Table 2. Relevant standard vocabularies and ontologies or guidelines.....	16
Table 3. Recommended and acceptable data formats for sharing, re-use and long-term preservation ⁴	19
Table 4. Distribution of responsibilities regarding the management of research data	21

Table of Abbreviations

Abbreviation	Definition
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
CC	Creative Commons
CERN	Conseil Européen pour la Recherche Nucléaire, European Organization for Nuclear Research
DMP	Data Management Plan
DOI	Digital Object Identifier
EC	European Commission
ENA	European Nucleotide Archive
EU	European Union
FAIR	Findable, accessible, interoperable and re-usable
IPR	Intellectual property rights
LCA	Life cycle analysis
MIBBI	Minimum Information for Biological and Biomedical Investigations
NCBI	National Center for Biotechnology Information
ORCID	Open Researcher and Contributor iD
PID	Persistent identifier
RDM	Research data management
URL	Uniform Resource Locator
UTF-8	Unicode Transformation Format – 8-bit

1 Introduction

Promoting open science is within the core values of the Horizon Europe Programme. A central pillar to the open science policy is the responsible management of research data, which should comply with the principles of 'findability', 'accessibility', 'interoperability' and 'reusability' (the 'FAIR principles')¹. According to the Grant Agreement, beneficiaries shall manage all research data generated in an action under the Programme in line with the FAIR principles and shall establish a Data Management Plan (DMP).

The project DMP is a crucial component for achieving a responsible management of research data. For the purpose of the DMP, research data is defined as any information that has been collected, observed, generated or created as a result of research activities. Therefore, the DMP outlines the data management strategy that will be used by the BIOSYSMO project. Research data typically follows a cycle composed of several consecutive and interdependent processes (Figure 1). Therefore, this strategy covers the full lifecycle (Figure 1) of the research data collected, processed and/or generated within the project.



Figure 1. Schematic representation of the typical research data management (RDM) lifecycle.
Image sourced from Bongaerts (2022)².

The research data lifecycle starts with the **design and planning** of the project and how data will be handled. This stage includes the drafting of the DMP. Some decisions that may be taken during the planning stage is the initial description of the planned data and metadata, the documentation that will accompany the data, the choice of repositories, the choice of file formats or the need for re-using data, among others.

The second stage of the data lifecycle is the **collection and creation** of data, which includes the actual data retrieval. During this stage, data must be organized, stored and backed-up. Metadata must be collected and/or created.

The third stage of the data lifecycle includes the **processing and analysis**. Until this point, the data collected may be considered as raw data. The third stage includes the data curation, analysis and quality control. Analysis workflows may be developed. Provenance and versioning of data must be recorded.

The last step of the data lifecycle is the **sharing and re-use**. First, data is securely preserved. Data ownership and intellectual property rights (IPR) are clarified. Then, the conditions for sharing outside the project are chosen, for example, by selecting the most appropriate sharing license. This last step of the data lifecycle closes the loop and connects with the start of a new cycle.

The DMP includes information on:

- how research data will be handled during and after the end of the project
- what data will be collected, processed and/or generated
- which methodology and standards will be applied
- whether data will be shared/made open access
- how data will be curated and preserved, including after the end of the project

In order to cover this information, the DMP has been structured according to the template provided by the EC in the following sections:

- **Data summary** (Section 2), describes the data that will be collected, generated or re-used within BIOSYSMO and that has been identified so far by the consortium.
- **FAIR data** (Section 3), defines the strategy to make the research data findable, accessible, interoperable and re-usable (FAIR) and is sub-divided in the following sections:
 - Findable data (Section 3.1), includes the measures to making data findable, including provision for metadata.
 - Accessible data (Section 3.2), includes information about the deposit of data and open access to data.
 - Interoperable data (Section 3.3), includes aspects related to standards, vocabularies, ontologies and file formats.
 - Reusable data (Section 3.4), includes information about conditions for data re-use such as clarifying licenses.
- **Other research outputs** (Section 4), describes the strategy for the management of research outputs other than research data, such as publications, deliverables, presentations, etc.
- **Allocation of resources** (Section 5), lists the distribution of responsibilities regarding research management.
- **Data security** (Section 6), describes the measure to ensuring that data is securely stored.
- **Ethics** (Section 7), lists any ethical issue that may arise regarding the handling of data within the project.
- Other issues (Section 8)

The DMP is a living document that evolves throughout the project. The first version, submitted as Deliverable D8.2 in M6, outlined potential scenarios for updates and improvements, including:

- new data or updates to existing data are identified
- changes in consortium policies (e.g., innovation potential, decision to file for a patent)
- changes in consortium composition (e.g., new consortium members joining or current members leaving)

In this context, this updated version includes newly identified relevant data categories in Section 2, an updated data summary in Annex II, and improvements to the data management strategy in Section 3, such as the development of a guiding protocol for uploading datasets to Zenodo (Annex IV: Protocol for uploading datasets to Zenodo). Additionally, it provides a description of the initial open data generated and shared within the BIOSYSMO collection on Zenodo and other repositories, detailed in Section 9.

As with the initial version of the DMP, this intermediate version has been developed in collaboration with all project partners. Information gathered in the early stages through the data management questionnaire (Annex I: Data management questionnaire) was shared with partners to facilitate review and updates to the content:

- Identification of the person responsible for data management for each partner.
- Description of the data: each partner identified the data that they expect to collect, generate or re-use.
- Questions about FAIR data, e.g., regarding institutional repositories or other discipline-specific repositories that the partners may use; any relevant data and metadata vocabularies, standards, formats or methodologies; metadata that may be created for the data; data quality assurance processes.
- Other questions, e.g., regarding any national, funder, sectorial or departmental procedures for data management; any foreseen costs for making data or other research outputs FAIR.

The initial version of the DMP established a comprehensive framework for data management in BIOSYSMO, particularly regarding the research data generated. This updated version reflects the project's engagement with data transfer and the evolving needs that have emerged, such as the development of protocols to ensure data protection until publication while still enhancing the BIOSYSMO collection and promoting open access best practices.

Additionally, this version incorporates the integration of specialized repositories with generalist platforms like Zenodo, highlighting both the opportunities and challenges in managing diverse data sources. The final version will include all data generated within the project, ensuring their availability in accordance with open science principles. Any data not yet published by that stage would be subject to an embargo period until their official release. This document presents the most up-to-date information available at the time of writing while acknowledging areas for further expansion in future iterations.

2 Data summary

The first step for the proper management of data is the identification and description of the data that will be generated in the project. Therefore, this section addresses the goals of data collection/generation and its relevance to the objectives of the project. The utility of the data (i.e., to whom it might be useful) is also addressed. The description of data also includes information needed for the categorization of the data (source, form, format) as well as the identification of any re-used data (e.g., data generated previously for a different project that might impact or influence this one).

Since the initial version, this section has been updated and reviewed, particularly regarding communication, dissemination and exploitation data, reflecting the growing importance of these aspects within BIOSYSMO. Given the project's strong emphasis on dissemination and communication, highlighted by its participation in the [ALL4BIOREM cluster](#) and bilateral collaborations with other projects, this category has been expanded to better capture the data generated in these activities.

The BIOSYSMO project produces various types of data, categorized based on the collection method: observational, experimental, simulation, derived/compiled, reference/canonical, and communication and dissemination data. Each category has distinct characteristics and degrees of reproducibility:

- **Observational data** are captured in real-time, typically outside the laboratory, through observation of a behavior or activity. The collection methods vary depending on the observed system and the desired variable and may include, e.g., sensor readings, telemetry, images, surveys, etc. Because observational data are captured in real time, it would be very difficult or impossible to re-create if lost.
- **Experimental data** are collected under controlled conditions, e.g., in laboratory experiments specifically designed to produce and measure change or to create difference when a variable is altered. This type of data is often reproducible.
- **Simulation data** are generated by mimicking specific processes or systems by using computer test models. These data are reproducible if the model and inputs are preserved.
- **Derived or compiled data** are generated by using existing data points to create new data through some sort of transformation. Existing data are usually compiled from different data sources or datasets and can usually be replaced if lost.
- **Reference or canonical data** belong to static or organic collections of datasets, usually peer-reviewed, published and/or curated. Examples are gene sequence databanks or chemical structures.
- **Communication, Dissemination and Exploitation data** include materials generated for project outreach, such as presentations, reports, publications, newsletters, social media content, and engagement metrics. Also data related to KERs, exploitation potentials, business plans, IPs and ownership aspects are reported. These datasets help track and evaluate the project's impact and outreach efforts.

Likewise, the BIOSYSMO project generates data in different forms such as text, numeric, audiovisual, models, computer code, discipline-specific and instrument-specific.

The description of the data that will be generated or re-used within the project and that has been identified so far is included in Annex II: Data summary. This table will be updated throughout the project as soon as new data is identified.

This annex describes the data that will be generated, collected, processed, analyzed and re-used in the BIOSYSMO project. The data is described in terms of the following characteristics:

- **Partner:** partner generating, collecting, processing, analyzing and/or re-using the data.
- **Task:** task where the data is generated.
- Brief **description** of data.
- **Purpose of the data** generation or re-use and its relation to the objectives of the project.
- **Types of data**, according to the following categories: observational field data (e.g., sensor readings, telemetry, survey results, images); experimental data collected under controlled conditions (e.g., gene sequences, chromatograms, biochemical measurements); simulation data from test models (e.g., climate models, economic models, biological models); derived/compiled data generated from existing datasets (e.g., text/data mining, compiled database); reference (peer-reviewed) datasets (e.g., gene sequence databanks, chemical structures, spatial data portals).
- **Form of data**, according to the following categories: text, numeric, images, audiovisual, models, computer code, discipline-specific (e.g., geospatial data) and instrument-specific (e.g., spectroscopy data).
- **Origin/provenance of the data**, either generated or re-used. If you will re-use any existing data, indicate the source and what you will re-use it for.
- **Data utility** outside the project: to whom might the data be useful outside the project.

Annex II: Data summary has been updated since its initial version and will continue to be revised throughout the project as new data is identified.

3 FAIR data

The FAIR data concept was formulated in 2016 to support the reuse of research data¹. FAIR data are those data that meet the requirements of findability, accessibility, interoperability and reusability. The BIOSYSMO consortium will follow the FAIR Guiding Principles^{1,3}.

3.1 Making data findable, including provisions for metadata

The first element of the FAIR Guiding Principles intends to facilitate the findability of the data generated in the project, i.e., make sure that both humans and computers are able to easily find the data³. Data repositories facilitate this task by, e.g., assigning persistent identifiers (PID) or creating metadata describing the data. Metadata are an essential component to facilitate automatic data findability.

All data deposited in a repository will be identifiable and accessible by means of a standard identification mechanism, i.e., by means of a persistent and unique identifier. The consortium will use Zenodo (<https://zenodo.org/>) as the reference repository of the project. Zenodo automatically assigns a Digital Object Identifier (DOI) to all the deposited items. Nonetheless, other repositories, such as

institutional or subject-based repositories, may be used. If these repositories assign additional persistent identifiers (e.g., Handle Id used by several repositories or HAL Id by Archive ouverte HAL; Table 1), the partners will be careful not to duplicate the identifiers (e.g., making sure a single item is not assigned two DOIs) and will make sure that identifiers are properly cross-referenced. Moreover, the authors of outputs of the project will be identified with the Open Researcher and Contributor iD Identifier Schema (ORCID iD Identifier Schema), which is a persistent identifier of researchers.

Table 1. Repositories that may be used by the BIOSYSMO consortium

Repository	Type of repository	Type of data	Persistent identifier
Zenodo (https://zenodo.org/)	General purpose	Any kind of data	DOI
Archive ouverte HAL (https://hal.archives-ouvertes.fr/)	General purpose	Any kind of data	HAL Id
Dryad (https://datadryad.org/stash) ¹	General purpose	Any kind of data underlying peer-reviewed scientific and medical literature	DOI
RIUBU (https://riubu.ubu.es/)	Institutional (UBU)	Any kind of data	Handle (hdl)
dat@UBFC (https://search-data.ubfc.fr/)	Institutional (UMLP, formerly UBFC)	Any kind of data	none
CLARIN.SI repository (https://www.clarin.si/repository/xmlui/)	Disciplinary ²	Any kind of data	Handle (hdl)
e-cienciaDatos (https://edatos.consorciomadrone.es/)	Institutional (UPM ³)	Any kind of data	DOI
European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/browser/home)	Disciplinary	Nucleotide sequences	none
NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/)	Disciplinary	Nucleotide sequences	none
NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/)	Disciplinary	Array- and sequence-based data	none

¹ Dryad's requirement of the CC0 open access waiver may preclude selection of this repository under certain conditions

² Slovenian CLARIN repository

³ Consortium of Universities of the Region of Madrid and the UNED for Library Cooperation

Silva (https://www.arb-silva.de/)	Institutional, disciplinary	Nucleotide sequences	none
UNITE (https://unite.ut.ee/)	Disciplinary	DNA based fungal species	none
EMPODAT (https://www.norman-network.com/nds/empodat/)	Disciplinary	Chemical occurrence data	none
GitHub (https://github.com/) ⁴	Other	Software code	Digital Object Identifier (DOI), if linked with a Zenodo entry
BIOSYSMO Internal Sharepoint	General purpose	Any kind of data	none
BIOSYSMO Website (https://www.biosysmo.eu/)	Dissemination and communication data	File formats (e.g., PDF, PNG, JPEG)	none

In order to deposit the data, repositories usually create or request the user to add metadata to the data. Therefore, any data item deposited in a trusted repository will be accompanied by appropriate metadata. These metadata are machine-readable and can be harvested and indexed. For example, any entry in Zenodo will at least contain these metadata: type of item (e.g., publication, dataset, software, etc.), title, date, creator(s)/author(s), abstract, access rights (i.e., open, embargoed, restricted or closed), license and DOI. Zenodo's metadata is compliant with DataCite Metadata Schema (<https://schema.datacite.org/>) minimum and recommended terms. Other repositories may have different requirements of minimum metadata, which will be complied with at the time of each deposit. Besides the minimum required metadata of the selected repository, the BIOSYSMO partners will add additional metadata such as project acronym and number, keywords, author ID (e.g., ORCID), affiliation, etc.

Data shall include standard vocabularies for all data types present in the datasets wherever possible. In case of using uncommon or project-specific ontologies/vocabularies is unavoidable, mappings to more commonly used ontologies shall be provided. When existing and relevant, standard vocabularies and ontologies will be followed such as those of Minimum Information for Biological and Biomedical Investigations (MIBBI), an initiative that develops guidelines of minimum information for various biological disciplines. Some relevant standard vocabularies and ontologies or guidelines used and endorsed by recognized repositories and databases are listed in Table 2.

⁴ GitHub is an Internet hosting service for software development and version control using Git; although not a repository perse, it is included here because the possibility of linking items with Zenodo makes it a useful resource for archiving, versioning and sharing software code

Table 2. Relevant standard vocabularies and ontologies or guidelines

Standard	Relevant for	Example databases
Gene ontology (GO)	Genes and gene product attributes	ENA, NCBI GenBank, NCBI GEO, GOLD, DDBJ
Minimum Information about any (x) Sequence (MIxS)	Any sequence	ENA, NCBI GenBank, NCBI GEO, GOLD, DDBJ
Minimum Information about a (Meta)Genome Sequence (MIxS - MIGS/MIMS)	(Meta)Genome sequence	ENA, NCBI GenBank
Minimum Information about a high-throughput SEQuencing Experiment (MINSEQE)	High-throughput sequencing experiment	ENA, NCBI GenBank
Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE)	qPCR and RT-qPCR	ENA, NCBI GenBank
Minimum Information About a Microarray Experiment (MIAME)	Microarray experiments	NCBI GEO
Minimum Information Required in the Annotation of Models (MIRIAM)	Biological models	BioModels Database
Synthetic Biology Open Language (SBOL)	Genetic parts and plasmids	ENA, NCBI GenBank, NCBI GEO
Just Enough Results Model (JERM)	Microbiology, molecular biology, biomedicine	Any that requires MIBBI guidelines
Chemical Methods Ontology (CMO)	Data in chemical experiments	-
Chemical Entities of Biological Interest (ChEBI)	Chemicals	PubChem, ChEBI
extended-ILCD (eILCD) Data Format	LCA elements and parameters	European Platform on Life Cycle Assessment (EPLCA)
European NORMAN network guidelines	Pollutant chemical analyses	NORMAN network database
ISO guidelines	Chemical analysis	-
OECD guidelines	(Eco)toxicology tests	-

In addition, existing standard ontologies and nomenclatures will be used as much as possible for every concept/measurement in the datasets, e.g., Enzyme Commission Number (EC Number), IUPAC International Chemical Identifier (InChI or InChIKey), PubChem Chemical ID (CID), globally unique and persistent identifiers given by the UniProt database, NCBI accession numbers, etc.

Finally, all outputs and in particular publications resulting from the BIOSYSMO project will acknowledge the financial support by EU using the statement: *“The project BIOSYSMO has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101060211”*, which will facilitate the aggregation of all publications by the OpenAIRE infrastructure.

3.2 Making data accessible

The second element of the FAIR Guiding Principles aims to enabling the access to the data. Repositories not only facilitate the findability of data but also their accessibility. First, the selected repositories provide access to data and metadata over standard protocols such as HTTP and OAI-PMH, which are free and standardized and thus globally implementable to facilitate data retrieval. Anyone with a computer and an internet connection will be able to access at least the metadata. Second, the metadata accompanying the data will provide information on the access conditions to the data in case it is not openly accessible (e.g., in case authentication, authorization or personal request will be needed).

The consortium has agreed to use Zenodo—a general-purpose open repository developed under the European OpenAIRE program and operated by CERN—as the central trusted repository to deposit data generated in the project. A project community was created to collect all the deposited outputs related to BIOSYSMO (<https://zenodo.org/communities/biosysmo/>). Nevertheless, partners may deposit their data in other repositories such as institutional or subject-based. In fact, the accessibility to some types of data (e.g., genomic data) is greatly improved by the deposit in specific consolidated thematic repositories in the field (e.g., ENA, GenBank; Table 1). Depositing certain data in dedicated repositories also facilitates the automatic mining of data. Regardless of the additional repositories used, all BIOSYSMO datasets must also be uploaded or linked to Zenodo to ensure comprehensive data management.

At the time of deposit, datasets can be assigned open, restricted, or embargoed access. For instance, an embargo may apply until the associated publication is accepted. For research data that has not yet been published, Zenodo offers a restrictive mode, limiting access to IDENER (as BIOSYSMO community owner) and the corresponding partner. While partners can decide whether to upload data in restricted mode before publication or open mode post-publication, early uploads are strongly encouraged to ensure proper data handling. Zenodo provides strong protection, including:

- Ownership: Uploading data does not transfer property rights, all content remains the property of the submitting parties.
- Publishing support: Research materials for peer review can be uploaded in restricted records, with protected links shared with reviewers. Content can also be embargoed and automatically opened once the associated publication is released.
- Integration: Zenodo provides a rich API, enabling integration with external workflows and third-party tools.

If certain datasets cannot be shared (or need to be shared under restrictions), the partner will explain the motivation, clearly separating legal and contractual reasons from voluntary restrictions. The

reasons for not making data openly available will comply with the Grant Agreement. In case that the access to specific data will be restricted (e.g., for legal reasons), the metadata will state the contact protocol to request such access and will indicate an email of a contact person who can discuss access to the data. A dataset index is included in Annex V: Dataset index and is updated regularly. This index includes the following information: PID of the data, type of PID, description of dataset, whether the dataset is available in open access (and if not, reason for restricted access) and URL to repository.

All data deposited in a repository will be provided with appropriate metadata. The metadata will be openly accessible and licensed under a public domain dedication CC0. Moreover, if appropriate, the data entry will be accompanied by a .txt file providing documentation or reference about any software needed to access or read the data or, if possible, the relevant open source code will be included. Although each partner is responsible for uploading their datasets and ensure that the data is properly annotated and well documented, as explained in Table 4, the Exploitation and Dissemination team of BIOSYSMO prepared a dataset upload protocol with clear guidance (Annex IV: Protocol for uploading datasets to Zenodo), and IDENER, as the project coordinator, will be responsible for reviewing and accepting datasets to be published in this community.

Depositing data in a repository also guarantees long term availability. Zenodo guarantees the retention of data and metadata during the lifetime of the repository, which is currently 20 years. Moreover, metadata is accessible, even when the data is no longer available.

The use of the selected repositories (Table 1) is open and free to any user; therefore, no special arrangements are needed to deposit the data, other than complying with the general terms of use of the particular repository. As indicated in Table 1, the selected repositories will assign an identifier to the data, which is resolved to a digital object.

3.3 Making data interoperable

The third element of the FAIR Guiding Principles aims to improve data interoperability. Data interoperability can be greatly improved by the use of standard vocabularies, ontologies or methodologies as well as using standard file formats for data preservation. Section 3.1 already mentioned provisions to be followed such as the use of standard vocabularies in the field and minimum information schemas (Table 2). In addition, when existing, relevant and possible, the datasets will provide qualified references to other data. Specifically, references will be provided when a dataset builds on another dataset from the project or previous research, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset. Datasets will be referred to by including their unique and persistent identifiers.

In order to facilitate interoperability and long-term preservation of data, the partners will choose and favor file formats with the following characteristics: non-proprietary; open, documented standards; in common usage by the research community; using standard character encodings (ASCII, UTF-8); unencrypted; and uncompressed. Table 3 is a non-exhaustive compilation of preferred file formats that will be used as a reference by the consortium partners to define the most appropriate file format. If the data file is in a non-preferred format, migration to the preferred file format will be recommended, while preserving a copy in the original file format.

Table 3. Recommended and acceptable data formats for sharing, re-use and long-term preservation⁴

Type of data	Recommended formats	Acceptable formats
<i>Tabular data with extensive metadata</i> - variable labels, code labels, and defined missing values	- SPSS portable format (.por) - delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) - structured text or mark-up file of metadata information, e.g., DDI XML file	- proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)
<i>Tabular data with minimal metadata</i> - column headings, variable names	- comma-separated values (.csv) - tab-delimited file (.tab) - delimited text with SQL data definition statements	- delimited text (.txt) with characters not present in data used as delimiters - widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)
<i>Textual data</i>	- Rich Text Format (.rtf) - plain text, ASCII (.txt) - eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	- Hypertext Mark-up Language (.html) - widely-used formats: MS Word (.doc/.docx) - some software-specific formats: NUD*IST, NVivo and ATLAS.ti
<i>Documentation and scripts</i>	- Rich Text Format (.rtf) - PDF/UA, PDF/A or PDF (.pdf) - XHTML or HTML (.xhtml, .htm) - OpenDocument Text (.odt)	- plain text (.txt) - widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) - XML marked-up text (.xml) according to an appropriate DTD or schema, e.g., XHTML 1.0 - Jupyter Notebook (.ipynb) for interactive documentation and computational workflows
<i>Image data</i>	- TIFF 6.0 uncompressed (.tif)	- JPEG (.jpeg, .jpg, .jp2) if original created in this format - GIF (.gif) - TIFF other versions (.tif, .tiff) - RAW image format (.raw) - Photoshop files (.psd) - BMP (.bmp) - PNG (.png) - Adobe Portable Document Format (PDF/A, PDF) (.pdf)
<i>Audio data</i>	- Free Lossless Audio Codec (FLAC) (.flac)	- MPEG-1 Audio Layer 3 (.mp3) if original created in this format - Audio Interchange File Format (.aif) - Waveform Audio Format (.wav)
<i>Video data</i>	- MPEG-4 (.mp4) - OGG video (.ogv, .ogg) - motion JPEG 2000 (.mj2)	- AVCHD video (.avchd)
<i>Geospatial data vector and raster data</i>	- ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx, .sbn optional) - geo-referenced TIFF (.tif, .tfw)	- ESRI Geodatabase format (.mdb) - MapInfo Interchange Format (.mif) for vector data

- CAD data (.dwg) - tabular GIS attribute data - Geography Markup Language (.gml)	- Keyhole Mark-up Language (.kml) - Adobe Illustrator (.ai), CAD data (.dxf or .svg) - binary formats of GIS and CAD packages
---	---

3.4 Increasing data re-use

The last principle of FAIR data focuses on actions intended to facilitate the re-use of data either inside or outside the project or by third parties, particularly after the end of the project.

The partners will define the possible licensing terms and restrictions for each dataset, particularly for third party use. During the project, the data shall be made accessible to the consortium partners under the terms of the Grant Agreement and the Consortium Agreement. All deposited data will be assigned a clear and accessible data usage license, in line with the obligations set out in the Grant Agreement. License is one of the mandatory terms in Zenodo's metadata and is referring to an Open Definition license. Data downloaded by the users is subject to the license specified in the metadata by the uploader.

The data will include readme files in order to record the provenance, methodology of data collection and analysis, so as to facilitate data re-use. In addition, the metadata will make reference to any code of software needed to access or possibly re-use the data. Metadata will always describe the original authors of the published work.

Quality assurance processes

To ensure high-quality data, strict quality assurance processes are implemented. These include:

- Proper training of all personnel involved in testing for data quality management.
- Evaluation of the experimental protocols used or to be used, reproducibility, the tool's source code and abilities, the metadata completeness, and the presence of potential gaps in datasets. This is especially significant in the case of data reuse originating from pre-existing datasets that need to be imported into the BIOSYSMO repository.
- Experimentation conducted using well-established and validated procedures to ensure high quality, robust and reliable data (e.g., the use of protocols published in peer-reviewed journals).
- Proper equipment calibration and verification practices, including equipment-provided quality assessment.
- Data analyzed by the partner generating/collecting the data in order to identify any issues such as missing values, anomalies, or inconsistencies.
- Each experiment will be characterized in terms of accuracy and precision of the significant data measured during the test. Inconsistent, inaccurate and/or missing data will be managed by repeating the experiment.
- No data without information on consistency will be considered for technical purposes and/or modelling.
- Used of controlled vocabulary.

- Processes of data validation (upon entry and a posteriori) and data cleaning (e.g., removing outliers, handle with missing values).
- All the processes concerning data integrity will be documented and available to all users of the project.

Templates will be used for project documentation, and an internal peer review is performed for the main project deliverables to guarantee the deliverable is developed with a high level of quality. Each WP leader has to submit all the produced documents to another partner assigned as internal reviewer as well as to the project coordinator to check for the quality of the documents produced. The process will be reviewed on a timely basis at biannual BIOSYSMO General Assemblies to ensure ongoing suitability.

4 Other research outputs

Unless confidentiality issues are required, otherwise, other research outputs such as presentations, workflows, analysis scripts, etc. will be deposited in Zenodo or other appropriate repositories. Scientific peer-reviewed publications will be deposited in Zenodo and will be made openly available according to the open access mandate. A specific procedure for the management of publications according to the open access mandate is outlined in the deliverable *D8.1 – Project Management Handbook*.

All the deposited items will follow the FAIR Principles Guidelines described in Section 3. All outputs and in particular publications resulting from the BIOSYSMO project will acknowledge the financial support by EU using the statement: “*The project BIOSYSMO has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101060211*”, which will facilitate the aggregation of all publications by the OpenAIRE infrastructure.

5 Allocation of resources

IDENER R&D is the partner responsible for supervising the data management and updating the DMP within the BIOSYSMO project. Nonetheless, each partner of the BIOSYSMO project is responsible for ensuring that their own data is handled according to the FAIR principles detailed in the DMP. Each partner is responsible for the quality control and safe storage of their own data. Each partner will be responsible for the identification of new data or datasets that will be handled within their own tasks, clearly describing the new data according to Section 2 and reporting them to IDENER R&D in order to update the DMP. IDENER R&D will be responsible for updating this new information in the DMP.

The following Table 4 summarizes the distribution of responsibilities within the BIOSYSMO project.

Table 4. Distribution of responsibilities regarding the management of research data

Action	Responsible partner
Oversee of data management	IDENER R&D
Update of DMP	IDENER R&D
FAIR data management	Data owner(s)
Quality control and safe storage of data	Data owner(s)

Report any new data or changes to existing data	Data owner(s)
Deposit of data in the BIOSYSMO community in Zenodo	Data owner(s)
Review and approval of data uploaded to the BIOSYSMO community in Zenodo	IDENER R&D
Upload of data to other repositories (institutional, thematic, etc.)	Data owner(s)

The consortium has agreed to use Zenodo as a central repository for the project data (without precluding the further use of other repositories, e.g., institutional, thematic; Table 1). Zenodo and other chosen thematic repositories are free to use, while institutional repositories are maintained through general institution's budgets, making them free to users. Nevertheless, if the selected repositories do not provide the desired storage length, additional resources for long-term preservation of data will be evaluated based on the type and the size of the data requiring long-term preservation.

Each partner will be responsible for depositing their own data in the BIOSYSMO community in Zenodo (<https://zenodo.org/communities/biosysmo/>), while IDENER R&D will be in charge of reviewing and approving the uploaded datasets.

6 Data security

Each consortium partner will be responsible for maintaining the confidentiality of their own datasets and storing them securely in accordance with internal standards. To guarantee the backup operations, the datasets will be kept in digital format and knowledgeable IT system administrators will handle each organization's storage systems. As a general rule, the partners will limit the usage of USB flash drives, the data will be periodically backed up to prevent data loss, and the files will be systematically labelled to ensure consistency of the final datasets.

7 Ethics

Ethical issues

No ethical issues that can have an impact on data sharing have been identified for now. These issues will be monitored during the execution of the project and, if needed, taken into account in the next versions of the DMP.

Legal issues

Some data may be subject to sharing restrictions due to legal issues. Some of the identified issues are:

- Sharing of data collected at the sites managed by TAUW is restricted due to contractual agreements between TAUW and its clients. Those data are shared with the consortium within the contractual limits signed with its clients. Sharing of these data outside the consortium (e.g., for dissemination activities) will only be allowed upon explicit authorization of the involved client.
- Data available in public databases are usually shared under specific licenses chosen by the data owner. The use and sharing of these data are subject to the fulfilment of those licenses. For example, data available under a license CC BY-NC-SA (Attribution-Non Commercial-Share Alike) can only be shared under the same license.

8 Other issues

Other national, funder, sectorial or departmental procedures for data management

Besides the procedures for data management described in the BIOSYSMO DMP, the following partners will additionally follow other national, funder, sectorial or departmental procedures for data management as follows:

- **JSI:** Internal rules of data management are adapted d by Fair by design from Unlock project (Wageningen University and Research, TUDelft, Dutch Research Council). Briefly, the metadata schema is based on ISA Standard which is a combination of the Just Enough Results Model (JERM) and the Minimum Information About a Plant Phenotyping Experiment (MIAPPE). Metadata is generated using a template and validation is performed using same template to fit the requirements. Some templates for generating metadata are already available, but it can be customized, depending on the type of experiments and data collection. The format is standardized and obligatory to ensure machine-actionability.
- **UMLP (formerly UBFC):** Raw data are stored onto personal laptops, and automatically backed-up onto the Chrono-environment data center. Back-up are made daily through the "ATEMPO Live Navigator". Storage of data for further sharing will be done on secure laboratory servers (backup, secure access (DMZ)). The data will be made accessible through the OSU THETA search metadata portal, once deposited in the established Data-Chrono repository. Terms of use of Data-chrono are available at <https://data-chrono.univ-fcomte.fr/web/app.php/credits>.
- **UPM-CBGP:** the CBGP has implemented a data and knowledge asset management system (SEEK, from the FAIR-DOM project) which is being maintained by Dr. Wilkinson (CBGP), one of the founders of FAIR principles (<https://seek4science.org>). SEEK allows tracking and versioning of datasets, their use by research teams, and the outputs of analytical operations or workflows, throughout the research lifecycle. The SEEK installation is running at CBGP Cluster and can be used by CBGP researchers. Metadata will become part of the data deposition using the RO-Crate specification (<https://www.researchobject.org/ro-crate/>), a FAIR-enabled data archival standard that facilitates "virtual aggregation" of datasets dispersed over the Internet and provides a means of explicitly associating detailed domain metadata with its data record. In addition, publication of formal data descriptors in journals such as Nature Scientific Data or GigaScience will be encouraged.
- **ICL:** Additional guidance on data management will be taken from the ICL training available on Research Data Management Plans.

9 Available resources on the BIOSYSMO collection

9.1 Open data

As the BIOSYSMO project progresses, a considerable amount of research data has already been generated and made available through open-access repositories, ensuring compliance with FAIR principles. These datasets, publications, and deliverables are key outputs that contribute to the project's objectives and broader scientific community. Annex VI: Overview of data available in the BIOSYSMO collection for Zenodo provides a structured overview of the data already available in the BIOSYSMO collection within Zenodo. Which includes:

- Communication and Dissemination Data (C&D data): Newsletters and public outreach materials that facilitate engagement with stakeholders.
- Reference Data: Project deliverables, such as reports on microbial consortia, bioremediation strategies, genetically enhanced microorganisms, and other critical findings.
- Experimental Data: Datasets derived from laboratory and field experiments, including genomic and metagenomic sequencing and microbial studies.
- Scientific Publications: Peer-reviewed research articles and book chapters that contribute to the wider bioremediation and environmental science fields.

9.2 Restricted-access datasets

In addition to these datasets, two derived/compiled datasets have been generated and are currently uploaded to Zenodo under restricted access mode.

To maintain the highest standards of data integrity, all datasets are sourced from scientific literature databases come from peer-reviewed and trusted sources. The data management process includes regular internal reviews to verify accuracy and ensure alignment with the project's objectives. Before use, all databases undergo careful evaluation to confirm their suitability. While most of these original databases are publicly available under CC BY 4.0 licenses, IDENER has proactively sought explicit authorization from data providers to ensure transparency, validate their use, and inform them about the BIOSYSMO project. The licensing terms set by the original authors are strictly respected, with proper citation and attribution in all related outputs, even when the connection is indirect.

Once the necessary permissions are obtained, the datasets will be openly accessible within the BIOSYSMO collection, reinforcing the project's commitment to open science. However, they will remain under restricted access until a potential publication is finalized to ensure proper dissemination and scientific impact. Despite their current status, these datasets are already playing a crucial role in supporting computational modeling efforts and will continue to evolve as additional data is integrated.

- Proteins Involved in Metal Resistance: This dataset consolidates curated protein data related to metal resistance in bacteria, fungi, and plants, sourced from publicly available repositories such as BacMet, UniProt, and PlantPres. It provides key insights into resistance mechanisms, which are essential for understanding microbial and plant interactions in contaminated environments.
- BIOSYSMODb – Integrated Biodegradation Database. BIOSYSMODb is an integrative resource centralizing data on metabolic pathways, reactions, enzymes, and degradative organisms to support soil bioremediation strategies. It aggregates data from major repositories such as EAWAG BBD, MibPOPdb, MetaCyc, KEGG, and UniProt, enabling connections between computational predictions and experimental validation. The dataset also includes a Jupyter Notebook to facilitate structured data queries.

At this stage, BIOSYSMODb only includes compiled and derived data from public sources. Current work is devoted to the incorporation of experimental and simulation data generated within the project, progressively refining its predictive capabilities and validation potential. This database is designed to:

- Harmonize and unify data from major public repositories, ensuring consistency and interoperability.

- Expand with experimental and simulation data, progressively enhancing its predictive capabilities.
- Serve as a central resource for computational modeling in the development of optimized bioremediation strategies.

9.3 Future focus

As new data continues to be generated, Zenodo will be updated to ensure comprehensive coverage of all project outputs. Future updates will focus on the continuous inclusion of research advancements by incorporating:

- New scientific publications emerging from BIOSYSMO's research activities.
- Additional datasets from computational models, field experiments, and laboratory studies.
- Updated project deliverables, ensuring accessibility to key findings and methodologies.
- Expanded Communication & Dissemination materials, supporting outreach and stakeholder engagement.

By progressively integrating these elements, the BIOSYSMO community will evolve into a comprehensive, open-access reference for the scientific community, policymakers, and industry stakeholders involved in bioremediation and environmental management.

10References

- (1) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, 3 (1), 160018. <https://doi.org/10.1038/sdata.2016.18>.
- (2) Bongaerts, N.; Della Chiesa, S. *Research Data Management Lifecycle*; Zenodo, 2022. <https://doi.org/10.5281/zenodo.6602006>.
- (3) *FAIR Principles*. GO FAIR. <https://www.go-fair.org/fair-principles/> (accessed 2022-10-26).
- (4) *Data formats for preservation*. OpenAIRE. <https://www.openaire.eu/data-formats-for-preservation/> (accessed 2022-12-28).

Annex I: Data management questionnaire

Version 1.0 of the data management questionnaire



BIOremediation systems exploiting SYnergieS
for improved removal of Mixed pOllutants

Questionnaire of Data Management

Questionnaire information

Partner	
Date	
Responsible person for data management in your group (name and email)	

1 Description of the data

Please describe the data that you will generate or re-use during the project. For each WP/task provide the following information using the format below:

- **Purpose** of the data generation or re-use and its relation to the objectives of the project.
- **Types of data**, according to the following categories: observational field data (e.g., sensor readings, telemetry, survey results, images); experimental data collected under controlled conditions (e.g., gene sequences, chromatograms, biochemical measurements); simulation data from test models (e.g., climate models, economic models, biological models); derived/compiled data generated from existing datasets (e.g., text/data mining, compiled database); reference (peer-reviewed) datasets (e.g., gene sequence databanks, chemical structures, spatial data portals).
- **Form of data**, according to the following categories: text, numeric, images, audiovisual, models, computer code, discipline-specific (e.g., geospatial data) and instrument-specific (e.g., spectroscopy data).
- **Origin/provenance** of the data, either generated or re-used. If you will re-use any existing data, indicate the source and what you will re-use it for.
- To whom might your data be useful ('**data utility**') outside the project.

- **WP/Task:** Task xxx....
- **Brief description of data:** ...
- **Purpose of the data:** ...
- **Types of data:** ...
- **Form of data:** ...
- **Origin of the data:** ...
- **Data utility outside the project:** ...

- **WP/Task:** Task yyy....
- **Brief description of data:** ...
- **Purpose of the data:** ...
- **Types of data:** ...
- **Form of data:** ...
- **Origin of the data:** ...
- **Data utility outside the project:** ...

- **WP/Task:** Task zzz....
- **Brief description of data:** ...
- **Purpose of the data:** ...
- **Types of data:** ...
- **Form of data:** ...
- **Origin of the data:** ...
- **Data utility outside the project:** ...

2 Questions about FAIR data

List any repository (with link) that you may use to deposit your data (e.g., institutional repository or relevant subject-specific repositories such as gene sequence repositories).

List any data and metadata vocabularies, standards, formats or methodologies relevant in your field and that you will follow to make your data interoperable. Examples: metadata requirements of sequence repositories, Minimum Information for Biological and Biomedical Investigations (MIBBI), Gene Ontology, ISA-Tab schema, relevant ISO standards, etc.

Describe the metadata that may be created for your data (e.g., according to the previously listed standards).

Describe your data quality assurance processes.

3 Other questions

Will you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

Do you foresee any costs for making data or other research outputs FAIR in your project (e.g., direct and indirect costs related to storage, archiving, re-use, security, etc.)? How will these be covered?

Annex II: Data summary

This annex describes the data that will be generated, collected, processed, analyzed and re-used in the BIOSYSMO project. The data is described in terms of the following characteristics:

- **Partner:** partner generating, collecting, processing, analyzing and/or re-using the data
- **Task:** task where the data is generated.
- Brief **description** of data
- **Purpose of the data** generation or re-use and its relation to the objectives of the project.
- **Types of data**, according to the following categories: observational field data (e.g., sensor readings, telemetry, survey results, images); experimental data collected under controlled conditions (e.g., gene sequences, chromatograms, biochemical measurements); simulation data from test models (e.g., climate models, economic models, biological models); derived/compiled data generated from existing datasets (e.g., text/data mining, compiled database); reference (peer-reviewed) datasets (e.g., gene sequence databanks, chemical structures, spatial data portals).
- **Form of data**, according to the following categories: text, numeric, images, audiovisual, models, computer code, discipline-specific (e.g., geospatial data) and instrument-specific (e.g., spectroscopy data).
- **Origin/provenance of the data**, either generated or re-used. If you will re-use any existing data, indicate the source and what you will re-use it for.
- **Data utility** outside the project: to whom might the data be useful outside the project.

WP1. Field sampling and characterization of polluted and treated samples

Partner: UBU

- **Task:** Task 1.4
- **Brief description of data:** (Eco)toxicity-based pollution assessment samples before and after bioremediation.
- **Purpose of the data:** To assess the toxic effect of bioremediation technologies in the prokaryotic and eukaryotic organisms.
- **Types of data:** Derived data from experimental data or literature data.
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** From field trials and experiments conducted to track bioremediation
- **Data utility outside the project:** Hazard and risk assessment

Partner: JSI

- **Task:** Task 1.2
- **Brief description of data:** Data of samples analyzed for the pollutant degradation.
- **Purpose of the data:** To prepare the metadata of all the samples analyzed by CNRS.
- **Types of data:** The data will be discipline-specific for sample characteristics and instrument specific for chromatograms
- **Form of data:** numeric data
- **Origin of the data:** Generated (analytical equipment).
- **Data utility outside the project:** Future proposals and publications.

Partner: JSI

- **Task:** Task 1.3
- **Brief description of data:** Data generated by collection of the cells from the samples, cell aggregation and microfluidic isolation of pollutant degrading microorganisms (spectrophotometric data, flow cytometry data, microscopic imaging).
- **Purpose of the data:** To establish the efficiency of cell aggregation, microbial isolation, cell viability, cell growth, metabolic activity of the cells.
- **Types of data:** Experimental data (numeric data from spectrophotometry, zeta-potential reader, flow cytometry and image data from microscopy and flow cytometry).
- **Form of data:** numeric and image data
- **Origin of the data:** Data is generated from experiments at the JSI facilities
- **Data utility outside the project:** Future projects and publications.

Partner: LEITAT

- **Task:** Task 1.2
- **Brief description of data:** Physicochemical characterization of samples from polluted sites
- **Purpose of the data:** To characterize polluted samples and to study the degradation pathways of the target pollutants, biotransformation intermediates and final products.
- **Types of data:** Experimental data collected under controlled conditions.
- **Form of data:** Instrument-specific.
- **Origin of the data:** Generated.
- **Data utility outside the project:** This data will be useful for further comparison studies in other research projects

Partner: LEITAT

- **Task:** Task 1.4
- **Brief description of data:** (Eco)toxicity data of the samples before and after the application of the remediation technologies (EC50, NOEC/LOEC, dose-response curve)
- **Purpose of the data:** Ecotoxicological characterization of the samples before and after remediation.
- **Types of data:** Experimental data collected under controlled conditions.
- **Form of data:** Numeric and graphs.
- **Origin of the data:** Generated.
- **Data utility outside the project:** This data will be useful to the remediation technology developers to characterize the ecotoxicity of their samples and be able to compare samples before and after treatment. And, for the developing of QSAR models.

Partner: UMLP, formerly UBFC

- **Task:** Task 1.1 & Task 1.2
- **Brief description of data:** Soil analyses performed at the BIOSYSMO French sites 1, 2 and 10.
- **Purpose of the data:** Physicochemical characteristics of the studied soils for better understanding of the pollutant fate during bioremediation and re-use in Tasks 1.3 and 1.4
- **Types of data:** The data will be acquired during laboratory and field experiments (chemical measurements)
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** newly generated data from sites 1, 2 and 10 managed by the UBFC partner.
- **Data utility outside the project:** stakeholders in charge of polluted site management and polluted wider soil DB

Partner: UMLP, formerly UBFC

- **Task:** Task 1.3
- **Brief description of data:** Microbial analyses performed at the BIOSYSMO French sites.
- **Purpose of the data:** The molecular data will be included in an integrative database (task 2.1) and used in assisting for the selection of the most promising samples to be used to establish a collection of bacterial and fungal isolates/consortia
- **Types of data:** The experimental data will be collected under controlled conditions (e.g., gene sequences, biochemical measurements)
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.

- **Origin of the data:** newly generated data from sites 1, 2 and 10 managed by the UMLP (formerly UBFC) partner. Data will be re-used in Tasks 2.1 and 3.1
- **Data utility outside the project:** scientists

Partner: CIIMAR

- **Task:** Task 1.1
- **Brief description of data:** Data generated from the collection of sediments and plants that will be performed during the sampling campaigns.
- **Purpose of the data:** Collect and provide samples (non-vegetated sediments, rhizosediments, plant shoots and plant roots) throughout the project for physicochemical characterization, pollutants levels, microbial analysis, and microbial isolation.
- **Types of data:** Observational field data (images) and sample characterization: sediments and plants, for posterior analysis (sample identification)
- **Form of data:** text, numeric and images.
- **Origin of the data:** generated.
- **Data utility outside the project:** data can be used by researchers and environmental agencies to evaluate the presence or patterns of contaminants in the environment.

Partner: CIIMAR

- **Task:** Task 1.2
- **Brief description of data:** Physicochemical data such as water content for sediments and plants, organic matter and grain size analysis for sediments present on selected sites. Protocols for pollutants characterization in sediments and plant tissues as well as data from that characterization: atomic absorption spectrometry data for metals and liquid chromatography data for pharmaceutical compounds.
- **Purpose of the data:** Cross-validation of sample preparation and analytical methods that will be applied for routine physicochemical and pollutants analysis. Development/implementation of analytical techniques: atomic absorption spectrometry for the analysis of metals and liquid chromatography for the analysis of pharmaceuticals. Analysis of the characteristics and pollutants present in the selected sites is essential for the development of a solution for their removal and recovery of contaminated sites.
- **Types of data:** Experimental data from water content, organic matter and grain size analysis and instrument-specific data from atomic absorption spectrometry and liquid chromatography analysis.
- **Form of data:** text (e.g., DOC, XLS), numeric, instrument-specific data (e.g., chromatograms).
- **Origin of the data:** generated and re-use (the protocols for water content, organic matter, grain size and atomic absorption spectrometry analysis of metals are already established at CIIMAR).
- **Data utility outside the project:** Contaminant analysis can be used by policymakers to implement laws on environmental discharges, the usage of compounds, and contaminants concentrations in environmental matrixes.

Partner: CIIMAR

- **Task:** Task 1.3
- **Brief description of data:** Molecular identification (16S and ITS) data from the microbial communities present on the sediment of contaminated sites. Data on the microorganisms (16S) that can tolerate and degrade the contaminants present on site selected for laboratory experiments. Protocols for enrichment, selection and isolation of microorganisms capable of bioremediation of metals and pharmaceuticals.
- **Purpose of the data:** Characterization (diversity and function) of microbial communities (bacteria and/or fungi) associated with the sediments (non-vegetated and rhizosphere sediments) and plant roots (hemisphere and endosphere), present on the contaminated sites. The molecular identification (16S) data of the microorganisms obtained from the enrichment in the presence of contaminated matrices will be included in the integrated database and used for the selection of the most promising microorganisms to be used to establish efficient consortia. The consortia and isolated strains with the highest efficiency in the degradation of selected pollutants will be analyzed via genomic and metagenomic sequencing.
- **Types of data:** Experimental data such as gene sequences, molecular identifications, and alpha/beta diversity.
- **Form of data:** text (e.g., DOC, XLS), instrument-specific bioinformatics data (e.g., FASTQ Sequence and Sequence Quality Format).
- **Origin of the data:** re-used from the CIIMAR Microbial Culture Collection, generated in the project and/or cross-referenced with re-use data from public database
- **Data utility outside the project:** The scientific community can use the information regarding the organisms that can degrade different pollutants to also create solutions for the removal of contaminants in other polluted sites.

Partner: CNRS

- **Task:** Task 1.2
- **Brief description of data:** (1) Collection of the sample preparation and analytical methods applied by different partners for routine analysis of chemical pollutants. (2) Pollutant concentrations in the different sites (3) Chromatograms and mass spectra of degradation intermediates and metabolites
- **Purpose of the data:** (1) Exchange of analytical protocols and identify potential bottlenecks (2) Chemical characterization of the different sites (3) Identification of degradation intermediates and metabolites.
- **Types of data:** (1) Experimental data and instrument-specific data collected under controlled conditions (GC-HRMS, LC-(HR) MS data files) (2) Compiled data generated from existing datasets.
- **Form of data:** Text and numeric (Word and Excel) and instrument-specific (from chromatographic and mass spectrometer software)
- **Origin of the data:** generated data and re-used data from partners of the project to collect and compare the results of different chemical analysis.
- **Data utility outside the project:** Concentrations of pollutants may be useful to all interested stakeholders dealing with emerging substances – whether in studying their occurrence and effects or risk assessment and risk management (Competent authorities / Reference laboratories, government

institutions, standardization bodies). Chromatograms and mass spectra could be re-used to implement collaborative databases

WP2. Computational tools for biosystems design and implementation

Partner: IDENER

- **Task:** Task 2.1
- **Brief description of data:** This database collects chemical compounds extracted from some of the most extended regulations for safety and health. Among others, this data set includes xenobiotic compounds with significance for the BIOSYSMO propose.
- **Purpose of the data:** The purpose for the generation of this database is to provide significant chemical space diversity so that the computational tools for biodegradation developed in the frame of this project may be tested. Furthermore, this large chemical space would be clustered and grouped based on structural analogy or functional group correspondence. We will extract common metabolic pathways from the resultant groups, to degrade the most chemical space possible.
- **Types of data:** Quantitative, qualitative and ontology descriptors compiled from existing datasets. In addition, codes associated with various agencies and databases were collected for purposes of identification.
- **Form of data:** Most of data is either numeric or text related to the physicochemical properties of each compound. Text data for ontology descriptors and identification coding (EC number, PubChem ID, InChI, smiles etc.).
- **Origin of the data:** Ontology Data obtained from [Classyfire](#), a web-based application for automated structural classification of chemical entities; Quantitative and qualitative chemical descriptors obtained from [PubChem](#), the world's largest collection of freely accessible chemical information; [NORMAN](#) Network Priority Substance List; [POP](#) Stockholm Convention on Organic Pollutants; [UBAPMT](#) Potential Persistent, Mobile and Toxic (PMT) substances by the German Environment Agency, UBA, Germany; [SVHC](#) Candidate List of substances of very high concern for Authorization; PBT/vPvB: ECHA's persistence, bioaccumulation and toxicity (PBT) assessment [PBT](#); [WFD](#): Water Framework Directive list of priority substances; [NORMAN-SUSDAT](#), a "living database" compiling information provided by NORMAN network members and external contributors (NORMAN-SLE).
- **Data utility outside the project:** The collection, curation, and unification of compounds with negative impacts on human health and the environment will allow the creation of this complete chemical space database which may be exploited for many purposes. Following the Chemicals Strategy for Sustainability (CSS) (EC, 2020a), it will serve as a useful tool for the identification of chemicals substances to avoid. Mainly, it will allow the development of computational novel approaches to analyze chemical properties of promising less hazardous substitutes and compare its analogy with compounds within the database. Also, it will contribute to developing data-driven computational ecotoxicity models (LCA tools and QSAR methods).

Partner: IDENER

- **Task:** Task 2.1
- **Brief description of data:** The data used in this project task aims to create an integrated database that combines the maximum amount of information available on compound degradation pathways.

This mainly includes structures of initial degradation compounds; reactions involved in these pathways; substrates, products, and intermediates of these reactions; organisms capable of carrying out each pathway and genetic sequences regarding enzymes responsible for the reactions involved in each pathway.

- **Purpose of the data:** The objective of this data collection is the creation of an integrative database containing genetic and metabolic information on traits of interest obtained from publicly available databases in already characterized bacteria, plants, and fungi. During this process, several public databases have been consulted in order to extract the required information in an orderly and concise manner. In technical terms, data mining techniques have been used to allow the relative automation of the process. All collected information will be clustered and integrated into a cross-referenced SQL database.
- **Types of data:** The obtained data will be mainly derived/compiled data generated from existing datasets such as EAWAG, MetaCyc, Brenda, MibPOP or BKMs. Reference datasets and repositories especially for genome data (UniProt, NCBI Gene, NCBI Nucleotide and NCBI Genome) will be also survey.
- **Form of data:** Data format will be as text (tables in CSV format), numeric (tables, lists, IDs), discipline-specific (fasta and gbk files for genetic data) or computer code (Python, Shell script and R) for the data mining tools that should be developed or modified. The integrated database will be in SQL format.
- **Origin of the data:** The integrated database is based on the unification of information on compound degradation from different metabolic pathway databases such as EAWAG (<http://eawag-bbd.ethz.ch/>), MetaCyc (<https://metacyc.org/pwy-search.shtml>), Brenda (<https://www.brenda-enzymes.org/index.php>), MibPOP (<http://mibpop.genome-mining.cn/index/>), BKMs (<https://bkms.brenda-enzymes.org/download.php>), MetaNEtX (<https://www.metanetx.org/mnxdoc/mnxref.html>) and EnviPath (<https://envipath.org/pathway>). For genome and enzyme sequence information the data will come mainly from UniProt (<https://www.uniprot.org/>) and NCBI (<https://www.ncbi.nlm.nih.gov/>).
- **Data utility outside the project:** The creation of a unified database on metabolic pathways with the capacity to degrade compounds is of great use not only for the subsequent stages of the project itself. It also constitutes a source of knowledge that saves resources and time due to its integrative character in the field of biodegradation of both toxic and non-toxic compounds.

Partner: IDENER

- **Task:** Task 2.2
- **Brief description of data:** This task will focus on the generation of Genomic Scale Models (GEM) to computationally analyze the intricate network of biochemical reactions that determine the overall metabolic performance of the microorganism. These models will contain gene-protein-reaction associations that are present within the metabolic pathways of microorganisms, and in some cases, microbial communities, for the degradation of organic or inorganic pollutants to their respective products. In the context of BIOSYSMO, these models will be designed according to the composition of each biosystem (e.g.: plant-microbe/microbial consortium), that could be employed for the effective removal of these harmful contaminants from the ecosystem.
- **Purpose of the data:** Overall, the models will be used for predicting the metabolic capacity of microorganisms or microbial communities to degrade specific organic and inorganic contaminants. They will provide quantitative information on key parameters such as microbial growth rate,

identification of key genes and deriving optimal metabolic pathways. Once genomic and metagenomic information is converted into biodegradation pathways, they can be used to enhance the potential of the biosystems. In turn, the data provided from the operation of the biosystems in WP3 and WP4 can then lead to the validation of the models resulting in an iterative two-way data exchange.

- **Types of data:** (1) Experimental Data: Gene sequences, biochemical measurement; (2) Simulation data: growth rate, yield, metabolic fluxes, gene essentiality; (3) Derived/Compiled from existing datasets and reference datasets: genomic sequences and annotation, metagenomic data.
- **Form of data:** The data will be present in the form of text, numeric and will also consist of computer codes to display models (File types: .xml, .doc, .csv; Model data: .m (MATLAB file), .py (Python code), .mat (MATLAB object))
- **Origin of the data:** All the data for this task will come through multiple sources. The models will be optimized by the dataset prepared under Task 2.1. The models will be prepared from genomic and metagenomic information extracted from publicly available databases. They will be optimized and validated based on the experimental data from the biosystems in WP3 and WP4.
- **Data utility outside the project:** Apart from bioremediation, the GEM can benefit other applications involving microorganisms. When these models are provided with the appropriate dataset of biochemical reactions, they can be used to predict growth and production yield of industrially relevant compounds, utilization, and conversion of harmful contaminants, as well as explore the tools of genetic modification in existing biosystems for improving process design.

Partner: IDENER

- **Task:** Task 2.3
- **Brief description of data:** The data involved in this task come from experimental tests from field studies and validation (Task 5.2) and the data obtained from the mathematical models developed, to feed the roadmap for wider implementations (Task 5.3). Moreover, some data will be used in the parameters of the mathematical models concerning the multidisciplinary optimization (physical, environmental, technological and costs parameters), which will come from scientific and engineering literature.
- **Purpose of the data:** The experimental data from Task 5.2, and other public literature information will help to develop and validate the mathematical models. In addition, the data resulted from the mathematical models will help to perform the roadmap for wider implementations (Task 5.3).
- **Types of data:** It will be data from existing databases and scientific literature, original experimental data obtained in Task 5.2, and compiled/modelled data from the mathematical models which will be sent to Task 5.3.
- **Form of data:** All data, experimental and modelled, will be numeric in the form of tables, databases, lists, counts or measurements.
- **Origin of the data:** The data fed to the mathematical models will be collected from public databases or scientific literature and from newly generated experiments from task 5.2.
- **Data utility outside the project:** The data obtained from this task will be useful to create and develop models that emulate complex biodegradation biosystems, and how they interact with different environments, pollutants, and types of pollution.

Partner: UBU

- **Task:** Task 2.3
- **Brief description of data:** Development of bioremediation process models.
- **Purpose of the data:** To find the optimum parameters subjected to pre-defined physical, environmental, technological and cos constraints.
- **Types of data:** Experimental and observational collected data through the instruments used in the research.
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** Previous studies by new technologies and based on publications.
- **Data utility outside the project:** Development of new technologies related to bioremediation strategies.

Partner: UMLP, formerly UBFC

- **Task:** Task 2.1 & Task 2.2
- **Brief description of data:** Genome and metagenome characterization for the development of the GEM models
- **Purpose of the data:** The characterization data will be included in the integrated database (task 2.1) and used in assisting for the selection of the most promising samples to be used to establish a collection of bacterial and fungal isolates/consortia
- **Types of data:** The simulation data will be collected and use for modelling and clustering enzymes
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** newly processed data. Data will be re-used in Task 3.1.
- **Data utility outside the project:** scientists

Partner: CIIMAR

- **Task:** Task 2.1
- **Brief description of data:** All information collected from previous tasks will be clustered and integrated into a cross-referenced database.
- **Types of data:** Derived or compiled data from existing and reference datasets
- **Form of data:** text (e.g., DOC, XLS), numeric, instrument-specific bioinformatics data (e.g., FASTQ Sequence and Sequence Quality Format).
- **Origin of the data:** re-used from the CIIMAR Microbial Culture Collection, generated in the project and/or cross-referenced with re-use data from public database

- **Data utility outside the project:** The database can help the scientific community in future bioremediation projects.

Partner: CIIMAR

- **Task:** Task 2.2.
- **Brief description of data:** Data from the genomic and metagenomic analyses of the consortia and biosystems developed in previous and next tasks to help with the definition of the scope of Genome-scale models.
- **Types of data:** Simulation data to create genomic-scale models.
- **Form of data:** text (e.g., DOC, XLS), numeric, instrument-specific bioinformatics data (e.g., FASTQ Sequence and Sequence Quality Format), genome-scale models.
- **Origin of the data:** re-used from the CIIMAR Microbial Culture Collection, generated in the project and/or cross-referenced with re-use data from public database
- **Data utility outside the project:** The models created to implement and assess more efficient bioremediation methods are a great support database for the scientific community.

WP3. Development and optimization of consortia and plant-microbe biosystems

Partner: UBU

- **Task:** Task 3.1
- **Brief description of data:** Characterizing microbial isolates that tolerate the target pollutants, and promote plant growth, to be used during phytoremediation strategies.
- **Purpose of the data:** To establish microbial isolates and communities with plant symbiotic traits to improve the phytoremediation of water.
- **Types of data:** Experimental and observational collected data through the instruments used in the research. Besides, derived data from experimental data or literature data.
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** Samples collected in field trials (Task 1.1), public collections, provided by partners.
- **Data utility outside the project:** To determine a microbial consortium useful to understand better the symbiosis between organisms and phytoremediation strategies.

Partner: UBU

- **Task:** Task 3.4
- **Brief description of data:** Validation of the microbial engineering through auxotrophy test, prior to application in the soil bioremediation strategies.

- **Purpose of the data:** To validate the microbial engineered microorganisms aimed at improving tolerance and degradation of PAHs, TPHs, pesticides during later bioaugmentation strategies at a microcosms level.
- **Types of data:** Experimental and observational collected data through the instruments used in the research.
- **Form of data:** (1) Text and Numeric for experimental data (2) Instrument-specific for spectroscopic and analytical data
- **Origin of the data:** ICL, a BIOSYSMO partner, provided the engineering selected strains using tools validated in previous work. The strains were generated by ICL using the bioinformatics tools that were validated in their previous work.
- **Data utility outside the project:** Generate microorganisms modified genetically by biotechnology techniques with profits in bioaugmentation and soil bioremediation strategies.

Partner: JSI

- **Task:** Task 3.2
- **Brief description of data:** Data generated by aggregation of the cells, monitoring of the efficiency in pollutant degradation by various approaches (spectrophotometrically, metatranscriptomic and metabolomic tools, flow cytometry, microscopy).
- **Purpose of the data:** The purpose of this data is to establish the efficiency of aggregation/biofilms of the cells isolated in WP1 and to determine the efficiency of pollutant degradation. Based on this data, selection of the most optimal structure and cell combinations for pollutant degradation can be obtained.
- **Types of data:** Experimental data (data from spectrophotometry, zeta-potential reader, flow cytometry, microscopy, genome sequencing and metabolomic and transcriptomic data).
- **Form of data:** numeric and image data
- **Origin of the data:** Data is generated from experiments at the JSI facilities
- **Data utility outside the project:** Future projects and publications.

Partner: LEITAT

- **Task:** Task 3.2.
- **Brief description of data:** Studying electrostatic interactions and electrode redox potential to guarantee a successful inoculation of electroactive biofilms and/or consortia and the conditions to run the bioremediation trials using Bioelectrochemical systems
- **Purpose of the data:** to develop strategies for consortia aggregation and immobilization upon electrodes to use for biodegradation of target pollutants.
- **Types of data:** Experimental data collected under controlled conditions.
- **Form of data:** Numeric and images/graphs.
- **Origin of the data:** Generated.

- **Data utility outside the project:** This data will be useful for the enhancement of existing (bio)remediation technologies for target pollutants (metals and hydrocarbons) of groundwaters.

Partner: UMLP, formerly UBFC

- **Task:** Task 3.1
- **Brief description of data:** Biochemical and genomic characterization of microbial isolates from task 1.3
- **Purpose of the data:** The data will be used to construct and enhance biosystems composed of plants and microbes isolated in WP1 and/or selected in WP2 to achieve the pathways and traits required for effective bioremediation
- **Types of data:** The experimental data will be collected under controlled conditions (e.g., gene sequences, biochemical measurements).
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** Data generated with data input from Task 2.1. Data will serve as a baseline for ensuring reliable results in Task 3.2 and 3.3.
- **Data utility outside the project:** Future researchers in this topic through publications and information dissemination

Partner: UMLP, formerly UBFC

- **Task:** Task 3.3
- **Brief description of data:** Plant traits related to microbial colonization and pollutant tolerance
- **Purpose of the data:** Plant analyses will serve for measuring tolerance to metals and organic pollutants and capacity for colonization with selected bacterial and fungal mycorrhizal/endophyte strains
- **Types of data:** The experimental data will be collected under controlled conditions (e.g., microbial colonization; gene expression and biochemical datasets).
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG
- **Origin of the data:** Data generated from data obtained in Task 2.1.
- **Data utility outside the project:** Future researchers in this topic through publications and information dissemination

Partner: CIIMAR

- **Task:** Task 3.1
- **Brief description of data:** Data from the genomic and metagenomic analyses of the created consortia as well as data related to the capabilities for pollutants biodegradation by the same consortia in microcosms experiments.
- **Purpose of the data:** Molecular and degradation characterization of the consortia developed from microbial strains isolated from the polluted sites, and enhanced, or not, with microorganisms from culture collections to obtain optimized consortia for each combination of target contaminants.
- **Types of data:** Experimental data such as gene sequences, molecular identifications and alpha/beta diversity and instrument-specific data from atomic absorption spectrometry and liquid chromatography analysis for the degradation characterization.
- **Form of data:** text, numeric, instrument-specific bioinformatics data (e.g., FASTQ Sequence and Sequence Quality Format), instrument-specific data (e.g., chromatograms).
- **Origin of the data:** re-used from the CIIMAR Microbial Culture Collection, generated in the project and/or cross-referenced with re-use data from public database
- **Data utility outside the project:** Information on the best microbial consortia for the removal of contaminants can be used by the scientific community to develop new microbe remediation solutions for other polluted sites.

Partner: UPM

- **Task:** Task 3.3
- **Brief description of data:** Physiological parameters of plants treated with Zn and Cd in controlled conditions.
- **Purpose of the data:** Phenotyping poplar plants in response to Zn and Cd toxicity, which is an objective of the project itself.
- **Types of data:** Experimental data collected under controlled conditions: Numerical values of physiological parameters and images of height, root length, fresh and dry weight, stomata conductance, CO₂ assimilation, chlorophyll content, photosynthesis. Mostly obtained by the instrument CIRAS-3 Portable Photosynthesis System. Total metal content, measured by spectroscopy. Every parameter is measured from plants growing in controlled conditions in a biosafety P3 greenhouse.
- **Form of data:** Text, numeric, graphs and images
- **Origin of the data:** Original generated data
- **Data utility outside the project:** Bibliographic research for academics interested in poplar, metal toxicity and stress, abiotic stress and phytoremediation. Phyto and bioremediation companies. Publications

Partner: UPM

- **Task:** Task 3.3

- **Brief description of data:** Physiological parameters of poplar plants in symbiosis with microbes in real contaminated soil from site 10
- **Purpose of the data:** Phenotyping poplar plants in symbiosis with microbes in response to heavy metals toxicity.
- **Types of data:** Numerical values of physiological parameters and images of plants and microbes growing in controlled conditions. Total metal content.
- **Form of data:** Text, numeric, graphs and images.
- **Origin of the data:** Original generated data
- **Data utility outside the project:** Bibliographic research for academics interested in poplar, metal stress and toxicity, abiotic stress, mycorrhization and phytoremediation. Phyto and bioremediation companies. Publications.

Partner: ICL

- **Task:** Task 3.4
- **Brief description of data:** Data generated in this task will include genomic and transcriptomic data (e.g., genome sequences and expression profiles); physiological data relating to microbial phenotypes (e.g., growth curves, enzyme activity profiles, biomass yields); synthetic biology data (e.g., plasmid maps, gene sequences, metabolic models, population dynamics models); process data (e.g., methods and protocols, programming parameters); analytical chemistry data (e.g., spectrometry results, physiochemical property analysis); and visual data (e.g., microscopy images).
- **Purpose of the data:** Data will be used to address Objective#3 – To design, enhance, apply, and optimize synergistic biosystems for biotransformation and uptake of targeted mixed pollutants in water, soil, and sediments. Genomic data will be utilized to identify candidate metabolic pathways responsible for biotransformation activities and transcriptomic data will be used to highlight relative expression levels of different genes of interest. Physiological data will inform on the behavior of individual microbial strains and will permit the characterization of gene functions, and the optimization of biotransformation activities. Synthetic biology data will allow the construction of novel genetic circuits and a greater understanding of microbial interactions which will enable fine-tuning of population dynamics. Process data will lead to refined protocols to ensure the development of robust and reproducible systems. Data arising from chemical analysis will support hypotheses relating to degradation pathways and may also highlight process by-products of interest or concern. Visual data will contribute to a record of physical properties of microbial strains used in the project.
- **Types of data:** Types of data generated in this task will include experimental data collected under controlled conditions relating to biological and physiochemical properties (e.g., genomic data, Sanger sequencing, growth assays, enzyme assays, chromatograms); simulation data (e.g., biological and metabolic models); and data generated from existing datasets (e.g., comparative genomics).
- **Form of data:** Data generated in this task will take the form of text, images, models, and instrument-specific data (e.g., microplate reader assays, spectroscopy data, microscopy).
- **Origin of the data:** Data in this task will be generated primarily through experiments conducted as part of BIOSYSMO. Existing datasets anticipated to be used in the project will be reference microbial genomes obtained from public databases (e.g., NCBI, JGI).
- **Data utility outside the project:** Data generated in this task will be of interest to other members of the community researching the biotransformation of environmental pollutants. Genomic datasets

made publicly available can be mined for metabolic pathways and compared to other isolates with activities of interest. Genetic parts and plasmids developed in this task may have relevance to research projects in a similar field.

WP4. Application of improved biosystems and optimization of the bioremediation strategy

Partner: UBU

- **Task:** Task 4.1
- **Brief description of data:** Physiological, biochemical, microbiology and molecular data will be generated from microcosms experiments during PGPR-assisted phytoremediation. The second part consists of assessing the impact of bioelectrochemical stimulation on pollutants removal and PGPRs performance.
- **Purpose of the data:** To select the best performing tested aquatic species in terms of tolerance and remediation, towards a possible scaling up. Besides, to know more about the bioelectrochemical stimulation on decontamination as well as the role of PGPRs.
- **Types of data:** Experimental and observational collected data through the instruments used in the research.
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** Inoculation with bacteria/PGPRs from Task 3.1 and Task 3.2 or carriers provided by JSI.
- **Data utility outside the project:** To establish effective alternatives of biosystems in bioremediation strategy.

Partner: UBU

- **Task:** Task 4.5
- **Brief description of data:** Data derived from microcosms experiments with real contaminated soil samples and biostimulation/bioaugmentation treatments.
- **Purpose of the data:** To analyze, compare, and find the best performing bioremediation treatment for soil.
- **Types of data:** Experimental and observational data collected through the instruments used in the research.
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG
- **Origin of the data:** For bioaugmentation, inputs from Task 2.3 and Task 3.4 (ICL).

- **Data utility outside the project:** To understand the effectiveness of applied bioremediation and select the most effective for a possible scale-up.

Partner: JSI

- **Task:** Task 4.1
- **Brief description of data:** Data generated by preparation and characterization of the microbes immobilized on solid carriers for water treatment with aquatic plants.
- **Purpose of the data:** The purpose of this data is to establish the best conditions for carrier preparation and to evaluate the efficiency of metabolic activity of the cells. Further data on pollutant removal will be obtained by UBU.
- **Types of data:** Experimental data and instrument-specific data (data from spectrophotometry, flow cytometry, microscopy, genome sequencing and metabolomic and transcriptomic data).
- **Form of data:** numeric and image data
- **Origin of the data:** Data is generated from experiments at the JSI facilities
- **Data utility outside the project:** Future projects and publications.

Partner: JSI

- **Task:** Task 4.3
- **Brief description of data:** Data generated by preparation and characterization of the immobilized microorganisms for application of estuarine sediment phytoremediation.
- **Purpose of the data:** The purpose of this data is to establish the best conditions for carrier preparation and to evaluate the efficiency of metabolic activity of the cells. Further data on pollutant removal will be obtained by CIIMAR.
- **Types of data:** Experimental data and instrument-specific data (data from spectrophotometry, flow cytometry, microscopy, genome sequencing and metabolomic and transcriptomic data).
- **Form of data:** numeric and image data
- **Origin of the data:** Data is generated from experiments at the JSI facilities
- **Data utility outside the project:** Future projects and publications.

Partner: JSI

- **Task:** Task 4.4
- **Brief description of data:** Data generated by preparation and characterization of artificial electroactive biofilms for pollutants removal.
- **Purpose of the data:** The purpose of this data is to characterize the biofilm in terms of cell viability and metabolic activity, spatial orientation etc. and in association with LEITAT, establish pollutant removal efficiency.

- **Types of data:** Experimental data and instrument-specific data (data from spectrophotometry, flow cytometry, microscopy, genome sequencing and metabolomic and transcriptomic data).
- **Form of data:** numeric and image data
- **Origin of the data:** Data is generated from experiments at the JSI facilities
- **Data utility outside the project:** Future projects and publications.

Partner: LEITAT

- **Task:** Task 4.4.
- **Brief description of data:** Experimental data from bioelectrochemical systems for groundwater and wastewater remediation
- **Purpose of the data:** Bioelectrochemical characterization of the samples during the bioremediation treatment.
- **Types of data:** Experimental data collected under controlled conditions.
- **Form of data:** Numeric and images/graphs.
- **Origin of the data:** Generated.
- **Data utility outside the project:** This data will be useful for the enhancement of existing (bio)remediation technologies for target pollutants (metals and hydrocarbons) of groundwaters.

Partner: UMLP, formerly UBFC

- **Task:** Task 4.2.
- **Brief description of data:** Data for construction of biosystems in different bioremediation strategies and to validate them with polluted samples in the lab
- **Purpose of the data:** The data will be used in assisting for the application of poplar-microbe biosystems for soil phytoremediation.
- **Types of data:** The experimental data will be collected under controlled conditions (e.g., soil and microbial parameters, soil agronomic characterization, microbial respiration, C and N, enzymatic tests; biodiversity indicators)
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** newly processed data with data input from Task 3.1 and Task 3.3
- **Data utility outside the project:** Future researchers in this topic through publications and information dissemination

Partner: CIIMAR

- **Task:** Task 4.3

- **Brief description of data:** Physicochemical and Molecular data obtained from mesocosms experiments that simulate estuarine environment (with plants, sediments and enhanced microbial consortia obtained in previous tasks).
- **Purpose of the data:** The optimized plant-bacteria systems will be used in mesocosms experiments mimicking estuarine conditions, the data obtained from those experiments is crucial for the full assessment of the efficiency of the nature-based solution to be implemented.
- **Types of data:** Experimental data such as gene sequences, molecular identifications and alpha/beta diversity and instrument-specific data from atomic absorption spectrometry and liquid chromatography analysis for the degradation characterization.
- **Form of data:** text, numeric, instrument- specific bioinformatics data (e.g., FASTQ Sequence and Sequence Quality Format), instrument-specific data (chromatograms).
- **Origin of the data:** generated and cross-referenced with re-use data from public databases.
- **Data utility outside the project:** A solution that has been properly tested and can assist policymakers to address and tackle the growing pollution issue.

WP5. Field studies, validation and roadmap for wider implementation

Partner: IDENER

- **Task:** Task 5.3
- **Brief description of data:** The mathematical models developed in Task 2.3 will be used to generate modelling data. Finally, scientific literature data from public and open sources will be collected.
- **Purpose of the data:** The generated data from mathematical models developed in task 2.3 will be used to search for alternative scenarios with partner TAUW, to evaluate the feasible application of bioremediation systems performance and multicriteria analysis. The generated data will be used to select the optimal combination of strategies. Furthermore, with technological partners the data analysis will help to study alternatives for recovery and valorization of bioremediation by-products. Finally, a review of the most innovative techniques will be evaluated to assess the feasibility of innovative circular options.
- **Types of data:** (1) Simulation data from mathematical models, and extraction of state-of-the-art knowledge from (2) existing datasets or (3) referenced datasets
- **Form of data:** The data will be in form of numeric (Tables, lists, databases), for the modelling section of the task. For the review, will be numeric (Tables, lists, databases) and text.
- **Origin of the data:** The data from the modelling section will be newly generated from the models, and from the review, will be reused from other studies, collected from open and public sources.
- **Data utility outside the project:** The roadmap results data will serve to guide new potential research in the field of bioremediation, enhancing the knowledge and open possible bioremediation strategies.

Partner: UBU

- **Task:** Task 5.1

- **Brief description of data:** Evaluation of the bioremediation efficacy at the laboratory level using real samples.
- **Purpose of the data:** To select the bioremediation methods for further field test and their validation based on Key Performance Indicators (KPIs) at lower laboratory tests.
- **Types of data:** Simulation data, using experimental and derived data to assess BIOSYSMO technologies.
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** Data compiled from treatment strategy, the characteristics and needs of the specific sites.
- **Data utility outside the project:** To develop technologies without poor results at lower Technology Readiness Level (TRL).

Partner: UBU

- **Task:** Task 5.2
- **Brief description of data:** The best-performing selected treatment will be tested at the pilot scale-up. In addition to increasing synergies and the application of consortia from T4.2.
- **Purpose of the data:** To validate the field testing and the selected biotechnologies (TRL5).
- **Types of data:** Simulation data, using experimental and derived data to assess BIOSYSMO technologies. Observed data.
- **Form of data:** (1) Text and numeric data for recording experimental data and (2) computer code to store bioinformatic data in .mpr, .fasta, and .fastaq
- **Origin of the data:** Locations provided by TAUW, UMLP (formerly UBFC), or CIIMAR. Best-performing treatments from Tasks 5.1 and the consortia mix from Task 4.2.
- **Data utility outside the project:** Information grounding in the realities of a particular setting.

Partner: UBU

- **Task:** Task 5.3
- **Brief description of data:** Definition of the best-suited strategies based on the process conditions and outputs of the simulations.
- **Purpose of the data:** Report guidance on safe management practices.
- **Types of data:** Documents, spreadsheets
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** Results obtained in outputs of the simulations of bioremediation strategies.

- **Data utility outside the project:** Assess the feasibility of innovative circular options such as metal biorecovery, phytomining, or ecocatalysis.

Partner: CIIMAR

- **Task:** Task 5.2
- **Brief description of data:** Physicochemical and Molecular data that is going to be obtained from in situ experiments of microbial-assisted phytoremediation of salt marsh sediments.
- **Purpose of the data:** In a real contaminated environment the best microbe/-plant biosystem will be tested to ensure an effective solution.
- **Types of data:** Observational field data (images), experimental data such as gene sequences, molecular identifications and alpha/beta diversity and instrument-specific data from atomic absorption spectrometry and liquid chromatography analysis for the degradation characterization.
- **Form of data:** text, numeric, images, instrument-specific bioinformatics data (e.g., FASTQ Sequence and Sequence Quality Format), instrument-specific data (e.g., chromatograms).
- **Origin of the data:** generated and cross-referenced with re-use data from public databases.
- **Data utility outside the project:** A solution that has been properly tested can equip policymakers with better tools to address and tackle the growing pollution issue.

WP6. Environmental, socio-economic, risk and regulatory assessment

Partner: UBU

- **Task:** Task 6.1
- **Brief description of data:** Compilation of the information about the environmental Life Cycle Assessment (LCA) attending ISO 1440/14044 standards, datasets, or the EPLCA library.
- **Purpose of the data:** To evaluate the environmental impacts of the proposed bioremediation systems through an LCA.
- **Types of data:** It should contain experimental data and compiled/derived data from existing or referenced datasets
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** Re-used and generated data from observational field data at contaminated sites and experimental data under controlled conditions
- **Data utility outside the project:** Considerations of the environmental impact, the risk assessment and hazard assessment.

Partner: UBU: UNIVERSIDAD DE BURGOS, ES

- **Task:** Task 6.2

- **Brief description of data:** Considering the Life Cycle Costing (LCC) to assess the cost effectiveness of the proposed technologies.
- **Purpose of the data:** To determine the total ownership costs to assess the cost effectiveness of the proposed technologies.
- **Types of data:** Derived data, derived from the experimental data.
- **Form of data:** The data would be in the form of text, numeric and images. They will be compiled as Reports, spreadsheets, presentations - Microsoft Office formats, .pdf. Graphs – GraphPad .pzf. Images – JPEG (compressed), .GIF, .PNG.
- **Origin of the data:** Data for the LCC from WP4-WP5.
- **Data utility outside the project:** Reporting a socio-economic assessment.

Partner: BSY

- **Task:** Task 6.1
- **Brief description of data:** Environmental indicators and data of BIOSYSMO project
- **Purpose of the data:** The data will contain the environmental impact analysis of the project's biotreatments
- **Types of data:** processed data, calculations of the LCA, data derived from existing datasets, references
- **Form of data:** text, numeric, tables and graphs
- **Origin of the data:** ecoinvent 3.9 dataset, Generated data
- **Data utility outside the project:** LCA practitioners, LCA researchers, environmentalists, policy makers

Partner: BSY

- **Task:** Task 6.2
- **Brief description of data:** Economic data of BIOSYSMO project
- **Purpose of the data:** The data will contain the costs and economic analysis of the project's biotreatments
- **Types of data:** processed data, calculations of the LCC, data derived from existing datasets, references
- **Form of data:** text, numeric and graphs
- **Origin of the data:** Generated, surveys
- **Data utility outside the project:** Bioremediation industry, chemical industry, investors

Partner: BSY

- **Task:** Task 6.3

- **Brief description of data:** Social impacts and indicators data of BIOSYSMO project
- **Purpose of the data:** The data will contain the social impact analysis of the project's biotreatments
- **Types of data:** processed data, anonymized questionnaires and surveys, data derived from existing datasets, references
- **Form of data:** text, numeric and graphs
- **Origin of the data:** Generated, surveys
- **Data utility outside the project:** Government, NGOs, environmentalists, citizens and policy makers

WP7. Communication, dissemination and exploitation

Partner: EXE

- **WP/Task:** Task 7.1
- **Brief description of data:**
 - **Website and Social Media:** Social media accounts including their login details, as well as webpage admin login iii. List of news and posts, incl. pictures, banners, videos etc
 - **Communication material for BIOSYSMO:** Logo, Flyers, presentations, posters
 - **Communication activities mapping:** social media announcements and website news announced by partners, communication of the project in different events
- **Purpose of the data:** Communication of project results
- **Types of data:** NA
- **Form of data:** Text, numeric, images, audiovisual, XML files, Adobe Photoshop (.png, .jpeg), presentations
- **Origin of the data:** Data are generated by EXELISIS and can be reused by the consortium and beyond.
- **Data utility outside the project:** Data open to the public through BIOSYSMO website and social media

Partner: EXE

- **WP/Task:** Task 7.2.
- **Brief description of data**
 - Dissemination material for BIOSYSMO: Additional flyers, posters, videos, newsletters, press releases, and other dissemination materials to raise awareness of project outcomes among stakeholders.
 - Clustering: List of relevant projects and contact persons, meeting minutes, List of common areas of interest, agendas of common events and planning information.
 - Dissemination material for cluster including flyers, video, posters, policy brief, etc.

- Events participation. Lists of past and upcoming events that the consortium attended or is planning to attend.
- Publications. List of accepted publications and plan for upcoming ones.
- Events organisation: List of event organisers, List of presenters/invited speakers/ projects, Communication material for the events, Video recordings (if any), Evaluation forms, Participation certificates
- **Purpose of the data:** Dissemination of project results
- **Types of data:** NA
- **Form of data:** Text, numeric, images, audiovisual, XML files, Adobe Photoshop (.png, .jpeg), presentations
- **Origin of the data:** Data are generated by EXELISIS and can be reused by the consortium and beyond.
- **Data utility outside the project:** Data open to the public through BIOSYSMO website and social media

Partner: AXIA

- **WP/Task:** Task 7.3.
- **Brief description of data:**
 - **Monitoring of the KERs**
 - **Data about the targeted markets related to the project**
 - **SWOT and PESTLE analysis of the grouped KERs**
 - **Business plans for the commercial partners**
- **Purpose of the data:** Exploitation of project results
- **Types of data:** NA
- **Form of data:** Text, figures, tables, and graphs
- **Origin of the data:** Generated
- **Data utility outside the project:** The data will be useful for policy makers, innovators, environmentalists, citizens

Partner: AXIA

- **WP/Task:** Task 7.4. IPR and knowledge management [M1-M48]
- **Brief description of data:**
 - Background and foreground knowledge
 - Exploitation routes expected for the partners
 - Ownership expected per each KER

- **Purpose of the data:** Exploitation of project results
- **Types of data:** NA
- **Form of data:** Text, figures, tables, and graphs
- **Origin of the data:** Generated
- **Data utility outside the project:** The data will be useful for policy makers, innovators, environmentalists, citizens

Annex III: Metadata required by any Zenodo deposit

Minimum metadata for publications

Items marked with * are mandatory for any Zenodo upload; items marked with (*) will be mandatory for any BIOSYSMO upload (metadata of publication, conference, book/report/chapter or thesis, as relevant).

Upload type
required

Publication
Poster
Presentation
Dataset
Image
Video/Audio
Software
Lesson
Physical object
Workflow
Other

Publication type
Journal article
Publication type.

Basic information
required

Digital Object Identifier
e.g. 10.1234/foo.bar
Optional. Did your publisher already assign a DOI to your upload? If not, leave the field empty and we will register a new DOI for you. A DOI allows others to easily and unambiguously cite your upload. Please note that it is NOT possible to edit a Zenodo DOI once it has been registered by us, while it is always possible to edit a custom DOI.
Reserve DOI

Publication date *
2023-02-01
Required. Format: YYYY-MM-DD. In case your upload was already published elsewhere, please use the date of first publication.

Title *
Required.

Authors *
Family name, given names
Affiliation
ORCID (e.g.: 0000-0002-1825-0097)
Optional.
Add another author

Description *
Required.

Version
Optional. Mostly relevant for software and dataset uploads. Any string will be accepted, but semantically-versioned tag is recommended. See semver.org for more information on semantic versioning.

Language *

Optional. Primary language of the record. Start by typing the language's common name in English, or its ISO 639 code (two or three-letter code). See [ISO 639 language codes list](#) for more information.

Keywords *

[+ Add another keyword](#)

Additional notes

Optional.

License

required ▼

- Access right** *
- ☒  Open Access
 - ☐  Embargoed Access
 - ☐  Restricted Access
 - ☐  Closed Access

Required. Open access uploads have considerably higher visibility on Zenodo.

License *

Required. Selected license applies to all of your files displayed on the top of the form. If you want to upload some of your files under different licenses, please do so in separate uploads. If you cannot find the license you're looking for, include a relevant LICENSE file in your record and choose one of the *Other* licenses available (*Other (Open)*, *Other (Attribution)*, etc.). The supported licenses in the list are harvested from [opendefinition.org](#) and [spx.org](#). If you think that a license is missing from the list, please [contact us](#).

Funding

recommended ▼

Zenodo is integrated into reporting lines for research funded by the European Commission via [OpenAIRE](#). Specify grants which have funded your research, and we will let your funding agency know!

Grants

Optional. OpenAIRE-supported projects only. For other funding acknowledgements, please use the *Additional Notes* field.
Note: a human Zenodo curator will need to validate your upload - you may experience a delay before it is available in OpenAIRE.

[+ Add another grant](#)

Related/alternate identifiers

recommended ▼

Specify identifiers of related publications and datasets. Supported identifiers include: DOI, Handle, ARK, PURL, ISSN, ISBN, PubMed ID, PubMed Central ID, ADS Bibliographic Code, arXiv, Life Science Identifiers (LSID), EAN-13, ISTC, URNs and URLs.

Related identifiers

Optional. Resource type of the related identifier.

[+ Add another related identifier](#)

Contributors
optional

Contributors

Optional.

+ Add another contributor

References
optional

References

+ Add another reference.

Journal
optional

Journal title

Optional.

Volume

Optional.

Issue

Optional.

Pages

Optional.

Conference
optional

Conference title

Optional.

Acronym

Optional.

Dates

Optional.

Place

Optional.

Website

Optional. e.g. <http://zenodo.org>

Session

Optional. Number of session within the conference.

Part

Optional. Number of part within a session.

Book/Report/Chapter ^{*} optional ▼

For parts of books and reports.

Publisher
Optional.

Place
Optional.

ISBN
Optional.

Book title
Optional. Title of the book or report which this upload is part of.

Pages
Optional.

Thesis ^{*} optional ▼

Awarding university
Optional.

Supervisors

Family name, given names Affiliation ✕
Optional.

[+ Add another supervisor](#)

Subjects optional ▼

Specify subjects from a taxonomy or controlled vocabulary. Each term must be uniquely identified (e.g. a URL). For free form text, use the keywords field in basic information section.

Subjects ✕

[+ Add another subject](#)

Minimum metadata for datasets

Items marked with ^{*} are mandatory for any Zenodo upload; items marked with (^{*}) will be mandatory for any BIOSYSMO upload.

Upload type required ▼

 Publication ☐	 Poster ☐	 Presentation ☐	 Dataset ☒	 Image ☐	 Video/Audio ☐	 Software ☐	 Lesson ☐	 Physical object ☐	 Workflow ☐	 Other ☐
---	--	--	--	---	---	--	--	---	--	---



Funded by
the European Union

License
required

Access right *

☒ Open Access
☐ Embargoed Access
☐ Restricted Access
☐ Closed Access

Required. Open access uploads have considerably higher visibility on Zenodo.

License *

Creative Commons Attribution 4.0 International

Required. Selected license applies to all of your files displayed on the top of the form. If you want to upload some of your files under different licenses, please do so in separate uploads. If you cannot find the license you're looking for, include a relevant LICENSE file in your record and choose one of the *Other* licenses available (*Other (Open)*, *Other (Attribution)*, etc.). The supported licenses in the list are harvested from opendefinition.org and spdx.org. If you think that a license is missing from the list, please [contact us](#).

Funding

Zenodo is integrated into reporting lines for research funded by the European Commission via [OpenAIRE](#). Specify grants which have funded your research, and we will let your funding agency know!

Grants

European Commission (EU)

Start typing a grant number, name or abbreviation...

Optional. OpenAIRE-supported projects only. For other funding acknowledgements, please use the *Additional Notes* field.
Note: a human Zenodo curator will need to validate your upload - you may experience a delay before it is available in OpenAIRE.

+ Add another grant

Related/alternate identifiers
recommended

Specify identifiers of related publications and datasets. Supported identifiers include: DOI, Handle, ARK, PURL, ISSN, ISBN, PubMed ID, PubMed Central ID, ADS Bibliographic Code, arXiv, Life Science Identifiers (LSID), EAN-13, ISTC, URNs and URLs.

Related identifiers

e.g. 10.1234/foobar.56789

N/A

Optional. Resource type of the related identifier.

+ Add another related identifier

Contributors
optional

Contributors

Family name, given name

Affiliation

 ORCID (e.g.: 0000-0002-1)

Optional.

+ Add another contributor

References
optional

References

e.g.: Cranmer, Kyle et al. (2014). Decouple software associated to arXiv:1401.0080.

+ Add another reference.

Annex IV: Protocol for uploading datasets to Zenodo

1. Preparation of Datasets

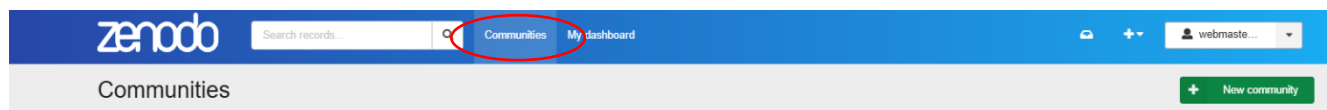
- **Compliance Check:** Ensure that the dataset complies with the project's data management plan (DMP) and any relevant legal or ethical guidelines.
- **Data Cleaning:** Ensure all datasets are cleaned and formatted according to the project's DMP.
- **Metadata Creation:** Prepare detailed metadata for each dataset, including title, description, authors, keywords, and funding information.

2. Zenodo Account Setup

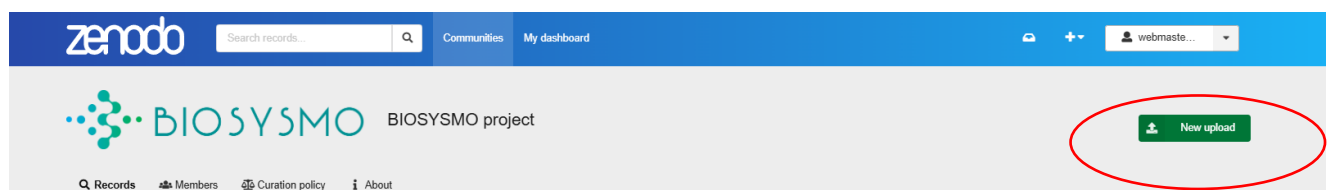
- **Account Creation:** Each partner must create a Zenodo account if they do not already have one (entity zenodo accounts can also be used).
- **Profile Completion:** Complete the profile with relevant information, including affiliation and ORCID iD.

3. Dataset Upload Process

- **Login:** Log in to your Zenodo account.
- **BIOSYSMO Community:** All datasets will be stored in the Zenodo BIOSYSMO community for centralized access.



- **New Upload:** Click on the "New Upload" button on the Zenodo dashboard.



- **File Upload:** Drag and drop the dataset files or select them from your computer.

- **DOI:** Request for a DOI to be added by zenodo via “get a DOI now”. Just note that when the “No DOI” option is selected, it is not mandatory to obtain a DOI. However, it is highly recommended.

Digital Object Identifier*

Do you already have a DOI for this upload? ☐ Yes ☒ No

- **Metadata Entry:** Fill in the metadata fields accurately. Ensure all required fields are completed.
- **Publication date:** As it is a compulsory item, in cases where the dataset is not yet published, they can enter a placeholder year such as 2025 or an interval like 2021/2026, representing the project duration.
- **Date:** Regarding the Date section, if the dataset has not yet been published or submitted, users can provide alternative dates, such as when the data was collected or created.
- **Licenses:** Licences are not mandatory, and until publication may be adequate not to assign one initially. However, once a dataset is released and considered open access, it must be distributed under a license permitting free use, modification, and sharing (e.g., Creative Commons Licenses such as CC0, CC BY, or CC BY-SA).
- **Link with other repositories:** There is an option to link the dataset to relevant repositories (e.g., NCBI) or corresponding publications.

Related works >

Specify identifiers of related works. Supported identifiers include DOI, Handle, ARK, PURL, ISSN, ISBN, PubMed ID, PubMed Central ID, ADS Bibliographic Code, arXiv, Life Science Identifiers (LSID), EAN-13, ISTD, URNs, and URLs.

Related works

Relation*	Identifier*	Scheme*	Resource type
Select relation...			

- **Funding Information:** Add funding information (under **Awards/Grants**), including grant numbers and funding agencies. Use both for EU funding and UKRI.

Add standard award/grant

☐ BIOSYSMO — BIOremediation systems exploiting SYnergieS for improved removal of Mixed pOllutants
European Commission

☐ BIOremediation systems exploiting SYnergieS for improved removal of Mixed pOllutants (BIOSYSMO)
UK Research and Innovation

4. Restricted option

- **Restricted button:** Make restrictions to data that should not be openly available

 **Visibility** *

Files only

☐ Public
 ☒ **Restricted**

The files of this record are restricted.



Public with restricted files
The record is publicly accessible. The files can only be accessed by users specified in the permissions.

5. Review and Submission

- **Review:** Carefully review the uploaded files and metadata for accuracy.
- **Submit for review:** Click the "Submit for review" button to submit the dataset. Note that once published, the dataset cannot be edited.

Draft 

6. Post-Submission

- **Internal review:** IDENER, as the project coordinator, will be responsible for reviewing and admitting datasets to be published in this community.
- **Sharing:** Share the DOI with the project lead and other relevant stakeholders.

7. Documentation

- **Documentation:** Document the upload process and any issues encountered in the project's mapping excel. Report these to the project coordinator for future reference.

Annex V: Dataset index

PID	Type of PID	Description of dataset	Is this dataset available in open access	URL to repository
MN128762	Other	<i>Pseudomonas</i> sp. _DPS1_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128762.1
MN128739	Other	<i>Bosea</i> sp. _DP2_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128739.1
MN128776	Other	<i>Shewanella</i> sp. _DPS3_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128776.1
MN128745	Other	<i>Chitinophagaceae</i> bacterium_DPS4_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128745.1
MN128733	Other	<i>Acinetobacter</i> sp. _DPS5_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128733.1
MN128764	Other	<i>Pseudomonas</i> sp. _DPS6_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128764.1
MN128740	Other	<i>Bosea</i> sp. _DPS7_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128740.1
MN128750	Other	<i>Leadbetterella</i> sp. _DPS8_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128750.1
MN128754	Other	<i>Microbacterium oxydans</i> _16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128754.1

MN128763	Other	<i>Pseudomonas</i> sp. _DPS10_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128763.1
MN128732	Other	<i>Acinetobacter</i> sp. _DBS1_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128732.1
MN128747	Other	<i>Herminiimonas</i> sp. _DBS2_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128747.1
MN128746	Other	<i>Dyadobacter</i> sp. _DBS3_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128746.1
MN128755	Other	<i>Ochrobactrum rhizosphaerae</i> _DBS4_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128755.1
MN128752	Other	<i>Leucobacter</i> sp._DBS5_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128752.1
MN128758	Other	<i>Pseudomonas</i> sp. _DBS6_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128758.1
MN128759	Other	<i>Pseudomonas</i> sp. _DBS7_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128759.1
MN128760	Other	<i>Pseudomonas</i> sp. DBS8_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128760.1
MN128761	Other	<i>Pseudomonas</i> sp. _DBS9_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128761.1
MN128753	Other	<i>Microbacterium oxydans</i> _DBS10_16S ribosomal RNA gene, partial sequence	yes	https://www.ncbi.nlm.nih.gov/nuccore/MN128753.1

10.5281/zenodo.14781136	DOI	Compilation of 16s RNA gene, partial sequence form MN128762-MN128753	yes	https://zenodo.org/records/14781136
PRJNA1130411	Accession number	whole-genome resequencing of <i>Metarhizium robertsii</i>	yes	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1130411
PRJNA1130442	Accession number	whole-genome resequencing of <i>Chalara bambusicola</i>	yes	https://www.ncbi.nlm.nih.gov/bioproject/1130442
PRJNA1130693	Accession number	whole-genome resequencing of <i>Hyalosciphia sp.</i> ,	yes	https://www.ncbi.nlm.nih.gov/bioproject/1130693
PRJNA1130711	Accession number	whole-genome resequencing of <i>Cladophialophora chaetospora</i>	yes	https://www.ncbi.nlm.nih.gov/bioproject/1130711
PRJNA1130799	Accession number	whole-genome resequencing of <i>Massarina sp.</i> ,	yes	https://www.ncbi.nlm.nih.gov/bioproject/1130799
PRJNA1130921	Accession number	whole-genome resequencing of <i>Coprinellus domesticus</i>	yes	https://www.ncbi.nlm.nih.gov/bioproject/1130921
PRJNA1131302	Accession number	whole-genome resequencing of <i>Penicillium ochrochloron</i>	yes	https://www.ncbi.nlm.nih.gov/bioproject/1131302
PRJNA1131310	Accession number	whole-genome resequencing of <i>Paraconiothyrium fuckelii</i>	yes	https://www.ncbi.nlm.nih.gov/bioproject/1131310
PRJNA1131796	Accession number	whole-genome resequencing of <i>Penicillium kongii</i>	yes	https://www.ncbi.nlm.nih.gov/bioproject/1131796
PRJNA1131974	Accession number	whole-genome resequencing of <i>Cadaphora sp.</i>	yes	https://www.ncbi.nlm.nih.gov/bioproject/1131974

PRJNA1132762	Accession number	whole-genome resequencing of <i>Leptodontidium</i> sp	yes	http://www.ncbi.nlm.nih.gov/bioproject/1132762
PRJNA1128693	Accession number	whole-genome resequencing of <i>Trichoderma simmonsii</i>	yes	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1128693
10.5281/zenodo.14886809	DOI	Compilation of whole-genome resequencing form PRJNA1130411-PRJNA1128693	yes	https://zenodo.org/records/14886810
10.5281/zenodo.14861167	DOI	Curated collection of protein datasets related to Metal resistance in bacteria, fungi, and plants.	No yet, pending authorization	https://zenodo.org/records/14861167
10.5281/zenodo.14795254	DOI	BIOSYSMODb: Curated Database for Biodegradation and Bioremediation	No yet, pending authorization	https://zenodo.org/records/14795254
10.5281/zenodo.14886810	DOI	Dataset from phytoremediation experiments with <i>P. australis</i> (+ <i>S. holoschoenus</i>) and water from Site 6. Bucket and mesocosms.	No yet, pending publication	https://zenodo.org/records/14886810

Annex VI: Overview of data available in the BIOSYSMO collection for Zenodo

Dataset Title	Category of Data	Type of Data	Format	Access Mode	Partner	Zenodo URL / DOI
BIOSYSMO Newsletter Issue 1	C&D data	Newsletter	PDF	Open	All	https://zenodo.org/records/10136343 DOI 10.5281/zenodo.10136342.
BIOSYSMO Newsletter Issue 2	C&D data	Newsletter	PDF	Open	All	https://zenodo.org/records/10136476 DOI 10.5281/zenodo.10136476
BIOSYSMO Newsletter Issue 3	C&D data	Newsletter	PDF	Open	All	https://zenodo.org/records/14235169 DOI 10.5281/zenodo.14235169
BIOSYSMO Newsletter Issue 4	C&D data	Newsletter	PDF	Open	All	https://zenodo.org/records/14235075 DOI 10.5281/zenodo.14235075
D1.2 Report on the identification and characterization of microbial isolates/consortia and related remediation competences	Reference data	Deliverable	PDF	Open	All	https://zenodo.org/records/14384749 DOI 10.5281/zenodo.14384749
D3.1 Intermediate report on microbial consortia for enhanced phytoremediation for each combination of target contaminants and matrices	Reference data	Deliverable	PDF	Open	All	https://zenodo.org/records/14361807 DOI 10.5281/zenodo.14361807
D3.2 Intermediate report on aggregated and immobilized consortia and biofilms - Outline	Reference data	Deliverable	PDF	Open	All	https://zenodo.org/records/14361881 DOI 10.5281/zenodo.14361881
D3.3 Intermediate report on genetically modified poplar lines with effective	Reference data	Deliverable	PDF	Open	All	https://zenodo.org/records/14361972 DOI 10.5281/zenodo.14361972

performances on metal uptake and endophyte colonization						
D3.4 Intermediate report on genetically enhanced and adapted microorganisms able to improve biodegradation of pollutants	Reference data	Deliverable	PDF	Open	All	https://zenodo.org/records/14362169 DOI 10.5281/zenodo.14362169
D7.1 Visual identity, website and social media accounts	Reference data	Deliverable	PDF	Open	All	https://zenodo.org/records/14362394 DOI 10.5281/zenodo.14362394
D7.2 Communication tools (poster, leaflet, roll-up, brochure)	Reference data	Deliverable	PDF	Open	All	https://zenodo.org/records/14362439 DOI 10.5281/zenodo.14362439
D7.3 PEDR including communication activities, reviewed version	Reference data	Deliverable	PDF	Open	All	https://zenodo.org/records/14362482 DOI 10.5281/zenodo.14362482
D8.1 Project Management Handbook	Reference data	Deliverable	PDF	Open	All	https://zenodo.org/records/14362191 DOI 10.5281/zenodo.14362191
D8.2 First version of the Data Management Plan	Reference data	Deliverable	PDF	Open	All	https://zenodo.org/records/14362313 DOI 10.5281/zenodo.14362313
Poplar as a woody model for the phytomanagement of trace element contaminated soils	Reference data	Chapter	PDF	Open	UMLP	https://zenodo.org/records/14235213 DOI 10.5281/zenodo.14235213
Pharmaceuticals in Water: Risks to Aquatic Life and Remediation Strategies	Reference data	Scientific publication	PDF	Open	UBU	https://zenodo.org/records/8391540 DOI 10.5281/zenodo.8391540
The potential of microorganisms as biomonitoring and bioremediation tools for mercury-contaminated soils	Reference data	Scientific publication	PDF	Open	UMLP	https://zenodo.org/records/8317975 DOI 10.5281/zenodo.8317975

Water potential governs the effector specificity of the transcriptional regulator XylR of Pseudomonas putida	Reference data	Scientific publication	PDF	Open	ICL	https://zenodo.org/records/7813379 DOI 10.5281/zenodo.7813379
Production of selenium nanoparticles occurs through an interconnected pathway of sulphur metabolism and oxidative stress response in Pseudomonas putida KT2440	Reference data	Scientific publication	PDF	Open	ICL	https://zenodo.org/records/7813373 DOI 10.5281/zenodo.7813373
Whole-genome resequencing of fungi isolated from an industrial wasteland in Vieux-Charmont, France	Experimental data	Dataset	FASTAQ	Open	UMLP	https://zenodo.org/records/14850145 DOI 10.5281/zenodo.14850145
Dataset of Proteins Involved in Metal Resistance	Derived Data	Dataset	FASTAQ CSV	Restricted	IDE	https://zenodo.org/records/14861167 DOI 10.5281/zenodo.14861167
BIOSYSMODb: Curated Database for Biodegradation and Bioremediation	Derived Data	Dataset	CSV, TXT, IPYNB	Restricted	IDE	https://zenodo.org/records/14795254 DOI 10.5281/zenodo.14795254
DPS_DBS Isolation Metadata	Derived Data	Dataset	PDF	Open	CIIMAR	https://zenodo.org/records/14781136 DOI 10.5281/zenodo.14781136
Phytoremediation data for buckets and mesocosm	Experimental data	Dataset	XLSX	Restricted	UBU	https://zenodo.org/records/14886810 DOI 10.5281/zeodo.14886810