Ref. Ares(2024)7744384 - 31/10/2024

D8.1. Initial Data Management Plan

VALORISH

GREEN VALORISATION CASCADE APPROACH OF FISH WASTE AND BY-PRODUCTS THROUGH FERMENTATION TOWARDS A ZERO-WASTE FUTURE

Grant Agreement Number 101135078

Deliverable name:	D8.1 Initial data management plan
Deliverable number:	8.1
Deliverable type:	DMP
Work Package:	WP8: Project coordination
Lead beneficiary:	IDENER
Contact person:	Igor de las Heras López; igor.delasheras@idener.ai
Dissemination Level:	Public
Due date for deliverable:	October 31st, 2024

Funded by the European Union

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

DOCUMENT CONTROL PAGE

Author(s):	IDENER Team
Contributor(s):	All
Reviewer(s):	All
Version number:	v.1.0
Contractual delivery date:	31-10-2024
Actual delivery date:	31-10-2024
Status:	Draft

REVISION HISTORY

Version	Date	Author/Reviewer	Notes
v.0.1	14-10-2024	IDENER Team/All	Creation, First Draft
V0.2	24-10-2024	IDENER Team/All	Reviewed version with comments from partners
V1.0	31-10-2024	IDENER Team/All	Final review; version to submit

ACKNOWLEDGEMENTS

The work described in this publication was subsidised by Horizon Europe (HORIZON) framework through the Grant Agreement Number 101135078.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Deliverable D8.1 is the first version of the Data Management Plan (DMP) of the VALORISH project. This document is an essential pillar of the open science priority of the Horizon Europe Programme. The DMP describes the overall research data management strategy to be followed in the project, including the provisions to be taken to make the research data findable, accessible, interoperable and re-usable (FAIR). The DMP will be updated throughout the project whenever significant changes arise and at least in months 24 and 40, when updated versions of the DMP will be submitted as deliverables D8.2 and D8.3, respectively.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

Abbreviation	Definition
ASCII	American Standard Code for Information Interchange
CC	Creative Commons
CERN	Conseil Européen pour la Recherche Nucléaire, European Organization for Nuclear Research
ChEBI	Chemical Entities of Biological Interest Ontologies
CID	Chemical ID
DMP	Data Management Plan
DOI	Digital Object Identifier
EC	European Commission
ENA	European Nucleotide Archive
EU	European Union
FAIR	Findable, Accesible, Interoperable and Reusable
GO	Gene Ontology
InChI	IUPAC International Chemical Identifier
IPR	Intellectual Property Rights
JERM	Just Enough Results Model
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCA	Life cycle analysis
LIMS	Laboratory Information Management System
MIBBI	Minimum Information for Biological and Biomedical Investigations
MIGS/MIMS	Minimum Information for Biological and Biomedical Investigations
MIQE	Minimum Information for Publication of Quantitative Real-Time PCR Experiments
MIRIAM	Minimum Information Required in the Annotation of Models
MIxS	Minimum Information about any (x) Sequence
NCBI	National Center for Biotechnology Information
ORCID	Open Researcher and Contributor iD
PID	Persistent Identifier
RDM	Research Data Management
REA	Research Europe Agency
SBML	Systems Biology Markup Language
SBOL	Synthetic Biology Open Language
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 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 VALORISH project. Research data typically follows a cycle composed of several consecutive and interdependent processes (Figure 1). Therefore, this strategy covers the full lifecycle of the research data collected, processed and/or generated within the project.

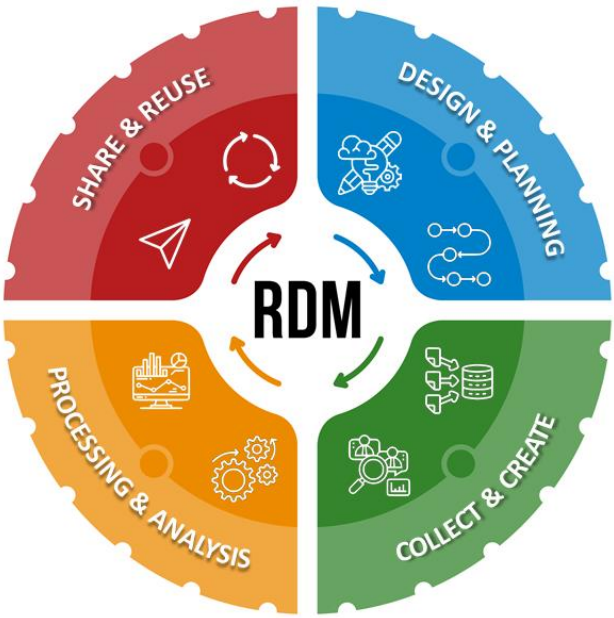


Figure 1. Schematic representation of the typical research data management (RDM) lifecycle. Image sourced from Bongaerts (2022) [2].

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 in 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 raw data. The third stage includes 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 VALORISH 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 will be updated throughout the project. This first version submitted as deliverable D8.1 in M6 will be updated in M24 and M40 and in between whenever significant changes arise, such as:

- 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)

The initial version of the DMP was prepared with the collaboration of all the partners. A data management questionnaire (Annex I: Data management questionnaire) was shared with the partners in order to collect their information regarding:

- 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 already covers a broad range of aspects related to the data management in VALORISH. Nevertheless, the upcoming versions will provide more details of particular issues such as data interoperability and practical data management procedures implemented by the VALORISH consortium. The present document, therefore, includes the available information at the time of writing and acknowledges some gaps of information that will be covered in later versions.

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).
The VALORISH project will generate different types of data that can be categorized into observational, experimental, simulation, derived/compiled or reference/canonical depending on the collection mode. The different types of data have different characteristics and degrees of reproducibility:

- **Observational data** are captured in real-time, typically outside the laboratory, through observation of a behaviour 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.

Likewise, the VALORISH project will generate 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 VALORISH 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 will be updated throughout the project as soon as new data is identified.

3 FAIR data

The FAIR data concept was formulated in 2016 to support the reuse of research data [1]. FAIR data are those data that meet the requirements of findability, accessibility, interoperability and reusability. The VALORISH 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 data [3]. 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 locatable 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; 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 VALORISH consortium.

Repository	Type of repository	Type of data	Persistent identifier
Zenodo (https://zenodo.org/)	General purpose	Any kind of data	DOI
Dryad (https://datadryad.org/stash)[1]	General purpose	Any kind of data underlying peer-reviewed scientific and medical literature	DOI
BioModels Database (https://www.ebi.ac.uk/biomodels/)	Biological mathematical models	Mechanistic models in standard formats	DOI
Laboratory information management system (LIMS)	Internal laboratory experimental result storage	Any kind of data	none
European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home)	Disciplinary	Nucleotide sequences	ORCID Data
NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/)	Disciplinary	Nucleotide sequences, organism and strain information	NCBI Taxonomy ID
NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/)	Disciplinary	Array- and sequence-based data	Sequence identifier
GitHub (https://github.com/)[2]	Other	Software code	Digital Object Identifier (DOI), if linked with a Zenodo entry

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. In the particular case of Zenodo, which will be the reference open access repository for scientific articles to be used for GA open

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

[2] 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

access mandate compliance, all the required metadata specified on the GA will be included.

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.

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
MIBBI (Minimum Information for Biological and Biomedical Investigations)	Comprehensive metadata of genomics and proteomics	ENA, NCBI GenBank
Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE)	qPCR and RT-qPCR	ENA, NCBI GenBank
NCBI taxonomy ID	Organism and strain identification	NCBI
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
UniProtKB Identifiers	Standardized protein sequence references	UniPRot
Systems Biology Markup Language (SBML)	Encoding computational models	SBML
MIRIAM Guidelines	GEM model annotation and curation	BioModels database
COBRA format	Constraints-based metabolic models	Bio.tools
SBML Level 3 FBC Package	Representing flux	SBML

	balance constraints models	
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)
ChEBI (Chemical Entities of Biological Interest) Ontologies	Chemical compound identification	CheBI
ISO guidelines	Chemical and sensory analysis	ISO Standards
KEGG Brite database	Pathway and enzyme information and ontological classification	KEGG
FASTA	Nucleotide and aminoacid sequences	UniProt

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 communication and dissemination materials resulting from VALORISH project will include the EU acknowledgement and disclaimer according to the GA.

3.2 Making data accessible

The second element of the FAIR Guiding Principles aims to enable 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 VALORISH (https://zenodo.org/communities/valorisheu). 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.
At the time of deposit, the data will be given open, restricted or embargoed access (e.g., an embargo may apply to data that will be published until the publication is accepted). 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 will be included in *Annex III: Metadata required by any Zenodo deposit* DMP and 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.
Deposit of 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 are accessible, even when the data are 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 preservation [4].

Type of data	Recommended formats	Acceptable formats
Tabular data with extensive metadata - 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)
Tabular data with minimal metadata - 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)
Textual data	- 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
Documentation and scripts	- 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., XHMTL 1.0
Image data	- 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)
Audio data	- 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)
Video data	- MPEG-4 (.mp4) - OGG video (.ogv, .ogg) - motion JPEG 2000 (.mj2)	- AVCHD video (.avchd)
Geospatial data vector and raster data	- ESRI Shapefile (.shp, .shx, .dbf,.prj, .sbx, .sbn optional) - geo-referenced TIFF (.tif, .tfw) - CAD data (.dwg) - tabular GIS attribute data - Geography Markup Language (.gml)	- ESRI Geodatabase format (.mdb) - MapInfo Interchange Format (.mif) for vector data - 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.

3.4.1 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 VALORISH 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 VALORISH General Assemblies to ensure ongoing suitability.

4 Other research outputs

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 Project Management Handbook and in the PDEC (D7.3).
All the deposited items will follow the FAIR Principles Guidelines described in Section 3. All outputs and in particular publications resulting from the VALORISH project will include the EU acknowledgement and disclaimer according to the GA.

5 Allocation of resources

IDENER is the partner responsible for supervising the data management and updating the DMP within the VALORISH project. Nonetheless, each partner of the VALORISH 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 report them to IDENER in order to update the DMP. IDENER will be responsible for updating this new information in the DMP.
The following Table 4 summarizes the distribution of responsibilities within the VALORISH project.

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

Action	Responsible partner
Oversee of data management	IDENER
Update of DMP	IDENER
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 VALORISH community in Zenodo	Data owner(s)
Review and approval of data uploaded to the VALORISH community in Zenodo	IDENER
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 VALORISH community in Zenodo (https://zenodo.org/communities/valorisheu), 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

7.1 Ethical issues

Some ethical issues may appear regarding research on humans and the handling of personal data for social activities. For this case, a specific data management protocol has been defined to solve and correctly manage such issues.
Apart of that, no further ethical issues that can have an impact on data sharing have been identified. These issues will be monitored during the execution of the project and, if needed, taken into account in the next versions of the DMP.

7.2 Legal issues

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

- 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

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

Besides the procedures for data management described in the VALORISH DMP, the following partners will additionally follow other national, funder, sectorial or departmental procedures for data management.

- BLUES: BLUES has developed a protocol to avoid any issue regarding research on humans and the handling of personal data (T6.3, T6.4, T7.4, T7.5):

Ethics consideration in the specified analysis
Collection and analysis of data are an essential component of the social-related analysis, clustering, networking and engagement activities in VALORISH project and will be undertaken in accordance with the data management plan. Data related to the specified analysis will be archived in suitable formats and held on secure servers located in Spain with secure access and appropriate backup to ensure data security and allow easy data processing. The following ethical issues will be considered during the implementation of VALORISH:

• *Data Privacy:* Privacy of the participants of the specified studies and activities of tasks 6.3; 6.4; 7.4 and 7.5 will be respected, no personal and organisational data will be shared publicly neither even outside the social, clustering and engagement activities managed by BLUES.
• Data Protection: Anonymity and confidentiality of the data will be ensured.
• Data Security: Data will be stored in BLUES's virtual secure servers.

• Informed consent:

- For the case of the online surveys, online Q-sorts participants will be asked to read and agree with the BLUES privacy policy before clicking the submit button on the online form. The privacy policy statement will also contain an explanation about the data management of the gathered information and the email operations@bluesynergy.eu for the participants to ask any change about their personal data.

- For the case of the focus group events, interviews and limited on-site activities, participants will be informed previously by email, using a defined template, about the BLUES privacy policy and about the data management of the information to be gathered. This email will also contain the email operations@bluesynergy.eu for the participants to ask any change about their personal data.

Participants searching and Data Collection

During the course of the project, it will be necessary to found participants for the social, networking and clustering tools to be implemented and collect certain data related to individuals and organisations to complete the specified studies of VALORISH's Technologies and final applications. These data can include the name, last name, email, organization, position, geographic location and phone number of participants.

BLUES will search for participants, related with VALORISH value chain, for the target studies. These potential participants will be contacted through different means: 1) Stakeholder database hold by the dissemination partner, 2) BLUES's contacts, 3) Information about related projects published in the different European Commission portals and Webs, 4) industry associations, 5) NGOs, 6) Universities and Research institutions, 7) VALORISH partners' references, 8) Public events organized by the VALORISH consortium.

Different data collection tools may be implemented in the course of VALORISH to gather required information from participants for the study. Some examples are the following:

A) *Online surveys:* Online customized surveys may be implemented by using Microsoft forms tool or another similar tool.

B) *Focus group events:* Relevant participants, in small groups, may be invited to participate in focus group sessions.

C) *Interviews:* Representatives of relevant sectors, industries, groups of opinion, etc; may be invited to an interview to know in more detail their feedback about the project's technologies, final applications and social issues.

D) *Workshops or Forums:* Verbal or written feedback may be gathered from attendees.

Data Storage

The duration of the VALORISH project will be of 42 months. During this period of 42 months and an additional period of 60 months, all the information processed by BLUES regarding the specified studies and analysis of target tasks will be safely stored in BLUES's secure servers summing a total of 108 months of data retention. After this period, the collected and processed information will be safely removed. The additional period of 60 months after the project finalization is in line with the procedures

established by the European Commission for Horizon Europe project execution regarding auditing and data backup procedures.

Technical and organisational measures to ensure personal data rights, security and protection

According to GDPR, technical and organizational measures will be implemented and executed to safeguard the data rights, data security and protection of the participants in the target tasks. To that end, at BLUES, the following measures (Table 5) will be driven.

Table 5. Technical and organisational measures to be followed by BLUES.

Technical Measures	Organizational measures
Privacy-oriented databases and filed for storage of confidential data (e.g., secure encryption).	Definition of a data-oriented hierarchy, where only authorized persons have access to the confidential data (access control).
Network communications security by applying the latest standards in cryptographic tunnels (e.g., use of VPN and https protocol).	Definition of personal data requests procedures in the defined data-oriented hierarchy.
No personal or organisational details collected during the described analysis and studies will be shared publicly. The deliverables D6.2, D6.6 and D7.9 are public, and the participants data will be duly anonymised before being processed and shared so that it cannot be traced back to the individuals. The deliverables D6.4, D7.4 and D7.6 are confidential, and the participants data will be also duly anonymised before being processed and shared so that it cannot be traced back to the individuals.	Small campaigns of awareness to all the involved users in accessing and processing of confidential data.
Authentication techniques to access the stored confidential data.	Definition of procedures for assuring the rights of the individuals as stated in GDPR.
Implementation of a security audit plan of stored confidential information.	Definition of Information security incident procedure.

9 References

[1] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Sci Data 3 (2016) 160018. https://doi.org/10.1038/sdata.2016.18.

[2] N. Bongaerts, S. Della Chiesa, Research Data Management Lifecycle, Zenodo, 2022. https://doi.org/10.5281/zenodo.6602006.

[3] FAIR Principles, GO FAIR (n.d.). https://www.go-fair.org/fair-principles/ (accessed October 14, 2024).
[4] Data formats for preservation, OpenAIRE (n.d.). https://www.openaire.eu/data-formats-for-preservation/ (accessed October 14, 2024).

10 Annex I: Data management questionnaire

VALORISH

GREEN VALORISATION CASCADE APPROACH OF FISH WASTE AND BY-PRODUCTS THROUGH FERMENTATION TOWARDS A ZERO-WASTE FUTURE

Grant Agreement Number 101135078

Data Management Plan

Questionnaire

Funded by the European Union

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

Main information

Questionnaire Information	
Partner	
Date	
Responsible person in VALORISH for data management in your organization (name and email)	

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	
Brief Description of data	
Purpose of the data	
Types of data	
Form of data	
Origin of the data	
Data utility outside the project	

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).

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.

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?

11 Annex II: Data summary

11.1 WP1

WP/Task	WP1/T1.1. Feedstock analysis, selection and preparation for processing
Partner	PESCANOVA
Brief Description of data	The data of this Task will consist on the collection, analysis and characterisation of the feedstock portfolio selected for VALORISH biorefinery.
Purpose of the data	The data will serve to feed information to almost all tasks of the project to perform valorisation objectives of the biorefinery, and to ensure a significant feedstock composition heterogeneity within the portfolio, representing the fish industry.
Types of data	Experimental data collected from processing the characterization of biomass that will be used as feedstock for valorisation.
Form of data	Text, numeric, images and audio-visual data.
Origin of the data	Almost all data will be novel, obtained in laboratory by analysing the feedstock properties and characteristics. If needed, re-used data will be collected from previous projects, market data, and peer reviewed publications.
Data utility outside the project	Data will be of benefit to seafood processors, product developers, scientific community, consumers, and public in general.

WP/Task	WP1/T1.1. Feedstock analysis, selection and preparation for processing
Partner	ANFACO
Brief Description of data	Data will be collected in the forms described below through the execution of T1.1. Experimental data collected from processing: characterization of waste and by-product biomass from fish industry through different technologies. Observational field data collected from sensor reading during fermentation processes in the biological reactors.
Purpose of the data	Data generation to fulfil the main objectives of ANFACO in the VALORISH project: valorisation of oily fraction from seafood by-products and protein fraction.
Types of data	Experimental data collected from processing the characterization of biomass that will be used as feedstock for valorisation.
Form of data	Text, numeric, images and audio-visual data.
Origin of the data	Almost all data will be novel, obtained in laboratory by analysing the feedstock properties and characteristics. If needed, re-used data will be collected from previous projects, market data, and peer reviewed publications.
Data utility outside the project	Data will be of benefit to seafood processors, product developers, scientific community, consumers, and public in general.

WP/Task	WP1/T1.2. Fish oil extraction techniques

Partner	ANFACO
Brief Description of data	Data will be collected through the execution of T1.2 within the VALORISH project by ANFACO. The data will consist of the results of the research on oil extraction techniques.
Purpose of the data	Data generation to fulfil the main objectives of T1.2: valorisation of fish by-products oily fraction to obtain fish oil. The purpose of data collection also includes the interaction with other WPs of the project.
Types of data	Experimental data collected from the laboratory employing novel oil extraction techniques
Form of data	Mainly, text, numeric, images and audio-visual data.
Origin of the data	The data will be generated.
Data utility outside the project	Data will be of benefit to seafood processors, product developers, scientific community, consumers, and public in general.

WP/Task	WP1/T1.2. Fish oil extraction techniques
Partner	NOFIMA
Brief Description of data	Experimental data collected under controlled conditions of oil extraction process.
Purpose of the data	Research
Types of data	Experimental data obtained as laboratory notes and reports
Form of data	Text and numeric
Origin of the data	Generated
Data utility outside the project	If relevant, the data will be published in peer-reviewed journals, so the data will serve for expanding the scientific knowledge on the field

WP/Task	WP1/T1.3. Fish protein fermentative hydrolysis
Partner	ANFACO
Brief Description of data	Data will be collected through the execution of T1.3 within the VALORISH project by ANFACO. The data will consist of the results of the research on fermentation process development.
Purpose of the data	Data generation to fulfil the main objectives of T1.3: valorisation of fish by-products protein fraction to obtain protein hydrolysates. The purpose of data collection also includes the interaction with other WPs of the project.
Types of data	Experimental data collected from the laboratory employing fermentation of fish by-products.
Form of data	Mainly, text, numeric, images and audio-visual data.
Origin of the data	The data will be generated.

Data utility outside the project	Data will be of benefit to seafood processors, product developers, scientific community, consumers, and public in general.

WP/Task	WP1/T1.4. Downstream processing of the protein hydrolysates
Partner	ANFACO
Brief Description of data	Data will be collected through the execution of T1.4 within the VALORISH project by ANFACO. The data will consist of the results of the research on the downstream of fish protein hydrolysates produced.
Purpose of the data	Data generation to fulfil the main objectives of T1.4: optimization of downstream of protein hydrolysates. The purpose of data collection also includes the interaction with other WPs of the project.
Types of data	Experimental data collected from the laboratory employing separation and purification technologies for obtaining fish protein hydrolysates isolates.
Form of data	Mainly, text, numeric, images and audio-visual data.
Origin of the data	The data will be generated.
Data utility outside the project	Data will be of benefit to seafood processors, product developers, scientific community, consumers, and public in general.

11.2 WP2

WP/Task	WP2/T2.1. Bioproducts production by hydrolysate-based fermentation
Partner	SINTEF
Brief Description of data	Data from cultivation of microbial strains e.g. biomass, fermentation parameters (e.g. temperature, pH, dissolved oxygen, respiration rate) and product yield
Purpose of the data	The data will be generated in relation to the objective of obtaining bioproducts by fermentation using fish protein hydrolysates as substrate and to valorise the side streams from the VALORISH biorefinery (WP2).
Types of data	Experimental data
Form of data	Text, numeric, instrument specific
Origin of the data	The data will be generated within the project
Data utility outside the project	The data generated will be relevant for publication in international journals therefore, will be interesting for scientific community, but also for other stakeholders regarding high-value bioproducts and microbial fermentation.

WP/Task	WP2/T2.2. Downstream activities for bioproducts purification
Partner	SINTEF
Brief Description of	Data from purification of the bioproducts (spectroscopy and mass spectrometry

data	data)
Purpose of the data	The data will be generated in relation to the objective of obtaining separated and purified bioproducts from the VALORISH biorefinery
Types of data	Experimental data
Form of data	Text, numeric, instrument specific
Origin of the data	The data will be generated within the project
Data utility outside the project	The data generated will be relevant for publication in international journals, therefore, will be interesting for scientific community, but also for other stakeholders regarding high-value bioproducts and microbial fermentation.

WP/Task	WP2/Task 2.3
Partner	RINA
Brief Description of data	Calcination process data. Two types of data are generated: raw and processed data.
Purpose of the data	Data process will be produced during the calcination process in order to monitor the operating parameters and establish possible relation with the results. Moreover, datasets will be used to compare results with data references.
Types of data	Experimental data collected under controlled conditions and datasets.
Form of data	Text; numeric; instrument specific; images.
Origin of the data	Laboratory measurement systems; available online database.
Data utility outside the project	Research centres; Hydroxyapatite production sector; side-stream valorisation sector; industries, e.g., energy, food, textiles.

WP/Task	WP2 / Task 2.3
Partner	NOFIMA
Brief Description of data	Experimental data collected under controlled condition for fishbone valorisation research
Purpose of the data	Research in the valorisation of fishbones to extract valuable products
Types of data	Experimental data (Laboratory notes and reports)
Form of data	Text and numeric
Origin of the data	Generated
Data utility outside the project	If relevant, the data will be published in peer-reviewed journals

WP/Task	WP2/T2.4. Side-stream biomass valorisation
Partner	ANFACO

Brief Description of data	Data will be collected through the execution of T2.4 within the VALORISH project by ANFACO. The data will consist of the results of the research on anaerobic digestion processes of residual biomass from fermentation.
Purpose of the data	Data generation to fulfil the main objectives of T2.4: optimization of downstream of protein hydrolysates. The purpose of data collection also includes the interaction with other WPs of the project.
Types of data	Experimental data collected from the laboratory employing separation and purification technologies for obtaining fish protein hydrolysates isolates.
Form of data	Mainly, text, numeric, images and audio-visual data.
Origin of the data	The data will be generated.
Data utility outside the project	Data will be of benefit to seafood processors, product developers, scientific community, consumers, and public in general.

11.3 WP3

WP/Task	WP3/T3.1. Fish oil and fish protein hydrolysates analysis
Partner	NOFIMA
Brief Description of data	Chemical analyses of fish-derived products. experimental data collected under controlled condition
Purpose of the data	Research regarding products analysis for application into food applications
Types of data	Experimental data (Laboratory notes and reports)
Form of data	Numeric
Origin of the data	Generated
Data utility outside the project	If relevant, the data will be published in peer-reviewed journals

WP/Task	WP3/T3.2. Fishbone derived bioproduct analysis
Partner	RINA
Brief Description of data	Fishbone derived bioproduct analysis: • Surface morphology (SEM images) • Microstructure of the powder (TEM images) • Surface chemistry (SEM-EDX images, XRF spectra) • Ca/P ratio (ICP-OES, ICP-MS)

Purpose of the data	• Data process will be produced during chemical and thermogravimetric analysis conducted in different condition of temperature heating rate, atmosphere, final temperature. • Data process will be produced after calcination/autoclave process, analysing the products obtained by SEM and TEM to study the microstructure of the powder.
Types of data	Experimental data collected under controlled conditions (raw data and processed data).
Form of data	Text; numeric; instrument specific; images.
Origin of the data	Laboratory measurement systems.
Data utility outside the project	Research centres; Hydroxyapatite production sector; side-stream valorisation sector; industries, e.g., energy, food, textiles.

WP/Task	WP3/T3.3. Bioproducts from hydrolysate-based fermentation analysis
Partner	SINTEF
Brief Description of data	Data for activity of bioproducts from fermentation
Purpose of the data	The data will be generated in relation to the objective to analyse, purify, formulate and validate the bio-based products.
Types of data	Experimental data
Form of data	Text, numeric, instrument specific
Origin of the data	The data will be generated within the project
Data utility outside the project	The data generated will be relevant for publication in international journals

WP/Task	WP3/T3.4. FPH organoleptic properties enhancement
Partner	NOFIMA
Brief Description of data	Data regarding organoleptic properties enhancement of fish protein hydrolysates and sensory assessments.
Purpose of the data	Research
Types of data	Experimental data collected under controlled condition (laboratory notes and reports).
Form of data	Text and numeric
Origin of the data	Generated

Data utility outside the project	If relevant, the data will be published in peer-reviewed journals

WP/Task	WP3/T3.5. Product formulation and validation
Partner	PESCANOVA
Brief Description of data	Data will consist on the collection of information regarding product formulation and validation activities of bioproducts developed in VALORISH
Purpose of the data	The data generated will be used to confirm the adequacy of the bioproducts tested to be used in food applications
Types of data	Experimental data
Form of data	Text, numeric, images and audio-visual data.
Origin of the data	All data will be generated within the Task
Data utility outside the project	Data will be of benefit to seafood processors, product developers, scientific community, consumers, food applications industries, and public in general.

11.4 WP4

WP/Task	WP4/T4.1. Computational proteolytic bacteria screening
Partner	IDENER
Brief Description of data	The Task will collect data of bacteria from public databases, such as NCBI or UniProtKB. The data will consist of bacterial strains, enzymes, DNA sequences, protein sequences, related to proteolytic activity. Peptide sequences and their related classification and biological activity roles.
Purpose of the data	The data will be used for selection of bacteria that can perform proteolytic fermentation with satisfactory performance, using the feedstock properties provided in T1.1. Selection will be also constrained to those that, as a result of the protein degradation, produces peptides of biological interest.
Types of data	The data will consist of canonical, derived/compiled data generated from existing datasets, datasets and simulation types: bacterial strains, enzymes, DNA sequences, protein sequences, related to proteolytic activity and theoretical peptides resulting from the proteolytic degradation.
Form of data	Most data will be collected in form of gene sequences, protein sequences, peptide sequences, lists, csv files, cleavage matrixes and proteolytic protein profiles in the form of Hidden Markov Models.
Origin of the data	The data will be collected from public databases, like NCBI, UniProtKB, MEROPS, etc. Such data comes from proved proteolytic activity from scientific publications, together with estimations based on computational calculations. The data will be in form of text, numeric,

	and computer coding.
Data utility outside the project	The creation of the capacity of selecting specific bacteria for proteolytic activity can be employed for many transversal biotechnology activities, where the bacteria have to hydrolyse a biogenic feedstock. This is a really useful know-how that may help many biotechnology organizations to approach biomass treatment in a very efficient way, having information of strain to use, hydrolysis yield and subproducts.

WP/Task	WP4/T4.2. Development of genome-scale metabolic models for bioprocessing routes
Partner	IDENER
Brief Description of data	The Task will focus on the use of Genome-Scale Metabolic models (GEMs) to evaluate the metabolic potential of the described microorganisms for using fish protein hydrolysates as a substrate for production of the target bioproducts. We will utilize already semiautomatically generated models by the Carveme protocol. When needed. additional GEMs will be developed for specific selected bacteria to be included in the GEMs compilation. to that purpose, genomic and metagenomic data will be obtained from public databases inspection and/or from experimental partners.
Purpose of the data	Data will be utilised to analyse, create, develop, optimise and refine GEMs that will be used for designing the best approach for the fermentation process using protein hydrolysates as substrate.
Types of data	Experimental data collected under controlled conditions (genomic and metagenomic), simulation data (generated by the models including growth rate, production yield, metabolic fluxes, gene essentiality), and derived/compiled from existing datasets and reference datasets: genomic sequences and annotation, metagenomic data.
Form of data	Data will appear in form of text, numeric and will also as computer coding to display models (file types examples: .faa, .gbk, .xml, .csv, .doc; .m, .py., mat)
Origin of the data	Experimental data will come from experimental partners of the consortium, database-derived/compiled data will come from public databases (NCBI, UniprotKB, EMBL genomes) and open-access genomic/metagenomic information, and simulation data will come from models generated in the task itself.
Data utility outside the project	The GEMs are useful for any bioproduction process where the metabolic fluxes can be analysed, and enough information of genomic/metagenomic data is available to analyse the specific process from substrate to products, enabling analysing any bioprocess by computational force, and simplifying the design of the process. This kind of tasks are starting to be present in large number of bioprocess development, showing overwhelming usefulness at the early stages of the development.

WP/Task	Task 4.3. Development of an integrated process model

Partner	IDENER
Brief Description of data	In this task, mathematical models will be developed. They will be fed by experimental tasks (WP1, WP2, WP3 and WP5), and the models will generate new data from simulation. Moreover, the generated data from models will feed other tasks in the project (WP4, WP5 and WP6).
Purpose of the data	The data generated in this Task aim to create knowledge on the processes of the biorefinery, to support the design of bioprocesses, assess them, and to help the experiments to be optimized. Specifically, the data of this Task will help to understand the processes, develop the optimization strategy, and support the basic engineering, upscaling activities and assessment activities.
Types of data	Experimental (from other WPs), Simulation (generated in this Task) and reference data (from scientific literature and public databases) will be employed in this Task.
Form of data	Most of the data, experimental, modelled and reference, will be numeric in the form of tables, databases, lists, counts or measurements. There will be also some data in form of computer code.
Origin of the data	Some data involved in this task will be reused from experimental results from WP1, WP2, WP3 and WP5, and reference data. And the data obtained from the mathematical models developed, will be generated within the Task.
Data utility outside the project	The data obtained from this task will be useful to create and develop models that create knowledge around bioprocess modelling, which has transversal utility in process modelling, a technical capacity that is widely used in research and industry.

WP/Task	Task 4.4. Multidisciplinary design optimisation
Partner	IDENER
Brief Description of data	In this task, a multidisciplinary design optimization is applied for the VALORISH process. Within the Task, simulation data will be created by using the mathematical models of T4.3 and reference data from scientific literature.
Purpose of the data	The data generated in this Task aim to obtain a set of parameters that are optimized for increasing the performance of the core processes of the biorefinery of VALORISH, in terms of process yields, environmental performance and costs. The data will be delivered to be used by experimental partners, but also for engineering and assessment activities, aiming to enhance the biorefinery performance in a multidisciplinary point of view.
Types of data	Experimental (from other WPs), Simulation (from T4.3 and generated in this Task) and reference data (from scientific literature and public databases) will be employed in this Task.
Form of data	Most of the data, experimental, modelled and reference, will be

	numeric in the form of tables, databases, lists, counts or measurements. There will be also some data in form of computer code.
Origin of the data	Some data involved in this task will be reused from experimental and simulation results from WP1, WP2, WP3, WP4 and WP5, and reference data. And the data obtained from the mathematical models developed, will be generated within the Task.
Data utility outside the project	The data obtained from this task will be useful to create and develop models that create knowledge around bioprocess modelling, which has transversal utility in process modelling, a technical capacity that is widely used in research and industry.

11.5 WP5

WP/Task	WP5/T5.1. Conceptual design and engineering
Partner	IDENER
Brief Description of data	The data will be the result of the generation of the conceptual design and basic engineering documentation of the core processes of the biorefinery of VALORISH.
Purpose of the data	Within this task, IDENER will use the data from the results from the holistic mathematical model and the optimisation results into a process design. IDENER will elaborate documentation constituting the conceptual and the basic engineering. The data will serve to be used in the implementation of the pilot-scale units of VALORISH and for the assessment and certification/standardization activities.
Types of data	The data generated in this task will be simulated using standard chemical engineering guidelines for process design.
Form of data	The data will come in form of text, numeric, images and technical drawings.
Origin of the data	The data will be generated in this task, however, some data used to develop the documentation will be reused from other WPs (WP1, WP2, WP3, WP4) of VALORISH, and other data will be collected from technical and scientific resources and public databased.
Data utility outside the project	Conceptual process design and basic engineering data are the basic documentation for any industrial deployment. However, for novel biotechnology processes there are not much knowledge or it is protected in IPR terms. Therefore, this king of documentation is very useful in order to generate and share knowledge for novel biotechnology processes deployment.

WP/Task	WP5/T5.2. Deployment and demonstration of pilot-scale units
Partner	BIOTREND
Brief Description of data	Data regarding experimental work on the upscale of VALORISH biorefinery core units: experimental data collected under controlled conditions; simulation data; reference; datasets
Purpose of the data	Process characterisation, development and optimisation. Process modelling and feed data to conceptual engineering design and process assessment and

	feasibility.
Types of data	Sensor readings, spectrophotometric and chromatographic analysis, parameters obtained from the literature.
Form of data	Text, numeric, images, and instrument-specific.
Origin of the data	Generated
Data utility outside the project	Non foreseen within the project duration.

11.6 WP6

WP/Task	WP6/T6.1. Evaluation of the environmental impact
Partner	BLUES
Brief Description of data	Technical data regarding the environmental impact assessment
Purpose of the data	Data needed to address evaluation of the environmental impact
Types of data	simulation data from test models, derived/compiled data generated from existing datasets, references
Form of data	Text, numeric, images
Origin of the data	Generated and re-used (from databases like Ecoinvent)
Data utility outside the project	Environmentalists and workers and researcher in environment related fields

WP/Task	WP6/T6.3. Evaluation of the social impact
Partner	BLUES
Brief Description of data	Technical information on quantitative and qualitative social-related data.
Purpose of the data	Data needed to address evaluation of the social-related studies.
Types of data	Simulation data from test models, derived/compiled data generated from existing datasets, references
Form of data	Text, numeric, images
Origin of the data	Generated and re-used (Social HotSpot database, STATISTA)
Data utility outside the project	Social researchers, ONGs or organizations with interests in social-related analysis.

WP/Task	WP6/T6.4. Social acceptance and perception analysis

Partner	BLUES
Brief Description of data	Technical information on quantitative and qualitative social-related data.
Purpose of the data	Data needed to address evaluation of the social-related studies.
Types of data	Simulation data from test models, derived/compiled data generated from existing datasets, references
Form of data	Text, numeric, images
Origin of the data	Generated and re-used (Social HotSpot database, STATISTA)
Data utility outside the project	Social researchers, ONGs or organizations with interests in social-related analysis.

WP/Task	WP6/Task 6.5
Partner	RINA
Brief Description of data	Data related to process standardisation and assessment, including results from bench tests on oil line, protein line, and side-stream residue valorisation.
Purpose of the data	To evaluate the efficiency of VALORISH solutions in valorising waste streams and to develop process and product standardisation benchmarks.
Types of data	Quantitative data from tests (e.g., carbon footprint, recycled material ratios), process efficiency metrics, compliance with existing standards.
Form of data	Numerical data, reports, KPIs, technical specifications, and possibly metadata describing test conditions and results.
Origin of the data	Generated from bench tests performed on VALORISH solutions; contributions from standardisation bodies and market analysis of existing standards.
Data utility outside the project	Policy makers, public authorities. Useful for establishing new standardisation documents, contributing to ongoing standardisation efforts, and providing benchmarks for market adoption of circular economy processes related to oil and protein valorisation.

11.7 WP7

WP/Task	WP7-Communication, Dissemination and Exploitation • Task 7.1. Communication activities • Task 7.2. VALORISH dissemination • Task 7.3. Exploitation plan and business model
Partner	KNEIA
Brief Description of data	Contact data of individuals outside the consortium interested in VALORISH activities (stakeholder data).
Purpose of the data	The purpose is to provide updates about the project progress, including invitations to participate in VALORISH open activities.

Types of data	Compiled data from website forms (name, email)
Form of data	Text (file format .xlsx)
Origin of the data	Website forms, project generic e-mail.
Data utility outside the project	Not applicable. Data subject to GPD constraints.

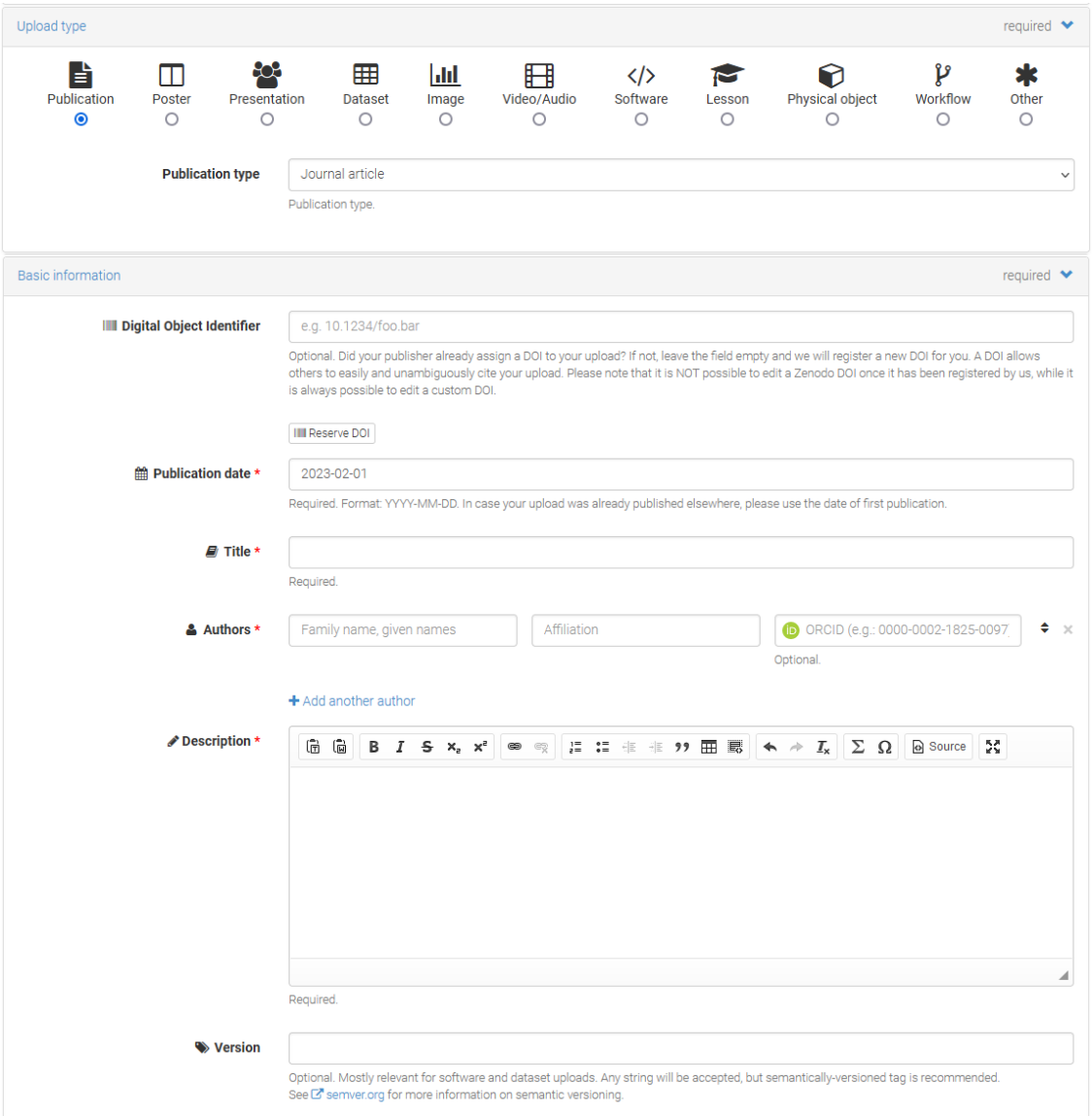
WP/Task	WP7/T7.4. Clustering and networking activities
Partner	BLUES
Brief Description of data	Data of clustering and networking activities.
Purpose of the data	Data needed to address the clustering and networking activities of the project.
Types of data	Derived/compiled data generated from existing datasets, references
Form of data	Text, numeric.
Origin of the data	Generated
Data utility outside the project	Organizations with interest on VALORISH main outcomes.

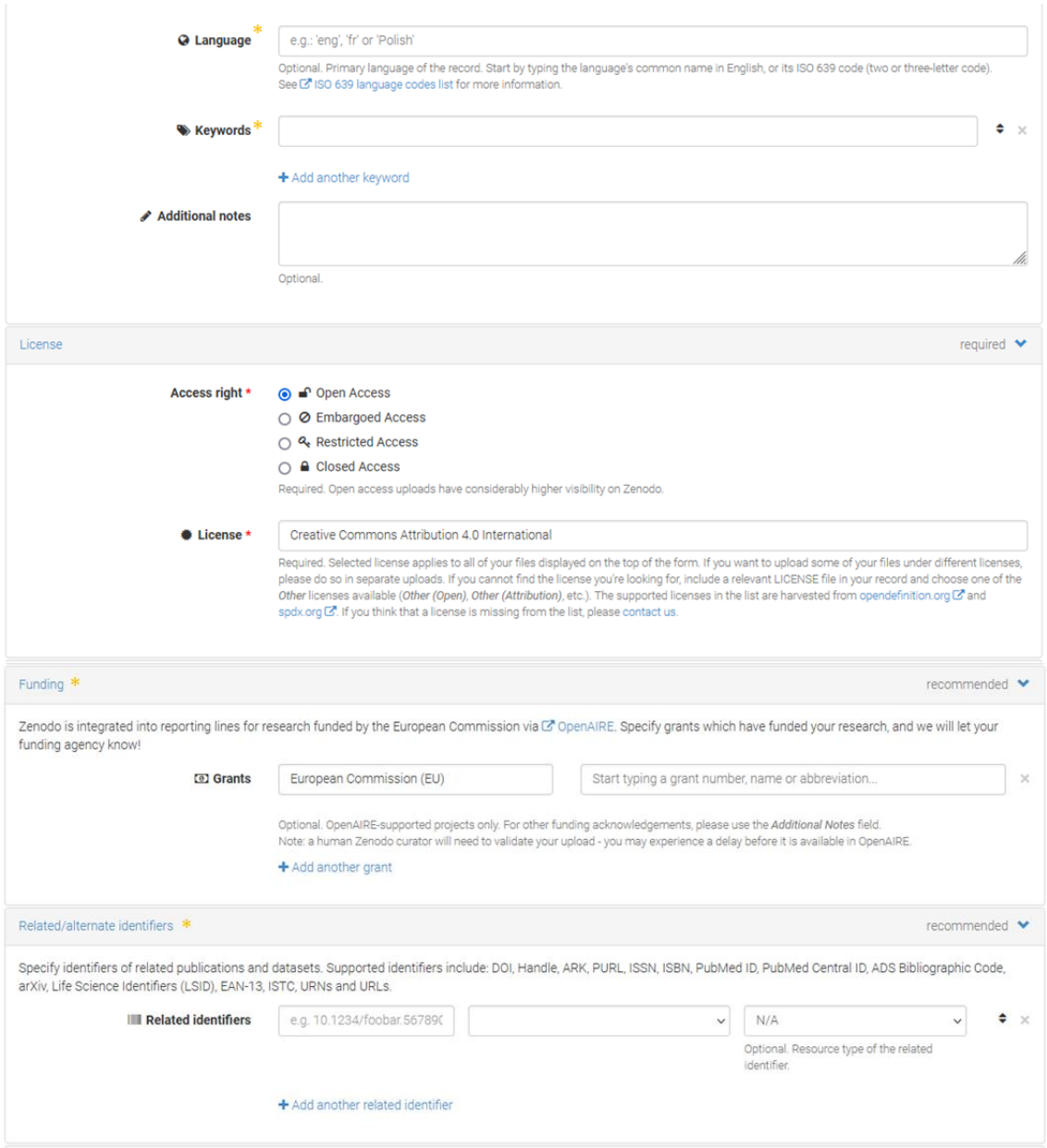
WP/Task	WP7/T7.5. Stakeholders identification and engagement
Partner	BLUES
Brief Description of data	Data of stakeholder management activities.
Purpose of the data	Data needed to address the stakeholder engagement activities of the project.
Types of data	Derived/compiled data generated from existing datasets, references
Form of data	Text, numeric.
Origin of the data	Generated
Data utility outside the project	Organizations with interest on VALORISH main outcomes.

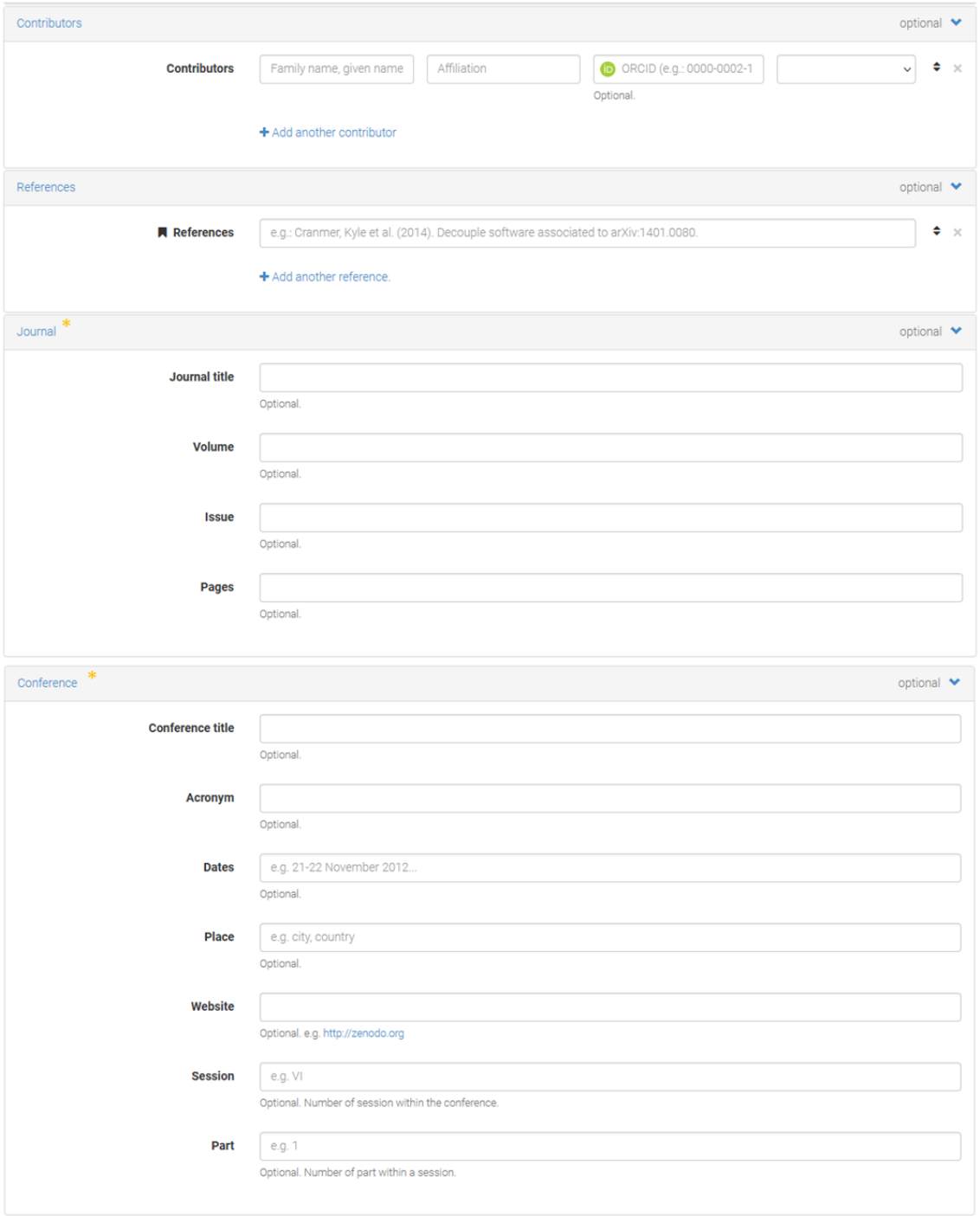
12 Annex III: Metadata required by any Zenodo deposit

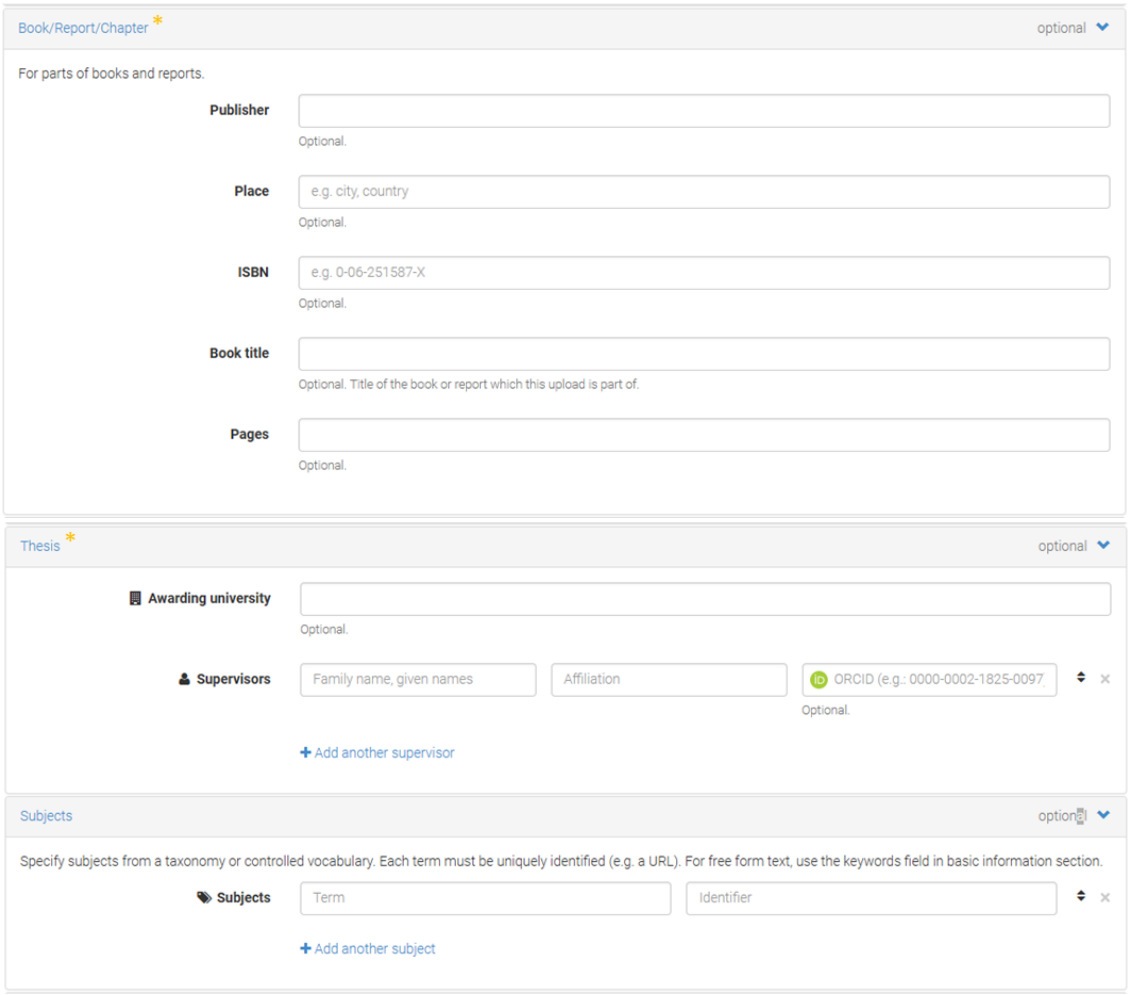
12.1 Minimum metadata for publications

Items marked with * are mandatory for any Zenodo upload; items marked with (*) will be mandatory for any VALORISH 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 *
Source
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.

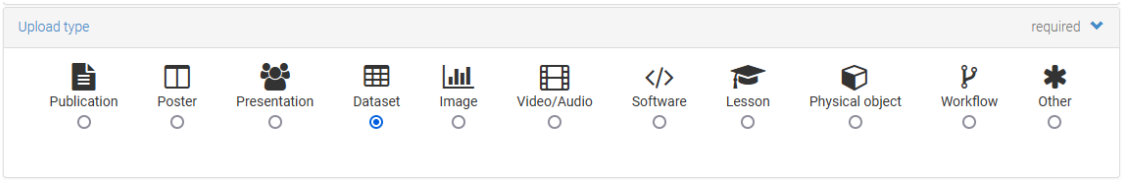

Language *
e.g.: 'eng', 'fr' or 'Polish'
Optional. Primary language of the record. Start by typing the language's common name in English, or its ISO 639 code (two or three-letter code).
See ISO 639 language codes list for more information.
Keywords *
+ Add another keyword
Additional notes
Optional.
License
required
Access right *
Open Access
Embargoed Access
Restricted Access
Closed Access
Required. Open access uploads have considerably higher visibility on Zenodo.
License *
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 *
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
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

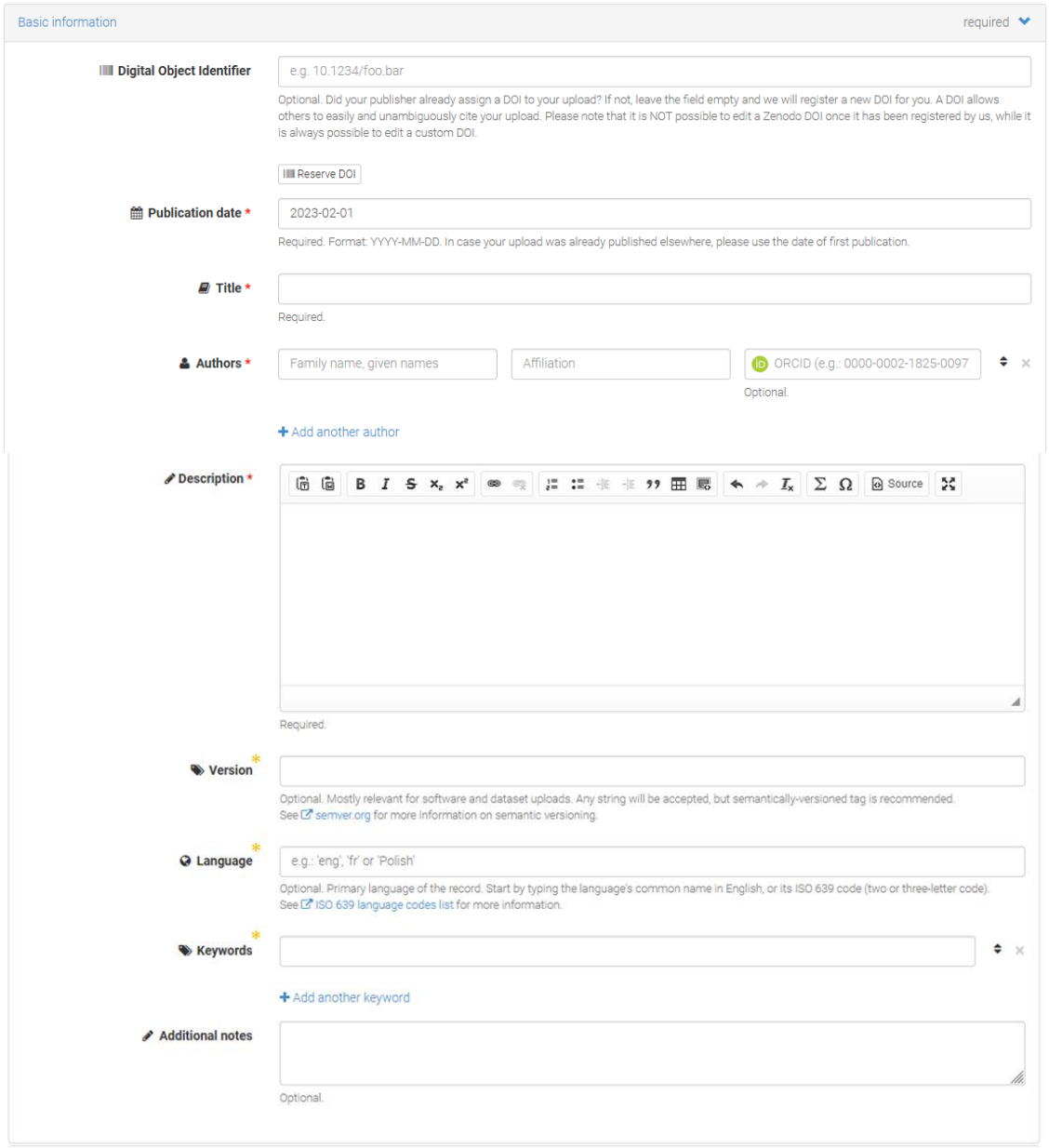

Contributors
optional
Contributors
Family name, given name
Affiliation
ORCID (e.g.: 0000-0002-1
Optional.
+ Add another contributor
References
optional
References
e.g.: Cranmer, Kyle et al. (2014). Decouple software associated to arXiv:1401.0080.
+ Add another reference.
Journal *
optional
Journal title
Optional.
Volume
Optional.
Issue
Optional.
Pages
Optional.
Conference *
optional
Conference title
Optional.
Acronym
Optional.
Dates
e.g. 21-22 November 2012...
Optional.
Place
e.g. city, country
Optional.
Website
Optional. e.g. http://zenodo.org
Session
e.g. VI
Optional. Number of session within the conference.
Part
e.g. 1
Optional. Number of part within a session.

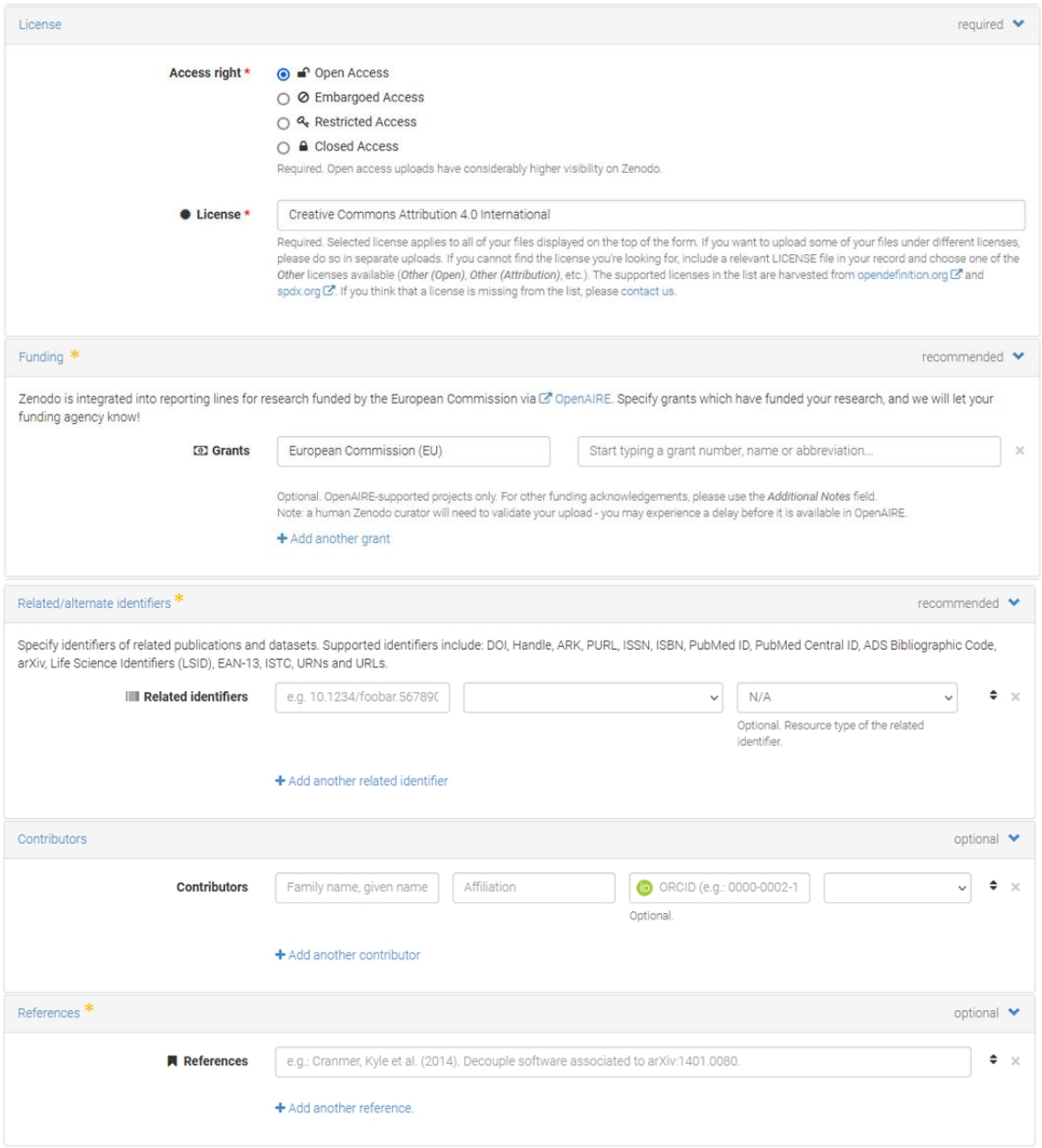


12.2 Minimum metadata for datasets

Items marked with * are mandatory for any Zenodo upload; items marked with (*) will be mandatory for any VALORISH upload.




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 *
Source
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
e.g.: 'eng', 'fr' or 'Polish'
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.



13 Annex IV: Dataset index

PID	Type of PID	Description of dataset	Is this dataset available in open access	URL to repository