Ref. Ares(2026)6245265 - 19/06/2026

D7.3. Data Management Plan

pHYBi
Phytomanagement for a bio-based textile industry

PHYtomanagement as a sustainable feedstock source of lignocellulosic-based high-value BIo-based products for textile applications

Grant Agreement Number 101156439

Deliverable name:	Data Management Plan
Deliverable number:	7.3
Deliverable type:	Report
Work Package:	WP7: Management
Lead beneficiary:	IDENER
Contact person:	Andrea Martos; andrea.dominguez@idener.ai
Dissemination Level:	Confidential
Due date for deliverable:	March, 31st

Circular Bio-based Europe Joint Undertaking

Bio-based Industries Consortium

Co-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 Circular Bio-based Europe Joint Undertaking. Neither the European Union nor the granting authority can be held responsible for them.

DOCUMENT CONTROL PAGE

Author(s):	**Andrea Martos**
Contributor(s):	IDENER Team
Reviewer(s):	All the consortium
Version number:	v.1
Contractual delivery date:	31-03-2025
Actual delivery date:	[dd-mm-yyyy]
Status:	Final

REVISION HISTORY

v.0	17-03-2025	Andrea Martos, Akanksha Mishra, Nuria Ferrera	Creation, First Draft
v.1	25-03-2025	Andrea Martos	Final comments
v.2	10-06-2026	Nuria Ferrera	RP1 revision implemented

ACKNOWLEDGEMENTS

The work described in this publication was subsidised by Horizon Europe (HORIZON) framework through the Grant Agreement Number 101156439.

DISCLAIMER

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 Circular Bio-Based Europe Joint Undertaking. Neither the European Union nor the granting authority can be held responsible for them.

TABLE OF CONTENTS

DOCUMENT CONTROL PAGE 2

REVISION HISTORY 2

ACKNOWLEDGEMENTS 2

DISCLAIMER 2

TABLE OF CONTENTS 3

EXECUTIVE SUMMARY 4

1. INTRODUCTION 6

1.1 DESCRIPTION OF THE DOCUMENT AND PURSUE 6

2. GOOD PRACTICES FOR DATA MANAGEMENT 7

3. DATA SUMMARY 8

5. FAIR DATA 24

5.1 Making data findable, including provisions for metadata 24

5.2 Making data accessible 30

5.3 Making Data Interoperable 41

5.4 Increase the Data Re-use 44

6. ALLOCATION OF RESOURCES 50

7. DATA SECURITY 50

8. ETHICAL ISSUES 50

9. BIBLIOGRAPHY 51

10. ANNEX 52

EXECUTIVE SUMMARY

This document presents the first version of the Data Management Plan (DMP) for the pHYBi project, developed under Horizon Europe Grant Agreement No. 101156439. The DMP establishes the framework for the generation, collection, storage, sharing, protection, and long-term preservation of data produced throughout the project, ensuring compliance with the FAIR principles (Findable, Accessible, Interoperable and Reusable) and the requirements of the European Commission.

The pHYBi project aims to demonstrate sustainable phytomanagement approaches for the production of lignocellulosic biomass from contaminated and marginal lands and its subsequent valorisation into high-value bio-based products for textile applications. Achieving these objectives requires the generation and management of a wide variety of datasets across multiple disciplines, including environmental monitoring, biomass characterisation, process optimisation, material development, sustainability assessment, modelling activities, communication, exploitation, and stakeholder engagement.

This first DMP identifies the main datasets expected to be generated by project partners and describes their purpose, origin, formats, expected size, ownership, accessibility conditions, and potential for reuse. The datasets encompass observational, experimental, simulation, derived, and reference data generated throughout the different work packages. These include soil and environmental characterisation data, geospatial information from field monitoring activities, biomass composition and fractionation data, organosolv process parameters, coating formulation and characterisation data, life cycle sustainability assessment results, modelling and simulation outputs, dissemination materials, stakeholder engagement records, and exploitation-related information.

Beyond describing individual datasets, the DMP also defines how data generated by different partners will be compiled, integrated, and used for joint analyses across the project. Experimental and monitoring data generated in WP1 and WP2 will provide the basis for biomass selection, site assessment, and phytomanagement evaluation. These datasets will subsequently be used by partners responsible for biomass processing and fractionation activities in WP3. The resulting characterisation and process data will then feed modelling, optimisation and upscaling activities, enabling the development of integrated process models and decision-support analyses. In parallel, technical, environmental, and socio-economic datasets will be combined to perform sustainability assessments, evaluate project performance, and support exploitation and replication strategies.

To facilitate this cross-partner collaboration, common procedures for data documentation, metadata generation, version control, and storage are established. The consortium will use shared repositories and agreed data management practices to ensure consistency, traceability, and accessibility throughout the project lifecycle. Data quality assurance measures will be implemented at partner level and coordinated centrally to support reliable data exchange and reproducibility of results.

The DMP further defines provisions related to data security, intellectual property protection, ethical compliance, and open access. While certain datasets may remain restricted due to confidentiality, commercial sensitivity, or intellectual property

considerations, the consortium is committed to maximising the availability and reuse potential of project results whenever possible.
As a living document**, this DMP will be periodically updated** throughout the project to reflect the evolution of research activities, the creation of new datasets, changes in data management procedures, and lessons learned during implementation. Future versions will provide additional details regarding data integration workflows, repository selection, publication strategies, and long-term preservation measures, ensuring that project data continue to generate value beyond the lifetime of pHYBi.

1. INTRODUCTION

1.1 DESCRIPTION OF THE DOCUMENT AND PURSUE

The pHYBi project's Data Management Plan (DMP), essential for its execution and mandated by the European Commission for projects taking part in the Horizon Europe framework, is presented in this report within the Grant Agreement. Task 7.3: Data Management is related to Deliverable 7.3, part of WP7: Management. This work will continue throughout the project.

Because of the intense research effort, a large amount of data will be generated during the pHYBi period. In order to handle the data effectively and produce data that is findable, accessible, interoperable, and reusable (FAIR), the data should be gathered in accordance with certain guidelines. Based on the template supplied by the European Commission, the responsible partner (IDENER) developed a questionnaire to gather data regarding the datasets to achieve this goal. Datasets templates are included as Annex.

The report is a living document that will be updated iteratively throughout the project's implementation in months M6 (March 2025, first iterations), M24 (September 2026), and M48 (September 2028). It contains the primary guidelines and initial basis for constructing a final and robust DMP.

2. GOOD PRACTICES FOR DATA MANAGEMENT

The components mentioned in the EU's data management guide (1) must be considered in the project development framework. Setting the proper guidelines for data management and being aware of the data generated by each partner within the pHYBi consortium is crucial for the correct execution of the project. Because of this, the information gathered in the tables below is crucial. Following a standard for data identification, selecting a version management system, or establishing a place or platform where the data may be accessed are just a few of the requirements that will be gathered in this document for the comprehension of all project participants.

One of the main objectives of data management is to make the data available to the public (OA). The two main elements of open access are publishing articles in repositories and granting them unrestricted access. When producing data or transferring files, it is essential to refer to the platform rules and recommendations supplied by the European Union to comply with the proposed standard.

The experimental activity in pHYBi will generate a large amount of data (lab test, surveys, modelling and simulation, etc.). These data will be converted into formats like tables or papers that are easier to read and use. IDENER will be responsible for making sure that the data generated adheres to the FAIR principles. Thus, the pHYBi project will address the basic information categories mentioned in Table 1:

Table 1. Type of data potentially appearing in pHYBi

OBSERVATIONAL	Gathered during the demo activities and in real-time (field data, operator, laboratory notes, images, etc.) (WP1, WP2, WP3, WP4)
EXPERIMENTAL	Collected during the tests for lignocellulosic biomass and in the fields (WP1, WP2, WP3).
SIMULATION	Model developed in WP4 will generate simulation data (prototypes, schemes, scripts, etc.). The data appearing from this activity will be replicable.
DERIVED	Reproducible derived data is created using pre-existing datasets. Data from experimental datasets can be utilized as inputs for modelling, techno-economic and environmental assessments, data computation and interpretation, data analytics, data visualization, outcomes analysis, and more. Additionally, the collaboration modifies data from its devices using internal coding.
REFERENCE/ CANONICAL	Reference or canonical data is a collection of published and/or curated (peer-reviewed) datasets, waste codes (EU/Basel), and data sets. Together with experimental data, these types of data will be used as input for modelling tasks (WP4) and sustainability evaluations (WP5).

3. DATA SUMMARY

The research and dissemination activities (publications, papers, conferences, posters, presentations, flyers, and website material) will be led by the data in the following tables. For the selected stakeholders (academics/students, scientists, society, industry, and policymakers), public information and publishable data will be accessible. This initial identification seeks to advance European scientific understanding and encourage dependability in the project outcomes. This will be accomplished after standards are made open and transparent. As agreed upon in the Grant Agreement, these tables will be updated on the dates specified in this document's Introduction during the project's implementation.

Table 2. IDENER Data Summary

IDENER Dataset 1	
Identifier	VRP_Model_vX_Date
Dataset description	Information about the model for soil and biorefinery scheme
Purpose of the data	The data generated by IDENER and collected from the technical partners will serve for the construction of the model
Type of data	Data generated from the simulation of the upscaling of the process
Form of the data	Text, numerical, pictures, presentations, website
Format of the data	.xlsx, .docx, .pptx, .jpeg, .jpg, .png, .txt, .data -- These are some of them, to be more defined during the project implementation
Origin of the data	
Data Utility	The data generated during pHYBi from our side will be useful for the activities of other partners, and has the potential to be interesting for end-users
Re-use of existing data	TBD
Data set is:	Expected to be increased and potentially publishable
Size of data	From kB to GB

Table 3. NTT Data Summary

NTT Dataset 1	
Identifier	ATR spectra
Dataset description	Absorbance (Y, arbitrary unit) as a function of wavelength (X, cm^{-1})
Purpose of the data	Identification of the presence of the coating on the fabric
Type of data	Raw spectra, processed data (ATR correction, Baseline correction, Smoothing), Figures
Form of the data	Images (figures), XY column text, LabSolutionIR Spectra
Format of the data	.txt; .xlsx; .irs; .ispd; .pdf

Origin of the data	Generated by micro-ATR analysis
Data Utility	Scientific community; industries
Re-use of existing data	Use of existing ATR spectra to compare them to the ones that we will obtain from our analysis
Data set is:	Not yet created
Size of data	1-10 Gb
NTT Dataset 2	
Identifier	SEM images
Dataset description	Images with microscopic level resolutions
Purpose of the data	Visual identification of the presence of the coating on the fabric
Type of data	Figures
Form of the data	Images (figures)
Format of the data	.png
Origin of the data	Generated by SEM analysis
Data Utility	Scientific community, industries
Re-use of existing data	N/A
Data set is:	Not yet created
Size of data	100-500 Gb
NTT Dataset 3	
Identifier	Coating formulation parameters
Dataset description	Characteristics of the coating formulations
Purpose of the data	Identify the optimal parameters, such as polymer concentration, organic and inorganic solvent used, solvent concentration, temperature, stirring speed, to obtain the most suitable formulation for coating application.
Type of data	Tables and texts
Form of the data	Excel files, text files
Format of the data	.xlsx; .docx
Origin of the data	Defined before the coating formulation production
Data Utility	Scientific community, industries
Re-use of existing data	Use of scientific publication to define the first parameters
Data set is:	Not yet created
Size of data	<1Gb
NTT Dataset 4	
Identifier	Mechanical test
Dataset description	Mechanical tests on coated fabric
Purpose of the data	Evaluation of mechanical characteristics of the coated fabrics
Type of data	Numeric graphic
Form of the data	Graphic, XY column text
Format of the data	.xlsx; .pdf
Origin of the data	Generated by Universal Tensile Machine

Data Utility	Scientific community, industries
Re-use of existing data	N/A
Data set is:	Not yet created
Size of data	1-10 Gb
NTT Dataset 5	
Identifier	LCA Dataset
Dataset description	LCA results generated by SimaPro software
Purpose of the data	LCA assessment of the project processes
Type of data	Numeric data, numeric graphic
Form of the data	Graphic, Excel files, Figures
Format of the data	.xlsx; .pdf; .jpeg
Origin of the data	Generated by SimaPro software
Data Utility	Scientific community
Re-use of existing data	Use of scientific publication data to compare them to the ones that we will obtain from our assessment
Data set is:	Not yet created
Size of data	<1Gb
NTT Dataset 6	
Identifier	S-LCA Dataset
Dataset description	S-LCA results generated by SimaPro software and SHDB
Purpose of the data	S-LCA assessment of the project processes
Type of data	Numeric data, numeric graphic
Form of the data	Graphic, Excel files, Figures
Format of the data	.xlsx; .pdf; .jpeg
Origin of the data	Generated by SimaPro software
Data Utility	Scientific community
Re-use of existing data	Use of scientific publication data to compare them to the ones that we will obtain from our assessment
Data set is:	Not yet created
Size of data	<1Gb

Table 4. SEZ Data Summary

SEZ Dataset 1	
Identifier	Communication and Dissemination Tools and Materials
Dataset description	Project Corporate Identity, Promotional Materials (Flyer, Poster, Roll-Up) and further communication materials
Purpose of the data	Data is generated to maximize visibility and impact of the project and distribute knowledge and project results.
Type of data	Project Logo, Flyer, posters, social media posts, website content
Form of the data	Text files, Presentations, images, graphical designs
Format of the data	.pdf, .png, .jpeg, .docx, .pptx, .html
Origin of the data	Created within the project by the communication team, design agencies and the consortium

Data Utility	Useful for stakeholders to understand project objectives, impact and outcomes
Re-use of existing data	Some elements (e.g., logos, templates) may be reused from institutional branding guidelines. Some datasets (e.g., previous reports or academic papers) may be reused as references
Data set is:	Growing (new materials created over time), Revisable (updates based on feedback)
Size of data	Estimated to be between 500MB - 5GB, depending on visual assets and multimedia files
SEZ Dataset 2	
Identifier	Stakeholder workshops
Dataset description	Participant lists, discussion summaries, audio/video recordings, interactive session results
Purpose of the data	Data is generated to facilitate collaboration, collect stakeholder insights, and support decision-making.
Type of data	Meeting minutes, video recordings, transcriptions, brainstorming session notes, attendance lists
Form of the data	Text, audio, video, interactive notes
Format of the data	.pdf, .docx, .xlsx, .mp4
Origin of the data	Collected during workshops with participants and interactive sessions
Data Utility	Useful for project partners, stakeholders, and researchers to analyse stakeholder input and improve project outcomes
Re-use of existing data	Previous workshop formats and stakeholder lists may be reused for consistency
Data set is:	Growing (new workshop data added over time), Revisable (edits and anonymization for privacy)
Size of data	Estimated to be between 5GB - 20GB, depending on audio/video content
SEZ Dataset 3	
Identifier	IP Management Plan and Exploitation Matrix
Dataset description	Databases on intellectual property rights, exploitation strategies, commercialization roadmaps
Purpose of the data	Data is generated during exploitation workshops to manage intellectual property, define exploitation strategies, and guide commercialization.
Type of data	Text, structured tables
Form of the data	Excel files, text files
Format of the data	.xlsx; .doc; .pdf
Origin of the data	Created within the project by all project partners
Data Utility	Useful for project partners to help with exploitation of results after the project

Re-use of existing data	Some elements may be based on prior IP management frameworks and best practices
Data set is:	Growing (new workshop data added over time), Revisable (edits and anonymization for privacy)
Size of data	Estimated to be between 500MB - 5GB

Table 5. CETIM Data Summary

CETIM Dataset 1	
Identifier	Biomass Characterisation Dataset – Functional and Elemental Analysis (WP1 Task 1.2)
Dataset description	Characterisation of raw biomass through: • FTIR and NMR spectroscopy to determine functional groups. • Elemental analysis to quantify carbon (C), hydrogen (H), nitrogen (N), and sulphur (S) content. • Lignin, cellulose and hemicellulose composition by quantitative acid hydrolysis. Metal presence in biomass by ICP-MS
Purpose of the data	Assess the chemical composition of biomass. Provide a baseline characterisation for further processing and extraction.
Type of data	Spectroscopic and elemental composition data.
Form of the data	Excel files, presentations, lab results, text files, images…
Format of the data	Numerical data (elemental composition percentages, metal presence, biomass composition). (xLS) Graphical data (NMR spectra). (jpeg) Tabular datasets (processed analysis results). (xLS)
Origin of the data	Generated from laboratory-based biomass characterisation experiments of the project samples.
Data Utility	Provides essential composition data for biomass processing for other researchers & technical experts in the field. Supports the optimisation of extraction and fractionation techniques.
Re-use of existing data	No pre-existing datasets; new experimental data is being generated.
Data set is:	Growing (adding data)
Size of data	About 1 GB
CETIM Dataset 2	
Identifier	Component Characterisation Dataset – Lignin, Cellulose, and Hemicellulose (WP3 Tasks 3.2 & 3.3)
Dataset description	Characterisation of lignin, cellulose, and hemicellulose after organosolv biomass extraction, including:

	• FTIR and NMR spectroscopy to determine functional groups. • Elemental analysis to assess composition. • Electron microscopy for morphological analysis. Chromatographic techniques for chemical composition evaluation (quantitative acid hydrolysis) and purity of extracted fractions.
Purpose of the data	To analyse the composition and structural properties of biomass fractions. To optimise the organosolv extraction methodology to improve yields. To support scalability and further processing of biomass components.
Type of data	Spectroscopic, morphological, elemental, and chromatographic data.
Form of the data	• Numerical data (elemental composition, chromatography peak areas). • Graphical data (FTIR, NMR spectra, chromatograms). • Microscopy images (electron microscopy, morphology). Tabular datasets (processed analysis results).
Format of the data	Raw instrument outputs (spectra, chromatograms, microscopy images). Processed data in spreadsheets, reports, statistical summaries.
Origin of the data	Generated from laboratory-based characterisation of biomass fractions.
Data Utility	Provides key insights into biomass fraction composition and structural properties for other experts in the field. Supports the optimisation of extraction methodologies. Facilitates scalability and application of biomass-derived components.
Re-use of existing data	No pre-existing datasets; new experimental data is being generated.
Data set is:	Growing (adding data)
Size of data	About 1 GB
CETIM Dataset 3	
Identifier	Experimental data – Organosolv Process Conditions
Dataset description	This dataset contains experimental conditions and process parameters used in the organosolv extraction of lignocellulosic biomass. It includes details on reaction conditions, solvent composition, temperature, agitation, extraction time, and equipment used.

Purpose of the data	To optimise the organosolv extraction process for enhanced yield and purity of cellulose, lignin, and hemicellulose. To assess the influence of different reaction parameters on biomass fractionation. To enable process scalability and reproducibility for industrial applications.
Type of data	Experimental process data. Reaction parameters and operating conditions. Equipment specifications and operational settings.
Form of the data	Numerical data (temperature, reaction time, solvent ratios, pressure). Textual descriptions (experimental setup, equipment used). Tabular datasets (structured logs of process conditions).
Format of the data	CSV, Excel (XLSX), JSON for structured process logs. PDF, DOCX for experimental reports and protocols. JPEG, TIFF, PNG for images.
Origin of the data	Generated through laboratory-based and pilot plant organosolv extraction experiments. Collected from instrument readings and process monitoring systems.
Data Utility	Provides essential insights into process optimisation for biomass valorisation. Supports scalability and reproducibility for industrial implementation. Enables comparative analysis across different biomass sources and solvent systems.
Re-use of existing data	No pre-existing datasets; this dataset is generated within the project based on experimental trials.
Data set is:	Primary experimental data collected for process development and optimisation.
Size of data	About 1 GB

Table 6. UBU Data Summary

UBU Dataset 1	
Identifier	Sites characterization: environmental context, and physicochemical, biochemical and microbiological soil characterization data
Dataset description	This dataset contains information related to the landscape structure, vegetation types, and ecological conditions (environmental context), as well as physicochemical, biochemical, and microbiological properties of soil samples (including pollutants). The dataset includes parameters such as pH, organic matter content, nutrient concentrations, microbial diversity and activity, enzymatic activity, available heavy

	metals, and pollutants (e.g., total petroleum hydrocarbons). Additionally, a landscape analysis will be performed, considering land use, vegetation structure, and other environmental descriptors.
Purpose of the data	The purpose of the data is to assess the evolution of landscape and soil health/ soil properties during the project. The data is used to evaluate soil conditions, the initial conditions (focusing on pollution), and how the different crops affect this pollutants and other soil-related properties.
Type of data	Primarily numerical (quantitative) data in tables, with some categorical descriptors (land use, vegetation type, etc.).
Form of the data	Spreadsheet files containing columns for different soil properties and sample identifiers.
Format of the data	.xlsx; .csv
Origin of the data	It is original, generated during the site visits and through lab analysis.
Data Utility	Farmers, landowners, municipalities, agronomists and environmental researchers.
Re-use of existing data	No existing data will be reused.
Data set is:	Growing and revisable.
Size of data	Small size (less than 2 GB).
UBU Dataset 2	
Identifier	Geospatial soil mapping data
Dataset description	his dataset includes spatial information related to soil properties, land use, and environmental conditions. It consists of vector and raster layers used in GIS applications. The data will be generated using drones and specialized equipment (e.g., electrical conductivity mapping) and supplemented with freely available satellite images (e.g., Sentinel and Copernicus). Additionally, Google Earth Engine will be used to analyse and process satellite imagery for enhanced monitoring.
Purpose of the data	The goal of this dataset is to create spatial representations of the experimental sites and to calculate/model key characteristics such as NDVI (Normalized Difference Vegetation Index), salinity indexes, moisture levels, and other relevant parameters. This will help identify polluted areas and monitor soil evolution throughout the experimental trials.
Type of data	Geospatial data in vector format (points, polygons, and shapefiles) and raster satellite images.
Form of the data	GIS-compatible files containing soil attribute maps and spatial metadata.
Format of the data	.gpkg, .cpg, .dbf, .prj, .qix, .snp, .shx, .tiff, .jpeg
Origin of the data	Derived from field sampling, dron flights, GIS processing, and satellite imagery.

Data Utility	Geographers, GIS analysts, environmental researchers, land-use planners, municipalities, landowners.
Re-use of existing data	Free satellite images.
Data set is:	Revisable and expandable during the project, as new spatial data is collected and processed from the same or new sites, or new properties/models are developed.
Size of data	Estimated 2-5GB.

Table 7. UMPL Data Summary

UMPL Dataset 1	
Identifier	Site History, soil & poplar data
Dataset description	Soil, poplar and microbiological analysis performed at the French case study
Purpose of the data	To understand how various soil, microbial and poplar properties interact with polluted environments, contributing not only to site remediation but also to the production of biomass suitable for valorisation in bio-based applications.
Type of data	Soil indicators (physico-chemical, microbiological), Plant phenotyping traits (growth, biomass production).
Form of the data	Excel files for soil analysis and plant indicators, Lab reports, images of poplar trees, images of case study site.
Format of the data	.csv or .xlsx for soil data, .jpg for images, .pdf and .docx for reports, .RData for statistical results.
Origin of the data	Historical data of the site and newly collected data from soil samples and poplar analysis taken at the polluted site. Cameras, lab equipment and UAV drones will generate the new data.
Data Utility	Scientific community (readers of publications), R&D environmental scientists, agricultural researchers, or soil remediation experts; stakeholders in charge of polluted site management and polluted wider soil databases
Re-use of existing data	We will re-use existing climate data from available databases We will re-use soil data from previous studies (PHYTOPOP & BIOFILTREE) to compare with current samples to assess changes over time.
Data set is:	All of the following: Growing (new data will be added as more samples are taken), Revisable (adding data and deleting or changing some data), Publishable, fixed (already published).
Size of data	Expected size of 1 GB.
UMPL Dataset 2	
Identifier	Poplar Plant Experiments and Fungal Optimization
Dataset description	Poplar plant phenotyping and growth data, including fungal treatment and amendment experiments for phytoremediation optimization.

Purpose of the data	To select the most appropriate poplar hybrids for phytomanagement and validate their effectiveness in polluted sites through experiments and fungal consortia.
Type of data	Experimental logs, Excel spreadsheets for plant measurements, images of plant growth, research presentations.
Form of the data	Excel files for soil analysis and poplar indicators, lab reports, images of poplar trees, images of case study site.
Format of the data	.csv or .xlsx for soil data and experimental logs, .jpg for images, .pdf and .docx for reports, .RData for statistical results.
Origin of the data	Data generated from controlled pot experiments using poplar hybrids and fungal consortia. Cameras, lab equipment will generate the new data.
Data Utility	Scientific community (readers of publications), R&D soil and phytoremediation researchers, fungal biologists, and environmental consultants.
Re-use of existing data	We will re-use existing poplar phenotype characteristics from Phytowelt to select the best hybrids. We will use fungi isolated in previous projects to assemble a consortium for the optimization experiments. We will re-use amendment efficiency and protocols to test on the selected poplar hybrids.
Data set is:	All of the following: Growing (new data will be added as more samples are taken), Revisable (adding data and deleting or changing some data), Publishable, fixed (already published).
Size of data	Expected size of 1 GB.

Table 8. PHY Data Summary

PHY Dataset 1	
Identifier	Gene Sequence Data
Dataset description	Sequence data of identified genes
Purpose of the data	The sequence data will be evaluated for the reconstruction of genome scale metabolic models
Type of data	Original sequence data (FASTA format), optimized sequence data for heterologous expression, integrated gene sequence in expression plasmid (.cm5 format), synthetic DNA-molecule and then cloned expression plasmid (DNA)
Form of the data	FASTA format, cm5 format; DNA-molecule
Format of the data	.txt; .cm5 ;vcf
Origin of the data	Original sequence obtained from database; optimized sequence generated by codon-harmonizer software
Data Utility	Scientific community
Re-use of existing data	Yes, we will re-use the data, to further improve our bacterial production strains
Data set is:	Revisable (adding data and deleting or changing some data)
Size of data	1MB

PHY Dataset 2	
Identifier	RNA Sequencing
Dataset description	RNA isolated from plant species analysed for specific gene regulation
Purpose of the data	Transcriptomic analysis of poplar to investigate gene regulation
Type of data	RNA Sequence
Form of the data	Clone manger files
Format of the data	.cm5; vcf
Origin of the data	Generated by Next Gen-Sequencing
Data Utility	Scientific community
Re-use of existing data	Insights of transcriptomic regulation can be compared with existing analysis in the same species or others to see possible variances in the poplar variations
Data set is:	Growing (adding data)
Size of data	
PHY Dataset 3	
Identifier	Photometric spectra
Dataset description	Absorbance (y-axis) at different wavelengths (x-axis)
Purpose of the data	Analysis of carotene content
Type of data	Raw spectra, processed data (normalisation), figures
Form of the data	Figures, XY columns t
Format of the data	.xlsx
Origin of the data	Generated by a photometer
Data Utility	R&D
Re-use of existing data	Comparison with previous carotene productions will provide insights of the usage capabilities of hemicellulose
Data set is:	Growing (adding data)
Size of data	
PHY Dataset 4	
Identifier	Fermentation data
Dataset description	Technical data (pH, Temp, stirrer speed, dissolved O_2) of the fermentation course with carotene-producing strains
Purpose of the data	Technical data of the fermentation to investigate the process parameter
Type of data	Raw data (table), plot (XY)
Form of the data	XY columns text, figures
Format of the data	.csv
Origin of the data	Generated by xCubio twin
Data Utility	Internal research team
Re-use of existing data	No
Data set is:	growing
Size of data	2MB

Table 9. UNIOVI Data Summary

UNIOVI Dataset 1	
Identifier	Sites characterization: environmental context; physicochemical, biochemical and microbiological soil characterization data
Dataset description	This dataset contains information related to the landscape structure, vegetation types, and ecological conditions (environmental context), as well as physicochemical, biochemical, and microbiological properties of soil samples (including pollutants). The dataset includes parameters such as pH, organic matter content, nutrient concentrations, microbial diversity and activity, enzymatic activity, available and total heavy metals (e.g., As, Cd, Cu, Hg, Pb). Additionally, a landscape analysis will be performed, considering land use, vegetation structure, and other environmental descriptors.
Purpose of the data	The purpose of the data is to assess the evolution of landscape and soil health/ soil properties during the project. The data is used to evaluate soil conditions, the initial conditions (focusing on pollution), and how the different crops affect this pollutants and other soil-related properties.
Type of data	Primarily numerical (quantitative) data in tables, with some categorical descriptors (land use, vegetation type, etc.).
Form of the data	Spreadsheet files containing columns for different soil properties and sample identifiers.
Format of the data	.xlsx; .csv
Origin of the data	It is original, generated during the site studies (mainly in WP1), in pilot-field experiments in WP2, and through lab analysis and previous reports.
Data Utility	Landowners, municipality and regional government, environmental researchers, environmental consultants.
Re-use of existing data	Some data have been collected from previous reports or studies, especially those related with pollutants distribution.
Data set is:	Growing and revisable.
Size of data	Small size (less than 2 GB).
UNIOVI Dataset 2	
Identifier	Geospatial soil and vegetation mapping data
Dataset description	This dataset includes spatial information related to soil properties, land use, and environmental conditions. It consists of vector and raster layers used in GIS applications. The data will be generated using drones and specialized equipment (e.g., electrical conductivity mapping) and supplemented with freely available satellite images (e.g., Sentinel and Copernicus). Additionally, Google Earth Engine will be used to analyse and process satellite imagery for enhanced monitoring.

Purpose of the data	The goal of this dataset is to create spatial representations of the experimental sites and to calculate/model key characteristics such as vegetation indices (coverage, tree density, canopy), spectral vegetation indices (NDVI, MPRI), spectral soil indices and color composition... This will help identify polluted areas and monitor soil evolution in the existing site (WP1) and also in the experimental trials (WP2)
Type of data	Geospatial data in vector format (points, polygons, and shapefiles) and raster satellite images.
Form of the data	GIS-compatible files containing soil attribute maps and spatial metadata.
Format of the data	.gpkg, .cpg, .dbf, .prj, .qix, .snp, .shx, .tiff, .jpeg
Origin of the data	Derived from field sampling, dron flights, GIS processing, and satellite imagery.
Data Utility	Geographers, GIS analysts, environmental researchers, land-use planners, municipality and regional government, landowners.
Re-use of existing data	Free satellite images and aerial view images taken by our team in the last years.
Data set is:	Revisable and expandable during the project, as new spatial data is collected and processed from the same or new sites, or new properties/models are developed.
Size of data	Several GB, some images could be very heavy

Table 10. DIH-LEAF Data Summary

DIH-LEAF Dataset 1	
Identifier	Data on sustainability assessment and interaction analysis between technical and social dimensions in agro and forestry sector.
Dataset description	Qualitative and quantitative data derived from surveys, interviews and stakeholder meetings, as well as sustainability impact analyses in the agro and forestry sector.
Purpose of the data	Assessment of the sustainability of pHYBi project bridging the technical and social dimensions to improve the acceptance of the project by industry and society.
Type of data	Primary data from surveys and interviews, as well as secondary data from technical reports and previous studies.
Form of the data	Reports in text format, spreadsheets, images, graphs…
Format of the data	.docx; .pdf; .xlsx; .pptx…
Origin of the data	Data collected through interactions and the workshops organized with stakeholders such as individuals, rural associations, research institutions, agro-industrial and forestry organisations.

Data Utility	Researchers, policy makers, agrobusiness and community organisations.
Re-use of existing data	Sectoral reports and public databases focusing on sustainability in both the textile and agro-food industries in order to exploit the knowledge acquired from previous research and interviews.
Data set is:	Revisable (adding data and deleting or changing some data).
Size of data	500 MB

Table 11. PARTICULA Data Summary

PARTICULA Dataset 1	
Identifier	.doc and .pdf files
Dataset description	Documents related to the projects
Purpose of the data	Drafting of deliverables
Type of data	.doc, .pdf and .xls files
Form of the data	Excel files, presentations, lab results, text files, images…
Format of the data	.xlsx; .doc; .ppt; etc…
Origin of the data	Created by us and partners
Data Utility	General public and possibly academia
Re-use of existing data	We will not reuse data
Data set is:	Growing (adding data), Revisable (adding data and deleting or changing some data), Publishable, fixed (already published), …
Size of data	5 GB

4. DATA INTEGRATION AND JOINT ANALYSIS FRAMEWORK

The pHYBi project relies on the generation of multidisciplinary datasets across several work packages, covering environmental monitoring, biomass production and characterisation, biomass processing, material development, modelling activities, sustainability assessment, dissemination and exploitation. While Section 3 provides an overview of the individual datasets generated by each partner, it is equally important to describe how these datasets will be integrated and used collectively to support the project's objectives.

The project follows a progressive data integration approach, where outputs generated in early-stage activities become inputs for subsequent tasks and analyses. Data generated during site characterisation and phytomanagement activities (WP1 and WP2), including soil physicochemical properties, contamination levels, microbial activity, vegetation performance, biomass yields and geospatial monitoring data, will provide the baseline information required to assess the suitability of the selected sites and biomass feedstocks.

These datasets will subsequently be used by partners involved in biomass processing and fractionation activities (WP3). Experimental data related to biomass composition, lignin, cellulose and hemicellulose content, extraction yields, organosolv process conditions and material properties will be combined with the site and biomass characterisation results to optimise biomass valorisation pathways and identify the most promising processing conditions.

The outputs generated in WP3 will provide critical input data for modelling and simulation activities. Process parameters, biomass composition data, extraction efficiencies and material performance indicators will be compiled and transferred to the modelling activities to support process optimisation, scale-up assessments and scenario analysis. In this context, data generated by multiple partners will be harmonised using common formats and metadata descriptions to ensure consistency and comparability.

In parallel, environmental, technical and socio-economic datasets generated throughout the project will be integrated to perform sustainability assessments. Life Cycle Assessment (LCA), Social Life Cycle Assessment (S-LCA) and other sustainability evaluation methodologies will combine experimental, modelling and operational data to quantify environmental impacts, social performance and resource efficiency of the proposed value chains. These analyses will support decision-making regarding technology selection, process optimisation and future industrial implementation.

Joint analyses will therefore involve the combination of experimental, observational, simulation and derived datasets originating from different partners and work packages. Data exchange will be facilitated through the consortium's shared data management infrastructure, ensuring that authorised project partners can access the information required for their activities while maintaining confidentiality and compliance with intellectual property requirements.

To guarantee traceability and reproducibility, all datasets used in joint analyses will be documented through metadata, version control procedures, and standardised naming conventions. Data quality checks will be performed by the generating partner before sharing, while the receiving partners will be responsible for validating the suitability of the datasets for their specific analyses.

This integrated approach (see Figure 1) ensures that data generated throughout the project are not only properly managed and preserved but also effectively combined to maximise scientific value, support cross-disciplinary collaboration and enable evidence-based evaluation of the pHYBi phytomanagement and biomass valorisation pathways.

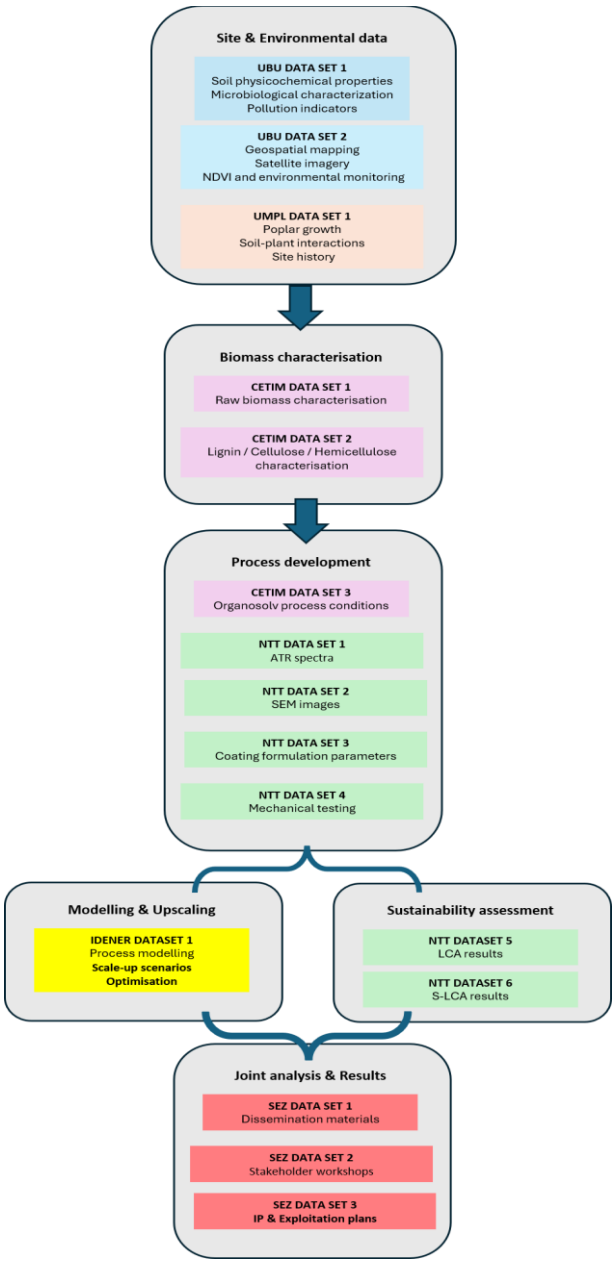


Figure 1: Data Integration and Joint Analysis Framework in pHYBi

5. FAIR DATA

To meet the FAIR data premises, the information gathered from each pHYBi partner in the surveys will be presented in the ensuing subsections (Figure X). It is possible to specifically mention the strategies and tactics for making data findable (both humans and computers should be able to locate the metadata), publicly accessible (once the data is located, users must understand how to access it), interoperable (the data must be compatible with applications or workflows for processing, analysis, and storage), and reusable (FAIR aims to maximise data reuse through thorough, well-described metadata).

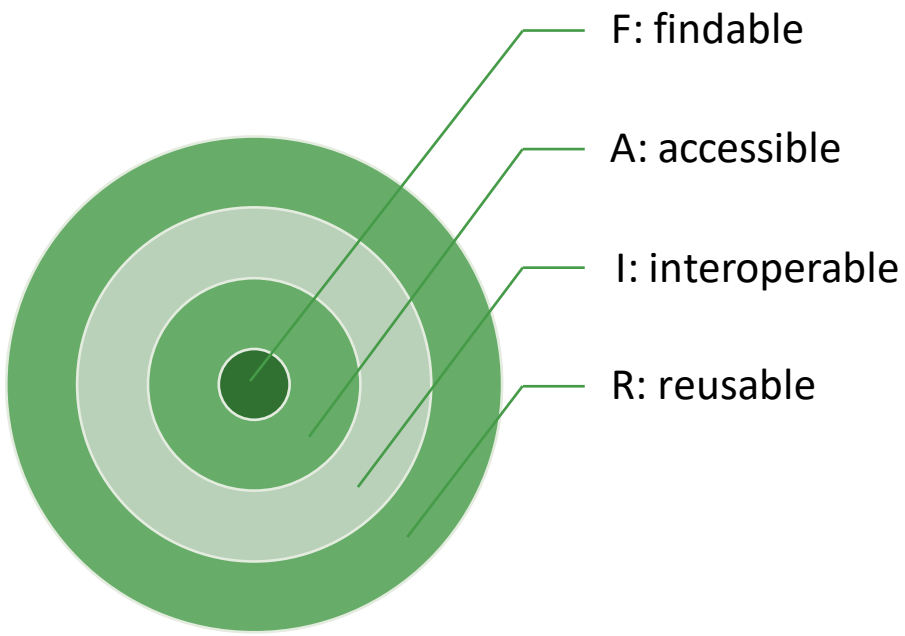


Figure 2. FAIR Data Principles

5.1 Making data findable, including provisions for metadata

The initial tentative approach that the partnerships will take in regard to the previously identified datasets is shown in the following tables. pHYBi will make the datasets accessible in this manner (F of the FAIR principles).

Table 12. IDENER information about "Making data findable, including provisions for metadata"

Partner: IDENER	
What naming conventions do you follow?	pHYBi_WPX_Description_Month-Year_Version --> Example: pHYBi_WP4_ firstversionLargeScale_Jan-2026_V1.0
Will search keywords be provided that optimize possibilities for re-use?	In following versions of the DMP

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. (If Applicable)	The data and results will be documented in a report
Clear versioning approach	N/A

Table 13. NTT information about "Making data findable, including provisions for metadata"

Partner: NTT	
What naming conventions do you follow?	DOI, ORCID
Will search keywords be provided that optimize possibilities for re-use?	NA
What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. (If Applicable)	NA
Clear versioning approach	NA

Table 14. SEZ information about "Making data findable, including provisions for metadata"

Partner: SEZ	
What naming conventions do you follow?	File names include project name, work package, and specific description: e.g. pHYBi_WP6_projectlogo
Will search keywords be provided that optimize possibilities for re-use?	No formal persistent identifier
What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. (If Applicable)	Basic metadata including title, author, creation date, and version history
Clear versioning approach	Version numbers will be maintained for tracking updates internally. Finalised materials will be marked as "final".

Table 15. CETIM information about "Making data findable, including provisions for metadata"

Partner: CETIM	
What naming conventions do you follow?	WPX_NumberX_Date
Will search keywords be provided that optimize possibilities for re-use?	Rich metadata will be provided to ensure the datasets are findable and reusable. The metadata will include details on:

	Sample origin and pre-treatment. Analytical techniques used (NMR, FTIR, elemental analysis, electron microscopy, chromatography). Experimental conditions and parameters (e.g., instrument settings, resolution, calibration details). Data processing and analysis methodologies. File formats and software compatibility.
What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. (If Applicable)	Descriptive metadata (sample information, methodology, instrument details) will be created. IUPAC-recommended nomenclature and metadata conventions for analytical chemistry will be followed where applicable, particularly for NMR, FTIR, elemental analysis, and chromatography. FAIR Data Principles to ensure data is Findable, Accessible, Interoperable, and Reusable. Dublin Core metadata elements for general dataset description.
Clear versioning approach	Metadata will be structured and formatted to ensure it can be harvested and indexed by relevant databases and repositories. The following approaches will be used: Standard metadata formats such as Dublin Core and ISA-TAB will be applied to ensure interoperability with data repositories. Persistent identifiers (e.g., DOIs) will be assigned to datasets to enhance their discoverability and citation.

Table 16. UBU information about "Making data findable, including provisions for metadata"

Partner: UBU	
What naming conventions do you follow?	SiteCode_SiteCharacterization/Index(or)property_WPX_Date_VX.X SiteCode: Is the code for the plot/area characterized. SiteCharacterization: This indicates that the folder includes the complete characterization with the metadata. So, just one file (with several sheets) to include the complete characterization of the soil. Index/Property: NDVI, salinity, moisture, conductivity, etc. In this case we need more than one file per site.
Will search keywords be provided that optimize possibilities for re-use?	Manually selected: Soil health, pollution, microbial activity, heavy metals, total petroleum hydrocarbons, land use, vegetation, enzymatic activity, agronomy, bioremediation, environmental monitoring, GIS, remote sensing, NDVI, soil mapping, Sentinel, Copernicus, Google Earth Engine, geospatial analysis, land-use change, environmental monitoring.
What metadata will be created? In case metadata	For the Site Characterisation: The first sheet of the file will include all the metadata: • General: Dataset title, description, authors, affiliations…

standards do not exist in your discipline, please outline what type of metadata will be created and how. (If Applicable)	• Technical: Abbreviations explanations regarding the info and properties included in the file, with the analysis methods and units (when necessary). We will try to follow the Dublin Core standard. For the Geospatial data: We will try to follow the ISO 19115 standard for geographic data where applicable. Including coordinates, data collection method, (pre)processing steps, author, affiliation, etc. Satellite images will maintain the original metadata.
Clear versioning approach	The versions will be identified in the tittle of the file. We will try to use the format of Zenodo, in order for the data to be easily indexed and reused.

Table 17. UMPL information about "Making data findable, including provisions for metadata"

Partner: UMPL	
What naming conventions do you follow?	Yes, the naming structure includes the project acronym, the year of data collection, the dataset type, a unique identifier and the version if applicable. For example: pHYBi_2025_Biomass_Measurements_WP1_v2.xlsx
Will search keywords be provided that optimize possibilities for re-use?	Yes, e.g. "phytoremediation" "TE polluted soil" "endophytic fungi" "microbial communities" "poplar hybrids"
What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. (If Applicable)	Rich metadata will be provided to ensure that datasets are easily discoverable and reusable. Descriptive metadata: Information about the dataset's content, objectives, and relevance to the project (e.g., type of data, sampling methods).
Clear versioning approach	We will adopt a clear versioning approach to manage updates and ensure dataset integrity over time. Each dataset version will be clearly labeled with version numbers (e.g., v1, v2, v3) and previous versions will be archived for reference. Metadata updates will be tracked, and the changes will be documented in a version control system.

Table 18. PHY information about "Making data findable, including provisions for metadata"

Partner: PHY

What naming conventions do you follow?	Date(YYYYMMDD)_project acronym_identifier
Will search keywords be provided that optimize possibilities for re-use?	We don't use data having ocr identification, therefor the focus will be on an accurate file name
What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. (If Applicable)	Data name; date of creation; date of changes; name of creator, all metadata is not directly created by us but is generated by software
Clear versioning approach	If distinct changes are being made, the older version will be saved separately

Table 19. UNIOVI information about "Making data findable, including provisions for metadata"

Partner: UNIOVI	
What naming conventions do you follow?	SiteCode_SiteCharacterization/Index(or)property_WPX_Date_VX.X SiteCode: Is the code for the plot/area characterized. SiteCharacterization: This indicates that the folder includes the complete characterization with the metadata. So, just one file (with several sheets) to include the complete characterization of the soil. Index/Property: NDVI, salinity, moisture, conductivity, etc. In this case we need more than one file per site.
Will search keywords be provided that optimize possibilities for re-use?	Manually selected: Soil health, pollution, microbial activity, heavy metals, total petroleum hydrocarbons, land use, vegetation, enzymatic activity, agronomy, bioremediation, environmental monitoring, GIS, remote sensing, NDVI, soil mapping, Sentinel, Copernicus, Google Earth Engine, geospatial analysis, land-use change, environmental monitoring.
What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. (If Applicable)	For the Site Characterisation: The first sheet of the file will include all the metadata: • General: Dataset title, description, authors, affiliations… • Technical: Abbreviations explanations regarding the info and properties included in the file, with the analysis methods and units (when necessary). We will try to follow the Dublin Core standard. For the Geospatial data: We will try to follow the ISO 19115 standard for geographic data where applicable. Including coordinates, data collection method, (pre)processing steps, author, affiliation, etc. Satellite images will maintain the original metadata.

Clear versioning approach	The versions will be identified in the tittle of the file. We will try to use the format of Zenodo, in order for the data to be easily indexed and reused.

Table 20. DIH-LEAF information about "Making data findable, including provisions for metadata"

Partner: DIH-LEAF	
What naming conventions do you follow?	pHYBi_DIH-LEAF_YYYYMMDD_Doc.Name
Will search keywords be provided that optimize possibilities for re-use?	To optimize re-use, datasets will include a curated list of descriptive keywords based on their relevance to enhance the discoverability of the datasets in search engines and data portals.
What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. (If Applicable)	Dublin Core standards will be followed.
Clear versioning approach	Sequential numbering with documentation of changes. Metadata will follow a clear versioning approach, with updates documented and previous versions retained for reference. It will be made available in readable formats to facilitate harvesting and indexing by data aggregators and academic search.

Table 21. PARTICULA information about "Making data findable, including provisions for metadata"

Partner: PARTICULA	
What naming conventions do you follow?	No
Will search keywords be provided that optimize possibilities for re-use?	If needed yes
What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. (If Applicable)	No metadata
Clear versioning approach	No metadata

5.2 Making data accessible

The accessibility of the datasets found in this part (A of the FAIR principles) will be assessed during pHYBi.

Table 22. IDENER information about "Making data accessible"

Partner: IDENER	
Will this dataset generated and/or used in the project be made openly available as the default? If positive, How?	Accessible to the consortium members for some of the outputs of large-scale model. The rest of the data will be studied in order to define the publication or not
What methods or software tools are needed to access the data?	TBD when the software will be selected
If there are restrictions on use, how will access be provided?	Sharepoint (restricted to the project partners)
Timing of re-used data	N/A
Protocols for access	Sharepoint
Restrictions	Model limited to the consortium; Virtual Replication Plan available for stakeholders
Identification	Workers personal accounts in the sharepoint
Committee for accessibility	N/A
Data publicly available	The data publishable will be available in ZENODO or similar (TBC)
Length of time	N/A

Table 23. NTT information about "Making data accessible"

Partner: NTT	
Will this dataset generated and/or used in the project be made openly available as the default? If positive, How?	*We don't have open data repository. Part of the dataset will be published in the Deliverables and potential scientific publications.*
What methods or software tools are needed to access the data?	*NA*
If there are restrictions on use, how will access be provided?	*NA*
Timing of re-used data	*NA*
Protocols for access	*NA*
Restrictions	*NA*
Identification	*NA*
Committee for accessibility	*NA*

Data publicly available	*NA*
Length of time	*NA*

Table 24. SEZ information about "Making data accessible"

Partner: SEZ	
Will this dataset generated and/or used in the project be made openly available as the default? If positive, How?	Yes, communication and dissemination materials (e.g., brochures, posters) will be made openly available via the project website
What methods or software tools are needed to access the data?	Standard office software (Microsoft Office, Adobe Reader) and web browsers.
If there are restrictions on use, how will access be provided?	No restrictions
Timing of re-used data	Materials are released as soon as they are finalised and checked by the consortium.
Protocols for access	Available via project website with open access.
Restrictions	Internal working documents may be restricted to project partners only.
Identification	No special identification needed for public materials.
Committee for accessibility	Not needed
Data publicly available	Yes
Length of time	Website will be available for at least 2 years after the project ends.

Table 25. CETIM information about "Making data accessible"

Partner: CETIM	
Will this dataset generated and/or used in the project be made openly available as the default? If positive, How?	The dataset will not be made openly available by default due to confidentiality considerations. The data generated in this project contains valuable insights related to biomass characterisation and process optimisation, which could impact the exploitation and commercial potential of the results. Before any data can be shared, a thorough review process will be conducted to identify and mitigate any risks associated with intellectual property (IP), competitive advantage, and commercial exploitation. Only non-sensitive and non-confidential data may be considered for controlled dissemination, following project consortium agreements and applicable regulations.
What methods or software tools are needed to	Accessing the dataset will require specialised software tools depending on the type of data generated: • Spectroscopic Data (NMR, FTIR): • Standard NMR processing software (e.g., MestReNova, TopSpin,

access the data?	ACD/Labs). • FTIR data viewers supporting JCAMP-DX format (e.g., Essential FTIR, SpectraGryph). • Elemental Analysis Data: • Spreadsheet software (e.g., Microsoft Excel, LibreOffice Calc) for tabular results. • Statistical or data processing tools (e.g., R, Python) for further analysis. • Chromatography Data (HPLC, GC): • Vendor-specific chromatography software (e.g., Chromeleon, OpenLab, ChemStation). • Microscopy Images (Electron Microscopy, Morphology): • Image analysis software such as ImageJ (Fiji) or proprietary SEM/TEM image viewers. • General Data Formats: CSV, Excel, and JSON formats can be accessed with common spreadsheet or programming tools. If structured metadata files are included, XML or RDF processing tools may be required. If data is to be shared, it will be provided in commonly used formats to facilitate accessibility across different platforms.
If there are restrictions on use, how will access be provided?	The data generated in this project is fundamental to the valorisation of lignocellulosic biomass and aligns directly with the objectives of pHYBi, particularly in the extraction, characterisation, and optimisation of biomass components for textile applications and pigment bioproduction. WP1 (Biomass Characterisation Data - NMR & Elemental Analysis): The characterisation of raw biomass through NMR spectroscopy and elemental analysis provides essential information on the functional groups and chemical composition (C, H, N, S content). This data ensures that the selected biomass is suitable for organosolv extraction and helps determine how different biomass sources influence process efficiency and product quality. WP3 (Lignocellulosic Component Characterisation - FTIR, Microscopy, Chromatography, etc.): The dataset supports the optimisation of the organosolv extraction process to enhance the purity and yield of cellulose, lignin, and hemicellulose (KI#6). Detailed FTIR, elemental analysis, electron microscopy, and chromatographic characterisation of the biomass fractions will ensure that cellulose extracted for textile coatings (KI#7) meets the required performance criteria, particularly in terms of structural integrity and chemical purity. The data will also validate the quality of C-5 sugars extracted from hemicellulose for pigment bioproduction (KI#8), ensuring that the biomass-derived sugars can support microbial fermentation without contaminants that could hinder bioprocess efficiency.
Timing of re-used data	If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

	An embargo period of 12 months will be applied to ensure that the generated data does not conflict with the exploitation strategy of the project results. This period allows for a thorough assessment of confidentiality, intellectual property (IP) protection, and potential commercialisation opportunities. The embargo will ensure that no data is publicly disclosed before: Intellectual property protection measures (e.g., patents) are assessed and, if applicable, filed. Publications related to the data are prepared and submitted. The consortium confirms that dissemination does not interfere with exploitation activities. In all cases, the embargo period will not be shorter than what is stipulated in Article 16 regarding the dissemination of project results. After this period, a review will determine whether certain datasets can be made available in a controlled manner while ensuring compliance with confidentiality and commercialisation requirements.
Protocols for access	The data will not be openly accessible by default due to confidentiality and exploitation concerns. Access to the data will be restricted, and any potential sharing will be subject to consortium agreements, intellectual property protection, and exploitation strategies. However, if any non-sensitive datasets are made available, they will follow standardised access protocols, such as: • Repository-based access – Data could be deposited in research data repositories (e.g., Zenodo, OpenAIRE, EUDAT) under specific access conditions (e.g., embargo, restricted access, or licensing). • FAIR principles compliance – Metadata will be structured to allow discoverability even if the raw data remains restricted. • Controlled access upon request – In some cases, data access may be granted upon request, subject to approval and signing of confidentiality agreements (e.g., through institutional data repositories or project-specific platforms). Publication of processed data – Summarised or aggregated data may be included in scientific publications, making key findings available without disclosing sensitive information.
Restrictions	Due to confidentiality concerns and the need to protect exploitation opportunities, access to the data will be restricted both during and after the project. The following approach will be applied: During the Project: Internal access: Data will be shared only among project partners through secure internal platforms (e.g., Microsoft Teams Sharepoint - Email). Controlled external access: Requests for access from third parties will be evaluated on a case-by-case basis, requiring prior approval from the consortium and, if applicable, the signing of Non-Disclosure Agreements (NDAs).

	After the Project: Confidential data remains restricted: Any datasets containing sensitive information related to process optimisation, intellectual property (IP), or commercial exploitation will remain confidential. Publication of non-sensitive data: Selected non-confidential datasets may be made available in scientific publications, reports, or research data repositories (e.g., Zenodo, OpenAIRE) with appropriate licensing conditions. Access under specific agreements: If relevant, external researchers or industry stakeholders may request access through bilateral agreements, ensuring compliance with exploitation strategies and IP protection. The access model will be aligned with the project's data management and exploitation plan, ensuring that results are protected while allowing controlled knowledge dissemination where appropriate.
Identification	During the Project: Access to project data will be managed through secure internal platforms (e.g., Microsoft Teams, institutional repositories, or encrypted cloud storage) where only authorised project partners have access. Authentication will be ensured via: Institutional credentials (e.g., project partners logging in through their organisation's accounts). Multi-factor authentication (MFA) for added security. Role-based access controls to restrict sensitive data to relevant work packages. After the Project: If certain datasets are made available under controlled access, different mechanisms can be used to verify the identity of requesters: Institutional Repositories with Authentication: Data can be hosted on platforms such as Zenodo, OpenAIRE, or EUDAT, requiring users to log in using institutional or ORCID credentials. Data Access Requests & Agreements: External access can be granted upon request, requiring individuals to submit a formal application including their institutional affiliation and intended use of the data. Access would be subject to approval by the consortium and, if necessary, a Non-Disclosure Agreement (NDA) or Data Use Agreement (DUA). DOI-Linked Access with Registration: If certain datasets are deposited in repositories with a DOI (Digital Object Identifier), access could be restricted to users who register and provide verifiable details about their research intentions. This approach ensures secure, trackable, and justified access while protecting sensitive data and allowing for knowledge-sharing under appropriate conditions.

Committee for accessibility	A formal Data Access Committee (DAC) will not be established; however, the project consortium will be responsible for evaluating and approving any data access requests. This approach ensures that data access does not conflict with the exploitation of results, intellectual property protection, or confidentiality agreements among project partners. Access to the data will be subject to: Confidentiality agreements (NDAs) and specific contracts to protect sensitive information. Evaluation of each request by the consortium, considering the proposed use of the data and its impact on the exploitation of results. Access restrictions aligned with the dissemination and exploitation strategy defined in the project's data management plan. This approach guarantees that data is managed securely and in line with the project's interests, allowing for controlled access under conditions that do not compromise commercial exploitation or innovation.
Data publicly available	Yes, metadata will be made openly available and will adhere to FAIR (Findable, Accessible, Interoperable, and Reusable) principles, ensuring that it can be indexed and discovered. However, while metadata itself will be publicly accessible, it may not always be licensed under CC0 (public domain dedication) if it contains references to confidential or commercially sensitive information related to the project's biomass valorisation processes. In such cases, metadata will be made available under a less permissive licence (e.g., CC-BY or CC-BY-NC) to ensure that commercially relevant aspects are protected. Yes, metadata will provide structured information about the datasets, including: Description of the data (e.g., type, format, methodology used). Analytical techniques applied (e.g., NMR, FTIR, chromatography, electron microscopy). Relevant parameters (e.g., instrument settings, sample preparation details). Data accessibility information, specifying whether access is restricted and the procedure to request it (e.g., through the project consortium, via NDA, or a controlled-access repository). If the dataset is not openly available, metadata will not provide direct access to the raw data but will include contact details or request procedures for potential users.
Length of time	The data will be available and findable for a minimum of five (5) years after the end of the project, in alignment with standard Horizon Europe data management practices. This period ensures that project results remain accessible for further research, validation, and potential exploitation. After this period, the availability of the data will be reassessed based on: Exploitation and intellectual property considerations – If the data remains

	commercially or strategically valuable, access restrictions may continue beyond five years. Repository policies – If data is deposited in a long-term institutional or open-access repository, availability may be extended accordingly. Project agreements – Any further extensions or archiving decisions will follow the consortium's data management and exploitation strategy. Metadata will be maintained beyond the availability of the data itself to ensure that records of the dataset remain findable and citable. Metadata will: Remain accessible in public repositories (e.g., Zenodo, OpenAIRE, institutional data repositories). Indicate whether the dataset is still available or archived. Provide references to related publications, reports, or alternative sources of information. Even if the dataset is no longer accessible, metadata will ensure that the research outcomes remain discoverable and reusable within the scientific and industrial community.

Table 26. UBU information about "Making data accessible"

Partner: UBU	
Will this dataset generated and/or used in the project be made openly available as the default? If positive, How?	Properties can be shared through ZENODO, for example, but the entire dataset cannot be openly available. Needs to be anonymized and shared without coordinates and personal data.
What methods or software tools are needed to access the data?	Standard spreadsheet software (e.g., Microsoft Excel, LibreOffice, R, Python (pandas)).
If there are restrictions on use, how will access be provided?	The access to personal data and coordinates should only be provided to member of the consortium. And they should know that not al that can be made public.
Timing of re-used data	The data can be re-used once published in scientific magazines by the authors or by request.
Protocols for access	Open-access anonymized datasets will be available via public repositories with a standardized CC-BY 4.0 license (or equivalent).
Restrictions	Sensitive data may be anonymized or aggregated before sharing.
Identification	For non-sensitive data, no identification required. For sensitive data, just consortium members that need the data, with name, affiliation and use purpose.
Committee for accessibility	Yes, it would be convenient.

Data publicly available	Yes, metadata will be openly available under a CC0 public domain dedication in compliance with the Grant Agreement.
Length of time	Non-sensitive data and metadata will remain accessible permanently in ZENODO

Table 27. UMPL information about "Making data accessible"

Partner: UMPL	
Will this dataset generated and/or used in the project be made openly available as the default? If positive, How?	Yes, datasets generated in the project will be made openly available as the default. Datasets will be uploaded to trusted public repositories such as • general repository : Zenodo (https://zenodo.org/) • general repository : Archive ouverte HAL (https://hal.archives- ouvertes.fr/) • disciplinariy repository : NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) for Nucleotide sequences, • disciplinariy repository : UNITE (https://unite.ut.ee/) for fungal sequences to ensure the data is findable and accessible. Each dataset will be assigned a DOI (Digital Object Identifier) for long-term citation and access. In general, all datasets will be shared openly unless there are specific legal or contractual restrictions, such as: Personal data that may be involved in the project (e.g., sensitive information about site locations is included). Confidential business data (e.g., proprietary data from industry partners or potential commercial patents). In this case, access will be restricted until the intellectual property is either published or protected under patent law.
What methods or software tools are needed to access the data?	The data can be accessed through common tools such as: - Microsoft Excel or CSV readers for spreadsheet data. - Statistical software such as R (for data analysis and visualization). - GIS software like QGIS for spatial data related to site mapping and landscape geomorphology. - Specific plugins or tools may be provided for handling the biomass characterization datasets. Detailed documentation will be provided, including: - Abbreviations - Data dictionaries to explain variable names and units of measurement. Links to relevant open-source code (for software used in data collection or analysis.

If there are restrictions on use, how will access be provided?	Access will be provided through data request forms for restricted datasets. These datasets will be made available under a controlled access protocol for those needing the data for research purposes and who agree to the terms and conditions outlined for use. If there are embargoes (e.g. for publication) or legal constraints, access will be granted once these restrictions are lifted.
Timing of re-used data	If an embargo is applied to give time for intellectual property protection or publication (e.g., related to patent applications or scientific publications), it will be clearly indicated. The typical embargo period will last for 12 to 24 months after data generation, depending on the nature of the data and any ongoing publication or patenting processes.
Protocols for access	The data will be made available through an open access protocol such as DOI-based access via repositories like Zenodo, which follows the OpenAIRE guidelines. Additionally, data will be compatible with standard data access protocols like OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), ensuring that the data can be easily harvested and indexed by other research databases.
Restrictions	If there are restrictions (e.g., due to sensitive data, intellectual property protection), the access to data will be controlled, with the following procedures: - Data access forms will be required to ensure compliance with usage restrictions. Specific data access requests will be reviewed by the data manager or project coordinator.
Identification	To ensure the appropriate individuals are accessing the data, the identity of the person accessing the data will be confirmed via: - Institutional login (for university or research-based access). - Email verification when accessing specific datasets. In some cases, users may need to sign a data access agreement if accessing restricted datasets.
Committee for accessibility	No, the formation of a data access committee is not foreseen.
Data publicly available	Yes, metadata will be made openly available, following the Grant Agreement requirements. Metadata will contain all necessary information to enable users to access the underlying data (e.g., links to repositories, dataset identifiers, descriptions).
Length of time	Metadata will remain accessible indefinitely, ensuring the long-term availability of the data. The data will be archived in a trusted repository, which guarantees the long-term availability of both data and metadata.

Table 28. PHY information about "Making data accessible"

Partner: PHY	
Will this dataset generated and/or used in the project be made openly available as the default? If positive, How?	All necessary data will be provided by us in form of presentations, in direct exchange, or in progress reports
What methods or software tools are needed to access the data?	Sequence manager software; Microsoft Office or other text processing software to read csv files
If there are restrictions on use, how will access be provided?	No restrictions
Timing of re-used data	Poplar varieties might be officially registered, thus specific variety information should be confidential until registration.
Protocols for access	Only persons involved in the project have access with interne access restrictions. The individual access is organized by project responsible person
Restrictions	see protocol for access
Identification	no
Committee for accessibility	no
Data publicly available	Yes, the metadata will be openly available under a CC0 domain dedication, in accordance with the Grant. It will include detailed descriptors to facilitate easy discovery of the data
Length of time	Internal Backup storage for at least 10 years

Table 29. UNIOVI information about "Making data accessible"

Partner: UNIOVI	
Will this dataset generated and/or used in the project be made openly available as the default? If positive, How?	Properties can be shared through ZENODO, for example, but the entire dataset cannot be openly available. Needs to be anonymized and shared without coordinates and personal data.
What methods or software tools are needed to access the data?	Standard spreadsheet software (e.g., Microsoft Excel, LibreOffice, R, Python (pandas)).
If there are restrictions on use, how will access be provided?	The access to personal data and coordinates should only be provided to members of the consortium. And they should know that not all that can be made public.
Timing of re-used data	The data can be re-used once published in scientific magazines by the authors or by request.

Protocols for access	Open-access anonymized datasets will be available via public repositories with a standardized CC-BY 4.0 license (or equivalent).
Restrictions	Sensitive data may be anonymized or aggregated before sharing.
Identification	For non-sensitive data, no identification required. For sensitive data, just consortium members that need the data, with name, affiliation and purpose of use.
Committee for accessibility	Yes, it would be convenient.
Data publicly available	Yes, metadata will be openly available under a CC0 public domain dedication in compliance with the Grant Agreement.
Length of time	Non-sensitive data and metadata will remain accessible permanently in ZENODO

Table 30. DIH-LEAF information about "Making data accessible"

Partner: DIH-LEAF	
Will this dataset generated and/or used in the project be made openly available as the default? If positive, How?	Due to sensitive information contained within the data, certain parts will be subject to restricted access to comply with privacy laws, data protection policy and ethical standards. To access the restricted data, researchers will need to apply through a formal request process, where requests will be reviewed by DIH-LEAF to ensure compliance with legal and ethical guidelines. The openly available portion of the dataset will be deposited in an open access repository.
What methods or software tools are needed to access the data?	To access to data collected by DIH-LEAF, users will need Microsoft 365.
If there are restrictions on use, how will access be provided?	For datasets with usage restrictions, we'll require researchers to request access, which we'll review to ensure compliance. This supports our mission to provide and exploit knowledge from stakeholders and DIH-LEAF's contact network, allowing the data to bridge the technical and social dimensions to improve the acceptance of the project by industry and society.
Timing of re-used data	Publication after project completion.
Protocols for access	Free access through License Creative Commons CC-BY.
Restrictions	During the project, data access is restricted to the research team as well as vetted collaborators on request. After the project, anonymized data will be placed in a data repository with a simplified access process for researchers. Restrictions implemented ensure privacy and respect proprietary information as

	well as data protection policy, being subjected to annual review and potential adjustments.
Identification	Personal account on the software/website used.
Committee for accessibility	DIH-LEAF will review requests for restricted access.
Data publicly available	Not answer yet.
Length of time	For a minimum of 10 years post-publication.

Table 31. PARTICULA information about "Making data accessible"

Partner: IDENER	
Will this dataset generated and/or used in the project be made openly available as the default? If positive, How?	IF agreed by the project board yes, If not then no
What methods or software tools are needed to access the data?	NO
If there are restrictions on use, how will access be provided?	NO restrictions
Timing of re-used data	N/A
Protocols for access	N/A
Restrictions	No restrictions
Identification	N/A
Committee for accessibility	N/A
Data publicly available	N/A
Length of time	N/A

5.3 Making Data Interoperable

The consortium will evaluate the data's interoperability (I of FAIR principles) in this step, which entails putting in place procedures for creating a standardised methodology for the data to guarantee accessibility for all parties.

Table 32. IDENER information about "Making data interoperable"

Partner: IDENER	
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	TBD

Table 33. NTT information about "Making data interoperable"

Partner: NTT	
What data and metadata	The chosen language should have a formal specification, i.e., the language's syntax and grammar are defined in a precise

vocabularies, standards or methodologies will you follow to make your data interoperable?	way. Another requirement is that the knowledge representation language specifications should be shared and accessible so others can read the specifications and learn the language. Finally, in other to support interoperability, the language should be designed to be used in more than one scenario.

Table 34. SEZ information about "Making data interoperable"

Partner: SEZ	
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	Data formats will be widely accepted ones such as PDF, DOCX, PNG, and PPTX to ensure compatibility.

Table 35. CETIM information about "Making data interoperable"

Partner: CETIM	
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	To ensure interoperability, the project will adopt standardised data and metadata vocabularies, formats, and methodologies commonly used in biomass valorisation, analytical chemistry, and bio-based materials research. The following standards will be followed: 1. Data Standards for Specific Analysis NMR & FTIR Spectroscopy: JCAMP-DX (IUPAC standard for spectroscopic data exchange). MestReNova/Bruker TopSpin format for raw NMR data. Elemental Analysis & Morphology Data: CSV, JSON, or XML-based structured formats to facilitate machine readability. Microscopy Data (Electron Microscopy, Morphology Analysis): OME-TIFF; DM4 (Open Microscopy Environment TIFF format)** for interoperability. 2. Metadata Standards & Vocabularies FAIR Principles Compliance – Metadata will be structured to allow discoverability, accessibility, and reuse. 3. Methodologies for Interoperability Machine-readable formats (XML, JSON, RDF) to enable automated data exchange. DOI (Digital Object Identifier for dataset referencing and citation. FAIR-aligned repositories (e.g., Zenodo, OpenAIRE, EUDAT) to ensure long-term accessibility and compatibility with research infrastructures. These measures will ensure that data from biomass characterisation, organosolv extraction, and bio-based material testing is interoperable across different platforms and disciplines,

	facilitating collaboration, reusability, and integration into larger research frameworks.

Table 36. UBU information about "Making data interoperable"

Partner: UBU	
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	We will use the the FAIR data principles. Soil parameters will be reported using standardized scientific units (e.g., mg/kg, µg/L, pH in standard scale). Data will be structured to ensure compatibility with common data analysis tools (e.g., R, Python, Excel, GIS platforms). Terms and expression (regarding properties and microbiological data) will be carefully explained.

Table 37. UMPL information about "Making data interoperable"

Partner: UMPL	
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	We will use standard vocabularies from FAO, ISO guidelines. We will follow the Dublin Core Metadata Initiative (DCMI) standards to describe the datasets in terms of authorship, rights, and content. The data will be structured according to the FAIR (Findable, Accessible, Interoperable, Reusable) principles to maximize data sharing, interoperability, and reuse. This includes providing open access to datasets through public repositories like Zenodo and ensuring datasets are machine-readable.

Table 38. PHY information about "Making data interoperable"

Partner: PHY	
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	Gene ontology (GO) will be utilized to ensure standardized biological annotations. Data management will be aligned with the FAIR principles to enhance usability and accessibility. Where applicable, data will be provided in CSV and XML formats, as these are widely compatible with various software tools.

Table 39. UNIOVI information about "Making data interoperable"

Partner: UNIOVI	
What data and metadata vocabularies, standards or methodologies will you follow to make	We will use the the FAIR data principles. Soil parameters will be reported using standardized scientific units following the SI (e.g., mg/kg, µg/L, pH in standard scale). Data will be structured to ensure compatibility with common data analysis tools (e.g., R, Python, Excel, GIS platforms).

your data interoperable?	Terms and expression (regarding properties and microbiological data) will be carefully explained in metadata.

Table 40. DIH-LEAF information about "Making data interoperable"

Partner: DIH-LEAF	
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	Not answer yet

Table 41. PARTICULA information about "Making data interoperable"

Partner: PARTICULA	
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	N/A

5.4 Increase the Data Re-use

The license and data reuse considerations will be evaluated case by case during the project's execution (R of FAIR principles).

Table 42. IDENER information about "Increase data re-use"

Partner: IDENER	
How will the data be licensed to permit the widest re-use possible?	TBD
When will the data be made available for reuse?	TBD
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?	N/A
How long is it intended that the data remains reusable?	TBD

Table 43. NTT information about "Increase data re-use"

Partner: NTT	
How will the data be licensed to permit the widest re-use possible?	N/A
When will the data be made available for reuse?	N/A

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?	N/A
How long is it intended that the data remains reusable?	N/A

Table 44. SEZ information about "Increase data re-use"

Partner: SEZ	
How will the data be licensed to permit the widest re-use possible?	Data will be published under a Creative Commons (CC BY 4.0) license, allowing re-use with attribution
When will the data be made available for reuse?	As soon as materials are finalized and published on the project website.
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?	Yes, public communication materials will remain available for stakeholders. The data is generated to maximize project visibility, ensure effective communication with stakeholders, and support dissemination activities.
How long is it intended that the data remains reusable?	At least 5 years after the project end.

Table 45. CETIM information about "Increase data re-use"

Partner: CETIM	
How will the data be licensed to permit the widest re-use possible?	Due to the confidential nature of the data and its commercial exploitation potential, it will not be openly available by default. However, if certain non-sensitive datasets can be shared, they will be licensed as follows: Confidential data (e.g., process optimisations, commercial applications) will remain restricted and accessible only under NDAs or specific agreements. Non-sensitive data may be made available under: CC BY 4.0 (allowing re-use with attribution). CC BY-NC 4.0 (for non-commercial use only). Metadata will be openly available under CC0 to ensure discoverability while indicating access restrictions for the datasets. Any external data requests will be evaluated by the consortium to ensure they do not compromise exploitation opportunities.
When will the data be made available for reuse?	The data will not be openly available by default due to confidentiality and exploitation considerations. However, if certain non-sensitive datasets can be shared, they will be made available after a minimum embargo period of 12 months following the end of the project. The release of any data will depend on:

	Intellectual property (IP) protection – Ensuring that patents or commercialisation strategies are not affected. Exploitation strategy – Preventing conflicts with industrial applications. Consortium approval – Each dataset will be reviewed before potential publication. After the embargo, non-confidential datasets may be deposited in open-access repositories with appropriate licensing, while sensitive data will remain restricted.
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?	The data generated in this project is essential for valorising lignocellulosic biomass for textile applications and aligns directly with pHYBi's objectives. The characterisation and analysis data will support the optimisation of the organosolv extraction process and ensure the quality of biomass-derived fractions for targeted applications. Biomass Characterisation (WP1 - NMR & Elemental Analysis) Provides key insights into chemical composition and functional groups of raw biomass. Ensures biomass suitability for organosolv extraction, guiding process adjustments based on feedstock variability. Lignocellulosic Fraction Characterisation (WP3 - FTIR, Microscopy, Chromatography, etc.) Optimises the extraction process to maximise the purity and yield of cellulose, lignin, and hemicellulose. Validates the quality of cellulose for textile coatings, ensuring compliance with industrial requirements. Supports the valorisation of hemicellulose-derived C-5 sugars for pigment bioproduction. By generating high-quality analytical data, the project ensures that the biomass valorisation strategy is efficient, scalable, and aligned with industrial needs, contributing to the development of sustainable bio-based textile solutions.
How long is it intended that the data remains reusable?	The project will generate experimental and analytical data related to biomass characterisation and valorisation, including: Spectroscopic Data: NMR, FTIR (raw instrument files, processed data in CSV, Excel, PDF). Chromatographic and Chemical Analysis Data: HPLC, GC, elemental analysis (instrument outputs, structured data in CSV, Excel, JSON). Microscopy and Morphology Data: Electron microscopy images and structural analysis (TIFF, PNG, JPEG). Processed Data and Reports: Experimental results, statistical analyses, and technical summaries (CSV, Excel, PDF, DOCX). These formats ensure compatibility and reusability across different research and industrial applications.

Table 46. UBU information about "Increase data re-use"

Partner: UBU	
How will the data be licensed to permit the widest re-use possible?	Most of the data could be used under a (CC BY 4.0) license, allowing unrestricted use with proper attribution (this does not include sensitive data).
When will the data be made available for reuse?	After publication.
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?	Maybe scientists, farmers, agronomists, environmental researchers, and policymakers could be interested, for soil health assessments, pollution analysis, and microbiological studies.
How long is it intended that the data remains reusable?	Permanently through public repositories. But private data should be erased after the end of the project (considering the specified in the CA and GA).

Table 47. UMPL information about "Increase data re-use"

Partner: UMPL	
How will the data be licensed to permit the widest re-use possible?	We will release the data under the Creative Commons Attribution 4.0 International (CC BY 4.0) license (commonly used for open-access publications in science and research). This license permits anyone to share, copy, distribute, remix, adapt, and build upon the data, even for commercial purposes, as long as proper attribution is provided to the original creators. This ensures that the datasets generated, which include soil, biomass, and plant phenotyping data, can be freely used by the global scientific community for research and practical applications, including those involved in phytoremediation, biomass valorisation, and environmental remediation.
When will the data be made available for reuse?	The data will be made available after quality control and validation processes have been completed. This includes the collection, cleaning, and processing of the data on site conditions (WP1) and experimental data on plant growth and biomass (WP2). We anticipate that the data will be deposited in trusted repositories such as Zenodo or OpenAIRE for public access by Q1 2028, with the option for immediate availability after publication of research results or upon final validation of data.
Are the data produced and/or used in the project useable by third	Data from both WP1 (site monitoring) and WP2 (phytomanagement) will be made available in open, non-proprietary formats such as CSV for tabular data and GeoJSON for geospatial data, ensuring compatibility with common analysis software and platforms. The datasets will be accompanied by comprehensive metadata, following the Dublin Core Metadata Standard and specific project guidelines for phytoremediation and biomass valorisation.

parties, in particular after the end of the project?	
How long is it intended that the data remains reusable?	We intend for the data to remain reusable indefinitely. The datasets will be deposited in a trusted repository like Zenodo, which ensures data archiving for the long term (up to 25 years or more). The data will be stored in standard, open formats (such as CSV and JSON) to ensure future compatibility with new tools and platforms. We will ensure the integrity and accessibility of the data by performing regular checks and updating the data formats if necessary. Additionally, we will collaborate with the repository to ensure the long-term sustainability and discoverability of the datasets. The project will generate and reuse various types of data, including: - Soil data: soil physicochemical and microbiological properties (CSV, Excel). - Plant phenotyping data: growth parameters, stress responses, biomass measurements (CSV, Excel, JSON, JPEG). - Geospatial data: site history, landscape geomorphology, terrain attributes (GeoJSON, KML). - Biomass characterization: chemical composition and calorific value (CSV, Excel). Experimental data: pot experiments, fungal consortia effects, amendments (CSV, RDATA).

Table 48. PHY information about "Increase data re-use"

Partner: PHY	
How will the data be licensed to permit the widest re-use possible?	The generated data can be made available in the form of a published paper. In cases where novel bacterial strains are developed and patented, these strains can be accessed and utilized under the appropriate licensing agreements
When will the data be made available for reuse?	As soon as patent or varieties are registered
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?	The research team will ensure that the data produced is accessible and usable by third parties, particularly after the project's conclusion. To maximize usability, the data will be provided in standard, non-proprietary formats such as CSV and NetCDF.
How long is it intended that the data remains reusable?	At least 10 Years

Table 49. UNIOVI information about "Increase data re-use"

Partner: UNIOVI	
How will the data be licensed to permit the widest re-use possible?	Most of the data could be used under a (CC BY 4.0) license, allowing unrestricted use with proper attribution (this does not include sensitive data).
When will the data be made available for reuse?	After publication.
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?	Maybe scientists, environmental researchers, and policymakers could be interested, for soil health assessments, pollution analysis, and microbiological studies.
How long is it intended that the data remains reusable?	Permanently through public repositories. But private data should be erased after the end of the project (considering the specified in the CA and GA).

Table 50. DIH-LEAF information about "Increase data re-use"

Partner: DIH-LEAF	
How will the data be licensed to permit the widest re-use possible?	Not answer yet.
When will the data be made available for reuse?	The dataset will be made available for reuse by the research team immediately following the publication of their initial research findings, which is anticipated to be in December 2024.
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?	Not answer yet.
How long is it intended that the data remains reusable?	Not answer yet.

Table 51. PARTICULA information about "Increase data re-use"

Partner: PARTICULA	
How will the data be licensed to permit the widest re-use possible?	N/A
When will the data be made available for reuse?	N/A
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?	No
How long is it intended that the data remains reusable?	5 years

6. ALLOCATION OF RESOURCES

The pHYBi project uploads and evaluates all outcomes (documents, reports, and other information) using a dedicated and private online storage space (MSTeams SharedPoint). Additionally, the collaborative space will function as an internal communication channel.

For public initiatives like OpenAire/EU and Zenodo/CERN, IDE will lead the consortium in choosing the best data management platform in the project's last year. The partnership will be able to disseminate scientific findings about the pHYBi project through this last way of data evaluation.

7. DATA SECURITY

Each organisation's knowledgeable IT system administrators and specialists will securely store and manage the project's digital assets and resources. Some datasets are stored on internal laptops, as observed in the previously described tables; hence, security protocols are already connected to their internal processes. Each partner will take the necessary security measures following their particular companies' standards. On the other hand, IDENER will be in charge of safeguarding the data kept on the platforms of the pHYBi common repository.

Additionally, by using established procedures, pHYBi partners will ensure data security:

- Prevention: To avoid data loss, keep the information in two places.
- Limiting the use of USB and other drives for storage
- Labelling: To maintain the uniformity of the final data set, classify the data methodically.
- Data sharing should adhere to ethical and legal norms, including data security and intellectual property.
- Preservation: Horizon Europe requires that research data be stored and made available for a minimum of ten years following the completion of the project.

8. ETHICAL ISSUES

An ethical dilemma needs to be resolved in the project's data handling. Processing of personal data is expected as a result of performing numerous duties. EU Regulation 2016/679 (GDPR) will control this personal data. The research data management will be conducted in the same way as the consortium's processing of personal data, which will be developed and discussed in detail in subsequent versions of this plan.

9. BIBLIOGRAPHY

1. EU's Data Management strategies https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en
2. European Commission. (2021). Model Grant Agreement for Horizon Europe Actions. Retrieved from https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/grant-agreement/horizon-europe_grant-agreement_en.pdf
3. European Commission. (2021). Guidelines on FAIR Data Management in Horizon Europe. Retrieved from https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/programme/guidance/guidelines-on-fair-data-management_en.pdf
4. European Commission. (2021). Open Science. Retrieved from https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science_en
5. European Commission. (2021). Horizon Europe: Frequently Asked Questions. Retrieved from https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/faq/faq-horizon-europe_en.pdf
6. European Union. (2021). Regulation of the European Parliament and of the Council establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination. Official Journal of the European Union, L 108/1. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2021.108.01.0001.01.ENG

10. ANNEX

9.1 Questionnaires to the partners

DMP Questionnaire

PHYtomanagement as a sustainable feedstock source of lignocellulosic-based high-value BIo-based products for textile applications

Grant Agreement Number 101156439

Partner: [ACRONYM]
Contact person: [Author / Contact person; email]

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 Circular Bio-based Europe Joint Undertaking. Neither the European Union nor the granting authority can be held responsible for them.

1. Data Summary

Please delete the cursive script (it serves as an explanation per line). If you plan to use different datasets, please, use as many tables as datasets you will produce.

Identifier	*Dataset group name*
Dataset description	*Description of the dataset*
Purpose of the data	*What is the purpose of the data generation or re-use and its relation to the objectives of the project?*
Type of data	*What types and formats of data will the project generate or re-use?*
Form of the data	*Excel files, presentations, lab results, text files, images…*
Format of the data	*.xlsx; .doc; .ppt; etc…*
Origin of the data	*What is the origin/provenance of the data, either generated or re-used?*
Data Utility	*To whom might your data be useful ('data utility'), outside your project?*
Re-use of existing data	*Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.*
Data set is:	*Growing (adding data), Revisable (adding data and deleting or changing some data), Publishable, fixed (already published), …*
Size of data	*What is the expected size of the data that you intend to generate or re-use?*

2. FAIR data

2.1 Making data findable, including provisions for metadata

Please delete the cursive script (it serves as an explanation per line). If you plan to use different datasets, please, use as many tables as datasets you will produce.

What naming conventions do you follow?	*Will data be identified by a persistent identifier? Example: Simulation_NumberX_WPX_Date*
Will search keywords be provided that optimize possibilities for re-use?	*Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?*
What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. (If Applicable)	*Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and*

	how.
Clear versioning approach	*Will metadata be offered in such a way that it can be harvested and indexed?*

2.2 Making data accessible

Please delete the cursive script (it serves as an explanation per line). If you plan to use different datasets, please, use as many tables as datasets you will produce.

Will this dataset generated and/or used in the project be made openly available as the default? If positive, how?	*If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*
What methods or software tools are needed to access the data?	*Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open-source code)?*
If there are restrictions on use, how will access be provided?	*What is the purpose of the data generation or re-use and its relation to the objectives of the project?*
Timing of re-used data	*If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.*
Protocols for access	*Will the data be accessible through a free and standardized access protocol?*
Restrictions	*If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?*
Identification	*How will the identity of the person accessing the data be ascertained?*
Committee for accessibility	*Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?*
Data publicly available	*Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*
Length of time	*How long will the data remain available and findable? Will metadata be guaranteed to*

	remain available after data is no longer available?
Only the responsible of the repository: Will the data be deposited in a trusted repository?	
Only the responsible of the repository: Have you explored appropriate arrangements with the identified repository where your data will be deposited?	
Only the responsible of the repository: Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?	

2.3 Making data interoperable

Please delete the cursive script (it serves as an explanation per line). If you plan to use different datasets, please, use as many tables as datasets you will produce.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	

2.4 Increase data re-use

Please delete the cursive script (it serves as an explanation per line). If you plan to use different datasets, please, use as many tables as datasets you will produce.

How will the data be licensed to permit the widest re-use possible?	
When will the data be made available for reuse?	
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?	*What is the purpose of the data generation or re-use and its relation to the objectives of the project?*
How long is it intended that the data remains reusable?	*What types and formats of data will the project generate or re-use?*