Ref. Ares(2026)3411246 - 31/03/2026

D4.1 INTERMEDIATE REPORT ON THE DATABASE CONSTRUCTION

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

Grant Agreement Number 101156439

Deliverable name:	Intermediate report on the database construction
Deliverable number:	4.1
Deliverable type:	Report
Work Package:	WP4: Modelling and Replication
Lead beneficiary:	IDE
Contact person:	Victoria Caballero – Idener Victoria.caballero@idener.ai
Dissemination Level:	Public
Due date for deliverable:	March 31st, 2026

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):	**M. Victoria Caballero**
Contributor(s):	**Miguel Cebrián Aldana**
Reviewer(s):	Andrea Martos
Version number:	v.3
Contractual delivery date:	[31-03-2026]
Actual delivery date:	[31-03-2026]
Status:	Final

REVISION HISTORY

v.0	[9-12-2025]	Victoria Caballero	Creation, First Draft
v.1	[27-02-2026]	Victoria Caballero	Comments/Ready for review
v.2	[18-03-2026]	Victoria Caballero	Incorporated partner review
v.3	[27-03-2026]	[Victoria Caballero /Andrea Martos]	Final version submitted

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 5

1. INTRODUCTION 6

1.1. DESCRIPTION OF THE DOCUMENT AND PURPOSE 6

1.2. WPS AND TASKS RELATED TO THE DELIVERABLE 7

2. REQUIREMENTS AND SPECIFICATIONS 9

2.1. DATA SOURCES AND INPUTS 9

2.2. DATA STANDARDS AND COMPLIANCE 14

3. OVERALL DATABASE ARCHITECTURE 16

3.1. DESIGN PRINCIPLES 16

3.2. DATA FLOW WITHIN THE PROJECT 17

4. DATA MODEL AND STRUCTURE 21

4.1. CONCEPTUAL DATA MODEL 21

4.2. LOGICAL DATA MODEL 25

5. CONCLUSIONS 29

6. ANNEXES 30

I. Soil data collection template 30

II. Plant data collection template 32

III. Extraction data collection template 36

EXECUTIVE SUMMARY

Deliverable 4.1 outlines the intermediate progress on constructing the central database for the pHYBi project. It describes the methods used to collect, organise, and store data generated from field experiments, laboratory tests, and biomass processing.

The database uses a modular structure based on Microsoft Excel templates to gather data from different project activities. These templates contain predefined fields and controlled vocabulary to ensure all technical partners submit consistent and comparable information regarding the pHYBi process. Once collected, the data from various sources are combined into a central repository that complies with standard open data principles.

A key feature of the database is a unique tracking code that follows every sample from field collection to final product testing. This specific code incorporates details such as the site location, sample type, date, and replicate number. Using this shared identifier allows researchers to reliably link soil conditions with plant growth and downstream processing results.

The platform integrates data across three main operational areas to support computer modelling and sustainability assessments. Work package (WP) 1 provides baseline measurements of soil quality, contaminant levels, and initial biomass characteristics. WP2 records information from field plant management trials, including plant growth and physiological responses. Furthermore, WP3 captures operational data from biomass extraction and textile processing steps.

This deliverable represents mid-term work within Task 4.1. The templates described will be completed by the end of the project. The final version will appear in Deliverable 4.4, which presents the complete database.

1. INTRODUCTION

The pHYBi project is founded on an integrated phytomanagement approach aimed at improving soil quality, enabling sustainable biomass production and supporting the valorisation of contaminated or marginal land resources. The project combines field-based experimentation, laboratory analyses and process-oriented assessments to generate robust scientific evidence supporting nature-based solutions for soil restoration and biomass utilisation. Achieving these objectives requires the coordinated collection, integration and analysis of heterogeneous datasets generated across different spatial scales, experimental conditions and technical domains.

Data within pHYBi are produced by multiple partners, each contributing specialised expertise and datasets derived from site investigations, monitoring campaigns, experimental trials and processing activities. This distributed data generation reflects the interdisciplinary nature of the project but also introduces challenges related to consistency, interoperability and traceability. Addressing these challenges is essential to ensure that project outputs can be effectively synthesised, analysed and reused.

The systematic collection and harmonisation of this information are key enablers for the digitalisation of pHYBi procedures. A structured and interoperable data framework underpins the development of digital tools, modelling activities and data-driven analyses, ensuring that project knowledge can be translated into scalable, reproducible and transferable solutions. In this context, Task 4.1 plays a central role in transforming fragmented datasets into a coherent digital asset that supports both internal research activities and long-term knowledge exploitation.

To achieve this aim, extensive coordination was carried out within Task 4.1 through close collaboration between WP leaders and pHYBi partners. This effort resulted in the co-development of standardised data collection templates and a coordinated planning framework for data sharing, validation and integration. These tools define common structures, metadata requirements and quality control procedures, ensuring that data generated by different partners can be consistently integrated into the pHYBi database and aligned with the project's digitalisation objectives and Data Management Plan.

1.1. DESCRIPTION OF THE DOCUMENT AND PURPOSE

This document presents the structure, content and functionality of the pHYBi project database developed under Task 4.1, "Collection of data and database management". It serves as the primary technical and organisational reference for how data generated across WP1, WP2 and WP3 are collected, harmonised, stored and made accessible throughout the project and serves as the reference framework for data collection, harmonisation and access across WP1–WP5. The report outlines the design principles guiding the database, the types of data it hosts, the workflows for data ingestion and curation, and the interfaces that enable users to retrieve and use project data.

The purpose of the document is to provide a comprehensive description of the database architecture and its operational framework, ensuring that partners have a clear understanding of how the system supports collaborative research and cross-WP integration. It specifies how the database aligns with the Data Management Plan (D7.3). The document also explains how indicators of soil and biomass quality are incorporated

and periodically updated to reflect new knowledge emerging from stakeholder engagement and the scientific outcomes of the project.
In order to harmonise inputs coming from field measurements, remote sensing and lab testing from the mentioned WP, sampling sheets to provide to the involved partners were developed. These templates aim to standardise data collection and reporting across all partners, ensuring consistent formats, units, metadata, and quality-control fields so that datasets can be directly compared, merged, and analysed without additional reformatting.
The information collected in the field and laboratory will later be entered into a Microsoft Excel-based template specifically created for this purpose. The template provides a common structure for all partners, including predefined fields, controlled terminology, and agreed units and reference formats, so that measurements and observations are recorded in a consistent and comparable way. Where relevant, it also captures essential metadata (e.g., site identification, sampling date, method/protocol, instrument details, and responsible partner), which improves traceability and facilitates quality control. It also includes an "Associated genomic sample (Y/N)" column to flag whether a given field or laboratory sample has a corresponding genomic sample available for WP2 collaboration. This simple identifier will make it easier to link physicochemical, site, and remediation-relevant measurements with genomic datasets when applicable. In turn, it will support the integration of genomic-based models with remediation modelling workflows whenever joint analyses or model coupling are required.
This association is possible thanks to a project-specific sampling coding system designed to uniquely identify each sample and preserve its traceability across all activities. This coding approach allows data generated in the field and in the laboratory to be consistently related and tracked throughout the entire pHYBi process, ensuring that measurements, analyses, and derived results can be linked back to the same sampling entity. In practice, it not only facilitates collaboration with WP2 by enabling reliable matching between environmental measurements and associated genomic samples, but also supports a holistic mathematical modelling approach by integrating heterogeneous datasets into a single coherent framework.

1.2. WPS AND TASKS RELATED TO THE DELIVERABLE

Deliverable 4.1 is directly linked to the activities carried out in WP4 and draws on scientific and technical inputs from several work packages. The database described in this report acts as the central repository for the project's research outputs and therefore supports the integration, analysis and long-term stewardship of data generated throughout pHYBi. WP4, through Task 4.1, is responsible for establishing, developing and maintaining this database, bringing together datasets from multiple sources, defining the database structure and implementing the procedures for data curation and alignment with the Data Management Plan under T7.4. This database will also be used across other WP4 tasks, particularly Task 4.2, to support model development, parameter fitting and model validation. IDE leads these efforts with contributions from all technical partners.

The data generated in WP1 and WP2 provide essential inputs to the system, including soil quality indicators, biomass characteristics and measurements collected during phytomanagement trials. These datasets are incorporated into the database following the agreed validation and harmonisation procedures and contribute to the iterative refinement of soil and biomass indicators as new knowledge emerges. WP3 adds process-related data, such as operational conditions, extraction yields and product quality, ensuring that information on biomass valorisation pathways is integrated with upstream and downstream assessments.

The data gathered and organized according to the database template will be instrumental for WP5 activities on sustainability assessment. Specifically, it will supply validated experimental inputs from WP1-WP3, such as energy/material usage, environmental releases, operational costs, and process yields, for Life Cycle Assessment (T5.1), Life Cycle Costing (T5.2), Social-LCA (T5.3), and the integrated evaluation (T5.4). This ensures comprehensive, harmonized data support the quantification of technical, economic, social, and environmental impacts across the pHYBi life cycle, to elaborate recommendations and guidelines.

The development of the database is also closely connected with the Data Management Plan in WP7, which provides the principles for data governance, FAIR compliance, metadata requirements and long-term preservation. The present deliverable ensures that the mechanisms defined in the DMP are operationalised and consistently applied across the project.

This deliverable is key for the continuation of activities within WP4. By establishing the database architecture, data flows and operational framework at an early stage, D4.1 provides the foundation required for subsequent modelling integration efforts, digital tools and the final database release. It is therefore a critical enabler for the coherent progress and overall execution of WP4.

2. REQUIREMENTS AND SPECIFICATIONS

2.1. DATA SOURCES AND INPUTS

2.1.1. WP1: SOIL QUALITY DATA, BIOMASS CHARACTERISTICS.

WP1 provides the core baseline datasets required for site investigation and feedstock characterisation, representing a fundamental input to the pHYBi database. Soil quality data collected under WP1 cover a comprehensive set of physical, chemical and biological parameters, generated using harmonised protocols as defined in Deliverable D1.1. These include indicators related to soil structure and physicochemical properties (e.g. texture, bulk and particle density, porosity, pH, electrical conductivity, cation exchange capacity and water holding capacity), nutrient availability, contamination profiles and microbial health.
In addition to soil parameters, WP1 generates detailed data on plant and biomass characteristics. These datasets capture indicators of plant health, physiological stress responses, biomass productivity and pollutant uptake, as well as lignocellulosic composition relevant for downstream processing and valorisation. Biomass quality parameters are recorded in a standardised manner to ensure comparability between sites and across monitoring campaigns.
Then, the data collection template is divided into two main tabs regarding WP1 activities: *soil* and *plant*. These correspond to the respective sample types collected under WP1. The *soil* tab captures all soil-related measurements, including physical, chemical, and biological parameters. This tab is further divided into multiple sections to organize the information systematically:

- **Sample information**: includes case study, sample code, treatment, sampling date, and soil type.
- **Soil structure and physicochemical properties**: contains total SOC, erosion, bulk density, particle density, porosity, electrical conductivity, cation exchange capacity, water holding capacity, water retention, pore water pressure, aggregate stability, organic matter content, pH, redox potential, and buffer index.
- **Soil microbial indicators**: includes basal respiration, microbial biomass, and fungi-to-bacteria ratio.
- **Soil nutrients**: covers calcium (Ca), sodium (Na), potassium (K), magnesium (Mg), nitrogen (N), and phosphorus (P).
- **Heavy metals**: includes silver (Ag), aluminium (Al), boron (B), barium (Ba), cadmium (Cd), cobalt (Co), chromium (Cr), copper (Cu), iron (Fe), mercury (Hg), manganese (Mn), nickel (Ni), lead (Pb), antimony (Sb), titanium (Ti), vanadium (V), and zinc (Zn).

- **Organic pollutants**: such as total petroleum hydrocarbons (TPHs) and polycyclic aromatic hydrocarbons (PAHs).
- **Salt content**: including sodium (Na+), potassium (K+), calcium (Ca+), magnesium (Mg+), chloride (Cl-), sulphate (SO_4^{2-}), nitrate (NO_3^-), and bicarbonate (HCO_3^-).

The data collection template ensures that all WP1 datasets are consistently structured and accompanied by sufficient metadata, enabling longitudinal monitoring, cross-site comparisons and direct linkage with downstream datasets from WP2 and WP3. Screenshots of these described template sections are provided in the annexes for reference.

A sampling code was established as a structured identifier to ensure each soil and plant sample can be uniquely tracked and cross-linked across field collection, laboratory analyses, and subsequent data integration steps throughout the pHYBi process. This common coding logic makes it possible to relate measurements coming from different sources (field observations, lab testing, and, when relevant, associated genomic samples) and supports both WP2 collaboration and a holistic mathematical modelling approach.

For soil, the code is built by concatenating five elements in a fixed order:

- Case study (country/site code): e.g., FRA (France), SPN (Spain), CRO (Croatia).
- Sampling location (following the LUCAS-type location convention): e.g., C (Central), N (North), E (East), S (South), W (West).
- Depth: TS (Topsoil) or SS (Subsoil).
- Date: YYYYMMDD.
- Replicate number: integer indicating replicate.

Format: CASE-LOCATION-DEPTH-YYYYMMDD-REP
Example: FRA-C-SS-20250328-3
(France – Central – Subsoil – 28/03/2025 – replicate 3)

This soil coding system keeps track of samples by giving each one a unique, structured identifier that captures the key descriptors needed to follow it end-to-end. This makes it possible to trace every soil record back to its exact sampling context and to consistently link all associated measurements (field notes, laboratory results, and any derived parameters) to the same physical sample throughout the pHYBi workflow.

Because every partner uses the same code structure, datasets coming from different teams and analytical pipelines can be merged reliably without ambiguity (e.g., avoiding mix-ups between similar sites, dates, or depth layers). In practice, it becomes the common “join key” in the global dataset, enabling integration across work packages, supporting quality control (duplicate checks, missing replicate detection, outlier screening by depth/location), and ultimately improving the consistency of inputs used for holistic mathematical modelling.

2.1.2. WP2: FIELD PHYTOMANAGEMENT DATA.

WP2 contributes operational and monitoring data derived from the implementation of phytomanagement strategies under field conditions. These datasets document the practical application of management interventions, including plant establishment, growth dynamics, land-use practices, amendment applications and temporal changes in soil–plant interactions. Field-based measurements are complemented by observational records and, where applicable, remote sensing outputs that capture spatial patterns and temporal trends.

The data collected in WP2 are essential for interpreting how site-specific conditions and management actions influence soil quality, plant performance and biomass production. Through the use of the harmonised data collection template, WP2 datasets are systematically linked to the baseline soil and biomass data from WP1, allowing changes over time to be clearly attributed to specific phytomanagement interventions.

Moreover, WP2 plays a key role in the iterative refinement of soil and biomass quality indicators. As field experience accumulates and stakeholder interactions progress, the indicators tracked within the database may be updated to reflect improved scientific understanding and practical relevance. The database structure supports this adaptive approach, ensuring continuity while allowing controlled updates to indicator definitions and monitoring priorities.

To capture this, the *plant* tab records information on biomass characteristics, physiological responses, and pollutant uptake. This dual-tab structure ensures that all relevant information from each sample type is systematically collected, clearly organised, and directly aligned with the metadata standards and quality checks defined in the template, facilitating integration into the central pHYBi database and supporting cross-WP analyses. This section of the template was structured into clearly defined sections to ensure consistency, traceability and interoperability across sites and monitoring campaigns. Screenshots of these described template sections are provided in the annexes for reference.

- **Sample information**: includes case study, sample code, plant species, treatment, sampling date, growth stage and sampling compartment (e.g. roots, stems, leaves or whole plant).
- **Plant growth and biomass productivity**: records indicators such as plant height, stem diameter, leaf area, fresh and dry biomass yield, above- and below-ground biomass allocation, and growth rates where applicable.
- **Physiological and stress indicators**: includes parameters related to plant health and stress responses, such as chlorophyll content, photosynthetic efficiency, transpiration, water use efficiency, visual stress symptoms and relevant biochemical markers.
- **Nutrient content**: captures macro- and micronutrient concentrations within plant tissues, including nitrogen (N), phosphorus (P), potassium (K), calcium (Ca), magnesium (Mg) and sodium (Na), supporting assessments of nutrient uptake and cycling.

- **Heavy metal uptake**: includes concentrations of metals accumulated in plant tissues (e.g. Cd, Pb, Zn, Cu, Ni, Cr, Hg and others as relevant), enabling the evaluation of phytoextraction, phytostabilisation and food-chain risk pathways.
- **Organic contaminants**: records plant-associated concentrations of organic pollutants, where applicable, supporting assessments of contaminant transfer and biomass safety for downstream valorisation.
- **Biomass quality and composition**: includes lignocellulosic composition (cellulose, hemicellulose and lignin), moisture content, ash content and other quality parameters relevant for biomass processing and valorisation pathways assessed in WP3.

A clear plant coding system was also established to ensure that every plant sample collected within the project can be uniquely identified and consistently traced from field sampling through laboratory analyses and final data reporting. By encoding essential information (case study, plant species, plant part, sampling date, and replicate) into a single standardized identifier, the code reduces ambiguity across partners and sampling campaigns and supports robust quality control. The code is built by concatenating four elements in a fixed order:

- Case study (country/site code): e.g., FRA, SPN, CRO.
- Plant species code: e.g., MICA (Miscanthus), etc.
- Plant part: LE (Leaf), ST (Stem), RO (Roots).
- Date: YYYYMMDD.
- Replicate number: integer indicating replicate.

Format: CASE-SPECIES-PART-YYYYMMDD-REP
Example: FRA-MICA-LE-20250328-3
(France – Miscanthus – Leaf – 28/03/2025 – replicate 3)

These codes allow plant samplings to be bonded to soil samples because they share the same core identifier elements (the case study/site code and the sampling date, and if sampling is co-located, the same location descriptor), which makes it straightforward to match records across matrices within the global dataset. By using these shared fields as a common “join key”, plant results (e.g., biomass, tissue concentrations, physiological traits) can be directly linked to the corresponding soil context (e.g., depth-resolved chemistry and physical properties) collected at the same site and campaign.

2.1.3. WP3: EXTRACTION PROCESS CONDITIONS, YIELDS, PRODUCT QUALITY.

WP3 delivers process-oriented datasets from biomass pretreatment, extraction, fermentation, and valorisation trials, capturing operational parameters, yields, and product characteristics essential for scaling phytomanagement outputs into viable bio-based products. These data document the transformation of WP1/WP2 biomass feedstocks through T3.1/T3.2 extraction (organosolv pretreatment, yields of

cellulose/hemicellulose/lignin, metal content) and T3.4/T3.5 validation/valorisation (fermentation conditions, product concentrations, coating performance). Key metrics include harvesting/debarking efficiency, drying parameters, particle size distribution, reaction conditions (temperature, solvent ratio, catalyst), biomass yields (%), and downstream testing (e.g., water/oil resistance, wash fastness).
These datasets enable performance evaluation of extraction efficiency, product purity, and application viability, while linking back to upstream soil/plant inputs via shared sample codes for full-chain traceability and optimisation.
The data collection template for WP3 is structured into two primary tabs: Extraction (T3.1; T3.2) and Validation/Valorisation (T3.4; T3.5). These tabs systematically organise process data from pretreatment through product testing. This section of the template aims to collect all this information while connecting it with information gathered in WP1 and WP2 by coding matching (e.g., CASE-SPECIES-PART-YYYYMMDD-REP or soil equivalents). These tabs include the following information. Screenshots of these described template sections are provided in the annexes for reference.

- **Sample information**: case study, soil sample code, plant sample code, valorisation sample code, and feedstock type.
- **Harvesting and debarking:** records the harvesting method applied, the debarking method used, the debarking duration in hours and efficiency as a percentage.
- **Drying**: records the initial moisture content as a percentage, final moisture content as a percentage, drying temperature in °C, and drying duration in hours.
- **Particle characteristics**: records bulk density, shredding method employed, particle size range in cm, and particle size distribution as a percentage.
- **Organosolv treatment parameters:** records solvent type, catalyst type, reaction temperature in °C, reaction time in hours, and solvent ratio as volume/volume.
- **Separation and recovery:** records cellulose yield as a percentage, hemicellulose yield as a percentage, and lignin yield as a percentage.
- **Metal content analysis:** records concentrations of silver (Ag), aluminium (Al), copper (Cu), iron (Fe), and zinc (Zn).

And information collected regarding further valorisation on the pHYBi project and validation includes:

- **Sample information:** case study, soil/plant/valorisation/lab sample codes, feedstock type, *E. coli* strain, inoculum concentration, replicate, culture volume (mL).
- **Fermentation**: records temperature in °C, fermentation duration in hours, initial pH, final pH, product concentration in mg/L, quantification method used, and growth rate in 1/h.
- **Coating**: records coating formulation ID, substrate type, substrate material, fabric construction, coating method applied, application thickness, curing temperature in °C, and curing duration in hours.

- **Coating performance testing:** records water resistance (Y/N), oil resistance (Y/N), and wash fastness grade.

A unified coding system is extended to WP3 samples, incorporating upstream codes (soil/plant) plus process-specific identifiers (e.g., valorisation sample code, LAB sample code) to track material flow from field to product. Format builds on prior logic: CASE-[UPSTREAM_CODE]-PROCESS-YYYYMMDD-REP (e.g., FRA-MICA-LE-20250328-EXT1-3 for extraction replicate 1). This "join key" enables end-to-end tracing, linking extraction yields to source biomass quality and soil context for holistic analysis.

2.2. DATA STANDARDS AND COMPLIANCE

This section outlines how the pHYBi database construction adheres to the data quality and regulatory compliance standards established in the project's Data Management Plan (D7.3) and the FAIR principles mandated by the European Commission.

2.2.1. FAIR PRINCIPLES.

The pHYBi database is architected to ensure that all data generated within WP1, WP2, and WP3 and integrated in WP4 and WP5 is Findable, Accessible, Interoperable, and Reusable (FAIR):

- **Findable**: every data entry is assigned a unique identifier following the project's standardized naming convention (e.g., CASE-COUNTRY-DEPTH-DATE-REP). This ensures full traceability from field sampling to digital modelling. Furthermore, datasets will be described using rich metadata based on the Dublin Core standards, as specified in D7.3.
- **Accessible:** during the project's execution, data is stored in the consortium's secure MS Teams/SharePoint environment. Upon validation and at the end of the project, public-facing datasets will be deposited in trusted repositories such as Zenodo, ensuring long-term accessibility.
- **Interoperable:** To facilitate data exchange between partners and external stakeholders, the database utilizes open, non-proprietary formats such as CSV, JSON, and XLSX. Data integration follows standardized vocabularies.
- **Reusable:** the inclusion of detailed methodology metadata ensures that third parties can understand and replicate the data generation process after the project ends.

2.2.2. INTEROPERABILITY CONSIDERATIONS.

To ensure the database acts as a central node for the project's modelling and replication tasks, the following interoperability measures have been implemented:

- **Semantic Interoperability**: database employs controlled vocabularies and standardized units of measurement. This prevents ambiguity when merging results from different laboratory sites.
- **Technical Interoperability**: modular architecture of the database allows for seamless integration with the simulation tools developed in WP4. By using standard structures or common file formats, the data can be directly integrated in Life Cycle Assessment (LCA) and Techno-Economic Assessment (TEA) software.
- **Metadata Alignment:** the database structure is aligned with the ISO 19115 standard for geographical and environmental information, ensuring that pHYBi's outputs can be harvested by European open science platforms and data aggregators.

3. OVERALL DATABASE ARCHITECTURE

The pHYBi database is conceived as a central, project-wide repository that organizes heterogeneous datasets from field campaigns, laboratory analyses and process trials into a coherent, analysis-ready structure. It is implemented as a structured, Excel-based template system for data entry (e.g. SOIL, PLANT, EXTRACTION tabs) combined with a consolidated master database that aggregates and harmonizes partner submissions. **The architecture is designed so that data generated in WP1–WP3 can be ingested with minimal reformatting, validated against shared rules, and then exposed in a form directly usable for modelling, replication and sustainability assessment activities** in WP4 and WP5.
For this, a common sampling and coding logic that provides unique identifiers for each soil, plant and process sample was designed, allowing all measurements associated with that sample to be linked across partners, workflows and time. Around this core, the database is structured into thematic modules (soil, plant/biomass, extraction/process, textile valorisation) that share a consistent layout and controlled vocabularies but can evolve independently as new indicators or methods are incorporated. IDE coordinates the overall architecture, template design and integration procedures, while contributing partners provide data in the agreed formats and participate in iterative refinements as use-cases and modelling requirements mature.

3.1. DESIGN PRINCIPLES

The architecture is guided by a set of design principles that respond to pHYBi's interdisciplinary nature and to the need of the project including modularity, integration and interoperability:

3.1.1. MODULARITY, INTEGRATION, INTEROPERABILITY.

- **Modularity.** The database is organised into logical modules that mirror the main data domains of the project: soil quality and contamination (SOIL sheet), plant and biomass characteristics (PLANT sheet), extraction and processing data (EXTRACTION and related sheets), and textile valorisation data. Each module has its own set of fields, sections and metadata tailored to the corresponding workflows. For example, the SOIL sheet is structured into "sample information", "soil structure and physicochemical properties", "soil microbial indicators", "soil nutrients", "heavy metals", "organic pollutants", and "salt content". This modular organisation allows new variables or sections to be added within a given domain without disrupting other parts of the system, and facilitates targeted updates driven by specific WP needs.
- **Integration**. Despite being modular, all sheets use a shared sampling code and harmonised metadata fields so that records from different modules can be seamlessly joined into an integrated project dataset. The soil and plant templates, for instance, both start with a common "case study / sample code /

treatment / sampling date / type" block, and extraction sheets reuse the same soil and plant sample codes as foreign keys. This ensures that a single soil sample can be consistently linked to its associated plant measurements, extraction trials and textile tests, **enabling end-to-end traceability from field conditions to valorisation performance.** Integration is further reinforced by aligned units, naming conventions and quality-control columns embedded directly in the templates.

- **Interoperability**. The database is designed to interoperate with other tools and standards at three levels:
    - Semantic interoperability: controlled vocabularies and standard units are enforced in the templates (e.g. mg/kg dry soil for nutrients and metals, meq/100 g soil for salt content, clearly specified lists of pollutants), reducing ambiguity when merging datasets from different labs.
    - Technical interoperability: the Excel-based templates follow stable, documented layouts so that they can be programmatically parsed and transformed into formats required by modelling, LCA and TEA tools (e.g. CSV, relational tables). The modular structure also aligns with the needs of WP4 simulation tools, which can ingest only the relevant modules without complex preprocessing.
    - Metadata alignment: the metadata fields in the database are aligned with recognised standards for environmental and spatial data (e.g. ISO-type descriptors for site, date, method, instrument), facilitating long-term integration in open-science platforms and ensuring FAIR compliance as defined in the project's Data Management Plan.

3.2. DATA FLOW WITHIN THE PROJECT

The database architecture is tightly coupled to a standardised data flow that specifies how information is generated, structured, validated and integrated across WPs. The flow starts at the point of data generation (field campaigns, lab analyses, extraction trials) and ends with curated records in the central database that are **ready for modelling, interpretation and sustainability assessment**. IDE coordinates this flow under Task 4.1, ensuring that templates, coding rules and quality checks are consistently applied and that partners have clear guidance on when and how to deliver their datasets.

3.2.1. HOW DATA TRAVEL FROM FIELD/LAB TO THE DATABASE.

1. Planning and template distribution
    - IDE develops and maintains the master Excel templates for soil, plant and extraction data, embedding predefined fields, units, controlled terminology and metadata requirements.

- These templates are circulated to WP leaders (WP1, WP2, WP3) together with concise instructions on data entry, coding rules and quality-control expectations.

2. Field and laboratory data generation.
    - WP1 and WP2 teams collect soil and biomass data in the field following agreed protocols (as defined indicators from D1.1), using field notebooks or internal lab systems to record raw measurements and observations.
    - Laboratory analyses (physicochemical parameters, microbial indicators, contaminant concentrations, lignocellulosic composition, among others) are performed according to harmonised methods, and results are stored in local lab formats.
3. Sample coding and traceability.
    - At the time of sampling and/or first data entry, each sample is assigned a project-specific code built from structured components (case study, location, depth or sample type, date, replicate), ensuring that every record can be uniquely identified and traced across all steps.
    - The same code is reused in all subsequent measurements and in different tabs of the Excel template, so that soil, plant, extraction and genomic data (where applicable) can be reliably cross-linked.
4. Template population by partners.
    - Contributing partners then transfer validated field and lab results into the official pHYBi Excel templates (SOIL, PLANT, EXTRACTION, etc.), filling in all mandatory metadata and measurement fields.
    - Dedicated columns are used to capture essential contextual information (e.g. method/protocol, instrument, detection limits, “Associated genomic sample (Y/N)”) to **support later integration with WP2 genomic datasets and with modelling workflows**.
5. Partner-side checks and submission.
    - Before submission, each partner performs internal consistency checks on codes, units, ranges and completeness using the structure and validation cues embedded in the templates.
    - The filled-in spreadsheets are then delivered to IDE following the schedule and file-naming conventions agreed in Task 4.1, typically via the project’s shared workspace or secure file transfer channels.
6. Central ingestion, validation and harmonisation.
    - IDE imports partner spreadsheets into the central database layer, running automated and manual quality controls (code uniqueness checks, duplicate detection, range/outlier screening, unit consistency).

- Where issues are detected, IDE opens a clarification feedback communication with the originating partner to correct or annotate records; resolved datasets are then integrated into the consolidated master tables.

7. Database consolidation and exposure.
    - Once harmonised, the data are stored in the structured database aligned with the conceptual model (soil, plant, process, metadata entities) and made available to WP4, WP5 and other WPs for modelling, replication and assessment.
    - Extracts tailored to specific uses (e.g. model input tables, LCA inventories, indicator dashboards) are generated as needed, always preserving the original sample codes and metadata for traceability.

3.2.2. ROLES OF IDE AND CONTRIBUTING PARTNERS.

IDE (database lead).

- Designs and maintains the overall database architecture, including the structure of the Excel templates, the sampling coding system and the mapping to the internal master database schema.
- Coordinates Task 4.1, including requirements elicitation from WP1–WP3, alignment with the Data Management Plan and updates to templates or procedures when new indicators or workflows arise.
- Manages data ingestion, validation and harmonization: runs central QC routines, resolves inconsistencies with partners, and ensures that integrated datasets remain consistent, traceable and FAIR-compliant.
- Provides curated data extracts and documentation to modelling, replication and sustainability assessment tasks in WP4 and WP5, and supports partners in using the database correctly.

Contributing partners (WP1, WP2, WP3 and others).

- Generate primary data in the field and laboratory according to agreed protocols and quality standards, and ensure that measurements are correctly associated with the project sampling codes.
- Populate the official pHYBi Excel templates with validated results, including all required metadata, and perform initial checks for completeness and consistency before submission.
- Collaborate with IDE during validation and harmonization, clarifying methods, resolving anomalies and, where necessary, updating past submissions to reflect improved protocols or corrected values.
- Use the central database as the reference source for cross-WP analyses, drawing on the integrated datasets for their own scientific objectives (e.g.

indicator refinement in WP1, genomic integration in WP2, process optimization in WP3) while feeding back any structural or semantic needs to IDE.

4. DATA MODEL AND STRUCTURE

The pHYBi data model translates the high-level architecture and coding system into a set of clearly defined entities, relationships and tables that allow heterogeneous measurements to be stored, queried and analysed in a consistent way. Conceptually, the model is organized around a small number of core entities (SOIL, BIOMASS, PROCESS, SUSTAINABILITY_METRICS, METADATA) that capture the main objects of interest in the project, complemented by satellite entities for specific parameter groups (e.g. soil structure, nutrients, heavy metals). These entities are implemented as structured tables in the Excel templates and in the consolidated database, with shared keys and controlled vocabularies ensuring that data can be joined and traced across WPs and workflows.

4.1.CONCEPTUAL DATA MODEL

4.1.1. ENTITY–RELATIONSHIP DIAGRAM.

At a conceptual level, the pHYBi database can be represented as an entity–relationship (ER) model[1] centred on the sample and experiment lifecycle. Although the ER diagram is implemented across multiple sheets and tables, its core relationships can be summarized as follows:

- SOIL is the foundational entity representing individual soil samples collected in field campaigns. Each SOIL record is uniquely identified by the project sampling code and is associated with one or more sets of physical, chemical and biological measurements.
- BIOMASS represents plant and biomass samples collected at the same sites and time points as the soil samples. BIOMASS entities are linked to SOIL either directly (through shared case study and sample codes) or via a dedicated linking key, capturing above- and below-ground relationships between soil conditions and plant responses.
- PROCESS captures the downstream valorisation steps applied to biomass and derived materials, including extraction, fermentation and textile coating. PROCESS records reference both the upstream BIOMASS (and, by extension, SOIL) from which the feedstock was obtained and the operational parameters and outcomes of the processing steps.
- SUSTAINABILITY_METRICS stores derived indicators used for environmental, economic and social assessments (aggregated emissions, resource use, cost items). These metrics are parameterised by PROCESS and, where relevant, by SOIL and BIOMASS characteristics used as input to sustainability models.

[1] Chen, Peter Pin-Shan. "The entity-relationship model—toward a unified view of data." *ACM transactions on database systems (TODS)* 1.1 (1976): 9-36.

- METADATA is a transversal entity providing contextual information (e.g. site descriptors, sampling protocols, analytical methods, instrument settings, quality flags) that is referenced by all other entities to ensure traceability and reproducibility.

In ER terms, the model can be summarized as:

- SOIL 1-to-many SOIL_MEASUREMENT subsets (structure, nutrients, contaminants, etc.).
- SOIL 1-to-many BIOMASS (multiple plant samples per soil context).
- BIOMASS 1-to-many PROCESS (multiple valorisation or processing trials per biomass type).
- PROCESS 1-to-many SUSTAINABILITY_METRICS (multiple indicators per process scenario).
- METADATA entities (e.g. SITE, METHOD, INSTRUMENT) referenced by foreign keys across SOIL, BIOMASS and PROCESS tables.

This conceptual structure reflects the logical progression from environmental conditions (SOIL) to biological response (BIOMASS) to technological transformation (PROCESS) and finally to integrated evaluation (SUSTAINABILITY_METRICS), with METADATA providing the necessary glue to keep the entire chain coherent. The entire diagram is presented in Figure 1. Alternatively, it can be accessed through the following link in *.svg* format.

4.1.2. KEY ENTITIES (SOIL, BIOMASS, PROCESS, SUSTAINABILITY METRICS, METADATA)

SOIL. This entity encapsulates all information related to a given soil sample at a specific location, depth and time. It includes:

- Identification attributes: case study, soil sample code, treatment, sampling date, soil type.
- Structural and physicochemical attributes: bulk and particle density, porosity, total SOC, water holding capacity, pH, redox potential, buffer index, aggregate stability, permeability, electrical conductivity, cation exchange capacity.
- Biological attributes: microbial basal respiration, microbial biomass, fungi:bacteria ratio.
- Contamination and nutrient attributes: macro-nutrients (Ca, Na, K, Mg, N, P), heavy metals (e.g. Ag, Cd, Cr, Cu, Fe, Pb, Zn), organic pollutants (TPHs, PAHs), and salt ions (Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, NO_3^-, HCO_3^-).

BIOMASS. This entity covers plant and biomass samples generated under the same phytomanagement conditions. It includes:

- Identification attributes: case study, plant sample code, treatment, sampling date, plant/biomass type.

- Physiological and stress indicators: chlorophyll content, photosynthetic efficiency, transpiration, water use efficiency, visible stress symptoms, biochemical markers.
- Growth and productivity indicators: plant height, stem diameter, leaf area, fresh and dry biomass yields, below-/above-ground allocation.
- Contaminant uptake and composition: heavy metal concentrations in tissues, organic contaminants where relevant, and lignocellulosic composition (cellulose, hemicellulose, lignin), moisture and ash content.

PROCESS. The entities describe the valorisation pathways applied to biomass and derived intermediates. They include:

- Sample linkage attributes: case study, soil sample code, plant sample code, valorisation sample code, feedstock type.
- Pre-treatment attributes: harvesting and debarking method, drying conditions, particle size distribution and bulk density.
- Extraction attributes: solvent and catalyst type, temperature, reaction time, solvent ratio, yields of cellulose, hemicellulose and lignin, residual metal content.
- Downstream processing attributes: fermentation parameters (strain, inoculum, temperature, pH, product concentration, growth rate) and textile coating parameters (formulation ID, substrate, coating method, curing conditions, performance tests).

SUSTAINABILITY_METRICS. This entity holds aggregated indicators computed within WP5 and related tasks, including:

- Environmental metrics (e.g. emissions, resource use, impact scores per functional unit) derived from LCA models using PROCESS and upstream data.
- Economic metrics (e.g. cost breakdowns, CAPEX/OPEX, unit costs) derived from LCC models.
- Social metrics (e.g. S-LCA indicators where applicable), linked to specific case studies or scenarios.

Though not fully implemented in the current template, the conceptual model anticipates these records as additional tables linked back to PROCESS and case study identifiers.

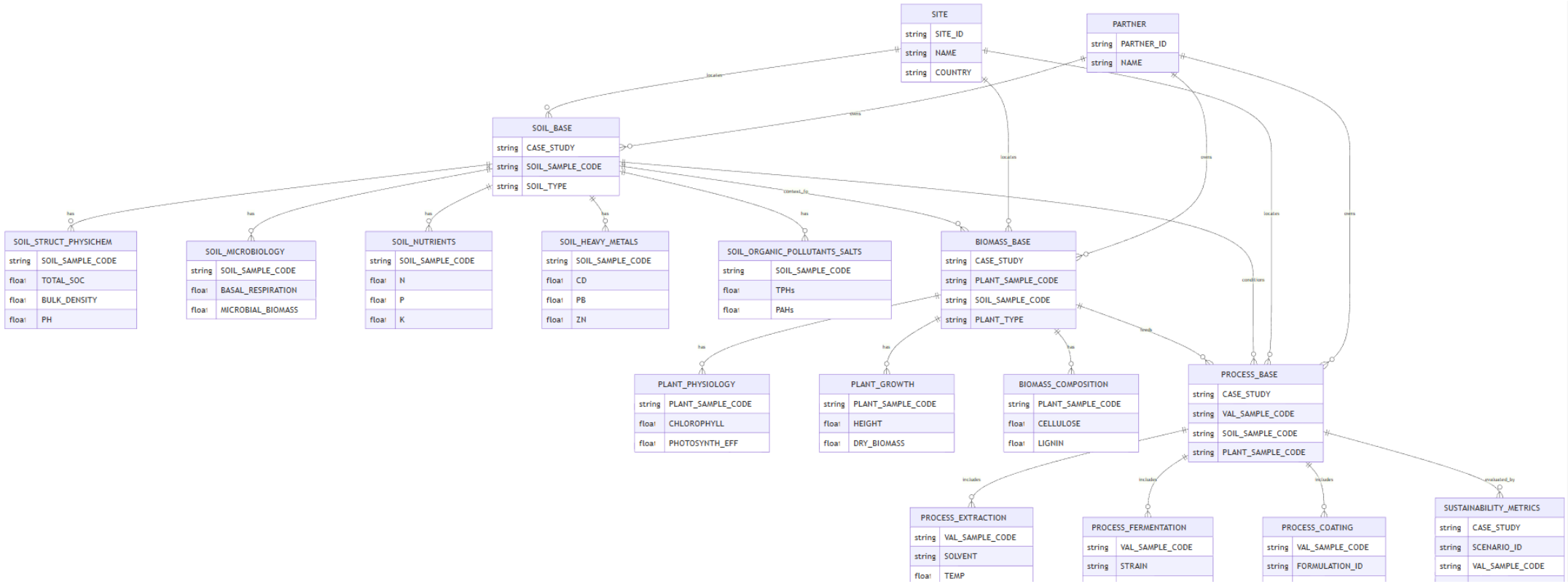


Figure 1. Conceptual entity–relationship diagram of the pHYBi database (soil, biomass, process, and sustainability entities).

METADATA. These entities store contextual and governance information such as:

- SITE descriptors: geographical location, land-use history, climate, contamination profile.
- METHOD descriptors: sampling and analytical protocols, standard references, detection limits.
- INSTRUMENT descriptors: equipment types, models, calibration dates.
- DATA_GOVERNANCE descriptors: responsible partner, data owner, versioning and quality flags.

These metadata records are referenced by codes in SOIL, BIOMASS and PROCESS tables to fully document how each measurement was obtained.

4.2. LOGICAL DATA MODEL

4.2.1. TABLES AND FIELDS.

The conceptual entities are implemented as a set of structured tables. In the Excel-based implementation used for data collection, these appear as different sheets and column groups, which are then mapped to relational tables in the consolidated database. The main logical tables are (see Figure 1):

- SOIL_BASE
  - Key fields: CASE_STUDY, SOIL_SAMPLE_CODE, TREATMENT, SAMPLING_DATE, SOIL_TYPE.
  - Purpose: core identification and general descriptors for each soil sample.
- SOIL_STRUCT_PHYSICHEM
  - Key fields: SOIL_SAMPLE_CODE (FK to SOIL_BASE).
  - Measurement fields: TOTAL_SOC, EROSION, BULK_DENSITY, PARTICLE_DENSITY, POROSITY, EC, CEC, WHC, WATER_RETENTION, PORE_WATER_PRESSURE, AGGREGATE_STABILITY, ORGANIC_MATTER, PH, REDOX_POTENTIAL, BUFFER_INDEX.
- SOIL_MICROBIOLOGY
  - Key fields: SOIL_SAMPLE_CODE (FK).
  - Measurement fields: BASAL_RESPIRATION, MICROBIAL_BIOMASS, FUNGI_BACTERIA_RATIO.
- SOIL_NUTRIENTS
  - Key fields: SOIL_SAMPLE_CODE (FK).
  - Measurement fields: CA, NA, K, MG, N, P (mg/kg dry soil).
- SOIL_HEAVY_METALS

- Key fields: SOIL_SAMPLE_CODE (FK).
- Measurement fields: AG, AL, B, BA, CD, CO, CR, CU, FE, HG, MN, NI, PB, SB, TI, V, ZN (mg/kg dry soil).

- SOIL_ORGANIC_POLLUTANTS_SALTS
  - Key fields: SOIL_SAMPLE_CODE (FK).
  - Measurement fields: TPHs, PAHs; NA_PLUS, K_PLUS, CA_PLUS, MG_PLUS, CL_MINUS, SO4_2MINUS, NO3_MINUS, HCO3_MINUS (meq/100 g soil).
- PLANT_BASE / BIOMASS_BASE
  - Key fields: CASE_STUDY, PLANT_SAMPLE_CODE, TREATMENT, SAMPLING_DATE, PLANT_TYPE.
  - Purpose: identification and general descriptors for plant/biomass samples.
- PLANT_PHYSIOLOGY, PLANT_GROWTH, PLANT_UPTAKE, BIOMASS_COMPOSITION
  - Key fields: PLANT_SAMPLE_CODE (FK to PLANT_BASE).
  - Measurement fields: physiological/stress indicators, growth metrics, tissue metal/organic contaminants, lignocellulosic composition, moisture, ash.
- PROCESS_BASE
  - Key fields: CASE_STUDY, VALORISATION_SAMPLE_CODE, FEEDSTOCK_TYPE.
  - Link fields: SOIL_SAMPLE_CODE, PLANT_SAMPLE_CODE (FKs).
  - Purpose: identification of each process/valorisation run and its linkage to upstream samples.
- PROCESS_PRETREATMENT, PROCESS_EXTRACTION, PROCESS_FERMENTATION, PROCESS_COATING
  - Key fields: VALORISATION_SAMPLE_CODE (FK to PROCESS_BASE).
  - Measurement fields: process-specific operational conditions and outputs (pre-treatment, extraction yields, fermentation performance, coating performance).
- SUSTAINABILITY_METRICS (future logical table)
  - Key fields: CASE_STUDY, SCENARIO_ID, VALORISATION_SAMPLE_CODE (FKs).
  - Measurement fields: environmental, economic and social indicators used in WP5 assessments.
- METADATA tables (e.g. SITE, METHOD, INSTRUMENT, DATA_OWNER)

- Key fields: SITE_ID, METHOD_ID, INSTRUMENT_ID, PARTNER_ID.
- Descriptive fields: site attributes, method descriptions, instrument details, partner role.

These logical tables correspond directly to the labelled sections in the Excel templates (the SOIL sheet sections map to SOIL_BASE, SOIL_STRUCT_PHYSICHEM, etc.), ensuring that the transition from partner-side spreadsheets to the integrated database is straightforward and traceable.

4.2.2. RELATIONSHIPS, KEYS AND LINKS TO INDICATORS.

Relationships in the logical model are implemented through shared keys that mirror the conceptual entity–relationship structure and the sampling code logic. The main mechanisms are:

- **Primary keys (PK).**
  - SOIL_BASE: (CASE_STUDY, SOIL_SAMPLE_CODE).
  - PLANT_BASE: (CASE_STUDY, PLANT_SAMPLE_CODE).
  - PROCESS_BASE: (CASE_STUDY, VALORISATION_SAMPLE_CODE).
  - SUSTAINABILITY_METRICS: (CASE_STUDY, SCENARIO_ID, VALORISATION_SAMPLE_CODE).
- **Foreign keys (FK).**
  - All SOIL_* measurement tables carry SOIL_SAMPLE_CODE as FK to SOIL_BASE.
  - All PLANT_* tables carry PLANT_SAMPLE_CODE as FK to PLANT_BASE.
  - PROCESS_BASE carries SOIL_SAMPLE_CODE and PLANT_SAMPLE_CODE as FKs, linking each process run to the corresponding environmental and biomass context.
  - PROCESS_* detail tables carry VALORISATION_SAMPLE_CODE as FK to PROCESS_BASE.
  - SUSTAINABILITY_METRICS references PROCESS_BASE (and, by extension, its upstream SOIL/PLANT) to associate each indicator with a specific process configuration and case study.
  - Metadata IDs (e.g. SITE_ID, METHOD_ID) are referenced in SOIL_BASE, PLANT_BASE and PROCESS_BASE to indicate the context and protocols used.
- **Links to indicators.**
  - The raw measurement fields in SOIL, PLANT and PROCESS tables are the direct inputs to indicator calculations in WP4 and WP5 (e.g. soil quality indices, biomass quality indices, process efficiency metrics).

- Derived indicators are either stored in dedicated indicator columns within the same tables or in SUSTAINABILITY_METRICS, depending on their scope and usage.
- Because all indicator calculations rely on the common sample and process keys, it is always possible to trace each indicator value back to the underlying measurements, methods and sites used to derive it.

5. CONCLUSIONS

The development of the pHYBi database under Task 4.1 is used to transform a fragmented set of field, laboratory and process datasets into a coherent, reusable digital asset that underpins the project's scientific and technological objectives and to allow interaction among WP. Standardised Excel templates (e.g. SOIL, PLANT, EXTRACTION) and a common sampling and coding system were co-designed, then, partners now share a unified framework for data collection, which ensures consistent formats, units, metadata and quality-control fields across WPs. This harmonisation enables cross-site comparisons, longitudinal analyses and seamless linkage between upstream environmental measurements and downstream valorisation data.

The overall database architecture, grounded on clear design principles of modularity, integration and interoperability, has been formalised into a transparent conceptual and logical model. Core entities (SOIL, BIOMASS, PROCESS, SUSTAINABILITY_METRICS and METADATA) and their relationships have been explicitly defined, and their implementation in structured tables ensures that heterogeneous data can be reliably joined and traced throughout the full pHYBi workflow. The resulting system is not only fit for the immediate needs of WP1–WP3 and WP4, but has been explicitly designed to supply validated and well-documented inputs to WP5's LCA, LCC and S-LCA activities, as well as to future model coupling and replication studies.

Data governance principles and FAIR compliance have been assured, while contributing partners have ensured that the templates and workflows remain grounded in experimental reality and operational constraints. The established data flow, from field and lab to curated central tables, provides clarity on responsibilities, timing and quality checks and ensures that new data can be integrated with minimal overhead as the project progresses. Importantly, the close linkage with the Data Management Plan guarantees that issues such as provenance, documentation and long-term stewardship have been addressed from the outset rather than as an afterthought.

In conclusion, D4.1 delivers a mature and operational database backbone that is already supporting ongoing data collection and early modelling work and that will remain central for subsequent WP4 tasks and WP5 sustainability assessments. By establishing the architecture, data model and standard operating procedures at this stage, the project has significantly reduced future integration risks and created the conditions for scalable digitalisation, reproducibility and post-project re-use of pHYBi data.

6. ANNEXES

I. Soil data collection template

Task 1.1: Definition of monitoring requirements and quality indicators
DATA COLLECTION TEMPLATE*
SOIL CHARACTERIZATION
*This tamplate is based on the information provided in D1.1.

Sample			
Case of study	**Soil Sample Code**	**Treatement**	**Sampling Date**

Figure 1A. Sample information: includes case study, sample code, treatment, sampling date, and soil type.

Soil Structure and Phisicochemical Properties																	
Soil Type	**Total SOC (g/kg)**	**Erosion**	**Bulk Density (g/cm3)**	**Particle Density**	**Porosity**	**Electrical Conductivity (dS/m)**	**Cation Exchange Capacity (me/kg)**	**Water Holding Capacity (gH2O/gSoil)**	**Water Retention**	**Pore Water Pressure**	**Aggregate Stability**	**Permeability**	**Organic Matter Content (g/kg)**	**pH**	**Total carbon (%)**	**Total nitrogen (%)**	**Redox Potential (mV)**	**Buffer Index**

Figure 2A. Soil structure and physicochemical properties: contains total SOC, erosion, bulk density, particle density, porosity, electrical conductivity, cation exchange capacity, water holding capacity, water retention, pore water pressure, and aggregate stability.

Soil Microbial Indicators		
Soil Basal Respiration	**Microbial Biomass (mg/kg)**	**Fungi:Bacteria ratio**

Figure 3A. Soil microbial indicators: includes basal respiration, microbial biomass, and fungi-to-bacteria ratio.

Soil Nutrients [mg/kg of dry soil]					
Ca	Na	K	Mg	N	P
9372.30	121.15	956.14	1153.15		2894.21

Figure 4A. Soil nutrients: covers calcium (Ca), sodium (Na), potassium (K), magnesium (Mg), nitrogen (N), and phosphorus (P) in mg of measured nutrient per kg of dry soil.

Heavy Metals [mg/kg of dry soil]																	
Ag	Al	A	B	Ba	Cd	Co	Cr	Cu	Fe	Hg	Mn	Ni	Pb	Sb	Ti	V	Zn
	5837.06	11.29	9.19		3.63	7.14	67.57	206.76	11202.05	2.76	232.98	27.76	542.61	4.01	93.57		672.91

Figure 5A. Heavy metals: includes silver (Ag), aluminium (Al), boron (B), barium (Ba), cadmium (Cd), cobalt (Co), chromium (Cr), copper (Cu), iron (Fe), mercury (Hg), manganese (Mn), nickel (Ni), lead (Pb), antimony (Sb), titanium (Ti), vanadium (V), and zinc (Zn)) in mg of measured nutrient per kg of dry soil.

Organic Pollutans		Salt Content [meq/100 g soil]							
TPHs	PAHs (mg/kg)	Na+	K+	Ca+	Mg+	Cl^-	SO_4^{2-}	NO_3^-	HCO_3^-

Figure 6A. Organic pollutants: such as total petroleum hydrocarbons (TPHs) and polycyclic aromatic hydrocarbons (PAHs) and Salt content: including sodium (Na+), potassium (K+), calcium (Ca+), magnesium (Mg+), chloride (Cl-), sulphate (SO_4^{2-}), nitrate (NO_3^-) measured in milliequivalents per 100 grams of soil.

II. Plant data collection template

Sample					
Case of study	Plant Sample Code	Soil Sample Code	Treatement	Assosiated genomic sample (Y/S)	Sampling Date

Figure 7A. Sample information: includes case study, sample code, treatment, sampling date, and soil type.

Plant Biochemical Stress Indicators																																							
									Hormones				Antioxidant Enzimes				Antioxidants (nmol/min/mg total protein)										Stress Biomarkers		Volatile Organic Compounds										
Plant Condition	Age	Genotype	Height (cm)	Diameter	Growth Rate	Plant Activity	Relative chlorophyll content	Total Protein Content (mg/gFW)	ABA	SA	CK	Auxins	SOD	CAT (µmol/min/mg total protein)	Apx (nmol/min/mg total protein)	POD	Vit C	Vit E	Flavonoids	Polyphenols	Carotenoids	Glutathione	Gallic acid	Chlorophyll a	Chlorophyll b	Catechins	Reactive Oxygen Species	Lipid Peroxidation	Isoprenes	Terpenoids	Acetone	Ethylene	GLVs	Others						

Figure 8A. Physiological and stress indicators: includes parameters related to plant health and stress responses, such as chlorophyll content, photosynthetic efficiency, transpiration, water use efficiency, visual stress symptoms and relevant biochemical markers.

For clarification, text on Figure 9A cells stands for:

- Plant Condition
- Age
- Genotype
- Height (cm)
- Diameter
- Growth Rate

- Plant Activity
- Relative chlorophyll content
- Total Protein Content (mg/gFW)

Hormones:

- ABA
- SA
- CK
- Axins

Antioxidant enzymes:

- SOD
- CAT (µmol/min/mg total protein)
- Apx (nmol/min/mg total protein)
- POD

Antioxidants (nmol/min/mg total protein) :

- Vit C
- Vit E
- Flavonoids
- Polyphenols
- Carotenoids
- Glutathione

- Gallic acid
- Chlorophyll a
- Chlorophyll b
- Catechins

Stress biomarkers:

- Reactive oxygen species
- Lipid peroxidation

Volatile Organic Compounds:

- Isoprenes
- Terpenoids
- Acetone
- Ethylene
- GLVs
- Others

Biomass Production and Contaminant Fate																					
Biomass (g)			Heavy Metals (µg/g (ppm))																	Organic Pollutans	
Above Ground	Below Ground	Moisture Content	Ag	Al	As	B	Ba	Cd	Co	Cr	Cu	Fe	Hg	Mn	Ni	Pb	Sb	Ti	V	Zn	TPHs	PAHs

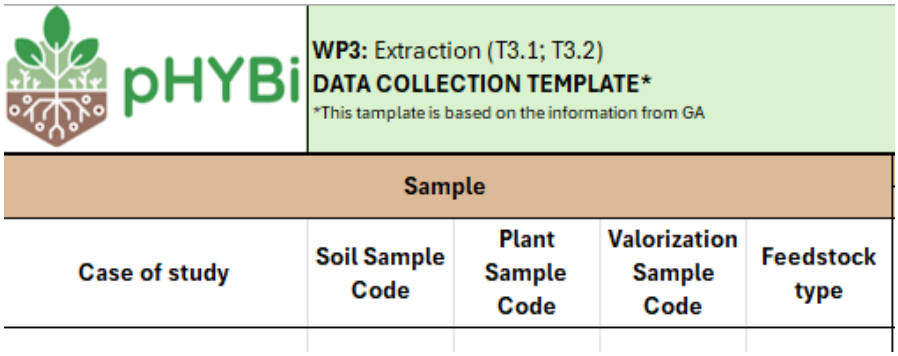
Figure 9A. Including plant growth and biomass productivity: records indicators such as plant height, stem diameter, leaf area, fresh and dry biomass yield, above- and below-ground biomass allocation, and growth rates where applicable; heavy metal uptake: includes concentrations of metals accumulated in plant tissues (e.g. Cd, Pb, Zn, Cu, Ni, Cr, Hg and others as

relevant), enabling the evaluation of phytoextraction, phytostabilisation and food-chain risk pathways; and organic contaminants: records plant-associated concentrations of organic pollutants, where applicable, supporting assessments of contaminant transfer and biomass safety for downstream valorisation.

Lignocellulosic Biomass Fractionation and Quality																		
Composition							Hemicellulose							Other Components			Fibre	
%C	%H	%N	%S	Ash Content (%)	Lignin (% (dry basis))	Cellulose (% (dry basis))	Xylan (% dry basis)	Mannan (% dry basis)	Galactan (% dry basis)	Arabinan (% dry basis)	Uronic acids (% dry basis)	Acid-soluble lignin (% dry basis)	Acetyl groups (% dry basis)	Waxes	Fats	Other (% dry basis)	Purity	Length

Figure 10A. Biomass quality and composition: includes lignocellulosic composition (cellulose, hemicellulose and lignin), moisture content, ash content and other quality parameters relevant for biomass processing and valorisation pathways assessed in WP3.

III. Extraction data collection template



WP3: Extraction (T3.1; T3.2)
DATA COLLECTION TEMPLATE*
*This tamplate is based on the information from GA

Sample				
Case of study	Soil Sample Code	Plant Sample Code	Valorization Sample Code	Feedstock type

Figure 11A. Sample information: case study, soil sample code, plant sample code, valorisation sample code, feedstock type

Pretreatment Process											
Pretreatment Process				Harvesting and Debarking			Drying			Particle Characteristics	
Harvesting Method	Debarking Method	Debarking Duration (h)	Debarking Efficiency (%)	Initial Moisture Content (%)	Drying Temperature (°C)	Drying Duration (h)	Final Moisture Content (%)	Bulk Density	Shredding Method	Particle Size Range (cm)	Particle Size Distribution (%)

Figure 12A. Organosolv sample pre-treatment information. Including harvesting and debarking: harvesting method, debarking method, duration (h), efficiency (%); drying: initial/final moisture content (%), temperature (°C), duration (h); and particle characteristics: bulk density, shredding method, particle size range (cm), distribution (%).

Extraction												
Organosolv Treatment Parameters					Separation and Recovery			Metal Content Analysis				
Solvent Type	Catalyst Type	Temperature (°C)	Reaction Time (h)	Solvent Ratio (v/v)	Cellulose Yield (%)	Hemicellulose Yield (%)	Lignin Yield (%)	Ag	Al	Cu	Fe	Zn

Figure 13A. Organosolv extraction collected information includes solvent type, catalyst type, temperature (°C), reaction time (h), solvent ratio (v/v); For separation and Recovery: cellulose yield (%), hemicellulose yield (%), lignin yield (%); and metal content analysis: Ag, Al, Cu, Fe, Zn (concentrations).

Sample								
Case of study	Soil Sample Code	Feedstock type	Plant Sample Code	Valorization Sample Code	LAB Sample code	E. coli starin	Inoculum concentration	Replicate

Figure 14A. For further textile valorisation it was included case study, soil/plant/valorisation/lab sample codes, feedstock type, E. coli strain, inoculum concentration, replicate, culture volume (mL).

Fermentation							
Culture Volume (mL)	Temperature (°C)	Fermentation Duration (h)	pH Initial	pH Final	Product Concentration (mg/L)	Quantification Method	Growth Rate (1/h)

Figure 15A. Fermentation: temperature (°C), duration (h), pH initial/final, product concentration (mg/L), quantification method, growth rate (1/h).

Textile coating											
Coationg							Coating Performance Testing				
Coating Formulation ID	Substrate Type	Substrate Material	Fabric Construction	Coating Method	Application Thickness	Curing Temp	Curing Duration (h)	Water Resistance (Y/N)	Oil Resistance (Y/N)	Wash Fastness Grade	

Figure 16A. Textile coating information, including coating: formulation ID, substrate type/material, fabric construction, coating method, application thickness, curing temp (°C)/duration (h); and coating performance testing: water resistance (Y/N), oil resistance (Y/N), wash fastness grade.