# D4.5 Repository Data and Knowledge Resources, v1

M21/MAY 2025

Funded by the European Union

This project has received funding from the Horizon Europe research and innovation programme under Grant Agreement No 101112838.

| | |
|---|---|
| **Acronym** | **SoilWise** |
| **Project Full Title** | **An open access knowledge and data repository to safeguard soils** |
| **GA number** | **101112838** |
| **Topic** | **HORIZON-MISS-2022-SOIL-01-01** |
| **Type of Action** | **HORIZON Innovation Actions** |
| **Project Duration** | **48 months** |
| **Project Start Date** | **1-Sep-23** |
| **Project Website** | **www.soilwise-he.eu** |
| **Deliverable Title** | **D 4.5 Repository Data and Knowledge Resources, v1** |
| **Delivery Time (DOA)** | **M21** |
| **Deliverable Submission Date** | **28/05/2025** |
| **Status** | **V1** |
| **Dissemination Level** | **SEN - Sensitive** |
| **Deliverable Lead** | **Nick Berkvens (EV ILVO)** |
| **Author(s)/Organisation(s)** | **Nick Berkvens (EV ILVO), Dajana Snopková (MU), Paul van Genuchten (ISRIC), Somakanthan Somalingam (WE), Cenk Doenmez (ZALF), Katerina Sachsamanoglou (GAIA), Lodovica Freyberg (ELO), Céline Blitz-Frayret (CIRAD), Giovanni L'Abate (CREA), Max Vercruyssen (DOMG - VL O)** |
| **Contributor(s)** | **EV ILVO, ISRIC, WR, ZALF, CREA, DOMG - VL O, INRAE, CIRAD, WE, GAIA, NP, ELO, MU** |
| **Peer-Reviewers** | **Fenny van Egmond (ISRIC), Dajana Snopková (MU),** |
| **Contact** | **nick.berkvens@ilvo.vlaanderen.be, radu.giurgiu@ilvo.vlaanderen.be** |
| **Work Package** | **WP4** |
| **Keywords** | **SoilWise repository, technology validation, repository population, data/metadata centralisation, demonstration phase, user cases, data & KM services** |
| **Abstract** | **This deliverable reports the activities validating the technology used to set up the first instance of the SWR. Each User Case extensively tested the tools available in this first instance during their preparatory activities for the Demonstration Phase of the first iteration. These activities equally validate the data and KM sources available in first instance and identified missing resources that still need to be harvested and made available in the SWR.** |

# Disclaimer

The content of this deliverable represents the views of the author(s) only and does not necessarily reflect the official opinion of the European Union. The European Union institutions and bodies or any person acting on their behalf are not responsible for any use that may be made of the information it contains.

In this document, the acronym 'DOMG – VL O' is used to refer to the Department of the Environment and Spatial Development, Flanders, Belgium, as per the partner's request for clarification. It's noted that in the grant agreement, the partner is identified by the acronym VL O (Vlaamse Gewest).

# List of Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| CIRAD | Centre De Cooperation Internationale En Recherche Agronomique Pour Le Developpement |
| CMS | Content Management System |
| CORDIS | Community Research and Development Information Service |
| CREA | Consiglio Per La Ricerca In Agricoltura E L'analisi Dell'economia Agraria |
| CSV | Comma Seperated Value |
| D | Deliverable |
| DG REA | Directory General Research Executive Agency |
| DM | Data Management |
| DOI | Digital Object Identifier |
| DOMG – VL O | Vlaamse Gewest |
| EEA | European Economic Area |
| ELO | European Landowners Organization |
| ESDAC | European Soil Data Centre |
| EU | European Union |
| EV ILVO | Eigen Vermogen Van Het Instituut Voor Landbouw- En Visserijonderzoek |
| FAIR | Findable, Accessible, Interoperable And Reusable |
| FAO | Food and Agriculture Organization of the United Nations |

| | |
|---|---|
| GAIA | Gaia Epicheirein Anonymi Etaireia Psifiakon Ypiresion |
| HTML | Hyertext Markup Language |
| INRAE | Institut National De Recherche Pour L'agriculture, L'alimentation Et L'environnement |
| INSPIRE | Infrastructure for Spatial Information in Europe |
| ISRIC | Stichting International Soil Reference And Information Centre |
| JRC | Joint Research Centre |
| KM | Knowledge Management |
| LLM | Large Language Model |
| ML | Machine Learning |
| MU | Masaryk University |
| NP | Neuropublic Ae Pliroforikis & Epikoinonion |
| SWR | Soilwise Repository |
| UC | User Case |
| WE | Wetransform Gmbh |
| WP | Work Package |
| WR | Stichting Wageningen Research |
| WU | Wageningen University |
| ZALF | Leibniz-Zentrum Fuer Agrarlandschaftsforschung |

# Table of Contents

# List of Tables and Figures

# 1 Executive Summary

- **Purpose:**
  The purpose of Deliverable D4.5 Repository Data and Knowledge Resources, v1 - v0.1 is to report on the validation of the technology used to establish the first instance of the SoilWise Repository (SWR) and to list relevant knowledge, data, and metadata necessary for user cases. This validation process aims to ensure that the functionalities of the SWR meet the requirements of the user cases by validating the tools and data available in the first instance during the preparatory activities for the Demonstration Phase. The scope of this deliverable includes the validation of the SWR's technology, the identification of missing resources, and the resolution of technical issues for future iterations. Within the larger project context, this deliverable supports the demonstration and evaluation of the SWR by external stakeholders, contributing to the overarching goal of improving soil health across Europe through informed decision-making and innovative solutions.

- **Intended audience:**
  This deliverable is primarily intended for technical stakeholders involved in soil data management and knowledge exchange within the European research and policy landscape. The audience includes data providers and repository managers who are responsible for maintaining and enhancing soil data infrastructure; researchers and data scientists working on interoperability and harmonization of soil information; software developers and IT specialists involved in developing data harvesting and knowledge management tools; and project managers overseeing user case implementation within the SoilWise project. Additionally, representatives from EU institutions (particularly JRC and REA) who are monitoring the technical implementation of soil data repositories, will find valuable information on the project's progress toward creating an integrated access point for soil-related data and knowledge. This document will be particularly relevant for stakeholders who need to understand the technical validation process, repository population strategies, and the current state of data and knowledge harvesting within the SoilWise Repository. Metadata curators and information specialists will also benefit from the comprehensive resources listing, which provides insights into approach of harvesting approximately 20,000 soil-related records from diverse repositories across Europe.

- **Description of the main activities:**
  The primary activities conducted during task 4.3 of the SoilWise project centred on two key areas: repository population and technology validation. For repository population, a comprehensive harvesting strategy was implemented to build a robust data repository supporting cross-sectoral and cross-institutional collaboration. This strategy utilized four distinct approaches: (1) combining CORDIS and OpenAire to discover and enrich metadata for European research outputs; (2) integrating project-specific portals such as PREPSOIL, EJP SOIL, and ISLANDR; (3) harvesting spatial and environmental data from government sources including the INSPIRE Geoportal and national catalogues; and (4) incorporating specialized data portals like ESDAC, ISRIC, FAO, and EEA. Additionally, technical components for metadata quality assessment were implemented. This resulted in the publication of two significant datasets on Zenodo: a CSV file containing approximately 20,000 metadata records of

datasets and knowledge sources related to Soil Health (DOI: 10.5281/zenodo.14851857), and multiple CSV files representing the Knowledge Graph integrated in the SWR (DOI: 10.5281/zenodo.14936020).

For technology validation, the project team initiated structured interaction between User Case (UC) partners and the development team to ensure further alignment between technical development and users about the prototypes developed towards the end of the development phase. The validation phase and the interactions for this began with the SoilWise 1st Annual Meeting in Florence (October 2025), which featured live demonstrations, collaborative sessions, and UC breakout sessions. Subsequently, bi-weekly online meetings were established throughout the Integration & Validation phase to address emerging needs and technical issues. A validation form was introduced for UC partners to submit technical bugs, blocking issues, and feature requests, during their preparations for the demonstration events.

- **Key results:**
  The validation and population of the SoilWise Repository (SWR) has yielded several important outcomes that demonstrate significant progress toward creating an integrated access point for soil data and knowledge in Europe.
  - **Result 1: Publication of a comprehensive soil data catalogue** - A collection of approximately 20,000 metadata records related to soil health has been successfully harvested, harmonized to a common metadata standard, and published on Zenodo with a permanent DOI (https://doi.org/10.5281/zenodo.14851857). This catalogue represents the first substantial centralization of scattered metadata on soil data and knowledge in Europe, drawing from over 15 different repositories including INSPIRE, CORDIS, OpenAIRE, ESDAC, and ISRIC. This achievement is significant, because it provides a unified access to soil information that was previously fragmented across multiple platforms, making it substantially easier for researchers, policymakers, and land managers to discover relevant soil data.
  - **Result 2: Development of a Soil Health Knowledge Graph** - A specialized knowledge graph focusing on soil health concepts has been created and published on Zenodo (https://doi.org/10.5281/zenodo.14936020). This graph structures insights from authoritative sources like the European Environment Agency's 2023 report on soil monitoring, creating relationships between different soil health concepts and measurements. The significance of this result lies in its ability to assist in unifying fragmented interpretations of soil health across research, policymaking, and agriculture, providing a common conceptual framework that enables more consistent approaches to soil health assessment and management.
  - **Result 3: Validation of repository technology through user testing** - The SWR's technical components have been tested by User Case partners, resulting in issues that have been systematically addressed through an agile development approach. This process has established a mechanism for continuous improvement and revealed insights about user requirements that can guide future iterations. This result is significant because it ensures the repository meets real-world needs and provides practical value to its intended users, while creating a foundation for scalable, sustainable repository development.

- **Research and practice implications:**
  The technology validation and repository population efforts documented in this deliverable have substantial implications for both research activities and practical soil management. For researchers, the centralized access to, currently, 20,000 metadata records from diverse sources eliminates significant barriers to find a wider range of relevant resources such as datasets, grey literature, policy briefs, project deliverables, enabling more robust scientific investigations. The structured Knowledge Graph provides a semantic foundation that can support more sophisticated data queries and knowledge discovery, potentially accelerating the pace of soil science research. Moreover, the validation findings reveal opportunities to further refine data and knowledge harvesting and integration approaches. The demonstrated technical integration between disparate data sources also provides a blueprint for similar efforts in other environmental science disciplines where data fragmentation remains a challenge. During the validation researchers highlighted certain needs, e.g. the need for improved temporal querying capabilities to identify studies from specific time periods and enhanced metadata augmentation that automatically extracts key research parameters like crop types, seasons, and measurement methodologies, the need for linking research publications with their underlying datasets, enabling researchers to trace from scientific papers back to the original data sources, and the need to better rank results by geographic relevance. The repository's capacity to surface previously hidden connections between datasets through semantic search and keyword linkages creates new opportunities for meta-analyses and cross-study comparisons that were previously difficult to achieve across fragmented data sources.

  For practitioners, including land managers, policymakers, and soil health professionals, these results offer practical value through improved discovery of relevant soil data and knowledge resources. The public availability of the SoilWise catalogue and Knowledge Graph creates new possibilities for developing practical soil management tools that draw on previously inaccessible or difficult-to-find information, such as the latest research findings. As the repository continues to evolve based on user feedback, it has the potential to significantly reduce the complexity of accessing soil information, enabling more informed decision-making and potentially lowering barriers to implementing soil health improvement practices across European agricultural, forest, and urban landscapes.

- **Policy implications:**
  The findings presented in this deliverable offer significant value for policymakers at both EU and national levels who are engaged with soil health governance. The centralized soil data catalogue and Knowledge Graph support the EU Mission 'A Soil Deal for Europe' by providing infrastructure to find and bring together results of projects funded by the Mission Soil, thus enhancing the use of the knowledge base generated by the Mission for a range of users, including policy makers on for example sustainable soil management practices. This includes in time easy access to summaries and policy briefs on policy relevant topics and results. This repository enables more efficient implementation of the Soil Monitoring and Resilience Directive by streamlining access to baseline and comparative data and knowledge across member states.

  The validation results also highlight the need for policymakers to continue supporting technical harmonization efforts and to consider how regulatory frameworks might better incentivize standardized data sharing. Particularly relevant to the Mission Soil objectives of reducing soil pollution and soil sealing, enhancing soil biodiversity, and improving soil structure, the repository's capacity to

bring together diverse datasets creates new opportunities for evidence-based policy development and more accurate impact assessment of soil protection measures across European territories.

- **Conclusion**:
  Deliverable D4.5 represents a milestone in the SoilWise project's journey toward establishing an integrated access point for soil data and knowledge in Europe. Through systematic harvesting of diverse data sources, creation of a structured Knowledge Graph, and validation of repository technologies with users, this work has laid a robust foundation for subsequent project phases. As SoilWise moves into its demonstration phase, the feedback collected during technology validation will directly inform future development priorities, ensuring that the repository evolves to meet stakeholder needs more effectively. The established validation mechanisms and collaboration patterns between technical teams and user case partners will continue to serve as a model for the next two development cycles, supporting the project's goal of assisting the improvement of European soil health through improved data accessibility and knowledge exchange. This deliverable not only documents significant technical achievements but also reinforces the project's commitment to co-creation and continuous improvement as elements of building sustainable digital infrastructure for soil health management.

# 2 Introduction

## 2.1 Project summary

Now more than ever, soil health is an issue that needs to be addressed urgently, as recent assessments state that 60-70% of European soils can be considered unhealthy (Bouma and Veerman, 2022). The EU Mission 'A Soil Deal for Europe', the EU Soil Strategy and the proposal for a EU Soil Monitoring and Resilience Directive (5 July 2023), aims to have 75% of EU soils healthy or significantly improved by 2030 and all soils healthy in 2050. Reaching such an ambition requires, among others, access to reliable, harmonised existing and new data and knowledge collected at local, national and EU levels to allow **informed decision-making at all scales to support the proposed EU Soil Monitoring and Resilience Directive and the EU Soil Strategy**.

The SoilWise project will provide an integrated and actionable access point to scattered and heterogeneous soil data and knowledge in Europe, making them FAIR (Findable, Accessible, Interoperable and Reusable) and improve trust, willingness, and ability to share and re-use soil data and knowledge. In three project development cycles, **co-creation and co-validation by multi-stakeholder groups are at the centre of project activities**. SoilWise recognises existing workflows and repositories for specific user needs and aims to work with them to enhance their discoverability, approachability, and interconnection. An open, modular, scalable, and extensible knowledge and data repository building on existing and new technologies will be provided while respecting data ownership, access policies and privacy. AI- and ML- techniques will be employed to interlink scattered data and knowledge, automatise the processes, infer new knowledge and increase FAIRness. **SoilWise applies infrastructure thinking instead of project thinking to design a repository for at least a decade to support EUSO evolvement accordingly**. The SoilWise repository and community are designed to be a joint starting point and common ground for countries, the European Commission, and other stakeholders to jointly guide soil and related spatial policy and informed decision-making towards the 2030 goals of the Green Deal, achieve healthy soils in 2050 and ensure broad uptake and implementation by land managers, policy, research, and industry.

All personal data acquired through SoilWise is processed in strict accordance with the relevant EU privacy regulations, highlighting our dedication to uphold to the highest standards of data privacy and security for our users.

## 2.2 Document scope

This deliverable describes the activities performed throughout T4.3 of the project, i.e. "Solutions & repository validation and population". This task was executed during the phase 03 of the project (Figure 1), i.e. the validation and integration phase, in the first iteration cycle. T4.3 validates the technical output of all the extensive development activities performed during phase 02 of the project (Figure 1), i.e. the development of the first prototype of the SoilWise Repository (SWR) in WP2, 3 and 4. The results of this task will be taken into account in preparation of phase 04 of the project (Figure 1), i.e. the demonstration of the SWR potential by all use cases (UC) and the evaluation by external stakeholders in WP5. Specific objectives of T4.3 that are reported in this deliverable are:

- Repository population: Centralizing metadata on all relevant knowledge, data, and metadata needed for UC user stories in the SWR.
- Technology validation: Ensuring the SWR's functionalities meet UC requirements.

- Iterative development: Identifying and resolving technical issues, with unresolved items listed as feature requests for future iterations.


*Figure 1 SoilWise process approach based on three development cycles (C#), each comprising four phases (P#).*

A high-level architectural diagram of the SWR is provided in Figure 2. The Harvester and Knowledge Graph, key components contributing to this deliverable, are also depicted in the diagram.

This deliverable is the first out of three foreseen versions. During phase 03 of each project development cycle, internal project partners in the UC will be consulted to validate the current tools and available data and knowledge in the SWR. The results of these validations will be reported at the end of each phase 03 in this or a next version of this deliverable (M21, M34 & M46). This deliverable coincides with Milestone 3 - End of integration and validation phase at 1st development cycle.

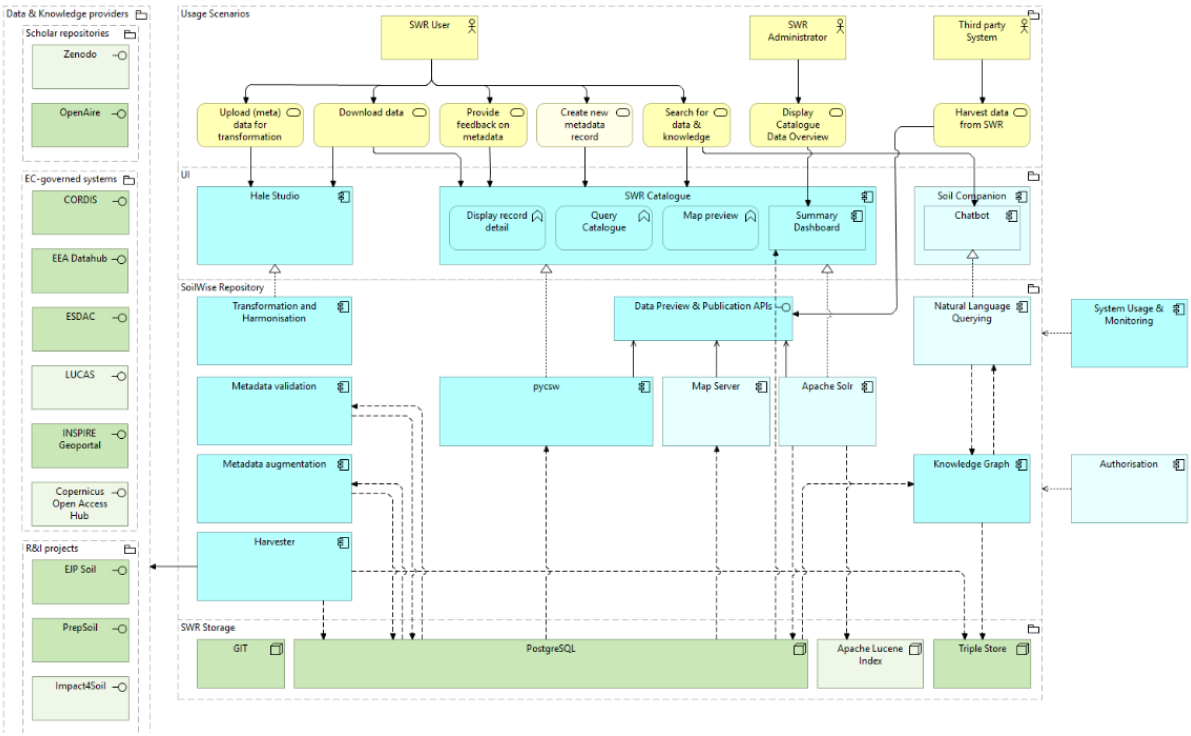



*Figure 2. High level overview of the SWR architecture (source: [https://prototype-2-0.soilwise-architecture.pages.dev/](https://prototype-2-0.soilwise-architecture.pages.dev/) )*

## 2.3 Document structure

This document is comprised of the following chapters:

- **Chapter 1** contains the executive summary
- **Chapter 2** introduces the project and the document
- **Chapter 3** explains the methodology used for harvesting and populating the repository with relevant data and knowledge
- **Chapter 4** describes the approach used to stimulating interaction between the development team and the UC, subsequently enabling the technical validation of the SWR by the UC

## 2.4 Relationship to other project deliverables

This deliverable relates to and complements the following deliverables:

- **D2.1, D2.2, D2.3, D2.4** – Developed & Integrated Data Management (DM) components, v2, v3, v4 (M18, M31, M47)

- **D3.1, D3.2, D3.3, D3.4** – Developed & Integrated Knowledge Management (KM) components, v1, v2, v3, v4 (M13, M18, M31, M47)
- **D4.1, D4.2, D4.3, D4.4** – Repository infrastructure, components and APIs, v1, v2, v3, v4 (M13, M18, M31, M47)
- **D1.3, D1.4** – Repository architecture, v1, v2 (M08, M42)
- **D1.1, D1.2** – Usage Scenarios, Requirements, v1, v2 (M6, M36)
- **D1.5, D1.6** – Repository Governance Model, v1, v2 (M21, M42)
- **D5.3, D5.4, D5.5** – Deployment and Evaluation Report, v1, v2, v3 (M21, M34, M46)
- **D7.2, D7.3, D7.4** – Open Science and Data Management plan, v1, v2, v3 (M6, M27, M48)

# 3 SWR population

## 3.1 Harvesting

### 3.1.1 Harvesting pipeline

The harvesting strategy was designed to build a comprehensive and reliable data repository on soil that supports cross-sectoral and cross-border collaboration. By systematically selecting and integrating diverse domain specific and actual data sources in consultation with the user groups, we ensure that the content collected is relevant. The approach was centered on harvesting the metadata of various types of data and knowledge resources—ranging from geospatial datasets to research publications and deliverables - to support users' needs. The harvesting approach constitutes of 4 different approaches.

1. The primary harvesting approach combines **CORDIS and OpenAire** to discover and enrich metadata for European research (project) outputs. CORDIS serves as the primary source for identifying project deliverables such as reports, articles, and datasets. Cordis itself provides limited metadata, so to address this, resources identified via CORDIS that include a DOI are further enriched using metadata from OpenAire. Project selection in CORDIS is based on two sources: a list of historic EU-funded projects maintained by ESDAC and a list of current Mission Soil projects. Currently only CORDIS records with DOIs are harvested, excluding items like progress reports (the information in those reports seems less relevant). For resources with DOIs, additional metadata is retrieved from OpenAire, which aggregates open-access content from repositories such as Zenodo and Dataverse. Metadata from OpenAire is provided in the OpenAire Research Graph (OAF) format and converted to Dublin Core. Not all DOIs found in CORDIS are listed in OpenAire, as OpenAire includes only open-access resources. Future plans include extending coverage by resolving additional DOIs via the DOI registry or Crossref.org.

2. In addition to the project results from academic repositories, the system taps into **project-specific portals**, notably those related to initiatives such as PREPSOIL, EJP SOIL, ISLANDR, BonaRes, and Impact4Soil. The PREPSOIL portal, for instance, operates via a headless CMS that occasionally provides an API for accessing

datasets, living labs, and knowledge outputs, albeit with limited metadata. When DOIs are present, they serve as entry points for enriching the data via OpenAire. Records without a DOI are provided an identifier and the minimal metatdata is harvested as is. Possible grey literature is treated like other sources but is generally filtered out by OpenAire and subsequently rejected from the catalogue. Furthermore, project updates are gathered through RSS feeds from the Mission Soil Platform's project websites, where a dedicated harvester scans and stores new entries. Projects for inclusion in this process are curated from lists maintained by ESDAC and the Mission Soil platform, both of which are scraped to populate the harvesting pipeline.

3. From the **government side**, metadata harvesting focuses on spatial and environmental data portals such as the INSPIRE Geoportal and various national catalogues. This method enables continued access to structured geospatial metadata. In some cases, these metadata records also reference DOIs, allowing for deeper integration and enrichment using OpenAire.

4. Lastly, other **specialized data portals** such as ESDAC, ISRIC, FAO, and EEA are also integral to the harvesting framework. ESDAC, for example, runs on Drupal CMS and contains valuable content like datasets, maps (EUDASM), and documents. A custom harvester scrapes this content directly from HTML, extracting metadata based on Dublin Core standards. When a DOI is found in these HTML pages, it becomes the primary identifier for the resource; otherwise, the URL is used.

More detailed information on the harvesting pipeline can be found at Harvester - Technical Documentation. We have noticed that the Mission Soil projects are not yet in the Cordis database harvested by the SWR (although they are displayed on the Cordis website), which is why no results appear in the SWR catalogue. We are discussing this issue with the Cordis team.

In addition, metadata quality was taken into consideration during the harvesting by adding certain technical components like a Link liveliness assessment and a Duplication identification (in the Harvester component). The first will assess whether the link is still 'alive', i.e. if a webpage or other digital resource will be reached when 'clicking' on the link and thus not receiving an error that the webpage or resource cannot be found. The SoilWise Catalogue reports this link status by providing a green field next to the link if it is active, a red field if it is inactive and a grey field if the link has not yet been checked. More detailed information can be found at Link Liveliness Assessment - Technical Documentation. The Duplication Identifier is designed to detect duplicate records based on identical DOIs. It also logs the source repositories from which each record originates. All source repositories that are found for a record are listed for each entry in the SoilWise Catalogue in the detailed record description page. Currently, when multiple source repositories are identified for the same record, only the first source is used for display in the catalogue, though future developments aim to merge identical records from different sources and identify additional record relations through content analysis. More detailed information can be found at Duplication identification - Technical Documentation.

In the grant agreement we envisioned the following list or repositories to harvest and this is the status of how it has been harvested (Table 1).

*Table 1. Current status of ingestion by SWR of repositories identified in Business requirement 1*

| repository | Already being harvested |
|---|---|
| INSPIRE | Inspire theme=Soil |
| EEA central data repository | Keyword=Soil |
| DataVerse | Yes, (via Cordis/OpenAire, Horizon Europe funded projects only) |
| DANS | Yes, (via Cordis/OpenAire, Horizon Europe funded projects only) |
| EJP SOIL | Yes, full |
| BonaRes | Yes, full |
| ORCaSa | Yes, full |
| ISLANDR | Yes, full |
| Prepsoil | Yes, full |
| ISRIC | Yes, full |
| AgriDataSpaces results | No |
| Sentinel products in Copernicus Open Access Hub | Partially, via EEA |
| EC (European Commission) DestinE (Destination Earth) | No |
| IACS data | INSPIRE geoporal; keyword=LPIS or GSAA |
| CORDIS | Yes, Soil related projects as indicated by JRC |
| Dutch National Georegistry | 10 records by uuid (related to usecase 1) |
| FAO | Keyword=Soil (currently offline) |

## 3.2 SWR data

On 12 February 2025 the metadata content of SWR was uploaded to Zenodo as CSV file, containing about 20.000 metadata records of datasets and knowledge sources related to Soil Health, which are imported from various repositories and, if needed, harmonised to the Dublin Core or ISO19139:2007 schema. The DOI for access is https://doi.org/10.5281/zenodo.14851857. Table 2 to Table 5 provide descriptive statistics about the distribution of the record types, topics, licenses and geographical scope for the dataset. At each project iteration a new snapshot will be added to the Zenodo record, so the population of the repository can be traced over time. We are aware there is overlap in some of categories in Table 2 to Table 5; These categories are based on the terms and classification used by the original repositories harvested. To facilitate the search functionality of the SWR we are cleaning and clustering certain categories withing the SWR using an element-matcher component (https://github.com/soilwise-he/metadata-augmentation/tree/main/element-matcher) that will be part of the metadata-augmentation module in the second development iteration

*Table 2 Number of records by record type (one type per record)*

| Record type | Number |
|---|---|
| Journalpaper | 9000 |
| Dataset | 7331 |
| Document | 1653 |
| Service | 505 |
| Series | 87 |
| Best practices and tools | 59 |
| Other | 41 |
| Education; Training material | 37 |
| Non Geographic Dataset | 37 |
| Publications ; reports | 36 |
| Journal-article | 18 |
| Scientific | 15 |
| Interview | 10 |
| Software | 10 |
| Conference deliverables | 9 |
| EU policy document | 9 |
| Policy documents | 8 |
| General articles | 7 |
| National policy document | 5 |
| Journalarticle | 5 |
| Computational notebook | 3 |
| Policy recommendations | 3 |
| Other | 2 |
| Vocational education & training | 2 |
| Image | 2 |
| Model | 1 |
| Datapaper | 1 |
| audiovisual | 1 |
| Secondary education | 1 |
| Video | 1 |

*Table 3 Number of records by topic category (records can lack a category or contain multiple categories)*

| Topic | Number |
|---|---|
| Geoscientific Information | 1987 |
| Farming | 816 |
| Environment | 380 |
| Biota | 98 |
| Climatology, Meteorology, Atmosphere | 92 |
| Imagery Base Maps, Earth Cover | 62 |
| Planning Cadastre | 35 |
| Boundaries | 31 |
| Economy | 29 |
| Elevation | 25 |
| Inland Waters | 14 |
| Society | 6 |
| Structure | 4 |
| Oceans | 3 |
| Location | 2 |
| Utilities, Communication | 2 |
| Transportation | 1 |

*Table 4 Number of records by license (records may lack a provided license or contain multiple licenses)*

| License | Number |
|---|---|
| Open Access | 5016 |
| Closed Access | 4058 |
| CC BY | 961 |
| Not Available | 651 |
| Restricted | 254 |
| Unspecified | 169 |
| No Conditions Apply | 168 |
| Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) | 138 |
| cc-by-4.0 | 113 |
| No limitations to public access | 109 |

*Table 5 Number of records by geographical scope (records can lack a scope or contain multiple scopes)*

| Scope | Number |
|---|---|
| National | 480 |
| Regional | 360 |
| European | 283 |
| Project | 154 |
| Regionale | 115 |
| Nacional | 14 |
| Lokal | 12 |
| Locale | 7 |
| Global | 6 |
| Național | 5 |
| Nazionale | 2 |
| Local | 2 |
| Nationell | 1 |
| Nacionalno | 1 |

In addition, on 4 March 2025, multiple CSV files were published representing the Knowledge Graph (KG) that is currently integrated in the SWR. A knowledge graph is a structured representation of interconnected information that enables semantic querying and reasoning across linked data sources. The SWR Knowledge Graph serves as a unified access point that links harvested metadata with various taxonomies, vocabularies, and ontologies implemented as RDF graphs, allowing users to discover connections and relationships between soil health resources that would otherwise remain fragmented. The KG is designed to assist in unifying fragmented interpretations of soil health—a concept variably defined across research, policymaking, and agriculture—by structuring insights from the European Environment Agency's 2023 report Soil Monitoring in Europe: Indicators and Thresholds for Soil Health Assessments. The DOI for access is https://doi.org/10.5281/zenodo.14936020. A draft namespace has been set for the knowledge graph as https://soilwise-he.github.io/soil-health. At this location an HTML representation of the graph has been published, to facilitate user-friendly URI lookup. The graph can also be accessed via the public SPARQL endpoint of the SoilWise triple store. More information about the knowledge graph can be found at Knowledge Graph – Technical Documentation

# 4 Technology validation of SWR

## 4.1 Interaction between UC and dev-team

At the start of the validation process, it was challenging for use case (UC) partners to envision how the SWR functionalities—originally derived from early user stories (D1.1) and integrated in the high level architecture (D1.3)—would apply in practice, as only loosely integrated components and proof of concept implementations were available at that point. To support the development of demonstration narratives and ensure better alignment between technical development and user needs, the SoilWise 1st Annual Meeting (Florence, 29–31 October 2025) was used as a platform to bring both groups together. The main activities included:

1. Live demonstration – A walkthrough of the SWR prototype showcasing its core functionalities.
2. Collaborative sessions – Discussions between the technical and user groups to clarify needs, methodologies, and roles.
3. UC breakout sessions – UC partners per stakeholder group collaborated with one of three technical teams (SWR Catalogue, Hale Transformation & Harmonization Tools, AI/LLM Functionality) for in-depth discussions on functionality, assumptions, and requirements.

Following the Annual meeting, a bi-weekly one-hour online meeting was established between UC partners and the dev team (developer team working on the technical development of SWR) to address emerging needs. This continued throughout the entire Integration & Validation phase 03 and the Demonstration phase 04 to support continued collaboration (Figure 1), address technical issues, and identify missing data sources and features for the next cycles. Topics covered included:

- Comprehensive walkthrough and explanation of the various functionalities (catalogue, harvester, Hale transformation/harmonization tools etc.), including a detailed breakdown of the components developed at that point, an analysis of the key attributes of each functionality, and an update on their implementation status.
- Metadata augmentation through keyword clustering/matching
- Technical bug reporting and SWR validation
- Demonstration preparations

## 4.2 Validation of SWR technologies

### 4.2.1 Internal validation within consortium

A reporting form (Figure 3) was introduced for UC partners to submit technical bugs, blocking issues for demonstration preparations, and feature requests that were identified during their preparations for the demonstration. At the time of finalizing the deliverable, 81 issues had been reported via the form of which half were technical bugs, 10 where issues blocking the demonstration preparations and the rest where feature requests for the further development of the SWR.

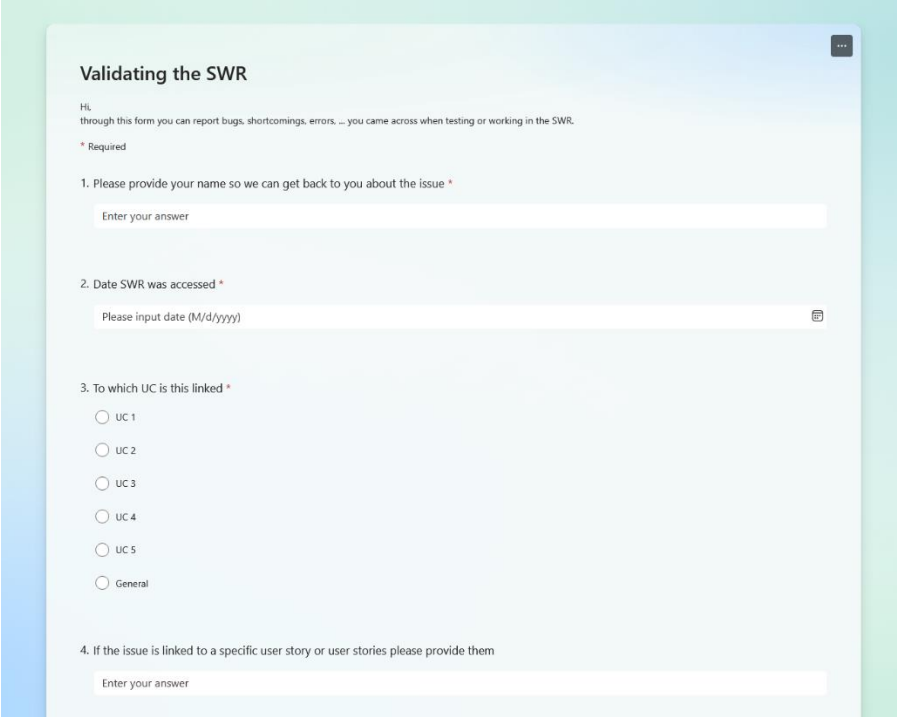
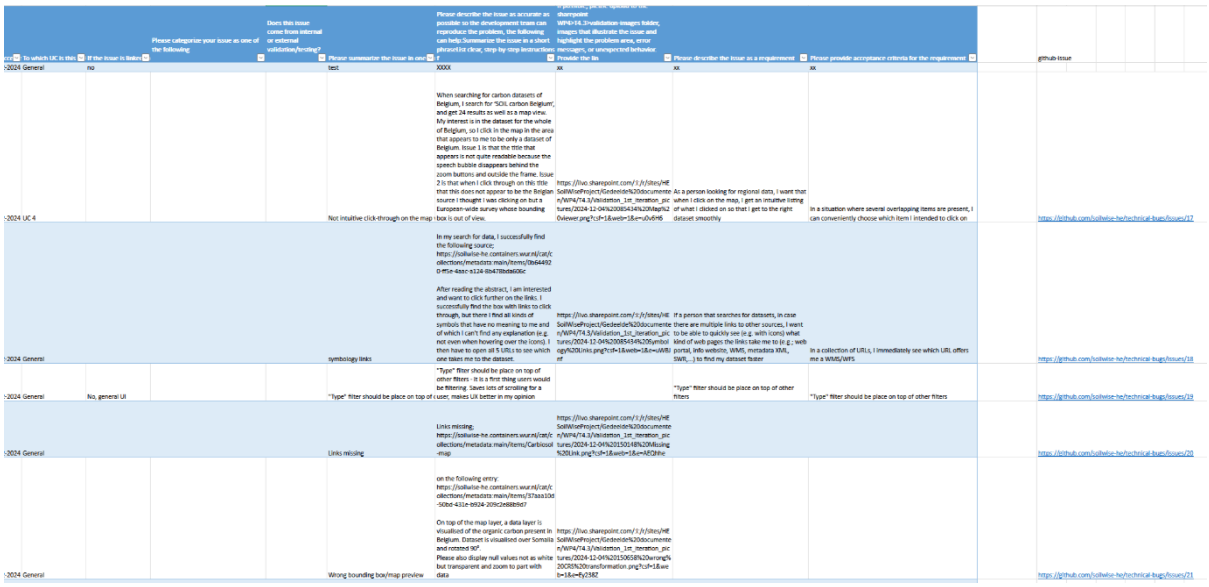



*Figure 3 Validation form used to collect feedback from UC during the validation and integration phase of the project.*

Reported issues were compiled into an Excel tracking file (Figure 4), regularly reviewed to:

- Convert technical bugs and blocking issues into GitHub issues for developers to address and resolve. Although development activities for phase 02 have officially concluded, ongoing efforts with an agile approach continued to support the demonstration activities of phase 04.
- Collect feature requests as input for WP1 activities in the next iteration



| | To which UC is this | If the issue is linked | Please categorize your issue as one of the following | Does this issue come from internal or external validation/testing? | Please summarize the issue in one | Please describe the issue as accurate as possible so the development team can reproduce the problem, the following can help: Summarize the issue in a short phrase/list clear, step-by step instructions | If possible, add the location in the sharepoint WP4>T4.3>validation-images folder, images that illustrate the issue and highlight the problem area, error messages, or unexpected behavior | Please describe the issue as a requirement | Please provide acceptance criteria for the requirement | github issue |
|---|---|---|---|---|---|---|---|---|---|---|
| -2024 | General | no | | | test | XXXX | xx | xx | xx | |
| -2024 | UC 4 | | | | Not intuitive click-through on the map | When searching for carbon datasets of Belgium, I search for 'SOIL carbon Belgium', and get 24 results as well as a map view. My interest is in the dataset for the whole of Belgium, so I click in the map in the area that appears to me to be only a dataset of Belgium. Issue 1 is that the title that appears is not quite readable because the speech bubble disappears behind the zoom buttons and outside the frame. Issue 2 is that when I click through on this title that this does not appear to be the Belgian source I thought I was clicking on but a European-wide survey whose bounding box is out of view. | https://ivo.sharepoint.com/:i:/r/sites/mE SoilWiseProject/Gedeelde%20documente n/WP4/T4.3/Validation_1st_iteration_pic tures/2024-12-04%20085434%20Map%20 Viewer.png?csf=1&web=1&e=u0v6H6 | As a person looking for regional data, I want that when I click on the map, I get an intuitive listing of what I clicked on so that I get to the right dataset smoothly | In a situation where several overlapping items are present, I can conveniently choose which item I intended to click on | https://github.com/soilwise-he/technical-bugs/issues/17 |
| -2024 | General | | | | symbology links | In my search for data, I successfully find the following source; https://soilwise-he.containers.wur.nl/cat/c ollections/metadata:main/items/0b644920 -ff5e-4aac-b124-8b478bdeb05c<br><br>After reading the abstract, I am interested and want to click further on the links. I successfully find the box with links to click through, but there I find all kinds of symbols that have no meaning to me and of which I can't find any explanation (e.g. not even when hovering over the icons). I then have to open all 5 URLs to see which one takes me to the dataset. | https://ivo.sharepoint.com/:i:/r/sites/mE SoilWiseProject/Gedeelde%20documente n/WP4/T4.3/Validation_1st_iteration_pic tures/2024-12-04%20085434%20Symbol ogy%20Links.png?csf=1&web=1&e=uWBJ nf | If a person that searches for datasets, in case there are multiple links to other sources, I want to be able to quickly see (e.g. with icons) what kind of web pages the links take me to (e.g.: web portal, info website, WMS, metadata XML, SWR,...) to find my dataset faster | In a collection of URLs, I immediately see which URL offers me a WMS/WFS | https://github.com/soilwise-he/technical-bugs/issues/18 |
| -2024 | General | No, general UI | | | "Type" filter should be place on top of | "Type" filter should be place on top of other filters - It is a first thing users would be filtering. Saves lots of scrolling for a user, makes UX better in my opinion | | "Type" filter should be place on top of other filters | "Type" filter should be place on top of other filters | https://github.com/soilwise-he/technical-bugs/issues/19 |
| -2024 | General | | | | Links missing | Links missing. https://soilwise-he.containers.wur.nl/cat/c ollections/metadata:main/items/Carbicool -map | https://ivo.sharepoint.com/:i:/r/sites/mE SoilWiseProject/Gedeelde%20documente n/WP4/T4.3/Validation_1st_iteration_pic tures/2024-12-04%20150148%20Missing%20Link.png?csf=1&web=1&e=MZQhhv | | | https://github.com/soilwise-he/technical-bugs/issues/20 |
| -2024 | General | | | | Wrong bounding box/map preview | on the following entry: https://soilwise-he.containers.wur.nl/cat/c ollections/metadata:main/items/37baa10d -50bd-431e-b924-209c2e88b9d7<br><br>On top of the map layer, a data layer is visualised of the organic carbon present in Belgium. Dataset is visualised over Somalia and created 90°.<br>Please also display null values not as white but transparent and zoom to part with data | https://ivo.sharepoint.com/:i:/r/sites/mE SoilWiseProject/Gedeelde%20documente n/WP4/T4.3/Validation_1st_iteration_pic tures/2024-12-04%20150658%20wrong%20Bbox%20transformation.png?csf=1&we b=1&e=Ey2582 | | | https://github.com/soilwise-he/technical-bugs/issues/21 |

*Figure 4. Excel file tracking all content uploaded to the validation form*

### 4.2.2 Validation with external stakeholders

To gather feedback from external stakeholders during phase 03 and phase 04 of the project development cycle, a Hotjar widget was added to the SWR (Figure 5). External stakeholders could provide their feedback by responding to the following questions posed.

- Did you find what you were looking for today?"
  - Yes / No
  - [If No] → "What were you hoping to find?"
- "How easy was it to find the data or knowledge item you needed?"
  - Scale: Very easy → Very difficult
  - Optional text box: "What made it easy or difficult?"
- How satisfied were you with the downloaded file(s)?"
  - Scale 1–5
  - Optional: "Was the format or content as expected?"
- "Is the information in the item pages (title, description, etc) clear and useful?"
  - Yes / No
  - [If No] → "What could be improved?"
- "How useful is this repository for your work or research?"
  - Scale 1–5
  - Optional text box: "What kind of work or use case are you using it for?"
- "What kind of data or content would you like to see added?"
  - Open text
- "If you could improve one thing about this repository, what would it be?"
  - Open text

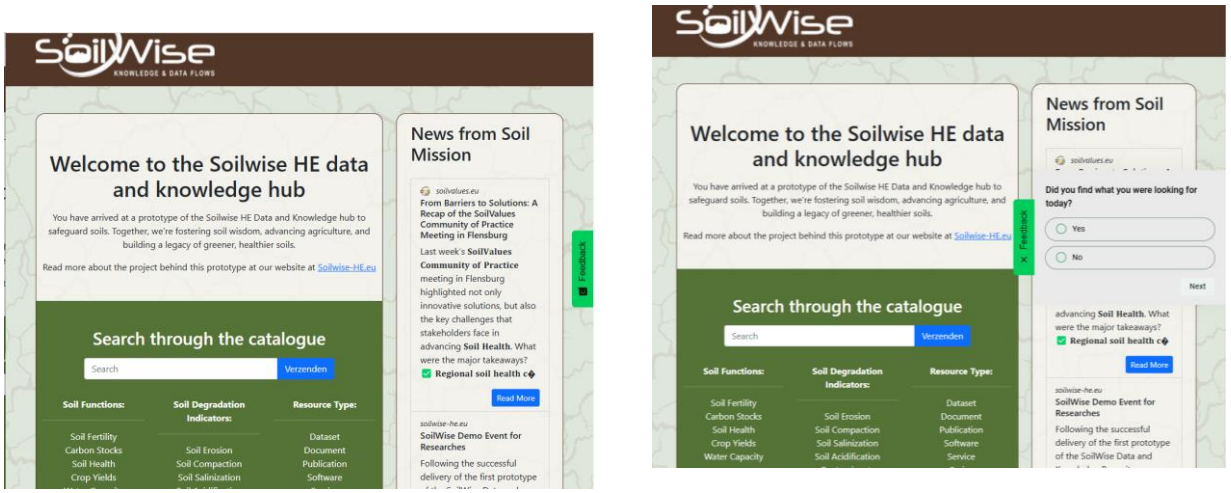



*Figure 5 Hotjar widget set up on the SWR catalogue to collect feedback from external stakeholders.*

### 4.2.3 Validation performed within Use Cases

The UC partners each validated the SWR in the following way.

**UC1 Soil health performance indicators for Land Managers**

As part of the validation process for UC01, ELO organized meetings with farmers, land managers and landowners to test the SWR. WR showcased the functions of the Catalogue and the Chatbot and our 'testers' were able to ask specific farming-related questions they wanted the repository to answer. The conclusion of these initial tests was then registered in the validation form. Once the development team was able to resolve reported blockers in the validation form, two public demonstrations were made. During these, the audience was able to see the functions and capabilities of the Catalogue and Chatbot (presented by WR) and ask questions or express their opinions on the functionalities.

**UC2 Leveraging a network of Soil R&I Knowledge and Data**

The validation of SoilWise repository within UC2 has been conducted during 3 phases of the use case activities:

- While creating narratives for demonstrations of Mission Soil Data and Knowledge management Cluster in Brussels (November 2024) and for the demonstration webinar for research community (April 2025), extensive tests on SWR (catalogue functionality) and Hale Studio (data transformation functionality) have been performed to ensure that a given resource or entry is suitable for demonstrating the functionalities.
- During the 5 hands-on sessions organized with Mission Soil projects stakeholders (February 2025): based on a testing form created in this use case and provided 1 week in advance, stakeholders were requested to test the SWR for their Mission Soil project with a facilitator and a developer of SoilWise. Both took notes of feedback, centralized by the use case leader.
- During the live demonstration webinar of April 2025: while showcasing SWR functionalities, live questions (multiple choice and open-ended) were addressed to the audience with a Mentimeter.

These validation activities lead to different kinds of outputs: improvement of the pre-existing functionalities and envisions of new functionalities. In the first case, bugs were reported in direct discussions with the development team or by writing GitHub issues on SoilWise and on Hale Studio or by filling the reporting form (Figure 3). In the second case, envisioned functionalities were written in the form of user stories, starting the second co-design phase.

**UC3 Policy Making & Evaluation to safeguard soil**

The validation activities for UC3 focused on evaluating how SoilWise supports policy makers and governmental bodies in improving soil-related data sharing, reporting efficiency, and transparency. The goal was to assess the system's capacity to reduce administrative burden and improve accessibility to standardized, policy-relevant soil data. The validation process included the following key components:

- Scenario Testing: Participants were asked to simulate these workflows using the SWR, focusing on functionalities such as data discovery, metadata inspection, export formats, and harmonization support.
  During these testing scenario's, bugs and improvement ideas were recorded in the reporting form.
- Feedback Collection: Specific attention was paid to how well SoilWise supported cross-border comparability and adherence to existing standards like INSPIRE and ISO 19115.
- This validation confirmed the relevance of SoilWise in supporting policy reporting but also highlighted key areas for improvement, such as clearer guidance for non-technical users, automated data transformation pipelines, and better integration with national open data portals.

**UC4 Enhanced capacities of Public Authorities and LLs actors**

UC2 contributed to the validation of the SWR (Soil Water Repository) by developing a test case evaluating key functionalities such as data discovery and download processes. The validation specifically assessed the search functionality within the SWR to improve dataset discovery and catalogue integration. While searching for datasets to support spatial modelling on agricultural productivity, the system successfully identified relevant data—primarily from the BonaRes Repository—demonstrating its potential to streamline access to curated resources. However, the process also revealed several limitations aligned with the predefined user story requirements, providing valuable insights for platform enhancement; Feedback was reported in the validation reporting form (Figure 3).

**UC5 Repository for new products, technologies and services**

In the context of establishing the demonstration scenario for the UC5 demo event, the SWR was extensively tested as part of the validation process. The primary goal was to identify any technical issues or bugs and to provide suggestions for potential improvements to enhance the SWR. The validation process was not limited to addressing only the specific requirements of the user story/demonstration scenario. It was further expanded to include a more organic and thorough exploration of the platform's functionalities. Testing was conducted by utilizing all the various pathways and components developed up to that point, while also adopting a more intuitive approach from a user's perspective. This allowed for a comparison between the two different methods,

helping to identify what works best and what is more efficient in each case. In addition, to validate transformation and harmonization of soil data, various activities were carried out, such as apprenticing sessions, end-to-end demonstrations of the transformation process of selected datasets to stakeholders using Hale Studio as well as validation against INSPIRE standards and format changes involving GeoPackage.

# 5 Conclusions

In conclusion, this deliverable has successfully validated the technology used to set up the first instance of the SoilWise Repository (SWR), centralizing metadata on relevant knowledge, data, and metadata to support user cases. The collaborative efforts between the development team and user case partners have ensured that the SWR's functionalities have been tested and start meeting user requirements, paving the way for future iterations and enhancements. Key activities included live internal showcasing of the SWR prototype, collaborative sessions to clarify needs and methodologies, and UC breakout sessions for in-depth discussions on functionality, assumptions, and requirements. Additionally, bi-weekly online meetings were established to address emerging needs, covering multiple relevant topics. Lastly, several digital forms where set up to collect and further manage feedback coming from the project partners during their validation activities of the SWR.

The implications of this deliverable are significant for the project's next steps, as it provides a solid foundation for the demonstration phase and external stakeholder evaluation. By addressing technical issues and identifying feature requests for future iterations, this deliverable contributes to the project's overarching goal of assisting to improving soil health across Europe through informed decision-making and innovative solutions. This deliverable not only validates the current technology but also sets the stage for continued collaboration and development, ensuring that the SWR remains an important tool for research, soil health management and policy-making in Europe.

# 6 References

Bouma, J., & Veerman, C. P. (2022). Developing Management Practices in: “Living Labs” That Result in Healthy Soils for the Future, Contributing to Sustainable Development. Land, 11(12), 2178.