D2.1 Developed & Integrated DM components

M13/SEPTEMBER 2024

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	D2.1 Developed & Integrated DM components Repository architecture
Delivery Time (DOA)	M13
Deliverable Submission Date	30/09/2024
Status	V1
Dissemination Level	PU - Public
Deliverable Lead	Tomáš Řezník (MU)
Author(s)/Organisation(s)	Dajana Snopková (MU), Tomáš Řezník (MU), Paul van Genuchten (ISRIC), Nick Berkvens (EV ILVO)
Contributor(s)	MU, WU, ZALF, EV ILVO, DOMG – VL O, CREA, ISRIC, WE, NP
Peer-Reviewers	Céline Blitz-Frayert (CIRAD) , Giovanni L`Abate (CREA)
Contact	tomas.reznik@sci.muni.cz
Work Package	WP2
Dissemination level	Public
Keywords	data management components, prototype, technical documentation
Abstract	The purpose of this deliverable is to describe the process of co-design and implementation of Data Management components of the SoilWise repository and the current version of technical documentation comprising a description of the component's functionalities after the first development iteration.

Disclaimer

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
API	Application Programming Interface
CI-CD	Continuous Integration and Continuous Delivery / Development
CIRAD	CENTRE DE COOPERATION INTERNATIONALE EN RECHERCHE AGRONOMIQUE POUR LEDEVELOPPEMENT
CREA	CONSIGLIO PER LA RICERCA IN AGRICOLTURA E L'ANALISI DELL'ECONOMIA AGRARIA
CSW	Catalog Service for the Web
DM	Data Management
EJP	European Joint Programme
ESDAC	European Soil Data Centre
ETS	Executable Test Suite
EU	European Union
EUSO	EU Soil Observatory
EV ILVO	EIGEN VERMOGEN VAN HET INSTITUUT VOOR LANDBOUW- EN VISSERIJONDERZOEK
FAIR	Findable, Accessible, Interoperable and Reusable
INSPIRE	Infrastructure for Spatial Information in Europe
ISO	International Organization for Standardization
ISRIC	STICHTING INTERNATIONAL SOIL REFERENCE AND INFORMATION CENTRE

KM	Knowledge Management
M	Month
ML	Machine Learning
MU	Masaryk Univerzity
NP	NEUROPUBLIC AE PLIROFORIKIS & EPIKOINONION
OGC	Open Geospatial Consortium
PSI	public sector information
PU	Public
RDF	Resource Description Framework
SOS	Sensor Observation Service
STAC	Spatio Temporal Asset Catalogue
SWR	SoilWise Repository
T	Task
DOMG – VL O	VLAAMSE GEWEST
WE	WETRANSFORM GMBH
WCS	Web Coverage Service
WFS	Web Feature Service
WMS	Web Map Service
WP	Work Package

WPS	Web Processing Service
WR	STICHTING WAGENINGEN RESEARCH
WU	WAGENINGEN UNIVERSITY
ZALF	LEIBNIZ-ZENTRUM FUER AGRARLANDSCHAFTSFORSCHUNG

Table of Contents

List of Tables and Figures

1 Introduction

1.1 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 (Bouman, 2022). The EU Mission 'A Soil Deal for Europe', the EU Soil Strategy and the proposal for a 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 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.

1.2 Document scope

This deliverable describes the methodology used in the process of co-designing of SoilWise repository (SWR) Data Management (DM) components and their implementation and delivery of their source code. As such, it documents the SWR developments within the first iteration, out of the four planned iterations. The first iteration has been designed as a proof-of-concept, based on mostly independent developments of SoilWise partners. Nevertheless, all the technical development achievements are tightly coupled with the SoilWise mission, scope and deliverables published so far. Among others, this deliverable (1) comprises achievements that are framed by the architecture defined in D1.3, (2) respects and follows stakeholders' demands/suggestions described in the D1.1, (3) has been shaped by a number of in-depth discussions with the European Commission, DG REA/JRC/AGRI, and (4) dozens of Mission Soil Horizon Europe projects.

Technical documentation was built in the form of an external web page accompanying this deliverable to capture the complexity of the SoilWise achievements. This documentation consists of a description of functionalities, detailed technical specifications, and interfaces between all modular components that will be employed in the

SWR. This fits with one of the work package (WP) 2 goals to detail the design and develop the DM components of SWR, more specifically, within the task (T) 2.1 Design of the data technology components (M7-M37) and T2.2 Implementation and deployment of data components (M7-M46).

Initially, a high-level architectural design was established, which was detailed in deliverable D1.3 and further refined through collaborative efforts in development clusters and monthly meetings. The implementation phase adopted an agile strategy (see Chapter 2), where the development process was broken down into six iterative cycles, each lasting four weeks. Key activities included populating and refining a Product Backlog, planning sprints, developing components, and conducting reviews and retrospectives to ensure tasks met the Definition of Done (DoD) criteria.

Delivery of the first SoilWise Repository prototype comprises the following Data Management components: (1) **Harvester** component responsible for harvesting metadata and knowledge resources, harmonizing metadata, and identifying duplicates, (2) **Catalogue** allowing users to query the harvested resources, download metadata, and access map previews, and also links to knowledge sources, (3) **Metadata Validation** component focusing on evaluating metadata completeness, structure validation, and link liveliness assessment, (4) **Transformation and Harmonization** component handling tasks such as metadata standards transformation, CRS transformation, data restructuring, and format conversion, (5) **Repository Storage** component responsible for the storage of raw and augmented metadata, user-enhanced content, and knowledge graphs. Detailed documentation, source code, and access points are provided for each component in Chapter 3.

This deliverable was created hand-in-hand with the other SoilWise technical Work Packages to seamlessly document all the technical developments and achievements so far. As such, this D2.1 document together with D3.1 Developed & Integrated KM components (M13) and D4.1 Repository infrastructure, components, and APIs (M13), serves as a delivery of the first prototype of SWR. The delivery of the first SWR prototype comprises (1) source code of developed technical components (DM + KM), (2) functionality description in the form of external technical documentation, (3) Release notes (listed in D4.1), (4) three separate accompanying methodological and management overview documents.

The second version of this D2.1 deliverable will be delivered in M18 (February 2025), elaborating on the results of the integration & validation project phase of the first project iteration cycle. The third version of this deliverable will be delivered in M31 (end of the second project iteration cycle), and the fourth version of this deliverable two months before the end of the project in M46 (end of the third project iteration cycle). All the mentioned deliverable versions will describe the technical developments of the SoilWise Data Management components. This technical level will be accompanied by a visionary level report, a complex mid- and long-term strategy, within the D2.5 Report on strategy for FAIRness on soil data in two iterations (M27 and M42).

1.3 Document structure

This document is comprised of the following chapters:

- **Chapter 1** provides an introduction to the project and the document
- **Chapter 2** explains the methodology used in the process of designing and implementing SoilWise Repository Data Management components
- **Chapter 3** presents a list of Data Management components that are present in the first delivered prototype. It includes an external link to the technical documentation that further details the component's functionality and technical specifications.
- **Chapter 4** details the technical documentation

Technical documentation is available at https://prototype-1-0.soilwise-documentation.pages.dev/, and a PDF exported version is also available on demand as a non-editable version saved at the date of the deliverable. The documentation includes links to the current versions of the components' code on Github, when available.

1.4 Relationship to other project deliverables

This deliverable relates to and complements the following deliverables:

- D2.2, D2.3, D2.4 – Developed & Integrated DM components, v2, v3, v4 (M18, M31, M47)
- D3.1, D3.2, D3.3, D3.4 – Developed & Integrated 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 – Repository architecture, v1, v2 (M08, M42)
- D1.1, D1.2 – Usage Scenarios, Requirements, v1, v2 (M6, M36)
- D1.5, D1.6 – Repository GM, v1, v2 (M21, M42)
- D4.5, D4.6, D4.7 – Repository Data and Knowledge Resources, v1, v2, v3 (M21, M34, M46)
- D5.3, D5.4, D5.6 – 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)

1.5 Relationship to project tasks

This deliverable relates to the following project tasks:

- T1.3 Requirements, Validation framework and Rolling plan – will feed and update the design and implementation of DM components (T2.1 and T2.2)
- T1.4 Define SoilWise Architectural Design – will be considered in the design and implementation of DM components (T2.1 and T2.2)
- T1.5 Define SoilWise Multi-Stakeholder governance model – will be considered in the design and implementation of DM components (T2.1 and T2.2)
- **T2.1 Design of the data technology components**
- **T2.2 Implementation and deployment of data components**
- T2.3 AI and ML for data findability and accessibility – will extend the design and implementation of DM components (T2.1 and T2.2)
- T2.4 Strategy for FAIRness on soil data – will be followed in the design and implementation of DM components (T2.1 and T2.2)
- T3.1 Design of the KM components – will consider and align with the design of DM components (T2.1)
- T3.2 Implementation and deployment of knowledge component – will consider and align with the implementation of DM components (T2.2)
- T4.1 Repository digital infrastructure for deployment and delivery – will integrate implementation of DM components (T2.2) and KM components (T3.2)
- T4.2 Interfaces for access, sharing, population and integration with EUSO – will be considered in the design and implementation of DM components (T2.1 and T2.2)

- T4.3 Solutions & repository validation and population – will validate designed and implemented DM components (T2.1 and T2.2)
- T5.2 User Cases implementation and demonstration – will demonstrate the functionality of implemented DM components (T2.2)
- T7.2 Technical and Scientific Management – prepares an Open Science and Data Management Plan (D7.2 – D7.4) describing the data management level of handled (meta)data resources.

2   Methodology

2.1   Co-design of Data Management components

The design of the Data Management component was preceded by a high-level design of the architecture, which took place within the framework of T1.4 Define SoilWise Architectural Design. Its first version was the subject of deliverable D1.3 Repository architecture, v1. Within the framework of these activities, the original plan was already exceeded, and the next level of architecture - the individual functions of the components - was partially detailed in the technical documentation. The architectural design resulted from the co-design project iteration that is described in Chapter 3 of deliverable D1.1. Usage scenarios, requirements.

After the submission of D1.3, work on the more detailed level of architecture design continued separately in the so-called "development clusters", each having its dedicated technical component. However, the original T1.4 working group continued to meet monthly at combined WP2+WP3+WP4 meetings to ensure the harmony of activities and easier integration of all technical components in the future. Meanwhile, the architectural design and the technical documentation were updated to reflect the recent design suggestions and development activities.

2.2   Implementation of Data Management components

The development process for the Data Management components was based on an agile project management practice in one-month sprints. This approach, illustrated in Figure 1 and detailed in Chapter 7 of D1.1 Usage scenarios, requirements, follows an iterative workflow:

1. Adding user stories, requirements, acceptance criteria, and functionality descriptions to the Product Backlog[1]. For the first iteration, the Product Backlog was populated with the vision scenarios (see Chapter 3).
2. Reviewing and prioritising tasks in the Product Backlog.
3. Planning the sprint and defining the Sprint Backlog[2].
4. Developing technical component functionality according to the tasks in the current sprint.
5. Verifying acceptance and DoD (Definition of Done) criteria during the sprint review.
6. Discussing improvement practices during the sprint retrospective meeting.
7. Repeating the process for the next sprint.

---

[1] a prioritized list of functionality which a product should contain and that drives the developmental work

[2] a list of work items or tasks a development team plans to complete during a project sprint
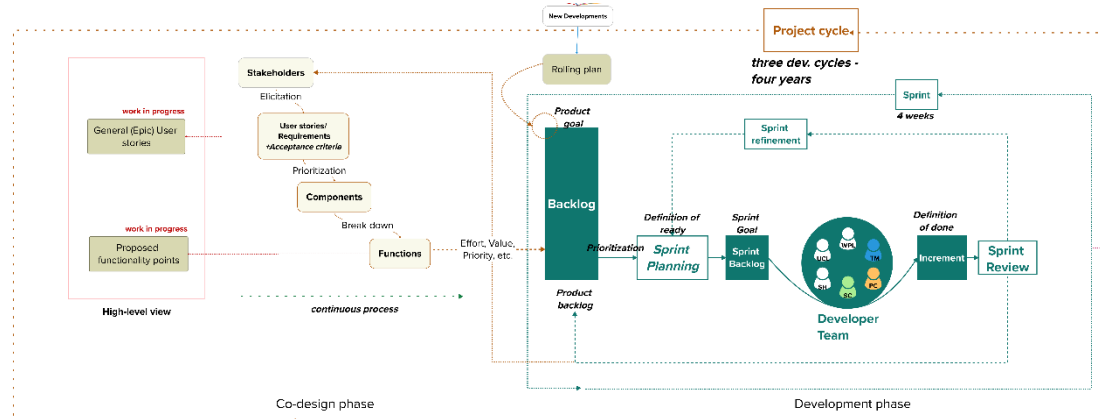
Figure 1. The agile process flow applied for the development of the Data Management components

Coordination of sprints and developments was managed using GitHub and a dedicated GitHub environment at https://github.com/soilwise-he. Separate repositories were created for each technical component, containing both the developed code and tasks (GitHub issues) addressed during the sprints. A GitHub project Soilwise Sprint Backlog (Figure 2) was used to plan and track the developments of each sprint, and another project Soilwise Sprint Refinement was established for group discussion topics during weekly sprint refinement meetings. A member was assigned responsibility for the development of each project component. Collaborators involved in the component's development were also identified, forming the developmental team for that component. A list of leads, collaborators, and repositories for each component is available at https://github.com/soilwise-he/Soilwise-Project-Backlog/blob/main/components_and_repo.md.
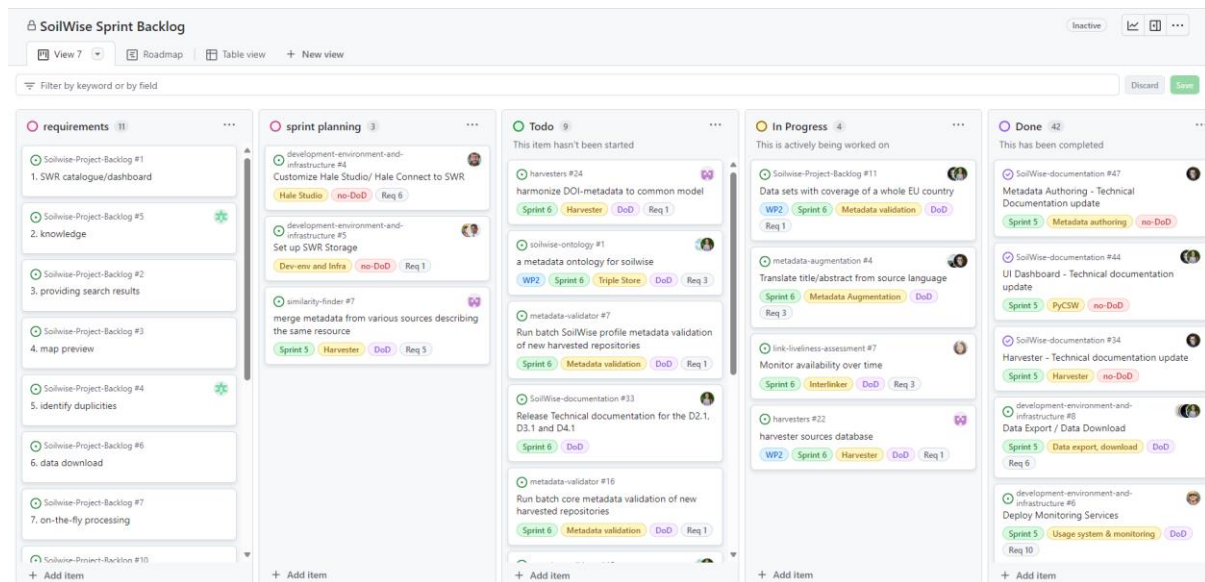

Figure 2 Illustration of the SoilWise Sprint Backlog board containing 11 functionalities (left requirements column) and associacted tasks the development team plans to work on or is working on during each sprint (other columns).

Each sprint lasted four or five weeks, with six sprints conducted for the first iteration of development activities (Figure 3). Each sprint began with a Sprint Planning meeting where the component leads selected tasks for the current sprint and added them to the Sprint Backlog. Each task included a description of the problem, the expected developmental work, and one or more DoD criteria to determine task fulfillment. Weekly Sprint

Refinement meetings involved component leads discussing developmental issues collectively. At the end of each sprint, a Sprint Retrospective meeting was held to evaluate the status of all developed tasks and confirm that completed tasks met the DoD criteria.
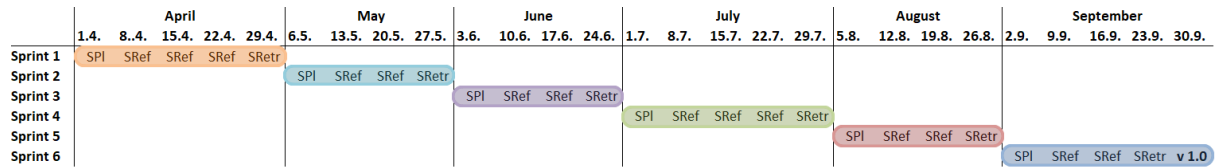

Figure 3 Gantt Diagram of first prototype development sprints, including Sprint Planning meetings (SPl), Sprint Refinement meetings (SRef), and Sprint Retrospective meetings (SRetr)

As project members continued to elicit and refine user stories, requirements, and acceptance criteria, a finalised Product Backlog was not available for the first prototype development in this iteration. The completed tasks in this first iteration pertain to the 10 functionalities defining the first prototype, described in chapter 2.5 of deliverable D1.3 Repository Architecture. An 11th functionality was added during the sprint. These 11 functionalities are documented in the GitHub project SoilWise Sprint Backlog (Figure 2) for guidance. Each functionality is further divided into sub-functionalities with criteria for evaluating their fulfillment during development. While the 11 functionalities and their derived sub-functionalities drive the development activities in this first iteration, in the next iterations, a fully-fledged Product Backlog comprising the user stories, requirements, acceptance criteria, and functionalities is foreseen.

3 Data Management components

This chapter lists the Data Management components that are subject to the delivery of the first SoilWise repository prototype, including the links to their source code and access point for users (if applicable). Knowledge Management components, which represent another part of the delivered prototype, are listed in Chapter 3 of parallel deliverable D3.1 Developed & Integrated KM components. Since data and knowledge in the SoilWise Repository are closely linked, in some cases, the components cannot be clearly assigned to only one Work Package as the developments have been shared by both WP2 and WP3. Therefore, in the list of Data Management component functions, we indicate that they belong to Knowledge Management components and vice versa.

A more detailed description of the components' functionality is described in the Technical Documentation, see more detail in Chapter 4. A summary in the form of release notes is also available in Deliverable D4.1. - Repository infrastructure, components, and APIs, v1.

Name:	Harvester
Version:	0.1.0
Functions:	Harvest metadata resources Harvest knowledge resources *(part of KM)* Metadata harmonization Metadata RDF Turtle serialization RDF to Triple Store Duplication identification *(part of KM)*
Documentation:	https://prototype-1-0.soilwise-documentation.pages.dev/technical_components/ingestion/
Source code:	https://github.com/soilwise-he/harvesters/releases/tag/v0.1.0
Access point:	not applicable
Name:	Catalogue
Version:	0.1.0
Functions:	Query catalogue CSW API OGC API Catalogue Data download (AS IS) Map preview User Engagement
Documentation:	https://prototype-1-0.soilwise-documentation.pages.dev/technical_components/catalogue/
Source code:	https://github.com/soilwise-he/pycsw/releases/tag/v0.1.0
Access point:	https://soilwise-he.containers.wur.nl/cat/
Name:	Metadata Validation
Version:	0.1.0

Functions:	Metadata profile validation Link liveliness assessment
Documentation:	https://prototype-1-0.soilwise-documentation.pages.dev/technical_components/metadata_validation/
Source code:	Metadata profile validation: https://github.com/soilwise-he/metadata-validator/releases/tag/v0.1.0/ Link liveliness assessment: https://github.com/soilwise-he/link-liveliness-assessment/releases/tag/v0.1.0
Access point:	https://data.soilwise.wetransform.eu/#/home (authorisation needed, account can be provided on demand) Link liveliness assessment: https://api.soilwise-he.containers.wur.nl/linky/docs
Name:	**Metadata Augmentation**
Version:	0.1.0
Functions:	Automatic metadata generation Translation module *(part of KM)*
Documentation:	https://prototype-1-0.soilwise-documentation.pages.dev/technical_components/metadata_augmentation/
Source code:	https://github.com/soilwise-he/metadata-augmentation/releases/tag/v0.1.0
Access point:	not applicable
Name:	**Transformation and Harmonization**
Version:	5.3.0
Functions:	Manual metadata upload Metadata transformation CRS transformation Data restructuring Format transformation Codelist mapping Units of measurements conversion Download interoperable metadata
Documentation:	https://prototype-1-0.soilwise-documentation.pages.dev/technical_components/transformation/
Source code:	https://github.com/halestudio/hale/releases/tag/v5.3.0
Access point:	https://data.soilwise.wetransform.eu/#/home (authorisation needed, account can be provided on demand)
Name:	**Repository Storage**
Version:	Postgres release 12.2 Virtuoso release 07.20.3239
Functions:	Storage of user-enhanced content – GIT *(part of KM)* Storage of raw harvested metadata – PostgreSQL / vector DB Storage of augmented metadata – PostgreSQL / vector DB

	Storage of augmented, linked metadata, knowledge graph - Triple Store *(part of KM components)*
Documentation:	https://prototype-1-0.soilwise-documentation.pages.dev/technical_components/storage/
Source code:	not applicable
Access point:	GIT: https://github.com/soilwise-he/soilinfohub/discussions PostgreSQL: https://pgadmin.isric.org Triple Store (Virtuoso): https://sparql.soilwise-he.containers.wur.nl/sparql

4 Technical documentation

The documentation of SoilWise Repository architecture is maintained in the public GitHub repository: https://github.com/soilwise-he/SoilWise-documentation. It comprises a description of functionality, detailed technical specifications, and interfaces between all modular components and is structured according to the main technical components. Note that the SWR documentation is also a living environment that is continuously updated during the development process. For this deliverable, a stable release is published at: https://prototype-1-0.soilwise-documentation.pages.dev/, and a PDF exported version is also available on demand as the non-editable version made at the date of the deliverable submission.

References

Bouman, 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.