# AI4SoilHealth

# Training datasets with manuals for application in different OS apps, including the AI4SoilHealth PWA & downstream mobile apps deployed

# D4.14

**Version 1.1**
**15.12.2025**

**Lead Authors: Giulio Genova (OGH), Octavian P. Chiriac (OGH), and Xuemeng Tian (OGH)**

**Reviewed by: Tomislav Hengl (OpenGeoHub), Fatemeh Hateffard (Stockholm University)**

**Action Number:** 101086179

**Action Acronym:** AI4SoilHealth
**Action title:** Accelerating collection and use of soil health information using AI technology to support the Soil Deal for Europe and the EU Soil Observatory

| HISTORY OF CHANGES | | |
|---|---|---|
| **Version** | **Publication date** | **Changes** |
| 1.0 | 10.12.2025 | Initial version |
| 1.1 | 12.12.2025 | Added soil datasets names, minor text corrections |

# Contents

Executive Summary 3
1. Introduction 4
1.1. Aim 4
2. Datasets and manuals 5
2.1. Training datasets 5
2.2. Soil properties maps and ancillary data 6
2.3. Manuals and application guidelines 7
2.4. Tools for spatial heterogeneity assessment 8
3. Final considerations 12
4. References 12

## Executive Summary

Deliverable 4.14 provides a comprehensive, ready-to-use package of training datasets, soil property maps, manuals, and tools to support soil health assessment across all13 platforms. Standardised datasets are publicly available, while data harmonisation is ongoing. Eleven soil property maps developed in this project and over 150 auxiliary environmental layers are hosted in EcoDataCube. The Soil Health Data Cube manual and harmonisation sheet provide clear documentation, tutorials and metadata, ensuring transparency and usability. To improve accessibility, several tools have been developed: the AI4SH dashboard for researchers, the AI4SH app for farmers, and the EcoDataCube portal for policymakers. Future updates will incorporate new data from pilot sites and broader user feedback, further improving data quality and accessibility.

# 1. Introduction

The AI4SoilHealth project aims to co-design, create and maintain an open access European-wide digital infrastructure for continuous monitoring of soil health metrics compiled using state-of-the-art Artificial Intelligence (AI) methods combined with new and deep soil health understanding and measures.
Specifically, WP4 delivers state-of-the-art science-based methods for soil health assessment and provides the best possible solutions to access in-situ data and observations following all the standards and procedures defined in the Data Management Plan. This deliverable is linked to T4.6 ("Integration and harmonization of in-situ and ancillary observations") and T6.4 ("Collect and summarize the user feedback on Soil Health tools, digital apps and implementation plan").

## 1.1. Aim

The purpose of Deliverable 4.14 is to provide a comprehensive, user-ready package of analysis-ready training datasets together with documentation and operational manuals that enable their seamless use across different operating systems, applications, and platforms within the AI4SoilHealth ecosystem. Specifically, it takes into account user feedback from task 6.4 and supports task 4.6 by addressing the following objectives:

- Produce analysis-ready training datasets

  Compile, harmonize, and format the datasets required for AI/ML model development so they are ready for direct use (e.g., in model training, validation, and benchmarking). These datasets should be fully documented, quality-checked, and accessible through a public repository.

- Provide harmonized soil properties maps and ancillary datasets

  Ensure that soil properties maps and curated ancillary data, such as land cover, land use, and climate, are made available in forms that can be consistently linked in space and time to soil observations. This enables robust integration across multiple data sources.

- Provide manuals and application guidelines

  Develop clear, practical documentation and training materials that explain how to use the datasets and tools

- Deliver tools for spatial heterogeneity assessment

  Provide tools, workflows, and examples that characterise the spatial variability of Earth observation (EO) satellite data around in situ sampling points. This helps to assess the degree of correspondence, representativeness, and reliability between remote sensing and field measurements.

## 2. Datasets and manuals

The available soil datasets were used to create soil properties maps, which were then integrated with other complementary layers and made available for use on different platforms. This process is summarised below.

### 2.1. Training datasets

The project addresses the challenge of integrating heterogeneous soil datasets collected across Europe. Due to differences in formats, metadata structures, measurement methods, and units, raw soil data cannot be directly compared or used for large-scale modeling. To overcome this, we employ a two-step workflow: standardization and harmonization. The standardization process combines different datasets in using only one common structure following next steps:

- Rename columns to common names (e.g., SOC, pH, clay).
- Convert coordinates to the same system (WGS84).
- Add metadata (dataset ID, site, country, year, depth).
- Save in one format (Parquet).

The harmonization process transforms the values in the same unit of measurement by following next steps:

- Align measurement methods (e.g., different lab techniques).
- Convert units to the same scale (e.g., % for SOC, % for clay).
- Remove unrealistic values.

The downloadable standardized data can be found in Soil Health Data Cube website and it contains data from:

- Bodemkundig Booronderzoek (BHR-P) - NL
- Bodemkundig Informatie Systeem (BIS) - NL
- Global Land Cover Estimation (GLanCE) - Global
- Geocradle - EU
- INFOSOLO - PL
- Land Use and Coverage Area frame Survey (LUCAS) - EU
- Precision Liming Soil Datasets (LimeSoDa) - Global
- Tidal Marsh Soil Organic Carbon (MarSOC) - Global
- SOils DAta Harmonization database (SoDaH) - Global

The dataset is not harmonized yet because the process is ongoing.

### 2.2. Soil properties maps and ancillary data

The layers for EcoDataCube are also available in SpatioTemporal Asset Catalogs (STAC) (Fig. 1) related to AI4SoilHealth project are:

- Soil clay content (%)
- Soil silt content (%)
- Soil sand content (%)
- Soil core bulk density (g/cm3)
- Soil organic carbon content (‰)
- Soil organic carbon density (kg/m3)
- Soil pH in CaCl2
- Soil pH in water
- Soil extractable potassium (ppm)
- Sol type dominant class
- Soil type probability

In addition, there are more than 150 layers available that comprises environmental, land cover, terrain, climatic, soil and vegetation layers covering the continental Europe at relatively fine spatial resolutions (30-m to 1-km).

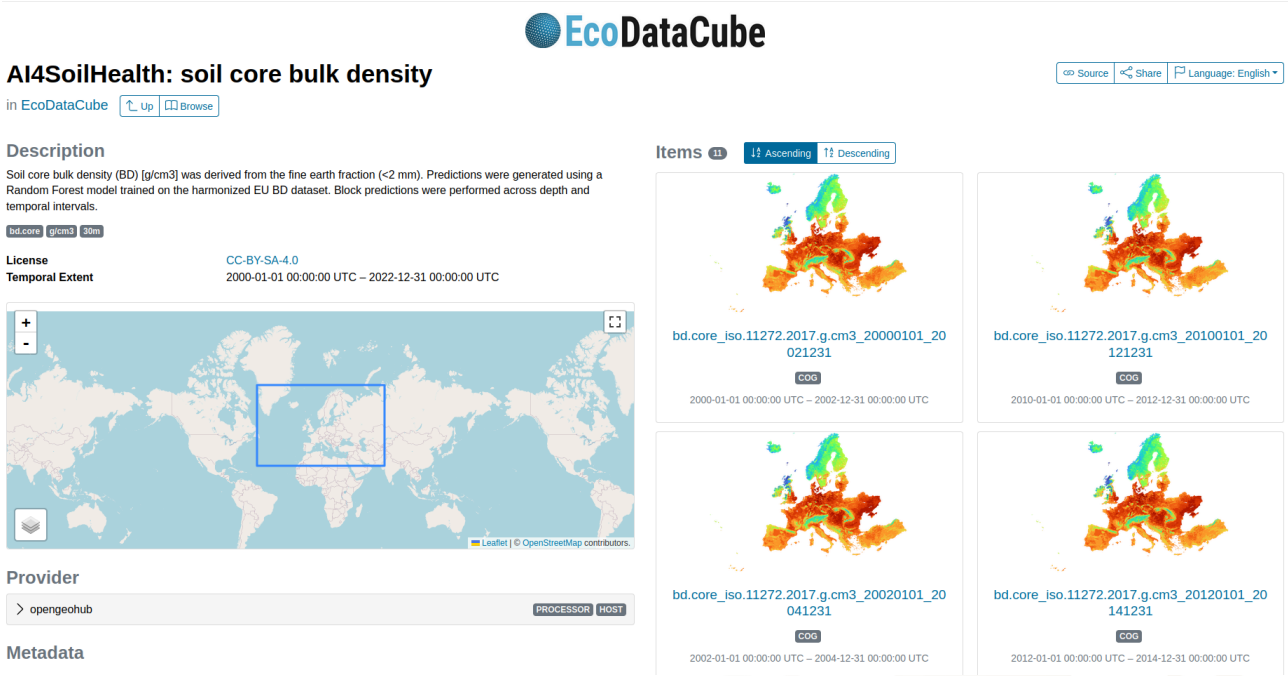


Fig. 1: Example of available map in STAC EcoDataCube (soil core bulk density)

## 2.3. Manuals and application guidelines

Soil Health Data Cube manual provides information about the available data, the downloadable formats, specifications regarding soil health assessment process, and data use tutorials (Fig. 2).

# Welcome

DOI 10.5281/zenodo.13838797

This document provides information about the **Soil Health Data Cube** for pan-EU (SHDC4EU). The main purpose of the SHDC is to be as a platform for more detailed computing i.e. to estimate trends in important soil health indicators (e.g. Bare Soil Fraction, vegetation cover, chemical soil properties and similar). The SHDC is available via S3 (Simple Storage Service) and STAC as open data, which means that any researcher across EU can access data directly using rstac or similar, and fetch values / aggregate per polygon or farm. list will be continuously updated and extended. **This document is continuously updated and new layers are continuously added.**

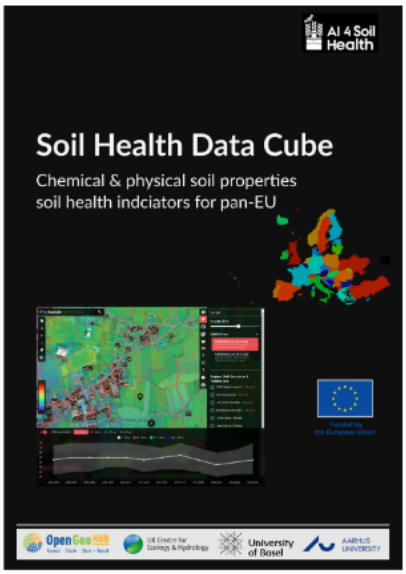


Fig. 2: Soil Health Data Cube manual overview

In the Soil Health Data Cube website the harmonization sheet was also included. The overview sheet provides high-level information including data sources, data availability across sources and properties, and the spatio-temporal distribution of the entire dataset. Each property sheet is named after a soil property involved in the harmonization process . The dataset comprises 25 sources and has more than one million data points (Fig. 3). Each one contains next columns:

- src: Indicates the origin or source from which the data is obtained.
- method description: Provides the original description of the measurement method as documented, detailing how the data was measured.
- data count: The number of data entries available for this measurement method from the specified source.
- quality score: A designated score indicating the reliability of the data, helping determine whether it is suitable for use.
- conversion formula: Specifies any necessary conversion formula to standardize the data.
- conversion reference: Cites the literature or source that supports the choice of the conversion formula.

- notes: Any supplementary information or remarks relevant to the data or its context.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Region | Property availability | Current state | soc (g/kg) | total.n (g/kg) | carbonates (g/k | clay (%) | silt (%) | sand (%) | ph.h2o | ph.cacl2 | bd.fe (g/cm3) | bd.tot (g/cm3) | cf.mass (%) |
| 2 | GO | SoDaH | Added | 15 | 11 | 0 | 40 | 40 | 40 | 16 | 0 | 4 | 0 | 0 |
| 3 | GO | MarSOC | Added | 9731 | 1788 | 0 | 0 | 0 | 0 | 0 | 0 | 8849 | 0 | 0 |
| 4 | GO | GLanCE | Added | 1131 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | EU | LUCAS | Added | 62882 | 62770 | 55042 | 26222 | 26222 | 26222 | 62789 | 62789 | 5846 | 5999 | 26083 |
| 6 | EU | Geocradle | Added | 1287 | 0 | 1186 | 1290 | 1227 | 1227 | 258 | 63 | 0 | 0 | 0 |
| 7 | EU | GEMAS | Added | 4012 | 0 | 0 | 4026 | 4008 | 4008 | 0 | 4025 | 0 | 0 | 0 |
| 8 | DE | BZE-LW | Added | 16334 | 16330 | 0 | 15867 | 15867 | 15867 | 16334 | 16334 | 16328 | 12720 | 0 |
| 9 | UK | UK-CS | Added | 2750 | 1072 | 0 | 0 | 0 | 0 | 2780 | 0 | 2706 | 0 | 2851 |
| 10 | UK | GMEP | Added | 1363 | 1363 | 0 | 0 | 0 | 0 | 1362 | 1362 | 1362 | 0 | 0 |
| 11 | CH | swiss.nabo | Added | 9115 | 55 | 1362 | 9947 | 9445 | 9312 | 2030 | 7214 | 4 | 44 | 0 |
| 12 | EE | estonia.kese | Added | 203 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 13 | SI | Pedoloski | Added | 6054 | 6054 | 0 | 6054 | 6054 | 6054 | 6054 | 6054 | 0 | 0 | 0 |
| 14 | PT | INFOSOLO | Added | 8063 | 6376 | 9647 | 9923 | 9923 | 9923 | 9721 | 0 | 0 | 1521 | 8346 |
| 15 | ES | ParcelasCOS | Added | 1600 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1600 | 1600 |
| 16 | ES | ParcelasINES | Added | 22158 | 0 | 0 | 21954 | 0 | 0 | 0 | 0 | 0 | 21954 | 21954 |
| 17 | ES | Castilla.y.Leon | Added | 21846 | 4798 | 17453 | 19697 | 19696 | 19697 | 0 | 0 | 0 | 145 | 0 |
| 18 | ES | LDBF | Added | 986 | 986 | 77 | 894 | 894 | 894 | 902 | 0 | 0 | 450 | 160 |
| 19 | NL | Netherland.BHR-P | Added | 323 | 275 | 0 | 47 | 0 | 148 | 148 | 148 | 0 | 0 | 0 |
| 20 | NL | BIS | Added | 855325 | 5651 | 0 | 631179 | 12911 | 12917 | 0 | 0 | 0 | 0 | 0 |
| 21 | DK | dk.pilot | Added | 131 | 0 | 121 | 0 | 0 | 0 | 0 | 0 | 131 | 131 | 0 |
| 22 | CZ | bmp | Added | 992 | 731 | 1424 | 0 | 0 | 0 | 1888 | 1888 | 0 | 0 | 0 |
| 23 | HR | MultiOne | Added | 1480 | 722 | 1165 | 1429 | 1429 | 1429 | 721 | 0 | 720 | 0 | 0 |
| 24 | HU | HunSSD | Added | 32 | 32 | 35 | 35 | 35 | 35 | 35 | 0 | 0 | 27 | 0 |
| 25 | HU | HSDOS | Added | 5487 | 5487 | 5487 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 26 | GO | LimeSoDa | Added | 428 | 0 | 0 | 428 | 0 | 0 | 36 | 352 | 0 | 0 | 0 |

Fig. 3: Soil data harmonization sheet overview

## 2.4. Tools for spatial heterogeneity assessment

### AI4SH Dashboard

The AI4SoilHealth dashboard allows the consortium partners to upload the data from the in situ pilot sites collected in task 4.1 (*General strategy, prioritization, inventory of in-situ data and analysis of data gaps*). Users can utilize advanced filtering and data manipulation techniques to customize their analyses, while various presentation options, such as maps and charts, enhance intuitive interpretation (Mornar and Bagić Babac, 2024).

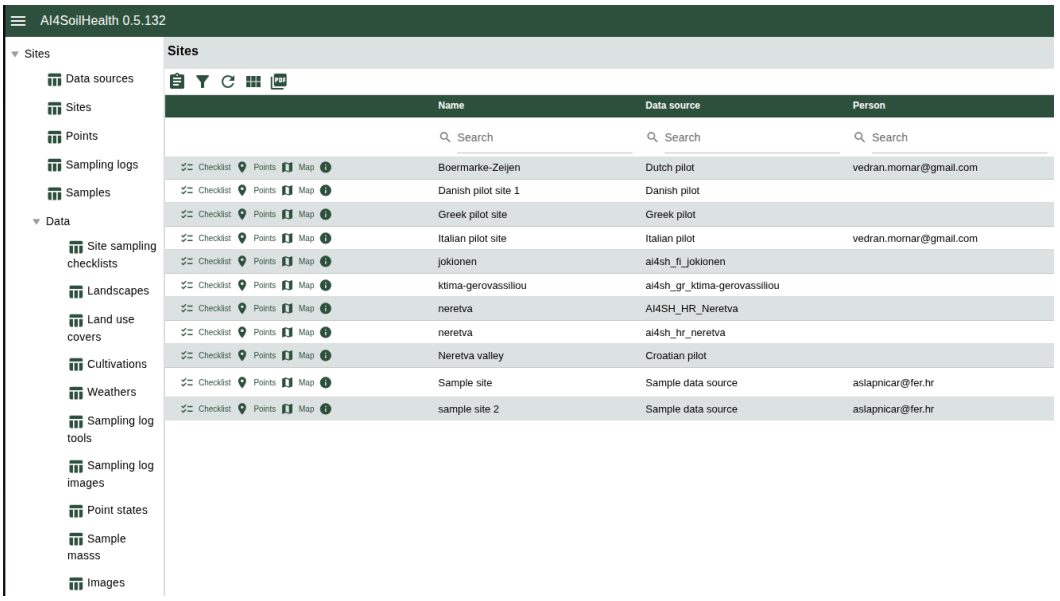


Fig. 4: Ai4SoilHealth Dashboard visualization

**EcoDataCube**

EcoDataCube portal hosts environmental layers representing dynamics of land cover, land use, climate, relief, potential and actual vegetation, forest cover and dynamics, soil variables from the Soil Health Data Cube and various long-term estimates of trends quantifying land degradation and land restoration processes (Witjes et al., 2023; Tian et al., 2024).


Fig. 5: EcoDataCube interface

**AI4SH App**

User feedback was taken into account during the development of the app (Fig. 6). This included data access, format, visualisation, and uncertainty estimation. In the next stage of users co-design and feedback, the project will broaden feedback collection to include a wider range of stakeholders, starting with farmers, researchers and advisors at pilot sites (Cordelia et al., 2024).

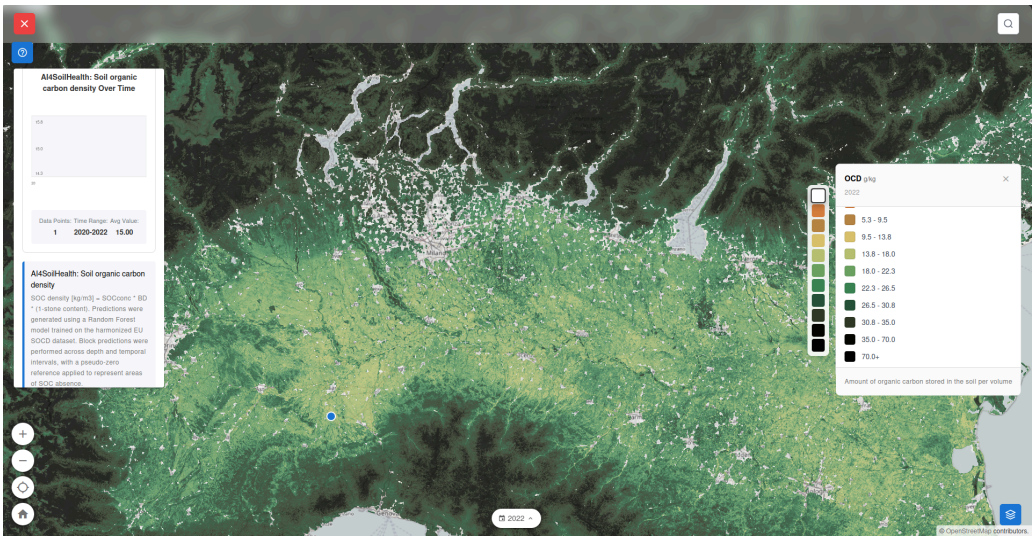


Fig. 6: AI4SoilHealth App interface

In addition the app allows users to easily retrieve trend analysis (Fig. 7) using either a point location or field boundaries from a number of dynamic soil properties and ancillary data such as biomass, tillage, crop, land cover, soil organic carbon, etc.

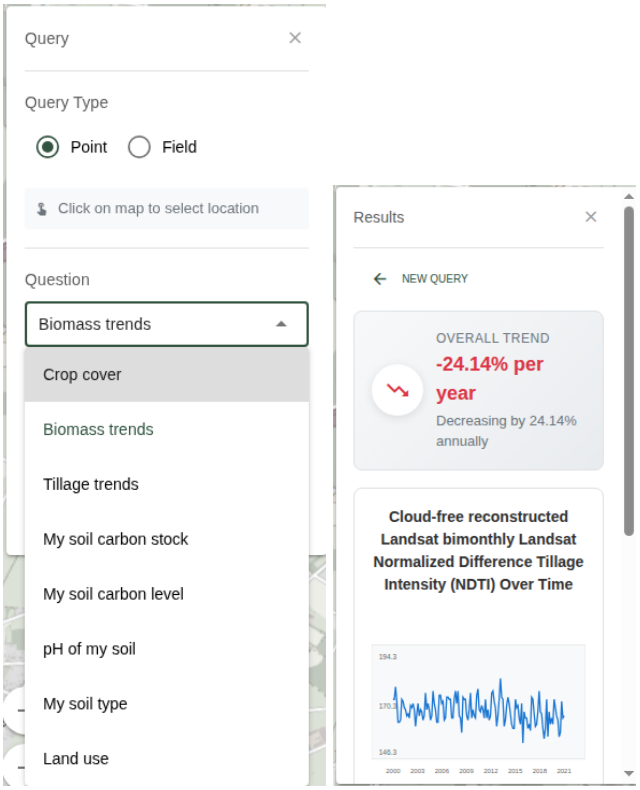


Fig. 7: AI4SoilHealth App query and results of trend analysis. Example of NDVI

## 3. Final considerations

This report has outlined the datasets and maps relating to soil properties produced as part of the AI4SoilHealth project, as well as the relevant documentation for their use. The tools available for visualising soil property layers have also been presented. In the next steps, the layers will be updated with new datasets from the pilots' sites and taking into account the users feedback. This will allow us to increase the quality of the data and share them on different platforms to maximize the number of end users.

## 4. References


Cordelia, H., Epelde, L., Stanton, K. J., Lo Moriello, C. S., Gumbricht, T., Morar, V., Minarik, R., Robinson, D. (2024). D6.7 Users Co-design and Feedback Report v1, AI4SoilHealth Horizon Europe project no. 101086179

Mornar, V. & Bagić Babac , M. (2024). D4.8 Interface to Soil Health Cube scientific data visualization module, AI4SoilHealth Horizon Europe project no. 101086179

Tian, X., Consoli, D., Witjes, M., Schneider, F., Parente, L., Şahin, M., ... & Hengl, T. (2024). Time-series of Landsat-based bi-monthly and annual spectral indices for continental Europe for 2000–2022. Earth System Science Data Discussions, 2024, 1-49. https://doi.org/10.5194/essd-17-741-2025.

Witjes, M., Parente, L., Križan, J., Hengl, T., & Antonić, L. (2023). Ecodatacube. eu: analysis-ready open environmental data cube for Europe. PeerJ, 11, e15478. https://doi.org/10.7717/peerj.15478.