EJP SOIL
European Joint Programme

Towards climate-smart sustainable management of agricultural soils

Deliverable 6.6

Geodatabase on agricultural properties including SOC and Agricultural soil functional properties related to water and nutrients – a metadata catalogue

Due date of deliverable: M48
Actual submission date: 31.01.2024

GENERAL DATA

Grant Agreement: 862695
Project acronym: EJP SOIL
Project title: Towards climate-smart sustainable management of agricultural soils
Project website: www.ejpsoil.eu

Start date of the project: February 1st, 2020
Project duration: 60 months
Name of lead contractor: INRAE

Funding source: H2020-SFS-2018-2020 / H2020-SFS-2019-1
Type of action: European Joint Project COFUND

DELIVERABLE NUMBER:	6.6
DELIVERABLE TITLE:	Geodatabase on agricultural properties including SOC and Agricultural soil functional properties related to water and nutrients – a metadata catalogue
DELIVERABLE TYPE:	Other
WORK PACKAGE N:	WP6
WORK PACKAGE TITLE:	Supporting harmonised soil information and reporting
DELIVERABLE LEADER:	F.M. van Egmond, Wageningen Research
AUTHOR:	P. van Genuchten, F.M. van Egmond, M. Fantappiè
DISSEMINATION LEVEL:	PU
DOI:	https://doi.org/10.18174/649307
	Publisher Wageningen University & Research
CC license:	CC-BY 4.0

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

Table of Contents

List of Tables

List of Figures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

List of acronyms and abbreviations

API	Application Programming Interface
CSV	Comma-separated values
CSW	Catalogue Service for the Web
DG AGRI	Directorate General for Agriculture and Rural Development
DG-ENV	Directorate General for Environment
DOI	Digital Object Identifier
EJP	European Joint Programme
ESDAC	European Soil Data Centre
EU	European Union
EUSO	EU Soil Observatory
FAIR	Findable, Accessible, Interoperable, Re-usable
GSP	Global Soil Partnership
HTML	Hyper Text Markup Language
INSPIRE	Infrastructure for Spatial Information in Europe
ISRIC	ISRIC-World Soil Information
JRC	Joint Research Centre
MCF	Multimedia Container Format
MDME	Model Driven Metadata Editor
MIT	Massachusetts Institute of Technology
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OGC	Open Geospatial Consortium
REA	European Research Executive Agency
TRL	Technical Readiness Level
WMS	Web Mapping Service
WP	Work Package
YML	Yet Another Markup Language

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

1. Aim

Geographical soil data (any data with a geolocation/coordinates) is essential for many reporting, mapping and monitoring applications to assess the past and current status of soil (health) and allow efficient and targeted soil research in Europe. However, at present it is not easy to find, access and re-use the relevant data across national borders or from the research community. This is because the data is either not registered somewhere or stored dispersed in various national and research repositories and is regularly missing the metadata needed to judge its usefulness. The datasets available under the theme Soil in the INSPIRE geoportal unfortunately do not resolve this issue. Most of these datasets do not really refer to the soil domain as described by the INSPIRE technical guidance: i.e. datasets on soil cover, that is, land use. Furthermore, several of the available soil datasets in EJP SOIL countries, published in the deliverable D6.1, D2.2, were not retrievable in the INSPIRE geoportal since they were not made accessible by the data owners in an INSPIRE compliant format. A first step to remedy this is therefore to **improve the findability and re-usability of soil geographical data by creating a dedicated soil metadata catalogue, including metadata on relevant soil datasets and improving on this metadata**. This activity will create one access point for various actors in the soil domain to search and find relevant soil data for policy support and research purposes.

This EJP SOIL soil metadata catalogue compiles metadata from national repositories, and in the future main, research repositories like ZENODO and Cordis could be added to enrich it.

The catalogue contains metadata and includes links to, and information about the datasets. It includes mechanisms for improvement of metadata through a participatory approach through 'Edit on Git' (see How does it work and how to operate, and the tutorials). The catalogue is provided with searching tools as described in the following paragraphs of the present deliverable. Tutorials are supplied on how to use the catalogue, and these tutorials were included in the EJP SOIL T5.5 capacity building courses on soil data assimilation (6-8 April 2022 and 24-26 January 2023).

The current catalogue can be accessed at: https://catalogue.ejpsoil.eu. Supplementary material is available at: https://ejpsoil.eu/soil-research/soil-data-monitoring-mapping-and-modelling. The tutorial is available at: https://ejpsoil.github.io/soildata-assimilation-guidance.

2. For whom

The catalogue provides an overview and allows better findability of relevant authoritative and research soil datasets that can be used by the EU Mission Soil, EU Mission Soil current and future projects, other soil research projects, single researchers and by local, regional, national and European authorities. There are no restrictions on access to the metadata and it is therefore of value to the world-wide soil community and adjacent domains. In 2025 (M60 of EJP SOIL), the EJP SOIL catalogue will be transferred to the European Soil Observatory.

3. Added value

The added value of the catalogue is that it (can) includes metadata of soil datasets that are stored in the main repositories in Europe, like ZENODO, Cordis, national repositories, ESDAC and the INSPIRE

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

data portal if the administrators choose to. It therefore provides quicker and more effective searches to users in Europe. The catalogue does not aim to compete with existing major (meta)data catalogues (like ISRIC, ESDAC, GSP, etc), but it is complementary and aimed specifically at soil data in Europe. The metadata content and software will be made available for reuse to ESDAC/EUSO after EJP SOIL programme completion (January 2025).

Metadata stored in different formats and annotated using different standards can be uploaded and is standardised to a single metadata format (the INSPIRE ISO19115) in an automated way. The accompanying metadata templates encourage the provision of additional metadata specific to the soil domain, so users can judge the suitability of a dataset for their purpose even better, stimulating re-use of data.

The setup of the catalogue specifically allows participation of users and producers of data to improve metadata in the catalogue. When a user identifies an error or needs more metadata for a dataset, he/she can indicate this to the data owner through the system. The data owner then receives a notification with the request to address this and thereby improve the metadata.

A preview of datasets is shown if the format of the dataset allows (e.g. preview image or WMS service) and searches are possible using keywords, geographic extent, by country or by the project title (for now, limited to EJPSOIL programme). When keywords are filled in, the catalogue allows to search for datasets containing data on e.g., selected soil properties.

For example, the catalogues allows answering questions like:

"As a national, regional or European governmental organization, I would like to know how many and which soil datasets are available containing soil parameters listed in the proposed Soil Monitoring and Resilience Directive, so I can use that as input for my soil monitoring strategy."

"As a funding agency, I would like to know on which important soil properties little data is available, so I can target my next call on understanding how these can be collected, interpreted and re-used better."

"As a researcher, I would like to find all available datasets in my region that contain data on soil organic carbon, so I can use these as input in my models and predict soil carbon sequestration potential for my region."

4. What is in it now

At present, the metadata catalogue contains metadata from the following sources:

- The EJP SOIL surveys on available soil and soil management datasets in EJP SOIL countries that was conducted in 2020 and published in D6.1 and D2.2.
- The MensMeu 1.0 survey conducted by JRC and DG-ENV
- An update on the GSP CountrySIS survey (2020) in EJP SOIL /European countries
- EJP SOIL produced data in WPs or in internal research projects
- Metadata of datasets harvested from ESDAC

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

All EJP SOIL partners have been regularly asked to check, update and where needed provide metadata of new datasets in the catalogue (last request was done in January 2024). Not all planned EJP SOIL results are available yet; new datasets will be added to the catalogue in the course of 2024. The final version in D6.8 will therefore include more data (due in November 2024).

So far, metadata of 519 datasets is included in the catalogue. In Table 1, metrics of datasets per country and in Table 2, the number of datasets per soil property, are provided. In Figure 1 and Figure 2, other characteristics of the included metadata are provided.

Table 1. Content of metadata catalogue at 26 January 2024, categorised per country.

Country	count	Country	count	Country	count
Austria	16	Hungary	39	Portugal	5
Belgium	34	Ireland	4	Romania	2
Bulgaria	5	Italy	121	Slovakia	9
Czech Republic	18	Latvia	16	Slovenia	24
Denmark	8	Lithuania	7	Spain	13
Estonia	7	Luxembourg	4	Sweden	29
Finland	6	Malta	1	Switzerland	1
France	29	Netherlands	9	Turkey	13
Germany	22	Norway	14	United Kingdom	67
Greece	1	Poland	7	N/A	49

Table 2. Content of metadata catalogue at 26 January 2024, categorised per soil property.

Soil property	count	Soil property	count
acidification	4	salinization	4
available phosphorus	2	soil compaction	6
available potassium	1	soil contamination	17
available water capacity	6	soil depth	9
calcium carbonate	6	soil erosion	13
organic carbon stock	9	soil map	59
cation exchange capacity	3	soil moisture	6
decline in soil organic matter	2	soil organic carbon	10
decline soil biodiversity	2	texture	19
pH	10	total nitrogen	3

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

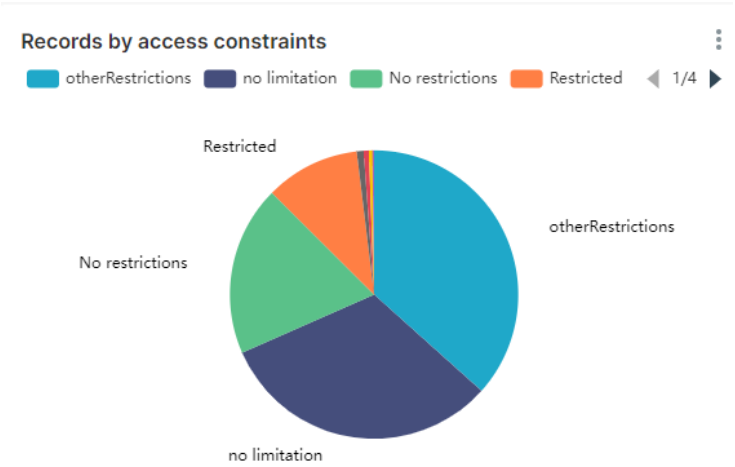

Figure 1. Overview of restrictions on use of datasets in the metadata catalogue.

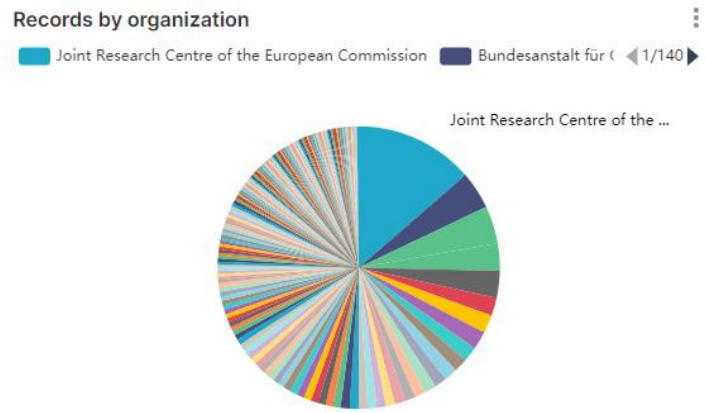

Figure 2. Overview of contributions per organisation of datasets in the metadata catalogue. Main contributor is the JRC, then the German government and many others.

5. How does it work and how to operate

Users can search the metadata catalogue by typing in keywords in the catalogue interface or search field (Figure 3). The interface is the OGC API – Records API, with HTML encoding. The application is based on pycsw (https://pycsw.org) which operates on a PostGres Database containing imported records from a Git repository (https://github.com/ejpsoil/ejpsoildatahub). In the search results, the catalogue allows easy filtering either by country, by EJP SOIL project or by type of resource (Figure 4). These categories can easily be changed in the future if needed, but they seem most adequate at present.

The Github repository is populated in three ways:

- Manual additions/mutations, for example using MDME (https://osgeo.github.io/mdme), YML
- Import from a spreadsheet (CSV)
- Import from an online (remote) location (harvest; CSW, WMS, OAI-PMH, DOI)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

The imported metadata is automatically standardised upon inclusion.

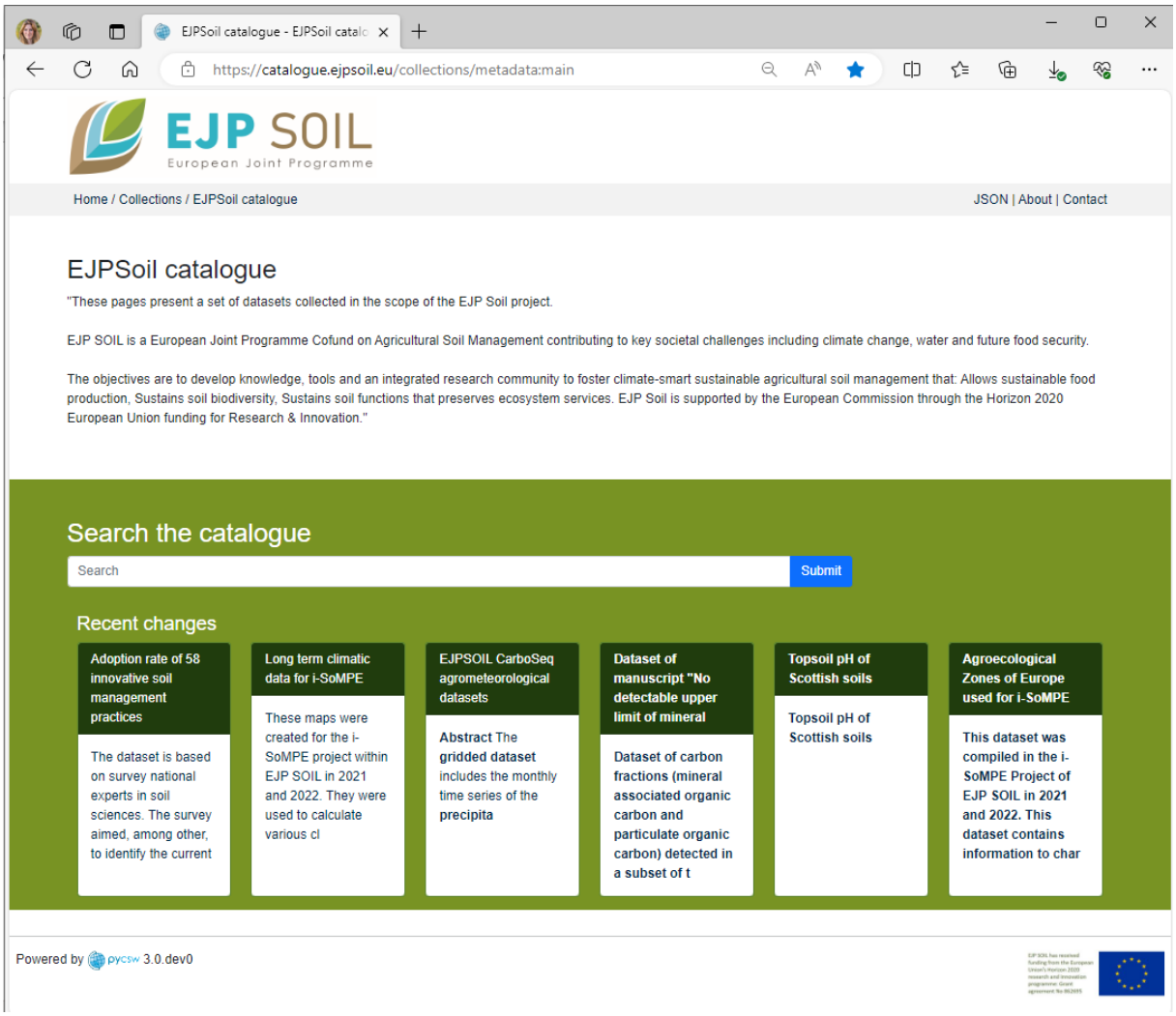

Figure 3. Search interface of the metadata catalogue.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

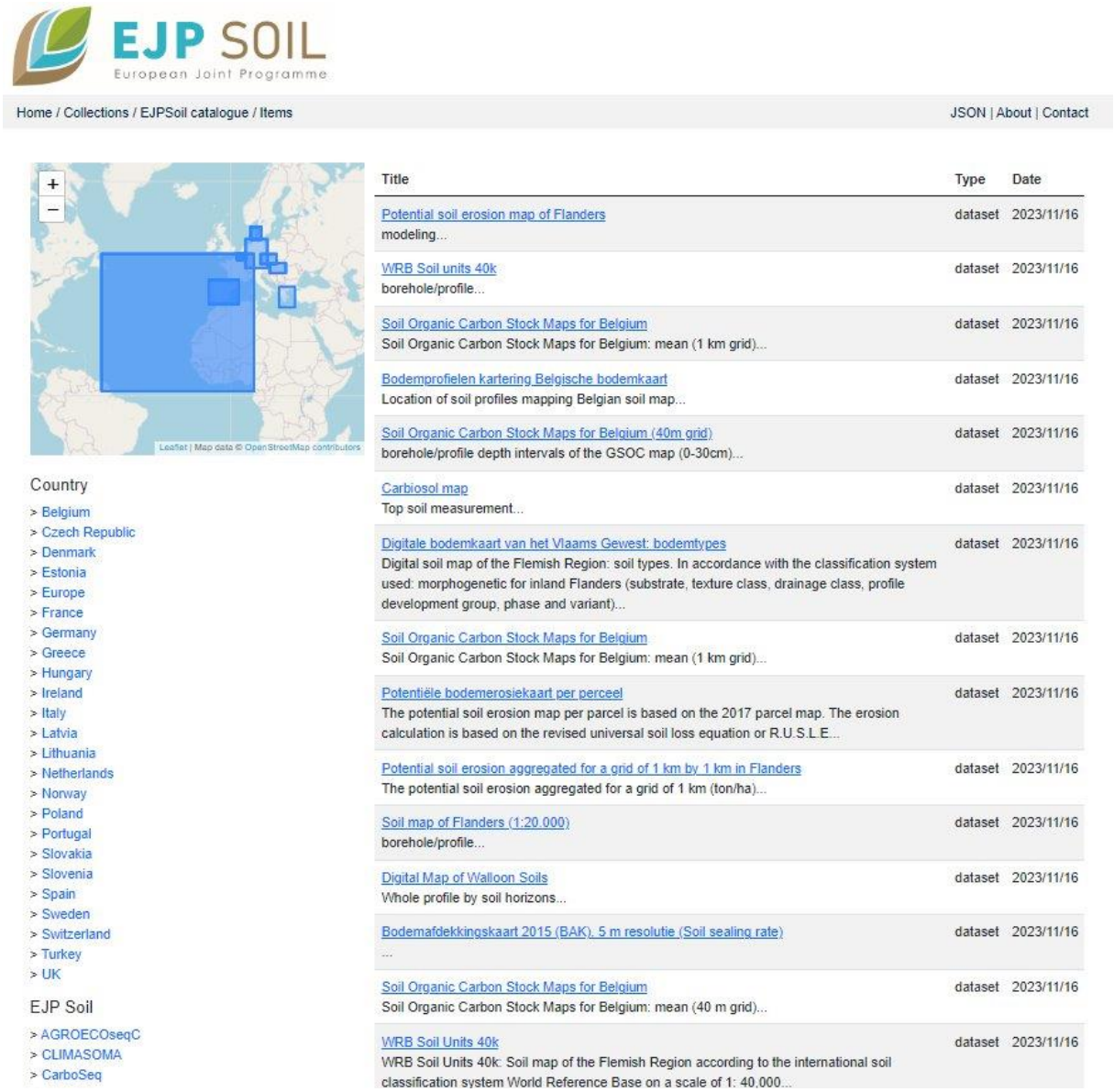

Figure 4. Returned search result and possibility to filter on country or EJP SOIL project.

The principles behind the development were (i) easy maintenance and (ii) preventing redundancy by importing metadata elements directly from the data sources or existing online sources. Open standards used are the ISO19115 metadata model, and Open Geospatial Consortium (meta)data exchange standards such as WMS, CSW, OGC API:Records. Interoperability with the academic community is facilitated by an OAI-PMH interface and the support of DOI/Datacite.

A GIT system as a backend, enables a full history of the contents. It also enables users to participate in the population of the platform. From the public interface, users can provide feedback on a record, or directly enter the GIT sources ('Edit me on Git') to ask questions or suggest additions to the community.

EJP SOIL encourages storage of new datasets in the H2020 eligible repositories like ZENODO, national and institutional repositories, based on their persistency (>25 years). Several Open Science webinars and a workshop at the EJP SOIL Annual Science Days (June 2023) have been held by EJP SOIL to this purpose. EJP SOIL consortium members are mandated to provide open access to their datasets through

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

a CC BY license (where possible), ensure a certain level of metadata, and to provide a DOI or other persistent identifier. The EJP SOIL metadata catalogue harvests relevant datasets from these repositories using the DOI's or relevant keywords (e.g. EJPSOIL).

The researchers are encouraged to prepare metadata already before or while developing a new dataset, by using EJP SOIL metadata checklist. Placing a metadata file (for example MCF) along every data file (known as the sidecar approach, https://en.wikipedia.org/wiki/Sidecar_file) is an easy way to compose and safeguard the metadata during the data life cycle. Only when the metadata has been largely filled in, researchers proceed with collecting or deriving the dataset itself. This ensures well considered data collection and full metadata description.

6. Architecture

Figure 5 shows an overview of the architecture of the system. At the top, three approaches of metadata ingestion are displayed. Imported metadata is first standardized before inclusion. At intervals, scripts can run on the metadata repository, identifying broken links for remote sources or alternate quality indicators.

When changes are detected, the records are converted to iso:19139:2007 (xml serialisation of iso:19115) and pushed to a PostGres Database which is queried by the pycsw catalogue application. Optionally, different frontends can be connected to the database for example a dashboard providing a quick overview of the content of the catalogue. An example is available here: https://dashboards.isric.org/superset/dashboard/29.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

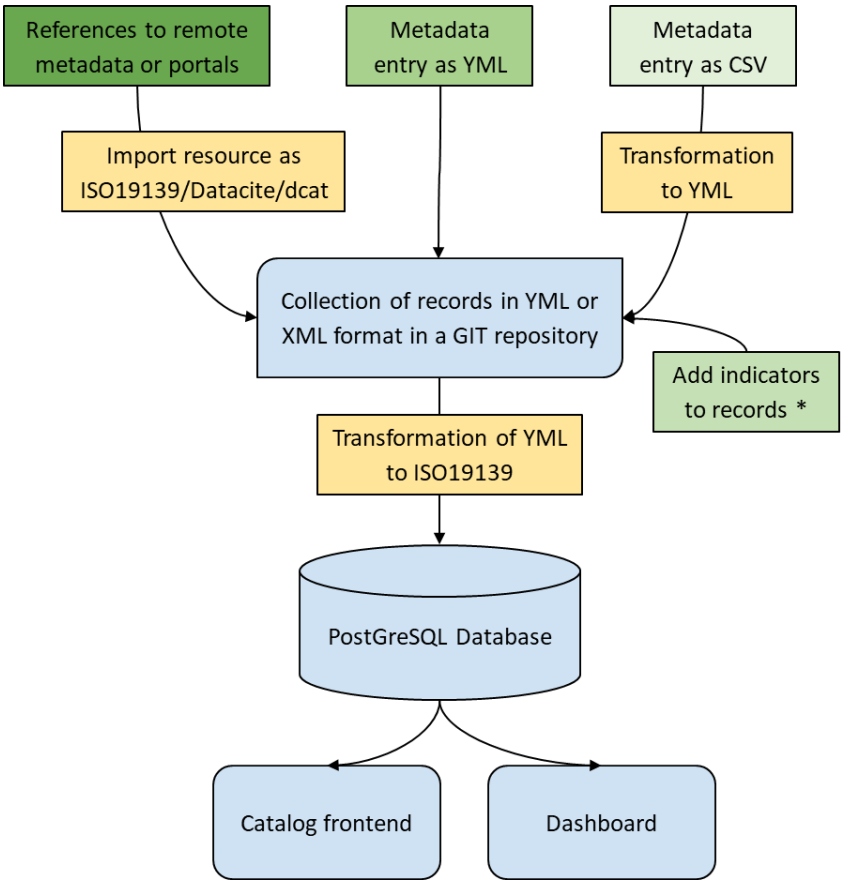

Figure 5. Architecture of the Catalogue.

Guidance on how to store data in ZENODO, annotate metadata and upload metadata to the catalogue is provided in the soil data assimilation wiki developed between EJP SOIL T5.5 and WP6: https://ejpsoil.github.io/soildata-assimilation-guidance. It provides technical guidance on all aspects of data assimilation, incl. annotation and upload.

All software modules are open source (MIT). Where available, modules are used with a high TRL and an active user community. The csv template has been developed by EJP SOIL in collaboration with the EUSO Working Group on Data Integration, the Hub SubGroup, and are available at https://ejpsoil.github.io/soildata-assimilation-guidance/metadata.html#metadata-in-ejp-soil. The MCF format (yaml template) is maintained by the geopython group at https://geopython.github.io/pygeometa/reference/mcf. A tool for metadata extraction is under development as part of this initiative at https://pypi.org/project/geodatacrawler. Components have been selected based on ease of maintenance and alignment with modern IT practices, such as scalability, containerisation and microservices. Aspects of the architecture as described in this document are also used in other research projects, such as LSC-hubs[1], Soils4Africa[2] and Space2Place.

[1] https://lsc-hubs.org/
[2] https://www.soils4africa-h2020.eu/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

7. Going forward

The current content of the metadata catalogue will be enriched with the metadata of all EJP SOIL produced datasets until the end of the project in January 2025 (D6.8). The GIT repository containing the metadata records is persistent and open, and will continue to exist after EJP SOIL. The catalogue can be used by other initiatives, projects and EUSO/JRC to upload metadata or use it as a source. The approach of the EJP SOIL metadata catalogue and the catalogue itself (its content, possibly including the processes and software), will be taken over by the SoilWise[3] EU Soil Mission project that aims to support building the EU Mission Soil's knowledge repository and advancing the European Soil Observatory. SoilWise will work on enrichment with metadata on non-geographical data and knowledge sources. Because EJP SOIL and SoilWise are very well aligned, the lessons learned in EJP SOIL and the legacy of EJP SOIL's metadata catalogue will not be lost.

During the last year of EJP SOIL, the additional harvesting options from major repositories (ZENODO, Cordis, INSPIRE GeoPortal, ESDAC, etc.) will be explored and small improvements to the catalogue interface will be made. Depending on the time and budget available, an ambition for the future is to enrich the metadata records with derived quality indicators that will incentivise improvement of metadata and facilitate better data harmonisation. This way, European soil datasets will be better reused in the future.

The EJP SOIL metadata catalogue has the potential to serve the wider European soil community (researchers, policy makers and practitioners) if datasets generated by initiatives/projects from the EU Mission Soil are included. This is not part of the EJP SOIL work programme, but the discussion between EJP SOIL, SoilWise and the EU Mission Soil (including Soil Mission Board) is welcome to identify the way forward. EU Mission Soil projects are recommended to:

- use the EJP SOIL metadata catalogue and improve its quality and searchability
- enrich the catalogue with new datasets,
- use EJP SOIL requirements (metadata template) for metadata annotation and persistent storage by including them in the project's Grant Agreement

Compliance checks by REA and collaboration with EUSO/JRC and SoilWise will be important incentives that can accelerate the development of the catalogue and thereby the usefulness of if to the EU Mission Soil.

Further inclusion and labelling of authoritative datasets in the catalogue will also be beneficial to DG ENV and DG AGRI as this will allow an overview of datasets that are or can be used for reporting obligations and monitoring in the framework of the proposed EU Soil Monitoring and Resilience Directive.

When the catalogue grows and is used for different applications, the governance, selection criteria and obligations for inclusion of metadata in the catalogue should be defined in a more future-proof way since at the moment this is performed by EJP SOIL as a project with an end date. We recommend to set up a dedicated group of editors (representatives of relevant stakeholders) that responds to questions and review requests.

[3] https://soilwise-he.eu/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695

8. Links to other deliverables

Enriched metadata supports the activity of transforming legacy vector data to a standardized data model. It supports for example the INSPIRE Good Practice of storing soil vector data in an INSPIRE Soil compliant GeoPackage format, as described in D6.4 (due in January 2024), which presents an easy-to-use approach for structuring soil datasets in a standardised way. If many soil datasets, of which the metadata is stored in the metadata catalogue, use this standardised data model and format it will be (much) easier to combine and re-use datasets that are returned in a search result.

Together, these two deliverables therefore can be a big step towards making soil data in Europe more FAIR (Findable, Accessible, Interoperable, Re-usable). D6.4 promotes interoperability and re-usability. D6.6 and D6.8 will contribute to findability and accessibility, and therefore also re-usability. The current content of the metadata catalogue will be enriched with the metadata of all EJP SOIL produced datasets until the end of the project in January 2024 (D6.8).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 862695