www.soilprom.eu

Organisation VITO
Main Authors Frie Van Bauwel & Maarten Van Loo

SOILPROM

Blueprint for harmonized databases

Date: 21/04/2025
Doc. Version: final

This project has received funding from the European Union's Horizon Europe Research and Innovation program under Grant Agreement No. 101156589. This output reflects only the author's view, and the European Union cannot be held responsible for any use that may be made of the information contained therein.

SOILPROM

Document Control Information

Settings	Value
Deliverable Title	Blueprint for harmonized databases
Work Package Title	WP1 – Setting the scene for soil pollution processes modelling
Deliverable number	D1.3
Description	Blueprint to guide databases owner to harmonise existing databases or develop new databases to make them compliant with EUSO and DestinE.
Lead Beneficiary	VITO
Lead Authors	Frie Van Bauwel, Maarten Van Loo
Contributors	Michiel Maertens, Ingeborg Joris
Submitted by	Ingeborg Joris
Doc. Version (Revision number)	Final
Sensitivity (Security):	Public
Date:	21/04/2025

Document Approver(s) and Reviewer(s):

NOTE: All Approvers are required. Records of each approver must be maintained. All Reviewers in the list are considered required unless explicitly listed as Optional.

Name	Role	Action	Date
Giannis Seremetis	IT	*<Approve>*	18/04/2025

Document history:

The Document Author is authorized to make the following types of changes to the document without requiring that the document be re-approved:

- Editorial, formatting, and spelling
- Clarification

To request a change to this document, contact the Document Author or Owner.
order (latest version first).

Revision	Date	Created by	Short Description of Changes

Configuration Management: Document Location

The latest version of this controlled document is stored in <location>.

Nature of the deliverable		
R	Report	X
DEC	Websites, patents, filing, etc.	
DEM	Demonstrator	
O	Other	

Dissemination level		
PU	Public	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

ACKNOWLEDGEMENT

This report is part of the deliverables from the project "SOILPROM" which has received funding from the European Union's Horizon Europe Research and Innovation program under grant agreement No 101156589

More information on the project can be found at: http://www.soilprom.eu.

TABLE OF CONTENTS

CONTENTS

EXECUTIVE SUMMARY

Purpose:
This deliverable under Task 1.3 of the SOILPROM project aims to streamline the integration and use of datasets necessary to support EU-wide soil pollution modeling. The focus is on aligning data requirements with those of the European Union Soil Observatory (EUSO) and Destination Earth (DestinE), facilitating model development, reproducibility, and policy support.

Key Objectives:

1. Outline data requirements from EUSO and DestinE.
2. Identify datasets needed for running SOILPROM's local models and explore potential EU-wide alternatives.
3. Evaluate the compliance of selected datasets with EUSO and DestinE standards.

Findings:

- **EUSO & DestinE Requirements:**
  - EUSO's requirements are still evolving, especially regarding data sharing under the upcoming Soil Monitoring Law.
  - DestinE has a structured data architecture centered around the STAC-compliant Data Lake, emphasizing harmonized, open access via API, and FAIR (Findable, Accessible, Interoperable, Reusable) principles.
- **SOILPROM Data Needs:**
  - Use-case leaders identified essential datasets required to run their models after which EU-wide dataset availability was evaluated for each dataset.
- **Compliance Evaluation:**
  - Most meteorological and surface datasets meet the required criteria for metadata availability, machine-accessibility (via API), and open sharing.
  - Several soil data parameters—notably nitrogen content—are missing from available EU-wide datasets.
  - Many available soil datasets lack direct accessibility or cannot be freely shared, limiting full integration into the modeling tool.

Recommendations:

- Monitor the development of EUSO's data framework to ensure future compliance.
- Prioritize datasets that meet FAIR principles but also strike a balance with scientific accuracy.
- Unavailable parameters could be derived from proxies that are available (e.g., soil description from texture/type data, P-AL from P-Olsen) where feasible.

- Highlight data access paths in the modeling tool even when automation isn't possible to improve user experience.

Conclusion:
The groundwork has been laid for harmonizing data inputs into SOILPROM's modeling platform, with most meteorological and surface data in place. However, challenges remain - particularly for soil data - which must be addressed through alternative data derivation, continued dataset exploration, and proactive alignment with evolving EU data standards.

1. INTRODUCTION

CONTEXT

Task 1.3 aims to provide guidelines to integrate and use existing databases or develop new databases that can effectively feed DestinE and EUSO. SOILPROM will develop and deploy multiple models that will provide the EU with insights on soil pollution. Since these models incorporate data from different sources, the onboarding process for soil researchers demands significant effort. This deliverable will streamline the process for all project partners involved in the project.

DELIVERABLE OBJECTIVES

This deliverable has 3 objectives:

1. Give an overview of the data requirements shared by EUSO & DestinE.
2. List all the data needed to run the local models, and options for EU-wide datasets that could replace these local datasets.
3. Cross-check of listed EU-wide datasets with the requirements found in (1).

2. Blueprint for Harmonizing Datasets

I. OBJ1: EUSO & DestinE requirements

Below the data requirements from EUSO and DestinE. The requirements have been shared and discussed with the WP4-leader in order to develop technical specifications to be considered in the development of the Modelling Platform and Decision Support Tool. Further below in the deliverable (OBJ3: Compliancy of datasets to EUSO & DestinE requirements), the compliancy of the collected SOILPROM datasets (OBJ2: SOILPROM models & data needs) will be checked with these requirements. As will become clear in the following paragraphs, the EUSO requirements are still in development, and the DestinE requirements are focussed on remote sensing data. To follow the requirements would hence be very difficult. We thus chose to focus on the general concept, rather than the exact technical specifications when checking compliancy with EUSO and DestinE.

EUSO requirements

The EU Soil Observatory (EUSO) is a platform that provides the Commission Services and the broader soil user community with the knowledge and data flows needed to safeguard and restore soils, including monitoring soil health and supporting EU research and innovation on soils. Bilateral communication was set up with EUSO. Currently they are still developing their own data sharing requirements, in the context for the Soil Monitoring Law and the EU Mission Soil The evolution will be followed closely to ensure compliance from the side of SOILPROM. The timeline of EUSO however doesn't match with the deadlines set for Task 1.3. We were therefore not able to include this into the deliverable. EUSO did share the already existing metadata and dataflows for the LUCAS topsoil dataset.

DestinE requirements

Destination Earth is a flagship initiative of the European Commission to develop a digital model of the Earth (digital twin) to model & monitor natural phenomena, hazards and the related human activities. It aims to assist users in designing accurate and actionable adaptation strategies and mitigation measures.

DestinE centralizes all data in their DestinE Data Lake. The data lake is one part of their 3 main products, the other 2 being the "DestinE Core Service Platform", and the "DestinE Digital Twin Engine". DestinE uses the following definition for their Data Lake:

> THE **DESTINE DATA LAKE** (DEDL) PROVIDES DISCOVERY, ACCESS, AND BIG DATA PROCESSING SERVICES. THE DESTINE DATA PORTFOLIO EVOLVES ACCORDING TO USER NEEDS AND FOLLOWS A DEFINED AND EVOLVING PROCESS FOR ENHANCEMENTS, WHICH TAKES INTO ACCOUNT REQUIRED DATA STORAGE NEEDS. IT PROVIDES A SEAMLESS ACCESS TO DATASETS VIA GUIS OR APIS TO DATA, IN ACCORDANCE TO THE DESTINE DATA PORTFOLIO, REGARDLESS OF DATA TYPE AND LOCATION.
>
> THE DEDL BIG DATA PROCESSING ALLOWS NEAR-DATA PROCESSING AND BY THIS CONCEPTUALLY SUPPORTS ML/AI APPLICATIONS EXECUTED ON THE DEDL. THE DESTINE DATA LAKE FEDERATES WITH EXISTING DATA HOLDINGS AS WELL AS WITH COMPLEMENTARY DATA FROM DIVERSE SOURCES LIKE IN-SITU, SOCIO-ECONOMIC, OR DATA-SPACE DATA.

In the following sections, the key elements essential for complying with their data lake requirements are oulined.

Harmonised Data Access

DestinE developed a REST[1]-based API that enables users to search, discover, and access data from the DestinE Data portfolio. This was done using a STAC compliant interface: this means that users can access datasets and services no matter how they're stored or accessed. This keeps the code for working with the data flexible and not tied to a specific source. More info on STAC can be found here.

Some tools that are able to communicate with the STAC API endpoint: PyStac, Postman, Swagger UI, cURL, etc… DestinE hence offers a vast number of ways to access the data. Documentation on how to use the API is provided here (open API documentation). DestinE also provides example Jupyter Notebooks that show how python code can be used to work with the API.

DestinE Data Portfolio

Destination Earth Discovery Service allows users to search for data among different collections. They can be found at https://hda.data.destination-earth.eu/ui/catalog.

[1] A REST API (RESTful API) is an application programming interface (API) that follows the principles of REST (Representational State Transfer) for communication between a client and a server, using standard HTTP methods to access and manipulate data.

Destination Earth - Data Lake - Data Portfolio

Digital Twins Collections are highlighted in green

	Dataset ID	Dataset Title	Keywords	Provider Role (Producer)	Provider Role (Processor)	Provider Role (Licensor)	Provider Role (Host)
1	EO.AERIS.DAT.IAGOS	In-service Aircraft for a Global Observing System	AERIS, AIRCRAFT, ATMOSPHERIC, IAGOS, L2	**IAGOS CORE** https://iagos.aeris-data.fr/ **FZ Jülich** https://www.fz-juelich.de/port...	**dedl** https://destination-earth.eu/	**IAGOS DC** https://iagos.aeris-data.fr/	**dedl** https://hda.data.destination-e... (EO.AERIS.DAT.IAGOS)
2	EO.CLMS.DAT.CORINE	CORINE Land Cover	CHA20012-2018, CLC, Copernicus, Corine, EEA39, European, Land Cover Change, Land cover, Satellite Image Interpretation, geospatial data, land cover, land use, landscape, landscape alteration	**European Environment Agency** https://www.eea.europa.eu/	**European Environment Agency** https://www.eea.europa.eu/	**European Environment Agency** https://www.eea.europa.eu/	**wekeo** https://www.wekeo.eu/ (EO:CLMS:DAT:CORINE)
3	EO.CLMS.DAT.GLO.DMP300_V1	Dry Matter Productivity 2014-present (raster 300 m), global, 10-daily - version 1	10-daily composite, Dekad, GLOBE, Orthoimagery, agricultural production, crops, dry matter, primary productivity	**VITO NV** https://vito.be/en	**VITO NV** https://vito.be/en	**European Commission Directorat...** https://www.copernicus.eu	**wekeo** https://www.wekeo.eu/ (EO:CLMS:DAT:CLMS_GLOBAL_DMP_300M_V1_10DAILY_NETCDF)
4	EO.CLMS.DAT.GLO.FAPAR300_V1	Fraction of Absorbed Photosynthetically Active Radiation 2014-present (raster 300 m), global, 10-daily – version 1	10-daily composite, Dekad, GLOBE, Orthoimagery, biogeophysical, fapar, geophysical environment	**VITO NV** https://vito.be/en	**VITO NV** https://vito.be/en	**European Commission Directorat...** https://www.copernicus.eu	**wekeo** https://www.wekeo.eu/ (EO:CLMS:DAT:CLMS_GLOBAL_FAPAR_300M_V1_10DAILY_NETCDF)
5	EO.CLMS.DAT.GLO.FCOVER300_V1	Fraction of Green Vegetation Cover 2014-present (raster 300 m), global, 10-daily – version 1	10-daily composite, Dekad, GLOBE, Orthoimagery, biogeophysical, geophysical environment, leaf area	**VITO NV** https://vito.be/en	**VITO NV** https://vito.be/en	**European Commission Directorat...** https://www.copernicus.eu	**wekeo** https://www.wekeo.eu/ (EO:CLMS:DAT:CLMS_GLOBAL_FCOVER_300M_V1_10DAILY_NETCDF)
6	EO.CLMS.DAT.GLO.GDMP300_V1	Gross Dry Matter Productivity 2014-present (raster 300 m), global, 10-daily – version 1	10-daily composite, Dekad, GLOBE, Orthoimagery, agricultural production, crops, gross dry matter, primary productivity	**VITO NV** https://vito.be/en	**VITO NV** https://vito.be/en	**European Commission Directorat...** https://www.copernicus.eu	**wekeo** https://www.wekeo.eu/ (EO:CLMS:DAT:CLMS_GLOBAL_GDMP_300M_V1_10DAILY_NETCDF)

Figure 1: Screenshot of the DestineE Data portfolio.

Technical data requirements

A step by step manual is provided on how to onboard own data onto the data lake. The main criteria is to prepare the data as a STAC collection, with its associated STAC items. Secondly, once reviewed and validated by a review board, the dataset needs to be accessible through the STAC API, in the same way as the rest of the DestinE Data Lake portfolio. Finaly, it needs to be discoverable through the DestinE Data Porfolio - Web Portal.

A **STAC Collection** provides information about a spatio-temporal collection of data and enables descriptions of things like the spatial and temporal extent of the data, the license, keywords, providers, etc.

A **STAC item** is a GeoJSON feature (Figure 2) and includes additional fields for:

- the time the asset represents
- a thumbnail for quick browsing
- asset links, links to the described data
- relationship links, allowing users to traverse other related STAC Items

A STAC Item can also contain more additional fields and JSON structures to enable data providers for any other purpose necessary.

```
{
    "stac_version": "1.0.0",
    "type": "Feature",
    "id": "20201211_223832_CS2",
    "bbox": [],
    "geometry": {},
    "properties": {},
    "collection": "simple-collection",
    "links": [],
    "assets": {}
}
```


Figure 2: Example GeoJSON feature of a STAC item.

DestinE Data Policy for DestinE Digital Twin Outputs

DestinE Digital Twin Outputs, created from freely available data, are shared under the free and open DestinE Data Policy (Creative CC-BY-4.0 licence, unless specified otherwise). Users worldwide can freely use, share, modify, and combine these outputs with other data. However, there are some conditions:

- Pre-set formats, timing, and distribution rules must be followed.
- Original data policies and licenses must be respected.
- Security rules and system protection measures must be upheld.
- Reliable access for users must be ensured.

These outputs come with no guarantees about data quality, accuracy, or suitability. More details can be found here.

II. OBJ2: SOILPROM MODELS & DATA NEEDS

Within this objective there is a clear focus on datasets that can be used in the modelling platform that will be produced within this project. Hence, only datasets spanning at least a big part of the EU are included in this section. Moreover, only parameters needed to run a trained model are considered, this means data specifically needed for model calibration/validation are not included. These data are typically in-situ measurements or analysis results from local samples. Parts of these data are already available to the use-cases, but a big part will be gathered during the project. This will thus be covered by Task 2.1 in WP2.

All 9 use-case leaders were asked to fill in a template table with the goal to list all data needs to run the models. The columns included in this template are listed in Table 1. Of these columns, the first four columns were mandatory to fill in, while the others were optional and could be filled in if the use-case leaders had knowledge of existing European or global datasets. The optional fields queried for relevant aspects of the datasets regarding harmonized data. The table derived from the filled in template table, can be found in Annex I. It was decided not to list the input

data of the TSMP model (FZJ use-case) as the output data of this model will only be used.

Table 1: Included fields in template table filled in by the use-case leaders. Bold column-names were mandatory to fill in.

Column	Explanation
Data requirement	Required input parameter.
Theme	General data theme, e.g. meteo, soil, water
Model	For which model will the data be used?
Mandatory	Is the data mandatory in the model?
Use-case	Which use-case needs this data?
Data source	Which instance ensures data availability?
URL	An URL where you can find the data or metadata on the dataset.
Access	How can the data be accessed? e.g. via API
Open access	Is the data freely accessible?
Spatial resolution	The resolution of the map. If the database is not a map, this field is not relevant and filled in with NA.
Temporal resolution	If the database includes timeseries, the measurement frequency. If it does not comprise timeseries, filled in with NA.
Depths	Measurement depths included in the data. If not relevant, filled in with NA.
Coverage	Geographical coverage database, e.g. global, EU, Belgium
Temporal coverage	Timespan included in the data. Relevant for data including timeseries, otherwise filled in with NA. e.g. January 1990 – June 2024.
Period source data	Time period from which data over multiple years are compiled to get one dataset (typically a map). Filled in with NA for datasets including timeseries. e.g. January 1990 – June 2024, or 2001.
License	The license under which the dataset is distributed. e.g. Open Data Commons Attribution (ODC-BY), or Attribution 4.0 International (CC BY 4.0). Filled in with NA when not stated
Can it be shared with 3rd parties?	Is it allowed to share the data without any restrictions with third parties?

Once the table was completed for each use-case, the reported European and global datasets were listed and for these, the information requested in the optional fields was completed. Furthermore, additional European and global datasets were sourced for the required parameters that initially lacked them, as well as for other parameters, to ensure that a range of options was available where possible.

An overview of the listed data needs can be found in the "data needs"-tab of Annex I and a list of the identified relevant European or global datasets can be found in "available data"- tab of Annex I.

III. OBJ3: COMPLIANCY OF DATASETS TO EUSO & DESTINE REQUIREMENTS

As mentioned in paragraph II.I (OBJ1: EUSO & DestinE requirements), we chose to focus on general concepts instead of the technical specifications in comparing the datasets to the EUSO and DestinE requirements. These general concepts essentially come down to the principles of FAIR data.

The main principles of FAIR data are that data are Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016). In broad terms, this means data need to satisfy following criteria: 1) Data can easily be accessed through a protocol like an API in a standard way; 2) The parameters are named following the conventions of an existing standard vocabulary; 3) Values are reported in SI units; 4) Terms and conditions on using and sharing the data are clearly specified, preferably by a license under which the data are distributed; and 5) Metadata are complete, clear and easily findable.

In practice, we are mainly concerned about the ability to use the data in the modelling platform. Hence slightly less strict criteria are considered here: 1) The metadata is easily accessible and is comprehensive enough to know exactly what the data entail; 2) The data can be easily accessed; and 3) The data can be shared with third parties. In Table 2, we map these "SOILPROM" requirements to existing DestinE or EUSO requirements.

Table 2: Mapping of SOILPROM requirements with existing DestinE/EUSO requirements.

ID	SOILPROM criteria	DestinE/EUSO requirement
R1	The metadata is easily accessible and is comprehensive enough to know exactly what the data entail	Combination of DestinE STAC Item & collection description
R2	The data can be easily accessed	DestinE Harmonized Data Access
R3	The data can be shared with third parties	DestinE Data policy

Table 3 shows for each data need how well the available datasets satisfy these three requirements. Requirements 1 and 3 show a pass/fail score, while requirement 2 is scored from 1 to 3. Each number represents a class of accessibility; A score of 1 means the data are not directly accessible (e.g. by form/request), score of 2 is given to data that need to be downloaded from a website, and datasets with a score of 3 can be accessed through an API or other machine-to-machine protocol that can be automated. A score of 3 is ideal, while a 2 is a sufficient option. A dataset with a score of 1 is to be avoided.

For the required meteorological datasets, at least one dataset was found that fulfils all three requirements for each separate data need. The same applies to the

required surface data. Only for the elevation data, manually downloading the data is required. Some data needs regarding soil-information are however problematic.

For soil-related data, there are several data needs for which no data are available on an EU-wide scale (median soil texture size, soil cohesion, soil description, Ca, Mg and N soil content, P-AL, and residual soil water content). For some of these parameters, this is not a major problem: Mg and Ca content, and residual soil water content are not strictly needed for running the models. Soil description and median soil texture size could not be found directly but might be derived from available data sets on soil texture and soil type, possibilities to do so need to be further investigated. Similarly, P-AL can potentially be derived from the available P-Olsen[2]. Thus, there are only 2 parameters for which an EU-wide data set is absent: soil cohesion and N content in the soil.

Further, for several available parameters (heavy metal concentrations, pesticide concentrations in topsoil, organic matter soil content and P-Olsen of soils), there are no datasets found that are easily accessible. Often, these cannot be shared with third parties as well. The only option for the model tool for these parameters is pointing out the link to the data description. While less ideal, this approach will still facilitate easier data-access for users of the modelling tool. For pesticide concentrations, the only available data source containing only point data and not a map poses an additional challenge.

Finally, an important note to make on the datasets to be selected to take up in the model tool: When multiple datasets are available for one parameter, the most optimal dataset from data harmonization point of view, is not necessarily the most optimal dataset from a scientific perspective.

Table 3: Overview of the available EU-wide data sources with their rating according to all three specified requirements. R1, R2, and R3 refer to the requirements as listed in Table 2. R2 has 3 categories: 1) not accessible directly (form/request); 2) Download website, 3) API/ machine-way to download data. Data needs that can be derived from other data needs (like daily min/max air temperature) are not explicitly mentioned. Where possible without confusion, parameters are put together for getting an as concise as possible overview.

Data need	Data set	R1	R2	R3
Meteo				
Atmospheric temperature	ERA5 – Copernicus: hourly data on single levels from 1940 to present	✓	3	✓
Potential evapotranspiration	ERA5 – Copernicus: hourly data on single levels from 1940 to present	✓	3	✓
	MAPPE model	✓	2	✓
	CRU TS	✓	2	✓

[2] "P-AL" refers to phosphorus extracted using an ammonium lactate solution, while "Olsen P" refers to phosphorus extracted using a sodium bicarbonate solution, both methods used to assess plant-available phosphorus in soil.

	GLEAM	✓	1	?
Precipitation	ERA5 – Copernicus: hourly data on single levels from 1940 to present	✓	3	✓
	ECAD – E-OBS gridded dataset	✓	2	✓
	EURADCLIM: The European climatological gauge-adjusted radar precipitation dataset (1-h accumulations)	✓	3	✓
	MSWEP: Multi-Source Weighted-Ensemble Precipitation	✓	1	✓
Surface pressure	ERA5 – Copernicus: hourly data on single levels from 1940 to present	✓	3	✓
Solar radiation	ERA5 – Copernicus: hourly data on single levels from 1940 to present	✓	3	✓
	ECAD – E-OBS gridded dataset	✓	2	✓
Incident radiation	ECAD – E-OBS gridded dataset	✓	2	✓
Relative humidity	ERA5 – Copernicus: hourly data on single levels from 1940 to present	✓	3	✓
	ECAD – E-OBS gridded dataset	✓	2	✓
Specific humidity (of ice, liquid, rain, snow, and vapor)	ERA5 – Copernicus: hourly data on pressure levels from 1940 to present	✓	3	✓
Wind speed	ERA5 – Copernicus: hourly data on single levels from 1940 to present	✓	3	✓
	ECAD – E-OBS gridded dataset	✓	2	✓
Wind direction **Wind speed in v-direction (meridional wind)** **Wind speed in u-direction (zonal wind)**	ERA5 – Copernicus: hourly data on single levels from 1940 to present	✓	3	✓
Soil				
Heavy metal concentrations in topsoil (As, Cd, Cr, Cu, Hg, Pb, Mn, Sb, Co and Ni)	ESDAC - Maps of heavy metals in the soils of the EU, based on LUCAS 2009 HM data	✓	1	✗
Arsenic in topsoil	ESDAC - Arsenic in European topsoils	✓	1	✗
Cadmium in topsoil	ESDAC - Cadmium in topsoils of the European Union	✓	1	✗
Copper in topsoil	ESDAC - Copper distribution in topsoils in the European Union	✓	1	✗
Mercury in topsoil	ESDAC - Mercury content in the European Union topsoil	✓	1	✗
Zinc in topsoil	ESDAC - Zinc concentrations in EU topsoils	✓	1	✗
(Sorbed) pesticide concentration in topsoil	ESDAC - Pesticides residues in European agricultural soils (point data)	✓	1	✗
Bulk density	ISRIC - WISE derived soil properties on a 30 by 30 arc-seconds global grid	✓	2	✓
	ISRIC - Harmonized World Soil Database (HWSD) version 2.0	✓	2	✓
	ESDAC - Soil Bulk Density in Europe	✓	1	✗
	ESDAC - Topsoil physical properties for Europe (based on LUCAS topsoil data)	✓	1	✗

Median soil texture size	/	/	/	/
Particle size distribution (sand, silt and clay)	ISRIC - WISE derived soil properties on a 30 by 30 arc-seconds global grid	✓	2	✓
	ISRIC - Harmonized World Soil Database (HWSD) version 2.0	✓	2	✓
	ESDAC - Topsoil physical properties for Europe (based on LUCAS topsoil data)	✓	1	✗
Soil cohesion	/	/	/	/
Soil description	/	/	/	/
Soil texture	ISRIC - Harmonized World Soil Database (HWSD) version 2.0	✓	2	✓
	ESDAC - Topsoil physical properties for Europe (based on LUCAS topsoil data)	✓	1	✗
Soil type	ERA5 – Copernicus: hourly data on pressure levels from 1940 to present	✓	3	✓
Ca content **Mg content** **N content**	/	/	/	/
CaCO3 content **pH**	ISRIC - WISE derived soil properties on a 30 by 30 arc-seconds global grid	✓	2	✓
Organic matter content	ESDAC - Soil Organic Matter (SOM) fractions	✓	1	✗
P-AL	/	/	/	/
P-Olsen	ESDAC - Phosphorus budget and P stocks	✓	1	✗
Depth of soil profile **Depth to groundwater**	ESDAC - European Soil Database v2 Raster Library 1kmx1km	✓	2	✓
Soil horizons	ISRIC - SoilGrids — global gridded soil information	✓	3	✓
Saturated hydraulic conductivity	ESDAC - 3D Soil Hydraulic Database of Europe at 1 km and 250 m resolution	✓	1	✓
Residual soil water content	/	/	/	/
Saturated soil water content **Water content at field capacity**	ESDAC - 3D Soil Hydraulic Database of Europe at 1 km and 250 m resolution	✓	1	✓
Water content at wilting point	ISRIC - SoilGrids — global gridded soil information	✓	3	✓
Soil moisture content **Soil temperature**	ERA5 – Copernicus: hourly data on single levels from 1940 to present	✓	3	✓
Surface				
Elevation data	GDEM V3 - ASTER Global Digital Elevation Map Announcement	✓	2	✓
Geopotential	ERA5 – Copernicus: hourly data on single levels from 1940 to present	✓	3	✓
Land use/Land cover	Copernicus Land Monitoring Service - CORINE Land Cover	✓	3	✓
Vertical velocity	ERA5 – Copernicus: hourly data on pressure levels from 1940 to present	✓	3	✓

3. Conclusions and Recommendations

I. Data gaps

For most required data, at least one EU-wide dataset can be accessed and shared, however, limited data gaps are present, mainly for soil data. The most problematic of these data gaps is the absence of nitrogen content in the soil and soil cohesion. Another point of attention is that some of the soil related datasets are only available through request and cannot be shared with third parties. The only possible way to share these in the modelling platform, is thus by informing the user of the existing link with the data description.

II. Roadmap to harmonization

In an ideal situation, following steps are proposed to harmonize all data:

1) Put all needed data in a uniform database.
2) Check if all data sets use the same vocabulary and units where relevant, and if not rename to existing standards.
3) Pad the database with rich metadata
4) Make data in database available through a standard API, similar to the STAC API developed in DestinE.
5) Make the data in the API available under a clear license.

However, a complete data harmonization within this project is not feasible thus a more pragmatic approach is proposed here.

III. Pragmatic approach for SOILPROM project

We propose following 4-step approach to harmonize the data:

Model selection

It was decided that not all models will end up in the modeling platform. Therefore, the model selection will lead already to a reduction of the datasets that need to be harmonized. The selection process of the models will also take into account the avalability of the data, based on the SOILPROM requirements

Use existing API endpoints

Some EU-wide datasets are available through API requests (ERA5, Copernicus). We would propose to use these as much as possible.

Provide database for datasets behind request form

Many datasets however come from webpages with a download link/request form. For the selected models, we propose to download this data and store them in a database where they can be retrieved for modelers using the SOILPROM platform. Alternatively, and less work demanding, would be to provide a list of weblinks where these datasets are to be downloaded/requested. This would however delay the modeler since they would have to manually request some data.

Bilateral communication EUSO

Given that EUSO currently lacks formal data sharing requirements and DestinE primarily focuses on remote sensing data, we favour a pragmatic approach. This solution allows us to address the immediate needs of the project efficiently while maintaining the flexibility necessary to adapt if further requirements emerge in the future. We propose to keep contact with EUSO and provide feedback on the current bottlenecks in the project, so that they can prioritize their work accordingly to the needs of EU projects.

Appendix 1: References and Related Documents

References

Wilkinson, M et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).

Annex 1

Annex 1 is this online sheet: https://docs.google.com/spreadsheets/d/1uhA-k-pH5RgFqxW57sx84HAjO-YDl-e6/edit?usp=sharing&ouid=108896081012526557552&rtpof=true&sd=true