SoilPulse – A pulse towards reusable Soil Process Data

Conrad Jackisch[1], Jonas Lenz[1,3] and Jan Devátý[2]
[1] TU Bergakademie Freiberg, Interdisciplinary Environmental Research Centre, [2] CTU in Prague, Department of Landscape Water Conservation, [3] IPROconsult GmbH, Dresden
Submitted on March 31, 2023

Abstract

1. *Data about soil functions and processes are fundamental for various applications to meet the SDGs, mitigate effects of climate change and to reverse biodiversity loss.* ***2.*** *Current experimental procedures, analytical methods and (meta)data standards have a high level of diversity, because of scale and spatio-temporal dependency and model specific demands. This leads to widespread absence of data interoperability and to hampered scientific progress.* ***3.*** *SoilPulse will propose means to facilitate data interoperability and target reusability by developing a soil-process related metadata generator and test for automated data reading and evaluation. In this, SoilPulse will guide users in aligning their (meta)data with standardized formats, while allowing them to keep their established work flows and data storage systems.* ***4.*** *SoilPulse will integrate and extend current (meta)data standards. Our tools will guide metadata creation and enable machine-readable data structures and automated tests for interoperability and reusability.* ***5.*** *The (meta)data standard and the provided tools will support data providers and curators, modelers, authorities, consultants and decision makers in land-use and soil related questions.* ***6.*** *SoilPulse eases metadata generation and data curation and is intended i) to be extended to further application in dynamic soil data and ii) to be incorporated into larger environmental RDM structures.*

I. Introduction

I.1 What is the scientific context of your pilot?

Soil functional data are essential to further develop the understanding of the complex and nonlinearly interacting land surface system, as well as to improve numerical models for hydropedological processes (e.g. infiltration, surface runoff and erosion) under fundamentally changing conditions [1]. Land degradation and soil erosion are very severe issues in this respect, which integrate several scientific domains (soil, agriculture, biodiversity, …), various scales and different levels of detail.

I.2 What is the data-challenge you face and what is the state-of-the-art?

Extensive primary data is generated in meticulous field work producing almost uniquely structured records for each experiment. For applications these data are aggregated to single values, thereby losing information required for other application cases (e.g. different evaluation

procedures, models, technical analyses). If primary data is published at all, it lacks common structures and uses various storage systems (i.e. relational databases or spreadsheets). Any uptake of such non-interoperable data requires laborious, dataset specific and manual processing. Recent attempts to create inter-institutional databases [2, 3] resulted in the creation of their own (meta)data structure, deviating from existing intra-institutional frameworks and not overcoming manual data curation requirements.

I.3 Vision for your community and their RDM workflows

Reusable, scale-aware, spatio-temporal discrete soil process data would allow for dynamic (dis)aggregation, for easy link to numerical models and for advancing from descriptive pedology to soil functional process understanding. This can become a foundation to rethink model and experimental concepts. SoilPulse helps experimentalists and guides modelers for fertilizing data exchange across domains, scales, teams and subdisciplines.

II. Pilot description

II.1 What is the proposed solution to your data-challenge?

Interoperability of existing [i.e. 2, 3, 4a, 4b, 5] and new datasets will be ensured by mapping their structure to a metadata standard, guided by the SoilPulse metadata generator. Curation overhead for researchers is minimized as they can maintain their established data structures and reuse mapping templates. The metadata standard will build upon existing standards (*see II.3*) and extend it for functional soil process analyses, in which we consider the experience from NFDI4Earth Pilot “Interoperability and Reusability of Geoscientific Lab Data'' in semantic mapping. Interoperability and reusability of the data will be demonstrated with automated tests, which implies that datasets can be standardized linked to earth system models (see objectives of the NFDI4Earth Pilot “Seamless Earth System Model-Data Integration“).

II.2 What is the technological backbone you rely upon?

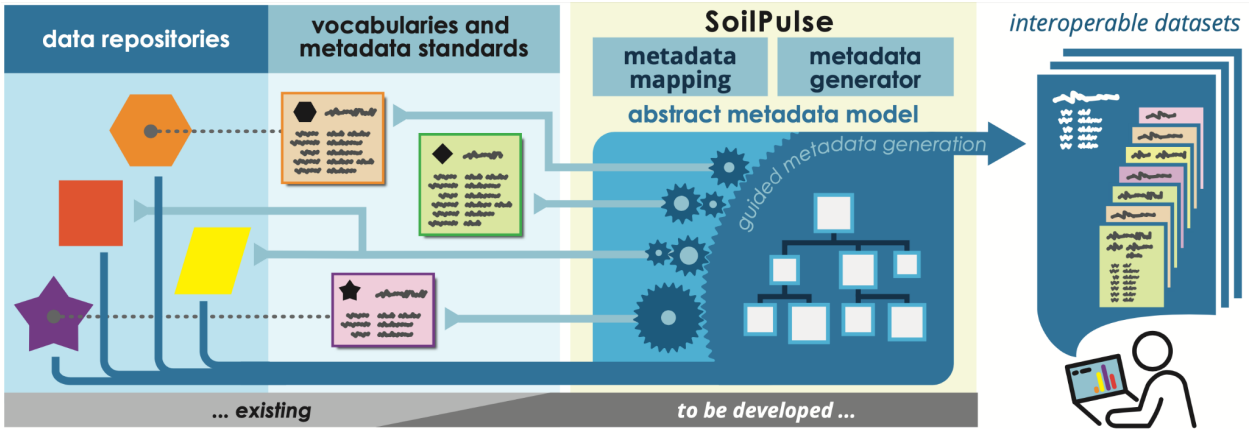
SoilPulse assists metadata generation and API-assisted upload for new and existing datasets through an open R/python-package and with a web app frontend, strongly extending current MetaEditors [6, 7, 8]. It will be publicly documented and hosted on GitHub. To improve findability of interoperable SoilPulse datasets, they will be listed by DOIs in the GitHub repo. A metadata-tag in the dataset linking to SoilPulse is foreseen. Defined FAIR MI tests and the automatic generation of human-readable reports will test data set interoperability and reusability.

II.3 Which are the standards and interoperability approaches used in the pilot’s context?

The semantically interoperable SoilPulse metadata standard will build upon existing standards (e.g. BONARES; ISCN, building on INSPIRE and Dublin Core), vocabularies (e.g. AGROVOC, upcoming GAIA Data GO FAIR) and data harmonization approaches (e.g. SoDaH; ESIP).

II.4 Which improvements are made compared to the status quo?

Interoperability of data repositories will be ensured by mapping their individual data structures to common vocabulary and metadata standards. Automated reusability tests will be developed. Data curation effort for individual researchers will be reduced through guided metadata generation and the ability to retain their established data structure and workflows.



III. Relevance for the NFDI4Earth

III.1 What are expected users and stakeholders and how do they benefit from your solution?

SoilPulse closely links researchers (experimentalists, data analysts and modelers) by providing an easy accessible entry point to enhanced data documentation and exchange. SoilPulse eases and standardizes access to research data also for system integrators (e.g. consulting companies and public authorities).

III.2 What measures are planned to support the uptake of your solution?

We will give presentations/workshops at national (DBG) and international (EGU) conferences demonstrating domain-specific data–metadata generation and domain-independent workflows. Open documentation and code will be published alongside respective manuscripts.

III.3 What is the potential for other sub-branches in the Earth System Sciences?

We provide an application of a metadata generator and reusability tester suitable for cases other than the development test-case (soil function and erosion). Our approach towards harmonized metadata standards for site, time, scale and application sensitive data will be documented for uptake in neighboring domains.

III.4 What elements of FAIR are particularly addressed?

SoilPulse ensures **I**nteroperability by mapping datasets to a common metadata standard, keeping a low metadata generation overhead for researchers. Data **R**eusability will be automatically tested and **F**indability is enhanced through metadata-tags and data registration.

III.5 What aspects of the research data life cycle are particularly addressed?

Soil Pulse addresses the research data life cycle at publishing stage (metadata generator) to the reuse stage (automated reusability tests and *FAIRification* of existing datasets).

III.6 Are there particular contributions that help the NFDI4Earth to engage with?

SoilPulse will demonstrate how to improve the reusability of data with high diversity in spatio-temporal detail and data structure. The metadata generator can be used as a blueprint for other domains and applications with low curation overhead.

IV. Deliverables

IV.1 Technical operability of the pilot (e.g., software shared, developed interface, implementation)

1. A metadata generator guides researchers to the creation of reusable data sets in a webapp. **2.** Automatic tests for reusability, including automatic generation of human readable reports of the data sets. **3.** SoilPulse metadata standard definition, which refers to existing standards.

IV.2 Material or actions for dissemination of knowledge/data

1. R/python package and documentation as entry point and demonstrator for the metadata generation on the example of data sets from several erosion research groups. **2.** Documented demonstration of data reusability in automatically generated machine and human readable reports or coupling to earth process models.

IV.3 Roadmap document for the community

The roadmap "SoilPulse+" projects how the pilot's development can be extended to other soil functions and processes and showcase the coupling of reusable data with earth system models.

V. Work Plan & Requested funding

We apply for 12 months (PM) funding TVL E13 (71163.36 €, including material resources).

Table 1: SoilPulse one-year implementation plan: **working process**, **milestone**

Task \ Month	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Prepare												
Generator												
Tester												
Evaluation												
Roadmap												
Outreach				←	/	→				←	//	→

- **Prepare** (estimated 2 PM): scoping vocabulary repositories, metadata standards and harmonization approaches → **Milestone:** define the SoilPulse metadata-standard
- **Generator** (est. 4 PM): development of the metadata generator as R/python package
- **Tester** (est. 1 PM): definition of Interoperability FAIR MI tests
- **Evaluation** (est. 2 PM): applying SoilPulse assisted metadata generation on database and file-based repositories + and showcase reusability by active interaction through metadata → **Milestone:** documentation demonstrating curation and interoperability
- **Roadmap** (est. 1 PM): inline docu for Roadmap "SoilPulse+" → **Milestone:** Roadmap
- **Outreach** (est. 2 PM): Promoting SoilPulse in relevant NFDI-Projects and Conferences (DBG, EGU, ...) → **Milestones:** Presentation/Workshop at two conferences