TUBAF
Die Ressourcenuniversität.
Seit 1765.

FACULTY OF CIVIL ENGINEERING
CTU IN PRAGUE

IPROconsult

NFDI4Earth
Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

NFDI4Earth Pilot:

SoilPulse

MAKE VARIOUS DATA ABOUT SOIL PROCESSES INTEROPERABLE WHILE MAINTAINING ESTABLISHED WORK FLOWS AND DATA STORAGE SYSTEMS

Jonas Lenz
Conrad Jackisch
Jan Devátý

SoilPulse @ NFDI4Earth Plenary
May 23th 2024

SoilPulse - Motivation

Process data from field experiments
→ very informative but highly specific
→ weakly structured, wild metadata

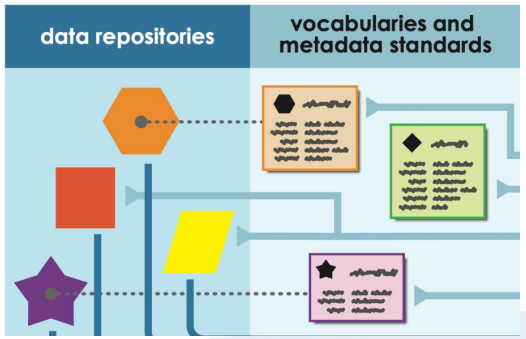
our example: rainfall-runoff simulations in field or lab

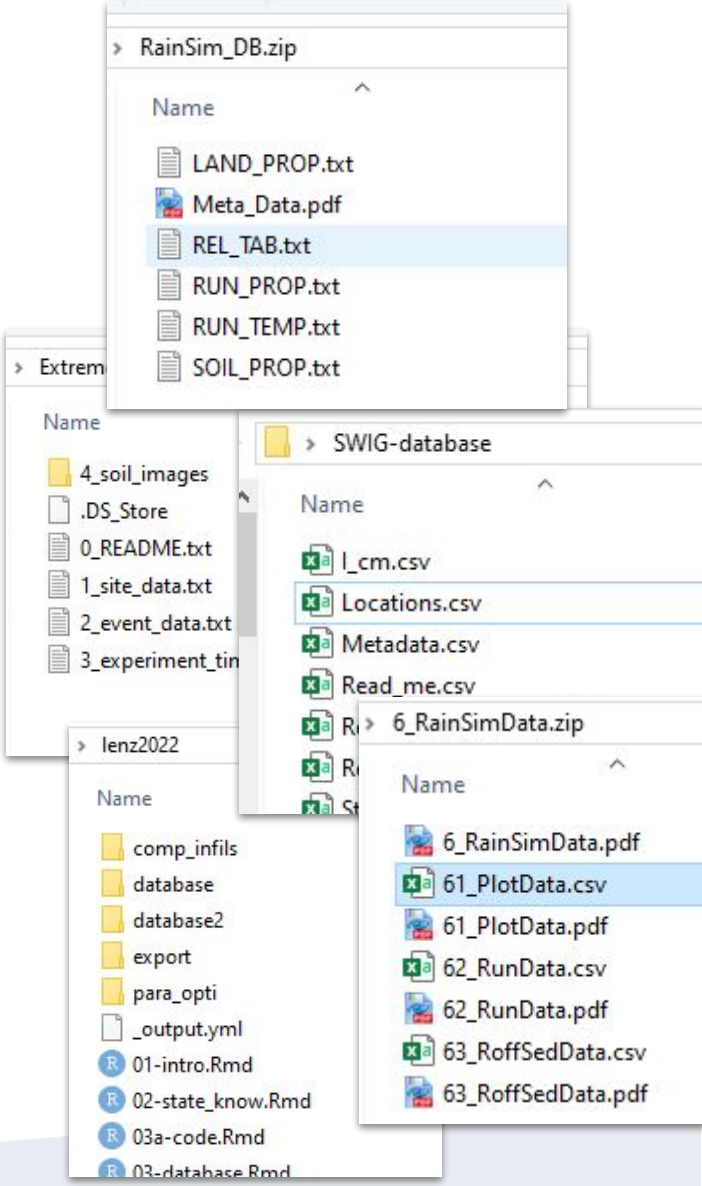
Issue: Missing standards, individual requirements, “traditional” procedures

- in experimental methods
- in recorded data and metadata
- in data management

State of data resources:
unFAIR, unpublished, incompatible

- limited awareness/competence
- no resources to revisit data





SoilPulse - Aims

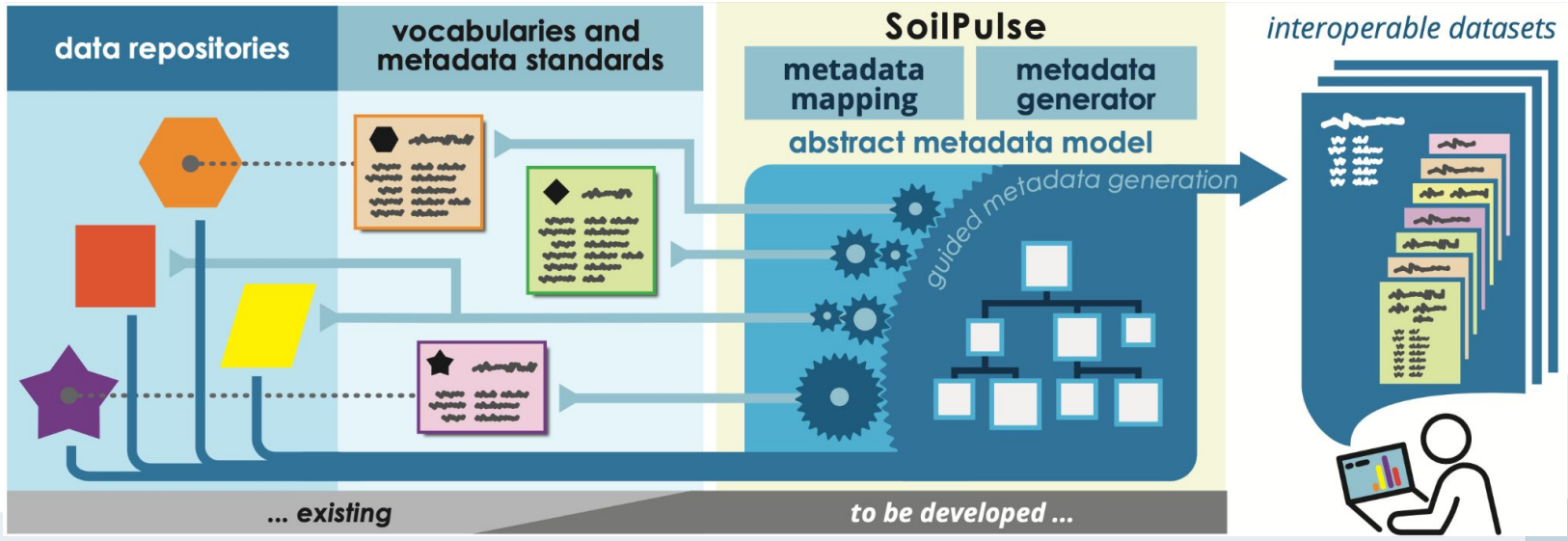
Make existing datasets FAIR – to save them and to make them (re)usable
→ tool for data harmonization
→ assisted, semi-automatic metadata generation/enrichment

Semi-automatic analysis (within context of rainfall experiments)
→ attempt to analyse converted datasets as prove of success/ quality control
→ prepare (meta-)data queries

Learn from (meta-)data needs
→ further/amend/revise existing metadata schemes
→ feedback deviations between datasets as easy to implement within respective procedures

On practical terms:

A tool to FAIRify existing data
→ raise awareness through demonstration at own data
→ direct positive effect for user w/ metadata queries and quality/interpretability checks
→ direct positive effect for community w/ interoperable data & streamlined schemes



Why does this need a pilot?

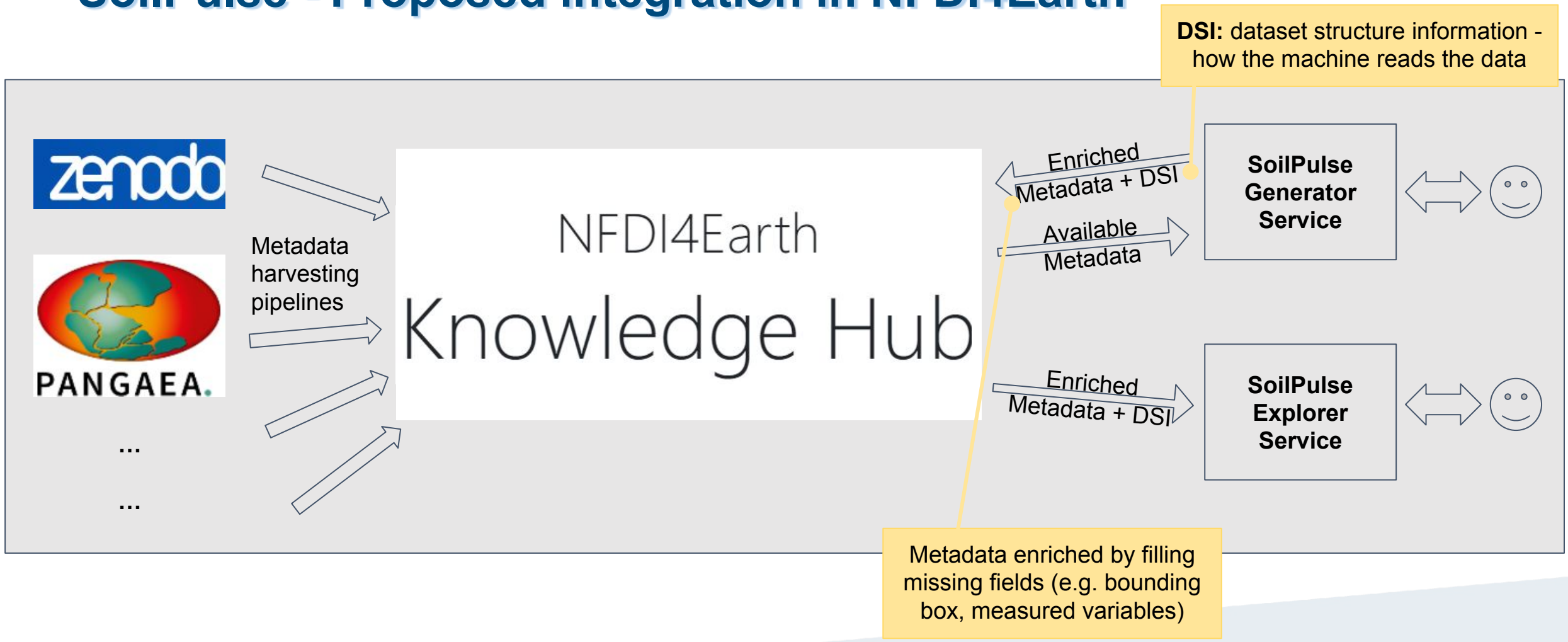
We are sure that there are many examples for valuable, existing data which require a lot of work to standardise.

1. This data is worth saving
2. There is little capacity (and capability) to prepare these data as FAIR
3. This can become a vehicle to define and promote open data and metadata schemes

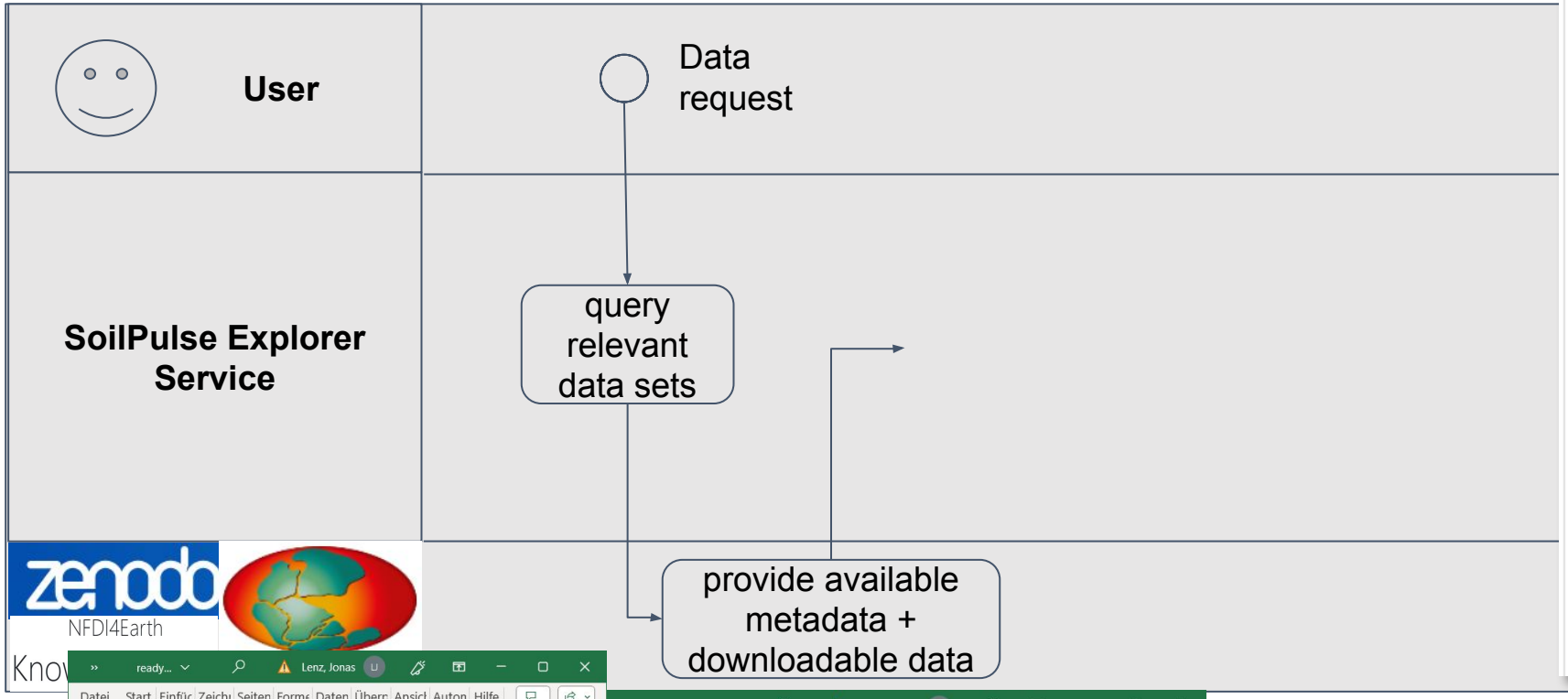
Rainfall experiments are a good example with respect to
→ missing standards,
→ process and site complexity,
→ varying vocabulary

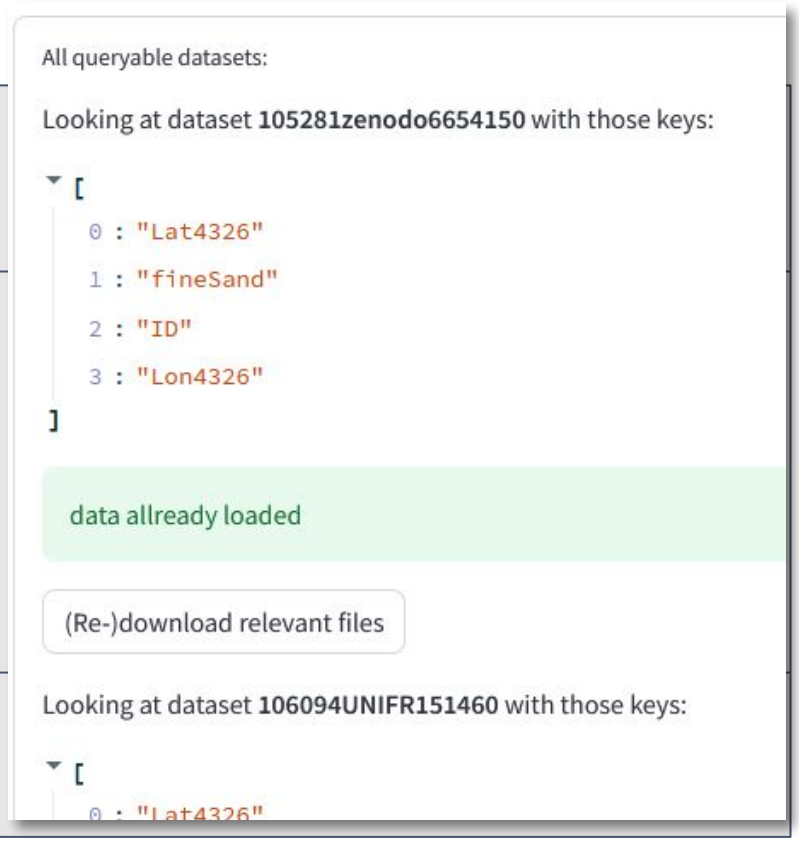
1. Place- and condition-dependent measurements
2. Functional characteristics, which require interpretation
3. Setups, instruments, procedures, naming conventions depend strongly on perceptual model and traditional approaches in groups
4. Contains point measurements, time series, images, links to monitoring of reference states,… which can be accumulated into one final measurement value (soil erosion rate) or analysed in more detail

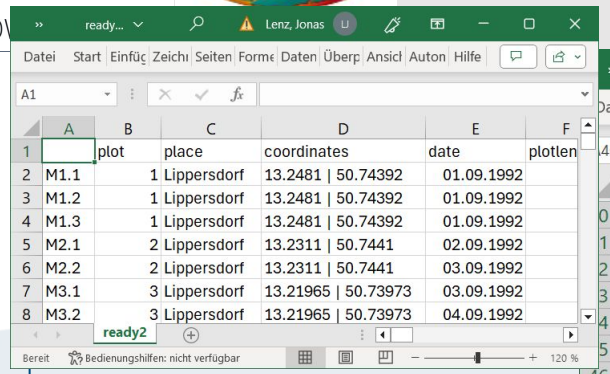
SoilPulse - Proposed integration in NFDI4Earth



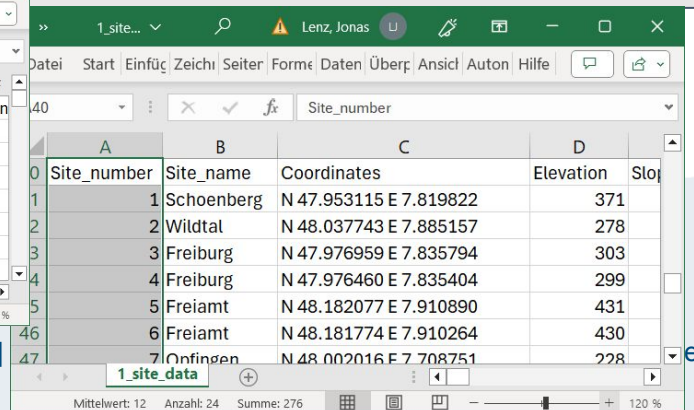
SoilPulse - Explorer







	plot	place	coordinates	date	plotlen
M1.1	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
M1.2	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
M1.3	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
M2.1	2	Lippersdorf	13.2311 | 50.7441	02.09.1992	
M2.2	2	Lippersdorf	13.2311 | 50.7441	03.09.1992	
M3.1	3	Lippersdorf	13.21965 | 50.73973	03.09.1992	
M3.2	3	Lippersdorf	13.21965 | 50.73973	04.09.1992	



Site_number	Site_name	Coordinates	Elevation
1	Schoenberg	N 47.953115 E 7.819822	371
2	Wildtal	N 48.037743 E 7.885157	278
3	Freiburg	N 47.976959 E 7.835794	303
4	Freiburg	N 47.976460 E 7.835404	299
5	Freiamt	N 48.182077 E 7.910890	431
6	Freiamt	N 48.181774 E 7.910264	430
7	Opfingen	N 48.002016 E 7.708751	228

SoilPulse - Explorer

DSI: dataset structure information - how the machine reads the data

User

Data request

SoilPulse Explorer Service

query relevant data sets

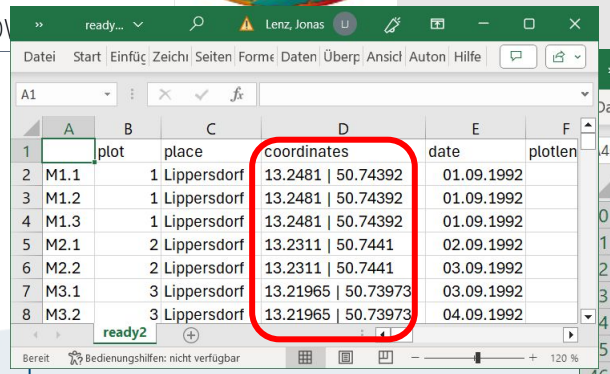
apply related DSI

provide available metadata + downloadable data

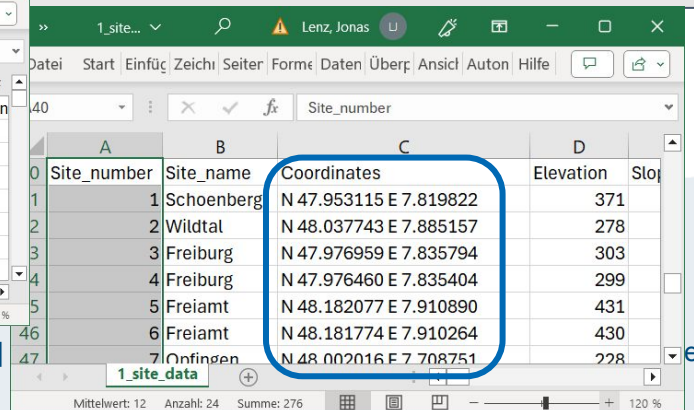
zenodo

NFDI4Earth

Know



	plot	place	coordinates	date	plotlen
M1.1	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
M1.2	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
M1.3	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
M2.1	2	Lippersdorf	13.2311 | 50.7441	02.09.1992	
M2.2	2	Lippersdorf	13.2311 | 50.7441	03.09.1992	
M3.1	3	Lippersdorf	13.21965 | 50.73973	03.09.1992	
M3.2	3	Lippersdorf	13.21965 | 50.73973	04.09.1992	



Site_number	Site_name	Coordinates	Elevation
1	Schoenberg	N 47.953115 E 7.819822	371
2	Wildtal	N 48.037743 E 7.885157	278
3	Freiburg	N 47.976959 E 7.835794	303
4	Freiburg	N 47.976460 E 7.835404	299
5	Freiamt	N 48.182077 E 7.910890	431
6	Freiamt	N 48.181774 E 7.910264	430
7	Opfingen	N 48.002016 E 7.708751	228

SoilPulse - Explorer

DSI: dataset structure information - how the machine reads the data

User

Data request

SoilPulse Explorer Service

query relevant data sets

apply related DSI

compile user requested data

provide available metadata + downloadable data

zenodo

NFDI4Earth

Know

Query for which colunns

ID × Lat4326 × Lon4326 ×

query these columns

	ID	Lat4326	Lon4326	dataset
122	113	51.1705	13.9772	105281zenodo6654150
123	114	51.1705	13.9772	105281zenodo6654150
124	115	51.1705	13.9772	105281zenodo6654150
125	116	51.1705	13.9772	105281zenodo6654150
126	1	47.9531	7.8198	106094UNIFR151460
127	2	48.0377	7.8852	106094UNIFR151460

ready2

	plot	place	coordinates	date	plotlen
M1.1	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
M1.2	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
M1.3	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
M2.1	2	Lippersdorf	13.2311 | 50.7441	02.09.1992	
M2.2	2	Lippersdorf	13.2311 | 50.7441	03.09.1992	
M3.1	3	Lippersdorf	13.21965 | 50.73973	03.09.1992	
M3.2	3	Lippersdorf	13.21965 | 50.73973	04.09.1992	

1_site_data

Site_number	Site_name	Coordinates	Elevation
1	Schoenberg	N 47.953115 E 7.819822	371
2	Wildtal	N 48.037743 E 7.885157	278
3	Freiburg	N 47.976959 E 7.835794	303
4	Freiburg	N 47.976460 E 7.835404	299
5	Freiamt	N 48.182077 E 7.910890	431
6	Freiamt	N 48.181774 E 7.910264	430
7	Opfingen	N 48.002016 E 7.708751	228

SoilPulse - Explorer

DSI: dataset structure information - how the machine reads the data

User

Data request

Do research task

SoilPulse Explorer Service

query relevant data sets

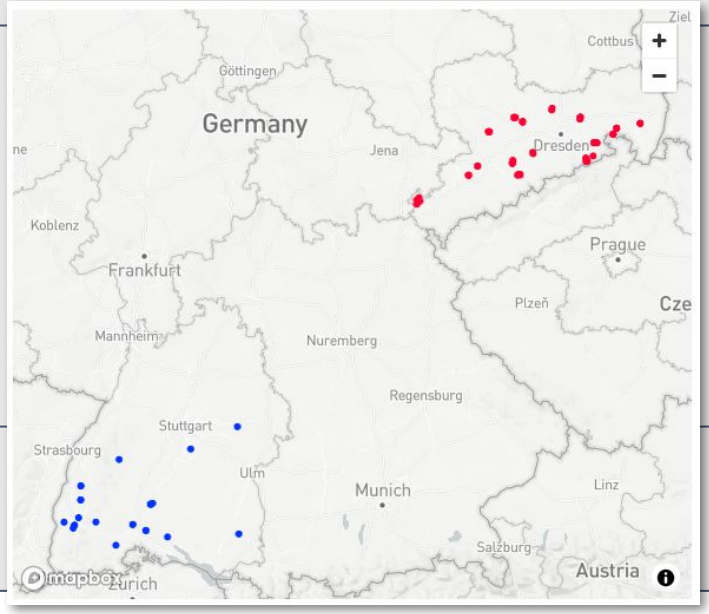
apply related DSI

compile user requested data

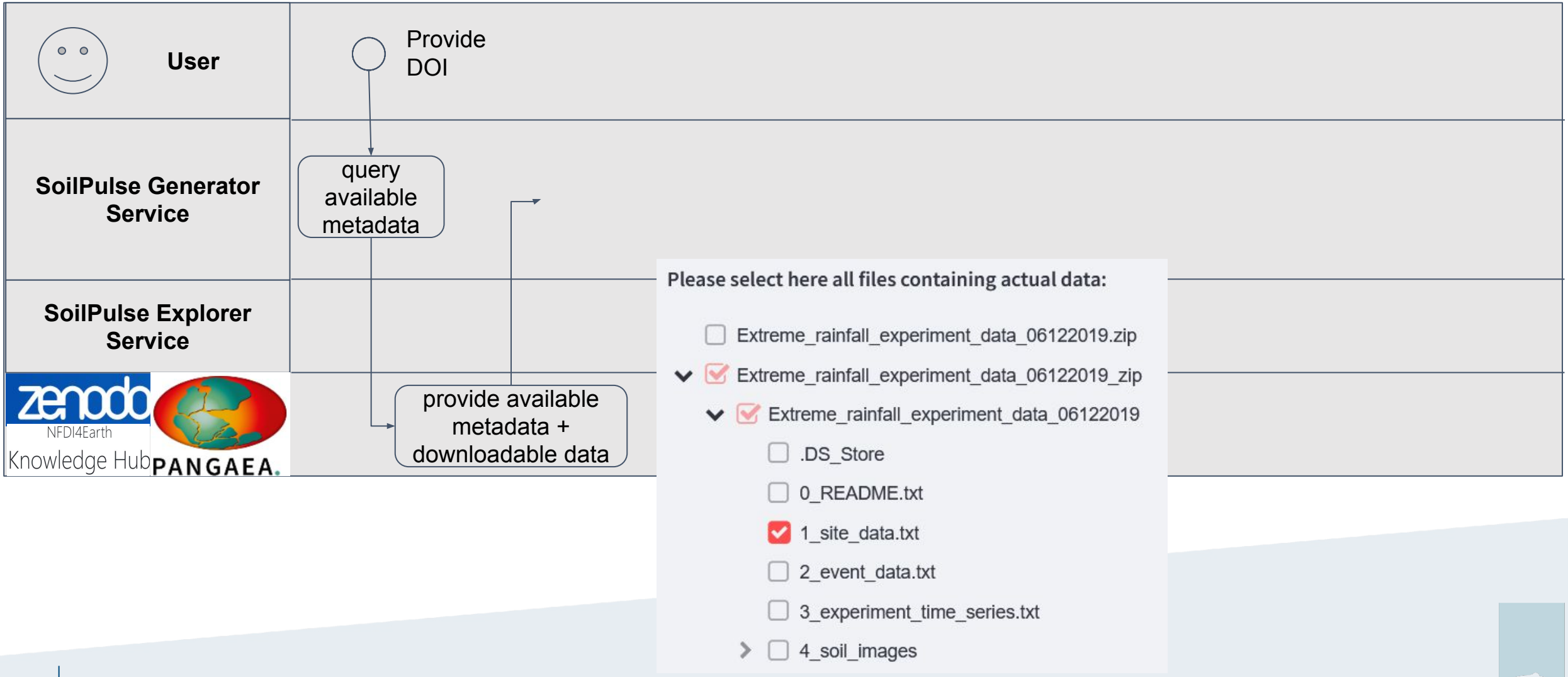
provide available metadata + downloadable data

zenodo NFDI4Earth Knowledge Hub

PANGAEA.

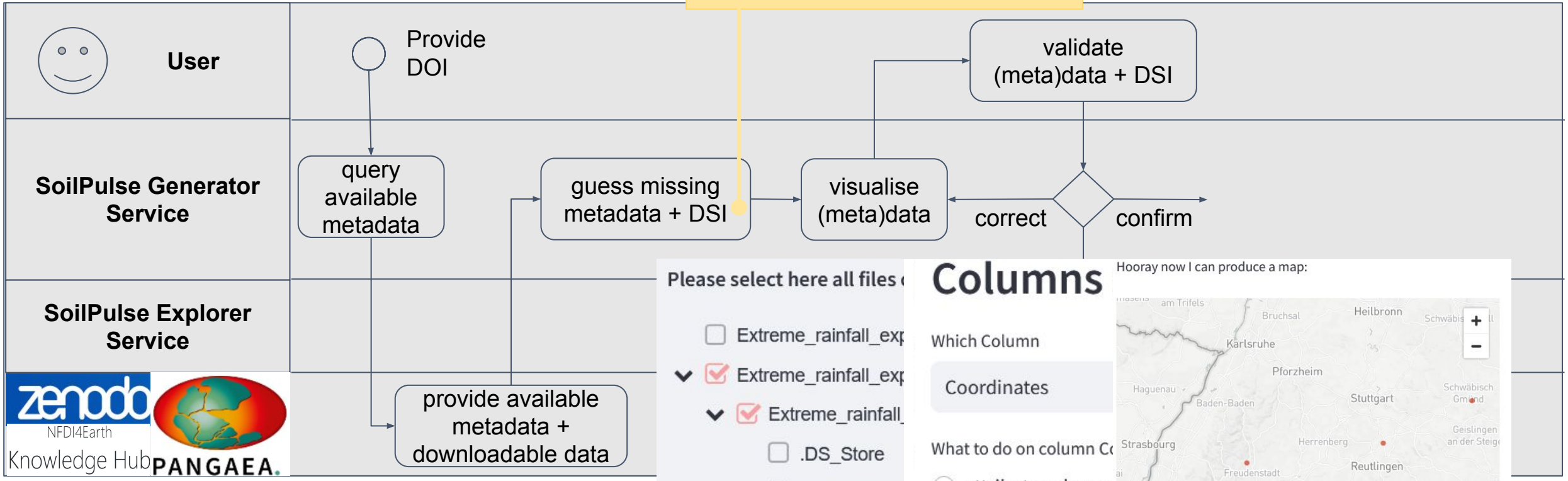


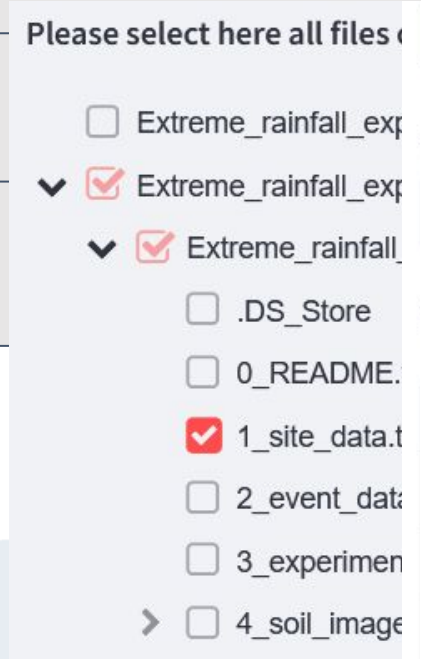
SoilPulse - Generator

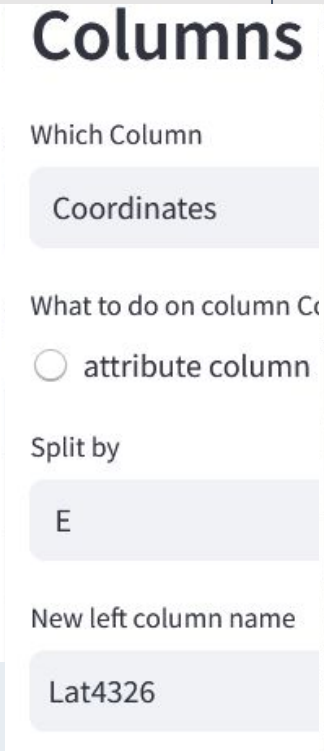


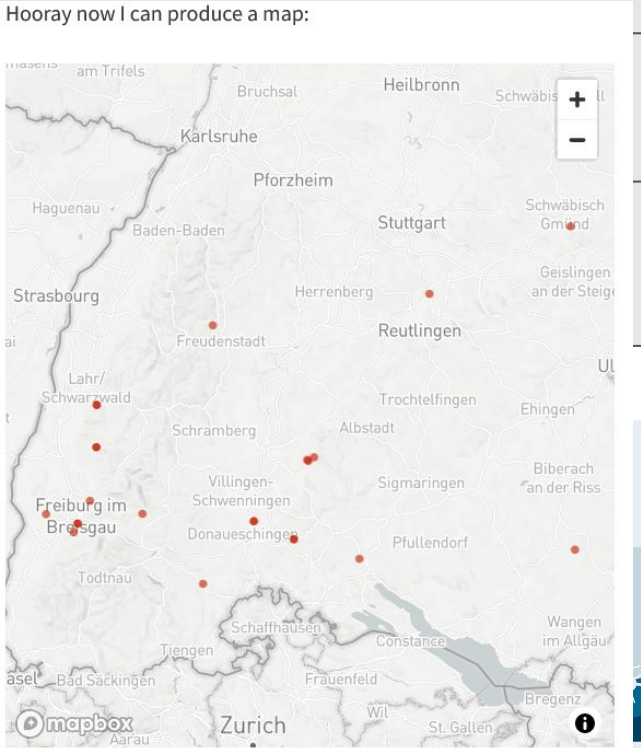
SoilPulse - Generator

DSI: dataset structure information - how the machine reads the data



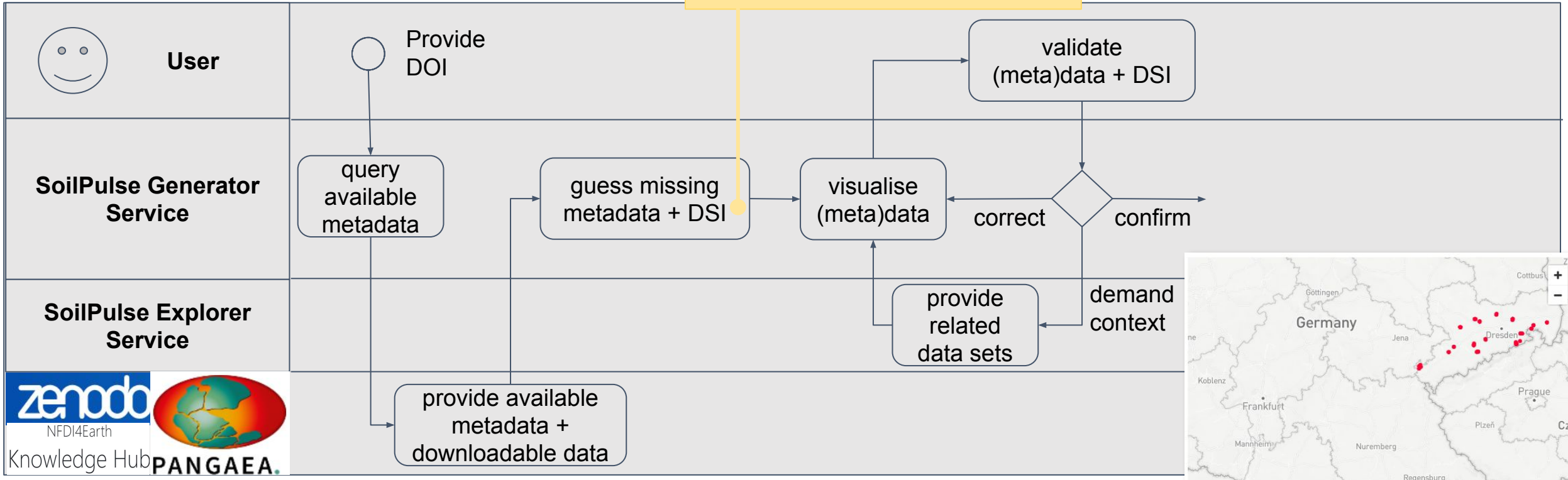


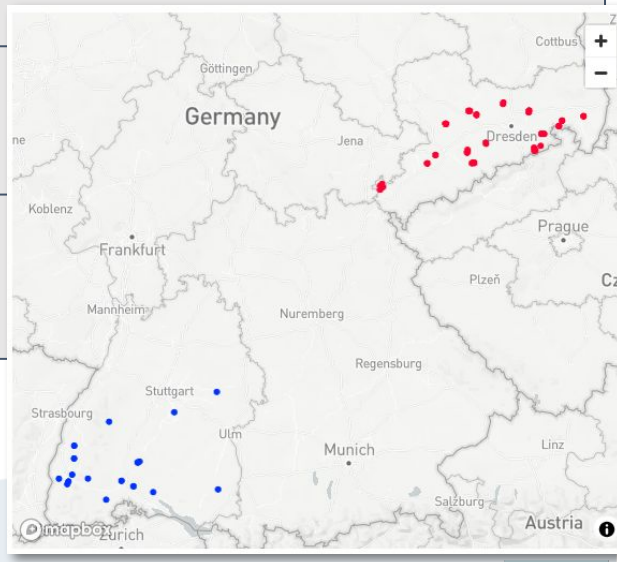




SoilPulse - Generator

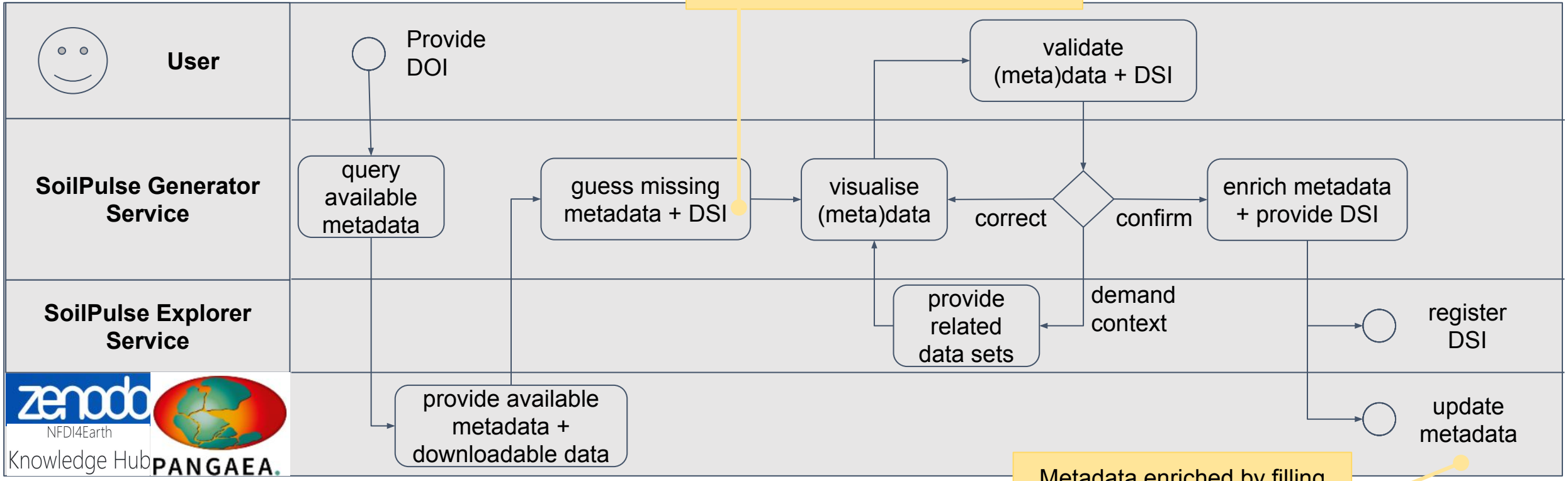
DSI: dataset structure information - how the machine reads the data





SoilPulse - Generator

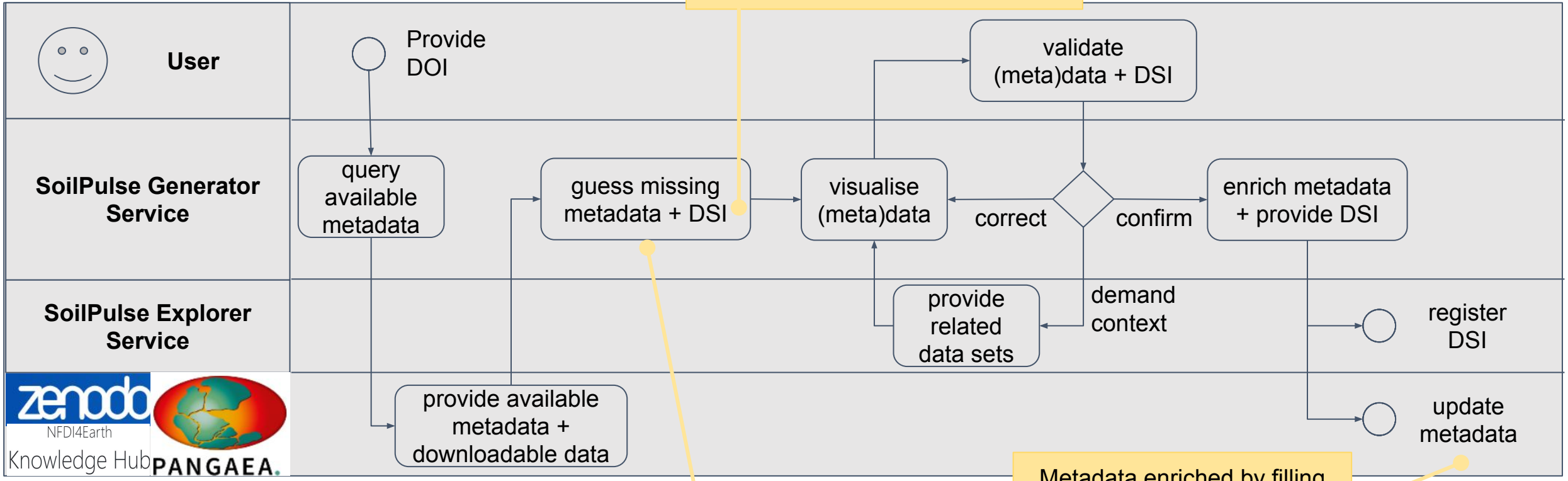
DSI: dataset structure information - how the machine reads the data



Metadata enriched by filling missing fields (e.g. bounding box, measured variables)

SoilPulse - Generator

DSI: dataset structure information - how the machine reads the data



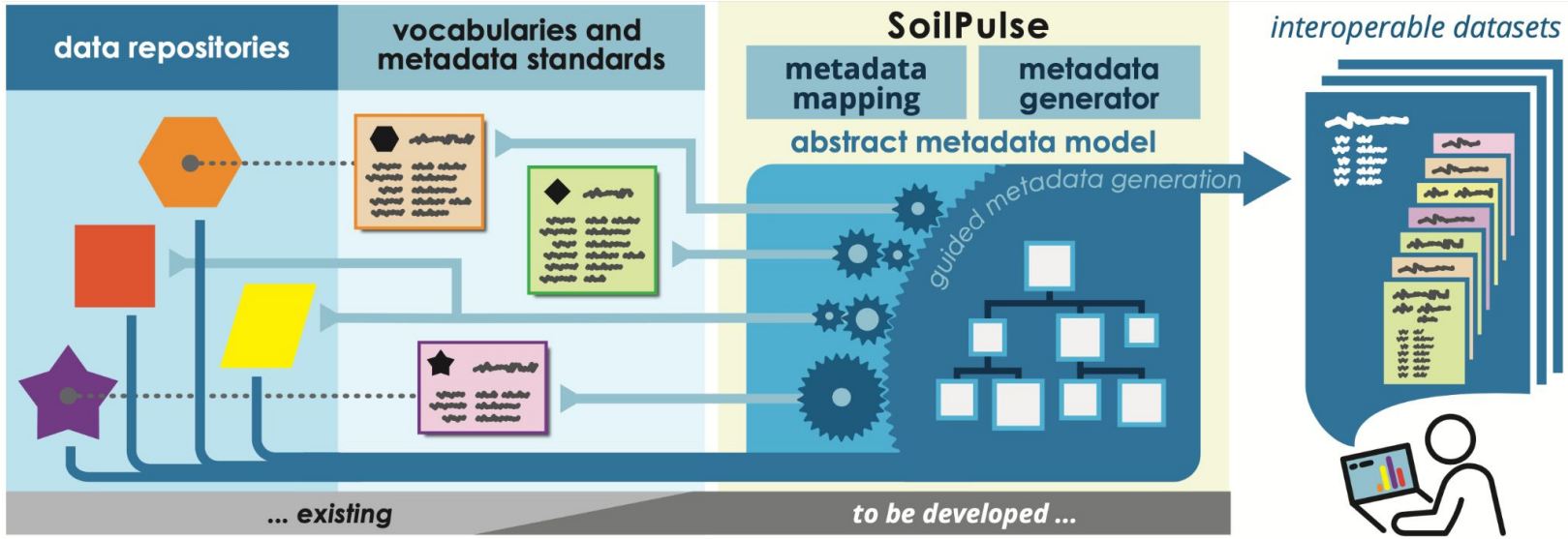
Initial guess:
- Planned: by algorithms
- Future: apply LLM

Metadata enriched by filling missing fields (e.g. bounding box, measured variables)

SoilPulse

MAKE VARIOUS DATA ABOUT SOIL PROCESSES INTEROPERABLE WHILE MAINTAINING ESTABLISHED WORK FLOWS AND DATA STORAGE SYSTEMS

see you at the poster…



More on the project:

https://soilpulse.github.io/

Demo on streamlit:

https://soilpulse-egu.streamlit.app

Get in touch:

mailto:conrad.jackisch@tbt.tu-freiberg.de

Thanks for funding:

NFDI4Earth
Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

SoilPulse - Components 1/2

Metadata generation/enrichment through a web interface (Demo: https://soilpulse-egu.streamlit.app/Metadata_retriever)

- User/Data creator provides files (**non standardized structure**), metadata shall be generated/enriched as automatically as possible, so the user “only” needs to approve and complete it.
  → Data structure needs to be mapped within metadata.
  → User gets feedback how the machine understands his data, while preparing the metadata.
  → User gets feedback if his data complies with data of other resources.
  → Semi-automatic generation of submission ready metadata to data files.

- Also applicable to **already published** resources (e.g. Datasets on Zenodo) -> Reference to the resource is then included in metadata to avoid republication.

SoilPulse - Components 2/2

(Meta-)Data exploration (Demo: https://soilpulse-egu.streamlit.app/Explorer)

- Metadata becomes access point for data
- Making data points queryable: “Get runoff values from all rainfall simulation experiments with total organic carbon content > 3% of a soil sample.”
- Feed data aggregates to models by defining model requirement templates.
- Combination with data from other resources.

- Requires (self-hosted) live system/ server holding (temporarily) all data.

Which Metadata do we need to generate in addition to existing metadata to increase data reusability?

SoilPulse - Data and issues 1/2

Soil, Erosion/Infiltration Experiments

- at the boundary between hydrology, agriculture and soil properties
- various process' observation
- state dependent (intial water content, plant development, …)
- functional characteristics of soils/ experimental sites

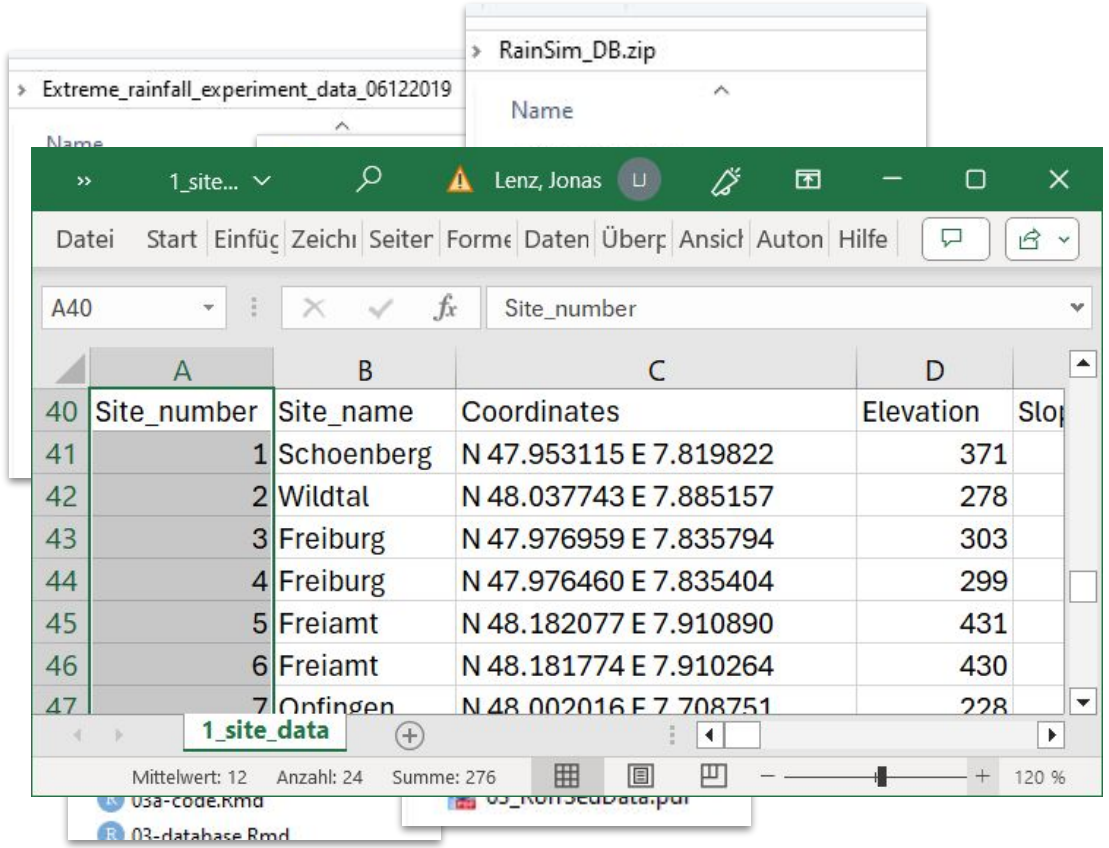
Data types

- single measurements of soil properties
- descriptions of treatment (last or history), plant development
- time series of processes (runoff, irrigation intensity)
- images
- …

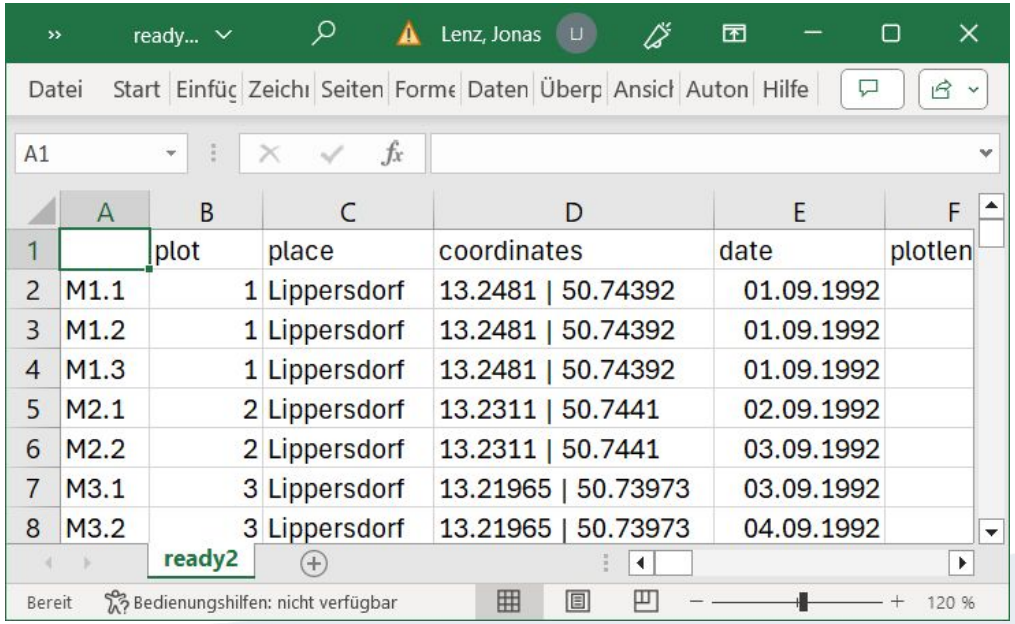
SoilPulse - Data and issues 2/2

Missing standards in data management

- Historic research data since
- Differing table structures → capture dataset structure information **(DSI)** in metadata
- Differing namings → Use controlled Vocabulary (e.g. Agrovoc)



	A	B	C	D	E
40	Site_number	Site_name	Coordinates	Elevation	Slop
41	1	Schoenberg	N 47.953115 E 7.819822	371	
42	2	Wildtal	N 48.037743 E 7.885157	278	
43	3	Freiburg	N 47.976959 E 7.835794	303	
44	4	Freiburg	N 47.976460 E 7.835404	299	
45	5	Freiamt	N 48.182077 E 7.910890	431	
46	6	Freiamt	N 48.181774 E 7.910264	430	
47	7	Opfingen	N 48.002016 E 7.708751	228	



	A	B	C	D	E	F
1		plot	place	coordinates	date	plotlen
2	M1.1	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
3	M1.2	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
4	M1.3	1	Lippersdorf	13.2481 | 50.74392	01.09.1992	
5	M2.1	2	Lippersdorf	13.2311 | 50.7441	02.09.1992	
6	M2.2	2	Lippersdorf	13.2311 | 50.7441	03.09.1992	
7	M3.1	3	Lippersdorf	13.21965 | 50.73973	03.09.1992	
8	M3.2	3	Lippersdorf	13.21965 | 50.73973	04.09.1992	

SoilPulse - Domain specific schemes

BONARES – a candidate scheme for soil data

- has INSPIRE, bla blubb
-
- Datenmanagement standard -> if existing, not consistent of time, specific to group
- Metadata Standard -> too general/technical for our issues

Keep FOCUS! We have to build the wrapper now – but this will be doen by LLMs soon. The issue will remain the science related things: Which information is expected? which perception about the processes and techniques is the foundation…

SoilPulse - Metadata schema 1/2

Adaptation of bonares metadata schema (Gärtner et al. 2017)

- Extensive soil specific schema, building upon INSPIRE and DataCite (https://doi.org/10.1016/j.cageo.2019.07.005)
- Bonares has metadata down to table structure and relation of tables

Extension:

- Assignment of controlled vocabulary concepts to single data points/columns:
  - e.g. “SOC”/”TOC”/”Corg” of original datasets becomes “total organic carbon” of AGROVOC (http://aims.fao.org/aos/agrovoc/c_c35fdd26).
- Make metadata within files readable for machines (e.g. table structure, timesteps, experiment ID). → Map down to single values.

SoilPulse - Metadata schema 2/2

Implementation (in progress):

- Devátý, J., Lenz, J., and Jackisch, C.: SoilPulse – A software package for semi-automated metadata management and publication, EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-18775, https://doi.org/10.5194/egusphere-egu24-18775, 2024.

- will be available as python package