Plan Overview

A Data Management Plan created using DMPonline

Title: Artificial intelligence for building open and accessible knowledge and data repositories to safeguard soils

Creator: Beichen Wang

Principal Investigator: Beichen Wang

Data Manager: Beichen Wang

Project Administrator: Beichen Wang

Affiliation: Wageningen University and Research (Netherlands)

Funder: European Commission

Template: Data Management Plan | Wageningen University and Research

ORCID iD: 0009-0008-5213-5114

Project abstract:

Soil degradation is currently a critical challenge in Europe, where 60-70% of soils are considered unhealthy. The project aims to enhance soil knowledge accessibility and usability through artificial intelligence, supporting the European Commission's goal of achieving 75% healthy soils by 2030. Current soil-related knowledge is often contained within unstructured texts, hindering its effective utilization for informed soil management and policymaking. To overcome this, the research focuses on four key areas: structuring soil knowledge from diverse textual sources, expanding this knowledge by interlinking it with external data, designing a natural language query system, and evaluating the real-world applicability of these tools in soil management decision-making. The project will leverage advanced technologies such as large language models, ontologies, and knowledge graphs to create a comprehensive soil health knowledge repository. By enhancing the integration and accessibility of soil knowledge, this research aims to significantly advance soil science and support sustainable soil management practices. The outcomes have the potential to impact a wide range of stakeholders, from researchers and policymakers to farmers, by providing a user-friendly knowledge base for informed decision-making in soil management.

ID: 178649

Start date: 01-04-2024

End date: 31-03-2028

Last modified: 22-05-2025

Grant number / URL: 101112838

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Artificial intelligence for building open and accessible knowledge and data repositories to safeguard soils

A. Describe the research project

1. Name researcher (please, add your full name):

Beichen Wang

2. What is the name of your department(s)?

- Plant Sciences

3. What is the name of your chair group(s) or business unit(s)? English name and abbreviation for chair groups from this page; business units from this page (expand to Wageningen Research and keep expanding to find your specific division / group). Examples: Bioprocess Engineering (BPE) or Contract Research Organization (CRO).

· Artificial Intelligence Group (AIN)

4. Describe the organisational context of your research project.

DMP version (or date last modified)	2025-05-22
Supervisor / (co-)promotors	Prof. dr. Anna Fensel, Dr. Luís Moreira de Sousa
Graduate School (WU only)	PE&RC
Start date of project	2024-04-01
End date of project	2028-03-31
Project number	3186400013
Funding body	European Commission

5. Give a short description of your research project.

Title	Artificial intelligence for building open and accessible knowledge and data repositories to safeguard soils
Summary	Soil degradation is currently a critical challenge in Europe, where 60-70% of soils are considered unhealthy. The project aims to enhance soil knowledge accessibility and usability through artificial intelligence, supporting the European Commission's goal of achieving 75% healthy soils by 2030. Current soil-related knowledge is often contained within unstructured texts, hindering its effective utilization for informed soil management and policymaking. To overcome this, the research focuses on four key areas: structuring soil knowledge from diverse textual sources, expanding this knowledge by interlinking it with external data, designing a natural language query system, and evaluating the real-world applicability of these tools in soil management decision-making. The project will leverage advanced technologies such as large language models, ontologies, and knowledge graphs to create a comprehensive soil health knowledge repository. By enhancing the integration and accessibility of soil knowledge, this research aims to significantly advance soil science and support sustainable soil management practices. The outcomes have the potential to impact a wide range of stakeholders, from researchers and policymakers to farmers, by providing a user-friendly knowledge base for informed decision-making in soil management.

6. List the individuals responsible for the following data management tasks.

Data collection	Beichen Wang (PhD candidate)
Data quality	Beichen Wang (PhD candidate) Anna Fensel (Professor AIN)
Storage and backup	Beichen Wang (PhD candidate)
Data archiving / publishing	Beichen Wang (PhD candidate)
Data stewardship / support	WUR Library RDM support (data@wur.nl)

7. I have requested a review of this data management plan from:

- WUR Library - Data Management Support (data@wur.nl).

8. Name of the data management support staff and / or data steward consulted during the preparation of this plan and date of consultation.

Dr. Sydney Jordan
WUR Library – Research Data Management Support
data@wur.nl
Date: 2025-05-22

B. Describe the data to be collected, software used, file formats and data size.

9. Will you use existing data for this project?

- Yes. Please specify below which data (e.g. DOI, URL, or storage location) and the terms of use (e.g. licence).

- European Environment Agency (EEA) report on soil health: This report will serve as the primary source for creating the initial soil health knowledge graph (KG).
- Existing ontologies and vocabularies:
  - AGROVOC (multilingual controlled vocabulary covering soil-related concepts).
  - GloSIS (Global Soil Information System Web Ontology).
  - ISO 11074 (Soil quality).
- External databases and literature:
  - Metadata from external repositories according to SoilWise project:
    - ESDAC (European Soil Data Centre);
    - The INSPIRE Geoportal soil domain related high-value datasets (soil, ortho-imagery, elevation, geology, natural risk zones, environmental monitoring facilities, agricultural and aquacultural facilities, area management / restriction / regulation zones & reporting units, land cover, land use, hydrography), as identified by EJP Soil project;
    - EEA central data repository;
    - national (soil) data and knowledge repositories that may not be accessible in the INSPIRE GeoPortal yet;
    - soil-related data & knowledge achievements of R&I projects including their interlinking with Cordis and OpenAire;
    - scientific knowledge and data repositories with R&I results such as DataVerse, EJP SOIL, BonaRes and ORCaSa;
    - Sentinel products in Copernicus Open Access Hub;
    - EC (European Commission) DestinE (Destination Earth);
    - IACS data;
    - Other data resources.

10. Will new data be produced?

- Yes.

- Soil health KG:
  - This will be the primary new data product and a key component of the SoilWise Repository (SWR), integrating structured information extracted from the EEA's report;
  - It will be represented in RDF format, using standardized ontologies (SKOS, Dublin Core, and further upper level ontolgies) and domain-specific ontologies and vocabularies (GloSIS, AGROVOC).
- Expanded soil knowledge repository:
  - An enhanced version of the soil health KG, interlinked with external data sources, literature, and web resources;
  - It will include metadata and keyphrases extracted from related scientific literature and datasets.
- Question-query-answer dataset:
  - A collection of natural language questions, their corresponding SPARQL queries, and answers derived from the soil health KG;
  - This dataset will be used for training and evaluating the question-answering system.

- Keyphrase extraction dataset:
  - A dataset of extracted and generated keyphrases from soil-related literature and metadata, along with their source texts.
- Recommendation system logs:
  - Anonymized logs of user interactions with the recommendation system, including queries and accessed resources.
- QA system interaction data:
  - Anonymized logs of user queries, system responses, and any feedback provided by users (e.g., farmers, land managers).
- Evaluation metrics and benchmarks:
  - Performance metrics for various components of the system (e.g., keyphrase extraction, query translation, recommendation accuracy);
  - Benchmark results comparing the developed methods against existing approaches.

11. Please describe the data you expect to generate and / or use in the table below. Include reused existing data as well (as these are files that you manage and store).

File contents	Data type	Software	(Open) file format	Estimated size of each file (range)	Estimated number of files (range)
EEA soil health report	Textual	PDF reader	PDF	~100 MB	1
Soil health KG	Structured data	RDF store (e.g., OpenLink Virtuoso)	RDF, Turtle	~100 MB	1
Metadata (data) from external databases	Structured data	Python, SQL database	JSON, XML, CSV, RDF	~1 KB	100k-1M
Expanded soil knowledge repository	Structured data	RDF store, Graph database	RDF, Turtle, TriG	~1 GB	1
Question-query-answer dataset	Structured data	SQL database	CSV, JSON	~10 MB	1-10
Keyphrase extraction dataset	Textual	Python	CSV, JSON	~50 MB	1-10
Recommendation system logs	Textual	SQL database	CSV, JSON	~10 MB	10-100
QA system interaction data	Textual	SQL database	CSV, JSON	~10 MB	10-100
Evaluation metrics and benchmarks	Numerical data	Python	CSV, JSON	~1 MB	10-50
Source code	Textual	Git	Various (.py, etc.)	~1 MB	100-1000
Project documentation	Textual	Git	MarkDown, .txt	~1 MB	5-10

12. Estimate how much data storage you require in total (e.g. by using the information in the table at question 11).

- 10-100 GB

C. Storage of data and data documentation / metadata during research

13. Where will the data,code and accompanying documentation / metadata be stored and

backed up during the research project (see the WUR Data Storage Finder)? Include platforms you use to share data, collect data on, or send data to for processing or analysis.

- WUR SharePoint / Teams - only when an up to date version of the research data is also safely stored on the W:drive or Yoda.
- Git@WUR (GitLab locally hosted at WUR)
- W:drive Enterprise File Storage (WUR network drive).

GitHub

D. Structuring your data and information

14. Give a (visual) representation of the folder structure you intend to use.

- phd_soilwise_project

  - raw_data
    - soil_health_report

  - metadata
    - harvested_metadata
    - generated_metadata

  - knowledge_graphs
    - soil_health_kg
      - initial_kg
      - expanded_kg

  - scripts
    - data_processing
    - kg_construction
    - nlp_scripts
      - keyphrase_extraction
      - question_answering
    - evaluation_scripts

  - processed_data
    - keyphrase_datasets

  - models
    - language_models
    - question_answering_models

- evaluation
  - metrics
  - benchmarks

- results
  - figures
  - tables
  - statistical_outputs

- documentation
  - data_dictionaries
  - methodology_docs
  - user_guides

- publications
  - manuscripts
  - presentations
  - posters

- project_management
  - timelines
  - meeting_notes
  - progress_reports

- external_resources
  - ontologies
  - vocabularies
  - reference_papers

15. Describe the file naming conventions you intend to use. Please give one or multiple example(s).

[subject_specifics]_[projectname]_[date]_[version].[extension]
The file name contains the project abbreviation so that it is clear to which project a file belongs. The date will be supplied in the format yyyymmdd to ensure proper sorting on date (i.e., 20250522) and conform to the international standard for using dates.
Example:
soil_health_KG_SHE_20240806_v01.ttl
NLQ_SHE_20260105_v02.json

16. How will you distinguish between versions of files (multiple answers possible)?

- The designation 'vRAW' is added to file names that contain raw unaltered data (before any processing and cleaning). Any alteration of RAW data is done on a copy of the RAW data and

appended with a version number which increases with each file modification (e.g. v01, v02, v03 etc.).
- Dates within file names are updated when files are modified.
- We will use Git versioning for code / scripts.

E. Data documentation and data quality

17. Describe below what data documentation and metadata will accompany the data to help make the data findable, understandable, and reproducible.

- The WUR codebook template (see template at https://doi.org/10.5281/zenodo.7701727).
- The WUR readme file template (see template at https://doi.org/10.5281/zenodo.7701727).

18. Describe what data and analysis quality controls will be used?

- Supervisors or peers will review the data and results for any anomalies (e.g. unexpected inconsistencies, outliers, correct labeling of data and / or treatments, correct and consistent coding applied, etc.).
- We will use standardised coding and terms of data throughout all experiments so that data descriptions are equal throughout various datasets created.
- We will use standard and validated protocols where appropriate.
- We will perform preliminary (pilot) experiments to validate intended experimental methods.

F. Working with sensitive data (personal data, ethics), data ownership, sharing and access

19. Who is the (rights)holder of the data (commonly known as the owner of the data)?

- WUR is the (rights)holder of the data.

20. What is the data classification for your project (for example as specified in SmartPIA) taking into account the (privacy) sensitivity of the data?

- Negligible.

21. Is this project registered in SmartPIA?

- Yes.

22. Please specify the (sensitive) data and privacy protection measures. Note that any measures undertaken should be consulted with the Information Security Officer (ISO) and Privacy Officer (PO).

- Data is classified as negligible and standard WUR security measures are undertaken.

23. Are there other ethical issues that need to be taken into account which may include approval from ethical committees?

- No.

24. Will there be any intellectual property (IP) rights or alternative applications or routes to impact (such as commercial interests) associated with the data?

- No.

G. Data archiving and publishing

25. Are there reasons to restrict access to the data or limit which data will be made publicly available?

- No.

26. Describe what data from question 11 will be archived internally (e.g. WUR network drive / Yoda@WUR) and not published, for a minimum of 10 years? Include the exact name for the storage medium chosen (see the WUR Data Storage Finder).

- Not applicable as data will be published.

27. What data will be published and made available for reuse via a data repository?

- Data underlying publications or reports. Please specify below which data listed in question 11.

All data will be published and made available for reuse via data repositories.

28. When will the data be available for reuse, and for how long will the data be available?

- Publication of data not underlying an article or report will be considered at the end of the project.
- Data will be available for at least 10 years upon completion of the project.
- Data will be available for at least 10 years as soon as the article or report is published and not required for any other article publication.

29. Which data repository do you intend to use to make the data findable and accessible (see the WUR Repository Finder)?

- Zenodo.

30. Which metadata standard will be used to describe the data during internal archiving and / or depositing in a data repository?

- Metadata standard from DANS Data Stations, 4TU.ResearchData and / or Zenodo (which often are the DublinCore or DataCite standard).

31. Which licence/terms of use will be applied to the data?

- Open access (Creative Commons Attribution licence (CC BY); anyone can access and reuse with attribution).

H. Data management costs

33. What resources (in time and / or money) will be dedicated to data management, data archiving or publication, and ensuring that data is reusable? Indicate as well how these costs will be covered.

- The PhD candidate and supervisor will spend at least 10% of their time on research data

management to approach the FAIR principles as much as possible.
- All costs for 10 year data storage and access management to that data after journal publication or report are covered by the research group / project.