https://www.ijecs.in
International Journal of Engineering and Computer Science
Volume 15, Issue 03, March 2026, Page No. 28310-28316
ISSN: 2319-7242 DOI: 10.18535/ijecs/v15i03.5447

GeoGraphRAG-Soil: An Open-Source Graph-Based Retrieval-Augmented Generation Framework for Automated Geospatial Soil Modelling

Dr Ranga Rao Velamala[1*]

Professor of Practice, Department of Computer Science and Engineering Visakha Institute of Engineering and Technology, Visakhapatnam ,Andhra Pradesh, India

Abstract:
Digital Soil Mapping (DSM) has advanced substantially over the past two decades, with predictive models becoming increasingly sophisticated. However, the methods and steps involved in constructing these models are not always fully documented and often rely on expert judgement, which can make modelling decisions difficult to reproduce or interpret. There is therefore a need to integrate DSM with modern technologies such as large language models (LLMs) to generate deeper insights and support more timely analytical outcomes. This paper introduces GeoGraphRAG Soil (GGRS), an open-source conceptual framework that automates and formalises DSM using graph-based retrieval-augmented generation and large language models.GGRS is built on a four-layer architecture: (1) a PostGIS spatial database for managing soil and environmental data; (2) a geospatial knowledge graph in ArangoDB encoding soil-environment relationships and geoprocessing operations; (3) a hybrid retrieval system in Milvus that identifies task-relevant subgraphs; and (4) a LangChain-based orchestration layer that translates natural language queries into executable geospatial workflows using open-source LLMs.GGRS advances DSM toward a more explicit, knowledge-driven approach that prioritises automation, interpretability, and methodological transparency by integrating structured domain knowledge with generative modelling techniques. The framework provides a cost-effective, open foundation to support soil health management in public sector initiatives and enables empirical evaluation in agricultural practice. GGRS can translate user queries into actionable, site-specific recommendations, such as identifying zones with low nitrogen and suggesting appropriate NPK fertilizer rates, predicting micronutrient deficiencies and recommending corrective crops, or assessing terrain and climate constraints to guide safe fertilizer application and crop selection. As an open-source reference architecture for researchers and policymakers, GGRS employs state-of-the-art technologies to support systematic workflow design and lays the foundation for future empirical validation and large-scale deployment by both public and private organisations.

Keywords: Digital Soil Mapping, Geospatial Workflow Automation, Retrieval Augmented Generation, Knowledge Graph, Open Source Geospatial Systems.

1. Introduction

Soil is a key natural resource that supports agricultural production, maintains ecosystem functions, and contributes to environmental sustainability. Detailed information on soil properties, such as nutrient content, organic carbon, pH etc., essential for precision agriculture, land management, and climate change mitigation (Wadoux, Molnar, & Lacoste, 2020). Traditional soil surveys rely on polygon-based maps, which are derived from field observations and expert interpretation. These maps are inherently static, costly to update, and limited in representing the continuous spatial variability of soils (Rossiter & Poggio, 2025). Digital Soil Mapping (DSM) addresses these limitations by integrating field observations with environmental covariates from digital elevation models, remote sensing, and climate data using statistical and machine-learning approaches (McBratney, Mendonca Santos, & Minasny, 2003). Advances in remote sensing, GIS, and predictive modeling have substantially improved resolution and accuracy, supporting applications from local farm management to national-scale soil inventories (Adeniyi, Bature, & Mearker, 2024; Hengl & MacMillan, 2019; Hussain et al., 2024). Designing effective DSM workflows poses substantial challenges. Covariate selection, model training, spatial prediction, uncertainty quantification, and post-processing are often insufficiently defined and primarily informed by expert judgment, custom scripts, and undocumented decisions, limiting reproducibility and regional transferability (Roger Bivand, 2021; Nussbaum et al., 2018). Recent developments in geospatial knowledge engineering and artificial intelligence provide mechanisms to formalize DSM workflows. Geospatial Knowledge Graphs (GKGs) explicitly link soil data, environmental variables, and modeling procedures, enabling users to query and compare alternative analytical pathways (Huang & Zhu, 2025). Retrieval-augmented generation (RAG) demonstrates how structured knowledge can guide complex, multi-step analyses (Lewis et al., 2020; Gao et al., 2024). These approaches support workflow systems that formalize expert knowledge through maintaining methodological adaptability within geosciences, which remains limited.

This paper presents GeoGraphRAG-Soil (GGRS), a knowledge-driven framework for DSM workflows. Integrating spatial databases, geospatial knowledge graphs, hybrid vector-graph retrieval, and generative components, GGRS enables reproducible and systematic workflow. It can translate user queries into actionable, site-specific recommendations, including identifying low-nitrogen zones, suggesting NPK fertilizer rates, predicting micronutrient deficiencies, or evaluating terrain and climate constraints to guide crop selection. GGRS provides a flexible, open, and structured foundation that supports reliable and adaptable soil information systems and structured integration of expert knowledge. To provide context for the scientific and technological foundation of GGRS, the following section reviews the centrality of soil, developments in DSM, and emerging knowledge-driven approaches for workflow automation.

2. Background and Related Work

Soil health plays a key role in global sustainability, and it supports multiple United Nations Sustainable Development Goals (SDGs), including zero hunger (SDG 2), clean water and sanitation (SDG 6), climate action (SDG 13), and life on land (SDG 15) (Lal et al., 2021). Effective soil stewardship improves food security, alleviates poverty, and strengthens environmental resilience. Accurate and accessible soil information is essential for evidence-based policy and land management decisions (Yin et al., 2022). National programs, such as India's Soil Health Card scheme (SHC,2026), highlight the demand for scalable and standardized soil assessment methods. Conventional soil surveys are based on manual field measurements and polygon-based mapping. These approaches are static, costly to maintain, and limited in capturing continuous spatial and temporal variability (Nikiforova et al., 2020; Rossiter & Poggio, 2025). Digital Soil Mapping (DSM) addresses these limitations by integrating field observations with environmental covariates derived from digital elevation models, remote sensing, and climate data through statistical and machine-learning techniques (McBratney, Mendonca Santos, & Minasny, 2003). Advances in remote sensing, computational resources, and predictive modeling have facilitated the development of high-resolution products, including SoilGrids and national digital soil mapping (DSM) initiatives, which support applications from farm-level management to national soil inventories (Hengl et al., 2017; Adeniyi, Bature, & Mearker, 2024; Hussain et al., 2024).

Despite these developments**,** DSM faces persistent challenges, such as model performance that depends on data quality, coverage, covariate selection, and modeling approach (Arrouays et al., 2020b). Despite the increasing complexity of predictive models, workflow processes such as covariate selection, model training, spatial prediction, uncertainty quantification, and post-processing remain insufficiently structured. Decisions are informed by expert knowledge, custom scripts, and undocumented practices, which constrain reproducibility, comparability, and transferability across regions (Nussbaum et al., 2018). Scripted pipelines enhance computational reproducibility but do not document the rationale behind methodological decisions, which is critical for audit and reuse.Recent advances in geospatial knowledge engineering and artificial intelligence provide mechanisms to formalize and automate complex DSM workflows. Large Language Models (LLMs) have been applied in scientific domains for tasks such as code generation, documentation, and advisory support (Yifan Yao et al., 2024; Kuska Ekukhat et al., 2024). Their use in soil science, however, remains limited due to incomplete domain knowledge, the risk of generating inaccurate outputs, and challenges in reasoning over structured relational data (Gong et al., 2024)

Retrieval-Augmented Generation (RAG) addresses these limitations by grounding LLM outputs in authoritative knowledge sources, improving factual accuracy and relevance (Lewis et al., 2020). Document-based retrieval may not fully capture the relational logic present in environmental and geospatial data. GraphRAG extends RAG through structured knowledge graphs as the retrieval source, enabling relation-aware reasoning and explainable outputs (Diamantini et al., 2024; Zhixin e.t al., 2025). Knowledge graphs represent entities (nodes) and their relationships (edges) in a machine-interpretable format, supporting advanced reasoning for workflow design. Geospatial Knowledge Graphs (GKGs) **represent** spatial entities and their interactions, as demonstrated by systems such as KnowWhereGraph. Existing workflow management platforms and provenance models offer structured support for scientific processes; however, a gap remains in applying graph-constrained RAG to automatically generate reproducible DSM workflows that integrate soil properties, environmental covariates, and geoprocessing rules in a structured, machine-interpretable format. Although DSM methods and predictive models progressed significantly, the formalization and automation of workflows remain underdeveloped. Existing practices, primarily guided by expert knowledge and frequently informal or poorly documented, limit reproducibility, transferability, and the systematic evaluation of alternative approaches. Graph-based AI techniques have not yet been fully applied to produce domain-specific, executable workflows. DSM workflows that represent pedological relationships and geospatial operations in a structured, reproducible manner are needed. There is a critical need for a framework that systematically translates expert knowledge and environmental data into reproducible, site-specific DSM workflows with transparency, flexibility, and machine interpretability.

Structured and reproducible DSM workflows that capture pedological relationships and geospatial operations are essential. Thus, a framework is needed to systematically translate expert knowledge and environmental data into site-specific workflows that are transparent, flexible, and machine-interpretable. Such a framework would support soil information, evidence-based agricultural and environmental management, and bridge the gap between predictive modeling and practical implementation. The GeoGraphRAG-Soil (GGRS) framework (Section 3) addresses this need by integrating geospatial data, expert knowledge, and generative modeling in a structured, reproducible approach.

3. Methodology

The GeoGraphRAG-Soil (GGRS) Conceptual Architecture

GeoGraphRAG-Soil (GGRS) is proposed as a technical architecture for knowledge-driven, graph-constrained generative Digital Soil Mapping (DSM) workflow design. The framework formalizes and automates DSM workflow construction, addressing limitations of traditional expert-driven approaches by integrating geospatial data management, knowledge graphs, hybrid retrieval, and constrained generative synthesis. GGRS is guided by four core principles: automation, interpretability, reproducibility, and open-source accessibility. Automation reduces manual workflow assembly, interpretability is achieved through structured knowledge graphs, reproducibility is ensured via formalized relationships and operations, and open-source accessibility leverages mature tools such as PostGIS, ArangoDB, Milvus, and LangChain with LLMs like LLaMA and Mistral. The architecture is conceptual, emphasizing system organization, knowledge flow, and interaction logic, rather than a single empirical implementation. GGRS provides a structured foundation for formalizing DSM domain knowledge, procedural rules, and workflow logic while supporting semi-automated synthesis of scientifically valid DSM workflow plans. These plans are intended for expert validation and execution.

3.1 Layered Architecture Overview

The GGRS framework consists of four interconnected layers (Figure 1 and Table 1), each building upon the previous to process data, represent knowledge, retrieve task-specific information, and generate executable workflows. The modular design supports efficient processing of complex geospatial soil modelling tasks, including the prediction of soil properties from environmental covariates.

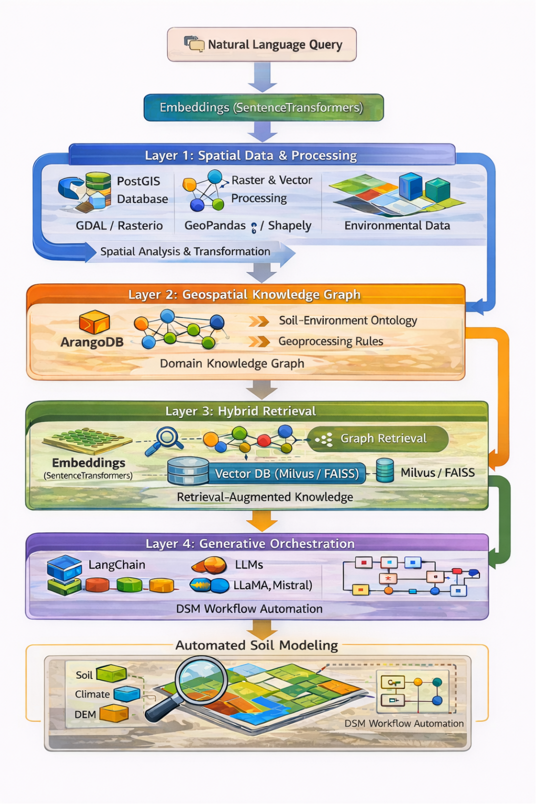

***Figure** 1. Architectural overview of GGRS illustrating the end-to-end workflow from spatial data processing to generative orchestration.*

Table 1. Four-Layer System Architecture for Semantically Constrained DSM Workflow Generation

Layer	Description	Key Technologies	Role in Workflow
1 Spatial Data & Processing	Stores and manages soil and environmental data, performs spatial analysis and transformations	PostGIS, GDAL/Rasterio, GeoPandas /Shapely	Provides the foundational data infrastructure and ensures efficient query and feature processing
2 Geospatial Knowledge Graph	Represents relationships between soil properties, environmental covariates, and geoprocessing operations	ArangoDB, Soil-Environment Ontologies, Geoprocessing Rules	Captures expert knowledge formally and constrains workflow synthesis using semantic and procedural relationships
3 Hybrid Vector-Graph Retrieval	Extracts task-relevant subgraphs using combined vector embeddings and graph traversal	Milvus/FAISS, SentenceTransformers	Identifies relevant nodes and edges for workflow synthesis, bridging knowledge representation and generative reasoning

Layer	Description	Key Technologies	Role in Workflow
4 Generative Orchestration	Produces executable DSM workflows from retrieved knowledge	LangChain, LLMs (LLaMA, Mistral)	Generates structured workflow specifications, including code, scripts, or configuration files, within subgraph-defined constraints

3.2 Layer 1: Spatial Data and Processing

This layer provides the geospatial data infrastructure required for DSM workflow execution (Arrouays et al., 2020b). Open-source tools such as PostGIS manage vector and raster data, GDAL/Rasterio supports raster processing, and GeoPandas/Shapely enable vector data manipulation. Environmental covariates, soil attributes, and spatial boundaries are stored and processed with indexing, feature derivation, and coordinate transformations (Warmerdam, 2008). Layer 1 of GGRS ensures that data is accessible, queryable, and ready for integration with upper layers.

3.3 Layer 2: Geospatial Knowledge Graph

The knowledge graph formalizes DSM expertise as a machine-readable semantic network (Shirvani-Mahdavi et al., 2025; Wang et al., 2025; Shimizu et al., 2025).Nodes represent soil properties, environmental variables, data sources, and processing operations, while edges represent relationships, dependencies, and constraints. Open-source graph databases (ArangoDB, Neo4j) and geospatial ontologies (Perry & Herring, 2012) are employed. Procedural knowledge, including operation order and method constraints, is captured, transforming tacit expert knowledge into explicit, reusable, and computable guidance. A subgraph for soil nitrogen modeling, for instance, might include Soil N,OC,NDVI, Slope, Zonal Statistics, and Random Forest, with edges representing their functional relationships.

3.4 Layer 3: Hybrid Vector-Graph Retrieval

Task-specific knowledge is retrieved using a combination of semantic embeddings and graph-based traversal (Siddharth & Luo, 2024). Natural-language queries are embedded via Sentence Transformers and matched against a vector-indexed knowledge database such as Milvus/FAISS (Reimers & Gurevych, 2019; Johnson, Douze, & Jegou, 2021). Graph traversal expands the context to include related nodes and edges, producing a subgraph that defines relevant variables, permissible operations, and logical dependencies for workflow synthesis.

3.5 Layer 4: Generative Orchestration

The orchestration layer interprets the retrieved subgraph and generates executable DSM workflows using open-source LLMs such as LangChain with LLaMA or Mistral (Wang, Yang, & Liu, 2025; Linders & Tomczak, 2025; Song et al., 2025). Graph constraints ensure that generation respects scientific and procedural rules. Outputs may include Python scripts, configuration files, or stepwise workflow instructions, which are ready for expert validation and execution.

3.6 End-to-End Workflow

The proposed methodology implements a sequential, knowledgegrounded pipeline comprising five functionally distinct stages. In the **Query Processing (Step 1)** phase, natural language user requests are transformed into semantic embeddings to facilitate subsequent retrieval processes. These embeddings inform the **Hybrid Retrieval (Step 2)** process, which integrates vector similarity metrics with structural graph traversal to identify the most relevant knowledge fragments. Retrieved fragments are then consolidated in **Subgraph Assembly (Step 3)** to construct a coherent taskspecific knowledge structure that constrains the **Constrained Synthesis (Step 4)** stage, ensuring the derivation of an ordered and executable workflow specification. In the final **Output Delivery (Step 5)** phase, workflow plans are generated in formats that are interpretable by both human researchers and automated execution systems, thereby supporting transparency, reproducibility, and implementation.

3.7 Architectural Contribution

GGRS conceptualizes DSM as a graph-constrained generative planning problem, integrating data infrastructure, knowledge representation, retrieval intelligence, and reasoning in a structured manner. By representing domain knowledge and constraining generative outputs to scientifically defined boundaries, the framework ensures reproducibility, interpretability, and modularity, addressing key limitations of traditional DSM, including expert dependency, lack of transparency, and limited scalability (Arrouays et al., 2020b; Nikiforova et al., 2020; Yin et al., 2022; Lal et al., 2021).

4. Discussion

The GeoGraphRAG Soil (GGRS) architecture represents a conceptual shift in Digital Soil Mapping (DSM), transforming workflows from expert-driven heuristic procedures into structured, query-driven, knowledge-constrained planning systems. While empirical validation is beyond the scope of this conceptual study, its potential can be illustrated using an assumed dataset representative of operational DSM. This dataset includes georeferenced soil laboratory measurements capturing macro and micro-nutrients (N, P, K, Fe, Mn, Zn, B, S); terrain derivatives from digital elevation models, including slope, aspect, and curvature; remotely sensed vegetation indices such as NDVI, EVI, and SAVI; land use and land cover data; climate variables; and lithology. These multidimensional covariates reflect standard practice in modern DSM, where soil-environment relationships are modeled through environmental correlation frameworks (McBratney et al., 2003; Minasny and McBratney, 2016). GGRS restructures the DSM workflow design rather than altering its scientific foundations. Conventional DSM relies heavily on tacit knowledge for covariate selection, preprocessing, and modeling sequences (Arrouays et al., 2020). In contrast, the GGRS framework organizes these decisions within a geospatial knowledge graph, representing soil-terrain, soil-vegetation, and soil-climate relationships as machine-interpretable entities and procedural rules.

Dependencies such as terrain derivation, normalization, feature stacking, and model training are captured explicitly, enhancing reproducibility, transparency, and interpretability (Lal et al., 2021). The framework harmonizes heterogeneous datasets and treats DSM as a retrieval-augmented reasoning task. Natural language queries trigger the extraction of task-specific subgraphs that link relevant covariates to compatible geoprocessing operations. This approach aligns with emerging paradigms that integrate structured knowledge with large language models to enable explainable, knowledge-intensive reasoning (Lewis et al., 2020). These features collectively set the stage for operationalizing GGRS workflows.

4.1 From Query to Actionable Insight

Given the assumed dataset, GGRS can interpret user queries and generate operational workflows for tasks ranging from targeted diagnostics to complex temporal analyses. Table 2 illustrates examples, demonstrating the framework's ability to move from static map consumption to dynamic, knowledge-guided workflow synthesis for decision support. These illustrative workflows highlight the feasibility of GGRS and provide a foundation for broader applications in diverse agricultural contexts.

Table 2: Illustrative User Queries and Corresponding GGRS-Synthesized Workflow Plans

User Query	GGRS-Synthesized Workflow Plan
Which areas exhibit low nitrogen, and what NPK fertilizer should be applied?	Identify spatial zones where soil nitrogen is below 0.1 percent. Evaluate slope and NDVI to flag erosion risk. Recommend site-specific NPK (20, 10, 10 kilograms per hectare) constrained by lithology and soil texture rules.
Where is zinc deficiency likely, and which crop could improve soil health?	Highlight regions with plant-available zinc below 0.5 milligrams per kilogram. Recommend legume cover crops, considering rainfall, slope, and leaching risk rules.
Can sulfur be safely applied in hilly areas with high rainfall?	Identify sulfur-deficient areas below 10 milligrams per kilogram and slopes below 15 degrees. Flag steeper slopes as high risk for leaching or runoff based on terrain and precipitation.
Which crop is optimal for a specific location?	Integrate soil fertility indices, terrain, NDVI, climate, and land use and land cover data. Generate crop suitability recommendations using knowledge graph models.
Model the change in topsoil organic carbon over the past five years using Landsat and climate data.	Preprocess multitemporal Landsat data to derive annual NDVI composites. Integrate spatiotemporal climate grids. Apply space-time geostatistical modeling such as STARFM and quantify prediction uncertainty.

4.2 Proposed Applications and Impact Pathways

GGRS offers transformative capabilities for managing diverse agricultural soils, from red sandy loams to salinity- or nutrient-affected black cotton soils, as shown in Table 2. Public organizations can automate soil health mapping initiatives such as the Soil Health Card program. Farmers and agronomists can receive expert-validated, tailored guidance through decision support tools. Researchers can extend GGRS for predictive modeling in drought-prone regions or integrate it with satellite-based remote sensing data, including Sentinel-2. A core innovation of GGRS is its use of large language models, which act as planners over retrieved knowledge graph substructures rather than independent sources, ensuring outputs are scientifically consistent, traceable, and explainable (Bubeck et al., 2023). Its transparent, adaptable, and open-source architecture facilitates adoption, supports long-term sustainability, and enables equitable access to digital soil information systems (Arrouays et al., 2020).

5. Limitations and Future Research

As a conceptual architecture, GGRS has limitations that define its current scope and guide future research toward operational use. Its effectiveness depends on the quality, completeness, and consensus of the Geospatial Knowledge Graph (GKG). Developing a coherent ontology, integrating diverse data, and encoding procedural rules requires substantial effort, and the GKG must evolve with new knowledge to prevent propagation of errors. Future work should explore semi-automated knowledge graph construction, collaborative expert curation, and versioning protocols to manage updates. GGRS is designed for standard digital soil mapping tasks, but its performance on complex, multi-step scenarios, such as combining real-time sensor data with long-term climate models, remains untested. Improving retrieval and reasoning capabilities, including subgraph assembly, workflow chaining, and meta-reasoning, is essential to handle such complexity. Although built with open-source components, GGRS requires expertise in geospatial databases, knowledge engineering, vector search, and LLM orchestration, which may limit accessibility for public agencies and extension services. User-centered interfaces, including natural-language chat systems and visual graph editors, are important to enable adoption and effective use. Large language models can produce errors, inconsistent reasoning, and variable outputs (Gong et al., 2024), which may affect reliability. Addressing these risks requires validation through automated consistency checks and systematic comparison with expert-designed workflows. Since GGRS automates workflow design rather than soil mapping, output accuracy depends on the quality, resolution, and coverage of input data. Empirical evaluation and uncertainty quantification are therefore critical to assess GGRS-generated workflows against expert workflows in terms of efficiency, reproducibility, methodological rigor, and overall accuracy (Minasny & McBratney, 2016).

6. Conclusions

This paper introduces GeoGraphRAG Soil (GGRS), a conceptual open-source framework designed to formalize and automate the synthesis of Digital Soil Mapping (DSM) workflows through graph-based retrieval-augmented generation. GGRS addresses a critical gap in current DSM practices: the reliance on informal, tacit knowledge in workflow design, which limits reproducibility, scalability, and methodological transparency. The proposed four-layer architecture integrates spatial data management, explicit knowledge representation in a geospatial knowledge graph, hybrid retrieval, and constrained LLM orchestration. GGRS provides a framework for transitioning DSM from an expert-dependent practice toward a structured, knowledge-driven, and automated approach. Workflow construction is formulated as a graph-constrained generative planning problem.

The framework explicitly identifies key challenges, including knowledge graph curation, scalable reasoning, and effective AI-human collaboration. Addressing these challenges is essential for the development of automated and reliable geospatial analysis systems. The potential applications of GGRS include public-sector soil health monitoring, agricultural extension, and methodological research. These limitations establish a clear research agenda. Addressing the identified challenges through community-driven knowledge engineering, human-centric interface design, and thorough validation of AI-assisted reasoning allows GGRS to provide a foundation for the development of transparent, reproducible, and accessible tools for soil science and sustainable soil health management.GGRS enables systematic access to soil health knowledge to support informed decision-making. As a cost-effective and open-source framework, GGRS can be employed by public and private organizations in developing and underdeveloped countries. These organizations can apply GGRS to gain actionable insights and implement LLM-based question-answer systems. This approach provides practical, technology-driven solutions for soil and health management, supporting both policymakers and farmers and contributing to the achievement of the United Nations Sustainable Development Goals.

References

1. Adeniyi, O. D., Bature, H., & Mearker, M. (2024). A systematic review on digital soil mapping approaches in lowland areas. Land, 13(3), 379. https://doi.org/ 10.3390/land13030379

2. Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., Richer-de-Forges, A. C., Morgan, C. L. S., Roudier, P., Poggio, L., & Mulder, V. L. (2020b). Impressions of digital soil maps: The good, the not so good, and making them ever better. Geoderma Regional, 20, e00255. https://doi.org/ 10.1016/j.geodrs.2020.e00255

3. Arrouays, D., Poggio, L., Salazar Guerrero, O. A., & Mulder, V. L. (2020a). Digital soil mapping and GlobalSoilMap. Main advances and ways forward. Geoderma Regional, 21, e00265. https://doi.org/ 10.1016/j.geodrs.2020.e00265

4. Bivand, R. S. (2021). Progress in the R ecosystem for representing and handling spatial data. Journal of Geographical Systems, 23(4), 515-546. https://doi.org/ 10.1007/s10109-020-00336-0

5. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv. https://arxiv.org/abs/2303.12712

6. Diamantini, C., Mele, A., Mircoli, A., Potena, D., Rossetti, C., & Storti, E. (2024). A Graph RAG approach to enhance explainability in dataset discovery. Data Science and Engineering. https://doi.org/ 10.1007/s41019-025-00313-x

7. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv. https://arxiv.org/abs/2312.10997

8. GDAL/OGR contributors. (2024). GDAL - Geospatial Data Abstraction Library [Software]. Open Source Geospatial Foundation. https://gdal.org

9. Gong, R., & Li, X. (2025). The application progress and research trends of knowledge graphs and large language models in agriculture. Computers and Electronics in Agriculture, 235, 110396. https://doi.org/ 10.1016/j.compag.2025.110396

10. Gu, Z., Long, T., Wang, S., Shang, X., Shen, W., Wei, X., & Xu, K. (2025). Construction of Q&A; methods based on knowledge graphs and large language models--improving the accuracy of landscape pest and disease Q&A.; Smart Agricultural Technology, 12, 101094. https://doi.org/ 10.1016/j.atech.2025.101094

11. Hengl, T., MacMillan, R. A., & OpenGeoHub foundation. (2019). Predictive soil mapping with R. OpenGeoHub foundation. http://www.soilmapper.org

12. Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE, 12(2), e0169748. https://doi.org/ 10.1371/journal.pone.0169748

13. Huang, W., & Zhu, R. (2025). Geospatial knowledge graphs. In J. P. Wilson (Ed.), The Geographic Information Science & Technology Body of Knowledge. University Consortium for Geographic Information Science. https://doi.org/ 10.22224/gistbok/2025.1.1

14. Hussain, S. S., Ganie, A., Dar, W. A., Wani, T., Hadayatullah, M., Baba, J., & Dar, R. (2024). The role of digital soil mapping in soil survey and agricultural planning. International Journal of Plant & Soil Science, 36(9), 438-449. https://doi.org/ 10.9734/ijpss/2024/v36i94993

15. Janowicz, K., Yan, B., Regalia, B., Zhu, R., & Mai, G. (2020). Deploying spatial data infrastructures for data science and semantic interoperability. In The Routledge Handbook of Geospatial Technology and Society (pp. 159-175). Routledge.

16. Johnson, J., Douze, M., & Jegou, H. (2021). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535-547. https://doi.org/ 10.1109/TBDATA.2019.2921572

17. Kasraei, B., Schmidt, M. G., Saurette, D. D., Bulmer, C. E., Zhang, J., Pennell, T., John, K., & Heung, B. (2024). Advancing digital soil mapping with multi-year crop cover data: Impacts on model accuracy and soil interpretation. Geoderma, 461, 117481. https://doi.org/ 10.1016/j.geoderma.2025.117481

18. Kuska, M. T., Wahabzada, M., & Paulus, S. (2024). AI for crop production - Where can large language models (LLMs) provide substantial value? Computers and Electronics in Agriculture, 221, 108924. https://doi.org/

10.1016/j.compag.2024.108924

19. Lal, R., Bouma, J., Brevik, E., Dawson, L., Field, D. J., Glaser, B., Hatano, R., Hartemink, A. E., Kosaki, T., Lascelles, B., Monger, C., Muggler, C., Ndzana, G. M., Norra, S., Pan, X., Paradelo, R., Reyes-Sanchez, L. B., Sanden, T., Singh, B. R., Spiegel, H., & Zhang, J. (2021). Soils and sustainable development goals of the United Nations: An International Union of Soil Sciences perspective. Geoderma Regional, 25, e00398. https://doi.org/ 10.1016/j.geodrs.2021.e00398

20. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459-9474). https://proceedings.neurips.cc/paper/2020/file/ 6b493230205f780e1bc26945df7481e5-Paper.pdf

21. Linders, J., & Tomczak, J. M. (2025). Knowledge graph extended retrieval augmented generation for question answering. Applied Intelligence, 55, Article 1102. https://doi.org/ 10.1007/s10489-025-06885-5

22. McBratney, A. B., Mendonca Santos, M. L., & Minasny, B. (2003). On digital soil mapping. Geoderma, 117(1-2), 3-52. https://doi.org/ 10.1016/S0016-7061(03)00223-4

23. McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 445, 51-56.

24. Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. Geoderma, 264, 301-311. https://doi.org/ 10.1016/j.geoderma.2015.07.017

25. Moore, I. D., Grayson, R. B., & Ladson, A. R. (1991). Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. Hydrological Processes, 5(1), 3-30. https://doi.org/ 10.1002/hyp.3360050103

26. Nikiforova, A. A., Fleis, M. E., Nyrtsov, M. V., Kazantsev, N. N., Kim, K. V., Belyonova, N. K., & Kim, J. K. (2020). Problems of modern soil mapping and ways to solve them. Catena, 195, 104885. https://doi.org/ 10.1016/j.catena.2020.104885

27. Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., & Papritz, A. (2018). Evaluation of digital soil mapping approaches with large sets of environmental covariates. SOIL, 4(1), 1-22. https://doi.org/ 10.5194/soil-4-1-2018

28. Perry, M., & Herring, J. (2012). OGC GeoSPARQL - A Geographic Query Language for RDF Data. Open Geospatial Consortium. (OGC 11-052r4)

29. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982-3992. https://doi.org/ 10.18653/v1/D19-1410

30. Rossiter, D. G., & Poggio, L. (2025). Representing soil landscapes from digital soil mapping products - helping the map to speak for itself. SOIL, 11, 849-881. https://doi.org/ 10.5194/soil-11-849-2025

31. Shimizu, C., Stephen, S., Barua, A., et al. (2025). The KnowWhereGraph ontology. Journal of Web Semantics, 84, 100842. https://doi.org/ 10.1016/j.websem.2024.100842

32. Shirvani-Mahdavi, N., Wingfield, D., Gutierrez, J. G., et al. (2025). A knowledge graph informing soil carbon modeling. LNCS. https://doi.org/ 10.1007/978-3-031-97207-2_18

33. Siddharth, L., & Luo, J. (2024). Retrieval augmented generation using engineering design knowledge. Knowledge-Based Systems, 303, 112410. https://doi.org/ 10.1016/j.knosys.2024.112410

34. Soil Health Card Scheme [SHC]. (2026). Soil health scheme. Government of India. https://www.soilhealth.dac.gov.in/aboutus

35. Song, J., et al. (2025). Graph retrieval augmented large language models for improved diagnostic decision making. npj Digital Medicine, 8, Article 19. https://doi.org/ 10.1038/s41746-025-01955-x

36. Wadoux, A. M. J.-C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, 210, 103359. https://doi.org/ 10.1016/j.earscirev.2020.103359

37. Wang, B., Moreira de Sousa, L., & Fensel, A. (2025). soilwise-he/soil-health-knowledge-graph v0.2.4 Zenodo. https://doi.org/ 10.5281/zenodo.15600197

38. Wang, S., Yang, H., & Liu, W. (2025). Research on the construction and application of retrieval enhanced generation (RAG) model based on knowledge graph. Scientific Reports, 15, Article 40425. https://doi.org/ 10.1038/s41598-025-21222-z

39. Warmerdam, F. (2008). The Geospatial Data Abstraction Library. In M. B. Hall & G. M. Leahy (Eds.), Open Source Approaches in Spatial Data Handling (pp. 87-104). Springer. https://doi.org/ 10.1007/978-3-540-74831-1_5

40. Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. High-Confidence Computing, 4(2), 100211. https://doi.org/ 10.1016/j.hcc.2024.100211

41. Yin, C., Zhao, W., & Pereira, P. (2022). Soil conservation service underpins sustainable development goals. Global Ecology and Conservation, 33, e01974. https://doi.org/ 10.1016/j.gecco.2021.e01974