# Towards a FAIR Knowledge Graph-Based Conversational Agent for Exploring Agroforestry Research[*]

Joseph Kivisi[1,*,†], Sufiet Erlita[1,†] ,Thomas Zschocke[1,†]

[1] *Center for International Forestry Research (CIFOR) – World Agroforestry (ICRAF)*

**Abstract**

Recent work on FAIR knowledge graphs has advanced the representation of scientific work in a structured and semantic way by rendering the information as readable by both machines and humans. Paired with a conversational agent, the discovery of this information can be further enhanced by integrating diverse data and large language model techniques. The proposed FAIR knowledge graph aggregates metadata about digital assets and research outputs harvested from textual and structured data sources in repositories of two international agricultural research institutions of CGIAR, based in Indonesia and Kenya, respectively. We document a preliminary framework on making these research outputs available through a FAIR knowledge graph and increasing their accessibility through a conversational agent while enhancing the visibility of this scientific work in Africa, Asia and Latin America more globally.

**Keywords**

FAIR data, knowledge graph, NLQ, LLM, graph RAG, conversational agent

## 1. Introduction

Knowledge graphs (KG) provide an interconnected representation of the current state of domain-specific knowledge while offering the options for extensions to answer natural language questions [1]. Further, knowledge graphs are also well suited to provide the infrastructure for managing FAIR scientific information, such as the Open Research Knowledge Graph [2]. This becomes even more relevant considering the limitations countries in the Global South are facing in accessing scientific information. International bodies such as the African Union (AU) want to address this issue by promoting data sharing to ensure that data are accessible to African researchers [3] while mainstreaming artificial intelligence (AI) in high-impact areas such as agriculture and climate change [4]. The Center for International Forestry Research (CIFOR) and World Agroforestry (ICRAF), two international agricultural research institutions and members of CGIAR, based in Indonesia and Kenya, respectively, work closely with local partners in Africa, Asia and Latin America to make their research outputs available as open access [5] while observing the FAIR data principles, that is, Findability, Accessibility, Interoperability, and Reusability [6].

However, this scientific information is more often than not siloed in institutional repositories in unstructured formats and is not interlinked, thus preventing the surfacing of new insights from research. Traditional data retrieval methods such as sequential search and index-based retrieval often fail when handling intricate and interconnected data structures, resulting in incomplete or misleading outputs. To address these constraints, CIFOR and ICRAF are working towards establishing a FAIR knowledge graph to facilitate the availability and accessibility of their scientific information esp. on agroforestry. Based on similar research on constructing agricultural knowledge graphs [7], [8] and sharing FAIR agricultural research information [9], we are building a corpus of domain-specific scientific information with machine-readable, structured research outputs to allow

[*] *SEMANTiCS'25: 21st International Conference on Semantic Systems, September 3-5, 2025, Vienna, Austria*

[1*] Corresponding author.
[†] These authors contributed equally.
j.kivisi@cifor-icraf.org (J. Kivisi); s.erlita@cifor-icraf.org (S. Erlita); t.zschocke@cifor-icraf.org (T. Zschocke)
0009-0003-5629-8352 (J. Kivisi); 0000-0002-2181-6274 (S. Erlita); 0000-0001-8367-1915 (T. Zschocke)

researchers esp. in Africa, Asia and Latin America to search for relevant scientific work or get an overview of the research in the field of international agroforestry. At the same time, we are leveraging large language models (LLMs) paired with Natural Language Querying (NLQ) to provide for a more intuitive information retrieval process.

In this paper, we present a framework for constructing a FAIR knowledge graph over the outputs in agroforestry research based on a local LLM, and integrating a conversational agent to enable NLQ and perform query-focused summarizations (QFS) over the data.

# 2. Methodology

Large language models (LLMs), like Claude by Anthropic or the GPT series by OpenAI, are generative pretrained transformers (GPTs) that are capable to understand and generate text like a human for performing a wide range of tasks, such as, generate content, summarize text, or answer questions. However, LLMs have limitations handling knowledge-intensive tasks especially when responding to questions requiring specialized expertise [10]. While retrieval-augmented generation (RAG) helps to adapt LLMs for specific domains by leveraging external knowledge from text corpora, it faces challenges when processing text corpora collected from diverse sources that vary in accuracy and completeness [11]. As we face a similar situation with the data sources on agroforestry, we employ a graph-based RAG approach to traverse the entirety of a large text corpus consisting of journal articles, research papers, technical reports, etc. for improving the capability on the contextual comprehension of the LLM, which employs a two-step approach by indexing the data from the source documents to create an LLM-derived KG and then utilizing the pre-built indeces to enhance the retrieval process [12]. We envision a solution for generating evidence-based responses on agroforestry similar to graph RAG-based approaches in the medical field [13] or in customer services [14].

## 2.1. Constructing the FAIR knowledge graph

We envision a knowledge graph question answering (KGQA) system [15] on agroforestry that provides factual answers to natural language questions by leveraging knowledge graphs (KG). In general, KGs encode domain knowledge as a network of nodes and edges, with the nodes representing real-world entities and the edges representing relationships between the entities [16].

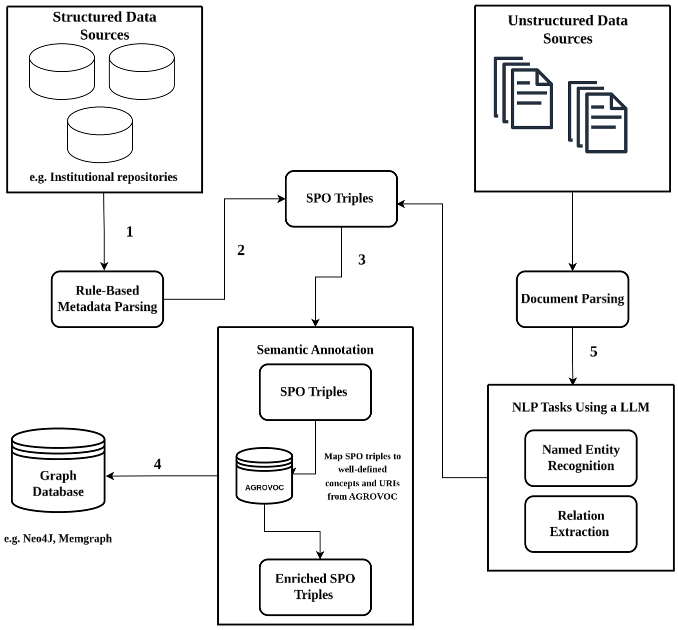


**Figure 1:** Overview of the Knowledge Graph Creation Workflow

In order to ensure FAIRness of the KG components (nodes, edges, extracted knowledge), we use a FAIRitization framework [17]. We adapted the overall process model for KG generation ranging from knowledge acquisition and hosting to knowledge curation and deployment [18] to develop the KGQA system in order to create a FAIR knowledge graph as follows (see Figure 1) :

1. Extract metadata from the data sources (e.g., bibliographic records in the institutional repository).
2. Use rule-based parsing to generate the nodes and edges of the knowledge graph as Subject-Predicate-Object (SPO) triples, that is, the head node (subject [e.g., researcher-001; publication-001; donor-001]), the edge (predicate [e.g., type; authoredBy; fundedBy]), and the tail node (object [e.g., Researcher; Publication; Donor]), as a declarative approach by defining a target set of entity types for extraction from the system along with the implicit relationships binding them.
3. Annotate the triples semantically by mapping the SPO triples to well-defined concepts and URIs from AGROVOC, a multilingual and controlled vocabulary covering concepts and terminology in the agricultural and related domains, maintained by the Food and Agriculture Organization of the UN [19].
4. Store the triples in the graph database.
5. Run Named Entity Recognition (NER) and relationship extraction from the document, while restricting the types of entities identified to meet the needs of agroforestry research domain.
6. Together with the relationships extracted, generate subject predicate object triples by the LLM.
7. Enrich the triples with semantic annotations with concepts from AGROVOC.

## 2.2. Building the conversational agent

As users benefit from a more intuitive way to interact with and search over the data in natural language, we employ a conversational agent to provide users with relevant information related to agroforestry. A conversational agent is a system designed for conversations with human users in natural language, either more informally as a chatbot, or more specific to user queries as a task-oriented agent [20], [21].

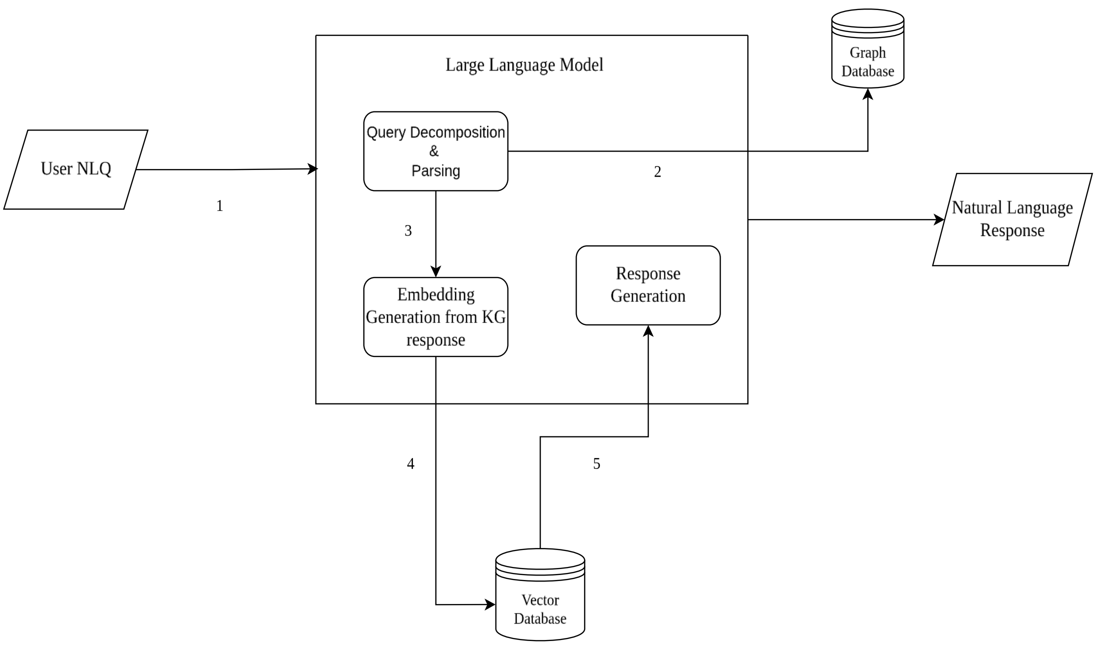


**Figure 2:** Overview of the conversational agent workflow

As illustrated in Figure 2, the (task-oriented) conversational agent generates responses on agroforestry based on a multi-step process:

1. Parse the NLQ using the LLM to extract concepts (e.g., what is agroforestry).
2. Query the knowledge graph based on the extracted concepts and performing relevance expansion to pull connected concepts (e.g., agroforestry applied in Kenya).
3. Enrich the NLQ and pass it back as input to the LLM.
4. Conduct a semantic search over the vector database to surface relevant documents using the enriched query.
5. Merge the relevant documents and original query as additional context to be used by the LLM to craft an appropriate response to the user (e.g., summary about agroforestry in Kenya together with a list of documents).

## 3. Conclusion

The framework to organize and query over domain-specific data in a KGQA system helps to enhance the discovery of implicit knowledge hidden within the published text resources. Layering the conversational agent on top of the knowledge graph simplifies the querying process. Enabling users to interact with the data in natural language helps them find information faster. As the framework is still under active development, we are in the process of building a proof-of-concept to validate the effectiveness of our approach as well as to evaluate the accuracy of knowledge extracted and the relevance and informativeness of the responses by involving subject matter experts and employing benchmark tests such as Needle in the Haystack (NIAH).[22]. Currently, the system is still restricted to organizing and managing textual knowledge on agroforestry in a structured representation. With the increasing availability of audiovisual resources collected by scientists in the field, we intend to expand the system towards a multimodal knowledge graph [23] by incorporating pictures, video and audio as data sources. All components, including scripts, FAIR-compliant datasets and guidelines will be hosted on a publicly accessible repository (Github) in line with the institutional open access policy [5].

## Declaration on generative AI

The author(s) have not employed any generative AI tools.

## References


[1] S. Auer, A. Oelen, M. Haris, et al., Improving access to scientific literature with knowledge graphs. Bibliothek 44 (2020), 516-529. doi: 10.1515/bfp-2020-2042.

[2] M. Stocker, A. Oelen, M. Y. Jaradeh, et al., FAIR scientific information with the Open Research Knowledge Graph. FAIR Connect 1 (2023), 19-21. doi: 10.3233/FC-221513.

[3] African Union (AU), AU Data Policy Framework. AU, Addis Ababa, Ethiopia, 2022. URL: https://au.int/en/documents/20220728/au-data-policy-framework.

[4] African Union (AU), Continental Artificial Intelligence Strategy. AU, Addis Ababa, Ethiopia, 2024. URL: https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy.

[5] CGIAR System Management Office, CGIAR Open and FAIR Data Assets Policy. CGIAR System Management Office, Montpellier, France, 2021. URL: https://hdl.handle.net/10568/113623.

[6] M. Wilkinson, M. Dumontier, I. Aalbersberg et al., The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3 (2016) 160018. doi: 10.1038/sdata.2016.18.

[7] H. Qin, Y. Yao, Agriculture knowledge graph construction and application. Volume 1756 of Journal of Physics: Conference Series (2021) 012010. doi: 10.1088/1742-6596/1756/1/012010

[8] C. Roussey, C. Guéret, M.-A. Laporte, Editorial: Knowledge graph technologies: the next Frontier of the food, agriculture, and water domains. Front. Artif. Intell. (2023) 6:1319844. doi: 10.3389/frai.2023.1319844

[9] J. D'Souza, Agriculture name entity recognition—Towards FAIR, resuable scholarly contributions in agriculture, Knowledge 4 (2024) 1-16. doi: 0.3390/knowledge4010001.

[10] J. Z. Pan, S. Razniewski, J.-C. Kalo, et al., Large language models and knowledge graphs: Opportunities and challenges, Transactions on Graph Data and Knowledge (TGDK) 1 (2023) 2:1-2:38. doi: 10.4230/TGDK.1.1.2.
[11] Z. Xiang, C. Wu, Q. Zhang, S. Chen, Z. Hong, X. Huang, J. Su, When to use graphs in RAG: A comprehensive analysis for graph retrieval-augmented generation. arXiv (Cornell University), 2024 (v2). doi: 10.48550/arXiv.2506.05690.
[12] D. Edge, H. Trinh, N. Cheng, et al., From local to global: A graph RAG approach to query-focused summarization. arXiv (Cornell University), 2025 (v2). doi: 10.48550/arXiv.2404.16130.
[13] J. Wu, J. Zhu, Y. Qi, J. Chen, M. Xu, F. Menolascina, V. Grau, Medical graph RAG: Towards safe medical large language model via graph retrieval-augmented generation. arXiv (Cornell University), 2024 (v2). doi: 10.48550/arXiv.2408.04187.
[14] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, Z. Li, Retrieval augmented generation with knowledge graphs for customer service question answering. arXiv (Cornell University), 2024 (v2). Doi: 10.1145/3626772.3661370.
[15] D. Banerjee, N. Hu, Y. Tan, D. Min, Y. Wu, R. Usbeck, G. Qi, Knowledge graph question answering and large language models. in: P. Hitzler, A. Dalal, M. S. Mahdavinejad, S. Saki Norouzi (Eds.), Handbook on Neurosymbolic AI and Knowledge Graphs, IOS Press, Amsterdam, Netherlands, 2025, pp. 466 – 531. doi: 10.3233/FAIA250220
[16] A. Hogan, C. Gutierrez, M. Cochez, et al., Knowledge graphs. Springer, Cham, Switzerland, 2022. doi: 10.1007/978-3-031-01918-0.
[17] D. Welter, N. Juty, P. Rocca-Serra, et al., FAIR in action - a flexible framework to guide FAIRification. Scientific Data 10 (2023) 291. doi: 10.1038/s41597-023-02167-2.
[18] D. Fensel, U. Şimşek, K. Angele, et al., Knowledge graphs: Methodology, tools and selected use cases. Springer, Cham, Switzerland, 2020
[19] Food and Agriculture Organization of the United Nations (FAO), AGROVOC: Semantic data interoperability on food and agriculture. FAO, Rome, Italy, 2021. doi: 10.4060/cb2838en.
[20] M. Wahde, M. Virgolin, Conversational agents: Theory and applications, in: P. P. Angelov (Ed.), Handbook on Computer Learning and Intelligence, Volume 1: Explainable AI and Supervised Learning, World Scientific, Singapore, 2022, pp. 497-544. doi: 10.1142/9789811247323_0012.
[21] S. Schöbel, A. Schmitt, D. Benner, M. Saqr, A. Janson, J. M. Leimeister, Charting the evolution and future of conversational agents: A research agenda along five waves and new frontiers. Information Systems Frontiers 26 (2024) 729–754. doi: 10.1007/s10796-023-10375-9.
[22] E. Nelson, G. Kollias, P.Das, S. Chaudhury, S. Dan, Needle in the Haystack for Memory Based Large Language Models. arXiv, doi: 10.48550/arXiv.2407.01437.
[23] Y. Chen, X. Ge, S. Yang, L. Hu, J. Li, J. Zhang, A survey on multimodal knowledge graphs: Construction, completion and applications, Mathematics 11 (2023) 1815. doi: 10.3390/math11081815.