A GraphRAG system for reliable Question Answering on Forestry-Related Topics

Felipe Vargas-Rojas[1,*,†], Tsiory Fanomezana Randriamisaina[1,†], Damien Arvor[2] and Isabelle Mougenot[1,†]

[1] *ESPACE-DEV, IRD, Univ. de Montpellier, Univ. Antilles, Univ. de Guyane, Univ. de la Réunion, Montpellier, France*
[2] *Université Rennes 2, UMR LETG CNRS 6554, Rennes, France*

Abstract

Forest policy and decision-making require timely and reliable access to large amounts of heterogeneous knowledge, including technical definitions, evaluation criteria, geographic data, and scientific references. However, this information is often dispersed across unstructured documents, making it difficult for computer systems to exploit directly. Lund (2018)—an official report produced by forestry institutions that consolidates numerous forest-related concepts while accounting for country-level differences—provides multiple definitions of forest, deforestation, afforestation, reforestation, and related terms. Based on this report, which comprises more than 200 pages of unstructured information, we develop a knowledge model using SKOS and complementary vocabularies to construct a knowledge graph. The resulting graph contains more than 3,000 definitions represented by 34,599 RDF triples. To demonstrate the usefulness of this knowledge graph, we developed a GraphRAG system that answers questions about forest-related definitions with greater precision and reliability.

Keywords

Forest definitions, unstructured data, heterogenous data, Knowledge Graph construction, Retrieval Augmented Generation (RAG)

1. Introduction

Providing a comprehensive and dynamically updated map of the world's forests is essential for numerous institutions, countries and policy makers. There are initiatives in this area, such as those led in Europe by the European Union and the JRC (Joint Research Centre), including GFC 2020 (a map of global forest cover for the year 2020 [1]) and GFT 2020 (Global Map of Forest Types for the year 2020), which form the basis of the European Observatory on Deforestation and Forest Degradation (EU Observatory on Deforestation and Forest Degradation[1]). These open-access maps of global forest cover, with a spatial resolution of 10 metres, can also support the implementation of the European Union Regulation on Deforestation (EUDR). The EUDR, which has not yet been implemented, will seek to ban the import into the EU of products (such as wheat or soya) that have contributed to deforestation or forest degradation after 31 December 2020. However, it seems equally essential: (i) to ensure that these maps are not subject to any dispute, (ii) to agree on a definition of what constitutes a forest and on what constitutes deforestation or reforestation. These maps are, in fact, based on a definition of "forest" that varies depending on the organizations responsible for the mapping. In the case of GFC 2020, the FAO's definition prevails; however, there are dozens of definitions of 'forest'. Depending on this definition and related terms (reforestation, deforestation, afforestation, etc.), the resulting maps may differ significantly.

The aim here is, therefore, to develop a knowledge graph that focuses on the definitions of forestry concepts rather than the concepts themselves, as is usually the case in related onto-terminologies

RAGE-KG 2026: The Third International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs, co-located with ISWC 2026, October 25–26, 2026, Bari, Italy

*Corresponding author.
†These authors contributed equally.
felipe.vargas-rojas@ird.fr (F. Vargas-Rojas); t.randriamisaina@ird.fr (T. F. Randriamisaina); damien.arvor@univ-rennes2.fr (D. Arvor); isabelle.mougenot@umontpellier.fr (I. Mougenot)
https://cv.hal.science/luis-felipe-vargas-rojas (F. Vargas-Rojas)
0000-0002-3530-7964 (F. Vargas-Rojas); 0000-0002-3017-9625 (D. Arvor); 0000-0002-1287-3269 (I. Mougenot)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1] see https://forest-observatory.ec.europa.eu/forest/

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-04

CEUR-WS.org/Vol-4312/paper41.pdf

(including ENVO [2], EUNIS Habitats [3] and AGROVOC [4]). This study aims to organise the conceptual pluralism of forest-related concepts, and focuses primarily on the definitions used to assess forest cover based on Earth observation, and in particular satellite imagery. The ultimate aim is to be able to compare maps produced using different definitions in order to achieve greater consensus and reliability in the assessment of forest resources. In this initial study, we present the key principles underpinning the construction of the knowledge graph, which utilises the SKOS standard [5] in a unique way, as we construct a SKOS concept for each definition of 'forest' and align these different definitions with one another. We then assess the relevance of the constructed vocabulary using a GrapRAG approach, based on a set of expert questions that demonstrate the added value of this alignment of definitions. Our main contributions are: (i) a knowledge model based on SKOS to structure definitions, their metadata, and the notion of "many definitions for one concept"; (ii) a curated knowledge graph of forest-related definitions that uses the knowledge model; and finally, (iii) a GraphRAG system that enables the exploitation and the access to the knowledge graph data.

2. Materials & Methods

2.1. Preliminaries

The Semantic Web [6] extends the traditional Web and aims to support a Web of data in which data are Findable, Accessible, Interoperable, Reusable (FAIR) [7]. The data are not only interpretable by humans but also by machines, thereby facilitating automation and interoperability. The W3C consortium has played a pivotal role in establishing several standards that underpin the evolution and adoption of the Semantic Web . We describe those that are most relevant to this work.

The Resource Description Framework (RDF) [8] is a standard for modelling data, in which information is represented as triples (subject, predicate, object). RDF includes Uniform Resource Identifier (URI) nodes and blank nodes. This model forms a directed graph, in which a triple represents an edge between two nodes. As an example, the following triples associate a concept with its definition:

```
:Forest  rdf:type  skos:Concept.
:Forest  skos:definition  "Plant community dominated by trees...".
```

Although multiple serialization formats are available, we primarily use the Turtle format (.ttl), which is human-readable and supported by many RDF libraries. Whilst "Knowledge Graph" is a broad concept, in this document it refers specifically to an RDF graph.

The Simple Knowledge Organization System (SKOS) [5] is a widely adopted standard for representing thesauri, classification schemes, and other controlled vocabularies. In Table 1, we present the properties that constitute our knowledge graph.

Property	Description
`skos:prefLabel`	designates the preferred label of a concept, ex. : « Forest ».
`skos:definition`	provides a textual definition describing a concept.
`skos:note`	allows supplementary annotations to be associated with a given concept.
`skos:broadMatch`	relates a concept to a more general concept generally in another concept scheme through a broader mapping relationship.
`skos:member`	relates a concept to a collection, indicating that the concept is a member of that collection.
`skos:related`	relates two concepts through an associative semantic relationship that is not hierarchical. For example, Forest and Deforestation are related concepts.

Table 1: Common SKOS properties associated to the classes skos:Concept and skos:Collection

SKOS constitutes the core of our knowledge model, as it is well suited to the requirements of our domain, which involves definitions, concepts, categories, hierarchies, and semantic relationships. It

provides sufficient expressiveness for representing this knowledge without introducing the additional ontological complexity associated with languages such as the Web Ontology Language (OWL) or RDF Schema (RDFS). Furthermore, SKOS is fully compatible with SPARQL, enabling efficient querying and exploration of the knowledge base.

2.2. Data sources

The main source for this work is a digital human-readable document published in Lund (2018) [9]. This document compiles forest-related definitions issued by various countries and international organisations. It is worth noting that the Food and Agriculture Organization of the United Nations (FAO) [10] and the United Nations Framework Convention on Climate Change (UNFCCC) [11] provide related definitions of forests, primarily considering them as collections of trees. This perspective may appear reductive from an ecosystem-based point of view. These definitions were published in response to the need for greater clarity and consistency in the terminology used in climate change reports, particularly within the framework of the UNFCCC. The document consists of 212 pages and follows a conventional structure composed of sections, headings, paragraphs, tables, and other textual elements. Although it is human-readable, its content is not directly machine-processable without further structuring and semantic annotation. A substantial effort has been undertaken to transform and structure this document into a Knowledge Graph (KG), as described in detail in the pipeline section 2.4.

2.3. Knowledge Model

The model is designed incrementally. First, basic notions are addressed using simpler SKOS terms: forest definitions are represented as *skos:Concept* instances with unique URIs, and their textual descriptions are added using the *skos:definition* property. Furthermore, since definitions are linked to their corresponding sources, we use the *dct:source* property —from the Dublin Core Vocabulary [12]— to represent this information, thereby extending the knowledge model beyond the SKOS vocabulary.

Categories. Definitions can be enriched with contextual information, such as the country of application or the domain perspective from which they are defined, including ecology, land use, cartography, and other relevant fields. To represent this contextual information, we use *skos:Collection*, a collection groups all the possible values of a specific category. Consequently, a definition can be linked to multiple categories, enabling a multidimensional contextual representation.

Concept hierarchies. A forest definition is linked to a top concept using *skos:broadMatch*. Within this model, we distinguish between two main groups of concepts: those associated with UNFCCC definitions and those derived from FAO definitions. For concepts that cannot be linked to an existing referential top concept, we introduce a local concept, *ex:Forest*. This concept acts as a generic reference point, representing the general notion of a forest shared across all definitions. Referential definitions can also be updated over time. The proposed model enables the introduction of new top concepts linked to specific dates, thereby supporting the representation of evolving reference definitions (e.g., UNFCCC 2006 and UNFCCC 2018).

Most common applications of SKOS assume that a concept is associated with a single definition. However, our use case is more challenging, as hundreds of definitions exist to describe what constitutes a forest. Our modelling decisions aim to capture and represent this inherent complexity.

The hierarchical organisation of concepts provides an initial strategy for structuring these definitions, but the model should evolve towards more fine-grained relationships between forest definitions through increasingly automated processes. Currently, we infer top concepts based on specific patterns; however, more sophisticated approaches are required to automatically identify and characterise relationships among definitions. Figure 1 presents a UML diagram of the proposed knowledge model.

2.4. Knowledge Graph construction pipeline

We adopt a strategy of generating the graph from mapping rules and regular expressions. Although LLMs could be used for this task, we preferred a more controlled and reliable approach, which favours

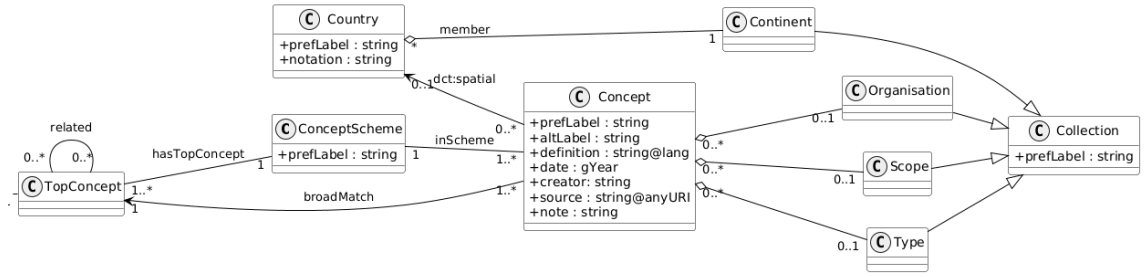


Figure 1: UML class diagram of the knowledge model for forest definitions

3. (Albania 2006 KP) *Cover Min. Area (ha) = 0.1 Min. Crown Cover (%) = 30 Min. Tree Height (m) = 3 Min. Strip Width (m) = ns* http://cdm.unfccc.int/DNA/index.html *and* https://cdm.unfccc.int/DNA/bak/allCountriesARInfos.html
4. (Albania 2007 UNFCCC) An area with 30% minimum tree crown cover, minimum 0.1 ha land area value, and 3m of tree potential height. (Fida, 2007) http://www.environmental-auditing.org/LinkClick.aspx?fileticket=fbx%2FbAIVeFU%3D&tabid=128&mid=568
5. (Albania 2008). 2 Forest is defined as an area covered at least 30% by dense trees on more than one tenth of a hectare [Law no. 9385 "On Forests and Forest Service]. 3 Open forest or forest land is defined as an area covered by 5-30% by forest vegetation, unregistered in another land use cadastre, [Law no. 9385 "On Forests and Forest Service]. 4. Forest is defined as dense woodland with a density of trees and shrubs not less than 30%. 5. Forest land

(a) from the document Lund (2018) [9]

```
ex:Forest_620 a skos:Concept ;
    dct:date "2007"^^xsd:gYear ;
    dct:source "http://www.environmental-auditing.org/LinkClick.aspx?fileticket=fbx%2FbAIVeFU%3D&
        tabid=128&mid=568"^^xsd:anyURI ;
    dct:spatial ex:Country_ALB ;
    skos:broadMatch ex:Forest_500 ; # UNFCCC concept
    skos:definition "An area with 30% minimum tree crown cover, minimum 0.1 ha land area value, and 3m
        of tree potential height."@en ;
    skos:inScheme ex:ForestScheme ;
    skos:note "Albania 2007 UNFCCC " ;
    skos:prefLabel "FOREST/FOREST LAND"@en .
```

(b) Converted into RDF Turtle

Figure 2: Example extracted from the official document, illustrating how a forest definition is transformed into RDF statements to construct the Knowledge Graph.

KG quality. Introducing ambiguity into the graph construction process can be counterproductive. Here, we detail the different pipeline steps.

Data extraction: The document does not follow a uniform structure; therefore, the first step consisted of analysing recurring patterns within the text. For instance, a definition may be introduced by a colon or formatted in bold. Similarly, the source indicating the original document from which a definition was extracted follows specific patterns, such as being enclosed in parentheses.

We developed a script to parse the document and extract its sections and paragraphs. The different functions process each paragraph using regular expressions (regex) to identify relevant information, including terms, their definitions, their sources, and additional contextual data such as countries or categories. These steps are iterative, and a human validator ensures that the output graph is correct and complete. Figure 2 presents a concrete example of how unstructured information is transformed and mapped into the knowledge graph.

Data organization and RDF serialization: In this step, the extracted information is structured and enriched before integration into the KG. Countries are identified and linked to the corresponding concepts to represent geographical context. Extracted concepts are also mapped to top-level references where possible, including referential concepts from organisations such as UNFCCC or FAO. Manual curation steps are performed to correct extraction errors and remove irrelevant content introduced

during the parsing process. Finally, the resulting structure is serialized in RDF Turtle format.

2.5. RAG Architecture

Retrieval-Augmented Generation (RAG) [13] has emerged as an architecture that enriches the responses of Large Language Models (LLMs) with information retrieved from external knowledge sources, typically unseen during the training phase. RAG models aim to improve the reliability, contextual relevance, and traceability of responses by augmenting LLMs with retrieved external knowledge. The workflow of a conventional RAG system can be described as follows:

1. A user's query is encoded as a vector in a semantic embedding space.
2. The top-k most relevant documents are retrieved based on vector similarity, typically using cosine similarity as the retrieval metric.
3. The retrieved documents are provided to the LLM as contextual information. Combined with the original user query, they enable the model to generate a more accurate and contextually enriched response.

Although effective, classical RAG systems exhibit several limitations. In particular, vector-based retrieval cannot natively exploit the structured knowledge encoded in Knowledge Graphs (KGs). Queries involving geographic constraints (e.g., a specific country) or hierarchical relationships between concepts may therefore retrieve documents that are semantically similar but factually inappropriate for the user's request. This limitation can contribute to hallucinations, where the generated response is plausible but not grounded in the correct knowledge.

To address these limitations, GraphRAG architectures have been proposed. By integrating Knowledge Graphs into the retrieval process, GraphRAG systems leverage explicit semantic relationships and graph structures to improve the relevance, factual accuracy, and explainability of the retrieved context.

GraphRAG extends the classical RAG architecture by incorporating a Knowledge Graph (KG) as a complementary source of context, rather than relying solely on unstructured documents. This approach combines the explicit semantic structure of RDF graphs with vector-based retrieval, enabling the exploitation of both semantic similarity and graph relationships. As a result, GraphRAG has demonstrated improved performance on tasks involving complex semantic structures, such as hierarchical relationships and geographic constraints.

We adopt the GraphRAG framework proposed by [14], which organizes the architecture into five main components, as follows:

- **Query Processor.** A user query is transformed from unstructured text into a structured representation using techniques such as named entity recognition, relation extraction, query structuring, query decomposition, and query expansion.
- **Retriever.** The processed query is used to retrieve relevant information from external knowledge sources, which in the case of GraphRAG may include graph-based representations in addition to textual documents. The retrieved content is then used to enrich the query and support the generation of a more informed response. A RAG system assumes "Text-In Text-Out" workflows, whilst GraphRAG may include "Text-In, Graph-out" and other combinations.
- **Organizer.** The retrieved data, which may take various forms such as entities, RDF triples, paths, or subgraphs, are preprocessed together with the structured query. The primary objective of this stage is to refine and organize the retrieved content, making it more suitable for consumption by the generator.
- **Generator.** The resulting outputs are adapted to the specific task requirements. For instance, Large Language Models (LLMs) can be used to generate text-based responses, even when the underlying retrieval process produces graph-based representations.

- **Graph data sources.** Unlike traditional RAG workflows, which primarily rely on document-based sources, GraphRAG exploits information represented in graph-based formats. In this work, we use Knowledge Graphs expressed in RDF as the structured knowledge source.

2.6. GraphRAG for forestry definitions

We implement a GraphRAG workflow tailored to the extracted forest-related definitions. The proposed approach relies on a hybrid retrieval strategy that combines vector-based and graph-based sources, allowing the system to leverage both unstructured semantic information and structured knowledge encoded in the Knowledge Graph.

Vector retriever. We use ChromaDB[2] as the vector retrieval engine. It performs cosine similarity searches over the 3,000 indexed concepts and returns the top-k most similar results. This step does not consider structural constraints from the Knowledge Graph, as it is based solely on the semantic proximity of vector representations.

KG retriever. This retriever analyzes the user query to identify and extract different categories of entities. These include the target concept type (e.g., forest, deforestation, reforestation), relevant international organizations (e.g., UNFCCC, FAO, IPCC), the geographical scope of interest (e.g., international, national, local), and geographic regions such as continents (e.g., Africa, Europe). A SPARQL query is executed in parallel with the vectorial retriever, using the detected entities as filters. For questions involving quantitative criteria (minimum area, canopy cover ratio, tree height), an additional enrichment step is performed using a dedicated threshold store.

Considering the two retrievers, a ranking function was implemented in order to decide which sources could be more relevant for a given query. The Reciprocal Rank Fusion (RRF) is as follows:

$$\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k + \mathrm{rank}_r(d)} \tag{1}$$

where $k = 60$ is a standard smoothing parameter and R is the set of retrievers. The *RRF* score produces a unified ranking of relevant concepts, regardless of whether they originate from graph- or vector-based retrieval.

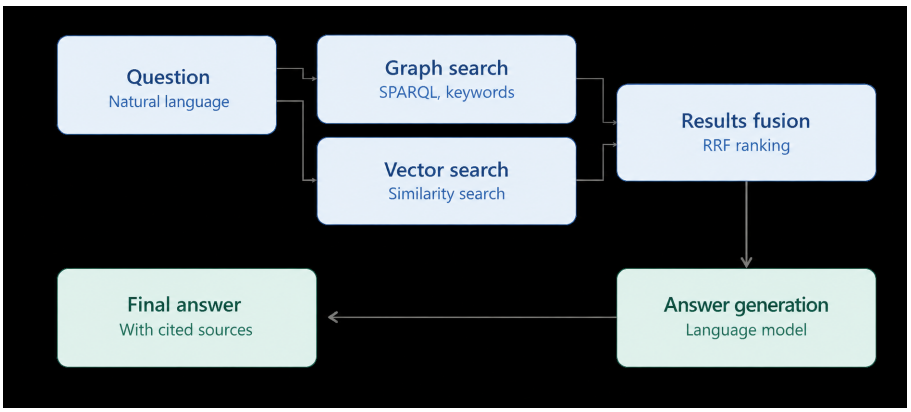


Figure 3: GraphRAG Pipeline

3. Demonstration

This section presents a proof of concept, including a qualitative analysis of the answers obtained for six different Competency Questions (CQs). Besides, we develop a prototype based in web tools in order to enable user interaction with the KG and the GraphRAG system.

[2]https://www.trychroma.com/

3.1. Competency Questions

- **CQ1**. Which countries in Africa define forest with a minimum crown cover of 30%?
- **CQ2**. How many African countries use a minimum forest area of 0.5 hectares or less?
- **CQ3**. Which Asian countries define forest with a minimum crown cover below 20%?
- **CQ4**. Can you compare the forest definition thresholds of Madagascar and Ivory Coast?
- **CQ5**. Can you compare deforestation definitions across FAO, UNFCCC and IPCC?
- **CQ6**. What are the UNFCCC criteria for afforestation?

We can divide the questions into two broad categories: CQ1–CQ4 involve numerical facts, such as crown cover and forest area, whereas CQ5 and CQ6 involve taxonomic and textual information as well as a broader context. Moreover, CQ5 and CQ6 do not involve countries, but global definitions.

3.2. Empirical preliminary evaluation

Table 2 presents the evaluation results for the six given CQs. These results distinguish between the execution of a RAG and a GraphRAG workflow, as indicated in the *Method* column. Besides, we use as baseline an one-shot prompting for the LLM, providing as context the raw textual document (not the graph). This workflow is refereed in the evaluation as LLM+Doc. The *Grounded?* column shows whether the information provided to the LLM only contains real facts from the contexts and there is not hallucinations, whereas the *Missing data* column compares the expected results (e.g., countries or definitions) with the obtained results; additional notes are included in this column to facilitate interpretation. Additionally, the *Sources* column indicates the number of sources retrieved for a given question. This value is relative to the total number of available sources: 3,207 sources for the RAG approach and 6,457 sources for the GraphRAG approach (3,207 document-based sources plus 3,250 graph-based sources). Finally, we report the execution time in seconds.

For each competency question (CQ), we construct a gold standard by manually inspecting the source document and specifying the expected answers. We then evaluate the answers produced by each method (LLM+Doc, RAG, and GraphRAG) against this gold standard. Discrepancies are recorded in the column *Missing data*. This evaluation is carried out using the Gemini model `gemini-3.6-flash`.

For **CQ1**, the RAG method produces a grounded answer (i.e., without fabricated facts or hallucinations), however according to our gold standard 9 out of 11 expected countries were not display in the answer. This gap appears because some source data is found in tables and the model has limitation to interpret such structure. Conversely, the GraphRAG execution, although slower in terms of processing time, retrieved almost all expected country-specific definitions, missing only one case (Morocco). Further analysis is required to determine the causes of this missing result; however, this represents a significant improvement compared with the classical RAG approach. For each method we report the number of sources, it is expected that the GraphRAG uses more sources since it enhances the RAG alternative, at this stage we still require further analysis to identify if the additional sources are all of them relevant, although we observe an important improvement in the obtained answer while keeping groundedness. The baseline (LLM+Doc) also has many missing countries (7/11) whilst it includes some hallucinations (false positives).

CQ2 exhibits behaviour similar to **CQ1**. The GraphRAG execution retrieved all expected definitions, whereas the RAG evaluation retrieved no country-specific definitions. Besides, the LLM-Doc method do not found information in the text for the thresholds. GraphRAG is convenient for this question since our method interprets tabular data and numerical attributes in a more semantic fashion.

CQ3 can also be interpreted using the same arguments. These three queries, which involve numerical facts, exhibit a substantial improvement in answer quality, with an improvement rate of more than 90% concerning our gold standard. That is a promising result for this preliminary evaluation. We observe

CQ	Method	Grounded?	Missing data & Notes	Sources	Time (s)
CQ1	LLM+Doc	No	7/11. Ten countries were returned, six were hallucinations	-	-
	RAG	Yes	9/11 (Congo, DRC, Kenya, Madagascar, Mali, Morocco, Nigeria, South Africa, Uganda)	6	2.400
	GraphRAG	Yes	1/11 (only Morocco missing)	39	31.423
CQ2	LLM+Doc	No	11/11 "threshold of 0.5 hectares or less is not stated in the text"	-	-
	RAG	Yes	All 11 countries missing	6	6.285
	GraphRAG	Yes	Complete	36	31.528
CQ3	LLM+Doc	No	12/12. "document does not contain the specific national definitions"	-	-
	RAG	Yes	10/12 missing (only Cambodia and Philippines found)	6	5.260
	GraphRAG	Yes	2/12 only Yemen and Malaysia missing	48	42.247
CQ4	LLM+Doc	No	"document does not contain the specific forest definition thresholds"	-	-
	RAG	No	Ivory Coast entirely missing, although one exists with the thresholds (0.1/30/5)	6	3.497
	GraphRAG	Yes	Complete	106	9.425
CQ5	LLM+Doc	No	"Comparison cannot be provided because the text provided is incomplete and truncated"	-	-
	RAG	No	Claims that no UNFCCC/IPCC definitions exist, although four UNFCCC and two IPCC definitions are present.	6	10.323
	GraphRAG	No	All UNFCCC and IPCC definitions missing	6	12.713
CQ6	LLM+Doc	No	"The provided text does not contain the specific UNFCCC criteria or numerical definitions for afforestation."	-	-
	RAG	Yes	Complete and relevant. Correctly cites the two UNFCCC afforestation definitions.	6	7.289
	GraphRAG	Yes	Only partially relevant. Focuses on forest thresholds rather than afforestation definitions	10	5.643
Total sources for the Vector DB = 3207 indexed documents Total sources for the KG = 3250 concepts					

Table 2
Comparison of RAG, GraphRAG and LLM+Doc across the six competency questions.

as well that the LLM+Doc method cannot interpret national level definitions. Recall that our SKOS knowledge model literally specify countries and continents.

CQ4, which also involves numerical information, introduces an additional level of complexity, as answering the query requires not only retrieving relevant values but also performing threshold-based comparisons. In this setting, the RAG method provides an incorrect answer, as it states that definitions for Ivory Coast do not exist. The GraphRAG method, in contrast, produces a complete response but exhibits unexpected behaviour: more than one hundred sources are retrieved, which are probably irrelevant for the target question. We are still investigating the causes of these excessive outputs. On the other hand, the LLM+Doc method cannot even detect the notion of threshold, as a result this method

cannot detect both countries and therefore their comparison.

For questions **CQ5** and **CQ6**, which involve a broader scope and no numerical facts, the GraphRAG method still requires improvements. Notably, for CQ5 both methods fail, whereas for CQ6 only the RAG method provides a correct response. For both questions, the LLM+doc method indicates that there is not enough information to provide an answer.

3.3. Prototype

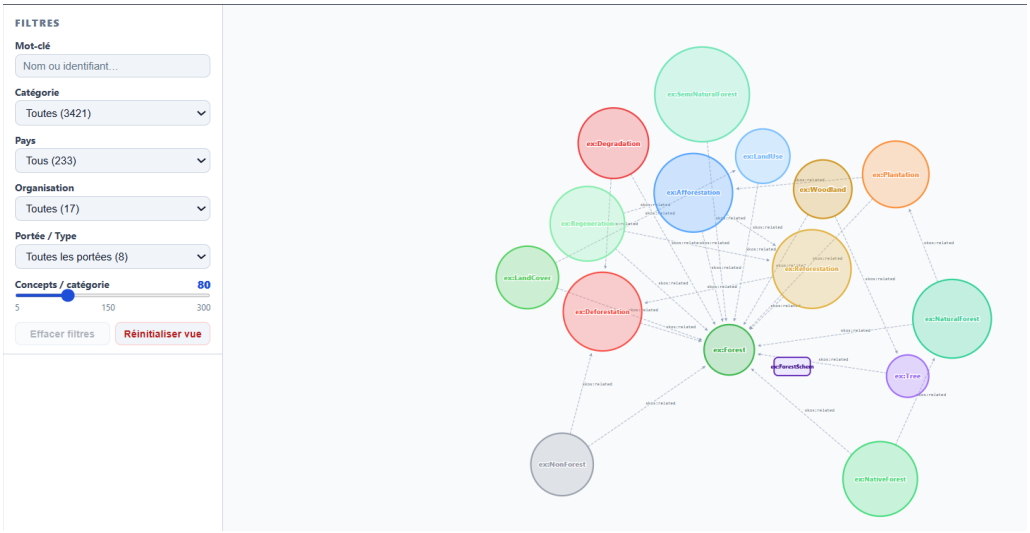


Figure 4: KG visualisation user interface

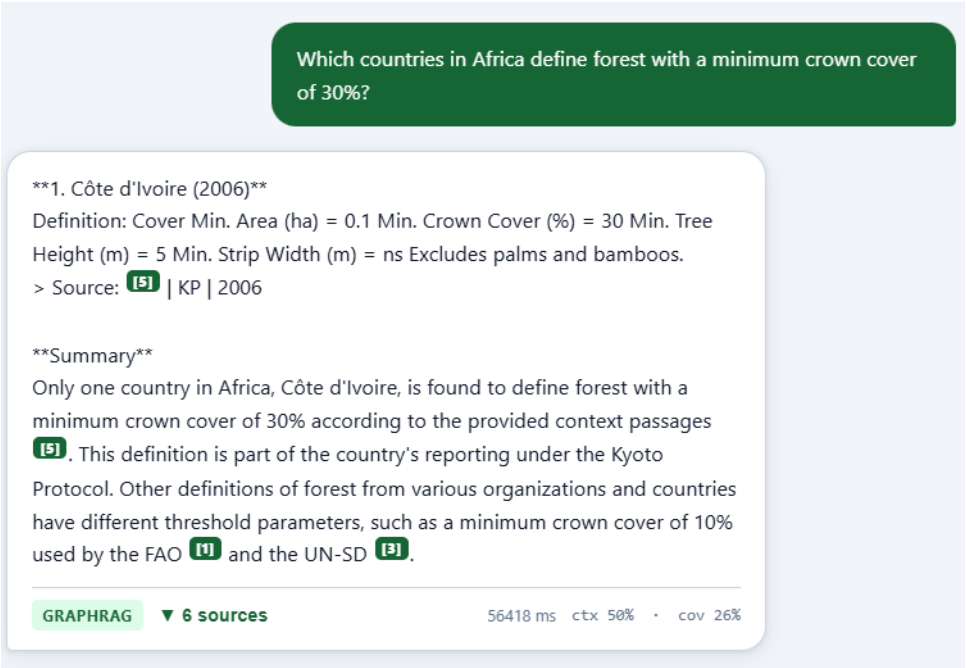


Figure 5: Conversational user interface

We implement a web-based frontend application to visualise the forest definitions Knowledge Graph and provide conversational access to the underlying knowledge. Figures 4 and 5 present a preview of these graphical user interfaces. For this frontend, we used **React 19** with **TypeScript**, together with **D3.js** for data visualisation. We also used the **n3.js** library to parse RDF in the browser. The interface

enables users to interactively navigate the graph by expanding and collapsing nodes, as well as to search for keywords across the entire Knowledge Graph.

4. Conclusions

This work is grounded in a socially relevant and challenging use case: the formal representation of ecological definitions retrieved from forestry documents, addressed in a systematic and comprehensive manner. To our knowledge, few or no works have applied Knowledge Graphs for this particular problem. This study presents a **Knowledge Graph** (KG) construction pipeline as a strategy for representing and managing unstructured information on forest definitions. The proposed approach addresses critical challenges in domains such as climate change, environmental sciences, and forestry, where consistent and machine-readable representations of concepts are essential.

We present a model that relies on SKOS as its main building block and leverages its expressive capabilities to represent hierarchies, categories, and semantic relationships. The resulting Knowledge Graph contains more than 3,000 definitions and 34,599 RDF triples. It captures numerous aspects from the original document, enabling the transformation of unstructured information into a formalised and machine-readable knowledge representation.

Both manual and programmatic curation steps were conducted to improve the quality and consistency of the generated graph. Our pipeline demonstrates the importance of formalising digital documents and guidelines into machine-readable formats. The resulting Knowledge Graph provides a valuable resource for further analyses and future research.

A **GraphRAG** system was implemented as a proof of concept to demonstrate the applicability of the proposed Knowledge Graph. We report preliminary results that provide an initial evaluation of how GraphRAG can improve LLM-generated responses. The GraphRAG approach achieved better performance for questions involving numerical facts; however, further analysis is required for broader textual queries requiring a global understanding of the context.

This tool could be particularly valuable in policy and decision-making scenarios, where reliable answers grounded in official documents are required and can be facilitated through a structured knowledge model.

As future work, we aim to conduct more comprehensive and quantitative evaluations of the GraphRAG system, while further enhancing the quality and reliability of the Knowledge Graph. This includes the integration of validation frameworks based on technologies such as SHACL [15] and ShEx [16] to enforce consistency constraints and improve data quality.

Supplementary Material

All software developments associated with this work are publicly available in the following GitHub repository https://purl.org/forest-rag/code.

Acknowledgments

This work is supported by the European Horizon Project “eco2adapt” [grant agreement ID 101059498] (eco2adapt: Ecosystem-based Adaptation and Changemaking to Shape, Protect and Maintain the Resilience of Tomorrow’s Forests).

Declaration on Generative AI

The authors used perplexity for grammar correction and improving conciseness while keeping the core message. Moreover, Deepl was used for translating ideas from French to English. The authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] C. Bourgoin, A. Verhegghen, S. Carboni, I. Ameztoy, L. Degreve, S. Fritz, M. Herold, N. Tsendbazar, M. Lesiv, F. Achard, R. Colditz, GFC2020: a global map of forest land use for year 2020 to support the EU Deforestation Regulation, Earth System Science Data 18 (2026) 1331–1365. doi:10.5194/essd-18-1331-2026.

[2] P. L. Buttigieg, E. Pafilis, S. E. Lewis, M. P. Schildhauer, R. L. Walls, C. J. Mungall, The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation, Journal of Biomedical Semantics 7 (2016) 57. doi:10.1186/s13326-016-0097-6.

[3] N. Moran, Classifying EUNIS Habitats using Ontologies and Data Mining Methods, Ph.D. thesis, 2016. doi:10.13140/RG.2.2.16643.30247.

[4] I. Subirats-Coll, K. Kolshus, A. Turbati, A. Stellato, E. Mietzsch, D. Martini, M. Zeng, AGROVOC: The linked data concept hub for food and agriculture, Computers and Electronics in Agriculture 196 (2022) 105965. doi:10.1016/j.compag.2020.105965.

[5] A. Miles, S. Bechhofer, SKOS simple knowledge organization system reference, W3C recommendation (2009).

[6] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Scientific american 284 (2001) 34–43.

[7] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, others, The FAIR guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[8] G. Klyne, J. J. Carroll, Resource description framework (RDF): Concepts and abstract syntax, 2004. URL: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, published: W3C Recommendation.

[9] H. G. Lund, Definitions of Forest, Deforestation, Afforestation and Reforestation, Forest Information Services, Gainesville, VA, USA, 2018. URL: https://www.researchgate.net/publication/324755790_2018_Definitions_of_Forest_Deforestation_Afforestation_and_Reforestation.

[10] E. Rametsteiner, T. Linhares Juvenal, M. Boscolo, FAO State of the World's Forests 2022, 2022.

[11] J. Kuyper, H. Schroeder, B.-O. Linnér, The evolution of the UNFCCC 43 (2018) 343–368. URL: https://www.annualreviews.org/doi/10.1146/annurev-environ-102017-030119. doi:10.1146/annurev-environ-102017-030119.

[12] S. L. Weibel, T. Koch, The dublin core metadata initiative, D-lib magazine 6 (2000) 1082–9873.

[13] P. Bellot, Optimiser l'IA générative avec le RAG - la génération de réponses à partir de LLMs, augmentée par la recherche d'information (2025). URL: https://www.techniques-ingenieur.fr/doi/10.51257/a/v1/h6042. doi:10.51257/a-v1-h6042.

[14] H. Han, Y. Wang, H. Shomer, K. Guo, J. Ding, Y. Lei, M. Halappanavar, R. A. Rossi, S. Mukherjee, X. Tang, Q. He, Z. Hua, B. Long, T. Zhao, N. Shah, A. Javari, Y. Xia, J. Tang, Retrieval-augmented generation with graphs (GraphRAG), 2025. URL: https://arxiv.org/abs/2501.00309. doi:10.48550/ARXIV.2501.00309, version Number: 2.

[15] H. Knublauch, D. Kontokostas, Shapes constraint language (SHACL), 2017. URL: https://www.w3.org/TR/shacl/.

[16] E. Prud'hommeaux, J. E. Labra Gayo, H. Solbrig, Shape expressions: an RDF validation and transformation language, in: Proceedings of the 10th International Conference on Semantic Systems, ACM, 2014, pp. 32–40. URL: https://dl.acm.org/doi/10.1145/2660517.2660523. doi:10.1145/2660517.2660523.