Umanistica Digitale – ISSN: 2532–8816 – n. 24, 2026
S. Khouri et al. – *Knowledge Graph for Materials Capitalization of Algeria's Built Heritage: Leveraging Automated Knowledge Extraction and Processing*
DOI: https://doi.org/10.60923/issn.2532-8816/22686

Knowledge Graph for Materials Capitalization of Algeria's Built Heritage: Leveraging Automated Knowledge Extraction and Processing

Selma Khouri

Laboratoire de la Communication dans les Systèmes Informatiques (LCSI), Ecole nationale Supérieure d'Informatique (ESI), Algiers, Algeria
s_khouri@esi.dz

Houda Oufaida

Laboratoire de la Communication dans les Systèmes Informatiques (LCSI), Ecole nationale Supérieure d'Informatique (ESI), Algiers, Algeria
h_oufaida@esi.dz

Sara Allaouchiche

Laboratoire Ville, Architecture et Patrimoine (LVAP), Ecole Polytechnique d'Architecture et d'Urbanisme (EPAU), Algiers, Algeria
s.allaouchiche@epau-alger.edu.dz

Sabrina Kacher

Laboratoire Ville, Architecture et Patrimoine (LVAP), Ecole Polytechnique d'Architecture et d'Urbanisme (EPAU), Algiers, Algeria
s.kacher@epau-alger.edu.dz

Izdihar Belmrabet

Ecole nationale Supérieure d'Informatique (ESI), Algiers, Algeria
ji_belmrabet@esi.dz

Yassmin Bachikh[1]

Ecole nationale Supérieure d'Informatique (ESI), Algiers, Algeria
jy_bachikh@esi.dz

Abstract

Ancient materials knowledge is a capital resource which ensures the effective preservation and restoration of historical structures. However, such knowledge is often complex to capture due to the nature of interrelated information related to heritage materials. Existing research proposing repositories dedicated for documenting ancient materials as first-class citizens, either

[1] Note: Authors are listed in their official order of contribution as displayed in the layout (left to right, top to bottom: 1–2, 3–4, 5–6).

proposes manual approaches, or is commonly centered around relational databases. Contrary to databases, Knowledge Graphs (KGs) are more intuitive, more expressive and more flexible allowing easier extensibility to other contexts (covering further regions, materials or periods) thanks to their graph structure.

Our study proposes the first KG dedicated to the Algerian ancient materials, and highlights its complete automatic construction pipeline. The generalizability of our approach lies in the modeling of a general KG structure (model) that integrates certain concepts inspired by the materials life-cycle model. Our proposal addresses the dual challenge of ensuring generalizability while managing the domain specificities which involve different challenges related to the existence of long-tail and technical entities, the characteristics of the used language and the content of input sources collected with varying granularity levels. To overcome these challenges, the defined KG required conducting various experiments for testing recent intelligent neural models to illustrate their performances and also their limits for our domain.

Keywords: Materials; built heritage; Algeria; Knowledge Graph Construction; Pretrained and Large language models (PLMs -- LLMs).

La conoscenza dei materiali antichi è una risorsa capitale che assicura l'efficace conservazione e il restauro delle strutture storiche. Tuttavia, tale conoscenza è spesso complessa da acquisire a causa della natura delle informazioni interrelate dei materiali. La ricerca esistente che propone repository dedicati a documentare materiali antichi come 'first-class citizens', o suggerisce approcci manuali, o si concentra comunemente su database relazionali. A differenza dei database, i Knowledge Graph (KG) sono più intuitivi, più espressivi e più flessibili, consentendo una più facile estensibilità ad altri contesti (che copre ulteriori regioni, materiali o periodi) grazie alla loro struttura a grafo.

Il nostro studio propone il primo KG per i materiali antichi Algerini e ne evidenzia l'intera pipeline di costruzione automatica. La generalizzabilità del nostro approccio risiede nella modellazione di una struttura KG generale (modello) che integra determinati concetti ispirati al modello del ciclo di vita dei materiali. La nostra proposta affronta la duplice sfida di garantire la generalizzabilità gestendo al contempo le specificità del dominio, che hanno comportato diverse sfide legate all'esistenza di entità "long-tail" e tecniche, alle caratteristiche del linguaggio utilizzato e al contenuto delle fonti di input raccolte con diversi livelli di granularità. Per superare queste sfide, il KG definito ha richiesto la conduzione di vari esperimenti per testare recenti modelli neurali intelligenti al fine di illustrarne le prestazioni e anche i limiti per il nostro dominio.

Parole chiave: Materiali; Patrimonio costruito; L'Algeria; Knowledge Graph Costruzioni; Modelli linguistici pre-addestrati; Modelli linguistici di grandi dimensioni

1. Introduction

Built heritage knowledge is a wealth that has to be collected, capitalized and disseminated in each region, state or country. Algeria is renowned for the diversity of its architectural heritage, reflecting the various historical eras that the country has known from prehistory to the present day (Roman, Arab-Berber, Spanish, Ottoman, etc.). The preservation of such heritage goes through the capitalization of its knowledge, which is the goal of our multidisciplinary research project launched with the aim of exploiting information and digital technologies for knowledge capitalization of Algerian architectural heritage (Khouri et al, 2023), (Amrani et al, 2024). The project involves Computer science and Architecture researchers belonging to two Algerian

higher education institutions. Our current study falls within the scope of this project and aims to document, integrate and capitalize the knowledge specifically related to ancient materials of the built heritage in Algeria, starting with the Algiers Casbah site as a first step (Allaouchiche and Kacher, 2025c). Ancient materials knowledge is a capital resource in the built heritage field, which ensures effective preservation and restoration of historical structures, enables informed conservation strategies and allows to maintain the integrity and authenticity of heritage sites. The Casbah of Algiers (Figure 1), listed as a UNESCO World heritage site since 1992, represents one of the most representative sites of the architectural history of Algeria, including the Ottoman period architecture, and is well documented in the literature. Figure 2 presents examples of ancient materials and their usage within different monuments in the Casbah of Algiers.

Figure 1: Casbah of Algiers localisation (Google Earth Captures)

A building material is defined as "a basic element of construction" (Roth et al, 2021). In our research context, we distinguish between "material" as a raw, unprocessed substance in its natural state (such as wood, stone, clay, etc.), and "product," which refers to any material that has been manufactured or processed into a form ready for use for specific application (such as brick, mortar, etc.). The knowledge related to such materials is fortunately available but accessing such information by heritage stakeholders can be very tedious since it requires the identification of relevant sources of information, their treatment, integration and analysis in order to answer specific requirements. Examples of typical architects' requirements are: *which techniques and materials are used for constructing the building elements? or What are the most common pathologies affecting ancient materials?* Note that accessing such specific knowledge via intelligent chatbots like the free versions Chatgpt or Gemini (relying on LLMs) can either give general information, or inaccurate information called "hallucination". The tendency of LLMs to generate "hallucination" is actually a well-known issue (Ji et al, 2023), which can be explained by the lack of explicit knowledge representation (Pan et al, 2024).

Furthermore, (Amin et al, 2008) underscore the challenges that heritage experts face in information acquisition. These challenges are primarily due to their advanced information-

seeking requirements, which involve complex search tasks (that transcend basic fact-finding) requiring comparisons, exploratory searches, and relationship analysis. Moreover, the imperative to integrate multi-source information and the crucial role of source trustworthiness further compound these difficulties. The primary goals of Knowledge Graphs (KGs) actually consists in providing an explicit and integrated representation of such knwoledge and in facilitating its acess by targeted users.

Existing literature of integrated repositories dedicated for documenting ancient Materials where materials are managed as a primary concept (first-class citizen), either proposes manual aproaches, or is usually centered around relational databases (to the best of our knowledge). Contrary to classical relational databases, Knowledge Graphs (KG) realize a physical data integration from different sources that are combined in a graph representation, which promotes convenient access to the desired information (Korro et al, 2025) and offers: an intuitive structure for end-users, a flexible schema and easy addition of new entities and their interlinking with other entities (Hofer et al., 2023). Drawing upon the "hierarchy of knowledge" (Ackoff, 1989), which typically progresses from Data to Information and then to Knowledge, databases are primarily concerned with the observed data and deduced information. Knowledge Graphs, on the other hand, are designed to capture and represent knowledge explicitly for reflecting expertise and actionable "know-how". KGs have thus emerged as suitable solutions for efficiently capturing, organizing and managing a domain knowledge. The domain of cultural heritage in general, has shown the development of KGs for different sub-domains around the world (Quattrini et al, 2017), (Dou et al, 2018), (Simeone et al, 2019), (Carriero et al, 2021), (Korro et al, 2025). Our study extends these efforts to the particular domain of Algerian ancient materials.

Figure 2: Illustrations of Materials and their usage within different Monuments in the Casbah of Algiers (Source:Authors' Captures)

Recent approaches for KGs construction promotes the usage of intelligent neural models (mainly Pre-trained Language Models -PLMs- and Large Language Models -LLMs-), and the exploitation of the knowledge encoded in the language models' internal parameters, called *parametric knowledge*, for the KG construction (Pan et al, 2024). Such models proved efficieny especially for resource-rich languages such as English, however, their efficiency have to be tested for each vertical domain (Sahbi et al, 2025). In our study, we propose a complete method for automatically constructing a KG dedicated to building materials (as first-class citizens) of the Algerian built heritage using recent intelligent models, which is a new research contribution to the best of our knowledge.

The KG definition is based on a set of sources rigorously collected through an extensive research and academic literature conducted by our architects experts in the context of an ongoing doctoral thesis. These sources are constructed after meticulous examination of reliable documents (books, academic thesis & publications, institutional sources, etc.). They are written in French mainly as the academic references chosen are defined mostly in French language (with some few sources in English). The input sources are various, they are mainly textual sources including different illustrations, tables and references.

Our approach addresses the dual challenge of: (i) ensuring generalizability reflected by the choice of a KG repository that can be easily extended and the design of a general KG model that draws inspiration from the materials life-cycle and the principles of heritage conservation, (ii) managing the domain specificity. The specificities of our overall multidisciplinary project (Khouri et al, 2023) are related to: (i) the lack of training datasets, (ii) the specificity of the used language which is Standard French involving different transliterated terms from other languages such as Arabic, Berber, Ottoman or Andalusian with variable spellings like Mihrab (*A niche in the mosque indicating the direction of Mecca - Qibla*), Qbou (*An overhanging bay*), Dar Aziza (*Aziza House*), etc. and (iii) the heterogeneities of the input sources format. The domain of Algerian ancient materials adds significant specificity since materials information are described using various technical terms, and presented with different granularity levels covering general information (eg. Material categories) and precise insights into construction techniques, pathology treatments, material provenance, etc. Our aim for the defined KG is to constitute a foundation of relevant and reliable knowledge reflecting a real expertise and know-how related to ancient materials; that can be

easily extended to other regions and periods and readily exploited for a wide spectrum of systems spanning from digital platforms (Allaouchiche et al, 2025b) to AI-driven applications.[2]

The rest of the paper is organized as follows: section 2 presents the related works. Section 3 describes the necessary background to understand our approach. Section 4 details our contributions for constructing the KG. Section 5 analyzes the results obtained. Section 6 concludes the paper and sketches some future works.

2. Literature Review

Material information plays a central role in different applications and domains (chemical, physics, engineering, robotics, construction, etc.). We found out that many terms can be related to material information (like material databases, material libraries, etc.) which constitutes a challenge for gathering the most relevant studies. For our literature review, we have focused on studies proposing repositories centralizing and integrating data collections of ancient construction materials for documentation and scientific purposes.

In the Algerian context, different studies documented ancient materials without developing digital tools. For example, (Rezkallah and Marmi, 2018) rely on petrographic analyses on sandstone and limestone used in the roman city of Thamugadi (Timgad) and documents the ancient quarries supplying them within a 25 km perimeter. (Kharfi et al. 2019) apply thermoluminescence TL dating and X-Ray fluorescence to investigate the age and clay sources of ancient bricks in the roman city of Cuicul (Djemila) in Algeria. (Djerad et al, 2022) use multiple analytical techniques to characterize roman lime mortars from Hippo Regius archaeological site (Annaba), identifying aggregate types and chemical properties, and linking them to local geological sources.

(Attari et al, 2019) focus on the physico-chemical characterization of an ottoman period lime-sand mortar used in the Rais Palace in Algiers, in order to assess material compatibility for restoration purposes, and investigate potential causes of the Bastion 23 façade deterioration. (Gheris A, 2023) proposes a multidisciplinary approach combining multiple analyses to characterize ancient lime mortars and mortar variability from Hippo Archaeological site. It introduces a refined method to confirm or suggest the dating of certain monuments of the site.

These studies cover different aspects like characterization, provenance, deterioration processes, etc. and reflect the richness and the complexity of the knowledge related to Algerian ancient materials. However, we note few digital technologies developed for centralizing this information scattered across multiple heterogeneous sources, dedicated to the Algerian context. We can cite (Allaouchiche et al, 2025a) proposing documenting materials used during the Roman period in Algeria,[3] and (Rezkallah, 2020) proposing a Geographic Information System (GIS) for the excavations at ancient Thamugadi. These studies contribute significantly to the documentation and dissemination of the knowledge about ancient materials for a better understanding of

[2] The construction of such systems does not fall within the scope of the paper, which focuses on the KG construction.

[3] We project to include this work and period for the alimentation of the KG under construction at a later stage.

Algeria's built heritage; however, they do not construct a knowledge graph for ancient materials information.

Regarding research beyond the Algerian context, many studies rely on database repositories. We have conducted in (Allaouchiche and Kacher, 2021) a survey on material libraries defined as "a virtual space to search, consult and learn about materials". We completed our review by conducting a lightweight Systematic Literature Review (SLR) to identify the closest studies aiming to develop a KG related to ancient materials as first-class citizens. The SLR is conducted on Google Scholar using the search queries ('Construction Material' AND 'Knowledge graph' AND 'Built Heritage') for the period (2020-2025). We varied the search query with the following synonyms Construction Material/Building Material, Knowledge graph/ Knowledge base, and Built Heritage/ Architectural Heritage, the different combinations provided (8) search queries. The SLR consisted on selecting the main studies by filtering non relevant studies according to their title, then their abstract if the title matches with the search goal.

From both reviews, the following works are considered as the most relevant. Regarding the development of digital material libraries dedicated to the built heritage domain, we can cite different studies. For example, (Kampfová and Richard, 2010) present a database of natural stones in the Czech republic by detailing its technical and content descriptions. (Frangipane, 2010) aims to define an electronic database of the historical stone resources of Friuli-Venezia Giulia (Italy). (Hyslop et al, 2010) present and describe a collection of building stones databases in the United Kingdom. (Delegou et al, 2013) propose to analyze building materials and decay themes in archaeology in Greece using a GIS system. (Turmel et al, 2016) propose a spatial and statistical analysis using GIS-database, for identifying the characteristics of sand origin of building materials of Pays Rémois (France), a region which experienced different historical changes.

More recently, a Roman Construction Materials Database has been designed specifically for a digital reconstruction system (Napolitano et al, 2019), the study covers two types of materials (stone and timber) of the Roman empire, it also cites other databases covering specific materials or specific regions. The Flame-D database (Perucchetti et al, 2021) is dedicated to ancient metallic objects, whose model is mapped to CIDOC-CRM (considered as a knowledge model) then linked to a GIS system.

In the same vein, (Correia and Silva, 2023) propose a database for building materials of the Portuguese built heritage, which is mapped to a defined ontology. Other studies propose cadastre repositories like (Schauppenlehner et al, 2021) presenting the development of historical earth building cadastre for an Austrian region (Weinviertel), that is alimented by local citizens and other actors like schools and museums. Among European projects, GeoERA program is dedicated for raw materials geological data including Eurolithos project for ornamental stone resources (Wittenberg and Oliveira, 2023).

In most of the cited studies, the automatic information extraction process is commonly not fully detailed. Additionally, they usually rely on a relational database structure to manage ancient material information, for example, MySQL system is used in (Napolitano et al, 2019) and PostgreSQL in (Perucchetti et al, 2021). Contrary to relational databases, a KG is recognized for its schema flexibility for integrating new data sources (Hogan et al, 2021) and its easy evolution at both levels (schema and instances) offering smooth addition of new classes, properties and relationships thanks to its graph structure (Weikum et al, 2020). Additionally, while the database model is *descriptive*, i.e. defined to serve a specific application, a KG model is *prescriptive* and aims to cover diverse application contexts (Pierra, 2008). Our study leverages the knowledge graph technology for Algerian ancient materials documentation and management.

We note that different KGs can be found in the literature related to cultural heritage in general like the UNESCO graph[4], (Dou et al, 2018) proposing a KG for Chinese intangible cultural heritage, (Carriero et al, 2021) proposing the ArCo knowledge graph dedicated to the Italian cultural heritage. We also note that knowledge-driven approaches to safeguard the architectural heritage are multiple. Some studies define or exploit KGs or just knowledge models to enrich the knowledge around the built environment artefacts and components ; to assist the conservation process ; to enhance the heritage knowledge management or to improve data retrieval, etc.

Examples of studies include, but are not limited to, (Acierno et al, 2017) proposing an ontology-based model designed to manage knowledge within architectural heritage conservation processes. (Wang et al, 2020) propose a KG for a grotto temple building in China, for facilitating the analysis, retrieval and navigation through the grotto's heritage resources. (Casillo et al, 2023) propose an ontology-based approach to enhance information retrieval of heritage artefacts by experts, in this study, the datasets used are mapped to the ArCo ontology. (Néroulidis et al, 2024) describe the development of a digital platform, exploiting a semantically enriched dataset, for the management of scientific data related to post-fire study of the cathedral Notre-Dame de Paris. (Korro et al, 2025) propose a framework for efficient information management of conservation-restauration projects, composed of a catalogue document, HBIM model and a knowledge base. Different other studies propose semantic extensions and enrichment of BIM environments like (Pauwels et al, 2013) proposing to enrich BIM models with semantic networks and technologies to enhance the documentation and interpretation of a built environment. (Quattrini et al, 2017) propose a pipeline for semantically enriching Heritage BIM (HBIM) data. (Simeone et al, 2019) propose to semantically enrich the BIM environment by using defined semantic bridges (mappings) between BIM database and a knowledge base. Finally, some other studies proposed knowledge extraction processes for built heritage artefacts like (Marchand et al, 2020) proposing a knowledge extraction approach from documents describing real estates in Quebec.

These studies highlight the importance of developing KGs for the built heritage domain. Their repositories may have materials related information, however, they are not presented as Materials repositories considering ancient material as instances of the primary class. Our aim is to develop a KG considering Materials as a first-class citizen, meaning that ancient materials knowledge is the focus of the repository, and is analyzed according to different dimensions.

Table 1 presents the most recent studies selected as the closest to our study reflecting research efforts in both Algerian and international contexts, and compares them according to these criteria: (1) knowledge content, (2) Input sources (Format and language), (3) knowledge extraction (NM: Not Mentioned/ Manual/ Automatic), Storage System (No storage/ Database / Knowledge Graph), Targeted application (i.e. automatic tool or software, "not applicable" for manual works), Automatic metadata Management ("Supported/ Not supported" if automatic approach is proposed, "Not applicable" otherwise).

[4] https://ich.unesco.org/en/dive

Works	Knowledge Content	Input sources	knowledge extraction	Storage system	Targeted application	Metadata
(Attari et al, 2019)	Characterization of ancient mortar in Algeria	Textual sources (French-English) and Laboratory tests	Manual	No storage	Not applicable	Not applicable
(Napolitano et al, 2019)	Two types of materials (stone and timber) of Roman culture	Format and Language (NM)	NM	Mysql Database	Digital reconstruction tool	Supported
(Rezkallah, 2020)	The excavations at ancient Thamugadi	Textual sources (French-English) and photographies	Data entered in the GIS	Database of the GIS	GIS	As supported by the GIS
(Perucchetti et al, 2021)	Ancient metallic artefacts	Format and Language (NM)	NM	Postgres Database	Web application - GIS	Supported
(Schauppenlehner et al, 2021)	Clay as building material	Data entered by users through a web application	NM	Mysql Database	Web application	NM
(Gheris A, 2023)	Dating of ancient materials (lime mortar, Annaba city in Algeria)	Textual sources (French-English), photographies and Laboratory tests	Manual	No storage	Not applicable	Not applicable
(Correia and Silva, 2023)	Building materials of the Portuguese built heritage	Format and Language (NM)	NM	Database	Web application	Supported
Our study	Algerian Ancient Materials (Casbah site, extensible to other sites)	French and various transliterated terms – Format (text, illustrations and tables)	Automatic (using intelligent model)	Knowledge Graph	Multiple applications targeted	Supported & Automatic (intelligent model)

Table 1: Comparison table with related works

Given this table, our aim is to provide a systematic and detailed approach for the construction of a KG dedicated to ancient materials in Algeria, using automatic knowledge and metadata extraction process, that can be used for multiple applications (such as digital libraries, building virtual education tools, chatbots, and many other AI-driven systems), where KGs have demonstrated significant efficiency gains over the past few years (Peng et al, 2023).

Dealing with "*domain-specific KGs*" requires adopting relevant knowledge extraction methods. The survey study (Abu-Salih, 2021) classifies such methods from simple methods relying on human-crafted rules to intelligent methods relying on deep learning techniques. The author also notes a serious lack of details concerning the information extraction techniques used in the literature related to domain-specific KGs. In our domain-specific KG, we will methodically detail the techniques used for the knowledge extraction, as well as their performances.We note that PLMs and LLMs models are currently explored for information and knowledge extraction in different fields (Arora et al, 2023), (Polak et al, 2024), (Giagnolini et al, 2025). However their support for KG generation has still to be analyzed and is still under consideration (Sahbi et al, 2025), especially for vertical domains like ours. The potential of language models present a great opportunity since they are easy to use, especially when targeted users do not necessarily have an IT profile for defining or using a complex knowledge extraction pipeline, but the usage of such models has to be evaluated for our particular context and input sources.

3. Theoretical Background

We define in this section the main concepts of knowledge graphs and knowledge extraction.

3.1. Knowledge Graphs (KG)

A knowledge graph is defined as "*a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities*" (Hogan et al, 2021).

Our KG is a domain specific KG (by opposition to a generic KG). Domain specific KGs are smaller but contain very reliable information (Pan, 2024). Two main approaches for KG development can be followed (Tamašauskaitė and Groth, 2022): top-down (where knowledge is extracted based on a data schema defined first) and bottom-up (where knowledge is extracted from data sources, and the KG schema is then defined). For our study, we opted for the first approach mainly because we defined a KG schema that contains the most relevant information of the domain (not only restricted to the set of sources). Furthermore, we see this KG with its future updates and improvements from new potential input sources, as a basis for different application requirements.

Three main typologies of KGs are defined (Hogan et al, 2021): directed edge-labelled graph (where labels are assigned to edges, like in the RDF model), Graph dataset (a set of named graphs identified by an ID and a default graph managing metadata about the graphs), and Property Graphs. Despite their lack of standardization compared to RDF graphs, we opted for this last typology for its flexibility of modelling allowing a set of properties (<property:value> pairs) and labels to be associated with both nodes and edges, and for its flexibility for handling path extraction queries (Hogan et al, 2021). These characteristics are important for our context., where our KG schema includes many necessary properties in the nodes and the edges, which augments the expressivity of entities of the domain and the expressivity of links between entities. Furthermore, path queries are important for allowing the heritage stakeholders to explore the graph and to identify connections between the concepts that were not initially stated in their requirements.

3.2. Knowledge Extraction

Knowledge acquisition (or extraction) is defined as "*the general process of collecting elements from multi-structured data to build a knowledge graph. It includes entity recognition, coreference resolution, and relation extraction*" (Zhong et al, 2023). Named Entity Recognition (NER) and Relation Extraction (RE) are the main operations that ensure the KG alimentation. These two tasks are inherited from the Natural Language Processing (NLP) field in which various models and toolkits were proposed and developed (Jehangir et al, 2023), (Detroja et al, 2023). Recently, deep-learning architectures (namely PLM and LLM) have leveraged the efficiency of NER and RE models especially for resource-rich languages such as English.

By definition, PLMs are based on the pre-training of the models on large-scale corpora then fine-tuning the models for specific and practical NLP tasks. They can encode significant amounts of linguistic knowledge from the pre-training corpora into their parameters allowing them to learn contextual representations of the language (Li J et al, 2024). Relatedly, LLMs are defined as computational models with massive parameter sizes and advanced learning capabilities, having the ability to predict the likelihood of word sequences or to generate new text based on given input (Chang et al, 2024) .

Let's take as example the following sentences that are literally reported from our input sources (see Figure 3):

- French: La pose d'un film imperméable (bitume, nylon) entre le support et la chape, permet de corriger le défaut d'étanchéité
- English: Laying an impermeable film (bitumen, nylon) between the substrate and the screed corrects the waterproofing defect

Entities to extract are: (impermeable film), (bitumen), (nylon), (substrate), (screed), (waterproofing defect). The relationship "restore" to extract is between the entities (impermeable film, substrate, waterproofing defect correction).

Another important knowledge to extract is the attributes values of entities (such as "natural" or "squared" for the attribute "form" in an entity "Cedar") and bibliographical references (such as author, title, year, etc.) defined in the sources.

We want to extract such fine-grained knowledge from the sources. Each language has its specificities (lexical diversity, syntactic proprieties, grammatical nature, etc.) which makes the applicability of existing models (usually tested for English) difficult to adapt. For NER, delimiting nominal segments and labeling each segment with the right entity type is challenging in a sense that various entity names are transliterated from other languages such as Arabic, Berber, Ottoman or Andalusian with variable spellings: "Dar Aziza" i-e "Aziza House" or "Djenan El Saydji" i-e "El Saydji Garden".

Other entities are very technical and highly ambiguous in terms of delimitation and semantic, leading to the non-efficiency of standard models when applied directly. For example "hydraulic mortar" or "cut stone" should be extracted as one entity and not separately, while each separate term can have another significance in other contexts. Furthermore, relationships to be extracted belong to the technical features of the materials used, which make them highly specific and difficult to extract even by using recent deep learning models. In our study, we will first explore the efficiency of using LLMs directly using a few-shot prompting approach (i-e, using prompts and given examples), then by relying on PLMs and LLMs for defining a customized pipeline (Figure 3).

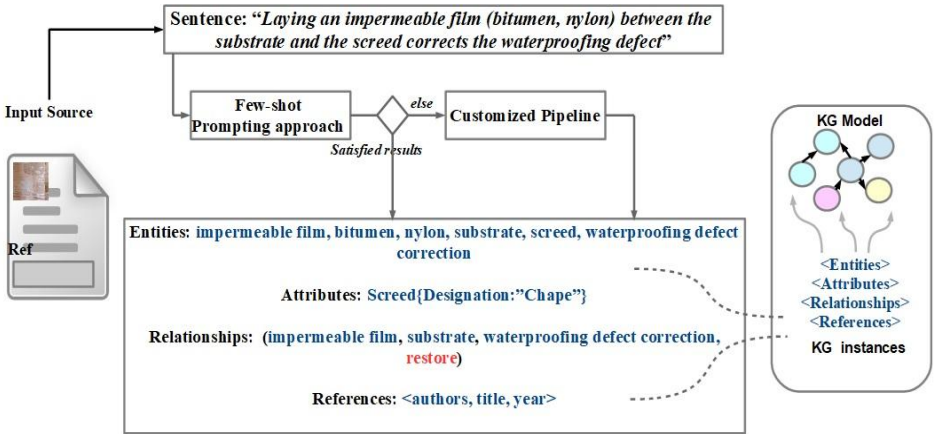

Figure 3: Knowledge extraction process illustration

4. Methodology: KG construction

Pursuing our goal of managing generalizability, we adopted a general KG construction process (Tamašauskaitė and Groth, 2022) that defined the key constituent stages of the overall process of KG development through a conceptual analysis of KGs literature and real-world case studies. As the authors highlight, this process is particularly well-suited for initial KG development. We projected and adapted the process to our context by different iterations for defining a suitable knowledge extraction pipeline. Different stages are provided which can be summarized as follows:

- Step1: Data sources identification (*section 4.1*),
- Step2: Construction of the KG model (*section 4.2*),
- Step3: Knowledge extraction and processing (*section 4.3*)
- Step4: KG construction and maintenance (*section 4.4*).

Figure 4 summarizes the global approach which steps are explained in what follows.

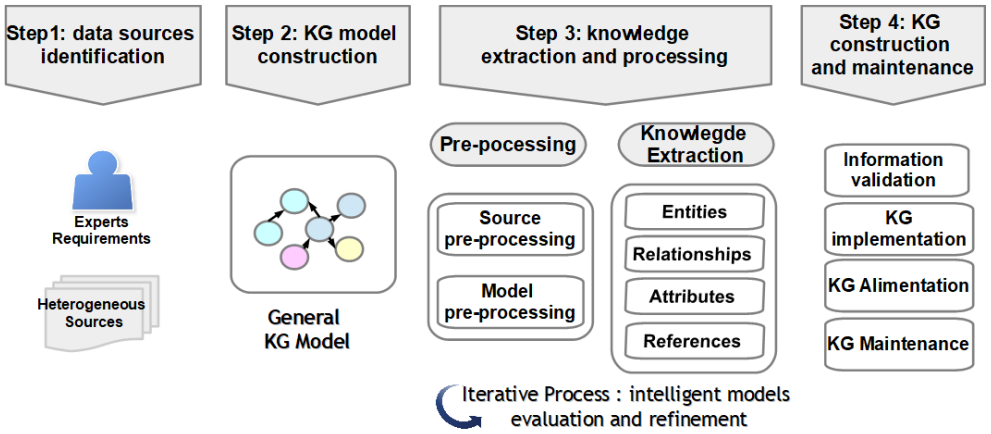

Figure 4: KG construction stages

4.1. Step 1: Data sources identification

As highlighted in (Weikum, 2021), input data quality is key for KG success, where prioritizing what is called "premium sources" (i.e. authoritative, high quality and high coverage sources) is recommended. The input data sources constructed by our domain experts constitute our premium sources. They are detailed in the following table (Table 2).

Source	*Format*
274 descriptive sheets describing materials	Textual format (doc files)
99 files Content describing pathologies and monuments	Excel files (containing structured and unstructured textual content)
482 bibliographical references	Inserted within the doc files. The cumulative number of references cited across the sheets cited is 482 (including repeated citations), the total number of references used is 69.
Geographical coordinates	1 Excel file
97 illustrations	PNG, JPG and JPEG files

Table 2: Input sources description

The set of sources have been created/elaborated by the architectes of our team, following a careful examination and cross referencing of reliable textual sources (mainly published books and thesis) and onsite observations about the materials and techniques used for the study corpus, checked and compared with the theoretical sources. The references collection, selection and examination have been a long and careful process during which an extensive amount of different sources have been gathered about the different aspects regarding the project, often following a snowballing method, as well as international and theoretical references (mainly published papers, books and thesis) to allow comparisons and interpretations. In addition to the complexity of the

source language and their format heterogeneity, they also do not follow a uniform structure and contain information of different granularity levels. Some sheets contain very specific information about materials, while others provide descriptions about monuments. These challenges augment the complexity of defining a suitable extraction pipeline.

4.2. Step 2: Construction of the KG model

The KG model provides a conceptual view of entities of the domain, their properties and their relationships. Different examples of ontologies in the general field of construction industry are cited in (Jäkel et al, 2025). However, recent studies like (Kaltenegger et al, 2024) point to the lack of an extensible data model for building material classification and the heterogeneity between different ontologies defining materials. This situation pushed us to define our own KG model, illustrated in Figure 5 using the standard language UML (Unified Modeling Language).

The KG model considers the principles of heritage conservation and draws inspiration from the material life-cycle model (Huang et al., 2020), which follows a circular process—from raw material acquisition to transformation into usable products, intended for different applications across the building components, integrated into an existing structure. Over time, these materials are subjected to diverse environmental and anthropogenic factors that contribute to their degradation throughout their service life. Such deterioration can compromise the structural stability of the building and, in extreme cases, result in partial or total collapse. In these situations, certain materials should be recovered and repurposed for reuse, in line with sustainable heritage practices.

From a methodological perspective, our approach for designing the KG model is based on the principles and activities of the well-known *Methontology methodology* (Fernández-López et al, 1997), namely: the requirements specification for defining a general KG model, knowledge acquisition, conceptualization of the model (using UML standard language), the integration where we considered reusing definitions already built in other ontologies, the implementation of the model (using Protege tool), its evaluation and documentation with architects experts. The KG model was defined using an iterative process to obtain a model fully validated by the architects.

The obtained KG model presents two central classes: *Material and Product*. The class **Material** have two reflexive relationships (*inherits_from* and *combined_with*), for example, materials (Lime, sand, water) can be combined to produce "lime mortar". A **product** is composed of a set of **materials**, for example "brick" is composed of clay, sand, etc. (*Made_of* relationship). A product may be composed of other products (*Composed_of* relationship). Products and materials provenance is represented by the **Place** entity. The class **Monument** indicates the set of monuments collected. For instance, "Dar Aziza", a famous historical monument in Algiers uses different products like tile, brick and mortar (*Prod_Used_in* relationship).

Each monument is made up of different **building elements** (*Elt_used_in* relationship), which ensure specific usages. Building elements are constructed from defined products using specific **techniques** ("construct" relationship). And each given product may be *shaped* using specific techniques or tools ("shape" relationship). Products and building elements are concerned by **restoration** processes (*restore* relationship), for example, the walls (eg. brick walls) are regularly resurfaced using lime mortar to prevent water infiltration. Different **pathologies** may *degrade* materials, products and building elements. For example, cracks and efflorescence can be observed in bricks and wall plaster. The classes "**Period**", "**Illustration**" and "**Source**" (i.e. bibliographical references) are linked to different entities. Note that these three classes are represented in the diagram without the links to other classes for the sake of readability of the diagram. All the entities are also characterized by a set of properties. For example, pathologies

may have different causes and should be treated using specific treatment techniques. We note that the knowledge of the domain is reflected through the concepts defined by these classes, and their inter and intra links with other concepts, offering a solid understanding of the domain than can be shared between different experts and read by the machine (Cursi et al. 2022).

It is also worth noting that we envision the alignment of the proposed ontology with foundational ontologies (like CIDOC-CRM), following a bottom-up approach (Trojahn et al., 2022), i-e. after extending the KG to cover other eras (eg. The Roman era). As noted in (Trojahn et al., 2022), matching domain and foundational ontologies is highly complex and requires the deep identification of the semantic context and exact meaning of the concepts included. Our objective is thus to ensure a holistic alignment with the foundational ontology, verifying that all mappings are based on the final, nuanced, and fine-grained definitions of the domain concepts.

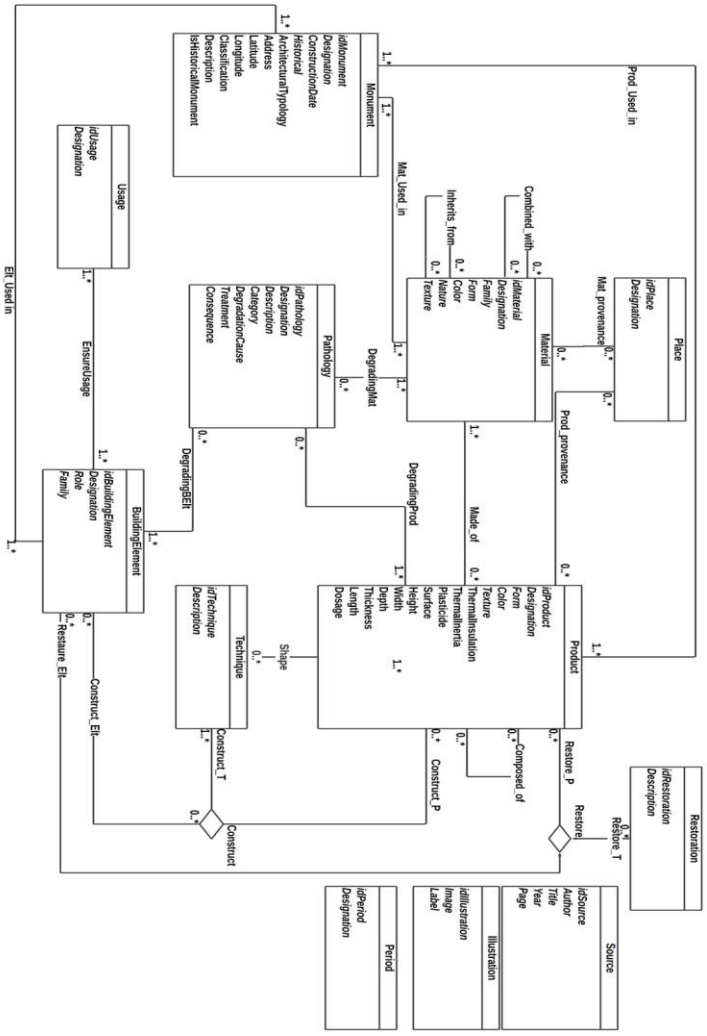

Figure 5: The knowledge graph model

4.3. Step 3: knowledge extraction and processing

The Knowledge Extraction (KE) process was performed through two main cycles. For the first cycle, a set of 117 descriptive sheets manually constructed by the architects experts of our team, were processed. A first KG version was developed from these sources by testing conventional knowledge extraction techniques that are adapted for domain specific KGs (Weikum et al, 2020), namely: CNN using a model constructed from Spacy,[5] StanfordNLP CRF Classifier,[6] and BiLSTM. We also relied on a dictionary of terms, which is known as a common input in domain-specific NER (Weikum et al, 2020), to identify specific and transliterated terms of the Algerian heritage domain. Its construction was based on local thesaurus such as (Chergui, 2015), completed by a semantic annotation service Wikifier (Brank et al, 2017).

A first set of 198 KG entries following the format <Source_Entity, Relationship, Target_Entity> were identified and validated by the experts for this first cycle. This first KG prototype has served as an initial dataset during the second cycle. However, the KG construction required the manual definition of different rules to be effective. This rule-based approach is not recommended for the generalization of the KE when new sources are included. Furthermore, the amount of information extracted was not sufficient. To overcome these issues, we considered in the second cycle, more advanced techniques based on language models. As stated in Figure 4, the KE phase includes two main tasks: Pre-processing (*section 4.3.1)* and Knowledge Extraction tasks (*section 4.3.2*). This last task includes: entities extraction (*section 4.3.2.1*) - relationships extraction (*section 4.3.2.2*) - Attributes extraction (*section 4.3.2.3*) and Bibliographical references extraction (*section 4.3.2.4*).

4.3.1. Pre-processing task.

The pre-processing of the initial dataset is classic in KE pipelines, it required:

Source_preprocessing. It includes these activities:

- LineExtraction: to separate the text/excel files into lines.
- TextCleaning: to eliminate special characters like "-", parenthesis... and supplementary annotations found in the sheets.
- CleanedFile_Generation: to generate an output text file containing the cleaned version of each input file.
- ImageEncoding: images were encoded in base64 format to facilitate their storage in Neo4J,[7] which is the storage system for the KG.

Model_preprocessing. Other pre-processing activities are specifically used for the language models:

- TextTokenisation: converts the input text given to the model into tokens understandable by the language models.

[5] https://spacy.io/

[6] https://www.nltk.org/api/nltk.tag.stanford.html

[7] https://neo4j.com/fr/

- Datasetsplitting: divides the dataset into two subsets: one subset for training (70%), a test dataset (20%) and a validation dataset (10%). This splitting is usual in machine learning in order to test the performances of the models' predictions and their capabilities to manage new information not seen in the training.
- ExtractionEvaluation. For evaluating the results of the knowledge extraction, we used usual standard metrics. Knowing that TP stands for True Positive, FP for False Positive, TN for True Negative and FN for False Negative, three metrics are used:
    - Precision= TP/(TP+FP) indicates the fraction of correct predictions among the overall predictions of the model.
    - Recall = TP/(TP+FN) indicates the capacity of the model to extract relevant information
    - F1-score= 2*(Precision*Recall)/(Precision+Recall) provides a mean between the two metrics of Recall and Precision.
    - We also used the "information_loss_Element" metric defined as: (number of erroneous elements + number of missing elements / Number of total elements expected) *100. Note that the annotations of "erroneous" and "missing" elements is conducted conjointly with the architects of our team.

4.3.2. Knowledge extraction task.

This task starts by the extraction of entities and relationships (section 4.3.2.1), then their attributes (section 4.3.2.2) and references (section 4.3.2.3). A global discussion of the proposed pipeline is also included (section 4.3.2.4).

Motivated by the accessibility of LLMs, our evaluation strategy consisted of first exploring the potential of a chosen LLM to extract relevant knowledge. We used a few-shot approach, where prompts are defined by providing the LLM a description of the information to extract and examples. We thus relied on prompt engineering consisting of prompt creation to elicit desired responses from the LLM for this specific task of knowledge extraction (Pan et al, 2024).

As the spectrum of LLMs is growing very fast and new LLMs are continuously emerging, choosing the most suitable LLM was not an easy task. We thus examined theoretically available open-source LLMs and analyzed established practices documented in various projects and demonstrations (available in e.g., GitHub repositories and technical tutorials).

We chose Llama LLM (Touvron et al, 2023) for its capabilities to effectively treat contextual information,[8] while ensuring optimal utilization of computational resources. This choice is also motivated by the fact that Llama is an open-weights language model. LlaMA is also recognized by the research community and is widely used by many research groups (Minaee et al., 2024). More precisely, we used the Llama model 'unsloth/llama-3-8b-bnb-4bit'.[9] [10]. This choice is motivated by our need to run the model locally under limited computing constraints, ensuring

[8] AI@Meta, "Llama 3 Model Card". (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[9] https://huggingface.co/unsloth/llama-3-8b-bnb-4bit

[10] Daniel Han, Michael Han, and Unsloth team. "Unsloth." (2023), https://github.com/unslothai/unsloth

the privacy of copyrighted material collected and elaborated through an ongoing thesis, by avoiding external LLM providers.

We found out that the LLM is very effective for extracting pairs of attributes and their values (attribute: value) and for extracting bibliographical references, as this will be shown in sections "*Attributes Processing.*" and "*Bibliographical references Processing.*". However, the extraction of entities and relationships using the LLM was not satisfying (at the time where experiments were conducted) and will be demonstrated in the following sections.

Our strategy consisted thus in defining a customized pipeline for knowledge extraction if the few-shot approach does not achieve its goal. For defining our extraction pipeline, a selection of relevant models to test was necessary:

- BERT (Devlin et al, 2019) is a reference large-scale pre-trained transformer-based language model which is used as a basis of different domain specific language models. In our study, we used 'bert-base-uncased' model.[11]

- MatSciBERT (Gupta et al, 2022) is a PLM that is effectively used in material science studies (Schrader et al., 2023). Even if our domain is not under the scope of material science field, its successful usage in this field can be seen as an opportunity for our study to understand the technical concepts related to materials.

We selected MatSciBERT after our exploration of different studies in the field, which has known a profusion of Natural Language Processing (NLP) driven research these last years, eg. (Polak et al, 2024).

- CamemBERT (Martin et al, 2019) is the widely adopted language model for French corpus, it is based on BERT. We have chosen this model for its relevance to understand the specificities and structure of French Language.

As illustrated in Figure 6, three iterations were necessary for defining the pipeline: the first iteration consisted in testing the three models on the input dataset (cleaned version), the second iteration consisted in testing the three models on an enriched version of the input dataset, and the third and last iteration consisted in testing the three models on the enriched dataset, augmented by relationships examples. Each iteration enhances the previous one.

We note that all experiments are performed using an 5th Gen Intel(R) Core(TM) i5-7200U @ 2.50GHz 2.71 GHz, 8GB memory, and with the O.S. Windows 10 Enterprise version 1607. The proposed algorithms are implemented in Python and Colab (to test different models and pipelines efficiently for choosing the most suitable models). The detailed results of the extraction of each element (NER, RE, attributes and references) are given the following sections.

[11] https://huggingface.co/google-bert/bert-base-uncased

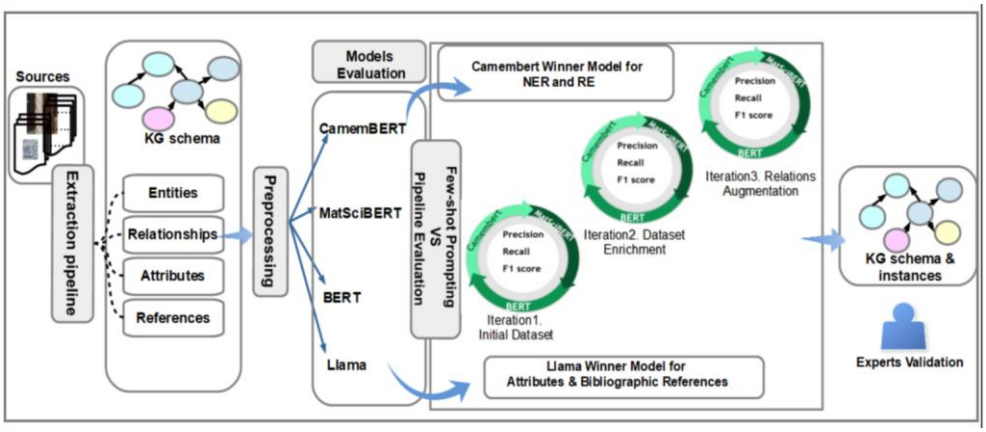

Figure 6: Strategy of exploration for the knowledge extraction task

4.3.2.1. Entities and relationships Extraction (NER and RE)

NER and RE are two main activities that are presented together since the results of one activity can influence the other. As stated previously, our strategy consists in first testing the few-shot approach with Llama LLM for NER and RE. We found out that this approach was not satisfying for this task (at the time where experiments were conducted) as many errors or omissions were noticed. More precisely, we used the "information loss" metric defined above to evaluate the extraction result using the few-shot approach with Llama LLM (Table 3).

Elements	*Information loss*
Entities	84,44%
Relationships	73,33%
Total loss	78.89%

Table 3: Information loss using the few-shot approach

Note that the *information loss* of the knowledge extraction of our defined pipeline will also be evaluated at the end of this section ("*Few-shot VS specific pipeline.*"), and the results will be compared with those of Table 3. We provide in Table 4 an example illustrating the loss of information with the Llama model. For the ease of reading, all the examples that will be provided for the rest of the paper are translated from French (original text) to English.

Error typology	Text	Result	Expected Result
Wrong entity and relationship (False positive)	The brick was regular in shape and appearance, made from clay of often variable quality, with a pinkish color, rarely yellowish	"entity" : " Shape", "relation" : "The brick was regular in shape ... yellowish",	Entity: Brick Relationship: made_of Entity: Clay

Wrong relationship (False negative)	Pierre/stone : cm<L<47 cm ; 15 cm<L<20 cm ; 20 cm<L<25 cm	37	"entity": "Stone", "relation": "Stone :",	No relationship to extract

Table 4: Example of erreneous results

This situation led us to consider the definition of a more specific and personalized extraction pipeline (rather than relying on few-shot prompting), where the three PLMs: Bert, MatSciBERT and CamemBERT were tested. To do this, a fine tuning is required for the three chosen models to adapt them to our specific context. The fine tuning required the specification of a set of hyperparameters to achieve the target adaptation. For this approach, it is important to choose a small learning rate to avoid the knowledge-forgetting problem. We have used the AdamW optimizer with a number of epochs equals to 9, a learning-rate of 5,00E-005 and a number of warm-up steps equal to (0.1 * epochs * train_size)/ batch_size to apply low learning rate during the first steps of training before applying more aggressive updates.

Different refinements were required in order to obtain satisfactory results. These refinements are decided based on the results of different experiments using the chosen languages models. The analysis of the models outputs allowed the refinement of the datasets until the obtention of satisfactory results reflected by the metrics: precision, recall and F1. Three iterations (rounds) allowed the identification of the "*winner*" model among the three models tested. These three rounds are detailed in what follows:

- Iteration 1. *Initial dataset.* Our domain is characterized by the lack of existing training datasets, which is a limit for machine learning models. A first series of tests is achieved on the initial dataset to evaluate the relevance of enriching it. Table 5 illustrates the results of the three models for the NER task. These mitigated results pushed us to consider the enrichment of the dataset before performing the RE activity. An example of erroneous results for this iteration is in Table 6.

Model	*Precision (%)*	*Recall (%)*	*F1 (%)*
Bert	22,35	31,24	26,06
Camembert	37,57	43,32	40,24
MatScibert	42,3	48,17	45,04

Table 5: NER Results (first iteration)

Error typology	*Text*	*Result*	*Result expected*
Wrong relationship extracted	….The vulnerability (low resistance) of earth/clay products (mortar, brick...) to weathering requires	Relationship predicted: Dated_in	Relationship: Restore

	constant maintenance, and in some cases, complete periodic restoration....		
No entity extracted	Form: used as a screed or layer	Entity predicted: /	Entities : screed, layer
Wrong entity extracted	The earth/clay covering is composed of: (check the order).	Relationship predicted: Composed_of Entity predicted: (check the order)	Relationship: Made_of Entity: earth/clay covering
No relationship extracted	Fired clay brick	Relationship predicted: /	Relationship: Made_of

Table 6: Erroneous result example for the first iteration

- Iteration 2. *Dataset Enrichment.* The Dataset-Enrichment considers the input sources and uses a grammatical matching to generate as output a table containing the following statements <Text_source, Relationship, SourceEntity, TargetEntity>. This is achieved following these steps:

- The algorithm prompts the three chosen language models in order to detect the verbs (denoting a relationship in the initial dataset or one of its grammatical variants).
- The algorithm checks whether the entities linked by this relationship are included in the initial dataset. If not, the dataset is enriched with these target and source entities.
- The set of detected entities is validated manually at the end of the process to constitute a reliable enriched dataset for the training of the language models. A total number of 747 new statements was enriched by this algorithm.

We note that the RE task is formulated as a sequence classification task rather than a token-level classification problem, and the chosen PLMs were fine-tuned using our annotated dataset using the set of relations already defined in the KG.

A new series of tests were performed on this enriched dataset, for both NER and RE. The results are presented in Table 7.

NER			
	Precision (%)	Recall (%)	F1 (%)
Bert	90,6	89,38	89,98
Camembert	90,86	89,7	90,28

MatScibert	99,71	99,72	99,71
RE			
Bert	65	71	67,87
Camembert	75	79	76,95
MatScibert	68,06	71,6	69,79

Table 7: KE results for the second iteration

This second iteration enhanced the extraction results w.r.t to the first iteration. However, the analysis of the extraction results showed that the three models presented some limits to identify all relevant entities due to the language specificities and the complexity of the sentence structure. We can cite these examples:

- The three models failed to identify the entity "thermal insulation" in the sentence "their level of thermal insulation is reduced".
- They also failed to identify the entities "Mortar" and "Earth" from the sentence "A thick earth mortar is spread, left for 3 to 5 days to harden".

The results revealed that MatSciBert provided the best results for NER, while Camembert provided the best results for RE, even though the three models did not perform well for RE. The detailed analysis showed that two main relationships were particularly difficult to extract "Restore" and "Shape" for the three models. For instance, the three models failed to identify Shape relationship expected in the following sentence:

- "Shape" relationship between entities "plaster" and its technique should be identified from the sentence: "This plaster is tamped for three consecutive days and nights using wooden mallets, with water and oil alternately poured at specified intervals until obtaining a thick consistency."

- Iteration 3. Relationship Augmentation. The lack of examples of the two relationships Shape and Restore in the dataset explains the low results obtained for RE ("Restore" and "Shape" represents respectively 3,61% and 5,09% of the dataset). This analysis pushed us to consider the augmentation of the dataset by new examples reflecting these two relationships.

As human-validated data can be resource and time consuming, the call to LLMs for data augmentation is an alternative to explore. The context of our language makes this analysis crucial. Our investigation revealed that using the LLMs to generate more training data and to enhance the generalization of extraction models for low-resource languages is successfully used in recent works on different NLP tasks (Yu et al, 2024). Following this direction, language models came again to the rescue, where we have tested:

- "facebook/m2m100_418M" (Fan et al, 2021) which is a translation model supporting more than100 languages.[12]

[12] https://huggingface.co/facebook/m2m100_418M

- The LLM "google/t5-large" (Raffel et al, 2020).[13]
- The LLM GPT "antoiloui/belgpt2" (Louis, 2024).[14]

These models are used for reformulating sentences from the initial dataset (the lines containing the two relationships "Restore" and "Shape") and for enriching it with more examples that can be used during the models training. However, while data augmentation can enhance the generalization of the extraction models, the relationships augmentation task required a manual verification of the proposed reformulations. We can resume the error typology of this data generation as follows: incorrect reformulations that do not give any meaning to the generated sentence, identical formulation of the input sentences or with very limited augmentation, some reformulations include the same input sentence with additional (out of context) words. Our verification of the reformulations given by the three models revealed that the first model provided the best results with 41 accepted reformulations for the relationship "Restore" and 40 for "Shape".

The augmented dataset is again considered for the retraining of the three language models and their performances are evaluated. Not surprisingly, the results of the NER were very slightly enhanced (in the third digit after the decimal point, Figure 7). Even the results of the RE were moderately augmented for the three models, but Camembert showed the best results with F1-score = 88,67% (Figure 8).

We can conclude that the reformulations enhanced the performances of the three models where LLMs offers a cost-effective approach to generating synthetic training samples, but it is not sufficient to identify all the relationships formulated in the dataset, where we notice that the amount of instances for "Shape" and "Restore" relationships is still very low. This fragment of the KG is still under development, and this situation convinced us that the input dataset has to be significantly augmented.

Finally, the trade-off between entities and relationship extraction makes ***Camembert*** the best model among the three models. We note that in our current setup, we performed end-to-end training for each model separately, which includes both entity and relation extraction components. However, more expriments can be driven for separating the NER and RE tasks using the more accurate entities extracted by MatSciBERT as input for CamemBERT's relation extraction module.

[13] https://huggingface.co/google/t5-large-lm-adapt

[14] https://huggingface.co/antoinelouis/belgpt2

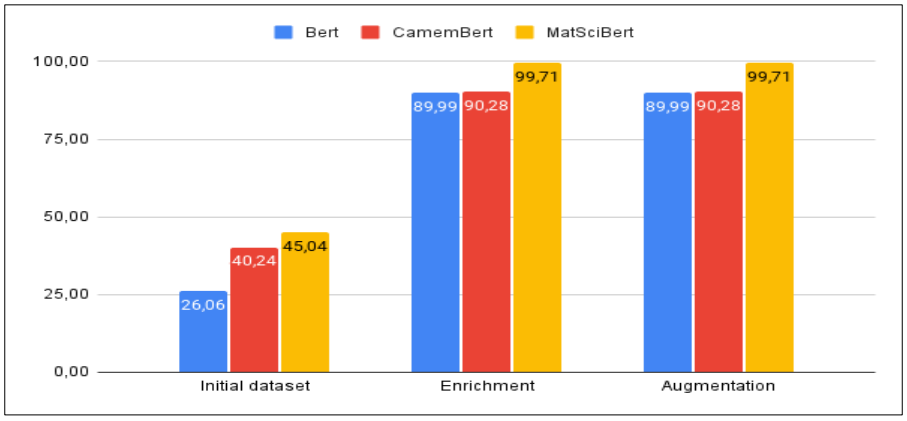

Figure 7: NER F1-scores for the three models along the three iterations

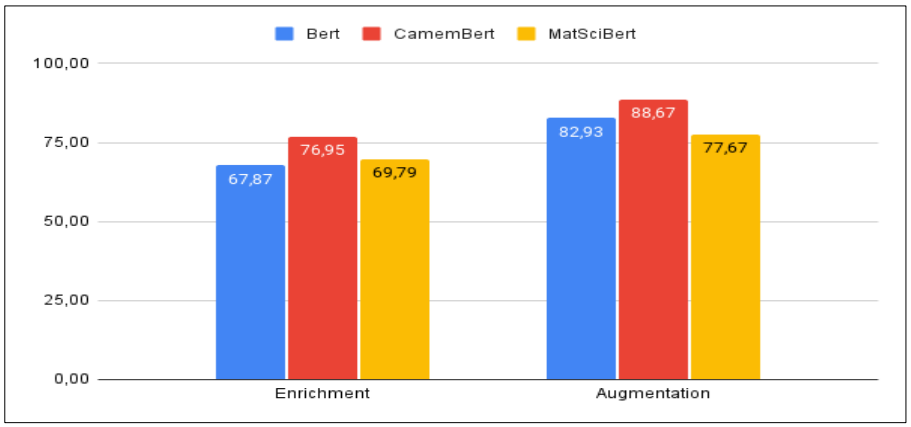

Figure 8: RE F1-scores for the three models along the last two iterations

Table 8 summarizes the NER and RE performances acorss the three iterations.

Models	Metrics	NER (*Initial dataset)*	NER (*Dataset Enrichment*)
Bert	Precision (%) Recall (%) F1 (%)	22,35 31,24 26,06	90,60 (+68.25) 89,38 (+58,14) 89,98 (+63,92)
Camembert	Precision (%) Recall (%) F1 (%)	37,57 43,32 40,24	90,86 (+53,29) 89,7 (+46,38) 90,28 (+50,04)
MatScibert	Precision (%) Recall (%)	42,3 48,17 45,04	99,71 (+57,41) 99,72 (+51,55)

	F1 (%)		99,71 (+54,67)
		RE (*Dataset Enrichment*)	RE (*Relationship Augmentation*)
Bert	Precision (%) Recall (%) F1 (%)	65 71 67,87	85,01 (+20,01) 80,95 (+9,95) 82,93 (+15,06)
Camembert	Precision (%) Recall (%) F1 (%)	75 79 76,95	88,06 (+13,06) 89,29 (+10,29) 88,67 (+11,72)
MatScibert	Precision (%) Recall (%) F1 (%)	68,06 71,6 69,79	74,65 (+6,59) 80,95 (+9,35) 77,67 (+7,88)

Table 8: Summary of NER and RE results across the three iterations

Figure 9 shows the learning curve (Accuracy/F1-score Vs Epochs) for the Camembert model over the nine (9) epochs.

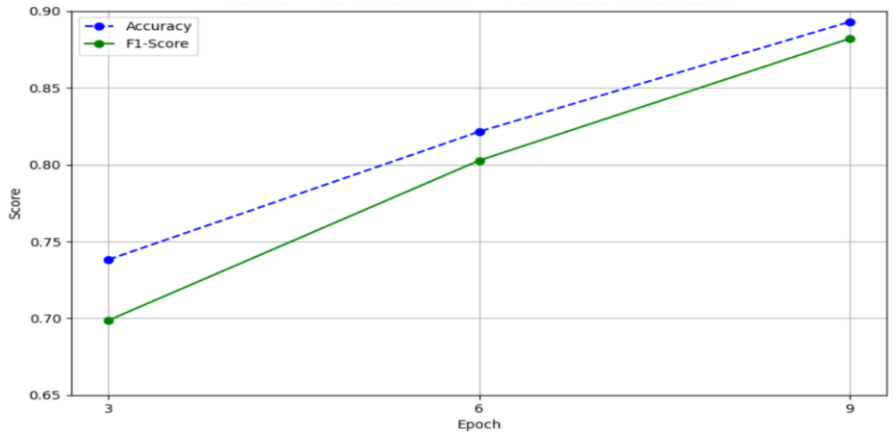

Figure 9: Learning curve for the Camembert model over the nine epochs

- Few-shot VS specific pipeline. To ensure the necessity of defining specific pipeline for the knowledge extraction task, w.r.t the few-shot approach that entirely relies on the LLMs, we compared the performances of Camembert (which showed the best results) and those of Llama LLM (presented in *cf.* Table 3 in section *4.3.2.1. Entities and relationships Extraction (NER and RE)*). Experiments were conducted on all the input sheets, using the information loss metric (Table 9).

Elements	Pipeline-Camembert	Few-Shot Llama
Entities	20,00%	84,44%
Relationships	33,33%	73,33%
Total loss	26.67%	78.89%

Table 9: Few-shot Vs specific pipeline results

These results illustrate that the loss of information is far less important with our defined pipeline. Even though the Camembert model showed very satisfied results, an information loss is also noticed and is explained by the difficulty of the model to manage some complex context or subtle sentence formulations. For instance, from the following sentence:

- Groin vault: in fact, they result from the compenetration and simultaneous construction of two barrel vaults of the same height, whose ridge lines remain whole.

Camembert model correctly identified entities "Groin vault" and "compenetration" but missed the entity "ridge lines".

4.3.2.2. Attributes Processing.

As for entities and relationships, the potential of Llama LLM was evaluated for the extraction of attributes and for references. Contrary to NER and RE, the obtention of very positive results pushed us to consider this model as the solution for these two tasks.

- Prompts_definition. We defined the following prompt (Figure 10) on Llama which contains the task to execute and a description of the attributes to extract.

```
Extract attributes and their values from the following text and format it as JSON

attributes=[
Text(id="form", description="The shape of the material"),
Text(id="color", description="The color of the material"),
…
],
```

Figure 10: Prompt definition for attributes extraction

Result. Figure 11 illustrates some results. The LLM was able to effectively extract the attributes, with a precision=90,6%, recall=100% and F1-score=95% (Figure 13).

```
Results =[
   ("Form : shape  : rough/raw or sawn, beam", [{"form": "rough/raw ; sawn; beam"}]),

   ("Color : Cedar wood color can vary/range from pale yellow to reddish brown",
      [{"color": ["pale yellow ", "reddish brown "]}]),
       …
)
```

Figure 11: Result example for attributes extraction

4.3.2.3. Bibliographical references Processing.

- Prompts_definition. Similarly, the following prompt (Figure 12) is defined containing the task, a reference model for structuring the references in records and examples to illustrate the expected result.

```
# Prompt
Extract the following information from the text and format it as JSON

# Reference model
attributes=[
    Text(
      id="title",
      description="The title of the reference.",
    ),
    Text(
      id="author",
      description="The author of the reference.",
    ),....
]

#Examples
examples=[
("1 Smith, J. (2020). The Art of Referencing. P18",
[{"title": "The Art of Referencing", "author": "Smith, J.", "year": "2020", "page": "p18"}]),

("2 Brown, A. (2018). References in Academic Writing.p65",
[{"title": "References in Academic Writing", "author": "Brown, A.", "year": "2018", "page":"p65"}])
 ]
```

Figure 12: Prompt definition for references extraction

- Results. The LLM was able to efficiently extract references with a precision=95,9%, recall=100% and F1-score=95,9% (Figure 13).

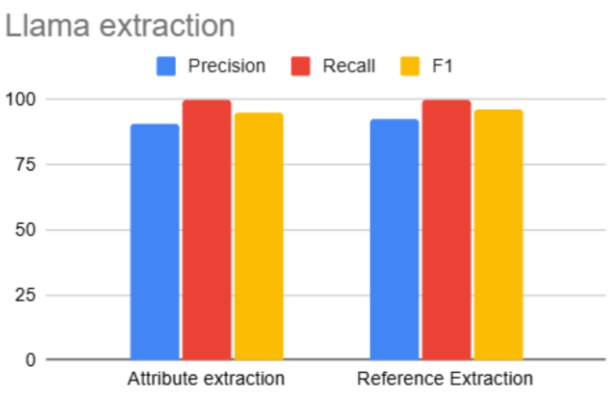

Figure 13: Extraction results for attributes and references

These overall results illustrate the relevance of constructing our global pipeline composed of steps relying on the LLM for Camembert model for NER and RE, and completed by Llama for extracting references and attributes effectively for each entity/relationship extracted.

4.3.2.4. Discussion.

This section discusses the challenges and broader implications revealed by our findings.

- Our pipeline required the generation of new data using translation models (iteration 3). Beyond the quality and accuracy of the generated data, the implications of model-driven data augmentation have been studied in different recent survey papers (Guo and Chen, 2024) (Xu et al, 2024), and can be summarized in these points: inherent biases and hallucinations in the LLMs and their pre-training data can introduce noise in the generated dataset, the lack of domain representativeness especially in specialized domains, and the ethical and privacy concerns either through the uploading of input data to LLMS APIs or through the risk that synthetic data generation might unintentionally expose sensitive elements of the underlying training data. While our current approach relies on manual verification of the model-generated data, a transition to advanced automated techniques is essential for ensuring scalability and managing increased data volumes.

- We emphasize that our implementation features a modular architecture designed for reuse. The separate definitions for training, validation, and extraction pipelines allow the system to generalize seamlessly across other models. For the Llama based-extraction, our approach utilizes an encapsulated pipeline that generates structured output via JSON parsing. By decoupling the generation, parsing, and structuring phases, we ensure the implemented solution is reusable. This modularity allows for the integration of future open-source models by maintaining the global architecture without requiring extensive code modifications.

- Finally, our definition of the complete KG construction process, led us to adopt distinct design decisions at several key development stages. However, it is important to recognize that the complexity of information extraction in this domain requires a dedicated and long-term benchmarking framework to evaluate and compare the large landscape of open and closed LLMs performance and suitability for this domain.

4.4. Step 4: KG construction and maintenance

This last phase in Figure 4 required a validation activity before alimenting the KG. The validation relied on domain experts of our team who annotated the results to identify correct and erroneous extractions (wrong or missing). Table 10 summarizes the obtained results by providing the accuracy (Number of correct elements / Number of elements extracted), the precision, recall and F1-score. The main errors information annotated were reviewed with the experts, however, the validation and maintenance of the KG is still ongoing and must be a continuous process.

Metrics	*Entities*	*Relationships*	*Attributes*	*References*
Accuracy	97,00%	84,30%	91,77%	95,20%
Precision	98,30%	84,49%	90,60%	95,20%
Recall	98,60%	99,60%	100,00%	100,00%
F1-score	98,40%	91,40%	95,00%	97,50%

Table 10: Extraction results (validation step with experts)

5. Results and Analysis.

The KG is implemented using Neo4J, a popular graph oriented DBMS chosen for its support of the graph property model and its performances. The total number of statements in the KG is 1387 relationship instances and 566 entity nodes. All the entities and relationships of the KG schema have been populated which provide a coverage rate of 100% for these two elements.

The *coverage rate* for attributes is 83,67% and for illustrations 50,78%. These two last rates are explained by the lack of these information in the input sources. We recall that our KG schema is defined to cover the domain of interest, and some concepts that are defined in the KG schema were not necessarily found in the input sources.

Concerning the *density of the KG*, the entities that present the higher rates of instances "BuildingElement" (13,07% among the overall instances), "Pathology" (13,96%), "Place" (8,13%), "Monument" (7,24%), "Material" (7,06%) and Product (5,12%).

Concerning the *connectivity of the KG*, the relationships that present the higher rates of instances are entering relationships to Source entity (39,15% instances among the total set of instances), relationship "DegradingMat" and "DegradingElt" (10,24% instances) and "Elt_used_in" (12,83% instances).

We note that *Cypher* is the query language used for querying a Neo4J KG, for example, the following Cypher query (*match(n:Materiau)-[r]-(p) return n, r, p*) returns the KG fragment that connects the Entity Material to other nodes. This fragment shows that the entity Material is at the heart of the KG, and is composed of 231 entity instances and 449 relationship instances. The relationships that are mostly populated in this fragment are: DegradingElt and DegradingMat (122), entering relationships to Source entity (70) and Made_of (57).

Other queries reflect examples of the requirements expressed by architects, that are easily provided by the KG, like:

- What are the most common pathologies with their category and degradation causes?
- Which pathologies are the most documented (according to the number of references/sources)?
- Which pathologies can affect each material?
- Classification of materials w.r.t their number of pathologies.
- Show the geographic position of the monuments.
- Return information from and to the material which designation is "Terre" (i.e. Earth/Clay).
- Which techniques and materials are used for constructing the building elements ?

Query (a)	*Query (b)*
MATCH()-[r]-(p:Pathologie) RETURN p.designation AS Pathology, COUNT(r) AS NB_Pathologies, p.Category AS Category,	MATCH(p:Pathologie)-[r]->(s:Source) RETURN p.designation, Count(r) As NB_sources ORDER BY NB_sources desc limit 10

collect (distinct p.cause_degradation) AS Degradation_Cause ORDER BY NB_Pathologies DESC	
Query (c) MATCH(m:Materiau)-[r:DEGRADER]-(p:Pathologie) RETURN m.designation AS Materials, collect(p.designation) AS Pathologies ORDER BY m.designation	*Query (d)* MATCH(m:Materiau)-[r:DEGRADER]->(p:Pathologie) RETURN m.designation AS Material, COUNT(r) AS NB_Pathologies ORDER BY NB_Pathologies DESC
Query (e) Match (m:Monument) RETURN m limit 25	*Query (f)* MATCH(nodes)-[relatedTo]-(:Materiau{designation:"Terre"}) RETURN nodes, type(relatedTo)
Query (g) MATCH(BuildingElt:Ouvrage)-[:EST_FABRIQUER]->(n:FabriquerRelation), (Pdt:Produit)-[:FABRIQUER_PAR]->(n), (T:Technique)-[:FABRIQUER_SELON]->(n), (Pdt)-[:COMPOSER]-(m:Materiau) Return BuildingElt, n,Pdt,T, m	

Table 11: Cypher queries translating the architects requirements

These requirements can require aggregations like requirements (a), (b) and (d). Other requirements allow us to collect information that was scattered among different sources like requirements (a) and (c). Other requirements allow to identify hidden connections between entities, like requirement (g). Finally, requirements can be exploratory aiming to return all the connected information of a given node like requirement (f). Table 11 shows the Cypher queries used that translate the architects' requirements cited.

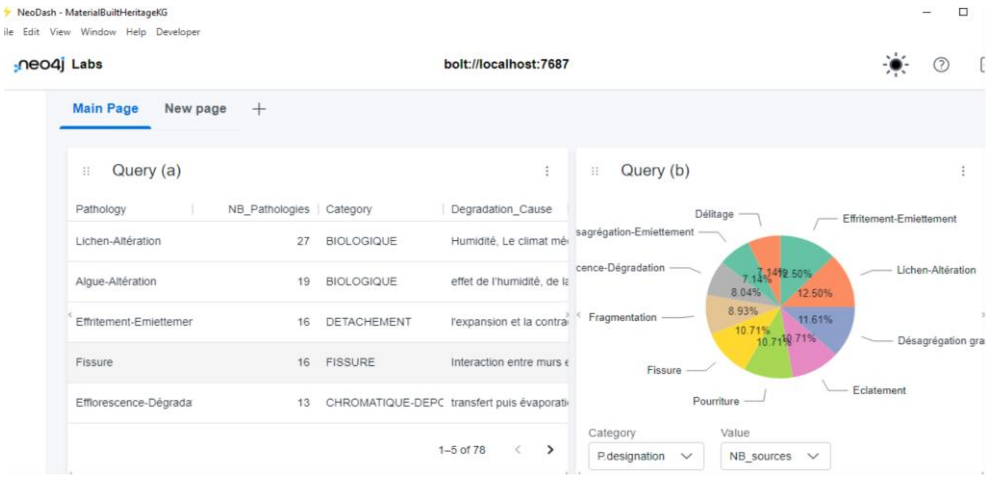

Figure 14: Results of Cypher queries (a) and (b) using NeoDash (Neo4J)

Figures 14,15, 16 and 17 present the results obtained for the Cypher queries related to requirements (a) to (f) and displayed using NeoDash which is a Neo4J API that allows to show the results of the query using different visualizations, in the form of a dashboard. Figure 18 illustrates the results of query (g) using Neo4J Browser which is the classic browser for running Cypher queries.[15] According to the nature of the query, we have chosen the adequate visualization type for displaying the results.

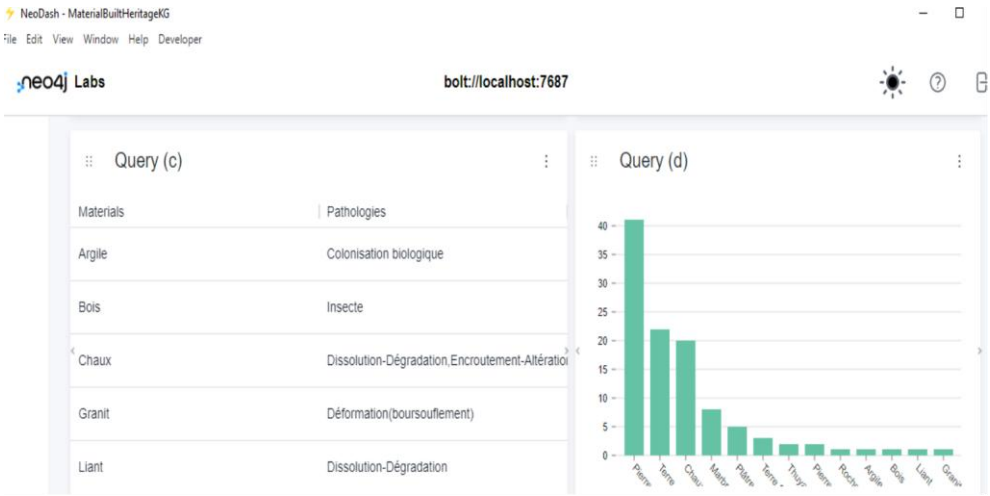

Figure 15: Results of Cypher queries (c) and (d) using NeoDash (Neo4J)

[15] Note that the concepts are displayed in French language in the cited figures because the KG is in French.

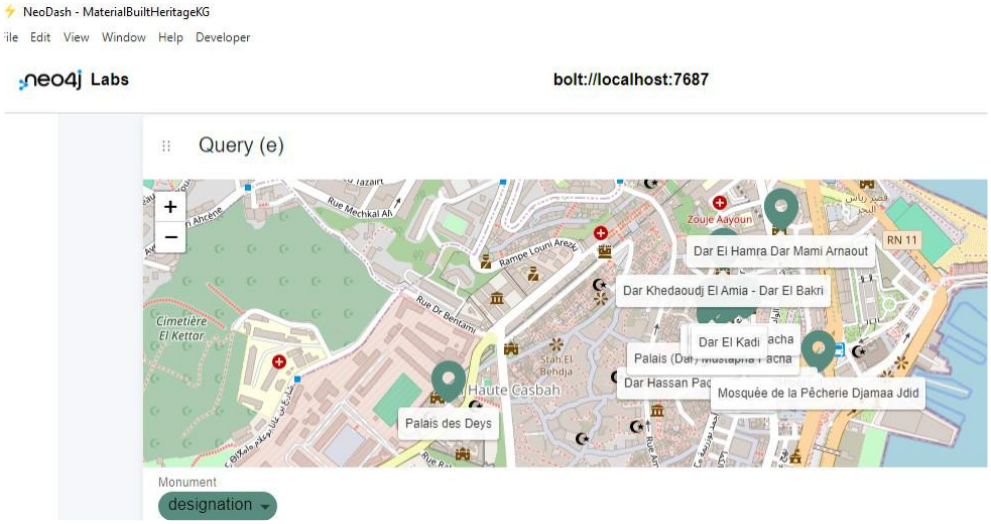

Figure 16: Results of Cypher queries (e) using NeoDash (Neo4J)

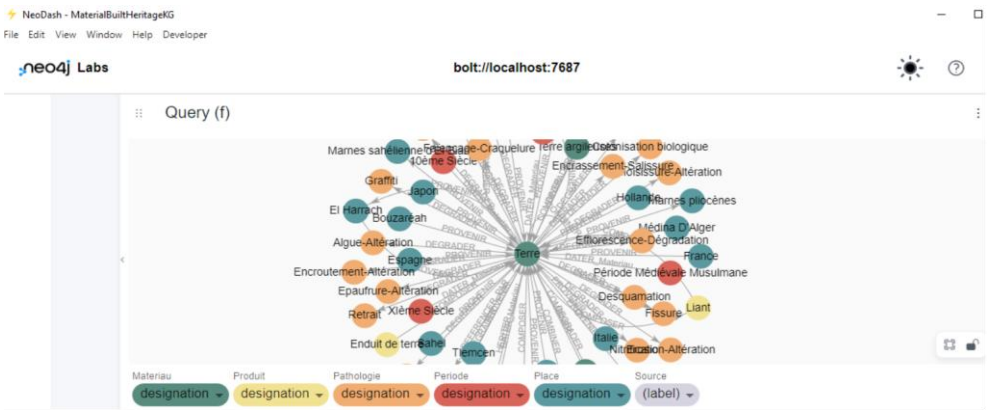

Figure 17: Results of Cypher queries (f) using NeoDash (Neo4J)

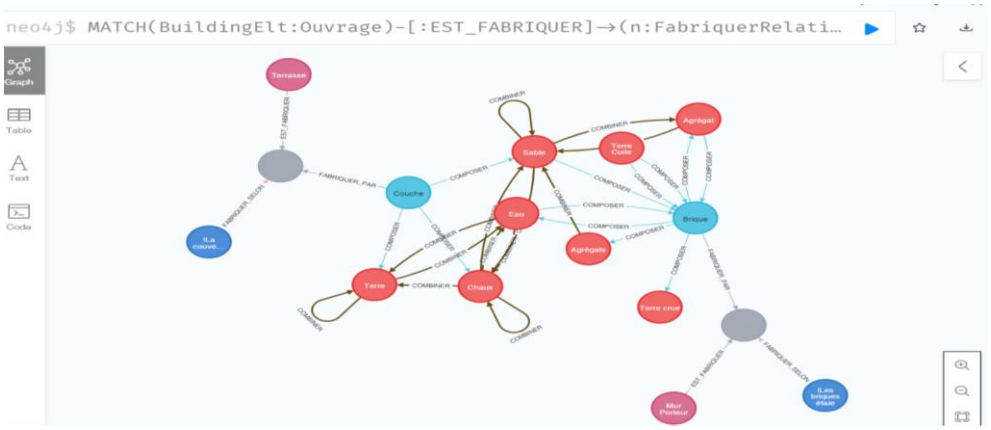

Figure 18: Results of Cypher query (g) using Neo4J Browser

6. Conclusion

Knowledge graph construction poses different challenges especially when dealing with specific and vertical domains. In our context "Ancient Materials of the Algerian built heritage", the sparsity of the input sources, their language, their reduced amount and the presence of long-tail and technical entities, posed various difficulties which pushed us to test different intelligent models that proved efficiency. Our aim is that the KG serves as a trustworthy basis to different applications like material platforms, building virtual education tools, chatbots, and many other AI advanced systems. The KG proposed is a first block that should be maintained and validated by different heritage stakeholders profiles. It can also be enriched with the knowledge related to other regions and other periods. The flexibility of the KG allows its broader extension to other KGs sharing similar characteristics eg. KGs dedicated to the heritage of Mediterranean regions (Carriero et al, 2021). Furthermore, the complexity of information extraction in this highly specialized domain requires a dedicated benchmarking framework to evaluate evolving LLMs' performance and suitability for this domain. Finally, connecting the current KG to physical materials will hopefully enrich the knowledge and documentation related to the usage, the management and the restoration of these monuments, and will constitute a wealthy resource for academic/teaching projects as well as for heritage restoration/conservation projects.

Acknowledgement

We thank Baiba Loubna for her contribution during the development of the first prototype of the study.

References

Abu-Salih, Bilal. 2021. "Domain-specific knowledge graphs: A survey." *Journal of Network and Computer Applications* 185 : 103076.

Acierno, Marta, Stefano Cursi, Davide Simeone, and Donatella Fiorani. 2017. "Architectural heritage knowledge modelling: An ontology-based framework for conservation process." *Journal of Cultural Heritage* 24 (mars): 124-133. https://doi.org/10.1016/j.culher.2016.09.010.

Ackoff, Russell L. 1989. "From data to wisdom." *Journal of applied systems analysis* 16, no. 1 : 3-9.

Allaouchiche, Sara, and Sabrina Kacher. 2021. "Digital Material Libraries: overview and application case." *Proc. of the Conference CIB W78*. Vol. 2021.

Allaouchiche, Sara, Sabrina Kacher, and David Sanz-Arauz. 2025. "Documenting Ancient Materials for a Digital Materials Library Project. The Case of Three World Heritage Sites in Algeria." *International Journal of Architectural Heritage* : 1-24. https://doi.org/10.1080/15583058.2025.2495960

Allaouchiche, Sara, Sabrina Kacher, Selma Khouri, Houda Oufaida, Yassmin Bachikh, and Izdihar Belmrabet. "Organizational approach for a digital materials platform applied to Algeria's built heritage." *Journal of Cultural Heritage* 76 (2025): 204-217.

Allaouchiche, Sara., and Sabrina Kacher, S. 2025. "A Digital Materials Library Applied to the Algerian Built Heritage: Identification of Common Materials". In *Conservation of Architectural and Urban Heritage CAH 2023*. (pp. 215-229). Advances in Science, Technology & Innovation. Springer, Cham. https://doi.org/10.1007/978-3-031-71145-9_16

Amin, Alia, Jacco Van Ossenbruggen, Lynda Hardman, and Annelies van Nispen. 2008. "Understanding cultural heritage experts' information seeking needs." In P*roceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pp. 39-47.

Amrani, Racha, Sabrina Kacher, Selma Khouri, Houda Oufaida, Safia Ouahab, and Mouna Cherrad. 2024. "Proposal of a knowledge capitalization process to construct Eco-Diars: A Knowledge-driven platform applied to traditional Algerian domestic architecture." *ACM Journal on Computing and Cultural Heritage* 17, no. 1: 1-28.

Arora, Simran, Brandon Yang, Sabri Eyuboglu, Avanika Narayan, Andrew Hojel, Immanuel Trummer, and Christopher Ré. 2023. "Language models enable simple systems for generating structured views of heterogeneous data lakes." *arXiv preprint* arXiv:2304.09433.

Attari, Nassereddine, Souad Laoues, and Soumia Chabane. 2019. "Characterization of an Archaeological Mortar from the Ottoman Period in Algeria." *International Journal of Materials, Mechanics, and Manufacturing* 7, no. 3: 154-159.

Brank, Janez, Gregor Leban, and Marko Grobelnik. 2017. "Annotating documents with relevant wikipedia concepts." *Proceedings of SiKDD 472*.

Carriero, Valentina Anita, Aldo Gangemi, Maria Letizia Mancinelli, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 2021. "Pattern-based design applied to cultural heritage knowledge graphs: ArCo: The knowledge graph of Italian Cultural Heritage." *Semantic Web* 12, no. 2: 313-357. https://doi.org/10.3233/SW-200422

Casillo, Mario, Massimo De Santo, Rosalba Mosca, and Domenico Santaniello. 2023. "Sharing the knowledge: exploring cultural heritage through an ontology-based platform." *Journal of Ambient Intelligence and Humanized Computing* 14, no. 9: 12317-12327.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen et al. 2024. "A survey on evaluation of large language models." *ACM transactions on intelligent systems and technology* 15, no. 3: 1-45.

Chergui, Samia. 2015. "Le lexique des patrimoines architecturaux dans la Régence d'Alger, support de thésaurus?." *Patrim. Maghreb À L'ère Numér* : 16.

Correia, Maria J., and António Santos Silva. 2023. "DB-HERITAGE Building Materials Data Aggregation in ARIADNE–challenges and opportunities." *Internet Archaeology* 64.

Cursi, Stefano, Letizia Martinelli, Nicolò Paraciani, Filippo Calcerano, and Elena Gigliarelli. 2022. "Linking external knowledge to heritage BIM." *Automation in Construction* 141: 104444.

Delegou, E. T., E. Tsilimantou, E. Oikonomopoulou, J. Sayas, C. Ioannidis, and A. Moropoulou. 2013. "Mapping of building materials and consevation interventions using GIS: The case of Sarantapicho Acropolis and Erimokastro Acropolis in Rhodes." *International journal of heritage in the digital era* 2, no. 4: 631-653. doi:10.1260/2047-4970.2.4.631.

Detroja, Kartik, C. K. Bhensdadia, and Brijesh S. Bhatt. 2023. "A survey on relation extraction." *Intelligent Systems with Applications* 19 : 200244.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "Bert: Pre-training of deep bidirectional transformers for language understanding." In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, volume 1 (long and short papers), pp. 4171-4186.

Djerad, Mohamed Sofiane, Khedidja Boufenara, Jacques Des Courtils, Nadia Cantin, and Yannick Lefrais. 2022. "Multianalytical characterisation and provenance investigation of natural pozzolana in Roman lime mortars from the archaeological site of Hippo Regius (Algeria)." *Mediterranean Archaeology and Archaeometry* 22, no. 3: 231-248. doi: 10.5281/zenodo.7412076.

Dou, Jinhua, Jingyan Qin, Zanxia Jin, and Zhuang Li. 2018. "Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage." *Journal of Visual Languages & Computing* 48 : 19-28.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines et al. 2021. "Beyond english-centric multilingual machine translation." *Journal of Machine Learning Research* 22, no. 107 : 1-48.

Fernández-López, Mariano, Asunción Gómez-Pérez, and Natalia Juristo Juzgado. 1997. "Methontology: from ontological art towards ontological engineering.". *Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series*, American Asociation for Artificial Intelligence.

Frangipane, Anna. 2010. "Working for an electronic database of historical stone resources in Friuli-Venezia Giulia (Italy)." *Geological Society London Special Publications*. 333(1), 197-209. https://doi.org/10.1144/SP333.19

Gheris, Abderrahim. 2023. "New dating approach based on the petrographical, mineralogical and chemical characterization of ancient lime mortar: Case study of the archaeological site of Hippo, Annaba city, Algeria." *Heritage Science* 11, no. 1 : 103. https://doi.org/10.1186/s40494-023-00942-3

Giagnolini, Lucia, Schimmenti, Andrea, Bonora, Paolo, tomasi, francesca. 2025. "Expliciting Contexts: Semantic Knowledge Extraction from Traditional Archival Descriptions". *Umanistica Digitale* no 20, p. 115-144. DOI: http://doi.org/10.6092/issn.2532-8816/21229

Guo, Xu, and Yiqiang Chen. 2024. "Generative AI for synthetic data generation: Methods, challenges and the future." *arXiv preprint arXiv:2403.04190*.

Gupta, Tanishq, Mohd Zaki, NM Anoop Krishnan, and Mausam. 2022. "MatSciBERT: A materials domain language model for text mining and information extraction." *npj Computational Materials* 8, no. 1: 102.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane et al. 2021. "Knowledge graphs." *ACM Computing Surveys (Csur)* 54, no. 4 : 1-37.

Hofer, Marvin, Daniel Obraczka, Alieh Saeedi, Hanna Köpcke, and Erhard Rahm. 2023. "Construction of knowledge graphs: State and challenges." *arXiv preprint* arXiv:2302.11509 .

Huang, Beijia, Xiaofeng Gao, Xiaozhen Xu, Jialing Song, Yong Geng, Joseph Sarkis, Tomer Fishman, Harnwei Kua, and Jun Nakatani. 2020. "A life cycle thinking framework to mitigate the environmental impact of building materials." *One Earth* 3, no. 5: 564-573. https://doi.org/10.1016/j.oneear.2020.10.010

Hyslop, Ewan, Andrew McMillan, Don Cameron, Alick Leslie, and Graham Lott. 2010. "Building stone databases in the UK: A practical resource for conservation." *Engineering Geology* 115, no. 3-4 : 143-148.

Jäkel, Jan-Iwo, Eva Heinlein, Hendrik Morgenstern, Hongjo Kim, and Katharina Klemt-Albert. 2025. "System-and Data-integrated linking of digital 3D models of existing bridge structures with Knowledge Graphs of non-destructive diagnostic methods." *J. Inf. Technol. Constr.* 30: 603-630.

Jehangir, Basra, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. "A survey on named entity recognition—datasets, tools, and methodologies." *Natural Language Processing Journal* 3 : 100017.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. "Survey of hallucination in natural language generation." *ACM computing surveys* 55, no. 12 : 1-38.

Kaltenegger, Julia, Kirstine Meyer Frandsen, and Ekaterina Petrova. 2024. "A data management perspective on building material classification: A systematic review." *Journal of Building Engineering* 92 : 109172.

Kampfová, Hana, and Richard Přikryl. 2010. "Electronic database of historical natural stones of the Czech Republic: structuring field and laboratory data." *Geological Society, London, Special Publications,* 333(1), pp.211-217. https://doi.org/10.1144/SP333.20

Kharfi, Fayçal, Lahcen Boudraa, Imene Benabdelghani, and Mahfoud Bououden. 2019. "TL dating and XRF clay provenance analysis of ancient brick at Cuicul Roman city, Algeria." *Journal of Radioanalytical and Nuclear Chemistry* 320, no. 2: 395-403. https://doi.org/10.1007/s10967-019-06491-z

Khouri, Selma, Houda Oufaida, Racha Amrani, Sabrina Kacher, Safia Ouahab, and Mouna Cherrad. 2023. "Knowledge base construction for the semantic management of environment-enriched built heritage: The case of Algerian traditional houses architecture." *Journal of Cultural Heritage* 63 : 217-229.

Korro, Jaione, José Manuel Valle-Melón, Ainara Zornoza-Indart, and Alvaro Rodriguez Miranda. 2025. "New perspectives and usual challenges: present technologies for document management in architectural heritage conservation-restoration works." *ACM Journal on Computing and Cultural Heritage* 18, no. 2 : 1-28.

Li, Junyi, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. "Pre-trained language models for text generation: A survey." *ACM Computing Surveys* 56, no. 9 : 1-39.

Louis Antoine. 2024. BelGPT-2: a GPT-2 model pre-trained on French corpora. *In:* https://github.com/antoilouis/belgpt2

Marchand, Erwan, Michel Gagnon, et Amal Zouaq. 2020. "Extraction of a Knowledge Graph from French Cultural Heritage Documents". In *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, edited by* Ladjel Bellatreche, Mária Bieliková, Omar Boussaïd, Barbara Catania, Jérôme Darmont, Elena Demidova, Fabien Duchateau, et al., 1260:23-35. Communications in Computer and Information Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-55814-7_2.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. "CamemBERT: a tasty French language model." *arXiv preprint* arXiv:1911.03894.

Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. "Large language models: A survey." *arXiv preprint arXiv:2402.06196* (2024).

Napolitano, Rebecca, Catherine Jennings, Sophia Feist, Abigail Rettew, Grace Sommers, Hannah Smagh, Benjamin Hicks, et Branko Glisic. 2019. "Tool Development for Digital Reconstruction: A Framework for a Database of Historic Roman Construction Materials". *Journal of Cultural Heritage* 40 (novembre): 113-23. https://doi.org/10.1016/j.culher.2019.05.007.

Néroulidis, Ariane, Thomas Pouyet, Sarah Tournon, Miled Rousset, Marco Callieri, Adeline Manuel, Violette Abergel et al. 2024. "A digital platform for the centralization and long-term preservation of multidisciplinary scientific data belonging to the Notre Dame de Paris scientific action." *Journal of Cultural Heritage* 65 : 210-220.

Pan, Shirui, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, et Xindong Wu. 2024. "Unifying Large Language Models and Knowledge Graphs: A Roadmap". *IEEE Transactions on Knowledge and Data Engineering* 36 (7): 3580-99. https://doi.org/10.1109/TKDE.2024.3352100.

Pauwels, Pieter, Rens Bod, Danilo Di Mascio, et Ronald De Meyer. 2013. "Integrating Building Information Modelling and Semantic Web Technologies for the Management of Built Heritage Information". In 2013 *Digital Heritage International Congress (DigitalHeritage),* 481-88. Marseille, France: IEEE. https://doi.org/10.1109/DigitalHeritage.2013.6743787.

Perucchetti, L., P. Bray, A. Felicetti, V. Sainsbury, P. Howarth, M. K. Saunders, P. Hommel, and M. Pollard. 2021. "FLAME-D database: An integrated system for the Study of archaeometallurgy." *Archaeometry 63*, no. 3: 651-667.

Pierra, Guy. 2008. "Context representation in domain ontologies and its use for semantic integration of data." In *Journal on data semantics* X, pp. 174-211. Berlin, Heidelberg: Springer Berlin Heidelberg.

Polak, Maciej P., Shrey Modi, Anna Latosinska, Jinming Zhang, Ching-Wen Wang, Shaonan Wang, Ayan Deep Hazra, and Dane Morgan. 2024. "Flexible, model-agnostic method for materials data extraction from text using general purpose language models." *Digital Discovery* 3, no. 6 : 1221-1235.

Peng, Ciyuan, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. "Knowledge graphs: Opportunities and challenges." *Artificial intelligence review* 56, no. 11: 13071-13102.

Quattrini, Ramona, Roberto Pierdicca, et Christian Morbidoni. 2017. "Knowledge-Based Data Enrichment for HBIM: Exploring High-Quality Models Using the Semantic-Web". *Journal of Cultural Heritage* 28 : 129-39. https://doi.org/10.1016/j.culher.2017.05.004.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of machine learning research* 21, no. 140 : 1-67.

Rezkallah, Younès, and Ramdane Marmi. 2018. "Building materials and the ancient quarries at Thamugadi (East of Algeria), case study: Sandstone And Limestone." In *The XI International Conference of ASMOSIA*, pp. 673-682. University of Split, Arts Academy in Split. https://doi.org/10.31534/XI.asmosia.2015/05.07

Rezkallah, Younes. 2020. "Le SIG des fouilles de l'antique Thamugadi: premiers résultats." *AOURAS 10* : 355-369. https://hal.science/hal-03219114/.

Roth, Hannah Rae, Meghan Lewis, and Liane Hancock. 2021. "The green building materials manual: a reference to environmentally sustainable initiatives and evaluation methods." *Springer Nature*.

Sahbi, Ava, Santini Cristian, Hussein Hassan, Neubert Kerstin, De Leo Vincenzo, Duan Xuemin, Celino Irene. 2025. "Kuba: Knowledge Graph Generation through User-Based Approach". In Ahmad, Raia Abu, Reham Alharbi, Roberto Barile, Martin Böckling, Francisco Bolanos, Sara Bonfitto, Oleksandra Bruns et al. "Semantic Web and Creative AI--A Technical Report from ISWS 2023." *arXiv preprint* arXiv:2501.18542.

Schauppenlehner, Thomas, Renate Eder, Kim Ressar, Hubert Feiglstorfer, Roland Meingast, and Franz Ottner. 2021. "A Citizen Science approach to build a knowledge base and cadastre on earth buildings in the Weinviertel region, Austria." *Heritage 4*, no. 1 : 125-139.

Schrader, Timo Pierre, Matteo Finco, Stefan Grünewald, Felix Hildebrand, and Annemarie Friedrich. 2023. "Mulms: A multi-layer annotated text corpus for information extraction in the materials science domain." *arXiv preprint* arXiv:2310.15569 .

Simeone, Davide, Stefano Cursi, et Marta Acierno. 2019. "BIM Semantic-Enrichment for Built Heritage Representation". *Automation in Construction* 97 (janvier): 122-137. https://doi.org/10.1016/j.autcon.2018.11.004.

Tamašauskaitė, Gytė, et Paul Groth. 2022. "Defining a Knowledge Graph Development Process Through a Systematic Review". *ACM Transactions on Software Engineering and Methodology*, avril, 3522586, 1–40. https://doi.org/10.1145/3522586.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière et al. 2023. "Llama: Open and efficient foundation language models." *arXiv preprint* arXiv:2302.13971.

Trojahn, Cassia, Renata Vieira, Daniela Schmidt, Adam Pease, and Giancarlo Guizzardi. 2022. "Foundational ontologies meet ontology matching: A survey." *Semantic Web 13, no. 4*: 685-704.

Turmel, Aurélie, Gilles Fronteau, Laurent Chalumeau, J-P. Deroin, Stéphanie Eyssautier-Chuine, Céline Thomachot-Schneider, Tim De Kock, Veerle Cnudde, and Vincent Barbin. 2016. "GIS-based variability of building materials towards the Île-de-France cuesta (Paris Basin,

France): inventory, distribution, uses and relationship with the environment." *Geological Society of London Special Publications 416, no. 1 (2016): 113-131*. doi: 10.1144/SP416.16.

Wang, Xiaoguang, Wanli Chang, et Xu Tan. 2020. "Representing and Linking Dunhuang Cultural Heritage Information Resources Using Knowledge Graph". *Knowledge Organization* 47 (7): 604-15. https://doi.org/10.5771/0943-7444-2020-7-604.

Weikum, Gerhard. 2021. "Knowledge graphs 2021: A data odyssey." *Proceedings of the VLDB Endowment* 14, no. 12 : 3233-3238. https://doi.org/10.14778/3476311.3476393.

Weikum, Gerhard, Luna Dong, Simon Razniewski, et Fabian Suchanek. 2020. "Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases". *Trends® Databases* 10 (2-4), 108–490.

Wittenberg, Antje, and Daniel de Oliveira. 2023. "Critical raw material resource potentials in Europe." *Materials Proceedings* 15, no. 1 : 24.

Xu, Derong, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. "Large language models for generative information extraction: A survey." *Frontiers of Computer Science* 18, no. 6 : 186357.

Yu, Junjie, Xing Wang, and Wenliang Chen. 2024. "Reliable data generation and selection for low-resource relation extraction." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, pp. 19440-19448.

Zhong, Lingfeng, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2023. "A comprehensive survey on automatic knowledge graph construction." *ACM Computing Surveys* 56, no. 4 : 1-62.