Generation of Multi-Modal Narratives of Cultural Objects from Knowledge Graphs and LLMs

Ishak Riali[1,2,*], Raphaël Gerssen[1], Elizabeth Rodríguez Estrada[1], Gabriel Spautz Vieira[1] and Martin Berger[1]

[1]*Faculty of Archaeology, Leiden University, The Netherlands*
[2]*LRDSI Laboratory, University of Blida 1, Algeria*

Abstract

This work proposes an automated system for generating contextualized narratives from structured museum data. In a first step, a tabular database containing object metadata (Identifier, Name, creator, culture, dating, material, etc.) is transformed into a knowledge graph based on the CIDOC CRM (Conceptual Reference Model) standard. This graph not only represents the objects themselves, but also models the events associated with them (creation, collection, acquisition, etc.), ensuring a rich and temporally anchored semantic understanding. In a second phase, targeted Cypher queries query this graph to extract contextual information related to each object. This information is then integrated into a structured prompt, sent to a pre-trained language model (LLM) via an API, to automatically generate a fluid, coherent, and historically informed narrative. Finally, the produced text is converted into an audio file, providing multi-modal playback accessible to a wide audience. The system has been empirically validated by domain experts who confirmed the accuracy and relevance of the generated narratives, as well as by members of the general museum public, who confirmed the texts' accessibility.

Keywords

LLMs, Knowledge graphs, RAG system, CIDOC CRM, Cultural Heritage

1. Introduction

Museums play a fundamental role in the preservation, transmission, and promotion of cultural heritage. However, the general public's access to the semantic and historical richness of museum collections is limited by the quality of documentation of objects in museum databases. Short, technical, and jargon-full database entries are poorly suited to non-specialist audiences or those with limited reading skills. This lack of contextualization reduces the potential educational and emotional impact of objects and (online) displays, particularly for younger generations or audiences with learning disabilities. At the same time, curators in archaeological and ethnographic museums are often responsible for thousands of objects. They are not equipped to manually update the contextual information of every individual object in their collection. As a result, undercontextualized and poorly documented objects tend to linger in depots away from the public eye. The use of advanced technologies such as artificial intelligence offers new ways to activate dormant collections, streamline curatorial work, and provide better access to museum collection for a broader audience [1, 2].

Among these technologies, LLMs stand out for their ability to automatically generate rich and coherent narrative texts that are adapted to diverse contexts [3]. However, to produce a historically relevant narrative faithful to museum knowledge, these models require a reliable, structured, and searchable information base. With this in mind, we designed a knowledge graph compliant with the CIDOC CRM standard, the international standard for modeling cultural information. This ontological model allows for the semantic and interoperable representation of key entities related to museum objects (creators, historical events, locations, periods, collections, etc.), while facilitating the inference and retrieval of complex information [4, 5].

SemDH 2026: Third International Workshop of Semantic Digital Humanities. Co-located with ESWC 2026, May 10, 2026, Dubrovnik, Croatia

*Corresponding author.
i.riali@arch.leidenuniv.nl (I. Riali); r.c.r.gerssen@hum.leidenuniv.nl (R. Gerssen); e.rodriguez.estrada@arch.leidenuniv.nl (E. R. Estrada); g.spautz.vieira@arch.leidenuniv.nl (G. S. Vieira); m.e.berger@arch.leidenuniv.nl (M. Berger)

The data source used to construct this graph consists primarily of tabular data extracted from records of the Latin American collections of the Museo delle Culture di Milano (MUDEC). The MUDEC houses a collection of around 7,000 artifacts from Africa, Asia, America, South-East Asia and Oceania. These collections include both ethnographic and archaeological material from different periods, regions, and cultural traditions. The diversity of objects in this collection provides an ideal ground for experimenting with the automatic generation of contextualized cultural narratives.

Retrieval-Augmented Generation (RAG) is an architecture that augments LLMs by enabling them to retrieve factual information from external sources before generating a response. This approach is recognized as a major success in generative AI for its ability to provide reliable and up-to-date context, reducing "hallucinations" and ensuring accuracy. Reducing hallucinations is fundamental in the context of museums, which aim to provide correct and truthful information about the objects that they care for. RAG is at the heart of many recent applications (such as [6, 7, 8, 9]), confirming its essential role in the practical integration of LLMs in enterprise environments.

In this work, we propose a comprehensive system for the multi-modal generation of narratives around museum objects, based on a Graph-RAG architecture. The system operates in three main stages: (1) Extraction of the historical and cultural context of a given object from the graph structured according to the CIDOC CRM, and (2) The implementation of a narrative generation process guided by a pre-trained LLM using structured prompts, exploiting the model's ability to produce fluent natural language while maintaining some control over the content. The quality of the narrative stories generated was evaluated by specialists in the field of archaeology, as well as by members of the general public. (3) Finally, this text is transformed into audio content using a text-to-speech module, enabling smooth and accessible oral reproduction for all audiences, including visually impaired visitors. Providing automated narratives about objects enriches the museum experience with dynamic, historically grounded, and customizable stories, in both online and offline settings. It offers a new way of "making objects speak", no longer solely through metadata, but via engaging stories automatically constructed from the information available in the museum database.

The remainder of the paper is organized as follows: the next section presents the related work and the theoretical foundations on which our approach is based. Then, we describe in detail the architecture of the proposed system, as well as the different data processing steps. We then present the evaluation approach conducted with experts in the field. Finally, we conclude with a discussion on the contributions of our work and future perspectives.

2. Related work

Several recent works have explored the integration of LLMs in the field of cultural heritage, from perspectives close to ours.

M. Reusens et al. [10] evaluated the use of the Llama model to automatically generate keywords from object descriptions in museum collections, to enrich metadata and improve collection accessibility. The approach proved to be effective in making the data more easily exploitable by researchers and the general public.

M. Mountantonakis et al. [11] proposed a system that combines the capabilities of ChatGPT with hundreds of RDF knowledge graphs via the LODsyndesis[1] to enable advanced services such as entity recognition and linking, data enrichment, or fact validation, with the aim of improving the discoverability of cultural resources.

Furthermore, G. Trichopoulos et al.[12] were interested in adapting GPT-4 as a personalized recommendation system for museum visitors, taking into account the context (location, date, temporary events) in order to improve the user experience.

The FolkRAG system [13] aimed at improving access to archival collections by integrating retrieval techniques such as sentence window retrieval, hypothetical document embedding (HyDE), and re-ranking strategies. It constructs a vector store using fine-grained text chunks (250 characters) to handle

[1] https://demos.isl.ics.forth.gr/lodsyndesis/

and retrieve documents.

C. C. Chang et al. [14] demonstrated that combining RAGs and LLMs represents a major advancement for the translation of minority languages, improving not only technical accuracy but also cultural fidelity. It also provides a methodological and ethical framework for supporting underrepresented linguistic communities.

Moreover, I. Vasic et al[15] proposed a local application integrating GPT-4o to guide users during 3D virtual tours of a museum, generating personalized responses to their queries in natural language, thus offering an immersive and interactive solution, particularly suitable for small museums with limited resources.

Z. Liang et al. [16] presented a multimodal knowledge graph dedicated to the indigenous Yanyuwa language, integrating textual, visual, and audiovisual data to preserve and contextualize linguistic and cultural knowledge. The authors use this graph to generate visual narratives in the form of data comics, making complex knowledge structures accessible to non-technical audiences. The storytelling process is guided by the graph's structure and human expertise, without resorting to generative models for the automatic production of textual content.

X. Wang et al. [17] explored the transition from structured data to narrative by identifying narrative patterns based on a knowledge graph dedicated to ethnic costumes to analyze how designers navigate through nodes and relationships to extract coherent scenarios along three axes: formal aesthetics, folklore, and regional context. Although their approach relies on manual selection and interpretation by designers, our study automates this synthesis process by using the generative capabilities of LLMs to produce complex multimodal narratives from similar data structures.

In parallel, I. Vasic et al. [18] analyzed the synergy between Knowledge Graphs and language models in the context of virtual museums. By comparing purely semantic approaches with solutions based solely on LLMs, they conclude that these technologies are complementary partners rather than competitors. Their study demonstrates that integrating the CIDOC-CRM ontology with the processing capabilities of LLMs mitigates the risk of hallucinations while facilitating access to complex data. Our research aligns with this hybrid vision by specifically using LLMs as a generation engine to transform these rigid data structures into fluid narratives.

Similarly, C. Palma et al.[19] developed a neurosymbolic architecture that transforms cultural knowledge graphs into logical narratives via a modular pipeline. their work prioritizes serendipity and human interaction to reveal surprising connections between heritage objects before they are narrated by a language model. Although this study focuses on the semantic exploration and narrative potential of structure, our work complements this approach by harnessing the power of LLMs to achieve automatic generation, transforming these relational discoveries into narrative experiences enriched by other media.

Recently, Blin [20] investigated the contribution of knowledge graphs to narrative understanding through a hybrid approach that combines symbolic rigor and neural methods. This work notably introduces ChronoGrapher, an automated system for constructing event-centered graphs that has been successfully applied to the analysis of historical narratives and hypothesis generation in the social sciences.

Works such as M. Reusens et al. [10] and G. Trichopoulos et al. [12] exploit LLMs directly on textual or relational data, without anchoring them in a formally structured semantic model. This limits the factual consistency and interoperability of the generated content. This is a particular problematic shortcoming in a museum context where historical rigor is non-negotiable. FolkRAG [13] adopts an RAG strategy based on text fragments, sacrificing the relational richness essential for event-oriented cultural reasoning, which reduces the system's ability to capture complex links of provenance or temporal contextualization. Although M. Mountantonakis et al. [11] combines LLMs and RDF knowledge graphs for enrichment and validation, this approach remains upstream of narrative generation and does not offer mediation accessible to the general public. C. C. Chang et al.[14] employs a RAG architecture combined with LLMs, but its focus on translating minority languages makes it a methodologically distinct contribution, with limited transferability to ontology-based heritage narrative generation. X. Wang et al. [17] explores the generation of narrative scripts from heritage knowledge graphs, but the reliance on manual selection by designers introduces a scalability bottleneck that compromises any large-scale institutional deployment. I. Vasic et al. [18] rigorously demonstrates the complementarity between knowledge graphs and LLMs in a museum context, but remains limited to a comparative analysis without proposing an operational system. The approaches of I. Vasic et al [15] and C. Palma et al.[19] support guided narrative experiences, but their reliance on user queries or manual curation makes them difficult to scale and unsuitable for collections of thousands of poorly documented objects. Furthermore, Z. Liang et al.[16] illustrates the potential of storytelling from multi-modal graphs, but the lack of generative language models significantly limits automation and the richness of the resulting narratives. Finally, while Blin [20] advances the field with a neuro-symbolic framework and the ChronoGrapher system for automated event-centric graph construction, the work remains focused on narrative understanding and analytical discourse rather than the end-to-end, multimodal dissemination required for inclusive museum mediation.

Our system uniquely integrates four simultaneous dimensions: (1) a CIDOC CRM-compliant knowledge graph as a structured foundation, ensuring ontological rigor and semantic interoperability; (2) a Graph-RAG architecture that uses relational paths—origin, temporal, spatial, and cultural dimensions rather than simple textual fragments; (3) automated and scalable narrative generation, evaluated against five state-of-the-art LLMs; and (4) multimodal dissemination via text-to-speech conversion, extending accessibility to visually impaired audiences. Furthermore, unlike most previous work, our system was deployed and evaluated on a real museum collection (MUDEC, 7,000 artifacts), using a multi-perspective evaluation framework involving domain experts, the general public, and LLMs used as automated judges. This combination of ontological grounding, automated generation, empirical validation, and operational deployment constitutes the core of the novelty of our contribution.

3. Knowledge graphs and CIDOC CRM

This section provides an overview of some background knowledge that will be used in the paper. It includes Knowledge Graphs and CIDOC-CRM.

Knowledge graphs are a framework to model data that represent entities (such as: Events, Objects, etc.) and capture relationships between them. They represent knowledge in a way that enables both humans and machines to reason about it [21].

The CIDOC CRM [22] is a high-level ontology for cultural heritage and museum documentation. It provides a shared semantic framework with the aim of integrating heterogeneous data from various sources. The CIDOC CRM provides a set of classes and predicates to model a given reality. The CIDOC CRM is an event-based model that focuses on events rather than objects. It models history as a set of events where people and things interact. It Links things, people, ideas, time, and places through shared events.

The ontology includes: **Actors** (E39) participate in **Temporal Entities** (E2), linking people or organizations to events. **Temporal Entities** (E2) are bounded by **Time-Spans** (E52), establishing the time frame of events. **Time-Spans** (E52) indicate the duration of events, helping to define the historical context. **Physical Things** (E18) and **Conceptual Objects** (E28) can be located at **Places** (E53) and are part of events. **Appellations** (E41) give a unique identifier based on convention or context to the entities, ensuring accurate reference. **Types** (E55) classify entities, making it easier to organize and to query data.

Figure 2 presents an example of a knowledge graph built following the CIDOC-CRM conceptual model, illustrating how cultural entities, events, and their relationships can be represented and connected within a standardized ontology for cultural heritage[23].

4. Proposed Methodology

This section presents the system we developed for the automatic generation of narrative stories around museum objects, based on a Graph-RAG architecture. Figure 2 illustrates the overall architecture of the

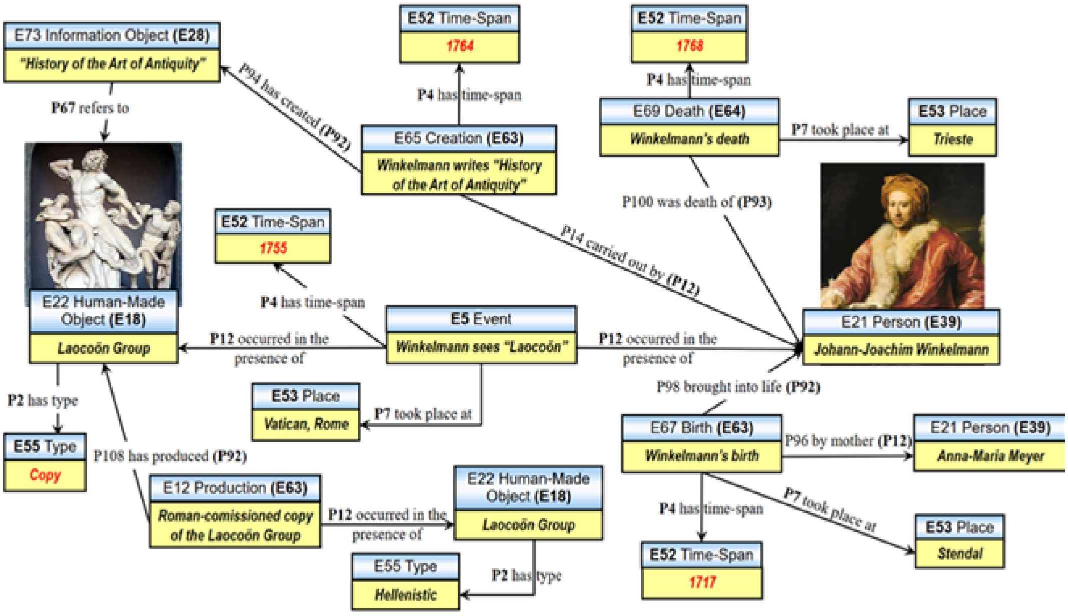

Figure 1: CIDOC CRM model overview. Adapted from the official CIDOC CRM version 7.3 specification, © CIDOC CRM Special Interest Group, licensed under CC BY 4.0.

proposed system, highlighting the relationship between the knowledge graph, the retrieval module, and the generation of narratives.

The methodology follows a structured *Knowledge-Graph-Augmented* pipeline. The process is initiated when a user submits a request for an object story. A Retriever module then executes predefined queries against a Neo4j knowledge graph to extract precise historical data.

This retrieval phase produces a semantic description of the object, including its provenance, timeline, and associated historical entities. This structured context is then fed into an LLM, which synthesizes the factual data into a coherent and historically grounded narrative.

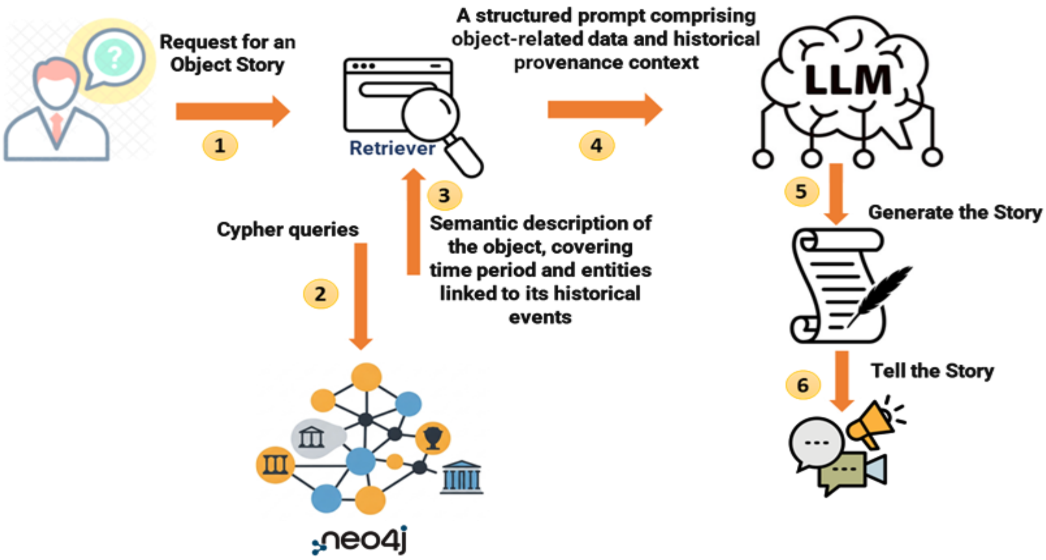

Figure 2: The proposed Graph-RAG architecture.

5. MUDEC KG Creation

In this section, we present our proposed methodology for building a CIDOC CRM-Based Knowledge Graph for the MUDEC collection. Our methodology encompasses data cleaning and automated KG construction.

5.1. Data Cleaning and Preparation

In order to ensure high-quality data, several cleaning tasks were performed on the MUDEC Museum database. These tasks included:

- **Language Standardization:** all entries in the database were translated into English to maintain a consistent language standard.
- **Correction of inconsistencies, standardization of formats** such as dates. Century ranges were transformed into numerical values based on predefined rules for better data representation. For example:
  - secc. IV/ V represented as : 301, 500
  - secc. I a.C./ I represented as : -100, 100
- **Data Consistency:** construction of a semantic dictionary for the cultures associated with the objects: (a) Grouping of equivalent or similar terms (e.g., "Atarco"; "Huari (Pachacamac)"; "Huari (Pativilca / Santa)"; "Huari (Santa)"; "Huari, Pativilca"; "Warí"; "Wari culture") into a single synset named 'Wari'. (b) Reduce lexical variability by applying normalization procedures to achieve greater semantic consistency within the graph.
- **Hierarchization of historical events**: identification, grouping, and structuring of historical events associated with objects. Establishment of a temporal and semantic hierarchy to organize events according to different levels of abstraction. Example: Acquisition → Transfer of ownership → Donation.

5.2. CIDOC CRM-Based Modeling

In this study, we mapped artifact data from the collections of the MUDEC to the CIDOC-CRM ontology to enable semantic interoperability. Hereafter, we describe how MUDEC data were mapped to CIDOC-CRM classes and properties.

1. **Artifact (E22_Man_Made_Object)**: Each artifact is modeled as an instance of the class `E22_Man_Made_Object`, identified by a unique ID generated from its inventory number.
2. **Ownership using P52_has_current_owner**: The *P52_has_current_owner* property links each artifact with its current owner. In our case, the Museo delle Culture di Milano, refers to the institution responsible for its custody.
3. **Object Naming (E41_Appellation)**: The name of each artifact is modeled using the *E41_Appellation* class. The link between them is established using *P1_is_identified_by*.
4. **Inventory Number (E42_Identifier)**: The original inventory number of each artifact is represented as an instance of `E42_Identifier`, with `P1_is_identified_by` linking it to the artifact.

To capture the collection history of artifacts, the collection event and associated details were modeled using CIDOC CRM as follows:

1. **Collection Event (E8_Acquisition)**: Each collection event was modeled as `E8_Acquisition`. A type is assigned to an acquisition by linking the `E8_Acquisition` event to an instance of `E55_Type` using the `P2_has_type` property (for example: Collection and Previous Collection). The artifacts are linked to the event using `P24_transferred_title_of`.

2. **Collection Date (E52_Time_Span)**: The collection date was modeled by a E52_Time_Span instance, associated with the collection event via P4_has_time_span, and capturing the date using P82a_begin_of_the_begin and *P82b_end_of_the_end*
3. **Collectors (E39_Actor)**: Collectors and previous collectors were represented as instances of E39_Actor(E74_Group or E21_Person), and their involvement linked to the collection event using P14_carried_out_by.

The modeling of the physical characteristics of each artifact, including its dimensions, material composition, and descriptive notes are modeled as described below:

1. **Material Composition (P45_consists_of)**: Each artifact's material is represented as an instance of E57_Material and linked to the artifact using the P45_consists_of property.
2. **Dimensions via P43_has_dimension**: The physical dimensions of an artifact such as height, length, and width are modeled as instances of E54_Dimension and linked to the artifact using P43_has_dimension property, with values specified using P90_has_value.
3. **Descriptive Note via P3_has_note**: A description of the artifact is modeled using the P3_has_note property, which links a textual note to the artifact.

The creation event of each artifact is represented as an instance of E12_Production, linked to the artifact via the P108_has_produced property. The temporal context is defined using P4_has_time_span, pointing to an instance of E52_Time_Span (e.g., 301–500 CE), while the geographical location is expressed via P7_took_place_at, referencing an instance of E53_Place (e.g., Peru). Creation techniques such as *Colombino* and *false lathe* are modelled as instances of E29_Design_or_Procedure, linked via P33_used_specific_technique.

The cultural context, the acquisition event, and the historical period in which an artifact was created are modeled as follows:

- **Cultural Attribution (P2_has_type)**: Each artifact is linked to the culture it likely originated from, utilizing the P2_has_type property, which connects the artifact to an instance of E55_Type.
- **Acquisition Event (E8_Acquisition)**: The acquisition event of artifacts is modeled via an instance of E8_Acquisition, which is associated to the artifact using the P22_transferred_title_to property. The temporal context of this event is captured using P4_has_time_span. Moreover, the actors involved in the acquisition are linked using the following two predicates:
  - *P23 transferred title from* → the former owner (actor)
  - *P24 transferred title to* → the new owner (actor)
- **Period Context (P8i_witnessed)**: Apart from historical years and dates, some artefacts were made during certain periods defined by archaeologists. These are linked to the object using the P8i_witnessed property, which links the artifact to an instance of E4_Period.

Furthermore, we modeled objects and their stories as a knowledge graph in Neo4j, illustrating the semantic connections between artifacts, their cultural origins, historical periods, materials used and the generated narratives. This graph allows for an intuitive visualization of complex relationships and facilitates thematic exploration by experts. Neo4j is a graph database designed to efficiently model and query complex relational data[2]. A representative part of the graph is shown in the Figure 3.

[2]Neo4j. https://neo4j.com

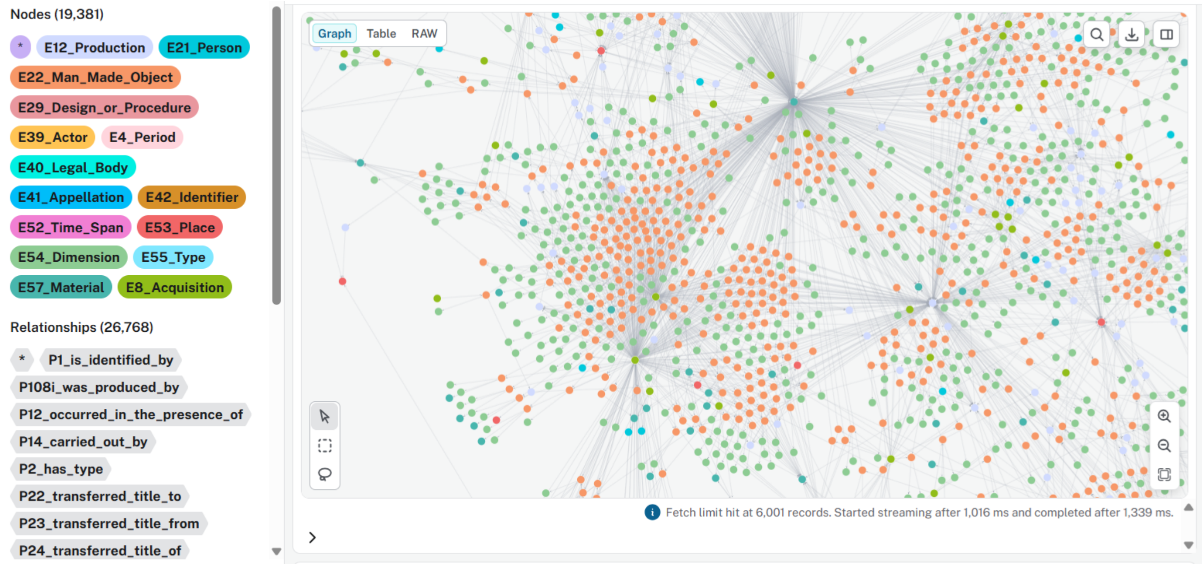

Figure 3: Extract from the semantic graph modeling the objects and their contexts generated in Neo4j.

5.3. Narrative Generation Driven by LLMs

This section details the narrative generation process at the heart of our system, which relies on pre-trained LLMs to transform extracted contextual information into coherent and engaging narratives.

To assess the ability of large language models to generate museum narratives from knowledge graphs, we selected five state-of-the-art models: GPT-4 (OpenAI), known for its narrative creativity and encyclopedic knowledge; Claude 4 Sonnet (Anthropic), which prioritizes factual accuracy and coherence over long passages; Llama 3.3 70B (Meta via Groq), a high-performance open-source alternative with fast inference times; Gemini 2.0 Flash Exp (Google), optimized for large-scale deployment and equipped with native multimodal capabilities; and Phi-4 (Microsoft), a compact model specializing in structured reasoning.

This selection reflects the diversity of paradigms (proprietary and open-source), the variability in size and specialization, and the different access modalities relevant to museum institutions with diverse resources. It allows us to explore the impact of these approaches on the quality and richness of the stories generated.

Each model was used to generate a narrative from the same object description. The general structure of a typical prompt includes the following elements: (A) Object Identification: A clear identification of the primary museum object on which the narrative should focus (e.g., title, inventory number). (B) Key Contextual Information: Information extracted via a Cypher query, presented in a logical and coherent manner. This may include details about the creator, culture of origin, period of creation, materials used, associated historical events, social, and cultural context in the knowledge graph. (C) Narrative Instructions: Specific instructions for the LLM regarding the style, tone, and purpose of the narrative.

A simplified example of a prompt might be: You are a knowledgeable and passionate museum curator with expertise in archaeology and cultural heritage. Below is structured information about an archaeological object from the MUDEC museum's collection. Your task is to narrate a vivid, emotionally engaging story of the object. Please speak in a respectful, and accessible tone — balancing storytelling with historical context.

GUIDELINES: -Write for the ear, not the eye: use flowing, conversational language that sounds natural when spoken -Include clear verbal transitions between ideas ("Moving forward in time..." "Centuries later...")
-LENGTH: 170–270 words

OBJECT INFORMATION: *- Object ID: Mudec4 - Title: bowl - Culture: Nazca culture - Historical*

Period: Early Intermediate Period - Materials: ceramic, mineral pigments - Creation Techniques: Colombino, false lathe, bladed, painting - Dimensions: 7 cm - Place of Origin: Peru - Date of Creation: between -100 and 300 - Primary Collector: Guillermo Schmidt-Pizarro - Collection Period: Unknown - Previous Collectors: Unknown - Acquired by MUDEC: 1934",

NARRATIVE STRUCTURE : *1. **Opening Hook** (25-35 words): Begin with an evocative scene or intriguing detail that immediately engages the Listener 2. **Physical Description** (20-30 words): Help listeners visualize the object using size comparisons, material descriptions, and tactile qualities 3. **Historical relevance** (50-70 words)inform visitors about the historical background of the piece and the cultural context it comes from? This will make the story rich and informative and make people learn from it 4. **Historical Journey** (60-80 words): Tell the story chronologically with clear time markers and smooth transitions between eras 5. **Present-Day Connection** (25-30 words):*

The output is:

Imagine an ancient Peruvian artisan grinding brilliant mineral pigments by lamplight, preparing to transform humble clay into a vessel that would outlast empires and speak across two millennia. This exquisite bowl, roughly the size of a large soup bowl at seven inches across, showcases the remarkable technical mastery of Nazca ceramicists. Its smooth ceramic surface, shaped using the ingenious false lathe technique, provides the perfect canvas for vibrant mineral pigments that still gleam with extraordinary intensity today. Created between 100 BCE and 300 CE during the Early Intermediate Period, this bowl represents the artistic pinnacle of Peru's Nazca culture. The Nazca people inhabited the arid coastal plains of southern Peru, where they developed sophisticated irrigation systems and created some of the world's most recognizable art, including the famous Nazca Lines. Their ceramic traditions reflected deep spiritual connections to the natural world, often featuring intricate designs that celebrated fertility, water, and agricultural abundance—essential themes for survival in their desert homeland. Moving through the centuries, this precious vessel eventually came into the collection of Guillermo Schmidt-Pizarro before finding its permanent home at MUDEC in 1934. For nearly ninety years now, it has served as a bridge between ancient Peru and modern Milan. Today, this humble yet extraordinary bowl continues to inspire wonder, proving that true artistry transcends both time and geography.

Full narrative examples for each model are provided in Appendix 9.

5.4. Multi-modal Diffusion via Text-to-Speech Conversion

We have integrated a multi-modal broadcasting module based on text-to-speech conversion (TTS), allowing us to enrich the user experience through the vocal narration of the generated stories. This module is based on Google Text-to-Speech technology, which transforms text into an audio signal using deep neural networks such as WaveNet [24], which are capable of producing a fluid and natural voice. The text is first linguistically analyzed (punctuation, rhythm, prosody), then synthesized in real time, generating accessible audio files. This functionality improves accessibility, facilitates cultural mediation, and allows for enriched and interactive dissemination of knowledge about museum objects.

6. Experimental study

To assess the quality of the narratives generated by our system from the museum knowledge graph, we opted for a hybrid experimental study involving three groups of evaluators: Domain Experts, Language Models, and the General Public. This approach aims to ensure a comprehensive evaluation that includes scientific validity, stylistic excellence, and accessibility to the target audience.

The study focused on a sample of thirty (30) distinct objects, selected to ensure diversity in terms of geography, period, culture, and artifact type. The stories were evaluated using a Likert scale from 1 (Very weak) to 5 (Very strong) on the following four criteria:

- **Factual correctness(C1)**: Is the story factually correct and consistent with the values, beliefs, symbols, traditions, or historical facts about the object's culture of origin?

- **Quality of literary style(C2)**: Is the style appropriate for a museum setting (metaphors, rhythm, figures of speech, etc.)?
- **Accessibility to the general public(C3)**: Is it easily understood by non-experts?
- **Hallucination(C4)**: Does the story contain invented or fictional information not supported by the object's description(A low value indicates good performance)?

The documents related to the evaluation are available in the GitHub repository https://github.com/Becaco/BecaNarratives. Please note that all evaluations were conducted completely anonymously. The experts, the general public, and the LLMs used as judges did not know which model had generated the narrative they were rating.

6.1. Composition of the Evaluation Panels

The evaluation was conducted by three distinct groups of evaluators, each with a specific role to play in the validation of the criteria:

- **Human Experts** (N = 6): Archaeologists, Curators, and Museologists. They evaluated criteria C1, C2, C3, and C4. Their judgment is authoritative in terms of scientific accuracy and professional compliance, and they represent the target user group for the developed tool. Human Experts judgments are used as the standard of reference (ground truth). The 6 experts formed 4 groups of evaluators, and each group evaluated all 150 narratives (30 objects × 5 models) according to the 4 criteria (C1, C2, C3, C4). This represents 4 groups × 4 criteria × 150 narratives = 2,400 expert evaluations in total.
- **Language Models (LLMs Judge)** (N = 3): Three large language models with varied structures were trained using specific prompts to assign scores on criteria C1, C2, C3, and C4. The models used are: DeepSeek-R1, Mistral-Large-Instruct-2407, Qwen3-32B.
  Their role is to evaluate the potential of automated rating and to compare the biases and performance of the different models in this judgment task.
- **General Public** (N = 7): Non-expert participants, they evaluated only criteria C2 and C3. Their evaluation is essential to validate the impact and clarity of the narrative, as they represent the target audience for the generated stories. 7 members of the general public formed 4 groups of evaluators, and each group evaluated all 150 stories (30 objects × 5 models) according to 2 criteria (C2, C3). This represents 4 groups × 2 criteria × 150 stories = 1,200 evaluations from the general public in total.

6.2. Results and Analysis

Figure 4 presents the average scores obtained by each generation model according to the four evaluation criteria (C1-C4) assessed by human experts (ground truth).

Performance analysis reveals a clear cross-sectional superiority of Gemini-2.0-flash across all criteria. The model achieves the highest scores in factual accuracy (C1: 3.77/5), literary quality (C2: 3.98/5), and accessibility (C3: 4.12/5), while exhibiting the lowest hallucination rate (C4: 1.32/5, best reliability control), demonstrating high performance that combines factual accuracy, narrative excellence, and veracity. Phi-4 also stands out for its excellent hallucination control (C4: 1.51/5, second-best score), partially compensating for its more modest literary quality (C2: 3.54/5). GPT-4 displays a balanced and solid performance across all criteria (3.35-3.99), without any notable weaknesses, while Claude-4-sonnet and Llama-3.3-70b exhibit comparable performance (averages 3.2-3.8) but with the highest hallucination rates (C4: 1.83 and 1.82), constituting a significant limitation for museum applications requiring absolute factual reliability.

Figure 5 shows an object-by-object granular analysis revealing that the Expert-Public gap varies significantly depending on the evaluated output, with extreme cases like Phi-4 where some texts receive 2.0-2.5 from experts but 4.5-5.0 from the public, exposing technical flaws invisible to non-specialists.

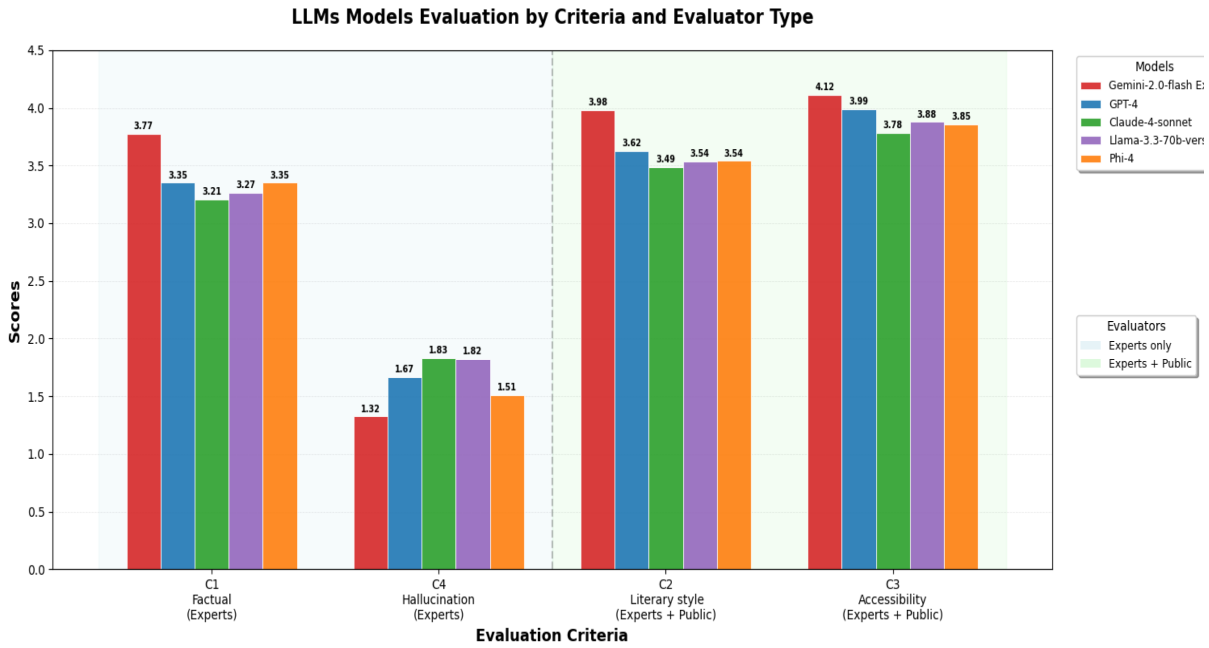

Figure 4: Performances Evaluations.

Experts discriminate finely between objects (dispersion 2.5-4.5) while the public converges towards 4.0-4.5, with only Gemini-2.0-flash generating a high consensus between both groups.

This visualization demonstrates that no model uniformly satisfies all evaluators on all objects, revealing that the perceived quality of an LLM system is fundamentally contextual and multi-dimensional, imperatively requiring multi-perspective evaluations rather than a single metric to grasp the complexity of excellence in AI text generation.

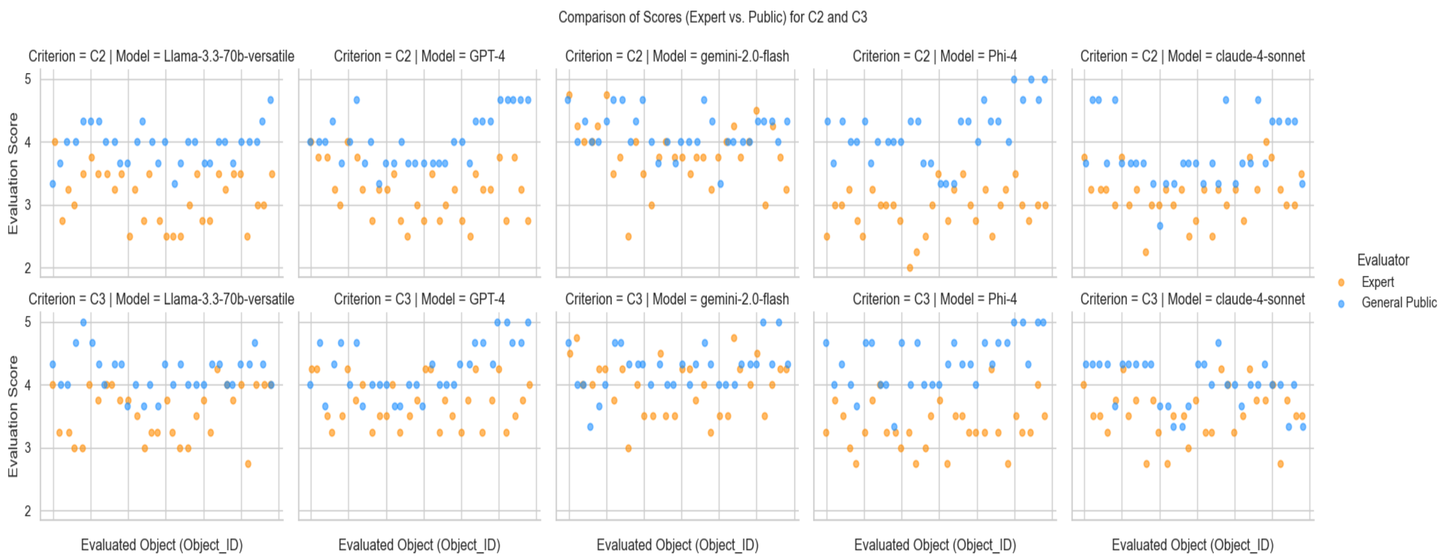

Figure 5: Comparison of Scores (Expert vs. Public) for C2 and C3.

6.3. Performance and Analysis of LLM Judges via Mean Absolute Error (MAE)

The table below summarizes the Mean Absolute Error (MAE) for three language models used as judges, across four criteria (C1 to C4). A lower MAE indicates better accuracy of the judge relative to ground truth.

Table 1
MAE of LLM Judges by Criterion

Criterion	DeepSeek-R1	Mistral-Large-Instruct-2407	Qwen3-32B
C1	0.40	**0.29**	0.74
C2	1.43	**0.87**	1.59
C3	1.28	**0.86**	1.25
C4	**1.04**	1.40	2.32
Mean	1.0375	**0.855**	1.475

Analysis of the results highlights the overall superiority of **Mistral** for this evaluation task. **Mistral Dominance:** This model achieves the lowest MAE on three of the four criteria (C1, C2, C3), displaying an overall MAE mean of $\mathbf{0.855}$, making it the most accurate judge. **Strength of Deepseek:** Deepseek performs best only on criterion C4 with an MAE of $\mathbf{1.04}$. However, its average performance (MAE $=$ 1.0375) ranks it second. **Performance of Qwen:** Qwen exhibits the highest error on all criteria except C1 and is the least accurate judge with an average MAE of 1.475. If accuracy is the priority, **Mistral_Judge** is the recommended choice, being closest to ground truth in the majority of criteria.

7. Discussion

In summary, the use of LLMs can be a valuable tool to help curators tell vibrant and accessible stories to different audiences, both within the walls of the museum itself and to visitors who only have access to museums in virtual spaces. The integration of text-to-speech technology also increases accessibility for individuals with reading or visual disabilities.

LLMs were able to significantly enrich the minimal information that was present in the original museum database. For instance, LLMs were provided with only very basic tabular information about the place of production, culture, and approximate dating of object PAM00005, a vase from the Nazca culture. The text generated by Gemini-2.0-flash stated that "Nazca potters mastered their craft in the river valleys of southern Peru. This was a time when communities thrived in one of Earth's driest deserts, creating elaborate textiles, geoglyphs, and ceramics like this vase". This text adds important contextual information about geography, climate, and cultural history, helping the (digital) museum visitor better understand the cultural background of this object and increasing both engagement and educational value. It is important to note here that LLMs only performed at this high standard for cultures that are relatively well-known. The Nazca culture from Peru, most famous for the Nazca Lines geoglyphs, has been studied extensively and featured in many popular science publications. In contrast, narratives generated around objects from lesser known cultures, did not contain this same level of contextualization.

However, we found that both the quality of the data and the quality of the output are essential and must be verified by a domain expert. In the stories created for this article, AI hallucinations were present and LLMs struggled with accurately presenting different aspects of the objects, such as size and texture. For example, the narrative created for PAM000016 by phi4 stated that a ceramic object "is about the height of a small child, standing 11 inches tall". Similarly, PAM000012 (8 cm. high) was said to "stand roughly the height of a modern-day child", and PAM000021 (16,5 cm. high) was compared to "roughly the size of a large apple". Apart from metaphors for size, LLMs at times used inappropriate metaphors for the material, textural, and tactile dimensions of objects and materials. PAM00009 was said to be "crafted from ceramic that feels smooth as the surface of a calm lake". Here the 'smoothness' of undisturbed waters in lakes was taken to have the same textural and tactile sensation as the smoothness of ceramics. Similarly, PAM000015 (a piece of textile) was said to be "as smooth as polished stone", again a misinterpretation of the tactile qualities of materials. Publishing this information without validation by the domain expert could undermine the trust that visitors have in the expertise and work of museums, emphasizing the need for a human-in-the-loop approach.

Finally, allowing LLMs to create object-based texts using different 'personas' could allow for the creation of differentiated texts for different audiences. In addition, the use of LLMs might also open up new possibilities to diversify the 'authoritative voice' that is often at the base of museum narratives, when exhibitions are curated by a single curator.

8. Conclusions

We proposed a Graph-RAG system designed to automatically produce contextualized narratives around cultural objects, based on a structured knowledge graph. By combining the generative capabilities of large language models with the semantic richness of the graph, our approach allows us to generate stories that are both factually accurate and culturally informed. The evaluation, conducted by both experts in the museum field, LLMs and members of the general public, demonstrated the relevance of our system, particularly when coupled with models such as Gemini-2.0-flash. These results highlight the potential of hybrid approaches that combine artificial intelligence and knowledge representation to promote and mediate cultural heritage.

This work also allowed us to identify several key lessons: (L1) the quality of the source data directly conditions the quality of the narratives; (L2) LLMs perform unevenly depending on the cultural notoriety of the objects; (L3) ontological anchoring reduces hallucinations but does not replace expert validation; (L4) the perception gap between experts and the general public highlights the need for a human-in-the-loop approach; (L5) LLMs as automated judges are promising but their reliability varies depending on the model used.

As an extension of this work, several avenues of research can be explored, (a) we plan to dynamically enrich the knowledge graph using open sources such as Wikidata, but also to integrate heterogeneous databases from several partner museums. This step is part of the Becaco project[3], which aims to study provenance research for cultural objects on a large scale. This scaling up aims to identify the links of circulation between objects, actors, and institutions, by mobilizing artificial intelligence to facilitate the exploration, comparative analysis, and storytelling of these complex trajectories, (b) We plan also to implement an adaptive story generation based on user profiles (visitors, researchers, children, etc.) to improve engagement and accessibility.

Acknowledgments

This research was funded by the European Union (ERC StG, BECACO, 101117234). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The authors wish to warmly thank Carolina Orsini, Anna Antonini, Alice Tadini, and Federica Villa of the Museo delle Culture di Milano for their valuable contribution, particularly in providing the database, as well as for their availability and collaboration throughout this work.

Declaration on Generative AI

During the preparation of this work, the author used Claude and ChatGPT for grammar and spelling checking, as well as to improve writing style and paraphrasing. The author then reviewed and edited the content as needed and assumes full responsibility for the final publication.

[3]https://www.universiteitleiden.nl/en/research/research-projects/archaeology/between-canon-and-coincidence-using-data-driven-approaches-to-understand-art-worlds-becaco

References

[1] H. Kim, The Impact of Digital Technologies on Contemporary Curatorial Practice, Master's thesis, Fashion Institute of Technology, State University of New York, 2024.

[2] P. Arantes, Museums in dispute: Artificial intelligence, digital culture, and critical curation, Arts 14 (2025) 65.

[3] S. Madakam, V. Lake, V. Lake, V. Lake, Internet of things (iot): A literature review, Journal of Computer and Communications 3 (2015) 164.

[4] G. Bruseker, N. Carboni, A. Guillem, Cultural heritage data management: The role of formal ontology and cidoc crm, in: Heritage and Archaeology in the Digital Age: Acquisition, Curation, and Dissemination of Spatial Cultural Heritage Data, 2017, pp. 93–131.

[5] F. Beretta, Semantic data for humanities and social sciences (sdhss): An ecosystem of cidoc crm extensions for research data production and reuse, arXiv preprint arXiv:2402.07531 (2024).

[6] J. Wu, J. Zhu, Y. Qi, J. Chen, M. Xu, F. Menolascina, V. Grau, Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation, arXiv preprint arXiv:2408.04187 (2024).

[7] J. Wu, J. Zhu, Y. Qi, J. Chen, M. Xu, F. Menolascina, V. Grau, Medical graph rag: Evidence-based medical large language model via graph retrieval-augmented generation, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 28443–28467.

[8] A. Masoudifard, M. M. Sorond, M. Madadi, M. Sabokrou, E. Habibi, Leveraging graph-rag and prompt engineering to enhance llm-based automated requirement traceability and compliance checks, arXiv preprint arXiv:2412.08593 (2024).

[9] Y. Ojima, H. Sakaji, T. Nakamura, H. Sakata, K. Seki, Y. Teshigawara, K. Aoyama, Knowledge management for automobile failure analysis using graph rag, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, pp. 6624–6631.

[10] M. Reusens, A. Adams, B. Baesens, Large language models to make museum archive collections more accessible, AI and Society (2025) 1–13.

[11] M. Mountantonakis, M. Koumakis, Y. Tzitzikas, Combining llms and hundreds of knowledge graphs for data enrichment, validation and integration: Case study cultural heritage domain (2024).

[12] G. Trichopoulos, M. Konstantakis, G. Alexandridis, G. Caridakis, Large language models as recommendation systems in museums, Electronics 12 (2023) 3829.

[13] P. Kelly, J. Schild, A. Jafari, Folkrag: A retrieval-augmented generation system for cultural heritage materials, Neural Computing and Applications 37 (2025) 20281–20297. doi:10.1007/s00521-025-11455-4.

[14] C. C. Chang, C. F. Li, C. H. Lee, H. S. Lee, Enhancing low-resource minority language translation with llms and retrieval-augmented generation for cultural nuances, in: Intelligent Systems Conference, Springer Nature Switzerland, 2025, pp. 190–204.

[15] I. Vasic, H. G. Fill, R. Quattrini, R. Pierdicca, Llm-aided museum guide: Personalized tours based on user preferences, in: International Conference on Extended Reality, Springer Nature Switzerland, 2024, pp. 249–262.

[16] Z. Liang, Z. Zeng, G. Fernandez Nieto, Y. Li, Y.-S. Tsai, G. Chen, Z. Swiecki, D. Gašević, J. Bradley, L. Sha, Data storytelling on multi-modal knowledge graph via data comics: A case study in yanyuwa language, in: Proceedings of the Joint Workshops of the 14th International Conference on Learning Analytics and Knowledge (LAK 2024), CEUR-WS, 2024.

[17] X. Wang, L. Lu, Q. Li, X. Yue, Patterns for generating narrative scripts through cultural heritage knowledge graphs for designers, Journal of Visualization 28 (2025) 661–680. doi:10.1007/s12650-025-01057-6.

[18] I. Vasic, H.-G. Fill, R. Quattrini, R. Pierdicca, Knowledge graphs vs. large language models: Competitors or partners in supporting virtual museums, ACM Journal on Computing and Cultural Heritage 18 (2025) 1–28.

[19] C. Palma, Neurosymbolic narrative generation for cultural heritage, IOS Press (2024).

[20] I. E. L. Blin, Narrative Understanding with Knowledge Graphs, Phd thesis, Vrije Universiteit Amsterdam, 2026. URL: https://doi.org/10.5463/thesis.1517. doi:10.5463/thesis.1517.
[21] Y. Duan, L. Shao, G. Hu, Specifying knowledge graph with data graph, information graph, knowledge graph, and wisdom graph, International Journal of Software Innovation (IJSI) 6 (2018) 10–25.
[22] D. Oldman, C. C. Labs, The cidoc conceptual reference model (cidoc-crm): Primer, 2014. CIDOC-CRM Official Website.
[23] CIDOC CRM Special Interest Group, Cidoc conceptual reference model, version 7.3.1, 2025. URL: https://cidoc-crm.org/Version/version-7.3.1, accessed: 2026-02-01.
[24] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, K. Kavukcuoglu, Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499 (2016).

9. Appendix: Example of LLM-Generated Museum Narratives

The following section presents five narratives generated by different Large Language Models (LLMs) based on the structured metadata of a single museum object: a Nazca ceramic vase held in the Mudec Museum (Object ID: Mudec1). The narratives are provided to illustrate variation in tone, style, and interpretive framing across different models.

Source Object Metadata (PAM 00001)

ID Mudec1

Name Vase

Culture Nazca culture

Materials Ceramic

Techniques Colombino, false lathe, bladed, painting

Dimensions Height: 8 cm; Length: 8.8 cm

Period Early Intermediate Period (301–500 CE)

Place of Production Peru

Acquisition 1934, Mudec Museum

Collector Guillermo Schmidt-Pizarro

Generated Narratives

Narrative A (llama-3.3-70b-versatile)

Imagine the vibrant landscapes of ancient Peru, where the sun casts golden rays over rolling hills, and skilled artisans are at work crafting beautiful objects. Among them is a stunning ceramic vase, roughly the size of a small table, its surface smooth and adorned with intricate designs that tell stories of the Nazca culture. This vase, created during the Early Intermediate Period, showcases techniques like Colombino and painting, reflecting a time when artistry flourished alongside daily life. The Nazca people, renowned for their impressive geoglyphs and advanced irrigation systems, infused their creations with spiritual significance, often depicting elements from their rich natural world. Moving forward in time to the early 1930s, this remarkable vase finds its way to the Mudec Museum, thanks to the passion of Guillermo Schmidt-Pizarro, who recognized its beauty and importance. Imagine it being carefully transported, its history woven into the very fabric of its ceramic body. Now picture this vase, standing proudly in our collection, a tangible link to a civilization that thrived over a millennium ago. It invites us to reflect on the creativity and resilience of the Nazca culture, reminding us of our shared human experience and the artistry that continues to connect us across the ages.

Narrative B (GPT-4)

Imagine a sun-baked valley in ancient Peru, the air alive with the murmur of artisans. Here, someone carefully shaped a small, ceramic vase, roughly the size of a clenched fist, intending for it to be both useful and beautiful. Feel its smooth surface in your mind, carefully painted with vibrant colours. This is not just a vessel; it is a window into the Nazca culture, a civilization that thrived during what we call the Early Intermediate Period, between 301 and 500 AD. The Nazca were master craftspeople. They did not have pottery wheels as we know them. Instead, they used techniques like *colombino*—building the vase by hand, coil by coil—and a *false lathe* to achieve its elegant form. The intricate designs were then applied with incredible precision, showcasing their understanding of colour and form. Centuries passed. The vase remained buried, a silent witness to the rise and fall of empires. Then, in 1934, it was brought to light by Guillermo Schmidt-Pizarro, becoming part of his collection. Soon after, it was acquired by the Mudec Museum, where it remains today. Now, as you stand before this Nazca vase, consider the journey it has taken. From the hands of an ancient artisan to our modern world, it speaks of a rich history and a shared human story. It is a testament to the enduring power of art and culture, connecting us to the past in a very tangible way.

Narrative C (gemini-2.0-flash)

Picture yourself in the arid coastal valleys of ancient Peru, around the fourth century, where skilled Nazca artisans are shaping clay with remarkable precision. Before you stands a small ceramic vase, roughly the size of a grapefruit, measuring just eight centimetres tall. Its smooth surface reveals the careful work of what we call a false lathe technique, where the potter rotated the vessel by hand while shaping and painting it with mineral pigments. The Nazca people flourished during what archaeologists call the Early Intermediate Period, between three hundred and five hundred CE. They inhabited the harsh desert landscape of southern Peru, yet created some of the most vibrant ceramics in the ancient Americas. This modest vessel represents their sophisticated understanding of clay preparation, firing techniques, and decorative artistry. Every painted line was applied with intention, every curve shaped with ancestral knowledge passed down through generations. Moving forward through the centuries, this small vase survived the collapse of Nazca civilization, the rise and fall of empires, and eventually caught the attention of Guillermo Schmidt-Pizarro, a collector who recognized its cultural significance. In 1934, it crossed oceans to join the Mudec Museum's collection, where it found a new purpose. Today, this humble vessel connects you directly to hands that worked nearly seventeen centuries ago, reminding us that human creativity transcends time and geography.

Narrative D (Phi-4)

In the heart of ancient Peru, a delicate ceramic vase was born. Roughly the size of a small table centrepiece, this vase is smooth as polished stone. The Nazca culture, known for their vibrant art, created this stunning piece during the Early Intermediate Period. Moving forward in time, between 301 and 500, skilled artisans in Peru carefully crafted it using Colombino, false lathe, and bladed techniques. Centuries later, in 1934, it was acquired by the Mudec Museum. Now, this beautiful vase stands as a testament to the ingenuity and craftsmanship of our ancestors, connecting us to the past and inviting us to appreciate its timeless beauty today.

Narrative E (claude-4-sonnet)

Imagine standing in the heart of ancient Peru, where the arid winds whisper tales of the Nazca, a civilization whose artistry has endured through time. Picture a vase, roughly the size of a small table, with a height of 8 cm and a length of 8.8 cm. Crafted from ceramic, its surface is as smooth as polished stone, adorned with intricate Colombino and bladed techniques, and vividly painted with the essence of its creators. This vase, a testament to the Early Intermediate Period of the Nazca culture, tells a story of a people who lived between 301 and 500 AD. The Nazca were known for their sophisticated pottery, which not only served utilitarian purposes but also acted as canvases for their expressions and beliefs. This particular vase, with its intricate designs, offers a glimpse into their world, showcasing their advanced artistic techniques and deep connection to their environment and spirituality. Centuries later, in 1934, the vase embarked on a new journey. Acquired by the Mudec Museum, its passage from the hands of the Nazca to those of the modern world was facilitated by Guillermo Schmidt-Pizarro, a dedicated archaeologist. This transition marked a new chapter in its life, bridging ancient culture with contemporary appreciation. Now, as it resides in the Mudec Museum, it invites us to connect with the past, offering insights into the lives of the Nazca people. It stands as a silent guardian of history, allowing us to appreciate the timeless beauty and ingenuity of a civilization long gone yet forever present.

A. Online Resources

The sources for the ceur-art style are available via

- GitHub,
- Overleaf template.