Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automatic extraction and structuring of cultural heritage analysis process documentation from audio and text files

Violette Abergel

violette.abergel@map.cnrs.fr

French National Centre for Scientific Research

Van Tuan Bui

Fondation des Sciences du Patrimoine

Amine Berbagui

Fondation des Sciences du Patrimoine

Research Article

Keywords: Heritage sciences, Knowledge extraction, Data provenance, Indexation, NLP, Speech recognition, Documentation, Knowledge graph

Posted Date: February 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8394677/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. Funding : This work was supported by the Paris Seine Graduate School Humanities, Creation, Heritage, Investissement d'Avenir ANR-17-EURE-0021 – Fondation for cultural heritage sciences This work benefited from State aid managed by the Agence Nationale de la Recherche (French National Research Agency) under the Investment for the Future Program integrated into France 2030, bearing the reference ANR 21-ESRE-0050 EquipEx+ ESPADON.

Automatic extraction and structuring of cultural heritage analysis process documentation from audio and text files

Violette Abergel[1, 2, *], Van Tuan Bui[3], Amine Berbagui[3]

1 CNRS, MAP UPR2002, Marseille, France
2 CNRS, INSA Lyon, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, Villeurbanne, France
3 Fondation des Sciences du Patrimoine, Cergy-Pontoise, France

* Corresponding author, violette.abergel@map.cnrs.fr

Abstract

Heritage science (HS) is an interdisciplinary field where collective knowledge emerges through an ongoing interplay between material objects and a wide range of research approaches that encompasses both the Humanities and experimental sciences. In this domain, the data management challenges are compounded by the strong heterogeneity of documentary sources, analytical data, and processes mobilized for condition reporting, analysis, monitoring or conservation purposes. Provenance metadata and paradata are essential for ensuring data reliability. Such documentation provides invaluable information on acquisition contexts and subsequent reuse possibilities. However, producing it rigorously is time-consuming, as the required information is diverse, context-dependent, and increasingly difficult to recover as time passes. In light of the massive daily data production in this field, developing methods to streamline data enrichment procedures is a clear priority. To address the risk of losing large amounts of undocumented data, the METAREVE project proposes a lightweight solution to help HS communities extract the key descriptive elements needed for minimal data understanding. Based on Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU), it takes the form of a web application that automatically documents scientific activities related to cultural heritage, drawing from common outputs such as expert reports, research articles, or even audio recordings of in situ acquisition processes. This approach has been implemented within the digital ecosystem developed by the EquipEx+ ESPADON project.

Keywords

Heritage sciences ; Knowledge extraction ; Data provenance ; Indexation ; NLP ; Speech recognition ; Documentation ; Knowledge graph

Introduction

Thanks to the development of new instrumental techniques and the spectacular increase of computational power, scientific investigations in cultural heritage are today driven by a wide variety of scenarios involving diverse acquisition and process modalities, depending on many factors from the object nature and complexity to restoration purposes. If such versatility proves to be precious to acquire fine knowledge on our cultural heritage, it also draws new challenges regarding the management of the resulting data production and our capacity to cross disciplinary viewpoints on a same object. In

particular, the strong heterogeneity of this everyday data production clearly hinders our capacity to properly disseminate it and share it with HS communities, and most of all, to ensure its reusability.

As a matter of fact, any acquisition process rooted in experimental science - such as 3D digitization, scientific imaging, or sampling for physical and chemical analysis - inevitably involves subjective decisions that can significantly affect both the nature of the data produced and its potential for future reuse. It is possible to keep track of these choices, provided that the acquisition conditions, operational workflows, and the hardware and software parameters used are carefully documented. Such documentation effort is at the heart of the FAIR principles (Findable, Accessible, Interoperable, Reusable), introduced back in 2014 to enhance the accessibility and usability of scientific data for both humans and machines. These principles have since been widely adopted, notably within the European Open Science Cloud initiative [1], [2]. However, the information to be provided is both diverse, context-dependent, and the field is often not conducive to such meticulous documentation tasks. Over time, this invaluable information tends to get lost, and proves difficult, if not impossible, to retrieve *a posteriori* for those who have not taken part in the operations. Thus, in order to limit the loss of this precious information, it appears essential to propose new approaches making the provenance documentation less burdensome for data producers.

Thankfully, the scientific production resulting from these acquisitions and processes often include many clues concerning the contexts of data production and analysis (e.g. scientific publications, expert reports, laboratory notebooks, etc.). Voice recordings also constitute a medium that could considerably lighten documentation processes. **Given this context, METAREVE intends to address the challenges of automatic extraction of provenance metadata in the heritage field.** The aim is to offer a versatile approach capable of taking advantage of the various digital data produced by heritage scientists during their studies - whether traditional text-based documents or audio sources - limiting the need for human intervention and ensuring a basic documentation of the processes behind data production. One major difficulty lies in the fact that the people, institutions, concepts, techniques, and protocols that may be mentioned regarding data manipulation are very specific to particular areas of expertise and contexts of analysis. They often refer to vocabulary terms or acronyms that are missing in standard learning corpora, yet are essential in our framework either for voice transcription or entity linking. For example, when a heritage scientist says “SOLEIL” in a sentence, he may be talking about the star (the Sun), or perhaps the French synchrotron facility. This semantic ambiguity can only be resolved by taking the context into account. Thankfully, although not exhaustive, there are many thesauri offered by heritage communities to support documentation initiatives by providing semantic references.

Since 2021, the ESPADON project, founded by the French National Research Agency (ANR) has been working to introduce new methods and technologies for heritage research. The project focuses, on one hand, on creating innovative instruments, and on the other, on designing a digital ecosystem to connect the many institutions involved in heritage sciences, along with the data and knowledge they generate. Heritage studies combine increasingly used highly technical instruments and expert insights from numerous field of expertise. Given its highly multidisciplinary nature, the project explores how data and knowledge produced from multiple disciplinary perspectives on the same object can be collected, correlated, and shared [3]. The ecosystem is built through the interconnection of several software components, covering each step of the data lifecycle (Figure 1). One of its main components is Opentheso [4], an open-source web application dedicated to thesaurus management. It includes several tools for vocabulary management, concept mapping, permanent identifier assignment, and structure lexicon creation. Its functionalities are accessible either through a web interface or through a REST

API. It should be noted that each concept can be aligned with freely available external vocabularies or semantic references (e.g. Art & Architecture Thesaurus (AAT) from Getty [5], Wikidata [6], etc). This tool is increasingly used in France in HS communities and has recently proven its undeniable usefulness in various scientific activities regarding the Notre-Dame cathedral, where it was "*the nerve center of the digital platform enabling complex cross-referencing between different tools*" [7]. The recent deployment of Opentheso in the ESPADON ecosystem now allows partner institutions to manage and share their specific vocabularies, like the TAPAC thesaurus shared by the *Laboratoire de Recherche des Monuments Historiques* (LRMH), which provides a useful structured list of concepts and definitions covering investigative techniques used in heritage sciences.

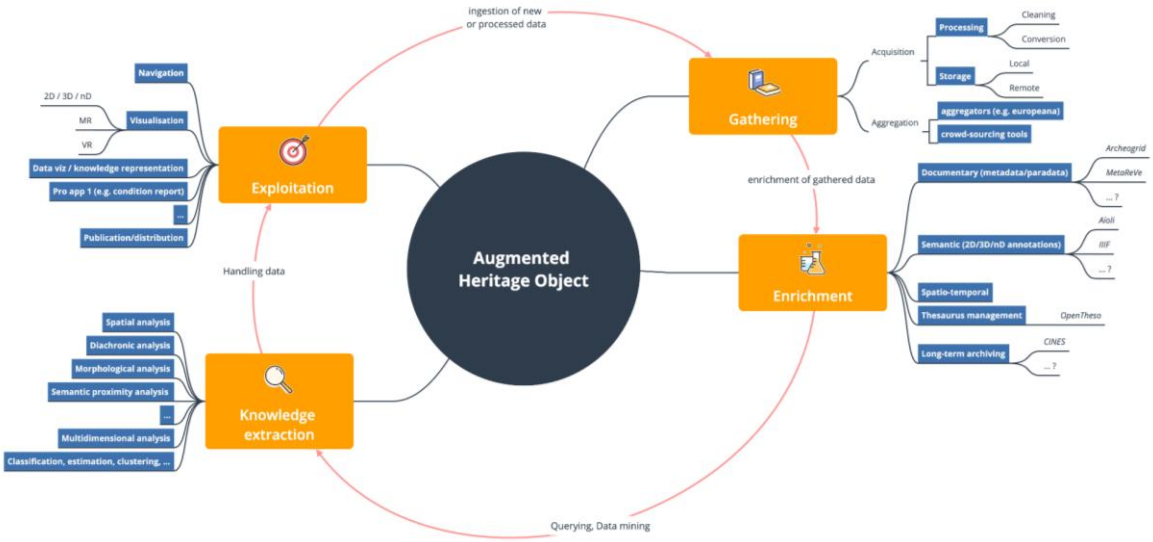


Figure 1:The ESPADON digital ecosystem, built around the concept of "Augmented Heritage Object".

In this context, our goal is to provide the ESPADON ecosystem with a new software component dedicated to data-driven metadata extraction, fueled by Opentheso's multidisciplinary thesauri, in order to gradually build a systematic and consistent documentation of HS analyses. This article aims to provide a comprehensive overview of the METAREVE workflow. In the section 2.1, we describe our Whisper fine-tuning and post-correction approaches, addressing the challenges of transcribing audio files in the field of HS. In the section 2.2, we evaluate different methods of document-level event extraction and describe the approach adopted for METAREVE and its results. In the section 2.3, we introduce our linking pipeline, which connects the extracted documentation to multidisciplinary thesaurus terms. Finally, in 3, we describe the implementation of the tool.

1. State of the art

The evolution of digital infrastructures over the last few decades has paved the way for data sharing practices and new methods for scientific information processing. The Semantic Web, one of the most significant developments since the Web's inception, gradually builds an immense knowledge base, promoting not only the interconnection of massive volumes of data, but also information retrieval, visualization and even analysis, although the subsequent information's quality is still somewhat uncertain. In this respect, the structuring of metadata according to current standards has been identified by the French Ministry of Culture as a major challenge for the cultural sector [8], conditioning the use of data by machines, and informing its reliability for users who have not participated in its production (quality, reproducibility, consistency, legal conformity, etc.). The FAIR principles [1], which provide

measures to make scientific data easy to find, accessible, interoperable and reusable, have been adopted in various European initiatives [2]. In the field of heritage sciences, in addition to the well-known CIDOC-CRM ontology [9] and its extensions (CRMsci, CRMdig, CRMgeo, CRMcr, etc.) [10], [11], [12], [13], numerous standards and conceptual models offer frameworks on which the necessary work of digital resources description can be based: from the most generic (DublinCore, PROV, OAI-ORE, etc.) [14], [15], [16], to the most specific (PARCOURS, OBOE, NeXus, etc.) [17], [18], [19], not forgetting the several international initiatives which have addressed (or are still addressing) these issues [20], [21], [22], [23], [24], [25]. Some initiatives like [26] even aim to build a global, flexible meta-model, compatible with CIDOC-CRM, and present interesting results concerning the description of analytical data. However, two main factors still hinder a proper systematic provenance documentation of HS data by its producers:

1) Most digital heritage initiatives from recent decades tend to prove that the adoption of new digital tools by HS communities largely depends on the ability of these tools to integrate with existing methodologies without fundamentally disrupting them. Despite its significant advantages in data management, analysis, and sharing, conceptual modeling remains challenging for HS communities due to the specialized knowledge in engineering that it requires.

2) The fuzzy variability of HS data production make it difficult to unify the description of scattered information and data, across different production contexts [27].

Given on the one hand, the diversity of disciplines mobilized in HS and the heterogeneous nature of data production methods and contexts, and on the other hand, the great value of knowledge engineering and/or semantic web complying approaches, some recent works [28], [27] show that the gap can be bridged by using a more intuitive metadata schema like the W7 model, aligned with the CIDOC and PROV ontologies, to easily document the provenance of heritage science data. Yet, to date, these interdisciplinary initiatives have not provided a truly intuitive tool for easily indexing analytic data and enabling their metadata alignment. Despite the existence of tools for manually managing the metadata of digital heritage resources [29], [30], this description work remains tedious for a knowledge producer, due to the amount of information required to precisely inform the nature of a resource, and the specificity of the related terms.

Artificial intelligence provides the means to tackle these issues, in particular through deep learning methods dedicated to the recognition of named entities (NER) or event extraction (EE) from unstructured data [31]. Event Extraction (EE) involves identifying structured information from text and generally consists of two subtasks: event classification and event argument extraction. In this work, we focus solely on cultural hertitage (CH) process events and identifying their arguments. In the literature, this corresponds to the task of Event Argument Extraction (EAE). Large language models (LLMs) have gained significant recognition for their capabilities, leading to increasing interest in their application to information extraction tasks including event argument extraction (EAE), within a generative framework [32]. This shift is primarily driven by LLMs' remarkable capacity for high-level, long-text comprehension and generation. Compared to token-based classification approaches, LLM-based methods can incorporate more semantic information and present queries in a more natural manner, enabling direct prediction of event arguments without the need to extract candidate arguments first [33], [34]. Document-level EAE (DocEAE) has caught much attention in recent years due to the challenge of arguments being dispersed across an entire document [35]. Many works intend to address this

challenge, such as [36] which introduced ChatIE, a zero-shot[1] EE framework built on top of ChatGPT following a dialog-style question-answering approach. In [37], the authors treated EAE as a conditional text-generation task, utilizing a template-based approach (e.g., "<arg1> attacked <arg2> using <arg3> at <arg4> place") and asking encoder-decoder models such as BART or T5 to fill in placeholders by copying tokens from the document. Furthermore, [38] explored EAE under prompt-based paradigms, focusing on how to best enrich prompts with additional information to boost performance. The authors showed that prompt-based EAE can be improved by enriching the prompt with more argument clues (especially from the same event).

In the heritage field, such approaches are increasingly applied on textual data [39], [40], [41], [42] but rarely applied to provenance and traceability challenges. They focus mainly on the extraction of information contained within the heritage objects themselves (e.g. extraction of mentions of localities, personalities, or dates in historical newspapers and archive resources, etc.) but exclude the many digital resources resulting from the multidisciplinary study of these objects, such as analytical data. The THESPIAN-NER tool, developed within the 4CH framework [43], proposes a very interesting approach based on the CIDOC-CRM and a deep neural network for named entity recognition through a web application which enables users to upload their Italian written report files and obtain labelled entities. Applied to collaborative semantic enrichment of textual datasets, the potential of this approach is undeniable. However, it is only available in the Italian language, and, for the moment, the extracted metadata are too fragmentary to meet the specific needs of provenance documentation.

Finally, these approaches focus on text-based datasets, but none of them explores the ability of spoken language to act as a more intuitive information carrier. Yet, oral exchanges between operators during data acquisition could constitute an extremely rich source of contextual information. For a busy heritage scientist, it could even be more convenient to take a few minutes to speak about his data acquisition workflows than filling endless dedicated forms or spreadsheets. Here again, artificial intelligence offers promising opportunities. Automatic speech recognition (ASR) methods are commonly used by voice assistants, automatic translators, and even the automatic subtitling of our precious video-conferencing tools. They have been considerably improved by recurrent neural networks [44], and now are part of our daily lives.

All these advances now offer significant opportunities for heritage sciences. The convergence of ASR and DocEAE methods opens considerable perspectives for multidisciplinary data management. In the following sections, we present a data-driven approach aimed at automatically extracting metadata in order to limit the loss caused by the growing amount of undocumented CH data.

2. Method

To assist with the documentation effort, our approach consists of providing an extraction tool capable of detecting and structuring information relevant to the operational contexts from simple audio or text files. As shown in (Figure 2), the file can be any of the following:

- An audio recording, whether a passive record of on-site operations (Figure 2.a), or an active record of a subsequent oral summary description (Figure 2.b).
- A digital written report (analysis report, extract from a laboratory notebook, etc.) (Figure 2.c).

[1] Zero-shot prompting refers to an approach where the prompt used to interact with the model does not include any labeled training examples for the target event.



a. On site informal discussions b. Subsequent audio report c. Expert reports, scientific papers…

Figure 2: METAREVE possible inputs.

From this file, we aim to identify any elements suggesting the use of HS analysis techniques and, for each of them, to answer the seven questions of W7 providing a basic understanding of the related data production process: Who? When? Where? Why? How? Which? What? Stored as a structured JSON file, such output offers three benefits:

- Synthetizing all the processed documents in a harmonized, comprehensive and intelligible way;
- Allowing to index the HS analyses performed on CH objects in a database by structuring valuable information about the processes involved;
- Facilitating corpus exploration through the construction of simple and comprehensive queries.

Many AI-based approaches enable information to be extracted from unstructured documents (see section 2.2). When it comes to audio documents, extraction usually falls within the scope of Spoken Language Understanding (SLU). In our case, we aim to support both audio and text documents. To offer a versatile tool capable of processing indifferently these two types of input data with similar performances, we have chosen to adopt a three-step approach, breaking down the process in complementary phases (Figure 3): Automatic Speech Recognition (ASR) – optional, depending on the input type (see section 2.1), Natural Language Understanding (NLU) – on the textual input or on the result of the transcription (see section 2.2), and Entity Linking (see section 2.3).

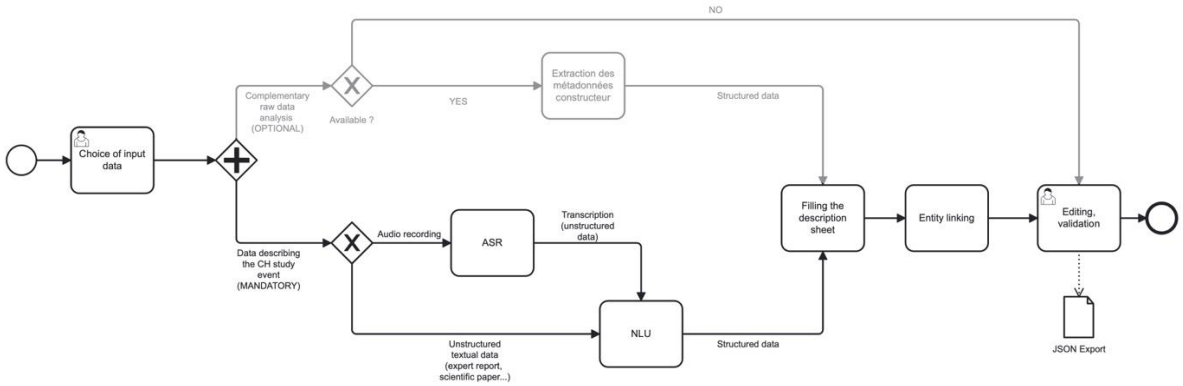


Figure 3: Diagram of METAREVE pipeline, with three main steps: ASR, NLU, and Entity Linking.

An optional pipeline – still under development – aims to complete this extraction by processing complementary raw analysis data (e.g. 3D point clouds, measurement series from sensors, etc.), getting the manufacturer parameters when they are available and stored in non-proprietary formats. For example, if an expert report contains several photographs, processing them separately would allow valuable information to be extracted from their EXIF data, such as the dates on which the data was acquired, the equipment used, etc. While such a process is straightforward when applied to photographic data, it becomes significantly harder in the case of more "sophisticated" analysis techniques.

2.1. Step 1 - Automatic Speech Recognition

This step aims to transcribe audio files, in order to generate an output compatible with the subsequent metadata extraction process (see section 2.2). Our Speech-To-Text (STT) model was designed to enhance the transcription accuracy of audio files referring to HS specific vocabulary.

2.1.1. Data gathering and model selection

We leveraged OpenAI's Whisper model [45], fine-tuning it on a custom dataset to address challenges in transcribing domain-specific vocabulary, analysis techniques, names and places. To do so, we curated a dataset of HS related audio files, provided by the *Laboratoire de Recherche des Monuments Historiques*, representing a variety of speakers, accents and speech contexts. In total, this amounts to approximately 11 hours of labeled audio data, split into 10 hours for training and 1 hour for testing. We ensured that the dataset covers multiple speaking styles, including conversational, formal, and spontaneous speech with different accents and challenging elements such as overlapping speakers and background noise to reflect as best as possible what we would have in real cases. The transcription performance of several prominent models on our test dataset was evaluated to establish baseline comparisons (Table 1).

Model	**WER (Word Error Rate)**	**Performance Insights**
MMS-1B-all	0.43	Observed good performance on general speech but struggled with specific names and rare terms
Wav2vec-FR-7k-large	0.33	Strong performance on clean speech, moderate drop in noisy conditions
Whisper-large-v2 (base)	0.15	Superior handling of multilingual data, globally very accurate but occasional errors with repetitions
Whisper large-v3	0.08	

Table 1: Table of transcription performance of different models.

While Whisper-large-v2 and Whisper large-v3 demonstrated robust transcription capabilities, several limitations emerged:

- **Specific Vocabulary:** Whisper struggled with uncommon or domain-specific terms that were not well-represented in its pre-training data.
- **Proper Names and Places:** Errors occurred in transcribing proper nouns, including unique names and less commonly mentioned locations.
- **Noisy Backgrounds:** Although the model handled moderate noise well, performance decreased in scenarios with significant background interference with also word repetitions.

These shortcomings emphasized the need for domain adaptation through fine-tuning.

2.1.2. Whisper Fine-Tuning

Audio input preprocessing

To overcome the aforementioned limitations, we proceeded to the fine-tuning of four variants of the Whisper model (**Whisper-small**, **Whisper-medium**, **Whisper-large-v2** and **Whisper large-v3**) to identify the best suited to our application domain. By tailoring each model to our dataset, we aimed to

improve their accuracy in recognizing domain-specific language and uncommon named entities. Fine-tuning not only aligned the Whisper models more closely with our audio domain but also helped reduce the **WER** significantly on the test dataset and make sure to get the best transcription possible for the second model (NLU) that comes afterward.

Prior to fine-tuning, the audio data underwent a preprocessing workflow. First, we used an audio **format standardization**, converting all the audio files to the **WAV** format to ensure compatibility. Then we adapted their **sampling rate**. Speech can be represented by a 1-dimensional array that varies with time, the value of the array at any given time step being the signal's amplitude at that point. From the amplitude information alone, we can reconstruct the frequency spectrum of an audio input and recover all acoustic features. However, since speech is continuous, it contains an infinite number of amplitude values, which can be an issue for computer devices expecting finite arrays. For this reason, we discretized our speech signal by sampling values at fixed time steps. The chosen interval is known as the **sampling rate** and is usually measured in samples/sec or Hertz (Hz). Sampling with a higher sampling rate results in a better approximation of the continuous speech signal, but also requires storing more values per second. The Whisper feature extractor expects audio inputs with a sampling rate of **16kHz** [45], matching our audio files to this value was crucial, as audio signals with different sampling rates have very different distributions.

Chunking

The Whisper feature extractor performs two operations: it first pads/truncates a batch of audio samples so that all samples have an input length of 30s maximum. Shorter ones must be padded to 30s by appending zeros to the end of the sequence (zeros being an audio signal corresponding to silence), longer ones must be truncated to 30s. Since all elements in the batch are padded/truncated to a fixed length, we didn't require an attention mask when forwarding the audio inputs to the Whisper model. Whisper is unique in this regard: with most audio models, it is expected to provide an attention mask detailing where sequences have been padded, and thus where they should be ignored in the self-attention mechanism. Whisper is trained to operate without one and instead identifies irrelevant inputs directly from speech signals. To transcribe audios longer than Whisper's 30s maximum receptive field, one of these two long-form algorithms are required:

1. **Sequential**: uses a "sliding window" for buffered inference, transcribing 30-second slices one after the other

2. **Chunked**: splits long audio files into shorter ones (with a small overlap between segments), transcribes each segment independently, and stitches the resulting transcriptions at the boundaries

In our case, we opted for the second approach, and each audio file was segmented into 30-second chunks with an overlap of one second between each to keep context information. Finally, in order to build our training dataset, we processed each chunk with the Whisper-large-v2 base model to get most of the transcription work done automatically. These transcriptions were then reviewed **manually** to fix any hidden spelling errors. Two CSV files were generated (Figure 4), one as training dataset (1048 samples) and the other for testing purposes (47 samples).

wav_filename	transcript
/home/ubuntu/data/ new_train_v2_10_sec_16000/00000104_25_.wav	lungs bilaterally clear to auscultation you go ahead check your neck here
/home/ubuntu/data/data_v4/old_audio_files/audio_3_65.wav	so if if the nine hundred
/home/ubuntu/data/ new_train_v2_10_sec_16000/00000111_28_.wav	no recent travel no recent sick contacts

Figure 4: Extract from the CSV dataset produced for Whisper fine-tuning. Each sample has two columns, the first one containing the file path, the second one its transcript.

Optimization and fine-tuning

To proceed to the fine-tuning, our computational resources (1x GPU Tesla T4 16Gb VRAM and 32 CPU & 512Gb RAM) were adequate but fairly limited considering that the largest Whisper version have more than one billion parameters. Fine-tuning all of these would have been prohibitively costly. To optimize this process, we used Hugging Face PEFT (Parameters-Efficient Fine-Tuning) [46], a library allowing to efficiently adapt large pretrained models to various downstream applications. PEFT supports Low-Rank Adaptation (LoRA), as well as several other variants. LoRA is a parameter-efficient fine-tuning technique that allows to adapt a large model without modifying all its weights [47]. Instead of updating directly the weight matrices, LoRA inserts small trainable "adapter" matrices into specific layers and learns only these low-rank updates while keeping the original weights frozen. This approach drastically reduces the number of trainable parameters (Table 2**Error! Reference source not found.**), and significantly decreases computational and storage costs, while preserving the knowledge of the base model. In practice, LoRA makes it feasible to specialize very large models on domain-specific data (e.g., technical speech or documents), yielding performance comparable to a fully fine-tuned model.

Model	**Params**	**Trainable params**	**Trainable%**
Whisper-medium	770M	9.4M	1.22%
Whisper-large-v2	1550M	15.7M	1%
Whisper-large-v3	1550M	15.7M	1%

Table 2: Parameters reduction for each model with LoRA method.

The (Figure 5) shows the evolution of the evaluation loss over time for the four Whisper variants. The green curve (Whisper-small) starts with the highest loss and remains clearly above the others, while the blue (Whisper-medium) and purple (Whisper-large-v2) models achieve progressively lower losses but exhibit a slight upward drift as training proceeds, suggesting the beginning of overfitting on the validation set, or suboptimal fine-tuning conditions (in our case, probably due to insufficient training data). In contrast, the orange curve (Whisper-large-v3) is remarkably stable and stays below all other models throughout training. This indicates that the model is inherently better suited to the task, either because of its larger capacity, its pre-training quality, or its resilience to overfitting under the current fine-tuning setup. Overall, this model demonstrates superior generalization performance and appears to be a promising candidate for deployment in our application domain.

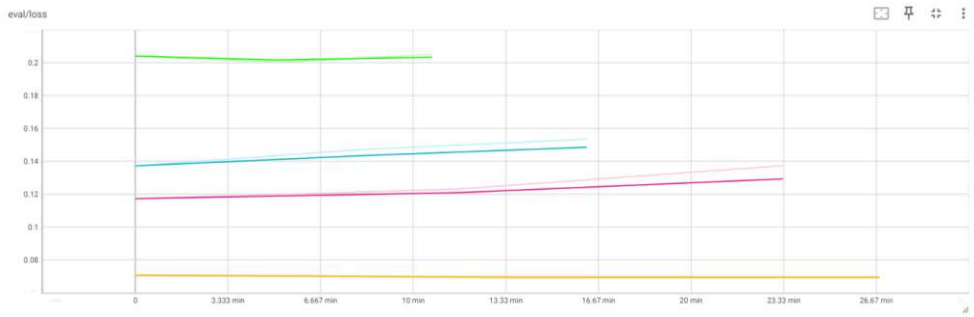


Figure 5:Evolution of the evaluation loss over time for Whisper-small (green), Whisper-medium (blue), Whisper-large-V2 (purple), and Whisper-large-V3 (orange).

The (Figure 6) shows the evolution of the training loss for the medium, large-v2 and large-v3 Whisper models. All three curves drop steeply in the first minutes, then continue to decrease more slowly, indicating that the models keep learning throughout training. The large-v2 model (purple) ends with the lowest training loss, slightly below both medium (blue) and large-v3 (orange), which suggests it fits the training data the most aggressively. In contrast, large-v3 converges quickly and then plateaus at a slightly higher loss, a behavior that is consistent with its better validation loss and indicates a more regularized fit that generalizes better to the evaluation set.

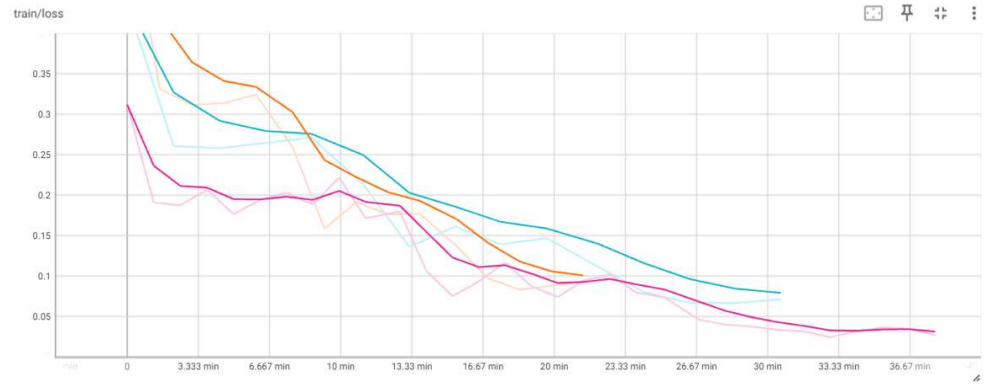


Figure 6: Evolution of the training loss over time for Whisper-small (green), Whisper-medium (blue), Whisper-large-V2 (purple), and Whisper-large-V3 (orange).

In (Table 3), the comparison of Whisper configurations confirms that, moving from the small baseline without optimization to the PEFT-optimized Whisper models systematically reduces both training and evaluation loss: from 0.24/0.20 for Whisper-small without optimization, down to 0.07/0.13 for Whisper large-v2 and 0.02/0.07 for large-v3. This indicates that parameter-efficient fine-tuning on top of larger backbones is very effective at fitting the domain data. In parallel, the raw and normalized WER decrease from 0.16/0.12 (Whisper-small without optimization) to 0.10/0.07 for Whisper-large-v2 with optimization, showing a clear gain in transcription accuracy. These quality improvements come with a computational cost: the runtime increases almost linearly with model size, from 14 minutes for the small model to 45 minutes for large-v3.

However, an interesting point is that **Whisper-large-v3, despite achieving the lowest losses, does not yield the best WER** (0.15 vs 0.10 for Whisper-large-v2), and its normalized WER is also higher. This suggests that further minimizing the training objective on a very large model does not automatically translate into better word-level accuracy, and highlights a trade-off between model size, optimization objective, and the practical metric (WER).

Sizes	PEFT	Training loss	Evaluation loss	Run time	WER	Normalized WER

	optimization					
small	No	0.24	0.20	14mn	0.16	0.12
small	Yes	0.14	0.17	13mn	0.14	0.10
medium	Yes	0.09	0.15	24mn	0.12	0.08
large-v2	Yes	0.07	0.13	34mn	0.10	0.07
large-v3	Yes	0.02	0.07	45mn	0.15	0.15

Table 3: Comparison of Whisper configurations, each line representing the best epoch of the training

Even though Whisper-large-v3 reaches the smallest training and evaluation losses, the metric it is trained to minimize is only an indirect surrogate for WER. With a very large capacity model and a relatively small fine-tuning dataset, it can overfit subtle token-level patterns without actually improving word-level alignment. Small differences in decoding behavior (e.g., punctuation, casing, segmentation) can also hurt WER even when the predicted token distribution is sharper. As a result, Whisper-large-v3 looks "better" from the loss perspective but does not translate this into the best WER on the evaluation set. For this reason, we decided to use the **Whisper-large-v2 model** for the following parts of this work, given its better WER and its overall good evaluation loss.

2.1.3. Results and post-correction

To adapt Whisper to our highly technical, multi-speaker domain, we applied LoRA-based PEFT fine-tuning on four model sizes (small, medium, large-v2 and large-v3), and the results clearly show the benefit of this approach. Compared to the small baseline without PEFT, the PEFT-optimized medium and large models achieve substantially lower training and evaluation losses, with large-v3 reaching the best overall fit and large-v2 yielding the lowest WER on our test set. The loss curves confirmed this behavior: training loss decreased steadily for all models, while the evaluation loss of large-v3 remained both the lowest and the most stable over time, indicating better generalization than the other configurations, even though its WER is not optimal. Fine-tuning was therefore essential in our setting: it allowed the models to internalize domain-specific vocabulary and speaker variability that the base checkpoints cannot fully capture, while LoRA keeps the computational and memory costs manageable. That said, the discrepancy between loss and WER suggests room for improvement, for example by tuning hyperparameters with WER as the primary early-stopping criterion, experimenting with alternative decoding strategies, or enlarging and diversifying the fine-tuning corpus to better align token-level optimization with the word-level accuracy required by the downstream metadata extraction task.

To improve the reliability of person names in the transcriptions, we introduce a dedicated post-correction stage based on a curated error dictionary and fuzzy matching (Figure 7). We maintain a CSV file with two columns: the correct spelling of each name (Name) and all its observed misspellings (Errors), which serves as a growing knowledge base that can be continuously enriched as new errors are discovered. After transcription, a SpaCy NER model (fr_core_news_md) is applied to detect all person entities in the text [48]. For each detected name, the system first checks for exact matches: if the name appears in the Name column it is left unchanged, whereas if it appears in Errors it is automatically replaced by the corresponding correct form. When no exact match is found, we compute fuzzy similarity scores between the detected entity and both columns, assign a confidence level based on predefined thresholds (low, average, high, very high), and record candidate corrections accordingly. At the end of this process, only entities whose confidence is greater than or equal to a user-specified threshold are actually replaced in the transcription. This design gives the user explicit control over how aggressive the post-correction should be, ranging from conservative to very soft, while leveraging a continuously updatable database of name variations to systematically correct recurrent transcription errors.

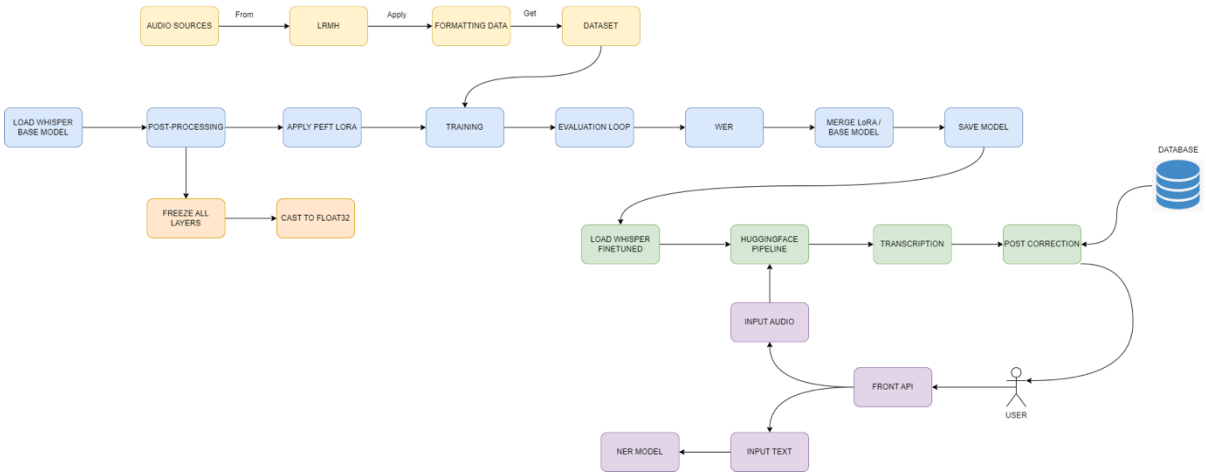


Figure 7: Synthesis of our ASR approach, including Whisper fine-tuning and post-correction to optimize audio-transcription.

2.2. Step 2 – Natural Language Understanding

The input of this process is either a transcription result from step 1 (see section 2.1), or natively textual contents (expert reports, scientific papers, laboratory notebook extract, etc.), often sent in PDF file format. From there, the goal is to extract relevant information regarding any use of HS analysis techniques, and structuring it to generate a JSON file. We formulated this task as a document-level event extraction [35]. Specifically, we define an event in CH, termed **CH Process**, which represents the use of scientific analysis methods, techniques, or procedures applied to CH objects within the scope of a project or mission.

2.2.1. Event extraction

To design the METAREVE Event Extraction approach, we began by testing different DocEAE methods based on recent open-source LLMs (e.g. Mixtral, Llama, and Qwen models), to evaluate and enhance their ability to produce metadata on CH Processes.

2.2.2. Metadata Schema

To structure the arguments related to a CH Process event, we employed the W7 model, which captures key contextual aspects: **What**, **Who**, **Where**, **When**, **Which**, **How**, and **Why**. Schemas derived from this model have been effectively used in diverse contexts, its adaptability and extensibility making it well-suited for representing provenance semantics across various scientific disciplines, such as biomedical research [49], [50], [51]. In the field of Cultural Heritage, [28] used a derivation of this schema to document multimodal acquisition workflows (e.g. photogrammetry, reflectance transformation imaging, laser scanning). More recently, [27] clearly illustrated its potential for HS data provenance documentation, demonstrating how it could be combined with CIDOC and PROV ontologies and help to answer highly specific competency questions. In our case, the W7 schema was used to define the arguments for the CH Process event within the study context. Specifically, the key arguments were defined as follows:

- **Who**: the individual experts involved, along with their names and roles/affiliations;
- **What**: the information about the CH object or the particular parts being documented, e.g. *Tree-ring width series from Notre-Dame's roof structure; 19th-century stained-glass windows*;

- **Where**: the location of the CH object, e.g. *Notre-Dame de Paris cathedral, Paris, France*;

- **When**: the time or period of the analysis processes, e.g. *between April 2019 and the summer of 2021*

- **Which**: the hardware or software components (e.g. instruments, devices, applications) and other supports used during the analysis, e.g. *New AGLAE facility; Opentheso software*;

- **How**: the methods, techniques, or procedures employed (the "action" sphere), e.g. *PIXE (Particle-Induced X-ray Emission); X-ray fluorescence spectrometry*;

- **Why**: the contextual details and purposes that motivated the performed actions, e.g. *Naming, relocating and dating the woods of Notre-Dame "forest" based on collated data and archaeological surveys of the remains*.

In a document-level CH Process event, multiple techniques are often applied to various CH objects or subsets, potentially using different instruments. To address this complexity, we adapted the W7 schema adding a depth level, to document each technique separately, specifying both **which** (instrument or tool) and **what** (the object or subset) for each technique. (Figure 8) presents an example of the expected structure of the output of such extraction. The results and implementation of this approach are presented in the next sections.



Figure 8: An example of the expected output of a CH Process event extraction using our W7-derived model.

2.2.3. Experiments and Results

This section presents the experiments done to assess the viability of our method. To assess the results, we used the ROUGE-L F1 as evaluation metric [52]. It was first calculated for each of the W7 question, after what we calculated the average score. ROUGE-L is determined by the longest common subsequence (LCS) between the LLM's output and the reference. For example:

- **LLM output:** La cathédrale Notre-Dame, Île-de-France
- **Reference:** La cathédrale Notre-Dame, 4e arrondissement de Paris, Île-de-France

Here, the LCS is "La cathédrale Notre-Dame, Île-de-France". Consequently, the ROUGE-L recall is 4/8 = 0.5 and the ROUGE-L precision is 1. The F1-score is then computed as:

$$ROUGE\text{-}L\ F1 = 2 * \frac{(precision * recall)}{(precision + recall)} = 2 * \frac{(0.5 * 1)}{(0.5 + 1)} \approx 0.67$$

To be noted that to perform DocEAE, the input PDF documents must be preprocessed. For this reason, our pipeline includes a conversion of their pages into Markdown-formatted text using PyMuPDF4LLM [53], containing both standard text and tabular data. During this conversion, any *Acknowledgments* and *References* sections are removed, as well as any *Appendix* content. However, when the input is a transcription coming from step 1 (see section 2.1), the text remain as is. In both cases, we finally check the whole text to ensure the information adheres to the desired format and remove any extraneous text.

We first conducted experiments on DocEAE using LLMs through two distinct methods (Figure 9):

1. Generating an answer for each W7 question using Retrieval-Augmented Generation (RAG) or Long-Context (LC). RAG is a technique that retrieves relevant information based on a query (question) and uses it to guide an LLM in generating a response. By directing attention to the most relevant segments, this approach helps the model avoid irrelevant details and reduces unnecessary computations. At the same time, large-scale pretraining enables LLMs to maintain strong long-context capabilities. When compared to the LC approach, RAG notably decreases the input length required for LLMs, resulting in lower computational costs.
2. Generating a structured Pydantic object that represents the metadata schema. Pydantic [54], a Python library for data validation, facilitates the creation of schemas ensuring that LLM responses conform to a predefined structure. By defining both the structure and properties of data, Pydantic models act as strict templates. Consequently, rather than returning free-form text, the model produces an object that precisely follows the Pydantic schema.

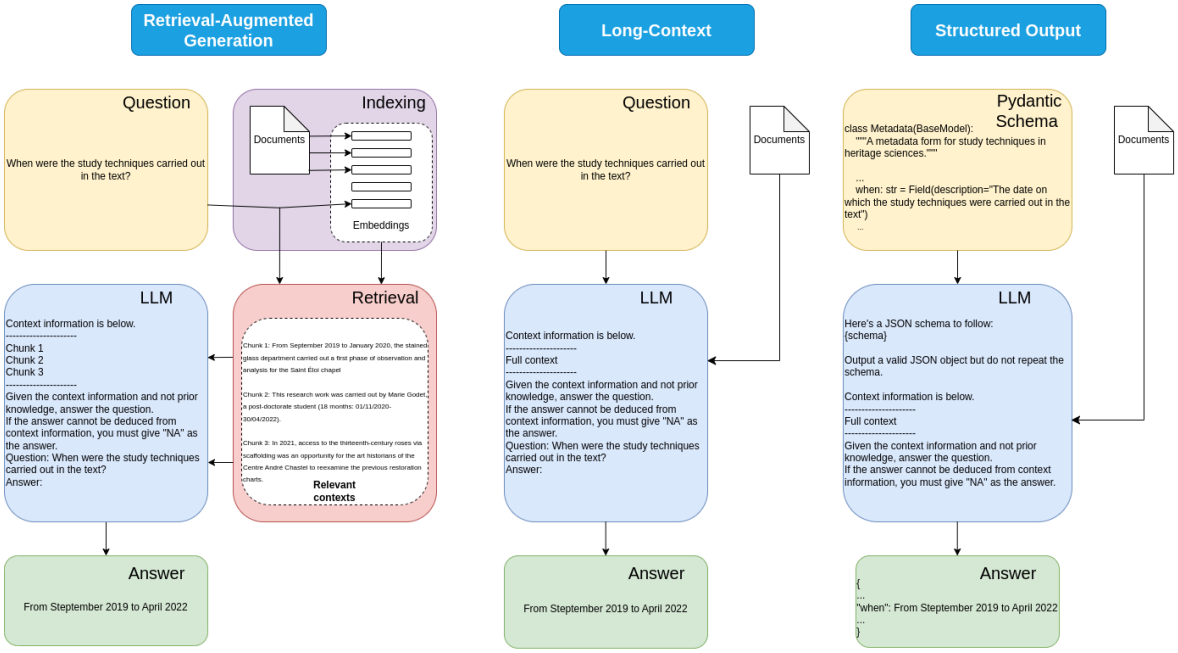


Figure 9: Overview of LLM Methods. Left: RAG consists of three main components: indexing, retrieval, and generation. Middle: LC includes only the generation stage. Right: The output of an LLM is structured as a Pydantic object.

Given our resource constraints, we selected llama.cpp for LLM serving [55], as it efficiently distributes layers between CPU and GPU memory, utilizing both for inference and minimizing GPU usage. All LLM models used in our experiments are in Q4_K_M quantization, as it provides a good balance between quality and model size [56]. We used the LlamaIndex framework [57] to build prompt-based LLM experiments. To compare which approach yields the best results, we used Mixtral 8x7B – Instruct [58] as the base LLM. For RAG, we employed gte-Qwen2-7B-instruct [59] as the embedding model, as it ranked first on the Massive Text Embedding Benchmark (MTEB) [60] for retrieval and reranking tasks (as of February 25, 2025).

Method	**ROUGE-L F1**	**Latency (seconds)**
LC	**0.56**	11
RAG	0.51	19
Structured Output	0.45	10

Table 4: Results of RAG, LC and Structured Output. LC outperforms RAG and Structured Output.

As shown in (Table 4), LC outperforms both RAG and Structured Output by a substantial margin, by 5% over RAG and 11% over Structured Output. This outcome underscores the exceptional long-context understanding capabilities of current LLMs. We observe that RAG underperforms partly because, for some documents, none of the retrieved passages contain the correct answer location. Additionally, a single prompt for Structured Output appears to be overly complex for the LLM, and the time required is nearly the same as using one prompt per question for LC (10s vs. 11s).

For *Who*, *Why*, and *Where* questions, answers are often more straightforward, as these details typically appear near the beginning of documents. In contrast, questions like *When*, *How* (*Which*, *What*) are more difficult, since their relevant information is usually scattered throughout the text. Addressing these requires the LLM to perform multi-step reasoning to reconstruct the complete answer—much like putting together the pieces of a puzzle. To enhance an LLM's reasoning ability, [61] suggests appending a simple prompt, "Let's think step by step", to encourage the model to engage in sequential thinking before generating a response. Recently, new internal reasoning models, such as DeepSeek R1 [62],

which achieves performance on a par with OpenAI-o1-1217, generate an internal chain of thought, effectively prompting themselves before producing a final answer. We conduct comprehensive LC evaluations on several strong, medium-sized models. Specifically, Mixtral 8x7B – Instruct, Qwen2.5-32B – Instruct and Qwen2.5-32B – Instruct [63], Llama-3.3-70B – Instruct [64], Gemma-3-27B-IT [65] employ the "Let's think step by step" prompt to facilitate step-by-step reasoning, while DeepSeek-R1-Distill-Llama-70B and DeepSeek-R1-Distill-Qwen-32B [62] adopt internal reasoning.

Method	**LLM**	**ROUGE-L F1**	**VRAM**	**Latency (seconds)**
Let's think step by step	Mixtral 8x7B – Instruct	0.57	32.51 GB	18
	Qwen2.5-72B – Instruct	**0.69**	54.86 GB	69
	Qwen2.5-32B – Instruct	0.57	28.91 GB	31
	Llama-3.3-70B – Instruct	0.57	53.47 GB	54
	Gemma-3-27B-IT	0.54	33.86 GB	29
Internal Thinking	DeepSeek-R1-Distill-Llama-70B	0.58	53.47 GB	87
	DeepSeek-R1-Distill-Qwen-32B	<u>0.62</u>	28.91 GB	57

Table 5: Results of the LC approach for two reasoning strategies. All models are in Q4_K_M quantization. The VRAM requirements take into account the model size, key-value (KV) caching, and compute buffer size. The best performance is shown in bold, and the second-best is underlined.

In (Table 5), we observe that Qwen2.5-72B – Instruct achieves the best performance, followed by DeepSeek-R1-Distill-Qwen-32B. Meanwhile, Mixtral 8x7B – Instruct shows only a 1% improvement over the non-reasoning approach, performing comparably to Qwen2.5-32B – Instruct and Llama-3.3-70B – Instruct. This finding suggests that reasoning is more effective in stronger (larger) models, aligning with previous study [61] that model size is critical for zero-shot reasoning.

These experiments allowed us to define our implementation framework. Given its performances, we opted for a long-context approach. In addition, despite its higher latency, the Qwen2.5-72B – Instruct good ROUGE-L F1 score inclined us to use it in the final implementation of the method. The implementation is described in section (see section 3). The output of this step is a JSON file, whose structure follows the W7-derived schema illustrated in (Figure 8). The (Figure 10) shows the result of this step, applied on a paper published by Mouaddib et. al. in Journal of Cultural Heritage Special Issue on Notre-Dame [66].

```
{
  "CH_Process_Event": {
    "WHO": [
      {
        "name": "El Mustapha Mouaddib",
        "role": "Laboratoire MIS-UPJV, 33, Rue Saint Leu, 80000 Amiens, France"
      },
      {
        "name": "Anthony Pamart",
        "role": "UMR 3495 MAP CNRS/MC, Campus CNRS, Bâtiment 31 chemin Joseph Aiguier, 13402, MARSEILLE cedex 20, France"
      },
      {
        "name": "Marc Pierrot-Deseilligny",
        "role": "LaSTIG lab at Université Gustave Eiffel, 73 Avenue de Paris, 94160 Saint-Mandé, France"
      },
      {
        "name": "Daniel Girardeau-Montaut",
        "role": "CloudCompare open-source project"
      }
    ],
    "WHY": {
      "Project": "2D/3D data fusion for the comparative analysis of the vaults of Notre-Dame de Paris before and after the fire"
    },
    "WHERE": "Notre-Dame Cathedral, Paris",
    "WHEN": "from April 2019 to 2023",
    "HOW": [
      {
        "technique": "Photogrammetry",
        "which": [
          "MicMac"
        ],
        "what": [
          "2D images of the vaults of Notre-Dame de Paris (before and after the fire)"
        ]
      },
      {
        "technique": "Laser Scanning (Lasergrammetry)",
        "which": [
          "Leica P20 scanner", "Faro focus 3D scanner", "CloudCompare"
        ],
        "what": [
          "3D point clouds of the vaults of Notre-Dame de Paris (before and after the fire)"
        ]
      },
      {
        "technique": "Data Fusion",
        "which": [
          "MicMac", "CloudCompare"
        ],
        "what": [
          "2D images and 3D point clouds of the vaults of Notre-Dame de Paris (before and after the fire)"
        ]
      },
      {
        "technique": "Visual and Quantitative Analysis",
        "which": [
          "MicMac", "CloudCompare"
        ],
        "what": [
          "Orthophotomosaics", "2.5D height maps", "contour plots", "3D meshes of the vaults of Notre-Dame de Paris (before and after the fire)"
        ]
      },
      {
        "technique": "2.5D Diagnoses",
        "which": [
          "MicMac", "CloudCompare"
        ],
        "what": [
          "2.5D height maps and contour plots of the vaults of Notre-Dame de Paris (before and after the fire)"
        ]
      }
    ]
  }
}
```


Figure 10: Result of the extraction process, performed on the paper from Mouaddib et. al., 2023 [66].

Although the extracted information is mostly accurate, our tests revealed occasional errors in the affiliation of the extracted individuals (*Who* field, *role* subfield). These discrepancies are likely due to ambiguities related to the document layout, which is variable across files. Additionnaly, the array structure of the values associated with the *What* field could be enhanced, as shown in (Figure 10). For example, the string "*2D images and 3D point clouds of the vaults of Notre-Dame de Paris (before and after the fire)*" could have been parsed into an array of length 4 as shown in (Figure 11). Nevertheless, the overall relevance of METAREVE results remain highly satisfactory, particularly given the considerable amount of completely undocumented HS data.

```
"what": [
  "2D images of the vaults of Notre-Dame de Paris before the fire",
  "2D images of the vaults of Notre-Dame de Paris after the fire",
  "3D point clouds of the vaults of Notre-Dame de Paris before the fire",
  "3D point clouds of the vaults of Notre-Dame de Paris after the fire"
]
```


Figure 11: The structuration of the extractions into comprehensive arrays still leaves room for improvement.

2.3. Step 3 – Entity linking

At this stage, the CH Events extracted from the previous step are well-documented enough to ensure at least a basic understanding of the analysis processes employed to study a CH object. This documentation can be used to proceed to the indexation of these analyses or of the data they generated, providing a comprehensive overview of HS workflows, approaches, and collaborations. However, to make this documentation interoperable and useful for research, it is important to remove possible ambiguities. To do so, we included a final step dedicated to entity linking. It aims to search for matches between the contents of the CH Events arguments and external references. In our case, we focused on specific thesauri provided by HS communities through the Opentheso instance of ESPADON ecosystem for vocabulary terms, and on the ORCID database for people.

Thus, for each person mentioned in a *Who* field, we first attempt to resolve their identity by querying the ORCID API [67] for an exact match. When a match is found, the corresponding unique identifier is added to the generated JSON metadata, enabling a direct link to the individual's public ORCID record. Similarly, elements appearing in the *How*, *Which*, *What*, *Where*, and *Why* categories can be checked against thesauri to identify possible correspondences with documented concepts. In order to avoid excessive processing time, this procedure originally relied exclusively on strict exact matching and was not executed automatically, but only on user request. Although a small number of valid matches were observed, the overall effectiveness of the method remained limited. The need to constrain runtime led to an overly rigid entity-linking mechanism that failed to capture many relevant associations. To address this issue, we redesigned the process by decoupling the linking step from the main extraction workflow and executing it asynchronously during the storage phase of the EAE output in the database. This restructuring significantly improved both performance and coverage.

The integration of METAREVE extractions into the knowledge base is based on a structured processing pipeline designed to produce a coherent, normalized graph aligned with stable semantic references (Figure 12). The CH Event, which constitutes the basic unit of the output, is extracted and then projected into an event collection as root node. An identifier is generated through a deterministic has (SHA-1) constructed from the invariant fields, which ensures that events are deduplicated and that existing entries can be automatically updated. This mechanism is crucial to ensure the structural stability of the graph and prevent the proliferation of redundant entities. Then, the seven dimensions of the event (*Who*, *Why*, *Where*, *When*, *How*, as well as the *Which* and *What* sub-structures), are analyzed and saved into separated dedicated collections. Each entity is normalized and filtered to remove instances lacking informative content, and saved through an upsert function using a deterministic identifier similar to that used for CH events. The relationships between these entities and the event root node are saved into an edge collection, with explicit typing.

In order to enable the semantic alignment, all thesaurus concepts from Opentheso are loaded and organized into a dictionary associating each normalized label with the list of corresponding concepts. This structure reduces the algorithmic complexity of matching processes and avoids the repetition of queries to the database. When a node contains textual contents (e.g. a technique, an instrument, a goal, etc.), its nominal fields are compared to the thesaurus dictionary. We explicitly distinguish between

exact matches (strict equality between the normalized text of the node and a concept label) and partial matches (presence of a concept label as a lexical unit inside a longer expression). To avoid false positives, the detection of partial occurrences is based on lexical boundaries, preventing the recognition of a label within a word (e.g. “arc” in “architecture”). Each match is saved as a thesaurus-typed relationship with metadata specifying the original field, the associated label, and the exact or partial nature of the match. In case of an exact match, the persistent identifier of the matched concept (usually an ARK ID) is reinjected into the corresponding JSON element as a new URI attribute. Conversely, partial matches result in the creation of relations without modifying the original node, in accordance with a cautious alignment policy that distinguishes between the categorical recognition of a stable concept and the mere presence of a relevant term in a broader expression.

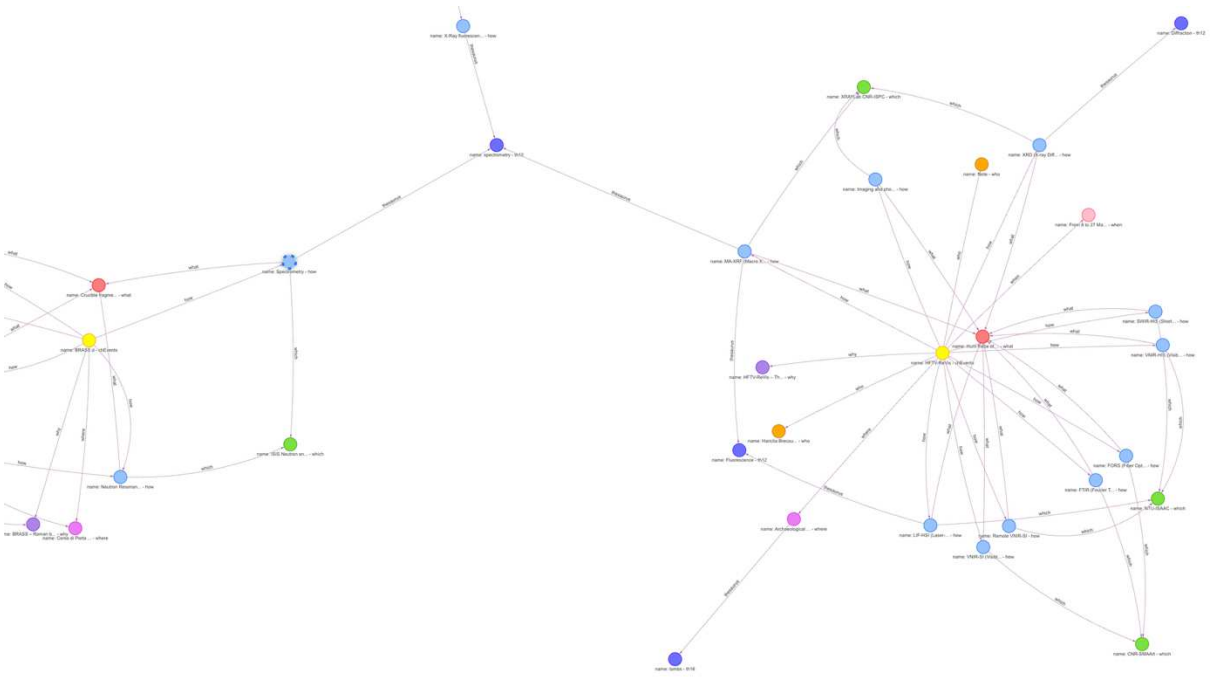


Figure 12: Example of results. Excerpt from the graph produced by processing several scientific publications from Journal of Cultural special issue “Notre-Dame de Paris: a multidisciplinary scientific site”. The CH events nodes (in yellow) are linked to Who (orange), Why (purple), Where (pink), When (light red), How (light blue), What (red), Which (green), and thesaurus concepts (dark blue) nodes.

3. Implementation

In terms of implementation, we opted for microservices-based architecture, with a dedicated orchestration layer (Figure 13). The application is delivered as a single-page web interface built with React [68] and Vite [69], where users can upload either audio or text files, trigger processing, and then monitor the status of each job through a dedicated “file tracking” page that lists completed and in-progress analyses together with an editable view to the resulting JSON metadata (Figure 14).

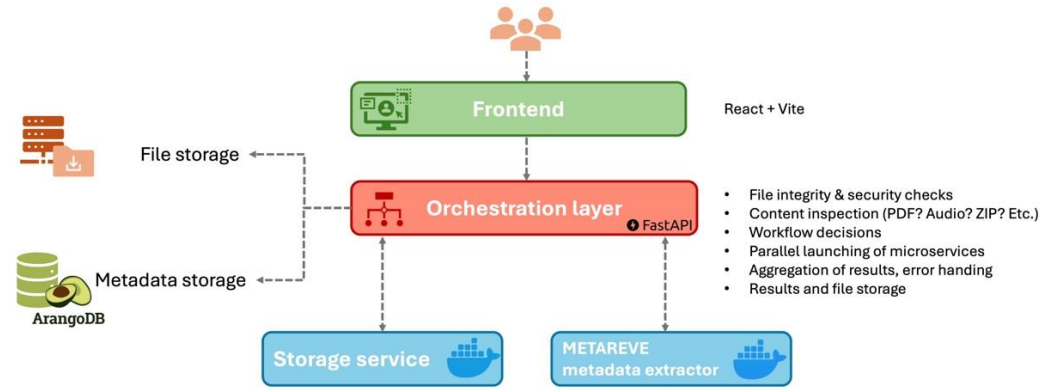


Figure 13: METAREVE architecture.

The orchestrator, powered by FastAPI, performs integrity and security checks, inspects file types, and dynamically triggers the appropriate processing pipelines. Depending on the uploaded content, it may launch multiple microservices in parallel—for example a CPU-based file conversion service and, for PDF or audio files, the METAREVE metadata extraction service. The orchestrator aggregates the results, stores structured outputs in ArangoDB, and moves the original files to a dedicated storage server. At a time when initiatives are flourishing to create digital tools and even entire ecosystems for HS, this design provides flexible, scalable, and resource-efficient processing across heterogeneous compute nodes, and makes it possible to anticipate the need to gradually integrate new processing pipelines.

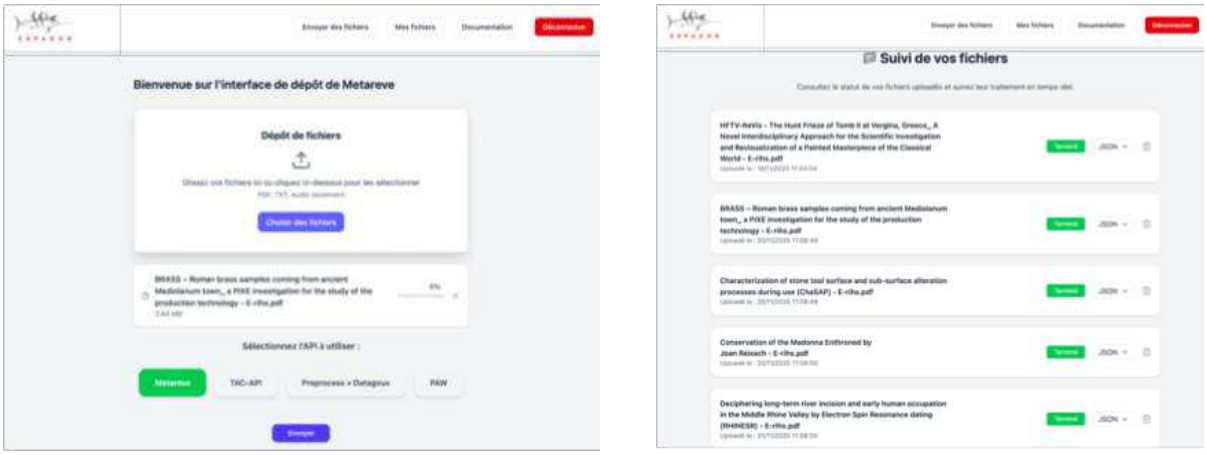


Figure 14: The single-page user interface, allowing to upload (left) and monitor the status of each job (right).

The METAREVE microservice is hosted on the same GPU server as the Whisper and Qwen models, so that file upload, transcription, metadata extraction and post-correction are orchestrated within a single deployment. Each extraction step can be called independently through a POST request:

- /*upload*: route allowing a file to be sent to the GPU server before processing.
- /*transcribe*: route launching the transcription of an audio file (.wav, .mp3).
- /*metadata*: route launching metadata extraction from a PDF or TXT file.

All extracted metadata are stored as JSON documents in an ArangoDB database [70], which exposes a simple job abstraction: for example, a route such as GET /jobs returns all jobs associated with the current user, while endpoints like GET /jobs/{job_id} or DELETE /jobs/{job_id} allow clients to retrieve or manage individual results. This architecture provides a clean separation between the interactive UI and the heavy computation, allowing to scale each element independently depending on needs, while making it straightforward to query, filter and reuse the produced JSON across downstream tools.

4. Perspectives and limits

In this paper we presented the METAREVE, the principles at the basis of its design, its main functionalities and current implementation. The application was tested using a data subset from Notre-Dame de Paris scientific action, mainly expert reports and scientific publications. The results are promising: the system reveals information about study events that had previously not been indexed or referenced in the project's databases. Naturally, the extracted information cannot match the level of detail provided by a manual metadata-enrichment process based on a schema tailored to a specific data type. However, the granularity achieved through METAREVE proves fully adequate for offering an overview of available resources, past analysis activities, and their interrelationships, thereby significantly enhancing their potential for subsequent exploitation. The intelligibility of the W7 schema, which directly encodes the seven most common questions, greatly simplifies the formulation of queries used to filter data across the corpus. The (Figure 15) illustrates examples of queries derived from simple natural-language questions. For heritage sciences, characterized by its wide range of disciplines and an even greater variety of investigative methods, such automatic extraction of metadata around this pivot schema is particularly valuable. It enables the harmonization of highly diverse resources while still preserving their specific features. The work of [27] clearly demonstrates how such approach can be articulated with finer granularities. Here, METAREVE effectively forms the level 0 of a Provenance-LOD.

Question	**AQL query**

List all CH events related to the concept "Spectrometry" in th12.	`FOR concept IN th12`
	`  FILTER LOWER(concept.name) == "spectrometry" OR "spectrometry" IN concept.labels[*].label`
	`  FOR v IN 1..2 INBOUND concept`
	`  GRAPH "linkedmetareve"`
	`  FILTER v._id LIKE "chEvents/%"`
	`  RETURN DISTINCT v`

On which CH object was used spectrometry?	`FOR concept IN th12`
	`  FILTER LOWER(concept.name) == "spectrometry" OR "spectrometry" IN concept.labels[*].label`
	`  FOR v IN 2..2 ANY concept`
	`  GRAPH "linkedmetareve"`
	`  FILTER v._id LIKE "what/%"`
	`  RETURN DISTINCT v`

Who was involved in studies on Notre-Dame?	`FOR loc IN where`
	`  FILTER loc.name != null`
	`  FILTER LIKE(LOWER(loc.name), "%notre-dame%", true)`
	
	`  FOR v IN 1..1 INBOUND loc`
	`  GRAPH "linkedmetareve"`
	`  FILTER v._id LIKE "chEvents/%"`
	
	`  FOR p IN 1..1 OUTBOUND v`
	`  GRAPH "linkedmetareve"`
	`  FILTER p._id LIKE "who/%"`
	`  RETURN DISTINCT p`

Figure 15: Example of AQL queries derived from simple questions.

Beyond the methodological and technical challenges, we consider this work as a first step towards the definition of bottom-up approaches allowing to lighten the workload of data documentation in multidisciplinary contexts. Of course, such data-driven methods are not intended to replace the hypothetical-deductive approaches of HS, but rather to complement them.

This project is still a work in progress and presents avenues for improvement. First, the calculation time (on average 17.2 seconds per 10 pages) is reasonable but could certainly be optimized. The ASR post-correction step could also gain in robustness, for example by replacing the local database with an approach relying on stable repositories (ORCID, IDREF, etc.). Furthermore, our roadmap includes the introduction of normalization functions, so that extracted information can be efficiently queried and exploited in digital environments. While a property graph clearly offers a valuable way to enhance the exploitation of such complex corpora (Figure 12), building the relationships linking data together require the ability to detect dimensional similarities. This is especially true regarding spatial and temporal information (*Where* and *When* fields), which are often expressed locally or even implicitly. The *Who* field also could be improved, partly because we currently limit the extraction to the individual experts involved in the CH Event, whereas it might be useful to extract other mentions as well (e.g. institutional partners, people being represented in the CH object, artists, etc.). Its linking is limited to ORCID and could be extended to other knowledge bases (e.g. an alignment to Wikidata would be useful regarding historical figures). Similarly, the *What* fields could be made less ambiguous if linked to the artwork catalogues of CH institutions. More generally, while the current linking strategy has proven its value by highlighting some similarities between the processed documents, it should now be partially redesigned to better comply with the principles of Linked Open Data [71]. Finally, although this point has been neglected for this first version, if available and stored in non-proprietary formats, a reading of the manufacturer parameters on raw analysis data would certainly highly enhance the automatic extraction results.

Conclusion

After decades of attempts to provide user-friendly interfaces for documenting increasingly massive and complex HS datasets, we decided to adopt the opposite approach. The heterogeneity of heritage science research topics, methodologies, and data has introduced a level of complexity that has resulted in proportionately lengthy metadata schemas. As a result, for data producers, the challenges of provenance documentation brought about by open science are hardly sustainable. Consequently, much HS data remains undocumented, making it impossible to share and reuse these valuable resources. Through METAREVE, we propose a complementary data-driven approach, which aims to simplify part of this work and limit the loss of undocumented data. In this paper, we presented our method for transcription, metadata extraction, and linking, which enables the construction of graphs from extracted knowledge. We first showed how fine-tuning Whisper and post-correction made it possible to address the challenges of transcribing audio files in the HS domain. Next, we presented our NLU method for metadata extraction. Our document-level event argument extraction approach relies on W7, a pivot metadata schema whose flexibility has proven useful for HS. We showed that a “let’s think step by step” long context approach yielded the best results on our data. Finally, we presented our linking method and the implementation of the tool, as a microservices-based architecture orchestrated with a FastAPI layer. Beyond all these methodological and technical aspects, METAREVE, now needs to be tested in real-world conditions by its future end-users, the HS data producers. The development of a user interface would simplify the learning curve and facilitate its adoption. A complementary tool allowing users to easily perform queries on extracted data through predefined widgets or conversational agent is currently under study.

Acknowledgements

This work is dedicated to the memory of our esteemed colleague Vincent Detalle, whose open-mindedness and scientific curiosity were a great source of inspiration. The authors would like to thank Olivier Malavergne, Ruven Pillay, Anaïs Guillem, Livio De Luca and Emmanuel Poirault for their valuable support.

CrediT Taxonomy

Contributors: Violette Abergel (VA), Van Tuan Bui (VTB), Amine Berbagui (AB), Vincent Detalle (VD), Olivier Malavergne (OM), Ruven Pillay (RP), Anaïs Guillem (AG), Fondation des Sciences du Patrimoine (FSP).

Contributions: Conceptualization: VA, VD; Data curation: VA, VTB, AB; Formal analysis: VA, AG; Funding acquisition: VA, VD, OM; Investigation: VA, VTB, AB; Methodology: VA, VTB, AB; Project administration: VA; Resources: FSP; Software: VTB, AB, VA; Supervision: VA, OM, RP; Validation: VA, VTB, AB; Visualization: VTB, AB; Writing – original draft: VA, VTB, AB

Funding statement

This work was supported by the Paris Seine Graduate School Humanities, Creation, Heritage, Investissement d'Avenir ANR-17-EURE-0021 – Fondation for cultural heritage sciences

This work benefited from State aid managed by the Agence Nationale de la Recherche (French National Research Agency) under the Investment for the Future Program integrated into France 2030, bearing the reference ANR 21-ESRE-0050 EquipEx+ ESPADON.

Authors contributions statement

Authors: Violette Abergel (VA), Van Tuan Bui (VTB), Amine Berbagui (AB).
VA conceived the theoretical framework. VA, VTB, and AB designed the methodology, performed the experiments, and analyzed the data. VA, VTB, and AB wrote the original draft of the manuscript. All authors reviewed and approved the final manuscript.

Data availability statement

The code and data related to this research are available at this link:
https://gitlab.espadon.net/Espadon/metareve

Clinical Trial Number

NOT APPLICABLE.

Ethics declaration

NOT APPLICABLE.

Consent to Publish declaration

NOT APPLICABLE.

Consent to Participate declaration

NOT APPLICABLE.

References

[1] M. D. Wilkinson *et al.*, 'The FAIR Guiding Principles for scientific data management and stewardship', *Sci. Data*, vol. 3, no. 1, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.

[2] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. Da Silva Santos, and M. D. Wilkinson, 'Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud', *Inf. Serv. Use*, vol. 37, no. 1, pp. 49–56, Feb. 2017, doi: 10.3233/ISU-170824.

[3] R. Thomas *et al.*, 'The augmented heritage object: An integrated paradigm for multi-disciplinary data and knowledge production in heritage science', *J. Cult. Herit.*, vol. Submitted, under review, 2025.

[4] M. Rousset, *Opentheso · GitLab*. (Sept. 27, 2018). Java, PLpgSQL, HTML, JavaScript. Maison de l'Orient et de la Mediterrannée (MOM), CNRS. Accessed: Dec. 12, 2025. [Online]. Available: https://gitlab.huma-num.fr/mrousset/opentheso

[5] P. Harpring, 'Development of the Getty Vocabularies: AAT, TGN, ULAN, and CONA', *Art Doc. J. Art Libr. Soc. N. Am.*, vol. 29, no. 1, pp. 67–72, Apr. 2010, doi: 10.1086/adx.29.1.27949541.

[6] D. Vrandečić, 'Wikidata: a new platform for collaborative data collection', in *Proceedings of the 21st International Conference on World Wide Web*, Lyon France: ACM, Apr. 2012, pp. 1063–1064. doi: 10.1145/2187980.2188242.

[7] A. Néroulidis *et al.*, 'A digital platform for the centralization and long-term preservation of multidisciplinary scientific data belonging to the Notre Dame de Paris scientific action', *J. Cult. Herit.*, vol. 65, pp. 210–220, Jan. 2024, doi: 10.1016/j.culher.2023.09.016.

[8] Ministère de la Culture et de la Communication - Secrétariat général, 'Métadonnées culturelles et transition Web 3.0.', Ministère de la Culture et de la Communication, Paris, France, N°2014-01, 2014.

[9] M. Doerr, 'The CIDOC CRM, an Ontological Approach to Schema Heterogeneity', p. 5p, 2005, doi: 10.4230/DAGSEMPROC.04391.22.

[10] M. Doerr, A. Kritsotaki, Y. Rousakis, G. Hiebel, and M. Theodoridou, 'CRMsci: the Scientific Observation Model An Extension of CIDOC-CRM to support scientific observation', *Prod. FORTH Collab. Version*, vol. 1, no. 1, 2014.

[11] M. Doerr and M. Theodoridou, '{CRMdig}: A generic digital provenance model for scientific observation', in *3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP 11)*, 2011. Accessed: Nov. 19, 2025.

[12] G. Hiebel, M. Doerr, and Ø. Eide, 'CRMgeo: A spatiotemporal extension of CIDOC-CRM', *Int. J. Digit. Libr.*, vol. 18, no. 4, pp. 271–279, Nov. 2017, doi: 10.1007/s00799-016-0192-4.

[13] I. Bannour *et al.*, 'CRM$_{CR}$ - a CIDOC-CRM extension for supporting semantic interoperability in the conservation and restoration domain', in *2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018)*, San Francisco, CA, USA: IEEE, Oct. 2018, pp. 1–8. doi: 10.1109/DigitalHeritage.2018.8810098.

[14] D. U. Board, 'Dublin Core Metadata Element Set, Version 1.1'. 2008. [Online]. Available: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[15] P. W. G. W3C, 'PROV: A Provenance Model'. 2013. Accessed: Dec. 12, 2025. [Online]. Available: https://www.w3.org/TR/prov-overview/

[16] C. Lagoze, H. Van De Sompel, S. Warner, and The Open Archives Initiative, 'OAI-ORE: A Model for Data-Intensive Services'. 2015. [Online]. Available: http://www.openarchives.org/ore/

[17] C. Niang *et al.*, ‘Supporting Semantic Interoperability in Conservation-Restoration Domain: The PARCOURS Project’, *J. Comput. Cult. Herit.*, vol. 10, no. 3, pp. 1–20, Aug. 2017, doi: 10.1145/3097571.
[18] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa, ‘An ontology for describing and synthesizing ecological observation data’, *Ecol. Inform.*, vol. 2, no. 3, pp. 279–296, Oct. 2007, doi: 10.1016/j.ecoinf.2007.05.004.
[19] M. Könnecke *et al.*, ‘The NeXus data format’, *J. Appl. Crystallogr.*, vol. 48, no. 1, pp. 301–305, Feb. 2015, doi: 10.1107/S1600576714027575.
[20] D. Myers, A. Dalgity, I. Avramides, and D. Wuthrich, ‘Arches: An Open Source GIS for the Inventory and Management of Immovable Cultural Heritage’, in *Progress in Cultural Heritage Preservation*, vol. 7616, M. Ioannides, D. Fritsch, J. Leissner, R. Davies, F. Remondino, and R. Caffo, Eds, in Lecture Notes in Computer Science, vol. 7616. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 817–824. doi: 10.1007/978-3-642-34234-9_87.
[21] A. Isaac and B. Haslhofer, ‘Europeana Linked Open Data – data.europeana.eu’, *Semantic Web*, vol. 4, no. 3, pp. 291–297, 2013, doi: 10.3233/SW-120092.
[22] A. D’Andrea and K. Fernie, ‘CARARE 2.0: A metadata schema for 3D cultural objects’, in *2013 Digital Heritage International Congress (DigitalHeritage)*, Marseille, France: IEEE, Oct. 2013, pp. 137–143. doi: 10.1109/DigitalHeritage.2013.6744745.
[23] C. Meghini *et al.*, ‘ARIADNE: A Research Infrastructure for Archaeology’, *J. Comput. Cult. Herit.*, vol. 10, no. 3, pp. 1–27, Aug. 2017, doi: 10.1145/3064527.
[24] I. Pallot-Frossard, ‘From CHARISMA to IPERION CH and E-RIHS: towards a European research infrastructure for heritage science’, *Technè*, vol. 43, pp. 46–49, 2016, doi: 10.4000/techne.641.
[25] A. Bucciero *et al.*, ‘H2IOSC Project: The Italian Federated Cluster for IoT-Based Monitoring and Digital Twinning of Cultural Heritage’, in *2024 9th International Conference on Smart and Sustainable Technologies (SpliTech)*, Bol and Split, Croatia: IEEE, June 2024, pp. 1–6. doi: 10.23919/SpliTech61897.2024.10612510.
[26] L. Castelli, A. Felicetti, and F. Proietti, ‘Heritage Science and Cultural Heritage: standards and tools for establishing cross-domain data interoperability’, *Int. J. Digit. Libr.*, vol. 22, no. 3, pp. 279–287, Sept. 2021, doi: 10.1007/s00799-019-00275-2.
[27] A. Guillem, V. Abergel, R. Roussel, F. Comte, A. Pamart, and L. De Luca, ‘Bridging the Provenance Knowledge Gap Between 3D Digitization and Semantic Interpretation’, *Heritage*, vol. 8, no. 11, p. 476, Nov. 2025, doi: 10.3390/heritage8110476.
[28] A. Pamart, L. De Luca, and P. Véron, ‘A metadata enriched system for the documentation of multi-modal digital imaging surveys’, *Stud. Digit. Herit.*, vol. 6, no. 1, pp. 1–24, Dec. 2020, doi: 10.14434/sdh.v6i1.33767.
[29] D. Myers, A. Dalgity, and I. Avramides, ‘The Arches heritage inventory and management system: a platform for the heritage field’, *J. Cult. Herit. Manag. Sustain. Dev.*, vol. 6, no. 2, pp. 213–224, Aug. 2016, doi: 10.1108/JCHMSD-02-2016-0010.
[30] S. Tournon, R. Vergnieux, and B. Dutailly, ‘Comment gérer les projets 3D collaboratifs en SHS ?’, in *Humanistica 2020*, Bordeaux, France, May 2020. Accessed: Nov. 19, 2025.
[31] J. Li, A. Sun, J. Han, and C. Li, ‘A Survey on Deep Learning for Named Entity Recognition’, *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.
[32] D. Xu *et al.*, ‘Large language models for generative information extraction: a survey’, *Front. Comput. Sci.*, vol. 18, no. 6, p. 186357, Dec. 2024, doi: 10.1007/s11704-024-40555-y.
[33] D. Lu, S. Ran, J. Tetreault, and A. Jaimes, ‘Event Extraction as Question Generation and Answering’, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 1666–1688. doi: 10.18653/v1/2023.acl-short.143.

[34] X. Du and C. Cardie, ‘Event Extraction by Answering (Almost) Natural Questions’, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2020, pp. 671–683. doi: 10.18653/v1/2020.emnlp-main.49.
[35] M. Tong *et al.*, ‘DocEE: A Large-Scale and Fine-grained Benchmark for Document-level Event Extraction’, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, 2022, pp. 3970–3982. doi: 10.18653/v1/2022.naacl-main.291.
[36] X. Wei *et al.*, ‘ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT’, 2023, *arXiv*. doi: 10.48550/ARXIV.2302.10205.
[37] S. Li, H. Ji, and J. Han, ‘Document-Level Event Argument Extraction by Conditional Generation’, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, 2021, pp. 894–908. doi: 10.18653/v1/2021.naacl-main.69.
[38] C. Liang, ‘Event Argument Extraction with Enriched Prompts’, 2025, *arXiv*. doi: 10.48550/ARXIV.2501.06825.
[39] S. van Hooland, M. D. Wilde, R. Verborgh, T. Steiner, and R. V. de Walle, ‘Named-Entity Recognition: A Gateway Drug for Cultural Heritage Collections to the Linked Data Cloud?’.
[40] S. Van Hooland, M. De Wilde, R. Verborgh, T. Steiner, and R. Van De Walle, ‘Exploring entity recognition and disambiguation for cultural heritage collections’, *Digit. Scholarsh. Humanit.*, vol. 30, no. 2, pp. 262–279, June 2015, doi: 10.1093/llc/fqt067.
[41] A. Felicetti, D. Williams, I. Galluccio, D. Tudhope, and F. Niccolucci, ‘NLP Tools for Knowledge Extraction from Italian Archaeological Free Text’, in *2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018)*, San Francisco, CA, USA: IEEE, Oct. 2018, pp. 1–8. doi: 10.1109/DigitalHeritage.2018.8810001.
[42] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, and A. Doucet, ‘Named Entity Recognition and Classification on Historical Documents: A Survey’, 2021, doi: 10.48550/ARXIV.2109.11406.
[43] A. Bombini, A. Alkhansa, L. Cappelli, A. Felicetti, F. Giacomini, and A. Costantini, ‘A Cloud-Native Web Application for Assisted Metadata Generation and Retrieval: THESPIAN-NER’, *Appl. Sci.*, vol. 12, no. 24, p. 12910, Dec. 2022, doi: 10.3390/app122412910.
[44] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, ‘Automatic speech recognition: a survey’, *Multimed. Tools Appl.*, vol. 80, no. 6, pp. 9411–9457, Mar. 2021, doi: 10.1007/s11042-020-10073-7.
[45] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, ‘Robust Speech Recognition via Large-Scale Weak Supervision’, in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds, in Proceedings of Machine Learning Research, vol. 202. PMLR, July 2023, pp. 28492–28518.
[46] S. Mangrulkar *et al.*, ‘PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods’. 2022. [Online]. Available: https://github.com/huggingface/peft
[47] E. J. Hu *et al.*, ‘LoRA: Low-Rank Adaptation of Large Language Models’, Oct. 16, 2021, *arXiv*: arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685.
[48] Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters, *explosion/spaCy: v3.7.2: Fixes for APIs and requirements*. (Oct. 16, 2023). Zenodo. doi: 10.5281/ZENODO.1212303.
[49] J. Liu, *W7 Model of Provenance and its Use in the Context of Wikipedia*. The University of Arizona, 2011. Accessed: Dec. 09, 2025. [Online]. Available: https://search.proquest.com/openview/36e7ec080d0e54c0c17203aef855ca61/1

[50] M. Huisman *et al.*, ‘A perspective on Microscopy Metadata: data provenance and quality control’, 2019, *arXiv*. doi: 10.48550/ARXIV.1910.11370.
[51] T. Auge, S. Genehr, M. K. and, F. Krüger, and M. Schröder, ‘Towards dimensions and granularity in a unified workflow and data provenance framework’, Apr. 15, 2025, *arXiv*: arXiv:2504.11278. doi: 10.48550/arXiv.2504.11278.
[52] C.-Y. Lin, ‘Rouge: A package for automatic evaluation of summaries’, in *Text summarization branches out*, 2004, pp. 74–81.
[53] J. X. McKie, *pymupdf/pymupdf4llm: PyMuPDF4LLM*. (2025). Python. Accessed: Dec. 09, 2025. [Online]. Available: https://github.com/pymupdf/pymupdf4llm
[54] S. Colvin *et al.*, *Pydantic*. (June 14, 2025). Python. Zenodo. doi: 10.5281/ZENODO.15662245.
[55] *ggml-org/llama.cpp*. (Dec. 09, 2025). C++. ggml. Accessed: Dec. 09, 2025. [Online]. Available: https://github.com/ggml-org/llama.cpp
[56] ‘k-quants by ikawrakow · Pull Request #1684 · ggml-org/llama.cpp’, GitHub. Accessed: Dec. 09, 2025. [Online]. Available: https://github.com/ggml-org/llama.cpp/pull/1684
[57] J. Liu, *LlamaIndex*. (Nov. 2022). Python. doi: 10.5281/zenodo.1234.
[58] A. Q. Jiang *et al.*, ‘Mixtral of Experts’, 2024, *arXiv*. doi: 10.48550/ARXIV.2401.04088.
[59] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang, ‘Towards General Text Embeddings with Multi-stage Contrastive Learning’, 2023, *arXiv*. doi: 10.48550/ARXIV.2308.03281.
[60] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, ‘MTEB: Massive Text Embedding Benchmark’, in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 2014–2037. doi: 10.18653/v1/2023.eacl-main.148.
[61] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, ‘Large language models are zero-shot reasoners’, *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 22199–22213, 2022.
[62] DeepSeek-AI *et al.*, ‘DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning’, 2025, *arXiv*. doi: 10.48550/ARXIV.2501.12948.
[63] Q. Team, ‘Qwen2.5: A Party of Foundation Models’. Sept. 2024. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5/
[64] A. Grattafiori *et al.*, ‘The Llama 3 Herd of Models’, 2024, *arXiv*. doi: 10.48550/ARXIV.2407.21783.
[65] Gemma Team *et al.*, ‘Gemma 3 Technical Report’, 2025, *arXiv*. doi: 10.48550/ARXIV.2503.19786.
[66] E. M. Mouaddib, A. Pamart, M. Pierrot-Deseilligny, and D. Girardeau-Montaut, ‘2D/3D data fusion for the comparative analysis of the vaults of Notre-Dame de Paris before and after the fire’, *J. Cult. Herit.*, vol. 65, pp. 221–231, Jan. 2024, doi: 10.1016/j.culher.2023.06.012.
[67] R. Blackburn *et al.*, *ORCID/ORCID-Source at development*. (2025). Java, JavaScript, FreeMarker, TypeScript. Accessed: Dec. 09, 2025. [Online]. Available: https://github.com/ORCID/ORCID-Source/tree/development
[68] *facebook/react*. (Dec. 09, 2025). JavaScript. Meta. Accessed: Dec. 09, 2025. [Online]. Available: https://github.com/facebook/react
[69] *vitejs/vite*. (Dec. 09, 2025). TypeScript. Vite. Accessed: Dec. 09, 2025. [Online]. Available: https://github.com/vitejs/vite
[70] *arangodb/arangodb*. (Dec. 09, 2025). C++. ArangoDB. Accessed: Dec. 09, 2025. [Online]. Available: https://github.com/arangodb/arangodb
[71] T. Berners-Lee, ‘Linked Data - Design Issues’, W3 - Design Issues, Linked Data. Accessed: Dec. 09, 2025. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html