Towards Semantic Retrieval over Conferences Recordings: Automatic Speech Recognition Results Analysis

Solenn Tual[1], Rodolphe Bailly[2], Rinaldo Lima[3] and Cassia Trojahn[1]

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France
[2]Philharmonie de Paris, Cité de la Musique, F-75019 Paris
[3]Univ. Federal Rural of Pernambuco, Department of Computing, Recife, Brazil

Abstract

The Philharmonie de Paris – Cité de la Musique maintains an extensive digital archive of transcribed conference recordings in a wide range of disciplines. The access to this archive has become a key objective. The Music Graph RAG (MGR) early stage project aims at developing a semantic retrieval tool to support fine-grained search of large-scale multimedia archives. This paper presents a qualitative analysis of the transcribed conferences produced with an automatic speech recognition framework. This step is crucial as the transcriptions errors directly influence the performance of all subsequent stages in the processing pipeline and technical choices.

Keywords

Music-related documents, ASR noise, Discourse structuring, Knowledge graph construction, Document retrieval

1. Introduction

The Philharmonie de Paris – Cité de la Musique is one of Europe's leading music institutions with its 3 concert halls, 2 orchestras, museum, education and knowledge departments. Alongside its artistic mission, it maintains an extensive and growing digital archive of conference recordings covering a wide variety of subjects, such as musicology, music history, composition, organology and social sciences. This collection is currently being expanded to include transcripts, with around 617 conferences having been processed to date, representing approximately 1,094 hours of audiovisual content. Improving discoverability to this large-scale multimedia archive is a key objective.

The Music Graph RAG – Retrieval-Augmented Generation – research project (MGR) aims at providing semantic retrieval tools that allow users to query the corpus and access excerpts corresponding to their information needs (e.g., specific topic, artist), thereby facilitating the exploitation of extensive collections of recorded lectures. The overall pipeline of processing and retrieval is composed of the following main steps: (i) transcription segmentation (using chunking and discourse structure recognition approaches); (ii) semantic interpretation (named entity recognition, entity resolution and linking); (iii) graph and embedding-based representation (including knowledge graph construction modelling discourse structure), and finally (iv) transcription excerpts retrieval using hybrid approaches (exploiting explicit entities and documents relations and similarity comparison in latent space). This paper overviews the different approaches and methods in the literature that can be exploited in each step of the pipeline, and adresses a qualitative and manual analysis of the corpus of transcribed recording conferences produced with an automatic speech recognition (ASR) framework. This analysis is crucial as the transcriptions errors of the corpora directly influence the performance of all subsequent stages in the processing pipeline and technical choices.

The paper is structured as follows. Section 2 discusses related work. Section 3 details the overall pipeline. Section 4 describes the corpus and the qualitative analysis of the transcriptions. Section 5 concludes the paper.

SWODCH'26: International Workshop on Semantic Web and Ontology Design for Cultural Heritage, Workshop days of ISWC 2026, Bari, Italy

solenn.tual@inria.fr (S. Tual); rbailly@cite-musique.fr (R. Bailly); rinaldo.jose@ufrpe.br (R. Lima); cassia.trojahn-dos-santos@univ-grenoble-alpes.fr (C. Trojahn)
0000-0001-8549-7949 (S. Tual); 0000-0002-1388-4824 (R. Lima); 0000-0003-2840-005X (C. Trojahn)

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

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-09

CEUR-WS.org/Vol-4320/paper4.pdf

2. Related Work

Chunking strategies Retrieval approaches benefit significantly from highly structured, interconnected input documents [1]. However, creating these connections via tree- or graph-based methods has limitations [2]. Specifically, window- or boundary-based chunking strategies fragment the discourse and disconnect coherent subparts, a critical flaw for domain-specific documents requiring complete and coherent excerpts. Recent works evaluate various chunking strategies and their retrieval effectiveness [3, 4]. *Zhou et al.* [3] categorize these approaches into two groups. Structure-based strategies encompass text-structure splitting (into paragraphs or sentences) and fixed-size-window splitting (into tokens or words). Meanwhile, semantic or LLM-guided strategies include sentence embedding similarity comparisons to detect shifts , alongside LLM-based approaches for detecting elementary propositions [5] or topic shifts [6]. The latter captures discourse structure, which is highly useful for document metadata or chaptering. Recent surveys conclude that chunking performance hightly depends on the specific task and corpus [3, 4, 7]. Furthermore, *Qu et al.* [7] highligth evaluation limitations, particularly the scarcity of appropriate metrics and real-world long-document evaluation corpora.

Topic-based discourse segmentation Works from the field of NLP aim to segment text preserving the rhetorical flow, which is crucial to structure documents like conference. This segmentation spans various scales, from annotating elementary discourse units (EDUs) [8, 9] to automated table-of-contents or chaptering generation for a macro-level structural overview [10, 11]. Intermediate structural levels also exists. In particular, topic segmentation divides documents into contiguous, topically unified sentence groups separated by topic shifts and speaker turns. It seems particularly relevant to create a graph that contains both the structure and the semantic relations between documents parts. This task involves either linear or hierarchical segmentation, and its outputs often feed larger summarization or information extraction pipelines. While early unsupervised approaches rely on lexical comparison [12], topic modeling [13], or embedding-based sentence similarity [14], modern supervised methods leverage deep neural networks. Entity coreference resolution can also help to identify sentences that refer to the same underlying subject [15]. , involving however significant computational costs, especially for long documents. Lexical chains based approaches could be employed to reduce such cost [16].

Graph construction from unstructured texts Graph-based indexing of documents is a relevant way to perform retrieval tasks. Graphs have been identified as a flexible and efficient way to represent both structure and knowledge of various types of documents (such as encyclopedic articles or scientific publications) along with creating connections between them [17]. Some methodologies generate document graphs using OpenIE approaches [18] which are also used in modern retrieval end-to-end approaches [19]. A wide range of graph construction or enrichment approaches rely on Named Entity Recognition (NER) [20, 21], resolution (ER) and linking (EL) [22, 23] to create relations between documents (and their subparts) based on those mentions and connect them to external knowledge [24]. The development of large language models has brought new possibilities to support those tasks [25].

Domain ontologies and knowledge graphs Recent retrieval approaches exploit the synergy between dense and symbolic representations to retrieve similar or topic-related content. The later representation implies the construction of suitable knowledge graphs adapted to the domain they represent as associated documents contain highly specialised vocabularies and rare entities that should be accurately modelled. In the musical domain, several projects have developed approaches to make documents indexable or resources searchable using graphs [26, 27, 28]. Domain ontologies, as DOREMUS[1] [29] or The Polifonia Ontology Network [30] – and their associated knowledge graphs (KGs) – cover a large area of concepts and entities (persons, instruments, musical works, interpretations of a work, etc.) related to the musical domain that could be retrieved in documents and included in graphs. Domain specific KGs can be supplemented by encyclopedic ones such as Wikidata [31] containing

[1]DOREMUS knowledge graph has been developed with data from three institutions including the Philharmonie de Paris.

both in-domain entities and entities associated with the musical domain regarding specific contexts (as nature sound studies or music in cinema).

Document retrieval approaches Retrieval approaches for RAG [32] have evolved along three main lines. Early approaches rely on keyword- and embedding-based search, where documents are independently encoded as dense vectors and retrieved through similarity matching with the query. While efficient and easy to generalize, these methods struggle to capture complex relationships that span multiple documents [32]. To address this, GraphRAG [33] methods construct an explicit graph structure over the knowledge extracted from documents, allowing retrievers to exploit global context and cross-document dependencies rather than treating isolated passages. Recent work on Graph Foundation Model (GFM) paradigm [34] extends GraphRAG by pre-training a query-dependent GNN retriever on large-scale KGs, drawing on advances in foundation models for KG reasoning. GFM also enables efficient single-step multi-hop retrieval that transfers to unseen datasets without domain-specific fine-tuning.

3. Overall Architecture

The overall architecture of the pipeline is presented in Figure 1.

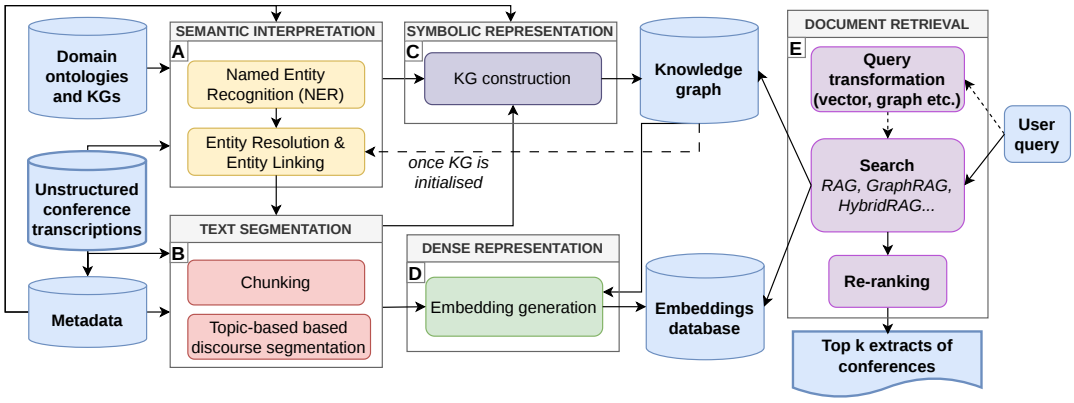


Figure 1: Architecture of the pipeline to perform conference transcription structuring and semantic document retrieval (conference excerpts), including a combination of symbolic and dense representation of the corpus.

Semantic interpretation (Step A) To enable KG construction and documents retrieval, the first step consists in annotating the transcriptions using a pipeline with NER, ER and EL. To perform the NER task lacking a labeled dataset, a possible approach is to use models allowing in-context learning like GLiNER [20] or LLMs [21]. To filter and enrich the extracted entities, the ER and EL step will resolve ambiguities and perform linking with external knowledge bases. This step will leverage off-the-shelf tools like Relik [22]. Framework like LELA [23] allows to combine NER and EL models. Linking will rely on domain-specific knowledge graphs (e.g., DOREMUS [29]), on encyclopedic graphs like Wikidata to provide music-related entities and entities from additional domains that might be important in a conference topic and on Philharmonie metadata (as speakers). Missing entities in KGs may be discovered. This first result can used to retrieve documents by querying the entities they contained.

Text segmentation (Step B) Transcription structuring faces a double challenge: provide fine-grained topic-consistent excerpts to the user at the end of the retrieval process and being sufficiently adapted to perform retrieval approaches efficiently. This aims to compare several chunking strategies [3] and topic-based discourse segmentation. Approaches based on co-reference resolution [15] seems relevant to preserve topic consistency to the excerpts.

KG construction (Step C) The resolved entities and structured documents will be added to a KG. Entities mentioned in several documents (or excerpts) and their real-world relations, as included in external KGs, should provide a dense network of connections that is useful for multi-hop reasoning. This can be supplemented with metadata (e.g. speakers, librarian keywords). Domain ontologies (DOREMUS or Polifonia Ontology Network) can be reused. Document structural relations should also be added to the graph to enable the use of discourse flow and logic in the retrieval process. Document structure could be completed by OpenIE extraction, which provide relations between the entities. The aim is to produce a KG representing the structure of the conference, its content including entities and domain keywords mentioned in it, the conference context (including speakers) and any other relevant information provided in the metadata (title, related events etc.).

Dense representation (Step D) Dense representations allow the retrieval of semantically similar content without exact keyword matches, helping discover missing connections within the graph. To evaluate their ability to handle noisy data, multiple embedding models should be tested during retrieval. Recent studies asses embedding robustness against noisy text recognised in images with OCR[2] engines, which shares similarities with ASR outputs. A viable initial development is the unsupervised fine-tuning of these models on conference transcriptions using a contrastive loss [35].

Document retrieval (Step E) Several retrieval architectures has to be evaluated to search for relevant conference excerpts, in particular to compare hybrid retrieval approach ability to combine the reliable, explicit relationships of knowledge graphs with the semantic flexibility of vector representations. Specifically, recent approaches will be compared including GFM against traditional RAG methodologies.

An important requirement of the Philarmonie de Paris – Cité de la musique pipeline is the ability to easily add new conferences and make them searchable by users.

4. Qualitative Analysis of the Transcripted Conference Corpus

This section presents the qualitative analysis of the transcribed conferences produced with an ASR framework. As stated above, it is crucial as the transcription errors of the corpora directly influence the performance of all subsequent stages in the processing pipeline.

4.1. Corpus of recordings and transcripted conferences

As stated, the MGR project examines a collection of conferences recorded by the Philharmonie de Paris – Cité de la Musique. Currently comprising 1,094 hours of audio and video content, these recordings include keynotes, round-tables, or interviews, optionally followed by Q&A sessions with the public, resulting in a variable number of speakers. Most conferences are in French, with occasional multi-language excerpts, and frequently feature musical performances that interrupt the speech flow. Conference topics cover both specific musical subjects (like studies of specific instruments or genres) and broader cross-disciplinary themes (as music in cinema or nature). Conference transcriptions have been realised with the ASR framework WhisperX [36]. Transcripts were produced following three phases. First, preprocessing steps include voice activity detection, segmentation, and batch creation from the audio stream. Second, the Whisper speech-to-text model [37] transcribes these segments in batches to parallelize the process, which are then aligned to the source audio down to phoneme units. Finally, the system detects separate speakers and tracks their conversational turns. The output is a structured file mapping sequential words to timestamps, sentence groups, and speaker identifiers.

Consequently, the transcripts contain domain-specific terms and entities (artists, musical works, movements, instruments, styles) alongside common named entities (people, places, organizations, historical periods), some recurring across multiple conferences. Professional librarians provide expert

[2]Optical Character Recognition

metadata for each recording, including title, subtitles, recording date, speaker names, descriptions, and keywords described in an associated 9-level taxonomy tree. Finally, most conferences have coarse-grained chaptering system dividing the content into coherent thematic subparts.

The analysis in this paper is based on a set of 39 conferences manually coarse-grained indexed by domain-experts of Philharmonie de Paris. It has been produced for a previous proof of concept of conferences indexation. This manual indexation process was discontinued due to its high time cost[3]. Annotations consist in two levels chaptering (chapters and tracks), identification of speakers turns and domain-related annotations classified in a dozen of categories. These annotations are associated with conference segments[4]. Keywords are distributed across four primary categories – *Themes* (used in 39 conferences), *Persons* (39), *Musical work* (36), and *Instruments* (15) –alongside four minor topic-specific classes: *Places* (3), *Factor* (3), *Performance technique* (1), and *Animals* (1). Four these 39 conferences, we aligned transcripts produced by WhisperX [36] with the metadata provided by the Philharmonie and with the librarians' test indexation using time stamps provided in these three sources.

4.2. Evaluation and discussion

As a gold standard of the indexed transcriptions (transcriptions without errors) is not available, a statistical analysis and empirical qualitative evaluation of the corpus was conducted, supported by the use of an application developed to explore the transcriptions (Figure 2). The 39-conferences corpus initially totals 564 530 word-level tokens. Following standard text normalization (lowercasing, punctuation stripping, stop-word/stop-locution exclusion, length filtering ≥ 2, and lemmatization), the corpus yields 206 497 tokens, 22 097 distinct words, and 10 983 hapax legomena [5]. Stylometric indicators reveals a tight distribution in Brunet's W[6] (11.36–12.59), signifying stable, high lexical diversity. 75% of conferences have a MTLD (Measure of Textual Lexical Diversity)[7] value above 79 (moderate-to-high lexical diversity) and 50% have MTLD value above 104 (high-to-very-high lexical diversity). However, it is not known what proportion of this lexical diversity is due to transcription errors that increase the size of the vocabulary.

Qualitative text evaluation also tends to show that named entities are particularly affected by transcription errors. The ASR model presents them as named entities (with the first letter uppercased, etc.), but the characters are incorrect, as can be seen in Figure 2 (for example *Zayas Iwani* instead of *Zahia Ziouani*). A relative phonetic similarity is observed between transcription and their exact writing. These errors degrade the accuracy of domain-specific terms, NER, ER and EL. Computing the Character Error Rate (CER)[8] on full transcription and entities using some corrected ASR outputs will provide an interesting insight into the impact of these errors on subsequent tasks.

Another type of mistake is the generation of false text. This error tend to happen during musical performances where no transcription is expected (see the red text in Figure 2). It may also be associated with desynchronisation between the text and audio, which can last for a few seconds.

The situations in which the model inaccurately generates the same sentence several times were also observed, sometimes introducing minor variations. These errors can potentially affect the quality of the returned results and have to be measured.

Distinct speakers recognition is the ability to distinguish real world persons talking in an audio or video. WhisperX provides this functionality in its speakers diarization output but its authors notes it might be unreliable[10] An initial assessment of this output was conducted, comparing WhisperX distinct speakers recognition results with the speakers indexed manually for each of the 39 conferences. WhisperX tends to recognise more speakers than declared in the manual indexation (average number of

[3]An example can be see here: https://pad.philharmoniedeparis.fr/player-conference.aspx#0998324
[4]Segments can span across multiple consecutive tracks.
[5]Detailed indicators and plots are available at https://gitlab.inria.fr/music-graph-rag/swodch2026-mgr-conferences-analysis.
[6]Defined in https://metricgate.com/docs/brunet-w-lexical-diversity/
[7]Defined in https://metricgate.com/docs/mtld-lexical-diversity/
[8]Defined in https://metricgate.com/docs/character-error-rate/
[9]https://pad.philharmoniedeparis.fr/pad/doc/CIMU/1158901/lilian-thuram-oser-la-musique?_lg=fr-FR
[10]See WhisperX documentation : https://github.com/m-bain/whisperX.



Figure 2: Example of WhisperX transcript. The generation of false text is observed during musical performances (highlighted in red) and transcription errors concerning named entities (highlighted in orange). The first false text sentence (*Sous-titres réalisés par la communauté l'Amara.org*) is a recurrent fabricated content in the corpus. The second false text sentence (*C'est la chine, la chine.*) is more complex to detect as it is plausible and not systematic. *Source : excerpt from "Lilian Thuram - Oser la musique", Philharmonie de Paris - Cité de la musique (2023)*[9].

speakers per conference is 4 according librarians and 7 according WhisperX outputs). This results should be taken carefully as librarians may have annotated only the speakers with significant contribution. Further analysis should be performed to assess precisely distinct speakers recognition and speaker changes as it could help to provide a structural segmentation of the conferences based on speakers turns (as in *Li et al. [15]* approach).

A more detailed analysis of the ASR results and their impact on the MGR architecture require manual correction of the transcriptions, including the alignment of text segments and entities for specialised metrics computation (overall and entity CER, Word Error Rate –WER, etc.). Depending on the results, systematic post-processing of the transcriptions may be performed with a focus on entities [38].

5. Conclusion and Future Work

This paper has provided an overview of the MGR project, which aims to develop semantic retrieval tools that enable users to query the corpus and access relevant conference transcriptions. It has presented a qualitative evaluation of ASR outputs, what reinforces the need for taking into account the specificities of the resulting transcriptions, such as mistranscribed or missing words, entities or unreliable speaker diarisation.Currently, the correction and annotation of few (4-5) conferences is conducted as an initial dataset for ASR output and NER and EL evaluation. We will then implement parts A and C of the pipeline, including NER, EL and KG construction. This initial setup will enable us to annotate more conferences more efficiently — annotating from scratch takes several minutes for every minute of audio — and so extend the evaluation dataset as well as evaluating a baseline architecture. Linked entities KG labels can be used to correct the associated text spans in transcriptions, which may have valuable effects on retrieval based on embeddings comparison, as ASR errors in text lead to generation of different embeddings for similar words. We are also considering evaluating the pipeline using state-of-the-art datasets in order to create a more generalised and robust architecture.

Acknowledgments

MGR is funded by the French Ministry of Culture and the Philharmonie de Paris – Cité de la Musique.

Declaration on Generative AI

Gemini Flash 3.1 and DeepL were used to check the writing. Gemini 3.1 Code was used as assistant to develop the visualisation interface. All scientific contributions and analyses are those of the authors.

References

[1] D. Liu, H. Ding, Q. Feng, X. Xie, Z. Xue, C. Wang, J. Li, J. Zhang, Y. Wang, Disco-RAG: Discourse-aware retrieval-augmented generation, in: Proc. of the 64th Annual Meeting of the ACL (Vol. 1: Long Papers), 2026.

[2] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, C. D. Manning, Raptor: Recursive abstractive processing for tree-organized retrieval, in: Proc. of ICLR, 2024.

[3] Y. Zhou, S. Wang, B. Koopman, G. Zuccon, Beyond chunk-then-embed: A comprehensive taxonomy and evaluation of document chunking strategies for information retrieval, arXiv:2602.16974, 2026.

[4] M. Śmigielski, M. Rajkowski, M. Zbrocki, M. Bernacki-Janson, K. Kunicki, J. Godziszewska, M. Piasecki, K. Wojtasik, Chunking methods on retrieval-augmented generation - effectiveness evaluation against computational cost and limitations, arXiv:2606.00881, 2026.

[5] T. Chen, H. Wang, S. Chen, W. Yu, K. Ma, X. Zhao, H. Zhang, D. Yu, Dense x retrieval: What retrieval granularity should we use?, in: Proc. of EMNLP24, ACL, 2024, pp. 15159–15177.

[6] A. V. Duarte, J. D. Marques, M. Graça, M. Freire, L. Li, A. L. Oliveira, LumberChunker: Long-form narrative document segmentation, in: Findings of EMNLP24, ACL, 2024, pp. 6473–6486.

[7] R. Qu, R. Tu, F. S. Bao, Is semantic chunking worth the computational cost?, in: Findings of NAACL25, ACL, 2025, pp. 2155–2177.

[8] M. Sediqin, S. E. Argamon, ESURF: Simple and effective EDU segmentation, in: arXiv, arXiv:2501.07723, 2025.

[9] L. Prevot, P. Muller, A few shades of supervision for discourse segmentation: Experiments on a french conversational corpus, Dialogue & Discourse 16 (2025) 35–73.

[10] S. Freisinger, P. Seeberger, T. Ranzenberger, T. Bocklet, K. Riedhammer, Towards multi-level transcript segmentation: LoRA fine-tuning for table-of-contents generation, in: Proc. Interspeech, 2025, pp. 276–280.

[11] F. Retkowski, M. Züfle, T. B. Nguyen, J. Niehues, A. Waibel, Beyond transcripts: A renewed perspective on audio chaptering, arXiv:2602.08979, 2026.

[12] M. A. Hearst, Text tiling: Segmenting text into multi-paragraph subtopic passages, Computational Linguistics 23 (1997) 33–64.

[13] Y. Lv, W. Tao, Q. Dai, Z. Chen, Q. Lu, N. Jiang, Dynamic topic segmentation in dialogues: Enhancing boundaries with topic-aware propagation, in: Companion Proc. of WWW '25, ACM, 2025, pp. 1172–1175.

[14] G. Glavaš, F. Nanni, S. P. Ponzetto, Unsupervised text segmentation using semantic relatedness graphs, in: Proc. of the Fifth Joint Conference on Lexical and Computational Semantics, ACL, 2016, pp. 125–130.

[15] X. Li, P. Li, A hierarchical framework for dialogue topic segmentation via global topic shifts and local entity coherence, in: Advanced Intelligent Computing Technology and Applications, Springer Nature, 2025, pp. 77–88.

[16] T. Wei, Y. Lu, H. Chang, Q. Zhou, X. Bao, A semantic approach for text clustering using WordNet and lexical chains, Expert Systems with Applications 42 (2015) 2264–2275.

[17] M. Mongiovì, A. Gangemi, GRAAL: Graph-based retrieval for collecting related passages across multiple documents, Information 15 (2024) 318.

[18] H. Kroll, J. Pirklbauer, J.-C. Kalo, M. Kunz, J. Ruthmann, W.-T. Balke, A discovery system for narrative query graphs: entity-interaction-aware document retrieval, Int. J. on Digit. Libr. (2023) 1–22.
[19] L. Luo, Z. Zhao, J. Liu, Z. Qiu, J. Dong, S. Panev, C. Gong, T.-T. Vu, G. Haffari, D. Phung, A. W.-C. Liew, S. Pan, G-reasoner: Foundation models for unified reasoning over graph-structured knowledge, 2026.
[20] I. Stepanov, M. Shtopko, D. Vodianytskyi, O. Lukashov, The million-label NER: Breaking scale barriers with GLiNER bi-encoder, arXiV, 2026.
[21] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, C. Guo, GPT-NER: Named entity recognition via large language models, in: Findings of NAACL25, ACL, 2025, pp. 4257–4275.
[22] R. Orlando, P.-L. Huguet Cabot, E. Barba, R. Navigli, ReLiK: Retrieve and LinK, fast and accurate entity linking and relation extraction on an academic budget, in: Findings ACL 2024, ACL, 2024, pp. 14114–14132.
[23] S. Haffoudhi, D. Nikola, F. Suchanek, N. Holzenberger, Lela: An end-to-end llm-based entity linking framework with zero-shot domain adaptation, in: Proc. of IJCAI-ECAI26, 2026.
[24] Q. Sun, Y. Luo, W. Zhang, S. Li, J. Li, K. Niu, X. Kong, W. Liu, Docs2kg: A human-LLM collaborative approach to unified knowledge graph construction from heterogeneous documents, in: Companion Proc. of WWW '25, ACM, 2025, pp. 801–804.
[25] H. Bian, LLM-empowered knowledge graph construction: A survey, arXiv:2510.20345, 2025.
[26] V. Eyharabide, V. Lully, F. Morel, MusicKG: Representations of sound and music in the middle ages as linked open data, in: Semantic Systems. The Power of AI and Knowledge Graphs, Springer, 2019, pp. 57–63.
[27] A. Ahola, E. Hyvönen, H. Rantala, A. Kauppala, Publishing and studying historical opera and music theatre performances on the semantic web: Case OperaSampo 1830–1960, in: Proc. of the SWODCH Int. Work., volume 3540, CEUR, 2023.
[28] A. Graciotti, N. Lazzari, V. Presutti, R. Tripodi, Musical heritage historical entity linking, Artif Intell Rev 58 (2025) 140.
[29] M. Achichi, P. Lisena, K. Todorov, R. Troncy, J. Delahousse, DOREMUS: A graph of linked musical works, in: ISWC 2018, volume 11137, Springer, 2018, pp. 3–19.
[30] J. de Berardinis, V. A. Carriero, N. Jain, N. Lazzari, A. Meroño-Peñuela, A. Poltronieri, V. Presutti, The polifonia ontology network: Building a semantic backbone for musical heritage, in: ISWC 2023, Springer, 2023, pp. 302–322.
[31] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85.
[32] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: NeurIPS 2020, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474.
[33] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, Graph retrieval-augmented generation: A survey, ACM Trans. Inf. Syst. 44 (2025).
[34] L. Luo, Z. Zhao, G. Haffari, D. Phung, C. Gong, S. Pan, GFM-RAG: Graph foundation model for retrieval augmented generation, in: NeurIPS 2025, 2025.
[35] A. Michail, J. Opitz, Y. Wang, R. Meister, R. Sennrich, S. Clematide, Cheap character noise for OCR-robust multilingual embeddings, in: Findings of ACL 2025, ACL, 2025, pp. 11705–11716.
[36] M. Bain, J. Huh, T. Han, A. Zisserman, Whisperx: Time-accurate speech transcription of long-form audio, in: Proc. Interspeech, 2023.
[37] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: Proc. of ICML23, 2023.
[38] S. Im, W. Lee, J. An, Y. Kim, J. Ok, G. Lee, DeRAGEC: Denoising named entity candidates with synthetic rationale for ASR error correction, in: Findings of ACL 2025, ACL, 2025, pp. 15181–15193.