

# Faire du neuf avec du balisé : Quand une édition TEI devient la mémoire d'un RAG

Clément Castellon<sup>2</sup> , Floriane Chiffolleau<sup>1,2</sup> , and Alina Miasnikova<sup>1,2</sup> 

<sup>1</sup> Observatoire des Textes, des Idées et des Corpus (ObTIC), Paris, France

<sup>2</sup> Sorbonne Université, Paris, France

## Abstract

This study investigates the integration of XML-TEI encoded digital editions of heritage collections into a Retrieval-Augmented Generation (RAG) system. Using the collections of the Valentin Haüy Association, we demonstrate that leveraging hierarchical divisions, titles, and paragraphs improves chunking and information retrieval. Results show that the RAG delivers precise, coherent answers faithfully reflecting the source documents while enabling traceability through TEI metadata. This approach highlights the potential of structured digital editions as a dynamic memory for RAG systems, offering enhanced exploration and accessibility of complex archival corpora within the digital humanities.

**Mots-clés:** édition numérique, génération à enrichissement contextuel (RAG), encodage XML-TEI, collection patrimoniale

**Keywords:** digital edition, retrieval-augmented generation (RAG), TEI XML encoding, heritage collection

## Introduction

Ces dernières années, les systèmes RAG ont connu un intérêt croissant dans des contextes variés, et des travaux récents se sont attachés à explorer les usages concrets du RAG et la polyvalence de la méthode, dans différents contextes applicatifs. Ainsi, un panorama des usages contemporains du RAG, présentant différents domaines d'application, tels que la recherche scientifique, le patrimoine ou la documentation technique, est présenté dans [1].

En parallèle, les humanités numériques observent le développement croissant de la production d'éditions numériques de collections patrimoniales, encodées en TEI et déployées en ligne. Ces éditions offrent généralement, pour explorer leurs documents, l'option de recherche plein texte ou le filtrage par le biais de métadonnées, comme cela peut s'observer par exemple dans les éditions du registre TEI Publisher<sup>1</sup>.

Au regard de ces différentes perspectives, nous nous sommes posées la question de savoir s'il serait possible de mettre à profit l'encodage TEI structurelle d'une édition numérique de collections patrimoniales, afin de générer la mémoire d'un RAG, favorisant ainsi l'exploitation d'éléments déjà mis à disposition.

Pour répondre à cette question, nous nous sommes appuyés sur le projet d'édition numérique des collections patrimoniales de l'association Valentin Haüy (AVH), consacrée dans un premier temps à un fonds dédié aux aveugles de guerre.

---

Clément Castellon, Floriane Chiffolleau, and Alina Miasnikova. "Faire du neuf avec du balisé : Quand une édition TEI devient la mémoire d'un RAG." *Actes de la Conférence Humanistica*, éd. par Serena Crespi, Simon Gabay, Martin Grandjean, Ariane Pinche, Marie Puren et Léa Saint-Raymond. Vol. 4. Anthology of Computers et the Humanities. 2026, 134–149. <https://doi.org/10.63744/UR8DCmRglj12>.

© 2026 par les auteurs. Sous licence Creative Commons Attribution 4.0 International (CC BY 4.0).

1. <https://www.e-editions.org/map>.

## 1 Mettre à disposition les collections de l'AVH

### 1.1 L'association Valentin Haüy et ses collections

Fondée en 1889 et reconnue d'utilité publique depuis 1891, l'association Valentin Haüy (AVH), créée par et pour les personnes aveugles, œuvre pour l'autonomie des personnes aveugles ou malvoyantes, que ce soit au sein même de l'association, ou grâce à sa bibliothèque et sa médiathèque [13]. La médiathèque est notamment reconnue comme la première au monde à avoir rassemblé une documentation aussi exhaustive sur la cécité. Elle possède également un fonds documentaire riche qui témoigne de la riche histoire de la communauté des personnes aveugles.

Pour les personnes aveugles, la lecture de textes se fait en braille, comme dans un système d'écriture tactile, ou grâce à des lecteurs d'écran qui restituent le contenu sous forme de parole ou de texte braille. À l'inverse, il existe aussi des textes « en noir », c'est-à-dire des écrits courants, donc inaccessibles aux personnes aveugles ou malvoyantes sans traduction en braille ni version lisible par machine [12].

### 1.2 Enjeux d'accès et de valorisation des collections patrimoniales

La bibliothèque patrimoniale de l'AVH contient une grande quantité de ces textes « en noir », soit plus d'un million de pages de documents. Ces textes sont très divers, allant de revues de diffusion d'informations et de connaissances destinées aux typhlophiles<sup>2</sup> à un fonds dédié aux aveugles de guerre, et bien que des chercheurs du monde entier consultent ces documents, leur inventaire est insuffisant et ils sont par ailleurs très fragiles.

La combinaison de ces différents éléments a motivé le lancement d'un projet de création d'éditions numériques scientifiques de ces documents, pour aller au-delà d'une simple transcription obtenue par reconnaissance automatique de texte, et proposer une version structurée et enrichie des documents de la bibliothèque patrimoniale. Afin de garantir un accès encore plus aisé aux connaissances contenues dans ces ressources, nous avons envisagé de mettre en place un système de RAG.

## 2 Exploiter son encodage XML-TEI pour mettre en place un RAG

### 2.1 Le RAG, un outil exponentiellement utilisé pour interroger des corpus d'humanités

La génération à enrichissement contextuel (*Retrieval-Augmented Generation*, RAG) désigne une famille d'architectures combinant des modèles de langage de grande taille (*Large Language Models*, LLM) avec des mécanismes de recherche d'information. Introduit notamment par [7], le principe du RAG s'appuie sur l'idée de ne pas faire reposer la génération de réponses uniquement sur les connaissances internes du modèle, mais de les enrichir dynamiquement à partir d'un corpus externe. Dans son schéma canonique, un système RAG comprend ces étapes principales : l'encodage de la requête utilisateur, la récupération de passages pertinents au sein d'un corpus indexé, la génération d'une réponse. Cette approche permet non seulement d'améliorer la factualité des réponses, mais aussi de rendre le raisonnement du modèle plus traçable [7].

Aujourd'hui, l'application du RAG à des corpus volumineux ou hétérogènes soulève des enjeux particuliers liés à la longueur et à leur diversité. Dans le contexte des débats parlementaires de la Troisième République française, la récupération de passages pertinents dans un corpus long et bruité est essentielle pour maintenir la cohérence contextuelle des réponses générées, comme observé dans [11]. De manière similaire, les politiques de numérisation massives des collections des institutions GLAM rendent nécessaire de repenser les modalités d'exploration et de valorisation des corpus, tant pour les chercheurs que pour les bibliothécaires et conservateurs [5]. Dans ce contexte, le RAG constitue un outil capable d'interroger ces corpus volumineux et d'extraire les

---

2. Personne qui aime et soutient les personnes aveugles et malvoyantes.

informations pertinentes pour la génération de réponses contextualisées. Enfin, un autre usage du RAG appliqué aux TEI Guidelines est présenté dans [6], afin de faciliter la création de documents conformes aux normes TEI grâce à un accès rapide et ciblé à la documentation normative.

Ces travaux convergent vers l'idée que la structure interne, qu'elle soit historique, patrimoniale ou normative, influence fortement la manière dont le contexte est constitué pour le RAG. Dans cette perspective, le présent article examine si des documents encodés en XML-TEI peuvent améliorer la pertinence et la précision du RAG.

## 2.2 Un encodage XML-TEI reprenant la structure et les spécificités des collections AVH

Dans le cadre de la mise en place de l'édition scientifique numérique des collections AVH, nous avons choisi de structurer et enrichir les documents en suivant le standard pour la gestion sous forme numérique des documents d'archives qu'est la *Text Encoding Initiative* (TEI) [3; 14]. Bien que ces guidelines contiennent des options pour une grande diversité de documents (correspondance, poésie, etc.), le premier échantillon des collections sur lequel nous travaillons se compose exclusivement de revues présentant une structure de titres et paragraphes, comme le montre la figure 1.

En conséquence, exceptés pour les cas où l'on trouve des tableaux (<table>, <row>, <cell>) ou des listes (<list>, <item>), l'encodage ne se construit qu'autour de divisions de parties (<div>), de titres (<head>) et de paragraphes (<p>), agrémenté de balises pour les sauts de page (<pb>) et la foliation des documents (<note>), comme le démontre la figure 2. La division du texte est clarifiée à l'aide de l'attribut *@type* dont la valeur sert autant à reconnaître l'unité et la hiérarchie du bout de texte encodé, à savoir des sections organisées comme ceci : section, subsection, subsubsection – ce dernier étant l'unité la plus basse utilisée dans l'encodage. Ce sont ces divisions que l'on va chercher à mettre à profit pour développer notre RAG.

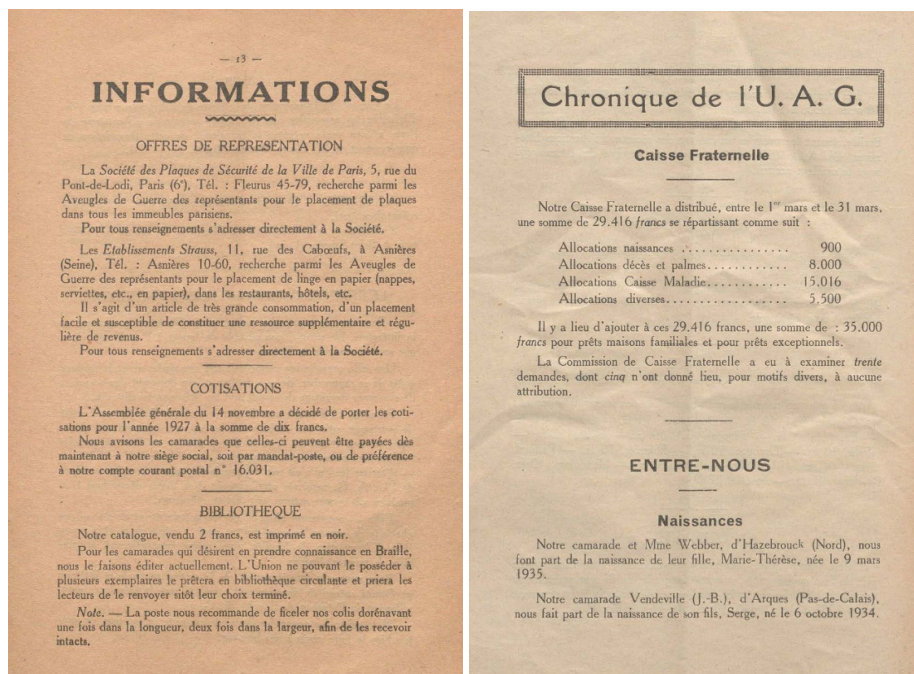


FIGURE 1 – Bulletin mensuel de l'Union des Aveugles de guerre et du Journal des blessés aux yeux, édition de janvier 1927 (page 15, gauche) et avril 1935 (page 20, droite).

```

<pb n="15"/>
<note type="foliation">— 13 —</note>
</div>
<div type="section">
  <head type="section" rend="align(center)" xml:id="h14">INFORMATIONS</head>
  <div type="subsection">
    <head type="subsection" rend="align(center)" xml:id="h15">OFFRES DE
      REPRESENTATION</head>
    <p>La Société des Plaques de Sécurité de la Ville de Paris, 5, rue du Pont-de-Lodi, [3 lines]
    <p>Pour tous renseignements s'adresser directement à la Société. Les Etablissements [3 lines]
    <p>Il s'agit d'un article de très grande consommation, d'un placement facile et [1 line]
    <p>Pour tous renseignements s'adresser directement à la Société.</p>
  </div>
  <div type="subsection">
    <head type="subsection" rend="align(center)" xml:id="h16">COTISATIONS</head>
    <p>L'Assemblée générale du 14 novembre a décidé de porter les cotisations pour l'année [1 line]
    <p>Nous avisons les camarades que celles-ci peuvent être payées dès maintenant à notre [2 lines]
  </div>
  <div type="subsection">
    <head type="subsection" rend="align(center)" xml:id="h17">BIBLIOTHEQUE</head>
    <p>Notre catalogue, vendu 2 francs, est imprimé en noir. Pour les camarades qui [3 lines]
    <p>Note. - La poste nous recommande de ficeler nos colis dorénavant une fois dans la [1 line]
  </div>
</div>
<pb n="20"/>
<note type="foliation">— 18 —</note>
</div>
<div type="section">
  <head type="section" rend="align(center)" xml:id="h21">Chronique de I'U. A. G.</head>
  <div type="subsection">
    <head type="subsection" rend="align(center)" xml:id="h22">Caisse Fraternelle</head>
    <p>Notre Caisse Fraternelle a distribué, entre le 1<hi rend="superscript">er</hi> mars [1 line]
    <p>Allocations naissances 900 Allocations décès et palmes..... 8.000 [4 lines]
    <p>La Commission de Caisse Fraternelle a eu à examiner trente demandes, dont cinq [1 line]
  </div>
  <div type="subsection">
    <head type="subsection" rend="align(center)" xml:id="h23">ENTRE-NOUS</head>
    <div type="subsubsection">
      <head type="subsubsection" rend="align(center)" xml:id="h24">Naissances</head>
      <p>Notre camarade et Mme Webber, d'Hazebrouck (Nord), nous font part de la naissance [1 line]
      <p>Notre camarade Vendeville (J. B.), d'Arques (Pas-de-Calais), nous fait part de la [1 line]
    </div>
  </div>
</div>

```

FIGURE 2 – Encodage XML-TEI des pages de 1927 (haut) et 1935 (bas) présentées dans la figure 1.

### 2.3 Adapter le *chunking* pour le RAG à l'encodage XML-TEI

Un point crucial dans l'implémentation d'un système RAG est la segmentation documentaire (*chunking*), c'est-à-dire la façon dont les documents sont découpés en unités qui seront indexées, recherchées et fournies au modèle génératif. Les stratégies classiques consistent à découper les textes en segments fixes, mais cette approche fragmente souvent le contexte et affaiblit la cohérence sémantique des passages récupérés. Des travaux récents examinent des alternatives pour améliorer ce processus. Par exemple, la comparaison des techniques de *late chunking* et de *contextual retrieval*, montre que la récupération enrichie par le contexte, bien que plus coûteuse en calcul, préserve mieux l'intégrité sémantique [10].

D'autres travaux s'intéressent plus spécifiquement à la relation entre la structure des documents et les stratégies de segmentation. Les auteurs de [9] proposent une approche de *chunking* hiérarchique, fondée sur l'idée que les documents possèdent une organisation multiniveaux qui peut être exploitée pour définir des unités de récupération de granularité variable. Leur méthode vise à respecter la structure interne des documents afin d'améliorer la récupération dans les systèmes RAG, indépendamment du domaine ou du format de représentation. Dans le cas de notre expérience, nous avons décidé de structurer le *chunking* en utilisant les balises structurales présentes dans notre corpus, afin d'extraire les métadonnées pour compléter le mécanisme de *retrieval*.

### 3 Quand le texte encodé devient texte généré : méthodes et résultats

#### 3.1 Implémentation et paramétrage du système RAG

Pour implémenter le système RAG, nous avons développé une interface interactive à l'aide de Streamlit<sup>3</sup>, permettant une manipulation flexible du corpus et des paramètres du système. Le pipeline repose sur LlamaIndex [8] pour l'orchestration de l'indexation et de la récupération, et sur le modèle d'*embeddings intfloat/multilingual-e5-large* [16], choisi pour ses performances sur des textes français et multilingues de registres variés. Plusieurs LLM *open source* ont été intégrés pour la génération, notamment *Mistral-7B-Instruct*<sup>4</sup> et *Zephyr-7B* [15], afin de ne pas dépendre d'API propriétaires. L'interface expose les paramètres principaux :

- le top-k<sup>5</sup>;
- la température<sup>6</sup>;
- le nombre maximal de tokens générés;
- le mode "*retrieval only*"<sup>7</sup>;
- des filtres post-récupération<sup>8</sup>.

Le point central du pipeline est la stratégie de *chunking*, entièrement pilotée par la structure TEI du corpus. Un module de détection automatique analyse par échantillonnage aléatoire un sous-ensemble de fichiers XML pour déterminer, sans intervention manuelle, le type documentaire majoritaire du corpus — correspondance, édition littéraire, édition parallèle, etc. — ainsi que la granularité de découpage appropriée. La décision repose sur une taxonomie des balises `<div>` selon leur rôle : unité de *chunk* directe (par exemple `<div type="chapter">` ou `<div type="letter">`), conteneur récursif à traverser (par exemple `<div type="section">`), ou élément à ignorer (pages de titre, tables des matières). La densité tokenique moyenne par unité est ensuite mesurée : si elle dépasse un seuil de 400 tokens, le module active une décomposition au niveau des `<p>`; si un paragraphe individuel dépasse encore ce seuil, il est soumis à un découpage par phrase. Cette cascade à trois niveaux garantit que chaque *chunk* reste dans une plage cohérente pour l'encodage sémantique, quelle que soit la structure interne du corpus. Enfin, les métadonnées extraites du `<teiHeader>` — auteur, date, titre de section, type de division — sont propagées vers chaque *chunk*, ce qui permet à la fois d'afficher des sources traçables dans l'interface et d'alimenter les filtres post-récupération.

#### 3.2 Des résultats prometteurs et particulièrement détaillés

Afin d'évaluer le fonctionnement du RAG<sup>9</sup>, expérimenté sur le corpus AVH contenant 284 fichiers et représentant 27 609 *chunks*, plusieurs questions de nature variée ont été soumises au système. L'ensemble de ces questions et réponses est présenté en annexe B. Les résultats obtenus montrent que le RAG fournit, de manière générale, des réponses claires et pertinentes. Il restitue fidèlement les informations présentes dans les documents sources (en mentionnant le document cité) ou, lorsque la question est formulée de manière plus large, il est capable d'en proposer une extrapolation raisonnée. Par ailleurs, le RAG démontre une bonne constance dans ses réponses. En effet, comme l'illustre le tableau B.1, lorsqu'une même question est posée à plusieurs reprises, les réponses produites sont similaires, voire identiques, tout comme les documents à partir desquels les informations sont récupérées. En outre, le système se distingue par sa capacité à ne pas produire

3. <https://streamlit.io>.

4. <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1>.

5. Nombre de passages récupérés soumis au modèle.

6. Paramètre du modèle de langage qui influence la diversité des réponses générées.

7. Interrogation directe de l'index sans passer par la génération.

8. Affinage des résultats par date, type de *chunk* ou mots-clés présents dans les métadonnées.

9. <https://github.com/obtic-sorbonne/rag-humanistica>.

d'informations infondées : lorsqu'une réponse ne peut être établie à partir des documents disponibles, le RAG n'invente pas de contenu et mentionne l'absence de résultat trouvé (B.3). L'intérêt principal du RAG proposé, ainsi que du *chunking* adopté, est notamment mis en évidence par les résultats détaillés figurant en annexe C. Ces résultats montrent en particulier que le *chunking* hiérarchique s'avère pertinent, dans la mesure où il permet d'identifier les réponses les plus adaptées au sein de différents types de sections (Annexe C Figure 3). De plus, dans les retours détaillés par source (Annexe C Figure 4), le RAG a été conçu de façon à fournir des informations contextuelles supplémentaires, notamment le type et le titre de la section dont la réponse a été extraite. Ce choix facilite ultérieurement la recherche de l'information dans le fichier XML correspondant.

## **4 Une solution fonctionnelle à renforcer, généraliser, déployer...**

### **4.1 Renforcer les capacités du RAG**

La solution proposée repose actuellement sur un fonctionnement essentiellement extractif, dans lequel les réponses générées correspondent à des informations explicitement présentes dans le corpus. Une première perspective d'amélioration consisterait à doter le RAG de capacités de raisonnement, lui permettant de traiter des questions nécessitant une inférence ou une agrégation d'informations distribuées dans plusieurs documents. Il deviendrait ainsi possible de formuler des requêtes transversales, par exemple de nature quantitative ou comparative, impliquant la combinaison de données issues de différentes entités ou sections du corpus.

Une autre piste d'amélioration concerne l'évaluation de modèles de langage de plus grande taille. Des modèles disposant de gestion du contexte plus avancée pourraient améliorer la qualité des réponses, en particulier pour des requêtes complexes mobilisant plusieurs unités de contexte. Une telle exploration permettrait de mieux mesurer l'impact du choix du modèle sur l'exploration d'éditions numériques structurées et d'identifier des compromis adaptés aux contraintes des projets en humanités numériques.

### **4.2 Généraliser l'approche à d'autres types d'encodage TEI**

L'expérience que nous présentons ici n'est qu'un simple *proof-of-concept*, l'idée ayant été de voir, avec un encodage basique et simplement structuré, si une édition numérique en TEI pouvait devenir mémoire de RAG. Si cela a effectivement été concluant, il est cependant essentiel d'approfondir notre recherche, pour garantir sa généralisation à d'autres types d'encodages. La TEI se caractérise par une quantité importante de balises et d'attributs, adaptés surtout aux types de documents travaillés, et il serait pertinent de vérifier que le *chunking* fonctionne aussi bien pour d'autres corpus. Un premier élément de réponse peut être apporté avec des tests que nous avons effectués sur deux autres corpus TEI, une correspondance de la Première Guerre mondiale<sup>10</sup> [4], et des documents relatant l'Holocauste<sup>11</sup> [2]. Comme le montrent les annexes D et E, le RAG semble répondre aussi bien qu'avec le corpus initial, ce qui est encourageant pour son adaptation à d'autres corpus TEI.

### **4.3 Déployer le RAG au sein d'une instance de publication de l'édition numérique**

Enfin, au regard du projet même sur lequel s'est appuyé le RAG pour mettre en place sa mémoire, la finalité de cette édition numérique est de la déployer sur un site web pour donner accès à tous aux collections patrimoniales. L'intérêt majeur de cette édition étant d'être accessible pour tous et notamment les principaux concernés par ces documents, le RAG pourrait être d'une grande utilité pour pouvoir faire des recherches dans les collections sans connaître ou voir son contenu.

---

10. 1 492 fichiers, 21 111 chunks.

11. 279 fichiers, 11 588 chunks.

Il sera donc nécessaire de réfléchir aux systèmes de gestions de contenus (CMS) qui existent pour découvrir si et comment cela pourrait s'intégrer, afin de proprement combiner les deux.

## Conclusion

Dans ce papier, nous avons montré qu'il était possible d'utiliser l'encodage XML-TEI d'éditions scientifiques numériques, comme mémoire pour un système RAG, utilisant donc la TEI comme outil de médiation entre une édition numérique et un RAG, augmentant ainsi ses capacités d'exploitation. Nos tests supplémentaires nous ont permis d'entrevoir que la technique pouvait effectivement s'adapter à d'autres corpus TEI, sans avoir à initier de changements trop conséquents sur la structure même du RAG. Il sera nécessaire néanmoins de l'appliquer à une plus grande quantité et diversité de fonds, pour tester de manière plus approfondie ses compétences et ses limites.

## Références

- [1] ARSLAN, Muhammad, GHANEM, Hussam, MUNAWAR, Saba et CRUZ, Christophe. « A Survey on RAG with LLMs ». In : *Procedia Computer Science* 246 (2024), p. 3781-3790. DOI : 10.1016/j.procs.2024.09.178.
- [2] BÉNIÈRE, Sarah, CHIFFOLEAU, Floriane et SCHEITHAUER, Hugo. « Streamlining the Creation of Holocaust-related Digital Editions with Automatic Tools ». In : *EHRI Academic Conference - Researching the Holocaust in the Digital Age*. EHRI-3. Varsovie, Pologne, 2024. URL : <https://inria.hal.science/hal-04594190>.
- [3] BURNARD, Lou. *Qu'est-ce que la Text Encoding Initiative ?* Trad. par Marjorie BURGHART. Encyclopédie numérique. Marseille : OpenEdition Press, 2015. DOI : 10.4000/books.oep.1237.
- [4] CHIFFOLEAU, Floriane et BAILLOT, Anne. « Le projet DAHN : une pipeline pour l'édition numérique de documents d'archives ». Document de travail. 2022. URL : <https://hal.science/hal-03628094>.
- [5] FACI, Adam et SILVESTRE DE SACY, Antoine. « RAG pour l'exploration de corpus en GLAM ». In : *La linguistique dans une ère nouvelle : discours, méthodes et technologies dans le paysage contemporain*. Nanterre, France, 2024. URL : <https://hal.science/hal-04829841>.
- [6] KHEMAKHEM, Mohamed, REKIK, Houcem Eddine et BOUAZIZ, Oussama. « Enhancing Technical Knowledge Acquisition with RAG Systems : the TEI use case ». In : *Texts, Languages and Communities - TEI 2024*. Buenos Aires, Argentine, 2024. URL : <https://hal.science/hal-04752713>.
- [7] LEWIS, Patrick et al. « Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks ». 2021. DOI : 10.48550/arXiv.2005.11401.
- [8] LIU, Jerry. « LlamaIndex ». Nov. 2022. URL : [https://github.com/jerryjliu/llama\\_index](https://github.com/jerryjliu/llama_index).
- [9] LU, Wensheng, CHEN, Keyu, QIAO, Ruizhi et SUN, Xing. « HiChunk : Evaluating and Enhancing Retrieval-Augmented Generation with Hierarchical Chunking ». 2025. DOI : 10.48550/arXiv.2509.11552.
- [10] MEROLA, Carlo et SINGH, Jaspinder. « Reconstructing Context : Evaluating Advanced Chunking Strategies for Retrieval-Augmented Generation ». 2025. DOI : 10.48550/arXiv.2504.19754.

- [11] PEREZ, Julien, PELLET, Aurélien et PUREN, Marie. « Évaluation automatique du retour à la source dans un contexte historique long et bruité. Application aux débats parlementaires de la Troisième République française ». In : *Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, sous la dir. de Frédéric BECHET, Adrian-Gabriel CHIFU, Karen PINEL-SAUVAGNAT, Benoit FAVRE, Eliot MAES et Diana NURBAKOVA. Marseille, France : ATALA & ARIA, 2025, p. 138-150. URL : <https://aclanthology.org/2025.jeptalnrecital-evalllm.12/>.
- [12] ROUSSEL, Céline et CHOTTIN, Marion. « The Forgotten Room of the Association Valentin Haüy in Paris, or How to Build an Inclusive Digitisation Project on the History of Blindness ». 2023. URL : <https://www.public-disabilityhistory.org/2023/10/the-forgotten-room-of-association.html>.
- [13] ROY, Noëlle. « Musée et Bibliothèque patrimoniale Valentin Haüy : cécité et culture de l'innovation ». In : *Canadian Journal of Disability Studies* 8, no. 6 (2019), p. 44-64. DOI : 10.15353/cjds.v8i6.579.
- [14] TEI CONSORTIUM. « TEI P5 : Guidelines for Electronic Text Encoding and Interchange ». Version 4.11.0. TEI Consortium. Fév. 2026. URL : <http://www.tei-c.org/Guidelines/P5/>.
- [15] TUNSTALL, Lewis et al. « Zephyr : Direct Distillation of LM Alignment ». 2023. DOI : 10.48550/arXiv.2310.16944.
- [16] WANG, Liang, YANG, Nan, HUANG, Xiaolong, YANG, Linjun, MAJUMDER, Rangan et WEI, Furu. « Multilingual E5 Text Embeddings : A Technical Report ». 2024. DOI : 10.48550/arXiv.2402.05672.

## A Le prompt du RAG

Tu es un assistant spécialisé dans les éditions numériques en humanités.  
Réponds uniquement en te basant sur les documents fournis.  
Cite toujours tes sources avec la date et le titre du document.

## B Les requêtes faites au RAG (Corpus AVH)

### B.1 Cohérence des réponses lors de répétitions de questions

**Question :** Peux-tu me citer les naissances en 1921 ?

**Réponse (Version 1) :**

1. Notre camarade et Mme Sibiril (Pierre), nous font part de la naissance de leur fille Suzanne, née le 2 octobre 1921.
2. Notre camarade et Mme Colleuille (Maurice), nous font part de la naissance de leur fils Robert-Maurice, né le 8 septembre.
3. Notre camarade et Mme Felgeirolles (Henri), nous font part de la naissance de leur fille Marie-Louise, née le 30 août.
4. Notre camarade et Mme Fossati (André), nous font part de la naissance de leur fils Maurice, né le 27 septembre.

5. Notre camarade et Mme Rascle (Edmond), nous font part de la naissance de leur fille Raymonde, née le 4 octobre.
6. Notre camarade et Mme Malgoire, nous font part de la naissance de leur enfant, né le 24 septembre.
7. Notre camarade et Mme Rouxel (E.), nous font part de la naissance de leur fille Blanche, née le 7 octobre.
8. Notre camarade et Mme Beve (E.), nous font part de la naissance de leur fille Marthe, née le 22 octobre.
9. Notre camarade et Mme Lamaury (Pierre), nous font part de la naissance de leur fils Fernand.
10. Notre camarade et Mme Hassen Ben Salem Kraiem, nous font pa. de la naissance de leur fille, née le 20 octobre.
11. Notre camarade et Mme Boucher, nous font part de la naissance de leur fils Raymond, né le 27 octobre 1921.
12. Notre camarade et Mme Gambert (François), nous font part de la naissance de leur deuxième garçon, né le 22 septembre.
13. Notre camarade et Mme Moteau (Jean), nous font part de la naissance de leur deuxième garçon, né le 17 octobre.

**Réponse (Version 2) :**

1. Notre camarade et Mme Sibiril (Pierre) ont eu une fille Suzanne le 2 octobre 1921.
2. Notre camarade et Mme Colleuille (Maurice) ont eu un fils Robert-Maurice le 8 septembre.
3. Notre camarade et Mme Felgeirolles (Henri) ont eu une fille Marie-Louise le 30 août.
4. Notre camarade et Mme Fossati (André) ont eu un fils Maurice le 27 septembre.
5. Notre camarade et Mme Rascle (Edmond) ont eu une fille Raymonde le 4 octobre.
6. Notre camarade et Mme Malgoire ont eu un enfant le 24 septembre.
7. Notre camarade et Mme Moteau (Jean) ont eu un deuxième garçon le 17 octobre.
8. Notre camarade et Mme Beve (E.) ont eu une fille Marthe le 22 octobre.
9. Notre camarade et Mme Lamaury (Pierre) ont eu un fils Fernand.
10. Notre camarade et Mme Hassen Ben Salem Kraiem ont eu une fille le 20 octobre.
11. Notre camarade et Mme Boucher ont eu un fils Raymond le 27 octobre 1921.
12. Notre camarade et Mme Gambert (François) ont eu leur deuxième garçon le 22 septembre.
13. Notre camarade et Mme Rouxel (E.) ont eu une fille Blanche le 7 octobre.
14. Notre camarade et Mme Lassalle (Auguste) ont eu une fille Augusta le 17 septembre.
15. Notre camarade et Mme Denos (Jean) ont eu un fils Pierre le 14 octobre.
16. Notre camarade et Mme Juif (Georges) ont eu une fille.
17. Notre camarade et Mme Masson (François) ont eu un fils Louis le 8

---

**Question :** Qu'est-ce que la guerre ?

**Réponse (Version 1) :**

1. La guerre est une action extérieure occasionnée par des événements de guerre et du fait de l'ennemi qui résulte en lésions graves et la destruction des vies humaines précieuses et des foyers (Bulletin mensuel de l'Union des aveugles de guerre et Journal des soldats blessés aux yeux - 17ème année 1935, n°2).
2. Selon l'Union des aveugles de guerre, la guerre offensive est immorale et criminelle (Bulletin mensuel de l'Union des aveugles de guerre et Journal des soldats blessés aux yeux - 13ème année 1931, n°7).

**Réponse (Version 2) :**

1. La guerre est l'action extérieure qui occasionne des lésions physiques à un individu pendant des événements de guerre et du fait de l'ennemi. (Source : Bulletin mensuel de l'Union des aveugles de guerre et Journal des soldats blessés aux yeux - 17ème année 1935, n°2)
2. Quelle est la position de l'organisme sur la guerre offensive ? La guerre offensive est immorale et criminelle. (Source : Bulletin mensuel de l'Union des aveugles de guerre et Journal des soldats blessés aux yeux - 13ème année 1931, n°7)

**B.2 Qualité des réponses à des questions diverses**

**Question :** Qu'est-ce que le vélo-car ?

**Réponse :** The Vélocar is a bicycle that is car-like in appearance, with a structure that resembles a small automobile. It is propelled by pedaling, and it has three gears that allow it to reach speeds of up to 50 kilometers per hour and climb hills easily. The occupants pedal at their own pace, each with their own independent pedal and free wheel. (1927-12-01, Bulletin mensuel de l'Union des aveugles de guerre et Journal des soldats blessés aux yeux - 9ème année 1927 n°10)

---

**Question :** Qu'est-ce qu'un aveugle de guerre ?

**Réponse :** According to the "Bulletin mensuel de l'Union des aveugles de guerre et Journal des soldats blessés aux yeux - 5ème année 1923, n°3" on 1923-03-01, an aveugle de guerre is not an invalid, but a mutilé; the difference is great. He is a man who, by virtue of a worldwide catastrophe, has lost the use of one of his senses, but who has not lost his intelligence, activity, skill, and the memory of everything he has seen. He can compensate for the sudden loss of his sight through new tactile and auditory"

---

**Question :** La vie ...

**Réponse :** The life of the blind veterans is discussed in the documents, particularly their pensions, rehabilitation, employment, and social aide. For example, in the document "Bulletin mensuel de l'Union des aveugles de guerre et Journal des soldats blessés aux yeux - 5ème année 1923, n°1", there is a questionnaire about Franceville, a home for the blind veterans ...

### B.3 Neutralité pour des questions non pertinentes

**Question :** Qui a volé l'orange du marchand ?

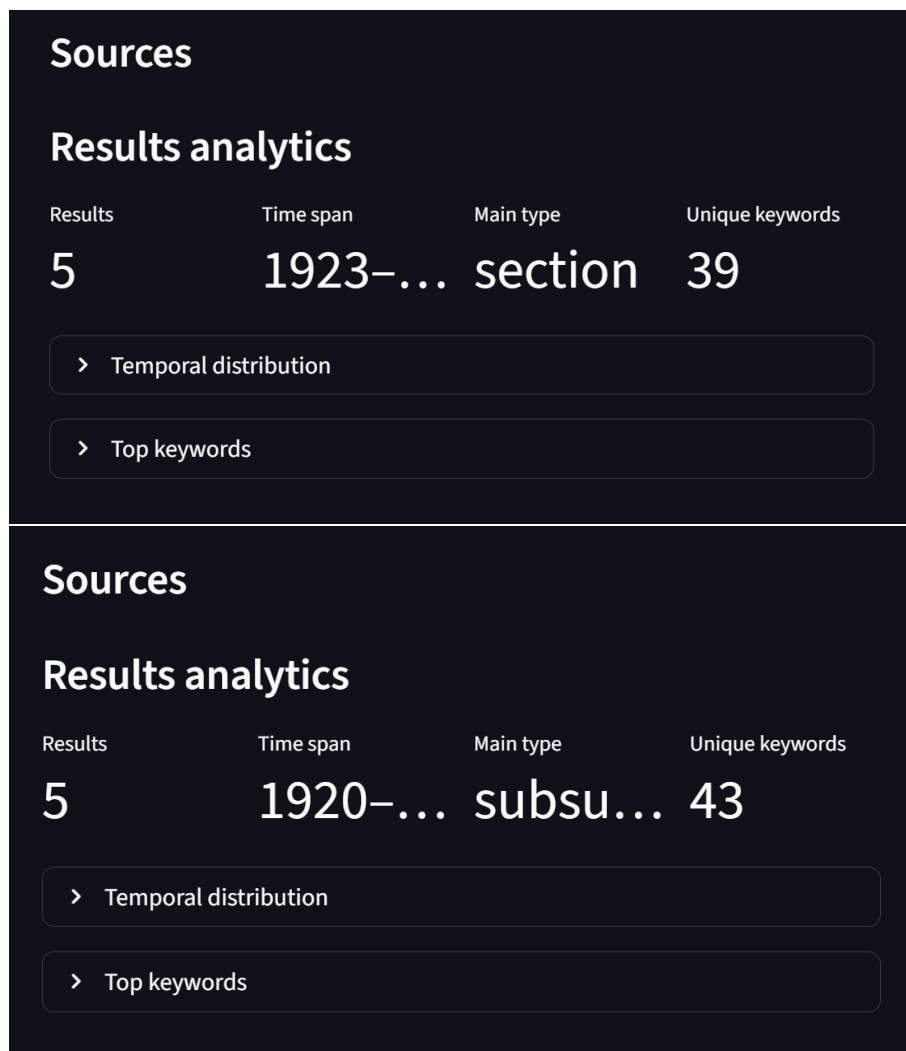
**Réponse :** The document does not provide information about someone stealing an orange from a merchant.

---

**Question :** Qu'est-ce que l'iphone ?

**Réponse :** L'iphone n'est pas mentionné dans les documents fournis, il est impossible de répondre à la question avec précision en se basant sur les documents.

### C Détails des retours de requêtes (Corpus AVH)



**FIGURE 3** – Retour général après requête

Temporal	Attribution	Structure																								
Date: 01/03/1923	Author: L'Union des Aveugles de Guerre	Type: section																								
Year: 1923		Section: NOTRE MAGASIN																								
Document: Bulletin mensuel de l'Union des aveugles de guerre et Journal des soldats blessés aux yeux - 5ème an...																										
File: bulletin_uag_jsby_1923_03.xml																										
<a href="#">Union des Aveugles de Guerre</a> <a href="#">Aveugles</a> <a href="#">Guerre</a> <a href="#">Mutilés</a> <a href="#">Pensions</a>																										
<a href="#">Article 64</a> <a href="#">Exposition de Gand</a> <a href="#">Radiotéléphonie</a>																										
<span>doc_type: archival</span> <span>strategy: adaptive</span>																										
<b>Content</b>																										
L'aveugle de guerre n'est pas un infirme, c'est un mutilé; la différence																										
Source 1 — score: 0.791 — Décès																										
<table border="1"> <thead> <tr> <th>Temporal</th> <th>Attribution</th> <th>Structure</th> </tr> </thead> <tbody> <tr> <td>Date: 01/03/1936</td> <td>Author: L'UNION des AVEUGLES DE GUERRE</td> <td>Type: subsubsection</td> </tr> <tr> <td>Year: 1936</td> <td></td> <td>Section: Décès</td> </tr> <tr> <td colspan="3">Document: Bulletin mensuel de l'Union des aveugles de guerre et Journal des soldats blessés aux yeux - 18ème a...</td> </tr> <tr> <td colspan="3">File: bulletin_uag_jsby_1936_03.xml</td> </tr> <tr> <td colspan="3"> <a href="#">UAG</a> <a href="#">Aveugles</a> <a href="#">Guerre</a> <a href="#">Assemblée</a> <a href="#">Maison</a> <a href="#">Boulleville</a> <a href="#">Loterie</a> </td> </tr> <tr> <td colspan="3"> <a href="#">Bibliothèque Braille</a> </td> </tr> <tr> <td colspan="3"> <span>doc_type: archival</span> <span>strategy: adaptive</span> </td> </tr> </tbody> </table>			Temporal	Attribution	Structure	Date: 01/03/1936	Author: L'UNION des AVEUGLES DE GUERRE	Type: subsubsection	Year: 1936		Section: Décès	Document: Bulletin mensuel de l'Union des aveugles de guerre et Journal des soldats blessés aux yeux - 18ème a...			File: bulletin_uag_jsby_1936_03.xml			<a href="#">UAG</a> <a href="#">Aveugles</a> <a href="#">Guerre</a> <a href="#">Assemblée</a> <a href="#">Maison</a> <a href="#">Boulleville</a> <a href="#">Loterie</a>			<a href="#">Bibliothèque Braille</a>			<span>doc_type: archival</span> <span>strategy: adaptive</span>		
Temporal	Attribution	Structure																								
Date: 01/03/1936	Author: L'UNION des AVEUGLES DE GUERRE	Type: subsubsection																								
Year: 1936		Section: Décès																								
Document: Bulletin mensuel de l'Union des aveugles de guerre et Journal des soldats blessés aux yeux - 18ème a...																										
File: bulletin_uag_jsby_1936_03.xml																										
<a href="#">UAG</a> <a href="#">Aveugles</a> <a href="#">Guerre</a> <a href="#">Assemblée</a> <a href="#">Maison</a> <a href="#">Boulleville</a> <a href="#">Loterie</a>																										
<a href="#">Bibliothèque Braille</a>																										
<span>doc_type: archival</span> <span>strategy: adaptive</span>																										
> Source 1 — score: 0.866 — NOTRE MAGASIN																										
> Source 2 — score: 0.854 — Qu'est-ce que la F. I. D. A. ?																										
> Source 3 — score: 0.853 — RÉGLEMENT DE FRANCEVILLE																										
> Source 4 — score: 0.853 — " L'œil qui voit "																										
> Source 5 — score: 0.853 — LISTE DES MEMBRES DU COMITÉ D'ACTION																										

FIGURE 4 – Retour par source après requête

## D Les requêtes faites au RAG (Corpus Paul d'Estournelles)

**Question :** Comment s'appelle la fille de Paul d'Estournelles

**Réponse** The daughter of Paul d'Estournelles is named Henriette.

Reference(s) : - doc\_title : Letter number 850 from Paul d'Estournelles de Constant to Nicholas Murray Butler (June 23, 1921) - file\_name : Lettre0850\_23juin1921.xml

---

**Question :** Les Italiens

**Réponse :**

1. Les Italiens ont été irrités contre les Français, les Anglais et les Américains (Letter number 724 from Paul d'Estournelles de Constant to Nicholas Murray Butler, November 26, 1920).
  2. Les Italiens ont reçu une acclamation pour le fascisme, qui réclame la Dalmatie, l'Adriatique, la Méditerranée, Nice, la Savoie, la Corse, la Tunisie, une partie de la Suisse, etc., et a tendu la main au féodalisme Hongrois contre les Slaves des Balkans (Letter number 1206 from Paul d'Estournelles de Constant to Nicholas Murray Butler, November 29, 1922).
  3. Les Italiens ne disent rien ni de Trieste, ni du Trentin, mais pensent aux sacrifices dont ils faut payer pour aboutir à des succès actuels. Ils ont également souligné que si les Allemands envahissent de nouveau la Péninsule, il n'est pas clair de savoir si il est la peine de se ruiner pour aboutir à ce résultat (Letter number 713 from Paul d'Estournelles de Constant to Nicholas Murray Butler, October 19, 1920).
- 

**Question :** Qui est Elihu Root ?

**Réponse :** Elihu Root is a person who is mentioned in the letters from Paul d'Estournelles de Constant to Nicholas Murray Butler, specifically in the documents "Letter number 1451 from Paul d'Estournelles de Constant to Nicholas Murray Butler (December 4, 1923)" and "Letter number 1438 from Paul d'Estournelles de Constant to Nicholas Murray Butler (November 22, 1923)".

Elihu Root was a prominent American statesman, lawyer, and academic who served as the United States Secretary of War and Secretary of State. He was also the President of the Carnegie Endowment for International Peace.

Reference(s) : - "Letter number 1451 from Paul d'Estournelles de Constant to Nicholas Murray Butler (December 4, 1923)", 1957 - "Letter number 1438 from Paul d'Estournelles de Constant to Nicholas Murray Butler (November 22, 1923)", 1957

---

**Question** Prix Nobel de la Paix

**Réponse** Based on the provided documents, the Nobel Peace Prize is not directly mentioned. However, in letter number 407 from Paul d'Estournelles de Constant to Nicholas Murray Butler dated August 9, 1918, it is mentioned that the Carnegie Endowment for International Peace is offering its services to the U.S. government to help treat international affairs related to the war. This organization was established by Andrew Carnegie, who received the Nobel Peace Prize in 1911. (1957, 1957)

---

**Question :** Traité de Versailles

**Réponse :**

1. The Treaty of Peace has been proclaimed and it will have to become a reality. (Letter number 14 from Paul d'Estournelles de Constant to Nicholas Murray Butler, December 3, 1914)
2. The Treaty of Versailles has been signed, but it remains to be applied and amended in such a way that this war truly ends and cannot resume. (Letter number 622 from Paul d'Estournelles de Constant to Nicholas Murray Butler, February 15, 1920)
3. The Treaty of Peace, as it was proclaimed in a letter from Paul d'Estournelles de Constant to Nicholas Murray Butler on December 3, 1914 (ref. Lettre0014\_3decembre1914.xml).
4. The Treaty of Versailles, which was signed but not yet executed, according to a statement in a letter from Paul d'Estournelles de Constant to Nicholas Murray Butler on September 18, 1923 (ref. Lettre1391\_18septembre1923.xml).
5. The Treaty of Versailles, which Germany has signed but is yet to be applied and amended to ensure the cessation of the war in a letter from Paul d'Estournelles de Constant to Nicholas Murray Butler on February 15, 1920 (ref. Lettre0622\_15fevrier1920.xml).
6. The Treaty of Versailles, which Germany has accepted without conditions in a letter from Paul d'Estournelles de Constant to Nicholas Murray Butler on June 24, 1919 (ref. Lettre0552\_24juin1919.xml).

## **E Les requêtes faites au RAG (Corpus EHRI)**

**Question :** La déportation des Juifs vers Auschwitz

**Réponse :** Les textes fournis décrivent des déportations de Juifs vers Auschwitz-Birkenau. Les témoignages de Otto Kalwo et Eduard de Wind font mention de cette destination. (Documents EHRI-DC-054\_DE\_EN.xml et EHRI-DC-066\_DE\_EN.xml)

---

**Question :** Comment est la vie dans le ghetto de terezin ?

**Réponse :** La vie dans le ghetto de Terezin était difficile. Les barracks étaient surpeuplés, une grande partie des ghettoisés devaient dormir sur le sol, car tout espace était utilisé comme lieu de sommeil. Sans lits, sans matelas ou sacs à paillasses, seulement quelques-uns étaient temporairement disponibles. Les campagnards dormaient sur des sols froids, beaucoup d'entre eux dans des salles telles que les écuries, les entrepôts et les lieux similaires sur le sol de ciment dur. Dans les baraques des Sudè

---

**Question :** attaque antisémite violente

**Réponse :** The documents provide multiple evidence of violent anti-Semitic attacks in Slovakia. In the memorandum by the Association of the Victims of Racial Persecution Košice chapter, they mention anti-Jewish riots (doc\_title : Memorandum on anti-Jewish riots in Slovakia, Memorandum zu antijüdischen Ausschreitungen in der Slowakei, Memorandum o protižidovských nepokojích na Slovensku, paragraph xml\_id : ). Furthermore, the Association of the Victims of Racial Persecution, eastern district of Košice, ...

---

**Question :** hitler

**Réponse :** The query "hitler" does not appear in the provided documents.

---

**Question** Aggression non juive

**Réponse** At the personal level, the persecution is even more severe. Association with Germans is strictly forbidden, and even kids are not allowed to play with German kids. Although in legal terms we are not expelled from the country, the method used goes against humanity and arouses (uncontrollable) anger. (Giuseppe Talamo Atenolfi on Hungarian actions against Jews in occupied Galicia in 1942, Giuseppe Talamo Atenolfi über ungarische Aktionen gegen Juden im besetzten Galizien im Jahr 1942, EHRI-DR-19

---

**Question :** Aggression non-juive

**Réponse :**

1. In the document "A fifteen-year-old youth, on the German invasion of Wyszaków and surrounding areas" written on an undetermined date, the author describes how Jews were severely beaten. (keywords : testimony, robberies, beaten)
  2. In the document "Ray Atherton on the reaction of the non-Jewish Bulgarian population towards antisemitic demonstrations in 1939", the author mentions that there is no sympathy in Bulgaria for the antisemitic demonstrations of the organization "Workers of New Bulgaria" ...
-

**Question :** Qui est persecuté outre les juifs

**Réponse :** Les juifs sont persecutés en plus de quelqu'un d'autre, mais il n'est pas spécifié dans les documents fournis qui sont les autres personnes qui sont persecutées.