Offre de stage (Ma ster/Ingénieur)

IA & Deep Learning : OCR et LLMs pour la segmentation de dictionnaires anciens

Travail de recherche appliquée	Combinaison de Vision par Ordinateur (VLM/OCR/OLR) et LLMs
Accès à des modèles récents	MistralOCR, LLaMA 3, GPT (API OpenAI), modèles de raisonnement
Environnement technique	Cluster PAGODA du LIRIS, outils d'annotation (LabelStudio, LayoutParser)
Impact	Valorisation scientifique et patrimoniale : contribution à des corpus historiques prestigieux (Dictionnaire de Trévoux, Grande Encyclopédie)
Valorisation	Possibilité de publications / valorisation académique

Mots clés

LLM, VLM, OLR/OCR, segmentation automatique d'entrées lexicographique

Contexte

Le projet OLR-LLM[1] financé par l'IXXI[2] (Institut Rhônalpin des Systèmes Complexes) et la MSH Lyon St-Etienne[3] s'inscrit dans le cadre d'une collaboration entre le laboratoire d'informatique LIRIS[4] et le laboratoire de science du langage ICAR[5].

Les corpus patrimoniaux numérisés, tels que les dictionnaires et encyclopédies anciennes, constituent des sources précieuses pour la recherche en humanités numériques, en lexicographie historique et en histoire des savoirs. Cependant, leur exploitation scientifique est souvent entravée par la complexité de leurs formats d'encodage, issus de pipelines de numérisation standards comme celui de la BnF, qui produit des fichiers XML METS/ALTO accompagnant des images haute définition. Ces formats mêlent informations structurelles, spatiales (position des blocs de textes sur l'image) et textuelles issues d'OCR, mais nécessitent un traitement avancé pour permettre une segmentation sémantique fiable des documents.

Dans le domaine de la **lexicographie numérique**, plusieurs initiatives ont porté sur la structuration automatique de dictionnaires historiques et ont montré l'importance d'un balisage structuré pour la réutilisation scientifique de dictionnaires anciens, en soulignant les difficultés liées à l'identification automatique des unités lexicographiques dans des contextes de typographie non normalisée (Galleron & Williams, 2022). La segmentation des entrées reste une tâche peu explorée à grande échelle.

[1] https://www.ixxi.fr/projets/projets-finances-en-2025#section-4
[2] https://www.ixxi.fr
[3] https://www.msh-lse.fr
[4] https://liris.cnrs.fr
[5] https://icar.cnrs.fr

Concernant l'analyse des **images de documents patrimoniaux**, les méthodes classiques d'OCR ont longtemps reposé sur des moteurs comme Tesseract (Smith, 2007), mais ceux-ci montrent leurs limites face aux variations typographiques des ouvrages anciens. Les approches plus récentes **d'OCR/OLR (Optical Layout Recognition)** intégrant des réseaux de neurones convolutifs ou des Transformers, comme **Donut** (Kim et al., 2022) ou **LayoutLMv3** (Huang et al., 2022), offrent des perspectives prometteuses en combinant reconnaissance de texte et compréhension de la structure de page. Néanmoins, ces modèles restent peu adaptés à des documents long format aux entrées longues et à hiérarchie lexicale, comme ceux du Dictionnaire Universel François-Latin (Romanello et al. 2021 ; Pinche & Stokes, 2024).

L'arrivée des **grands modèles de langage (LLMs)**, tels que GPT, LLaMA (Touvron et al., 2023), ou Mistral, a ouvert de nouvelles voies pour le traitement de textes patrimoniaux (Scius-Bertrand et al., 2024). Leur capacité à modéliser des séquences longues, à inférer des structures implicites et à s'adapter via le prompt engineering ou le fine-tuning les rend particulièrement pertinents pour la segmentation sémantique d'entrées dans des corpus lexicographiques. Les deux corpus sélectionnés dans ce projet — le *Dictionnaire de Trévoux* (1704–1771) (DUFLT) et *La Grande Encyclopédie* (1886–1902) (LGE) — posent des défis contrastés mais complémentaires. Le premier est marqué par une grande variabilité dans la structuration des entrées et un usage dense de subdivisions internes, ce qui en fait un cas emblématique pour tester la finesse d'analyse contextuelle des LLMs. Le second, plus récent, propose une organisation typographique plus régulière, mais présente un volume massif de données qui pose des enjeux d'échelle et de généralisation.

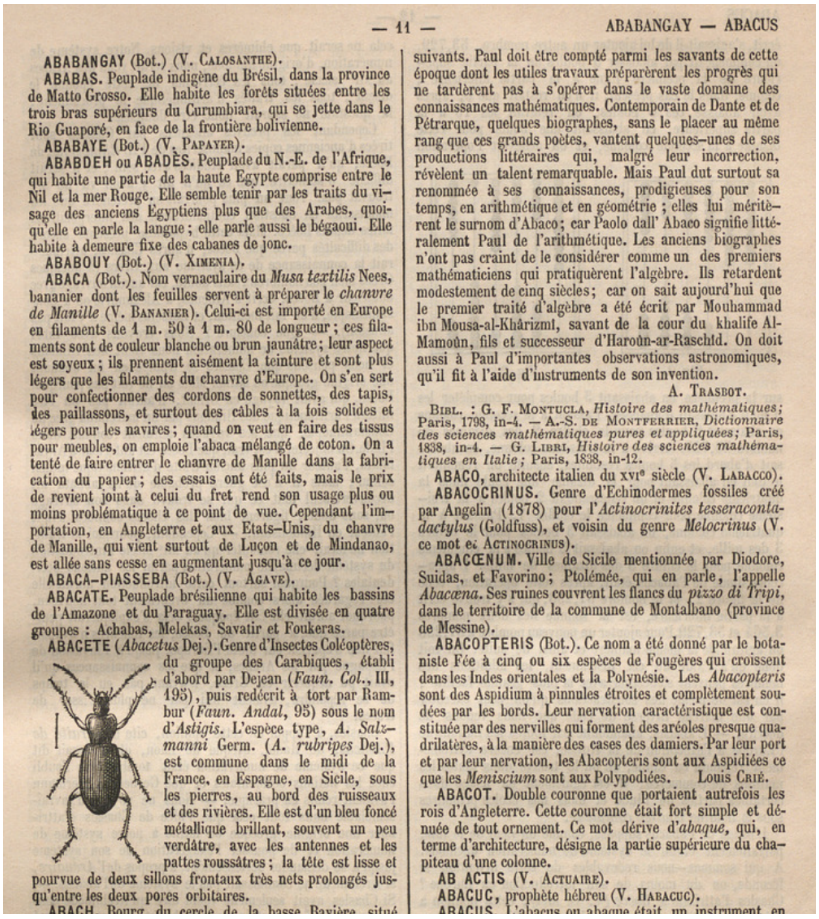
Ce projet se situe donc à l'intersection de plusieurs champs disciplinaires — traitement automatique des langues, vision par ordinateur, humanités numériques — et ambitionne de proposer une méthodologie innovante pour l'analyse structurelle de corpus patrimoniaux. En mobilisant à la fois des approches fondées sur la reconnaissance de structure via image (OLR) et sur l'interprétation linguistique contextuelle via texte (LLMs), il s'inscrit dans une dynamique de convergence entre IA et sciences du patrimoine.

Objectifs du stage

Le cœur de ce stage repose sur l'exploration des capacités des grands modèles de langage à segmenter des corpus lexicographiques anciens issus de la numérisation patrimoniale. Plus précisément, il s'agit de mettre au point une chaîne de traitement automatique permettant d'identifier et de structurer les entrées lexicographiques dans deux types de formats sources : d'une part, des images de pages numérisées au format PDF (voir Figure 1), et d'autre part, des fichiers XML METS/ALTO, produits par le pipeline de numérisation de la BnF.

Les modèles traditionnels d'analyse de documents anciens peinent à identifier la granularité sémantique pertinente dans ce type de textes, souvent composites, hiérarchisés, voire redondants. C'est notamment le cas pour le *Dictionnaire Universel François-Latin* de Trévoux, où une entrée lexicographique peut s'étendre sur plusieurs colonnes et intégrer définitions, exemples, citations, remarques ou variantes orthographiques. À l'opposé, *La Grande Encyclopédie*, bien que plus récente et typographiquement plus régulière, représente un défi par son ampleur et son niveau de détail.

Dans un premier temps, le stage proposera un traitement différencié selon le type de source (voir la Figure 2). Pour les documents au format image (PDF), l'approche reposera sur un VLM (Vision Langage Model) tel que le modèle **MistralOCR**, qui combine reconnaissance optique des caractères (OCR), reconnaissance de la structure de page (OLR), et compréhension linguistique par LLM. Ce modèle sera interrogé via l'API MistralAI afin de tester sa capacité à détecter les blocs lexicographiques, à segmenter correctement les entrées et à identifier les différentes sous-sections internes (définitions, exemples, étymologies, etc.). Des premières expérimentations d'OCR et de segmentation des entrées sur quelques pages numérisées de LGE ont montré des résultats très encourageants comparés à des approches plus classiques.



— 11 — ABABANGAY — ABACUS

ABABANGAY (Bot.) (V. Calosanthe).

ABABAS. Peuplade indigène du Brésil, dans la province de Matto Grosso. Elle habite les forêts situées entre les trois bras supérieurs du Curumbiara, qui se jette dans le Rio Guaporé, en face de la frontière bolivienne.

ABABAYE (Bot.) (V. Papayer).

ABABDEH ou **ABADÈS.** Peuplade du N.-E. de l'Afrique, qui habite une partie de la haute Egypte comprise entre le Nil et la mer Rouge. Elle semble tenir par les traits du visage des anciens Égyptiens plus que des Arabes, quoiqu'elle en parle la langue ; elle parle aussi le bégaoui. Elle habite à demeure fixe des cabanes de jonc.

ABABOUY (Bot.) (V. Ximenia).

ABACA (Bot.). Nom vernaculaire du *Musa textilis* Nees, bananier dont les feuilles servent à préparer le *chanvre de Manille* (V. Bananier). Celui-ci est importé en Europe en filaments de 1 m. 50 à 1 m. 80 de longueur ; ces filaments sont de couleur blanche ou brun jaunâtre ; leur aspect est soyeux ; ils prennent aisément la teinture et sont plus légers que les filaments du chanvre d'Europe. On s'en sert pour confectionner des cordons de sonnettes, des tapis, des paillassons, et surtout des câbles à la fois solides et légers pour les navires ; quand on veut en faire des tissus pour meubles, on emploie l'abaca mélangé de coton. On a tenté de faire entrer le chanvre de Manille dans la fabrication du papier ; des essais ont été faits, mais le prix de revient joint à celui du fret rend son usage plus ou moins problématique à ce point de vue. Cependant l'importation, en Angleterre et aux Etats-Unis, du chanvre de Manille, qui vient surtout de Luçon et de Mindanao, est allée sans cesse en augmentant jusqu'à ce jour.

ABACA-PIASSEBA (Bot.) (V. Agave).

ABACATE. Peuplade brésilienne qui habite les bassins de l'Amazone et du Paraguay. Elle est divisée en quatre groupes : Achabas, Melekas, Savatir et Foukeras.

ABACETE (*Abacetus* Dej.). Genre d'Insectes Coléoptères, du groupe des Carabiques, établi d'abord par Dejean (*Faun. Col.*, III, 193), puis redécrit à tort par Rambur (*Faun. Andal*, 93) sous le nom d'*Astigis*. L'espèce type, *A. Salzmanni* Germ. (*A. rubripes* Dej.), est commune dans le midi de la France, en Espagne, en Sicile, sous les pierres, au bord des ruisseaux et des rivières. Elle est d'un bleu foncé métallique brillant, souvent un peu verdâtre, avec les antennes et les pattes roussâtres ; la tête est lisse et pourvue de deux sillons frontaux très nets prolongés jusqu'entre les deux pores orbitaires.

ABACH. Bourg du cercle de la basse Bavière, situé

suivants. Paul doit être compté parmi les savants de cette époque dont les utiles travaux préparèrent les progrès qui ne tardèrent pas à s'opérer dans le vaste domaine des connaissances mathématiques. Contemporain de Dante et de Pétrarque, quelques biographes, sans le placer au même rang que ces grands poètes, vantent quelques-unes de ses productions littéraires qui, malgré leur incorrection, révèlent un talent remarquable. Mais Paul dut surtout sa renommée à ses connaissances, prodigieuses pour son temps, en arithmétique et en géométrie ; elles lui méritèrent le surnom d'Abaco ; car Paolo dall' Abaco signifie littéralement Paul de l'arithmétique. Les anciens biographes n'ont pas craint de le considérer comme un des premiers mathématiciens qui pratiquèrent l'algèbre. Ils retardent modestement de cinq siècles ; car on sait aujourd'hui que le premier traité d'algèbre a été écrit par Mouhammad ibn Mousa-al-Khârizmi, savant de la cour du khalife Al-Mamoûn, fils et successeur d'Haroûn-ar-Raschid. On doit aussi à Paul d'importantes observations astronomiques, qu'il fit à l'aide d'instruments de son invention.

A. Trasbot.

Bibl. : G. F. Montucla, *Histoire des mathématiques* ; Paris, 1798, in-4. — A.-S. de Montferrier, *Dictionnaire des sciences mathématiques pures et appliquées* ; Paris, 1838, in-4. — G. Libri, *Histoire des sciences mathématiques en Italie* ; Paris, 1838, in-12.

ABACO, architecte italien du XVI[e] siècle (V. Labacco).

ABACOCRINUS. Genre d'Echinodermes fossiles créé par Angelin (1878) pour l'*Actinocrinites tesseraconta-dactylus* (Goldfuss), et voisin du genre *Melocrinus* (V. ce mot et Actinocrinus).

ABACŒNUM. Ville de Sicile mentionnée par Diodore, Suidas, et Favorino ; Ptolémée, qui en parle, l'appelle *Abacœna*. Ses ruines couvrent les flancs du *pizzo di Tripi*, dans le territoire de la commune de Montalbano (province de Messine).

ABACOPTERIS (Bot.). Ce nom a été donné par le botaniste Fée à cinq ou six espèces de Fougères qui croissent dans les Indes orientales et la Polynésie. Les *Abacopteris* sont des Aspidium à pinnules étroites et complètement soudées par les bords. Leur nervation caractéristique est constituée par des nervilles qui forment des aréoles presque quadrilatères, à la manière des cases des damiers. Par leur port et par leur nervation, les Abacopteris sont aux Aspidiées ce que les *Meniscium* sont aux Polypodiées. Louis Crié.

ABACOT. Double couronne que portaient autrefois les rois d'Angleterre. Cette couronne était fort simple et dénuée de tout ornement. Ce mot dérive d'*abaque*, qui, en terme d'architecture, désigne la partie supérieure du chapiteau d'une colonne.

AB ACTIS (V. Actuaire).

ABACUC, prophète hébreu (V. Habacuc).

ABACUS. L'abacus ou abaque était un instrument en

Figure 1 : Extrait d'une page numérisée de LGE. La structure peut être plus complexe avec des tableaux ou figures sur toute la largeur de la page ou avec des articles longs de plusieurs pages, etc.

En parallèle, pour les documents encodés en XML METS/ALTO, le projet prévoit une étape de segmentation des entrées visant à exploiter pleinement les métadonnées spatiales et structurelles (coordonnées de blocs, hiérarchie de zones, styles typographiques). Ces fichiers contiennent souvent des zones textuelles fragmentées et mal alignées avec la structure logique du document. Le pipeline développera des algorithmes de fusion et de réorganisation des blocs, en vue d'obtenir une version "reconstruite" du document apte à une analyse linguistique fine. Cette reconstruction reposera sur l'utilisation de LLMs, en particulier **LLaMA 3 70B**, déployé sur la plateforme PAGODA du LIRIS et les derniers modèles GPT via l'API d'OpenAI (incluant les modèles de raisonnement). Des techniques avancées de *prompt*

engineering permettront d'orienter le modèle vers une compréhension structurée du texte, afin de détecter les débuts et fins d'entrées lexicographiques, les niveaux de hiérarchie interne, ainsi que les ruptures discursives. L'approche s'appuiera notamment sur des prompts contextuels, des annotations d'exemples (few-shot prompting), voire des tentatives de *chain-of-thought* pour affiner les délimitations dans les cas ambigus. Une première expérimentation du modèle GPT O4-mini a montré des résultats intéressants en prenant en entrée à la fois le PDF (texte) et le fichier ALTO correspondant.

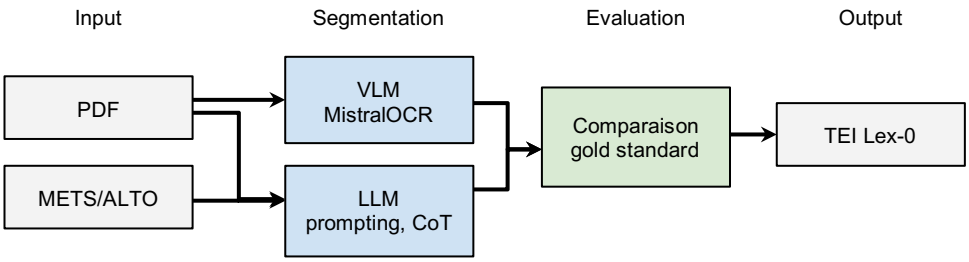


Figure 2 : Schéma du pipeline

Un volet essentiel du stage portera sur la comparaison des deux approches (PDF image + MistralOCR vs PDF texte & XML + LLaMA/GPT), à la fois du point de vue des performances quantitatives (précision, rappel, F1-score) et qualitatives (cohérence interne, respect de la hiérarchie lexicographique, lisibilité des extraits produits). Nous évaluerons en particulier la qualité de la transcription (dans le cas de l'approche ne reposant pas sur le format ALTO) et la bonne segmentation des entrées du dictionnaire (limites début et fin des articles correctement identifiées). Pour ce faire, un **corpus d'évaluation annoté manuellement** sera constitué à partir d'un échantillon représentatif des deux œuvres étudiées. Ce corpus "gold standard" permettra également de calibrer certains modules spécifiques, notamment en cas d'échec partiel des LLMs ou de nécessité de post-traitements. Pour la partie segmentation le ou la stagiaire pourra s'appuyer sur des échantillons de pages déjà annotées (voir Figure 3) et sur un modèle LayoutParser fine-tuné. L'architecture du pipeline se voudra modulaire et réplicable, afin de pouvoir être appliquée à d'autres dictionnaires anciens ou encyclopédies, y compris dans d'autres langues. À terme, le projet vise à produire une **chaîne de traitement semi-automatique** capable d'assister des bibliothèques patrimoniales, des chercheurs en humanités numériques ou des lexicographes et historiens dans l'exploitation structurée de vastes corpus imprimés.

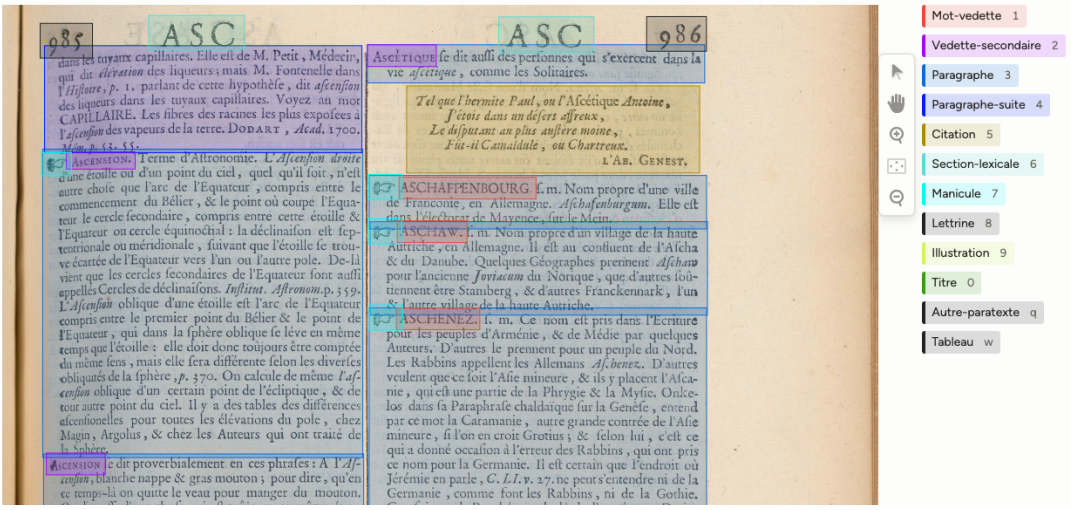


Figure 3 : Extrait d'une page du DUFLT annotée avec LabelStudio.

Enfin, tous les outils, jeux de données annotés et scripts développés dans le cadre du projet seront diffusés en open source, dans une logique d'interopérabilité avec les standards de la lexicographie numérique (TEI Lex-0) et de valorisation des collections numériques patrimoniales. Le projet ambitionne ainsi de contribuer au développement d'une méthode robuste, reproductible et transposable pour la segmentation de textes complexes dans le champ des humanités computationnelles.

Bibliographie

Francois, M., Eglin, V., & Biou, M. (2022). Text detection and post-OCR correction in engineering documents. In International Workshop on Document Analysis Systems (pp. 726-740). Cham: Springer International Publishing.

Galleron, I., & Williams, G. C. (2022). Tenir la promesse du Dictionnaire universel: l'esprit encyclopédique d'Henri Basnage de Beauval. Langue française, 214(2), 27-42.

Huang, Y., et al. (2022). LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. arxiv.org/abs/2204.08387

Kim, G., Hong, T., et al. (2022). Donut: Document Understanding Transformer without OCR. ECCV 2022.

Lairgi, Y., Moncla, L., Cazabet, R., Benabdeslem, K., & Cléau, P. (2024). itext2kg: Incremental knowledge graphs construction using large language models. In International Conference on Web Information Systems Engineering (pp. 214-229). Singapore: Springer Nature Singapore.

Moncla, L., Joliveau, T., & Vigier, D. (2024). Propositions pour une étude interdisciplinaire de la géographie dans un dictionnaire universel et une encyclopédie du XVIIIe siècle. In 1er colloque du réseau METALEX:«Lexicographie, Métalexicographie, nouveaux défis».

Pinche, A., & Stokes, P. (2024). « Historical Documents and Automatic Text Recognition: Introduction ». Journal of Data Mining & Digital Humanities. Documents historiques et reconnaissance automatique de textes. https://doi.org/10.46298/jdmdh.13247.

Romanello, M., Najem-Meyer, S., et Robertson, B.. 2021. « Optical Character Recognition of 19th Century Classical Commentaries: the Current State of Affairs ». Dans Proceedings of the 6th International Workshop on Historical Document Imaging and Processing. New York : ACM. https://doi.org/10.1145/3476887.3476911.

Scius-Bertrand, A. et al. (2024) Are Layout Analysis and OCR Still Useful for Document Information Extraction Using Foundation Models? ICDAR 2024

Smith, R. (2007). An Overview of the Tesseract OCR Engine. In Proceedings of the Ninth International Conference on Document Analysis and Recognition.

Touvron, H., Lavril, T., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. Meta AI. arxiv.org/abs/2302.13971

Zeghidi, H., & Moncla, L. (2024). Evaluating named entity recognition using few-shot prompting with large language models. arXiv preprint arXiv:2408.15796.

Déroulement du stage

Profils recherchés : Master 2 Informatique / Ingénieur

Des compétences sont attendues en programmation, en IA (Machine Learning et Deep Learning) et en traitement d'image. Des connaissances en traitement automatique de la langue (TAL) seront appréciées.

Bon niveau en français exigé.

Rémunération : environ 630€ par mois

Lieu : Laboratoire LIRIS – INSA Lyon, Bâtiment Blaise Pascal, Campus La Doua, Villeurbanne.

Date de début : février/mars 2026

Durée : 5 à 6 mois

Candidature : Envoyer un mail présentant votre parcours, vos motivations ainsi que votre CV et vos derniers relevés de notes à : ludovic.moncla@insa-lyon.fr et veronique.eglin@insa-lyon.fr

Date limite de candidature : 14 novembre 2025 (entretiens au fil de l'eau)