## Offre de stage (Master/Ingénieur)

# IA et encodage TEI de gros volumes de données textuelles

Un des défis pour l'étude quantitative de très gros volumes de données textuelles obtenues par numérisation + OCR/OLRisation est de parvenir à les encoder au standard international XML-TEI de la manière la plus automatisée possible tout en assurant un résultat le plus fiable et le plus fin possible.

**L'objectif de ce stage**, **financé par le LabEex ASLAN,** est de tester et comparer des stratégies d'encodage assistée par l'intelligence artificielle et en particulier les modèles génératifs. Il fait suite aux travaux exploratoires présentés à la 15è édition de l'*International Conference on Historical Lexicography and* Lexicology (ICHLL15) Lisbonne, 2025 (Vigier, Lesnova & al. 2025). L'œuvre qui sera l'objet des travaux conduits dans ce stage est le *Dictionnaire Universel François et Latin* de Trévoux (désormais DUFLT) édité entre 1704 et 1771 (huit éditions *in folio* successives). Ce dictionnaire a été relativement peu étudié jusqu'à présent et n'a pas fait l'objet de tentative aboutie de numérisation et de traitement informatique, de sorte qu'on ne dispose actuellement, pour ce qui le concerne, que de très peu de données textuelles numériques annotées et structurées.

Les laboratoires impliqués dans ce stage (à ICAR et LIRIS) disposent des versions (re)numérisé par la BNF *via* son infrastructure *Gallica* de six des huit éditions *in folio* (1704, 1721, 1732, 1743, 1752, 1771). Ils disposent aussi de la version numérisée, océrisée et balisée en XML de l'édition de 1743. Les expérimentations projetées dans le cadre de ce stage ont pour but de mettre au point un pipeline de traitement afin de faire évoluer, *a minima*, l'annotation XML dont nous disposons pour l'édition de 1743 vers un balisage XML-TEI en recourant au système de notation préconisé par la TEI Lex-0. Cet objectif passera par les étapes suivantes :

**Etape 1**. Consolidation du choix d'encodage XML-TEI « de surface » mis au point durant un précédent stage. Ce schéma vise à sélectionner parmi les préconisations de la TEI Lex-0 toutes celles dont le balisage met en jeu des traces linguistiques de surface : typo-disposition du texte, paradigme d'expressions figées ou d'abréviations indiquant le domaine traité dans l'article, la catégorie morphosyntaxique du mot figurant en vedette … Ce choix d'un encodage de surface vise à guider l'IA grâce à des marques relevant de systèmes de structuration textuelle explicitement présents à la surface du texte et à éviter une annotation qui serait basée uniquement sur des indices sémantiques.

**Etape 2**. Tests, comparaisons et évaluation de différentes méthodes d'encodage automatisée mettant en jeu l'IA générative. Cette étape mettra en place une méthodologie pour développer l'approche la plus performante afin d'intégrer l'IA générative dans le processus d'encodage XML-TEI. Le/la stagiaire devra mettre en place une expérimentation incrémentale allant d'un texte déjà segmenté et structuré vers du texte numérisé en sortie d'OCR « au kilomètre » non corrigé. Nos premières expérimentations (présentées à la conférence ICHLL 2015) montrent des résultats très encourageants. Notre proposition vise à implémenter une approche hybride reposant à la fois sur une approche de *prompt* ingénierie (intégrant une approche *few-shot*) et de post-traitements pour limiter les risques « d'hallucinations » et en particulier de modification ou d'altération du texte source. Le corpus d'expérimentation et d'évaluation utilisé sera composé de l'édition de 1743 du Trévoux et du contenu de différentes pages tirées aléatoirement dans les autres éditions du Trévoux pour lesquelles nous devrons construire à la main la vérité terrain.

**Etape 3**. Élargissement des limites de l'encodage de surface. La dernière phase du travail visera à « repousser » le plus possible les limites de l'encodage de « surface » en testant le balisage XML-TEI d'informations dont la détection nécessite une interprétation plus complexe (par exemple, l'identification des citations etc.).

Les retombées des résultats obtenus à l'issue de ce stage sont potentiellement très importantes pour la communauté des spécialistes de lexicographie numérique pour les textes anciens. Aucune publication à ce jour ne fait état d'expérimentations visant à mettre en jeu l'IA générative pour cette tâche cruciale d'encodage XML-TEI. Les derniers travaux de recherche traitant de très gros volumes de données en lexicographie, réalisés dans le cadre de l'ANR BASNUM, ont bénéficié du traitement par logiciel GROBID-dictionaries développé par M. Kehmakem durant sa thèse à l'INRIA. Or ce logiciel s'appuie sur des modèles d'apprentissage machine qui n'intègrent pas les derniers développements de l'IA. Nous avons montré à plusieurs occasions (Brenon & Vigier 2022, Vigier, Lesnova & al. 2025) que ses performances rencontraient des limites problématiques que les nouveaux développements de l'IA permettent précisément de dépasser. La publication des premiers résultats dans le cadre d'un article publié avec les encadrants du stage (Denis Vigier et Ludovic Moncla) constituerait donc un apport à la communauté.

**Bibliographie**


Brenon, A. & Vigier, D. (2021). « The specificities of encoding encyclopedias: towards a new standard? ». ICHLL11: 11th International Conference on Historical Lexicography and Lexicology, 16 juin 2021, Logroño, Universidad de La Rioja (Espagne). HAL : halshs-03266745.

Vigier, D., Lesnova, T., Moncla, L., Joliveau, T. (2025), "Universal dictionaries and ancyclopædias from the eighteenth century to the digital age". *ICHLL15 International Conference on Historical Lexicography and Lexicology*, Jun 2025, Lisboa, Portugal. ⟨hal-05135577⟩


**Déroulement du stage**

Profils recherchés : Master 2 TAL / Humanités numériques

- ✓ Des compétences minimales sont attendues dans le domaine de la TEI et en IA
- ✓ Bon niveau en français exigé.
- ✓ Rémunération : environ 630€ par mois
- ✓ Lieu : Laboratoire ICAR – Ecole Normale Supérieure de Lyon, Bâtiment Recherche, Site LSH.
- ✓ 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 : denis.vigier@ens-lyon.fr et ludovic.moncla@insa-lyon.fr

Date limite de candidature : 21 novembre 2025 (entretiens au fil de l'eau)