Offre de stage (Master/Ingénieur)

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

Un projet innovant à l'intersection de l'informatique, du traitement du langage et de la visio par ordinateur et pluridisciplinaire, pleinement orienté IA appliquée aux humanités numériques.

- **Vision par ordinateur & OCR avancé** : exploitation de modèles tels que MistralOCR ou Donut pour détecter et structurer les blocs de texte.
- **Traitement automatique du langage naturel** : utilisation de GPT, LLaMA ou Mistral pour segmenter les entrées lexicographiques et identifier leur hiérarchie interne.
- **Pipeline multi-source** : intégration et fusion de PDFs image et de fichiers XML METS/ALTO pour le traitement des documents.
- **Évaluation** : analyse de performances par précision, rappel et F1-score sur un corpus annoté

Les dictionnaires et encyclopédies anciens sont des trésors pour les humanités numériques, mais leur exploitation est limitée par la complexité des formats de numérisation. Ce stage s'inscrit dans une collaboration entre le laboratoire d'informatique LIRIS et le laboratoire de sciences du langage ICAR, avec pour objectif d'explorer l'apport des grands modèles de langage (LLMs) et de la vision par ordinateur dans l'analyse de corpus patrimoniaux.

Le projet porte sur la segmentation automatique d'entrées lexicographiques à partir de données numérisées (PDF et XML METS/ALTO). Le/la stagiaire explorera des approches récentes combinant OCR/OLR (MistralOCR, LayoutLM, etc.) et LLMs (GPT, LLaMA 3) pour identifier et structurer les articles de deux œuvres majeures : le Dictionnaire de Trévoux (date ?) et la Grande Encyclopédie (date ?). L'objectif est de développer une chaîne de traitement innovante alliant vision et langage, avec évaluation sur un corpus annoté et diffusion open source des outils.

Le/la stagiaire acquerra une solide expérience pratique en IA (Deep Learning, NLP, Vision-Langage Models), en traitement de documents complexes et en développement open source. Il/elle participera à un projet de recherche interdisciplinaire à fort impact scientifique, avec possibilité de valorisation (publications, collaborations).

Une version détaillée du sujet est disponible ici : [Offre de stage détaillée]

Infos pratiques

- Lieu : LIRIS – INSA Lyon (Campus La Doua, Villeurbanne)
- Début : février/mars 2026 | Durée : 5–6 mois
- Rémunération : ~630 €/mois
- Profil recherché : Master 2 / Ingénieur en informatique (IA, Machine Learning, Deep Learning, TAL, vision)
- Candidature avant le 14 novembre 2025 (entretiens au fil de l'eau)
- Envoyer CV + relevés de notes + lettre de motivation à : ludovic.moncla@insa-lyon.fr et veronique.eglin@insa-lyon.fr

LIRIS INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON msh Lyon St-Étienne IXXI