Offre de stage en traitement et analyse des carnets de fouilles de Gergovie



Intitulé du stage : « Fouiller la fouille. Quand les archives deviennent données »

Contexte

Entre 2022 et 2025, la Maison des Sciences Humaines de Clermont-Ferrand (UAR 3550, UCA/CNRS) et l'Institut de recherches archéologiques préventives (Inrap) ont initié un programme de recherches sur l'oppidum de Gergovie (Puy-de-Dôme), associant fouilles archéologiques et étude historiographique d'une pluralité d'investigations conduites sur le site entre le Second Empire et les lendemains de la Seconde Guerre mondiale. L'examen de ces travaux anciens visait à retracer les pratiques scientifiques des précédentes générations d'archéologues : d'une part, en refouillant leurs excavations, d'autre part, en collectant et en exploitant leurs archives de terrain (carnets de fouilles, photographies, documents graphiques), largement disséminées à travers l'Europe.

Les carnets de fouilles retrouvés ont donné lieu à un vaste chantier de retranscription de l'écriture manuscrite (HTR) en utilisant des technologies basées sur l'apprentissage machine qui ont permis de produire des modèles adaptés à chaque scribe. À l'instar d'autres projets récents d'analyse de données en archéologie, les opérations de transcription et d'entraînement des modèles ont été réalisées à l'aide d'eScriptorium, logiciel libre et open source développé et mis à disposition par l'équipe ALMAnaCH de l'INRIA. Les données produites aux formats TXT et XML-ALTO ont permis la consultation synchronisée des images numérisées et de leur transcription au sein d'une même visionneuse.

Contenu du stage

Dans le cadre de ce projet, nous souhaitons développer un outil expérimental qui permettra à un/une archéologue de poser une question en langage naturel afin d'obtenir des passages pertinents à partir des carnets de fouilles de Gergovie. L'intérêt de l'outil est de fournir des réponses sur la base d'une recherche sémantique et non pas d'occurrences textuelles. Pour cela, nous mettons en place une chaîne de traitement des textes qui comprendra les étapes de nettoyage de données (corrections des erreurs, normalisation, lemmatisation), de structuration (encodage en XML-TEI et associations des métadonnées) et de vectorisation (par le biais de l'architecture Transformers et un modèle multilingue).

Nous souhaitons mener une partie de ce travail avec un.e stagiaire en histoire, littérature, langues, archéologie ou humanités numériques. L'étudiant ou l'étudiante recruté(e) interviendra dans toutes les phases de la chaîne de traitement, en particulier la phase de prétraitement et d'analyse des résultats.

Une formation aux technologies utilisées sera dispensée en début de stage.

Missions

- Relecture et/ou correction des textes retranscrits ;
- Vérification et intégration des métadonnées dans les fichiers textuels ;
- Structuration des textes en suivant le modèle TEI (Texte Encoding Initiative) ;
- Identification et balisage des thématiques et/ou mots-clés en accord avec la responsable scientifique du projet.

Profils recherchés : Étudiant.e en master 2 avec une spécialité dans l'une des disciplines suivantes : histoire, archéologie, lettres, langues, humanités numériques.

Durée : 4 mois

Début du stage : entre février et mars 2027

Gratification : 4,50€ / heure

Lieu : Bibliothèque de la Maison des Sciences Humaines, 4 rue Ledru, 63001 Clermont-Ferrand

Encadrantes :

- Aurelia Vasile : ingénieure en humanités numériques
- Marion Dacko : responsable scientifique du projet « Gergovie – Archives de fouilles »

CV et lettre de motivation à transmettre au plus tard le 18 décembre

à l'adresse suivante : archives-gergovie.msh@uca.fr