



Stage de Master II

Collecte et alignement de données pour la traduction automatique dans les musées

Présentation

L'objectif de ce stage est de développer des ressources multilingues alignées spécifiquement pour la traduction automatique dans le domaine muséal. Les trois langues cibles sont le français, l'arabe et l'anglais. La mission inclut la collecte et l'alignement de terminologies dans ces langues pour créer des ressources parallèles. Celles-ci, enrichies de métadonnées, pourront ensuite être utilisées pour entraîner un modèle de traduction automatique adapté aux besoins des musées.

Approche méthodologique

La méthodologie repose sur la création de ressources bilingues ou multilingues pertinentes au sein des institutions culturelles. La collecte des données consistera à extraire automatiquement des termes issus d'archives institutionnelles ou de collections numériques (par exemple, descriptions d'œuvres, textes d'exposition, catalogues, cartels). L'alignement multilingue sera basé sur des approches classiques (statistiques) ou modernes (embeddings), avec une vérification et correction manuelles. Enfin, les données seront enrichies par des annotations incluant des métadonnées (contexte culturel, termes techniques...), structurées selon des standards tels que TEI ou RDF, en s'appuyant sur un glossaire disciplinaire existant.

Apports

Acquisition de compétences en gestion de corpus multilingues, extraction de terminologies, annotation enrichie et création de ressources adaptées au domaine muséal.

Profil recherché

Niveau Master 2 en TAL, humanités numériques, sciences de l'information ou dans un domaine connexe. Compétences attendues en gestion de corpus multilingues et en annotation de données. Une familiarité avec les modèles de langage et la programmation en Python est également nécessaire. Une maîtrise du français et de l'anglais est requise, l'arabe serait un atout.

Conditions du stage

Lieu de travail : ObTIC/SCAI, campus Pierre et Marie Curie – Sorbonne Université.

Temps de travail hebdomadaire : 35 h

Durée du stage : 4 à 6 mois

Début du stage : février ou mars 2025.

Gratification : la gratification perçue est conforme au taux légal en vigueur.

Modalités de recrutement

CV et bulletin de notes à envoyer par mail à motasem.alrahabi@sorbonne-universite.fr

Date limite de candidature : 17 janvier 2025.