Moyen-Orient et
Mondes Musulmans
Groupement d'Intérêt Scientifique

Post-doctorant en philologie numérique des textes en alphabet arabe

Référence : UMS 2000
Lieu de travail : Paris–Campus Condorcet
Type de contrat : CDD Scientifique
Durée du contrat *:* 6 mois éventuellement renouvelable une fois
Date d'embauche prévue : 1er juillet 2020
Quotité de travail : Temps complet
Rémunération : entre 3007,25 à 3615 € bruts mensuels
Niveau d'études souhaité : Doctorat
Expérience souhaitée : 1 à 4 années

Missions

La fouille et l'analyse automatiques de sources textuelles ont ouvert depuis quelques années des perspectives originales à la recherche historique ou littéraire, en rendant scientifiquement possibles des analyses statistiques quantificatives, ou qui combinent le quantitatif et le qualitatif, en matière de recherche d'information, de classification des textes, d'annotation sémantique ou d'extraction de renseignement. Ces avancées ont à ce jour en France marginalement concerné les textes en arabe, turc et persan, ou l'édition de corpus textuels outillés dans ces langues. Leur graphie non-latine soulève une difficulté particulière de lecture automatique que l'intelligence artificielle, par le biais de l'apprentissage-machine, est aujourd'hui en mesure de lever, pour peu que l'on s'en donne les moyens. L'enjeu est d'ordre à la fois technologique (mise à niveau internationale), scientifique (nouveaux fronts de la connaissance) et économique (maîtrise des coûts).

Le travail proposé s'inscrit dans le cadre des activités stratégiques du GIS Moyen-Orient et mondes musulmans (MOMM), et en particulier dans ses actions prioritaires soutenues par le plan SHS 2020 du MESRI. Il consiste à identifier les verrous qui restent à dépasser et les moyens d'y parvenir afin que la recherche française puisse se doter des moyens les plus adaptés pour l'océrisation des textes (en premier lieu imprimés, mais aussi manuscrits) en arabe, turc et persan et s'en saisisse à des fins de recherche. Le travail inclut une dimension expérimentale, soutenue par des moyens complémentaires en prestation informatique et encodage de textes. Il s'agira d'autre part de mettre sur pied avec les acteurs du champ (unités scientifiques ou universitaires spécialisées, services de formation continue) un programme pluriannuel de sessions d'études doctorales sur une semaine dédiées aux nouvelles philologies numériques moyen-orientales, africaines et asiatiques (cofinancé par le GIS MOMM).

Activités

- Produire un état des lieux sur les avancées récentes de l'océrisation des textes en alphabet arabe (imprimé et cursif) au plan européen et international, et les verrous qui restent à dépasser en vue de disposer d'un outil ouvert. Le livrable prendra la forme d'un rapport écrit.

- Répertorier les possibilités concrètes d'analyse ouvertes par l'océrisation à partir d'un exemple concret de corpus textuel : identification et extraction d'information, classification, analyse sémantique et génétique. Le livrable prendra la forme d'un démonstrateur réalisé avec un prestataire en informatique et doté d'un hébergement pérenne.
- Encoder en XML-TEI et outiller un corpus test à publier en ligne, choisi parmi la collection de manuscrits et imprimés arabes maghrébins conservés à la Bibliothèque des langues et civilisations (BULAC). Le livrable prendra la forme d'une publication en ligne réalisée avec un spécialiste de l'encodage.
- Concevoir et mettre en place un plan de formation national pluriannuel dans les nouvelles philologies numériques aréales (avec organisation d'une première initiative dès fin 2020).

Compétences

- La candidate ou le candidat doit être titulaire d'un doctorat en études arabes ou toute autre spécialité des sciences humaines dédiée à l'étude du Moyen-Orient et des mondes musulmans sur la base de sources en langue arabe (histoire, histoire de l'art et archéologie, science politique, islamologie, linguistique, études littéraires, sociologie, anthropologie).
- La candidate ou le candidat doit posséder une bonne connaissance du paysage français et européen des études sur le Moyen-Orient et les mondes musulmans.
- La candidate ou le candidat doit posséder une bonne familiarité avec les problématiques et les outils des humanités numériques
- La candidate ou le candidat doit être capable de communiquer efficacement de manière écrite et verbale, en français et en anglais : de nombreuses interactions à prévoir avec des scientifiques européens, dont la langue de travail est l'anglais.
- La candidate ou le candidat doit disposer de bonnes capacités rédactionnelles : la recherche doit aboutir entre autres à la rédaction d'un rapport à remettre au MESRI.
- Il est attendu également qu'elle ou il ait de bonnes capacités relationnelles : interactions nombreuses à prévoir avec les différents cercles concernés par la recherche : équipes du GIS MOMM, scientifiques étrangers, prestataires, interlocuteurs institutionnels.

Contexte de travail

La post-doctorante ou le post-doctorant disposera d'un bureau au Campus-Condorcet dans les locaux attribués au GIS MOMM, structure fédérative gérée par l'UMS 2000 (CNRS/EHESS). Le GIS MOMM rassemble actuellement 49 équipes dédiées pleinement ou partiellement à l'étude du Moyen-Orient et des mondes musulmans. Elle ou il travaillera en liaison étroite avec la direction du GIS MOMM (Eric Vallet, Elise Massicard, Mercedes Volait) et sa secrétaire générale (Cyrielle Michineau).