

FICHE DE POSTE

Libellé de Fonctions :

Ingénieur(e) de recherche en science des données/humanités numériques

Catégorie : A

Corps : IGR

BAP (le cas échéant) : D

Emploi-type (le cas échéant) : D1A41 - Ingénieur-e de recherche en production, traitement et analyse de données

Les activités qui composent la fiche de poste sont appelées à évoluer en fonction des connaissances du métier et des nécessités de service

Qui sommes-nous ?

Sorbonne Université est une université pluridisciplinaire et de recherche intensive. Poursuivant la tradition humaniste de la Sorbonne, elle s'attache à répondre aux enjeux scientifiques du 21^e siècle et à transmettre les connaissances issues de ses laboratoires et de ses équipes de recherche à ses étudiantes et étudiants et à la société tout entière.

Déployant ses formations auprès de 54 000 étudiantes et étudiants dont 4 700 doctorantes et doctorants et 10 200 étudiantes et étudiants étrangers, elle emploie 6 300 enseignantes et enseignants, enseignantes-chercheuses et enseignants-chercheurs, chercheuses et chercheurs et 4 900 personnels de bibliothèque, administratifs, technique, sociaux et de santé. Son budget est de 670 M€.

Sorbonne Université, principalement située au cœur de Paris, dispose d'un potentiel de premier plan et étend sa présence dans plus de vingt sites en Ile-de-France et en régions.

Présentation de la structure

Le CELLF :

Laboratoire de recherche en littérature, philologie et science des textes, allant du XVI^e au XXI^e siècle et incluant la francophonie, spécialisé notamment dans l'histoire littéraire et l'édition critique (papier et numérique), le Centre d'étude de la langue et des littératures françaises (UMR 8599 CELLF, CNRS - Sorbonne Université) compte 80 chercheurs et enseignants-chercheurs, 6 IT-BIATSS, et 150 à 200 doctorants. Des membres non-statutaires lui sont aussi rattachés en tant que « chercheurs associés ». Hébergé par Sorbonne Université, il est multi-site (rue Victor Cousin, rue Serpente). Le CELLF encourage les projets transdisciplinaires et trans-séculaires. Les humanités numériques sont au cœur de son programme scientifique.

L'ObTIC :

L'Observatoire des textes, des idées et des corpus (ObTIC) est une équipe-projet dédiée aux humanités numériques et rattachée administrativement au CELLF UMR 8599. Faisant suite à l'ancien LabEx OBVIL, l'ObTIC s'appuie sur une expertise acquise dans la production et l'édition des données, dans la conception et l'expérimentation des logiciels et des algorithmes pour les chercheurs en lettres et sciences humaines et sociales, et dans le développement de nouvelles méthodologies de recherche, fondées sur le numérique, au sein de ces disciplines.

Localisation : Maison de la Recherche, 28 rue Serpente, 75006 Paris ; SCAI, campus Jussieu, 4 place Jussieu, 75005 ; BnF DataLab, Bibliothèque nationale de France, Bibliothèque François-Mitterrand Quai François Mauriac, 75013.

Missions et activités principales

Mission :

Financé par le projet européen ERC ModERN, ce poste a pour objectif de renforcer le déploiement de la politique numérique de l'équipe du projet. Au sein du projet ModERN, la personne recrutée travaillera en étroite collaboration avec le PI et les autres membres de l'équipe. Son travail l'amènera aussi à collaborer avec divers services et partenaires de Sorbonne Université. La personne recrutée aura pour mission principale d'effectuer la gestion et l'enrichissement de la base de données du projet ModERN. En particulier, la personne aura une forte connaissance des méthodes d'analyse liées à la modélisation par graphe le « social network analysis ». Elle sera responsable du développement de la plateforme finale du projet qui donnera accès aux utilisateurs à la base de données ModERN. La plateforme mettra à disposition des chercheurs un moteur de recherche avec un système de filtre, des réseaux de co-présences textuelles interactifs, et d'autres outils de visualisation de la base de données en entier et des parties filtrées. Elle donnera également aux utilisateurs la possibilité d'effectuer des analyses différentes des données avec des méthodes numériques du traitement automatique de langues, aussi avec celles qui utilisent les LLMs (différents types de classification, analyses sémantiques, analyses thématiques, clusterisation etc.). La contribution de la personne recrutée représentera donc une intervention importante dans les domaines de la science des données et de la linguistique computationnelle.

La personne recrutée participera aussi de manière ponctuelle à la politique de traitement et d'exploitation des données contenues dans la base de données principale du projet ModERN. Elle fera des propositions en matière d'outils numériques pertinents et envisagera si besoin des développements spécifiques. Elle saura également accompagner techniquement l'équipe ModERN dans ses projets de recherche et de réaliser le développement et la maintenance de solutions venant des divers domaines telle la science des données, l'analyse des réseaux et l'intelligence artificielle. La personne recrutée jouera un rôle de formation vis-à-vis de l'ensemble des acteurs ModERN en matière des outils mis en place.

Cette mission sera décomposée de la manière suivante :

Activités principales :

- Développer une plateforme permettant un accès à la base de données du projet ModERN.
- Concevoir et développer une interface « user-friendly » pour la plateforme du projet ModERN.
- Développer des méthodes computationnelles et des outils d'analyse des données pour l'exploitation de la BDD, en s'y appuyant sur les analyses de réseaux et de graphes, ainsi que des analyses lexicales/sémantiques avec les LLMs.
- Proposer des solutions informatiques en matière d'outils numériques pour le traitement automatique des textes, la visualisation des réseaux et autres éléments techniques du projet.
- Accompagner le développement des outils numériques et de la production documentaire : environnement, données, outils et procédures.
- Participer à la production scientifique du projet (articles, white papers, etc.) ainsi que la rédaction des spécifications techniques pour les outils/méthodes développés.
- Participer à la vie scientifique de l'équipe : formations, séminaires et ateliers.

Conduite de projet : OUI

Encadrement : NON

Dans le cadre de vos fonctions, vous pourrez être amené à dispenser des formations internes en lien avec votre expertise métier.

Connaissances et compétences*

Savoir-faire transversaux :

Nous recherchons une personne de formation supérieure (BAC +5 min.), avec ce type de compétences :

- Connaissance démontrée de la science des données, du traitement automatique de langues et de l'intelligence artificielle appliquée.
- Maîtrise d'outils de visualisation des données, de l'analyse de réseaux et de graphes.
- Maîtrise d'un langage de script (Python) et des bases de données (SQL).

- Connaissance des standards patrimoniaux (XML, TEI, IIIF...), des langages d'interrogation et de transformation (Xpath, XSLT...), des normes d'interopérabilité, Bash, Git.
- Connaissance de l'apprentissage profond, des modèles de HuggingFace et expérience de travail avec des larges modèles de langue (apprentissage de transfert, fine-tuning) et avec des corpus littéraires sera un atout.

Connaissances transversales requises :

- Organisation et fonctionnement de la recherche et de l'enseignement supérieur.
- Organisation et fonctionnement de Sorbonne Université.
- Réglementation applicable à son domaine d'activité professionnelle.
- Statuts, règlements et procédures de gestion des personnels des établissements publics d'enseignement supérieur et de recherche.

Savoir-faire :

- Capacité à proposer et à élaborer des tâches en lien avec les projets en cours
- Capacité à définir et à implémenter des solutions informatiques agiles et sur mesure
- Capacité à interagir avec des interlocuteurs de différentes disciplines
- Capacité à hiérarchiser et à prioriser

Savoir être :

- Bon relationnel et sens du travail en équipe
- Disponibilité et réactivité
- Sens de l'analyse et force de proposition
- Adaptabilité et autonomie
- Organisation et rigueur
- Sens du service public
- Confidentialité

Exposition aux risques professionnels :

x Non

Oui : si oui, indiquer les informations relatives aux risques physiques (port de charge, machines dangereuses, vibrations...), biologiques, chimiques, rayonnements ionisants ou non ionisants. Si l'agent est exposé aux produits dangereux dont les CMR, il doit impérativement disposer d'une Fiche Individuelle d'Exposition téléchargeable sur intranet dans la rubrique « Prevention-des-risques-professionnels/fiche-individuelle-d-exposition-aux-agents-chimiques-dangereux ».

Conditions particulières d'exercice : néant.

Rémunération : Rémunération selon le barème de rémunération au recrutement des contractuels BIATSS de Sorbonne Université.

Prise de fonction : novembre 2024. **Contrat de 24 mois renouvelable.**

Documents requis :

1. Une lettre de motivation (en français et/ou en anglais) à l'attention de la responsable du projet (Glenn Roe) expliquant l'intérêt du/de la candidat(e) pour cette recherche et les compétences qu'il/elle envisage de mobiliser pour la mener à bien.
2. Un CV détaillé.
3. Une copie du document certifiant le niveau d'étude requis.
4. Les noms et adresses email de deux référents.

Critères de sélection :

1. Diplôme d'études supérieures ou doctorat en science des données, en informatique, en linguistique informatique ou dans un domaine connexe.
2. Compétences démontrées en linguistique computationnelle, machine/deep learning, modélisation des réseaux et programmation (Python).
3. Un intérêt pour la littérature française l'histoire intellectuelle.
4. D'excellentes capacités de communication.
5. Une bonne maîtrise du français et de l'anglais.
6. Un grand sens de l'organisation.

Procédure de sélection :

Les documents requis doivent être envoyés au responsable du projet Glenn Roe (glenn.roe@sorbonne-universite.fr).

Tous les documents doivent être soumis avant le 30 septembre 2024, au format PDF. Les candidatures incomplètes ou soumises après l'échéance ne seront pas prises en compte.

Les candidat(e)s retenu(e)s seront informé(e)s par email au plus tard le 15 octobre 2024 et invité(e)s à prendre part à des entretiens organisés à Paris (ou en distanciel) fin octobre 2024.

*Conformément à l'annexe de l'arrêté du 18 mars 2013 (NOR : MENH1305559A)

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council (ERC). Neither the European Union nor the granting authority can be held responsible for them.