



## Offre n°2024-08020

### Ingénieurs backend and bigdata

**Type de contrat :** CDD

**Contrat renouvelable :** Oui

**Niveau de diplôme exigé :** Bac + 5 ou équivalent

**Autre diplôme apprécié :** doctorat

**Fonction :** Ingénieur scientifique contractuel

### A propos du centre ou de la direction fonctionnelle

Le centre Inria de l'Université de Rennes est l'un des huit centres d'Inria et compte plus d'une trentaine d'équipes de recherche. Le centre Inria est un acteur majeur et reconnu dans le domaine des sciences numériques. Il est au cœur d'un riche écosystème de R&D et d'innovation : PME fortement innovantes, grands groupes industriels, pôles de compétitivité, acteurs de la recherche et de l'enseignement supérieur, laboratoires d'excellence, institut de recherche technologique

### Contexte et atouts du poste

#### Le projet CodeCommons

[Software Heritage](#) est une initiative française (Inria, soutenue par l'Unesco) pour archiver du code. Cette initiative a récolté du code open source publiquement accessible à partir de projets de développement logiciel ce qui a permis d'archiver 14 milliards de fichiers sources, 2 milliards de commits et plus de 200 millions de projets (Github, gitlab, etc.).

Le projet *Code Commons* prend appui sur Software Heritage pour positionner la France comme référence mondiale pour la base de données d'apprentissage sur le code. Pour cela, il va consolider et monter en puissance le bien commun numérique unique construit par Software Heritage depuis 2016, et construire l'infrastructure logicielle indispensable pour l'exploiter efficacement, tout en donnant un précieux avantage compétitif aux acteurs français de l'IA générative.

Le projet *Code Commons* se distingue par plusieurs innovations majeures :

- L'accélération de la collecte des codes sources, et l'élargissement du périmètre de Software Heritage permettront d'étendre et d'enrichir l'archive existante à un rythme sans précédent, incluant des tâches, commentaires, discussions, et métadonnées associées à des articles scientifiques, entre autres
- Un nouveau modèle de données unifié et une architecture scalable permettront de sélectionner aisément et d'extraire efficacement des sous-ensembles de données de l'archive, pour les adapter aux nouveaux besoins d'entraînement des modèles IA de nouvelle génération
- Une caractérisation poussée des fichiers sources, incluant la licence, les langages de programmation utilisés, des indicateurs de qualité du code (comme les motifs de conception et les CVE), et l'historique et les caractéristiques du projet (popularité, activités, dépendances).
- L'utilisation du [SWHID](#) (Software Heritage Identifiant), en cours de normalisation, qui offre une méthode unique et efficace pour la traçabilité, la transparence et la reproductibilité, facilitant ainsi l'identification des corpus d'apprentissage utilisés dans l'entraînement d'une IA.

Le projet s'appuie sur un partenariat stratégique entre [Inria](#), le [CEA](#), et [Tweag](#), réunissant des compétences complémentaires essentielles à la réussite de *Code Commons*. Inria apporte son expertise à travers les équipes de Software Heritage, DiverSE, Almanach, et

Cedar, offrant un large éventail de compétences en ingénierie des langages, génération de code, traitement des langues et IA générative, et analyse de données à grande échelle. Le CEA contribue avec son expertise en traitement automatique des langues et en ingénierie des systèmes et logiciels. Tweag, connu pour son approche innovante en matière de développement logiciel, complète ce consortium. [AboutCode](#) apportera son expertise pour la détection des licences des codes sources avec son logiciel Scancode, référence mondiale dans le domaine. Le projet bénéficie également du soutien de partenaires académiques internationaux, tels que les Universités de Pise et de Bologne, et de l'expertise de personnalités éminentes telles que Patrick Valduriez. Ce partenariat multidisciplinaire garantit une approche holistique et innovante, essentielle pour relever les défis complexes posés par l'IA générative.

## Mission confiée

### Vos tâches

Dans ce cadre, l'équipe DiverSE (en étroite collaboration avec l'équipe Software Heritage) recrute une équipe de **huit ingénieurs** sous la responsabilité scientifique et technique de permanents de l'équipe pour participer à ce projet. Dans ce cadre, l'équipe va travailler sur deux briques importantes concernant l'extraction efficace des données et des briques d'analyse de code efficace pour la construction de méta-données spécifiques. Concrètement, les deux premières tâches visent à reconstruire l'outil [GHTorrent](#) mais au-dessus de Software Heritage et reprendre l'ensemble des scripts d'entraînement de [starcoder](#) pour une intégration au-dessus de Software Heritage. Ces deux démonstrateurs serviront de cas nominal pour l'évaluation des l'ensemble des tâches effectuées dans ce projet par les partenaires. D'autres tâches d'analyse de code viendront compléter ces démonstrateurs, telles que la construction d'un graphe reliant les corrections des vulnérabilités logicielles dans le code avec les causes de leur apparition, entre autres.

## Principales activités

### Pourquoi nous rejoindre à INRIA Rennes chez DiverSE

Ce projet est unique par son ambition, son réseau de contacts, son impact potentiel. Il se retrouve au cœur des activités d'une équipe dynamique fortement intégré à l'équipe de Software Heritage.

#### Son ambition

Vous participerez à un projet open source d'envergure mondiale. Dans une époque où la maîtrise des données est un enjeu géopolitique stratégique pour les états, Code Commons inaugure au niveau national l'utilisation d'une archive unifiée fiable du code source mondiale. À vocation européenne, cette initiative s'intègre dans une ambition plus large visant à bénéficier au niveau européen d'un outil et d'une agence pour le pilotage de ces données associées au domaine de l'open source afin de garantir une souveraineté européenne dans le domaine de l'ingénierie logicielle, de l'IA, et de la cybersécurité (software supply chain attack; ...).

#### Son réseau de contacts

Vous serez au cœur d'un réseau d'utilisateurs dont le but est de faciliter l'adoption. Nous avons déjà assuré le soutien d'acteurs majeurs de l'IA en France, qui ont fourni des lettres d'engagement à collaborer: [Craft.ai](#), Hugging Face, Kyutai, LightOn, Mistral et Prairie.

#### Son impact potentiel

À l'heure où de nombreux projets open source deviennent méfiants du pillage par les

acteurs de l'IA de données/code qui n'ont pas été produits dans le but de servir de données d'apprentissage, reprendre la main au sein d'une initiative open source est un moyen de garantir une traçabilité de l'usage du code open source et ainsi permettre une confiance dans ces outils.

## En Bretagne au coeur d'une équipe jeune et dynamique

L'équipe de recherche DiverSE étudie les techniques de l'ingénierie logicielle pour la construction fiable et efficace d'applications. Notre expertise se place dans le domaine de l'ingénierie des langages, de la variabilité logicielle, du test, de l'architecture, etc. Avec une petite quinzaine de permanents (chercheurs Inria, CNRS, enseignants chercheurs INSA/Université de Rennes dont 3 IUFs), une quinzaine de doctorants et plusieurs ingénieurs, l'équipe est reconnue au niveau mondial dans ces domaines d'expertise. Elle est aussi reconnue pour son ambiance sur site, ses pauses café et ses séminaires aux verts mémorables. Nous avons en outre la chance d'héberger dans nos locaux deux ingénieurs de l'équipe de Software Heritage facilitant ainsi les liens entre les groupes.

Des connaissances en Python et Rust seront appréciées. D'une manière générale, nous nous attendons à recevoir des ingénieurs ayant la capacité de maîtriser plusieurs langages de développement.

## Compétences

Des connaissances en Python et Rust seront appréciées. D'une manière générale, nous nous attendons à recevoir des ingénieurs ayant la capacité de maîtriser plusieurs langages de développement.

## Avantages

- Prise en charge partielle des frais de transport en commun sur le trajet domicile-travail ou FMD
- Restauration subventionnée
- Prise en charge partielle des frais de mutuelle
- Possibilité de télétravail (à hauteur de 90 jours annuels) et d'aménagement du temps de travail

## Rémunération

Rémunération mensuelle brute de 2979 euros selon diplôme et expérience

## Informations générales

- **Thème/Domaine** : Programmation distribuée et génie logiciel  
Ingénierie logicielle (BAP E)
- **Ville** : Rennes
- **Centre Inria** : [Centre Inria de l'Université de Rennes](#)
- **Date de prise de fonction souhaitée** : 2024-10-01
- **Durée de contrat** : 2 ans
- **Date limite pour postuler** : 2024-09-21

## Contacts

- **Équipe Inria** : [DIVERSE](#)
- **Recruteur** :  
Barais Olivier / [Olivier.Barais@irisa.fr](mailto:Olivier.Barais@irisa.fr)

## A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut

fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

## L'essentiel pour réussir

- être vraiment enthousiasmé par notre projet
- être persévérant (se relever et continuer lorsque les choses ne se déroulent pas comme prévu - la vraie recherche se déroule rarement comme prévu)
- ne pas avoir peur (par exemple, ne pas hésiter à hacker une machine virtuelle, un compilateur, un noyau ou à implémenter un algorithme complexe)
- avoir une attitude de petit enfant (vouloir comprendre et apprendre tout ce qu'il rencontre)
- avoir une attitude d'ingénieur (ne pas prendre la première solution qui vient à l'esprit, mais examiner les principales alternatives)
- avoir une attitude de chercheur (vouloir vraiment comprendre quelque chose et ne pas se contenter de la première meilleure explication)
- vouloir examiner ce qui est simple et évident avant d'explorer ce qui est compliqué
- être capable de se concentrer (ignorer les nombreuses autres choses intéressantes que l'on pourrait faire)
- éprouver du plaisir à trouver un argument ou une explication logique et claire
- aimer lire (des livres, des documents, des documents, des documents)
- aimer convaincre les autres à l'aide d'arguments solides
- être d'accord pour travailler dur
- être heureux de rester en Bretagne pendant un certain temps
- être d'accord pour voyager de temps en temps sur de longues distances (par exemple, pour des conférences)

**Attention:** Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

## Consignes pour postuler

Merci de déposer en ligne CV, lettre de motivation et éventuelles recommandations

### **Sécurité défense :**

Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

### **Politique de recrutement :**

Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.