Offer #2024-08020

Backend and bigdata engineers

Contract type: Fixed-term contract
Renewable contract: Yes
Level of qualifications required: Graduate degree or equivalent
Other valued qualifications: PhD
Fonction: Temporary scientific engineer

About the research centre or Inria department

The Inria Centre at Rennes University is one of Inria's eight centres and has more than thirty research teams. The Inria Centre is a major and recognized player in the field of digital sciences. It is at the heart of a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc.

Context

The CodeCommons project

Software Heritage is a French initiative (Inria, supported by Unesco) to archive open-source code. This initiative has collected publicly accessible open-source code from software development projects, resulting in the archiving of 14 billion source files, 2 billion commits and over 200 million projects (Github, gitlab, etc.).

The Code Commons project is building on Software Heritage to position France as a global reference for the code learning database. To do this, it will consolidate and scale up the unique digital commons built by Software Heritage since 2016, and build the software infrastructure needed to exploit it effectively, while giving French players in generative AI a valuable competitive edge.

The Code Commons project aims to achieve several major innovations:

- Accelerating the collection of source code, and broadening the scope of Software Heritage will enable the existing archive to be extended and enriched at an unprecedented rate, including tasks, comments, discussions, and metadata associated with scientific articles, among other things.
- A new unified data model and scalable architecture will make it easy to select and efficiently extract subsets of data from the archive, adapting them to the new training needs of next-generation AI models.
- Extensive characterisation of source files, including licence, programming languages used, code quality indicators (such as design patterns and CVEs), and project history and characteristics (popularity, activities, dependencies).
- The use of the SWHID (Software Heritage Identifier), currently being standardised, which offers a unique and effective method for traceability, transparency and reproducibility, making it easier to identify the learning corpora used in training an AI.

The project is based on a strategic partnership between Inria, CEA, and Tweag, bringing together complementary skills that are essential to the success of Code Commons. Inria is contributing its expertise through the Software Heritage, DiverSE, Almanach, and Cedar teams, offering a wide range of skills in language engineering, code generation, language processing and generative AI, and large-scale data analysis. The CEA is contributing its expertise in automatic language processing and systems and software engineering. Tweag, known for its innovative approach to software development, completes the consortium. AboutCode will be contributing its expertise in source code licence detection with its Scancode software, a world reference in the field. The project also benefits from the support of international academic partners, such as the Universities of Pisa and Bologna, and the expertise of eminent figures such as Patrick Valduriez. This multidisciplinary partnership guarantees a holistic and innovative approach, which is essential for tackling the complex challenges posed by generative AI.

Assignment

Your tasks

The DiverSE team (in close collaboration with the Software Heritage team) is recruiting a team of eight engineers under the scientific and technical responsibility of permanent members of the team to take part in this project. As part of this project, the team will be working on two important building blocks:
efficient data extraction and efficient code analysis building blocks for the construction of specific metadata. In concrete terms, the first two tasks aim to rebuild the GHTorrent tool but on top of Software Heritage and to take over all the starcoder training scripts for integration on top of Software Heritage. These two demonstrators will serve as a nominal case for the evaluation of all the tasks carried out in this project by the partners. Other code analysis tasks will complement these demonstrators, such as the construction of a graph linking the corrections of software vulnerabilities in the code with the causes of their appearance, among others.

Main activities

Why join INRIA Rennes at DiverSE

This project is unique in its ambition, its network of contacts and its potential impact. It is at the heart of the activities of a dynamic team that is closely integrated with the Software Heritage team.

Its ambition

You will be taking part in an open source project on a global scale. At a time when control over data is a strategic geopolitical issue for states, Code Commons is inaugurating the use of a reliable unified archive of source code at national level. With a European vocation, this initiative is part of a wider ambition to benefit at European level from a tool and an agency to manage this data associated with the open source domain in order to guarantee European sovereignty in the field of software engineering, AI and cybersecurity (software supply chain attack, etc.).

Its network

You will be at the heart of a network of users whose aim is to facilitate adoption. We have already secured the support of major AI players in France, who have provided letters of commitment to collaborate: Craft.ai, Hugging Face, Kyutai, LightOn, Mistral and Prairie.

Its potential impact

At a time when many open source projects are becoming wary of AI players plundering data/code that has not been produced with the intention of being used as learning data, regaining control within an open source initiative is a way of guaranteeing traceability of the use of open source code and thus enabling confidence in these tools.

In Brittany at the heart of a young, dynamic team

The DiverSE research team studies software engineering techniques for the reliable and efficient construction of applications. Our expertise lies in the fields of language engineering, software variability, testing, architecture, etc. With around fifteen permanent staff (Inria and CNRS researchers, INSA/Université de Rennes lecturers, including 3 IUFs), around fifteen PhD students and several engineers, the team is recognised worldwide in these areas of expertise. It is also renowned for its on-site atmosphere, its coffee breaks and its seminars with memorable greens. We’re also lucky enough to be able to host two engineers from the Software Heritage team on our premises, facilitating links between the groups.

Required skills: Knowledge of Python and Rust will be appreciated. Generally speaking, we expect to welcome engineers with the ability to master several development languages.

Skills

Knowledge of Python and Rust will be appreciated. Generally speaking, we expect to welcome engineers with the ability to master several development languages.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Possibility of teleworking (90 days per year) and flexible organization of working hours
- Partial payment of insurance costs

Remuneration

Monthly gross salary from 2 979 euros according to diploma and experience

General Information

- Theme/Domain: Distributed programming and Software engineering
- Software engineering (BAP E)
- Town/city: Rennes
Inria Center: Centre Inria de l'Université de Rennes
Starting date: 2024-10-01
Duration of contract: 2 years
Deadline to apply: 2024-09-21

Contacts
- Inria Team: DIVERSE
- Recruiter: Barais Olivier / Olivier.Barais@irisa.fr

About Inria
Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

The keys to success
- be really excited about our project
- be persistent (get back up and continue when things don't work out as planned -- true research rarely works out as planned)
- be fearless (e.g., be ok hacking a virtual machine, a compiler, a kernel, or implementing a complex algorithm)
- have a small child's attitude (to want to understand and learn about everything they encounter)
- have an engineer's attitude (not to take the first solution that comes to mind, but to look at the key alternatives)
- have a researcher's attitude (to want to truly understand something, and to not be satisfied with the first best explanation)
- want to look at the simple and obvious before exploring the complicated
- be able to focus (to ignore the many other cool things one could also do)
- derive pleasure from coming up with a logical and clear argument or explanation
- like to read (books, papers, papers, papers)
- like to convince others using sound arguments
- be ok working hard
- be happy staying in Brittany for quite some time
- be ok traveling long distance from time to time (e.g., for conferences)

Warning: you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply
Please submit online: your resume, cover letter and letters of recommendation eventually

Defence Security: This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST). Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy: As part of its diversity policy, all Inria positions are accessible to people with disabilities.