



**Offer #2025-08795**

## **Ingénieur OCR et Analyse de mise en page**

*The offer description below is in French*

**Contract type :** Fixed-term contract

**Level of qualifications required :** PhD or equivalent

**Fonction :** Temporary scientific engineer

**Level of experience :** Recently graduated

### **Context**

Ce poste d'ingénieur IA est ouvert au sein de l'équipe-projet ALMAnaCH du Centre Inria de Paris. ALMAnaCH est une équipe de recherche d'une cinquantaine de membres, dont 7 membres permanents, spécialisée dans le traitement automatique des langues (TAL) et les humanités numériques (traitement des sources historiques et littéraires par l'informatique). Elle fait partie des 230 équipes-projets d'Inria, l'institut national de recherche en informatique et en automatique, établissement public de recherche regroupant 9 centres de recherche, dont le Centre de Paris auquel appartient ALMAnaCH.

Le poste prend place dans le cadre du projet européen [ATRIUM](#) (Advancing frontier Research In the arts and hUMANities), qui vise à relier les principales infrastructures de recherche dans les arts et les humanités (DARIAH), l'archéologie (ARIADNE), les langues (CLARIN) et la communication savante ouverte en sciences humaines et sociales (OPERAS). ATRIUM répond aux défis posés par la diversité des disciplines en sciences humaines et sociales, en proposant des services interopérables adaptés à des communautés aux méthodologies variées.

Dans ce contexte, le travail portera plus particulièrement sur la segmentation des documents complexes (formulaire) et la gestion des caractères rares dans le cadre de la reconnaissance de texte en ensemble semi-ouvert.

L'employé travaillera en collaboration directe avec Thibault Clérice, tout en interagissant avec différents membres de l'équipe impliqués dans la conception d'interfaces utilisateur et d'expériences utilisateur (UI/UX), la définition des lignes directrices pour la segmentation des documents, ainsi que d'autres aspects liés à la reconnaissance optique de caractères (OCR).

## Assignment

### Missions :

Sous la responsabilité de Thibault Clérice, la personne recrutée aura pour mission d'améliorer les capacités de moteurs de segmentation et de reconnaissance automatique de texte (ATR) dans le cadre de leur utilisation dans les plate-formes eScriptorium et dérivées. En particulier, sont concernés:

- la reconnaissance de texte en ensemble semi-ouvert pour les documents archéologiques;
- la reconnaissance en segmentation de formulaire;
- le few-shot learning de segmentation, pour des documents répétitifs.

### Collaboration :

La personne recrutée sera en lien avec l'ingénieur de BACK IN TIME pour l'intégration des systèmes dans l'interface eScriptorium, avec une autre ingénieure Inria d'ATRIUM pour l'implémentation des workflows designés dans le cadre du WP4 et de la mise en place des demonstrators du WP5. Plus largement, des collaborations internes avec l'ingénieur OCR du PIQ CLLG et les ingénieurs du projet COLaF sont attendues.

Au national et à l'international, des collaborations avec le groupe responsable d'eScriptorium sont attendues, dont des réunions hebdomadaires, ainsi qu'avec le reste du projet ATRIUM.

## Main activities

### Principales activités:

- Développer la reconnaissance de texte en ensemble semi-ouvert pour les documents archéologiques et l'intégrer à des moteurs compatibles avec eScriptorium ou ses dérivés;
- Développer la reconnaissance en segmentation de formulaire et l'intégrer à des moteurs compatibles avec eScriptorium ou ses dérivés;
- Développer le few-shot learning de segmentation, pour des documents répétitifs et l'intégrer à des moteurs compatibles avec eScriptorium ou ses dérivés;
- Participer aux réunions autour des plate-formes eScriptorium et ses dérivés dont Inria est membre ainsi qu'aux réunions du groupe ATRIUM;

- Assurer une documentation des fonctionnalités produites.

#### **Activités complémentaires:**

- Aider à la création de recommandation pour la segmentation de documents modernes liés à l'archéologie;
- Entraîner des modèles adaptés aux besoins du projet et les mettre à disposition;
- Aider et maintenir les moteurs concernés dans eScriptorium et ses dérivés.

**Missions collectives :** participation à la vie de l'équipe et des projets, y compris pour des conférences ou rencontres internationales à l'étranger.

## **Skills**

#### **Compétences techniques et niveau requis :**

- Maîtrise de PyTorch et PyTorch lightning;
- Intérêt pour les sciences humaines et leurs particularités;
- Maîtrise de Git;
- Connaissance des standards de la reconnaissance de texte (ALTO/PageXML).

#### **Langues :**

- Anglais B2 minimum
- Français B2 souhaité
- Une autre langue est appréciée, en particulier dans le cadre européen d'ATRIUM.

#### **Compétences relationnelles :**

- Bonnes capacités d'organisation.
- Bon relationnel.

Compétences additionnelles appréciées :

## **Benefits package**

- Restauration subventionnée
- Transports publics remboursés partiellement
- Congés: 7 semaines de congés annuels + 10 jours de RTT (base temps plein) + possibilité d'autorisations d'absence exceptionnelle (ex : enfants malades, déménagement)
- Possibilité de télétravail et aménagement du temps de travail
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.)

- Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria)
- Accès à la formation professionnelle
- Sécurité sociale

## General Information

- **Theme/Domain** : Language, Speech and Audio  
Data production, processing, analysis (BAP D)
- **Town/city** : Paris
- **Inria Center** : [Centre Inria de Paris](#)
- **Starting date** : 2025-07-01
- **Duration of contract** : 2 years
- **Deadline to apply** : 2025-05-04

## Contacts

- **Inria Team** : [ALMANACH](#)
- **Recruiter** :  
Clerice Thibault / [thibault.clerice@inria.fr](mailto:thibault.clerice@inria.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

- Se sentir à l'aise dans un environnement interdisciplinaire, aimer apprendre et écouter sont des qualités essentielles pour réussir cette mission.
- Intéressé par les problématiques de sciences ouvertes et la reproductibilité des sciences.
- Une thèse dans le domaine de l'OCR ou de l'analyse de la mise en page constitue un réel atout.

**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**

### **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.