2019-01374 - PhD Position F/M [Sub - 2019] - Interactive tools for the transcription of handwritten documents

**Type de contrat** : CDD de la fonction publique  
**Niveau de diplôme exigé** : Bac + 5 ou équivalent  
**Fonction** : Doctorant

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

The Inria Lille - Nord Europe Research Centre was founded in 2008 and employs a staff of 360, including 300 scientists working in sixteen research teams. Recognised for its outstanding contribution to the socio-economic development of the Nord - Pas-de-Calais Region, the Inria Lille - Nord Europe Research Centre undertakes research in the field of computer science in collaboration with a range of academic, institutional and industrial partners.

The strategy of the Centre is to develop an internationally renowned centre of excellence with a significant impact on the City of Lille and its surrounding area. It works to achieve this by pursuing a range of ambitious research projects in such fields of computer science as the intelligence of data and adaptive software systems. Building on the synergies between research and industry, Inria is a major contributor to skills and technology transfer in the field of computer science.

**Contexte et atouts du poste**

Public and private archives contain hundreds of precious ancient documents for historic and genealogical research: Civil records, church records, military records, population censuses. These archives are regularly scanned for preservation and sharing. However accessing information is tedious when the corpus is not indexed or not transcribed. Several project aim at indexing and transcribing such documents automatically, such as Transkribus [6] or Himanis [5]. Despite their efforts, transcribing documents automatically remain complicated and unsatisfactory for historians [4]. Automatic methods are not able to interpret complex writings (irregular, with overlaps, bad scan quality) [3]. Moreover these algorithms require hand-made learning database and machine learning experts to adapt the tool to a specific document corpus.

The candidate’s work will consist in designing, implementing and evaluating interactive tools for helping manual transcription of scanned handwritten documents. For this purpose, the candidate will study the historians and genealogists activities and workflow, in order to frame their needs and methods; search for existing manual and automatic methods for transcribing handwritten documents; and analyze problems that can be solved by humans, machines, and how to effectively combine these methods.

The first task on this thesis will be to continue previous work on advanced selection techniques for handwritten text[1]. In this study we designed a pixel selection tool similar to the magic wand [2]. The problem of regular magic wands is that it propagates to the whole adjacent pixels, which leads to incorrect selections in our case. We solve this problem with a combination of a magic wand and a brush. This work is in progress, and requires tuning, as well as a user evaluation. Later work will focus on annotation techniques that support collaborative work, as well as studying the combination of interactive and automatic approaches. This tool will be invaluable for bootstrapping the transcription of large corpuses, as well as helping transcribing small corpuses.

This thesis is part of a larger project in which we would like to combine interactive and automatic methods to get the best of both approaches. The candidate will collaborate with historians, computer scientists specialized in document analysis, and Geneanet, a leading genealogy website with thousands of users and millions of documents.

**References**


**Mission confiée**

The thesis will be carried out in the Loki team in Lille, France, joint between Inria – Lille Nord Europe and the CRIStAL (UMR CNRS 9189) laboratory of the University of Lille.

**Principales activités**

The candidate’s work will consist in designing, implementing and evaluating interactive tools for helping manual transcription of scanned handwritten documents. For this purpose, the candidate will study the historians and genealogists activities and workflow, in order to frame their needs and methods; search for existing manual and automatic methods for transcribing handwritten documents; and analyze problems that can be solved by humans, machines, and how to effectively combine these methods.

**Avantages**

- Subsidized meals
• Partial reimbursement of public transport costs
• Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours)
• Access to vocational training
• Social security coverage

Rémunération
1st and 2nd year: 1593.50€ Net monthly salary (after taxes)
3rd year: 1676.31€ Net monthly salary (after taxes)

Informations générales
• Ville: Villeneuve d'Ascq
• Centre Inria: CRI Lille - Nord Europe
• Date de prise de fonction souhaitée: 2019-10-01
• Durée de contrat: 3 ans
• Date limite pour postuler: 2019-04-22

Contacts
• Equipe Inria: LOKI AE
• Directeur de thèse: Huot Stéphane / stephane.huot@inria.fr

A propos d'Inria
Inria, l'institut national de recherche dédié aux sciences du numérique, promeut l'excellence scientifique et le transfert pour avoir le plus grand impact. Il emploie 2400 personnes. Ses 200 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3000 scientifiques pour relever les défis des sciences informatiques et mathématiques, souvent à l'interface d'autres disciplines. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 160 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
A successful candidate must hold a MSc in computer science or equivalent, and show a great interest in performing high quality research in Human-Computer Interaction. Knowledge in OCR and machine learning is helpful. He or she must speak and write English fluently, and experience or strong interest in software development. Creativity, independence, team working and communication skills are valuable advantages.

Consignes pour postuler
CV, application letter, one or more letters of recommendation and last school transcripts

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.

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.