Job offer

Engineer R&D: Computer Vision - Machine Learning

Centre d'études supérieures de la Renaissance | Laboratoire d'Informatique de Tours

ANR TypoReF : *Typographie de la Renaissance Française (1470-1640)* (*ANR-22-CE27-0009*)

Informations générales

- Location : LIFAT, Tours, France
- Publication date of the offer: March, 2023
- Type of contract: Fixed-term contract, technical and administrative support
- Sector of activity: research & development, digital humanities
- Contract period: 12 months
- Expected hiring date: September 1st, 2023
- Quota of work: full time
- Grade: Engineer or research engineer (Postdoc)
- Remuneration: according to previous professional experience, based on public salary levels (French public service)
- Education level: High level degree in Computer Science or PhD
- Deadline for application: May 22, 2023

Context

This position is part of the ANR TypoReF project (*ANR-22-CE27-0009*) directed by Remi Jimenes (CESR Tours). Conceived in an interdisciplinary approach, the project aims at inventorying, describing and studying French typographic materials of the Renaissance by developing tools for the analysis and exploration of digitized book corpora and by setting up a dedicated digital platform.

The digitization of old books began in the mid-1990s. Since then, teams of researchers combining computer scientists and specialists in written heritage have collaborated to develop tools dedicated to the exploitation of these digital corpora. For many years, document image analysis and recognition were considered complex tasks decomposed into two independent sub-tasks: layout analysis and text recognition. Layout analysis aims at detecting all image component such as text blocks, tables, images, graphics, signatures... Text recognition is devoted to the recognition of characters inside text blocks (paragraphs).

In the framework of a collaboration established in 2003, the CESR and the LIFAT have developed the Agora and Retro software. The developments were extended in 2011 and 2012 by two successive "Digital Humanities" grants funded by Google. These layout analysis and indexing tools adapted to the images of printed books of the Renaissance have made it possible to automatically extract and index some 13,500 engravings, lettering, headbands and fleurons from 600 facsimiles of the « Virtual Humanistic Libraries » research project (BVH).

Agora and Retro are already more than 15 years old and based on technologies that are now outdated given the developments of recent years. Recently, deep architectures have improved the state-of-the-art performance for both sub-tasks. Based on the very strong experience accumulated during the last twenty years, the TypoReF project intends to allow the development of new tools for the analysis of layout, identification and indexing of typographic materials adapted to the ancient, printed book, using the most powerful technologies of the moment, in particular those based on deep learning.

Missions to achieve

The goal is **the implementation of a web-based content labeling platform**. This labeling concerns both the collection of metadata produced by an expert user and the production of metadata resulting from an analysis of the pixels, their layout, and the relationships between the extracted elements of content. This double indexing (human/machine) must allow to search, compare, and group shapes, and thus to highlight links that would otherwise be difficult to perceive.

The platform requires developing the following functionalities:

- **Semantic segmentation of document images using deep learning architectures** (to replace old Agora). It consists of the localization and characterization of specific elements of content inside the digitized pages of an old printed book (text blocks, engraved decoration, etc.) with different levels of detail (within a text block, we will thus distinguish lines and characters). Characterization of these elements is also needed: lettering, bands, marks, characters, etc. A set of associated metadata allows one to characterize and retrieve precisely the location of each component.
- **Clustering of the extracted elements of content** (to replace old Retro) consists in making comparisons between elements, mainly engraved ornaments (such as lettering or headbands), to propose matches between similar forms. The platform will use different types of **machine learning algorithms, including the so-called "unsupervised" learning.**
- **Import/export data from/to the internal database according to standards such as ALTO or IIIF.**

Skills

- Software programing and software engineering in Python and web (HTML, CSS, and JS)
- Machine Learning, Computer vision
- Ability to work in a team, curious and rigorous spirit

Conditions of recruitment

- **Management structure and recruitment:** University of Tours, CESR, UMR 7323 - 59, rue Néricault Destouches BP 12050, 37020 Tours Cedex 1 - https://cesr.univ-tours.fr/
- **Workplace**: Laboratoire LIFAT, 64 avenue Jean Portalis 37200 Tours - http://lifat.univ-tours.fr

How to apply

- **Deadline for application:** May 22, 2023
- **Send CV and motivation letter to:** Thierry BROUARD, Jean-Yves RAMEL --> prenom.nom AT univ-tours.fr