



EA 4108  
www.litislab.eu

FR CNRS 3638  
www.normastic.fr



## *Engineer / PostDoc position*

### *Information extraction, Text Recognition in Historical Document Collections*

#### **LITIS**

LITIS (Laboratoire d'Informatique, Traitement de l'information et des Systèmes) is a research laboratory associated to the University of Rouen Normandie, Le Havre Normandie Normandie, and School of Engineering INSA Rouen Normandie. Research at LITIS is organized around 7 research teams which contribute to 3 main application domains: Access to Information, Biomedical Information Processing, Ambient Intelligence. LITIS currently includes 90 faculty staff members, 50 PhD students, 10 PostDoc and Research Engineers. The Machine Learning team of LITIS is developing research in modeling unstructured data (signals, images, text, etc...) with machine learning algorithms and statistical models. For more than two decades it contributes to the development of reading systems and document image analysis for various applications such as postal automation, business document exchange, digital libraries, etc...

#### **EURHISFIRM project**

EURHISFIRM aims at developing a research infrastructure to connect, collect, collate, align, and share reliable long-run company-level data for Europe to enable researchers, policymakers and other stakeholders to analyze, develop, and evaluate effective strategies to promote investment and economic growth. To achieve this goal, EURHISFIRM develops innovative tools to spark a "Big data" revolution in the historical social sciences and to open access to cultural heritage

EURHISFIRM is a project funded by the European Commission within the Infrastructure Development Program of Horizon 2020. The goal of the Program is to develop world-class research infrastructures lasting for decades ([https://ec.europa.eu/research/infrastructures/index\\_en.cfm?pg=home](https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=home) ). Research infrastructures are facilities, resources and services used by the science community to foster innovation and extend the frontiers of knowledge.

The first phase of the Infrastructure Development Program lasts for three years. It aims at developing an in-depth design study of the Research Infrastructure. After this phase, Development and Consolidation Phases follow if further applications will be successful. EURHISFIRM brings together eleven research institutions in economics, history, information technologies and data science from seven European countries.

#### **Position to be filled**

Position: Post-Doctoral fellow

Time commitment: Full-time

Duration of the contract: April 1<sup>st</sup> 2018 – October 2019, (renewable contract until March of 2021)

Contact: Prof. Thierry Paquet, [Thierry.Paquet@univ-rouen.fr](mailto:Thierry.Paquet@univ-rouen.fr)

Indicative salary: €36 000 gross annual salary, with social security benefits

Location: LITIS, Campus du Madrillet, Faculty of science, Saint Etienne du Rouvray, France



## Missions

Within the project, you will be in charge of developing text information recognition technologies (ICR) from historical document images (mostly printed), and information extraction from these data (such as person names, names of companies, dates, positions, stock prices etc...). The datasets are made of financial yearbooks and price lists of European companies, in different European languages. Your mission includes

- 1- the development of a machine learning based reading system of text lines composed of both deep optical models, and language models (statistical, and grammar based). Layout analysis falls out of the scope of the mission.
- 2- Data preparation for evaluation purposes
- 3- Benchmarking with other technologies (commercial products)
- 4- Integration of the system as a web service allowing its integration and deployment into a full system
- 5- Coordination with partners of the project regarding datasets preparation and collation of datasets, as well as software interoperability with other developments within the EurHisFirm consortium.

## Requirements :

The successful applicant should have a strong record in statistical machine learning and have experience in one popular platform and programming language in the field, so as to design, develop and make the prototype evolve.

- PhD, or Computer Engineer, with a good record in Machine Learning
- demonstrates ability to work in a team, curious and rigorous spirit
- Excellent written and verbal communication skills (French and English)

## Technical skills :

C/C++, Python, Tensor Flow, Keras, and other librairies (Numpy, OpenCV, Kaldi ..), knowledge about web technologies