

Call for applications

THE UCLouvain UNIVERSITY

&

STATE ARCHIVES OF BELGIUM

recruit

A specialist in **computer science** or **digital humanities** m/f/x

1 open-ended FTE: 50% PostDoc (UCLouvain) and 50% project manager (AGR)

with a particular interest in the automatic processing of
digital documents involving both text and images

for the FED-tWIN programme - Prf-2022-025 - ARKEY

#NLP #ComputerVision #HTR #MetadataExtraction #UX #InfoVis

Context

FED-tWIN is a research program of the Belgian Federal Science Policy Office (BELSPO) aimed at promoting sustainable cooperation between federal scientific establishments and Belgian universities, by funding common research profiles.

Today, the FED-tWIN program makes it possible to open a function for the ARKEY project (“A content-enriched user-oriented access key to digital archives”). It involves on the one hand the National Archives and State Archives in the Provinces (in short the State Archives, or AGR), a federal scientific institution, and on the other hand the Université catholique de Louvain (UCLouvain).

Goals

The main objective of the ARKEY research profile is to improve the digital valorisation of archive collections through long-term tools. It involves (1) the research and development of an enhanced access key to digitized content, and (2) the improvement of the navigation experience within archive collections. It builds on the expertise of a multidisciplinary team from AGR and from several research groups within UCLouvain (see “Partners” section below). ARKEY aims to bring added value for society and public service, by improving the accessibility and intelligibility of archives: a priority for many researchers and a foundation of democratic states.

Currently, AGR and UCLouvain own a large number of digitized documents from a wide variety of sources from different periods. This diversity offers a challenge to automated content analysis, particularly to optical character recognition (OCR) tools, which are not trained to such variation. Archives are also challenged by storage, format, metadata, and navigation of digitized documents: most of these documents are not sufficiently spotlighted. To respond to these challenges, ARKEY proposes a 3-step plan:

1. **AI-aided text and layout recognition.** ARKEY will develop and evaluate semi-automated content-analysis machine-learning techniques, specifically designed for handwritten documents and early printed books. They will rely on state-of-the-art OCR and Handwritten Text Recognition (HTR) methods, and focus on information extraction based on robust layout analysis.
2. **Content-enriched digital archival representation.** The data extracted from the content analysis will be used to enrich the representation of archive documents. This second challenge therefore aims to investigate and improve state-of-the-art Natural Language Processing methods to enrich Encoded Archival Description (EAD) files with automatically generated metadata based on semantic modeling, named entity recognition, and query expansion.
3. **User-oriented and context-aware navigation.** The third challenge of ARKEY is to allow archive users to effortlessly benefit from the content-enriched archive description described in the previous sections and, more generally, to improve their navigation experience within the archives. It implies the implementation of a user-oriented design method to elaborate efficient finding aids and visualization tools. In particular, we will contribute to alleviate the 2 following issues: (1) the lack of understanding of the available archival representations and the way they relate to each other, and (2) the difficulty of translating an initial question into a specific search and navigation scenario.

Partners

The **National Archives and State Archives in the Provinces** (www.arch.be) are a federal scientific institution that is part of BELSPO (Belgian Federal Science Policy Office). The institution is composed of the National Archives in Brussels, 19 State Archives repositories throughout the country, and the Center for Studies and Documentation War and Contemporary Societies (CegeSoma). The State Archives obtain and preserve (following appraisal) archive documents that are at least 30 years old from courts, tribunals, public authorities, notaries and from the private sector and private individuals. It ensures that public archives are transferred according to strict archival standards. Making these archival documents available to the public, while respecting the protection of privacy, is one of the primary missions of the institution. By means of its 19 reading rooms, the State Archives provides appropriate infrastructure to a wide public. Maintaining a direct service to the public via the Internet is another of the institution's key priorities.

The **Université catholique de Louvain** (www.uclouvain.be) is a Teaching and Research University of the Wallonia-Brussels Federation in Belgium. It has nearly 35,000 students and 3,000 researchers. Within it, the MiiL (Media innovation and intelligibility Lab) is an interdisciplinary platform for innovation in digital production and appropriation. Anchored within the "Language and Communication" Research Institute (ILC), this platform brings together information and communication experts, linguists, computer engineers, lawyers and economists, and offers adapted solutions to issues relating to the intelligibility of digital data and their standardized representation,

user experience (UX), and interactive communication. In the context of this project, the Researcher will work within the MiiL, in close collaboration with several other research teams and departments of the University, whose joint expertise covers the 3 challenges described above. First, the CENTAL, center for automatic language processing, which brings together many researchers specialized in the computer processing of textual data. Second, the UCLouvain Archives Service, which keeps the definitive archives of the university and oversees the proper management of all the documents produced or acquired by the university. Finally, the GEMCA, Group for Early Modern Cultural Analysis, which is an interdisciplinary research center bringing together researchers in literature, art history, history, and philosophy.

Work duties

The Researcher will directly work towards the achievement of the 3 project goals described above. Within the framework of the FED-tWIN program, he or she will conduct research, and develop methods, expertise and collaborations in order to achieve these goals. The work will involve in particular

- high-level scientific production and publication of the research results through the appropriate channels;
- a strong commitment to scientific communication, both oral and written, for a wide target audience. Given the social role of AGR, in addition to academic production, vulgarized publications are also expected, as well as organization of general public events;
- applications for projects allowing the development of interdisciplinary and international scientific research, and the promotion of cooperation between UCLouvain and AGR;
- an educational activity allowing the valorization of research achievements in teaching activities at UCLouvain.

Diploma

PhD degree in Computer Science , Digital Archives, Digital Humanities, Natural Language Processing, or another field directly relevant to the position.

The PhD must have been obtained a maximum of 12 years before the deadline for submitting applications¹.

Technical skills

The profile requires solid technical skills relating to the automatic processing of digital documents containing both text and images.

In particular, the candidate must be able to demonstrate:

¹The period covered is extended by one year per period of maternity leave of at least three months, or per uninterrupted period of at least three months of full-time parental leave or full-time adoption leave, taken by the candidate between obtaining the doctorate title and the final date for submitting application, without the total extension being able to exceed one year per child. The period referred to in the first paragraph is also extended by the actual duration of the certified periods of long-term illness of the candidate or of a close family member of the candidate to whom the latter has given medical treatment, insofar as it concerns uninterrupted periods of at least three months.

- recognized scientific expertise in the **processing of digitized documents**. This includes image processing (*pattern recognition* and *layout analysis*), optical character recognition (OCR) and automatic language processing (NLP),
- specific knowledge of **machine learning** techniques applied to the fields mentioned above (such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), transformers-type models),
- **programming** skills in Python or a similar programming language (experience with PyTorch or another machine learning library is required),
- at least a basic expertise in **user experience** (UX) and user-oriented design, and the flexibility to deepen these skills.

Soft skills

The candidate will have:

- project management skills to ensure the proper conduct of research, ensure coordination between work packages, and stimulate collaboration with colleagues from both organizations if necessary,
- an ability to seek additional funding and to build and manage a research team.
- a team-oriented attitude
- a strong experience in *Open Science*, in order to make scientific results widely accessible. The Researcher will develop a data management plan that complies with the FAIR principles (Findable, Accessible, Interoperable, and Reusable).
- adherence to the “Code of ethics for scientific research in Belgium” which establishes the main principles of ethical scientific practice.

Complementary skills

Constitute assets:

- a professional network in the field of archival science or digital humanities and knowledge of research topics in these fields,
- expertise in international standards for archival description (in particular EAD3 and RiC) and systems modeling (in particular the OAIS reference model), as well as their implementation in a real archives management system.

Languages

A basic knowledge of French is necessary for a good integration within the two institutions. If, at the time of recruitment, the candidate does not know French (enough), UCLouvain will offer him or her training in order to acquire the required level. The candidate must also be able to communicate in English. Given the bilingual environment of the AGR (Belgian federal institution), knowledge of Dutch is also an advantage. UCLouvain can also provide training support in these two languages.

Terms of employment

The proposed position is a 100% position structured as follows:

- A 50% open-ended contract to be performed at the AGR in Brussels (rue de Ruysbroeck 2) at the grade of “chef de travaux” (SW2). Baremic scale SW21 (EUR 31,880.00 – EUR 48,350.00 gross non-indexed salary, i.e. an indexed gross monthly salary of approximately EUR 2,860 for this part-time, with no other seniority than the required PhD). All the services provided in the public services as a researcher can be valued.
- A 50% open-ended contract to be performed at UCLouvain (as a post-doctoral researcher) in Louvain-la-Neuve (Ruelle de la Lanterne Magique, 14). Hiring is done according to the current salary scale rules at the University and depends on seniority.

The entry into office is scheduled for October 1, 2023.

Interested?

Candidates must send their application file (a single file in PDF format) before **May 1, 2023** by email to eddy.put@arch.be and to antonin.descampe@uclouvain.be (reference: FEDTWIN/ARKEY application).

The application file must consist of a letter of motivation, a detailed CV with a list of publications, a copy of the required diploma, and two letters of recommendation.

The authors of the letters of recommendation will be invited to send their mail directly to Antonin Descampe and Eddy Put, without going through the candidate.

Candidates selected on the basis of their application file will be invited for an oral hearing by the Joint Selection Committee that will be organized during the months of May or June .

For any further information, please contact Antonin Descampe (antonin.descampe@uclouvain.be) or Eddy Put (eddy.put@arch.be).

UCLouvain and the AGR want to create a work environment in which all talents can develop as much as possible, without distinction of gender, age, cultural origin, nationality or disability. For any questions regarding accessibility or the possibilities of support, please consult <https://jobs.uclouvain.be/content/ValeursRH/> and/or contact the HR department of the AGR (pers@arch.be).