Arkindex, an open-source platform for document image recognition applied to archeology

Christopher Kermorvant, TEKLIA, Paris
Mélodie Boillet, TEKLIA, Paris
Pablo Ciezar, Institut National de Recherches Archéologiques Préventives
Anne-Violaine Szabados, CNRS - ArScAn - Archéologies et Sciences de l'Antiquité
Julien Schuh, UPN - Université Paris Nanterre

Introduction

Computer vision and machine learning have found many applications to assist archaeologists in their fieldwork (Bickler 2021). However, these techniques can also be useful for the exploitation of archaeological documentation, be it traditional scientific publications (articles, monographs, catalogues) or grey literature (internal reports, excavation notebooks, dissertations, plans, sketches, photos). Recent advances in artificial intelligence, in particular multimodal generative artificial intelligence, open up great prospects for the exploitation of this textual and iconographic data. This resource is abundant but often poorly catalogued and only available in its original form: on paper or, after digitization, as an image of the document. The PictorIA consortium brings together 47 partners - museums, archives and academic laboratories - from the French-speaking humanities, arts and social sciences (HASS). Its aim is to study and develop AI based methods and tools to extract and structure textual and iconographic information from image corpora in the HASS, particularly in the field of archaeology.

The exploitation of digitized archaeological documents presents three challenges: first, the volume of these documents is often substantial. Secondly, the content and format of these documents is very diverse, ranging from handwritten notes to photographs. Finally, the configuration of the extraction of information requires the professional expertise of those who produced the documents or will use the data. For these reasons, it is necessary to develop tools that are scalable to mass processing, adaptable to many types of documents and information to be extracted, and directly usable by end users. As part of its goal to evaluate tools, the PictorIA consortium evaluated the open-source Arkindex platform for extracting information from archaeological documents.

Methods

Arkindex is a digital document processing platform designed with three objectives: to integrate any type of document and processing method, to process up to millions of documents, and to adapt to the needs of users.

In order to adapt to any type of project and information, Arkindex does not require the use of predefined document structure types. The structure of the project can be customised, from the organisation of the corpus down to the smallest textual or visual element, by specialising

a generic type for each desired level. These types can then be organised hierarchically to reflect the overall structure of the project. Secondly, Arkindex does not restrict the type of information that can be extracted from these elements. Each element can be enriched with an orientation, classes, transcripts and metadata. All these enrichments can be preconfigured, constrained (for classes) or free (for metadata).

Arkindex enables the integration of any type of processing algorithm with an API that covers the full functionality of the platform and a modular design that separates data processing and storage. All processing is integrated into Docker containers that run on independent compute nodes and interact with the platform through the API. This architecture makes it possible to take advantage of the rapid improvement of algorithms by easily integrating and training them into the platform. Arkindex's distributed processing architecture enables scaling by distributing processing across many compute nodes, whether on local servers, in the cloud or using supercomputing resources through the SLURM system.

Finally, Arkindex offers a graphical web interface that allows users and business experts to contribute to the definition of the data structure, configure the processing and verify the results. The source code of Arkindex is open-source, which allows the community to either develop new features or fund developments.

Results

In this section, we will present 2 use cases to illustrate how Arkindex's generic and modular design allows it to process very different documents in the field of archaeology.

The first use case is the textual and visual indexing of the handwritten excavation journals of the French archaeologist Bernard Bruyère, preserved at the French Institute of Oriental Archaeology in Cairo. A corpus of 4 notebooks (892 pages) was organised and processed in Arkindex in order to obtain an index of text lines, hieroglyphic lines, plans and illustrations. The hieroglyphs, plans and illustrations were localised and manually enriched with the date and location of find, inventory number, other names and owner metadata. The illustrations were manually enriched with 62 classes describing the type of object or motif. All lines of handwritten text were automatically transcribed using a handwritten text recognition system (Maarand et al. 2022). The result of this semi-automatic text-image indexing will be published on the IFAO website.
The second use case concerns the analysis of wheel decorations on Argonne pottery (Bakker et al. 2018). For a preliminary study, 212 photographs of fragments of Argonne pottery were selected and imported into Arkindex. These images of fragments were manually annotated to locate and identify the different decorations present according to a side / fragment / wheel / motif hierarchy. A classification was then added to the patterns to categorise them into 9 pre-defined classes. This set of annotated images was then used to train a pattern detection and classification model using the YOLO open-source library integrated into the Arkindex platform. This preliminary study has enabled rapid validation of the automatic wheel pattern detection approach using the YOLO library, paving the way for a larger project.

Discussion

Since the 2010s and the rise of deep learning algorithms, two fundamental trends in the machine learning community have enabled the adoption of these artificial intelligence technologies in many application areas: on the one hand, the development of state-of-the-art open source software libraries and, on the other hand, a methodological convergence that makes it possible to solve many problems with the same models, namely deep neural networks, whether for images or text. The "last mile" is to make these tools accessible to end users through simple, adaptable and efficient interfaces. Arkindex aims to achieve this by providing an open-source platform for processing scanned documents, both text and images, that allows the integration of best-in-class AI models as they are developed and provides scalable processing capacity. The PictorIA consortium will continue to evaluate Arkindex through other use cases, contribute to its development to adapt it to the needs of archaeologists, and provide an instance and computing resources to partners.

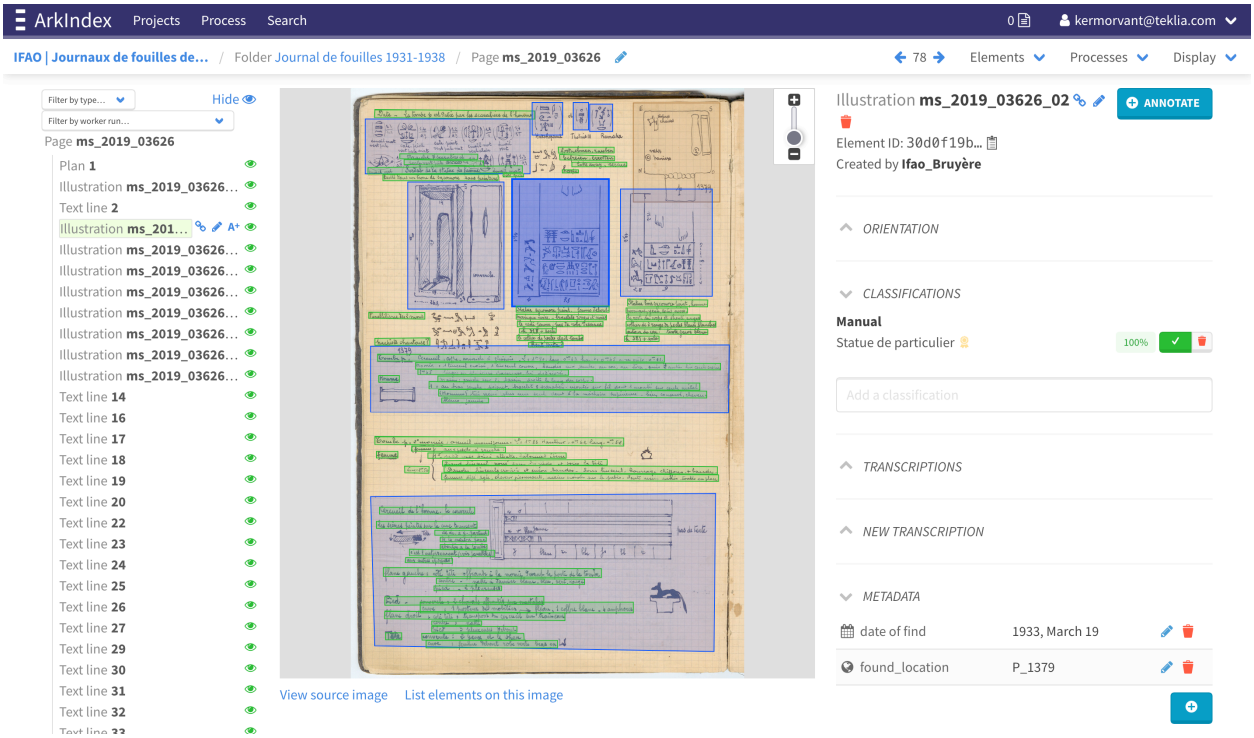

Figure 1: Text and illustration indexing of B. Bruyère's handwritten excavation journal using Arkindex.

References

- Bickler, Simon H. 2021. "Machine Learning Arrives in Archaeology." Advances in Archaeological Practice 9 (2): 186–91. https://doi.org/10.1017/aap.2021.6.

- Bakker, Lothar, Wim Dijkman, Paul van Ossel, and Pablo Ciezar. 2018. "Le corpus des décors à la molette sur céramique sigillée d'« Argonne » de l'Antiquité tardive." In La céramique en Champagne: production, diffusion et consommation, 211–22. Reims: SFECAG.
- Maarand, Martin, Yngvil Beyer, Andre Kåsen, Knut T. Fosseide, and Christopher Kermorvant. 2022. "A Comprehensive Comparison of Open-Source Libraries for Handwritten Text Recognition in Norwegian." In Document Analysis Systems: 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022, Proceedings, 399–413. Berlin: Springer-Verlag. https://doi.org/10.1007/978-3-031-06555-2_27.