From field notes to digital archive: AI-assisted indexing of Bernard Bruyère's excavation journals of Deir el-Medina (1922-1955)

Cédric Larcher, IFAO, Cairo, Egypte

Christopher Kermorvant, TEKLIA, Paris, France

Introduction

Artificial intelligence techniques are finding more and more applications in archaeology, whether to process tabular or geographical data, images or text. In the latter category, a distinction is made between scientific literature - indexed by bibliographic databases - and "grey literature", which includes all scientific documents that are not disseminated by following the academic publication process. This "grey literature", which includes excavation reports, constitutes an important part of the available archaeological data but remains under-exploited because it is difficult to index and analyse.

The excavation journals of Bernard Bruyère, archaeologist at the French Institute of Oriental Archaeology in Cairo, in the first half of the twentieth century, form a set of almost a thousand pages of texts and drawings that retrace the daily activities of the archaeological mission of Deir el-Medina in the field and list the discoveries made on site between 1922 and 1952. The archaeologist unearthed the village and the necropolis of the craftsmen who were responsible for digging and decorating the tombs of the pharaohs of the New Kingdom. Deir el-Medina is an exception in an archaeological landscape where elsewhere in Egypt, agglomerations have been superimposed over the centuries. The remains and objects, whose good state of preservation is undoubtedly due to the location of the site in the desert, have provided detailed information on the daily, social, professional and religious life of this community of craftsmen who lived there more than 3200 years ago. This quantity of documents explains why studies on Deir el-Medina constitute an important sub-field of Egyptology. Everything that the archaeologist has discovered on site is recorded in detail in his excavation journals, which are therefore regularly requested by researchers.

For conservation reasons, they have all been digitised and are now accessible from the IFAO website. However, the mere consultation of the images generated was not an appropriate method for searching for precise information. We therefore had to set up an interface that would allow us to search for text, index names, keywords and the many drawings of objects and plans made by the author. We have therefore developed an automatic enrichment chain for these digitized notebooks based on page analysis and automatic handwriting recognition algorithms.

Methods

The processing of Bruyère's excavation journals involved both automated and manual steps to ensure a high level of accuracy in the transcription and enrichment of the content. First, text line recognition was performed using a generic model based on Doc-UFCN (Boillet, Kermorvant and Paquet 2022), which identified individual lines of handwritten text within images of the pages.

A hybrid approach was used for text transcription. First, a training corpus was developed by manually transcribing 2,260 lines of text. This was supplemented by automatic alignment of a further 1,478 lines using pre-existing partial transcriptions. The resulting annotated dataset of 3,738 lines was used to train a recognition model based on the Kaldi library (Arora et al. 2019), using data augmentation techniques to increase model robustness.

The trained model achieved a character error rate of 4.5%. This model was used to automatically transcribe all the detected lines of text. To ensure the highest possible accuracy, all automatic transcriptions were then manually checked and corrected using the Callico interface (https://doc.callico.eu/).

The visual elements of the journals were too complex to be automatically detected with sufficient precision. Therefore, the Arkindex (https://doc.arkindex.org/) annotation interface was used to manually locate the illustrations, maps and lines of hieroglyphic text. These elements were then enriched with key metadata, including date of discovery, location, inventory numbers, associated names and ownership information. In addition, the illustrations were classified into 62 different categories based on object type or motif representation, creating a comprehensive taxonomic structure for the visual content.

Results

The processing workflow resulted in a comprehensive digital archive, accessible through the Arkindex platform, providing researchers with new access to Bruyère's excavation journals. The full dataset comprises 892 digitised pages from four notebooks, containing 36,506 lines of detected and transcribed handwritten text. The visual content includes 5,479 illustrations, 268 maps and 360 lines of hieroglyphic text, all precisely located and enriched with metadata.

The platform offers several ways to explore and analyse this rich archaeological documentation. Users can navigate through the four notebooks using a page-by-page interface, viewing both the original digitised pages and their enriched transcriptions. The search functionality offers several query approaches: full-text search within the handwritten content, which returns results with precise line locations on the original pages; filtered searches of the image database using the 62 predefined object classes; and targeted searches using structured metadata fields, including discovery dates, find locations, object owners and inventory numbers.

This multimodal search capability allows researchers to efficiently locate specific information within the rich documentation, whether they are tracking specific artefact types, investigating specific archaeological contexts, or following the chronological progression of discoveries. The system thus transforms a previously manual consultation process into an efficient digital research tool, while preserving the original structure and context of the Bruyère documentation.

Discussion

This project demonstrates the successful digitisation and enrichment of Bernard Bruyère's excavation journals, transforming these unique archaeological documents into a searchable digital resource. The workflow implemented, combining automated processing with manual validation, effectively processed nearly 900 pages of mixed handwritten text and illustrations, making this valuable documentation accessible to the research community.

While our approach relied on traditional document processing methods - text line detection, manual annotation of training data, and custom training of handwritten text recognition (HTR) models - recent developments in visual language learning models (LLMs) raise questions about possible alternative approaches. These powerful models could potentially transcribe handwritten text with minimal or no training data. However, current visual LLMs have significant limitations for this type of archival processing. Most critically, they do not provide precise text position information, which is essential for both manual verification and the implementation of location-aware search functionality. The ability to pinpoint the exact location of text in the original documents remains a critical feature for researchers consulting these materials.

Similarly, while automatic segmentation models such as Segment Anything show promise for illustration recognition, they currently lack the sophistication required to accurately interpret and segment the various types of archaeological documentation - from field sketches to engineering drawings and site plans. The nuanced understanding required for this task still requires human expertise and validation.

Nevertheless, these emerging technologies have the potential to improve the processing workflow. While manual validation is still necessary to ensure scientific standards, the integration of these tools could speed up the initial processing stages. Future work could explore hybrid approaches that leverage the strengths of both traditional document analysis techniques and emerging AI capabilities, potentially leading to more efficient processing pipelines for archaeological documentation.

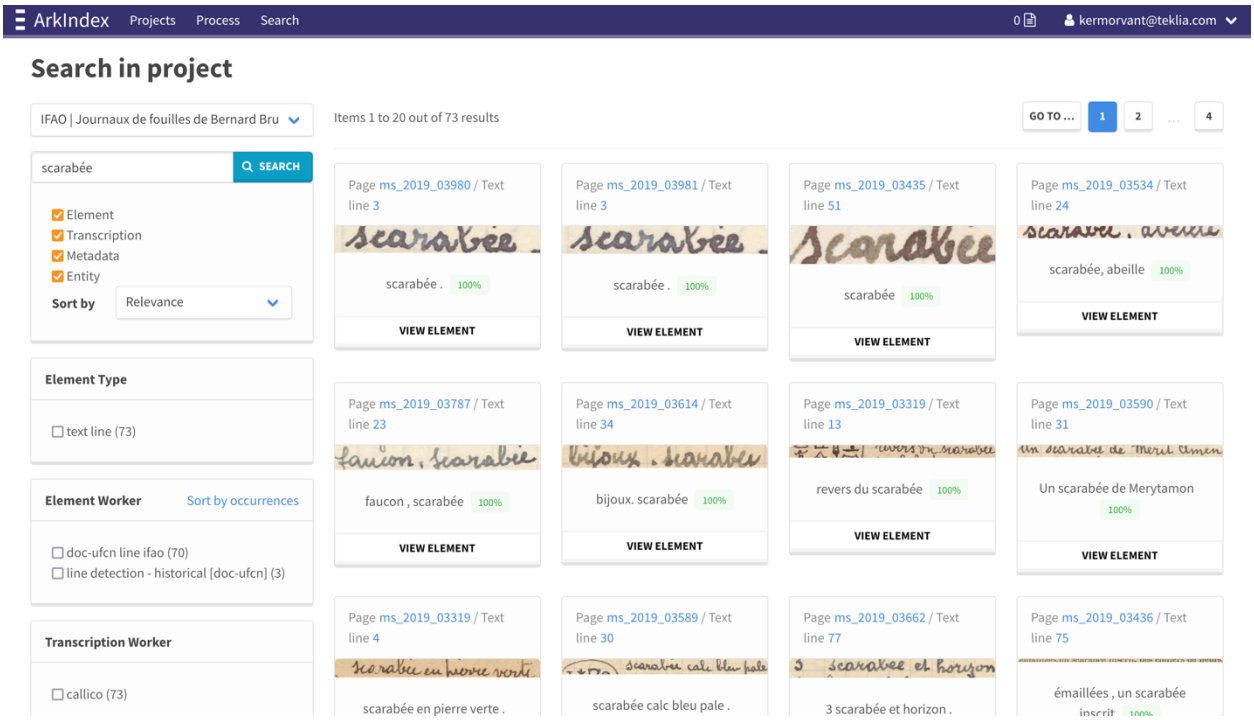

Figure 1: The Arkindex search interface on the excavation notebooks.

References

Larcher, Cédric. 2017. “Les archives de Bernard Bruyère concernant les fouilles de Deir el-Médina conservées à l'Ifao.” In *À l’œuvre on connaît l’artisan.. de Pharaon ! Un siècle de recherches françaises à Deir el-Medina (1917-2017)*, edited by Hanane Gaber, Laure Bazin Rizzo, and Frédéric Servajean, 311-330.

Arora, A., et al. 2019. "Using ASR Methods for OCR." In International Conference of Document Analysis and Recognition.

Boillet, Mélodie, Christopher Kermorvant, and Thierry Paquet. 2022. "Robust Text Line Detection in Historical Documents: Learning and Evaluation Methods." International Journal on Document Analysis and Recognition.