Conference Agenda

Overview and details of the sessions of this conference. Please select a date or location to show only sessions at that day or location. Please select a single session for detailed view (with abstracts and downloads if available).

Please note that all times are shown in the time zone of the conference. **The current conference time is: 26th June 2024, 02:05:39pm WEST**

Authors

Session Overview

Session

AI for DH

Time: **Wednesday, 19/June/2024: 11:30am - 1:00pm**
Session Chair: **Natalia Ermolaev**, Princeton University

Location: **Auditorium B3, 5th floor**

Auditorium B3, 5th floor, Avenida de Berna 26C, Berna Campus, NOVA FCSH, Lisbon, Portugal

Presentations

Building a Fichero: New Tools, Old Documents, and Machine Learning Workflows with an Endangered Afro-Colombian Archive

Andrew Janco[1], Kelly López Roldán[2], Daniel Tubb[3]
[1]University of Pennsylvania, United States of America; [2]Independent Researcher, Colombia; [3]University of New Brunswick, Canada

This paper describes outcomes and challenges in human-scale document processing. We discuss a workflow that begins with document preservation, moves through text recognition, and ends with a catalogue that demonstrates capabilities of LLMs for research and archival work, while remaining attuned to the vision of research partners.

Until 2022, the Istmina Circuit Court archive, with documents from the 1870s to 1930s, was rotting, disorganized, and in garbage bags. Yet, this archive is a crucial source of Afro-Colombian history in an often-marginalized region of the Chocó in Colombia. In 2023, seven young people from Istmina and Quibdó worked with the Muntú Bantú Foundation, a community center focused on Afro-diasporic memory. With researchers from various universities, they were able to digitize the archive, which is available online at the British Library. While the project was successful, the digitization has enabled new workflows to catalogue and interpret the archive. This paper explores these workflows.

Throughout, we are interested in a key problem of equity in knowledge production: How can new tools be used to the benefit of local knowledge-producers? Our paper focuses on the work of cataloguing archival materials, a first step in enabling local researchers (and others) to make meaning. Project interns catalogued 330 Case Files and wrote a book of micro-history. Yet, the task of cataloguing 470 more cases remains daunting. We reflect on machine learning pipelines to extract text, catalogue the archive, and understand 61,000 images using Weasel, a digital workflow system.

To extract text, we built a workflow which fetches images hosted on the British Library; uses Kraken to segment text in each image; deploys Google Vision to extract handwritten and typewritten text; sends the image, the polygonal representations, and text to eScriptorium, where users can review it and correct the text. Here, we discuss both positive outcomes and ongoing challenges in recognizing text in typewritten versus handwritten documents, in segmenting images into regions, and working with these tools.

From there, to create a catalogue, we built workflow that downloads transcriptions from eScriptorium; uses spaCy named entity recognition to extract and link the names of people, places, dates, events, organizations; and employs open-source large-language models running locally via Ollama (mistral:instruct and mistral:8x7b) to generate summaries, timelines, catalogue entries, and other catalogue material. We run experiments with RAGatouille, ColBERT, LangGraph powered agent-based workflows to do further work on the archive. Finally, we export this material as a catalogue of extracted text, summaries, named entities, keywords, etc. into a fichero, a box of linked and tagged digital index cards in Markdown format which Obsidian and similar tools make accessible to non-technical users. Additionally, we use Nomic's Atlas to map the data and metadata. The collection can be published to the web using 11ty or other static generators.

Here, we discuss steps to create a machine-generated catalogue, challenges to choose the right approaches, the costs of online commercial AI models, and the importance of the right tool.

Automatic Clustering of Hebrew Manuscripts

Daria Vasyutinsky Shapira, Berat Kurar-Barakat, Mohammad Suliman, Sharva Gogawale, Nachum Dershowitz
Tel Aviv University, Israel

This paper presents the interdisciplinary research conducted at Tel Aviv University in the framework of the ERC Synergy project, MiDRASH. Our work combines scholarly domains of Hebrew paleography and deep machine learning.
We aim to automatically cluster medieval Hebrew script types-modes beyond the limits of contemporary human paleography. Currently, we are working on Ashkenazi square script.
Successful algorithms will be applied to other medieval Hebrew script type-modes, allowing improved clustering of the least-studied script types, such as Byzantine and Yemenite, and deepening our understanding of the sub-clustering of Oriental, Sephardic, and Italian scripts. This, in turn, will lead to the discovery of new paleographic patterns, improved layout segmentation based on script types, and more.