



HAL
open science

Streamlining the Creation of Holocaust-related Digital Editions with Automatic Tools

Sarah Bénéière, Floriane Chiffolleau, Hugo Scheithauer

► To cite this version:

Sarah Bénéière, Floriane Chiffolleau, Hugo Scheithauer. Streamlining the Creation of Holocaust-related Digital Editions with Automatic Tools. EHRI Academic Conference - Researching the Holocaust in the Digital Age, EHRI-3, Jun 2024, Varsovie, Poland. hal-04594190v1

HAL Id: hal-04594190

<https://inria.hal.science/hal-04594190v1>

Submitted on 30 May 2024 (v1), last revised 24 Jun 2024 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Streamlining the Creation of Holocaust-related Digital Editions with Automatic Tools

Sarah Bénéière¹, Floriane Chiffolleau^{1,3}, Hugo Scheithauer^{1,2}

¹ALMAnaCH research team, Inria Paris

²École Pratique des Hautes Études (EPHE) - PSL University

³Le Mans Université

Conference paper proposal

Abstract

The [EHRI Online Editions](#) are collections of archival Holocaust-related documents kept in various institutions, and are each curated around a specific theme. The current editorial process requires a great amount of tasks to be done manually—some of which can be automated with the help of machine learning tools. In this conference paper proposal, we would like to propose a semi-automated workflow (fig. 1), building on automatic text recognition (ATR) and document layout annotation (DLA) technologies, as well as on a customized EHRI TEI schema, for streamlining the creation of future holocaust-related digital scholarly editions.

The text transcriptions available for all EHRI edited documents were aligned to their original scans to create a dataset suitable for training ATR models, which aim at automatically acquiring digital text from physical documents. The dataset is composed of typewritten documents, distributed in subsets sorted by language, and in a multilingual subset (table 1). Our experiments show that each Kraken ATR model achieves high accuracy levels, and could then be used to efficiently extract text from typewritten digitized documents, for instance with the help of an ATR graphical user interface such as [eScriptorium](#) (Stokes et al. 2021). We also propose to structure acquired textual data with the help of DLA models. DLA aims at detecting the layout hierarchy of a document, such as headers, titles, paragraphs, etc. We plan to sample the EHRI digital editions to create a dataset for training a DLA model in the upcoming year. We are planning to use an object detection model, as they demonstrated their reliability for segmenting the layout of textual documents and gained increasing popularity in digital humanities projects (Clérice 2022; Sven & Matteo 2022).

The structured transcriptions are then semi-automatically encoded in TEI-XML with Python scripts. We designed a two-fold process that creates, for each digitized

document, a TEI file with the metadata of the source document and a structured text body based on layout information. The semantic encoding is supervised by a customization of the TEI standard created for EHRI, allowing the creation of enriched texts that can be searched like databases.¹

On the very end of that workflow, we would like to present a tool specifically made for TEI-based editions, [TEI Publisher](#), by showcasing a dedicated and ready-to-use EHRI application we developed, (fig. 2) in order to offer the possibility of a centralized interface for future holocaust-related editions.

References

- Clérice, T. (2022). *You Actually Look Twice At it (YALTAi): Using an object detection approach instead of region segmentation within the Kraken engine*.
<https://hal-enc.archives-ouvertes.fr/hal-03723208>
- Stokes, P. A., Kiessling, B., Stökl Ben Ezra, D., Tissot, R., & Gargem, E. H. (2021). The eScriptorium VRE for Manuscript Cultures, in *Ancient Manuscripts and Virtual Research Environments. Classic @ Journal*, 18.
<https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/>
- Sven, N.-M., & Matteo, R. (2022). *Page Layout Analysis of Text-heavy Historical Documents: A Comparison of Textual and Visual Approaches* (arXiv:2212.13924). arXiv.
<https://doi.org/10.48550/arXiv.2212.13924>

¹ We use a customized TEI ODD, available at <https://gitlab.inria.fr/dh-projects/workflow-ehri/-/tree/main/ODD>

Figure 1. Schematization of the proposed workflow

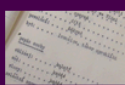


Figure 2. Homepage of the EHRI dedicated TEI Publisher application


Home Collections About Langue Français [Se connecter](#)

EHRI Online Editions

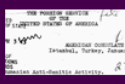
Trier par Titre Filtrer selon Titre Filtrer




Bordered Escape
This edition gathers documents testifying of the increasingly restrictive refugee policy in Czechoslovakia, following the "Anschluss".




Early Holocaust Testimonies
This edition gathers accounts of the persecution of Jews from the Nazi takeover of power in Germany (1933) to the Eichmann Trial (1961).




Diplomatic Reports
This edition gathers reports on the persecution and murder of European Jews.



Nisko Deportations
This edition gathers documents on the deportation of thousands of Jews to Nisko am San (Poland).



Indexes
Indexes of the EHRI Online Editions



Annotation Tool
Annotation Tool to work on the named entities recognition of the EHRI Online Editions

TEI Publisher 0.1 / web components 2.4.5 / API 1.0.0

This application was developed as part of the EHRI project with the help of TEI Publisher.



Table 1. Presentation of the EHRI automatic text recognition dataset and training results

| Language | Source | Number of documents | Number of lines | Automatic text recognition model accuracy |
|--------------|--|---------------------|-----------------|---|
| German | BeGrenzte Flucht; Die Nisko-Deportationen; Early Holocaust Testimony | 56 | 2287 | 97.9% |
| English | BeGrenzte Flucht; Early Holocaust Testimony; Diplomatic Reports | 54 | 1989 | 97.5% |
| Czech | BeGrenzte Flucht; Early Holocaust Testimony | 46 | 1713 | 96.7% |
| Danish | Diplomatic Reports | 36 | 1007 | 97.8% |
| Hungarian | Early Holocaust Testimony | 30 | 1334 | 95.7% |
| Polish | Early Holocaust Testimony | 15 | 468 | 93.1% |
| Slovak | BeGrenzte Flucht | 15 | 395 | 93.7% |
| Multilingual | BeGrenzte Flucht; Die Nisko-Deportationen; Diplomatic Reports; Early Holocaust Testimony | 252 | 9193 | 97.2% |