Transcribathon.eu: AI supporting collaborative transcription and enrichment of historical documents

Sergiu Gordea
Medina Andresel
<first>.<last>@ait.ac.at
AIT Austrian Institute of Technology GmbH
Vienna, Austria

Frank Drauschke
drauschke@factsandfiles.com
Facts & Files Digital Services GmbH
Berlin, Germany

Philip Kahle
p.kahle@readcoop.eu
READ-COOP SCE
Innsbruck, Austria

ABSTRACT

This paper presents the Transcribathou.eu tool and Europeana Transcribe, a citizen science initiative which aims at crowdsourcing transcription and enrichment of historical documents made available through the Europeana.eu platform. The public value of this work is internationally recognized through the honorary mention[1] awarded within the scope of European Union Prize for Citizen Science. The process for performing a complete digitization of historical documents is described together with AI technologies used for analyzing document content, generating semantic enrichments, clustering and visualizing documents by their topics.

CCS CONCEPTS

• **Information systems** → **Extraction, transformation and loading**; **Entity resolution**.

KEYWORDS

Transcribathon, collaborative transcription, semantic enrichment, Handwritten Text Recognition, cultural heritage

ACM Reference Format:
Sergiu Gordea, Medina Andresel, Frank Drauschke, and Philip Kahle. 2024. Transcribathon.eu: AI supporting collaborative transcription and enrichment of historical documents. In *International Conference on Advanced Visual Interfaces 2024 (AVI 2024), June 03–07, 2024, Arenzano, Genoa, Italy*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3656650.3656730

1 INTRODUCTION

Europeana, the EU Commission's platform for cultural heritage, provides online access to more than 50 million cultural heritage objects. Among those, a large number of handwritten or printed documents depicting historical events such as the First World War, The revolution from 1989, or societal transformations from the past centuries. However, the simple scans do not automatically enable barrier free access to the embodied information, especially when taking into consideration the large number of languages used, changes of the scripts and personal marks of the handwriting.

The Europeana Transcribe on Transcribathon.eu is a citizen science initiative aiming to fully digitize selected collections of documents. The initial prototype was implemented by Facts & Files Berlin, and was further developed to a fully featured product within the scope of EnrichEuropeana and EnrichEuropeana+ projects. While the initial prototype showcased a gamification approach for stimulating user engagement in transcription activities, subsequent development focused on integrating AI technologies to streamline the transcription and semantic enrichment processes. Furthermore, the extraction of machine readable texts enables the employment of natural language processing (NLP) technologies to analyze the information content, to cluster documents in specific topics and generate user friendly visualizations as shown in Section 4. The main contributions of this work is two fold. Firstly, the automatic handwritten text recognition, which was previously available only to professional users of Transkribus tool, is now made available for the large public, including both the active transcribers and the visitors of Transcribathon website. In the following subsections we showcase the advanced AI technologies and the collaborative process implemented within the tool to achieve the complete, high quality digitization and visualization of historical materials.

2 RELATED WORK

Different citizen science tools for manual transcription were introduced in the last decade with the aim to provide full text access to books and manuscripts digitized by public institutions and private persons, such as Transcribe Bentham [2] or Guttenberg Project. The development of the initial Transcribathon prototype was inspired by the European wide initiative of commemorating the Fist World War, for which a large number of memorabilia was digitized [4]. The machine learning research is a hot topic nowadays, due to the development of high performance computing and big data technologies. Within this context, the technology for Handwritten Text Recognition (HTR) was researched and successfully applied to large corpora of manually transcribed documents [5][6]. In turn, this enables the use of natural language processing technology to solve various tasks such as machine translation, enrichment, topic detection or document clustering [1]. The Transcribathon tool was further developed in the past years to integrate advanced AI technologies for automating and scaling up the process of crowdsourcing transcriptions, semantic enrichment and advanced visualization of publicly accessible manuscripts. The functionality integrated in the online tools is presented in the following sections.

[1] https://ars.electronica.art/citizenscience/en/europeana-transcribe-on-transcribathon-eu/

AVI 2024, June 03–07, 2024, Arenzano, Genoa, Italy

ACM ISBN 979-8-4007-1764-2/24/06
https://doi.org/10.1145/3656650.3656730

3 TRANSCRIBATHON COMPETITIONS

Transcribathon.eu facilitates user's engagement in crowdsourcing activities for the transcription and enrichment of heritage materials, following the goal of transforming historical handwritten documents in computer readable format, freely accessible for everyone. While the automated text recognition presented in Section 4 tremendously reduces the effort for generating document transcriptions, its use is still limited to certain types of documents and languages for which the HTR models are available. Therefore, the role of the end users is critical for this platform. Human abilities of deciphering hard-to-read handwritten texts and finding small details, nuances and connections between concepts are critical for the interpretation of handwritten materials (see Fig. 1). Consequently, the long term commitment of the user community is a key asset of the Transcribathon platform.

Figure 1: User engagement in Transcribathon competition

The Transcribathon frontend focuses on providing and enhanced user experience (UX) when interacting with the digitized documents, aiming to ensure continuous user engagement (see ref. [3] - Fig 2). The gamification component contributes to the achievement of this goal by stimulating user engagement trough time-bound online or on-site competitions, the so called Transcribathon Runs (see for example the 19th Century Run [2]). The participants compete with each other and get rewarded for their achievements. All contributions are open access, and the results made available immediately.

4 AUTOMATIC TRANSCRIPTION AND ENRICHMENT

The transcription of manuscripts from scratch is a time-consuming activity even for experienced paleography experts. Handwritten text recognition (HTR) technology tremendously reduces this effort (see ref. [3] - Fig 3). The Transcribathon tool integrates with the HTR services made available by READ-COOP through the Transkribus Metagrapho API[3], which makes available a set of public HTR models[4] trained on large corpora of documents. Custom models, trained in the Transkribus platform are also accessible through the tool's REST API. Moreover, these public models are continuously enhanced with additional (manually transcribed) documents and can be successfully reused to process collections of similar documents (i.e. using the same language and script from the same time period). When appropriate models are available, a typical average character error rate around 5% can be expected [5]. While the processing of selected document collections is triggered by system administrators, the end users can concentrate on proofreading, correcting recognition errors and approving the HTR text. The HTR editor provides an enhanced user experience by aligning the lines of text with the corresponding section in the original image.

Full text digitization of scanned materials enables further processing of transcribed documents. While the cultural heritage materials are available in one of the 40+ languages and dialects used in Europeana platform, machine translation is applied to make the content available in English language. Consequently, the information becomes accessible for a large proportion of online users. Moreover, these translations enable the use of a unified approach for computing automatic semantic enrichments for all transcribed documents (see ref. [3] - Fig 4). A custom Named Entity Recognition algorithm[5] was developed by combining the Stanford NER, Dbpedia Spotlight and Wikidata search to discover entities referenced within document descriptions and transcriptions. By linking places, persons and organizations discovered within the documents with Wikidata/Wikipedia articles, public users get access to additional information related to the narrative presenting historical events.

Transcribathon's vibrant user community was engaged up to now in the transcription of ca. 400K of document pages, accounting to an amount of more than 52 million characters and more than 150K semantic enrichments. The reviewed contributions are delivered and made accessible within the Europeana platform. This enables advanced visualization and search for or within individual full text data records[6]. Moreover, the contextualization of related documents in form of curated collections (see War Correspondence Gallery[7]) provides a meaningful browsing and exploration alternative to the default search based entry point. The process of creating curated collections is successfully supported by specialized algorithms for document clustering, as demonstrated in [1].

5 CONCLUSIONS

This paper aims to demonstrate the functionality of Transcribathon tool and the Europeana Transcribe initiative, which aim at providing transcriptions and rich semantic information for historical documents accessible through Europeana platform. Full-text digitization and translation of historical materials is an activity which mainly involves cultural heritage professionals and passionate citizens. The outcome of their work is made publicly accessible and freely reusable by any interested party. In particular, historians and digital humanities researchers get barrier free access to valuable datasets relevant for their work.

[2] https://europeana.transcribathon.eu/runs/19th-century/
[3] https://www.transkribus.org/metagrapho
[4] https://www.transkribus.org/public-models/
[5] https://github.com/EnrichEuropeana/Technical-Documentation?tab=readme-ov-file#semantic-enrichment-api
[6] https://www.europeana.eu/en/item/719/_8w334x59r
[7] https://www.europeana.eu/en/galleries/10831-wwi-war-correspondence

6 AKNOWLEDGEMENTS

The work presented in this paper is co-funded by the European Union under the projects "EnrichEuropeana: Enriching Europeana with user transcriptions and annotations" (Action number 2017-EU-IA-0142), "EnrichEuropeana+: Enriching Europeana through citizen science and artificial intelligence - Unlocking the 19th Century" (Action number 2020-EU-IA-0075) and “AI4Culture: An AI platform for the cultural heritage data space” (Action number 101100683).

7 CITATIONS AND BIBLIOGRAPHIES

REFERENCES

[1] Medina Andresel, Sergiu Gordea, and Srdjan Stevanetic. 2023. Curating User Galleries in Europeana using Topic Modeling Technology. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Greece) *(PETRA '23)*. Association for Computing Machinery, New York, NY, USA, 243–244. https://doi.org/10.1145/3594806.3596524

[2] Tim Causer, Justin Tonra, and Valerie Wallace. 2012. Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham*. *Literary and Linguistic Computing* 27, 2 (03 2012), 119–137. https://doi.org/10.1093/llc/fqs004 arXiv:https://academic.oup.com/dsh/article-pdf/27/2/119/2746713/fqs004.pdf

[3] Sergiu Gordea. 2024. Annex 1: Suplementary Materials - Schreenshots for demonstrating the functionality of Transcribathon tool within the context of the AVI2024 demo paper. https://github.com/EnrichEuropeana/Technical-Documentation/blob/master/papers/avi2024/suplements/Annex%201%20-%20Screenshots.pdf

[4] Jeremy Jenkins. 2018. The British Library, Europeana 1914-1918 and the Memorialization of the Great War. *Electronic British Library Journal* (2018). https://doi.org/10.23636/1085 arXiv:https://bl.iro.bl.uk/downloads/2e92acf1-60d0-45ed-941d-c9b42615b7a0?locale=en

[5] Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 04. 19–24. https://doi.org/10.1109/ICDAR.2017.307

[6] Gundram Leifert, Christel Annemieke Romein, Achim Rabus, Phillip Benjamin Ströbel, Benjamin Kiessling, and Tobias Hödel. 2023. Evaluating State-of-the-Art Handwritten Text Recognition (HTR) Engines; with Large Language Models (LLMs) for Historical Document Digitisation. https://doi.org/10.5281/zenodo.8102666