AI4Culture platform: upskilling experts on multilingual / -modal tools

Tom Vanallemeersch and **Sara Szoc** and **Marthe Lamote** and **Frederic Everaert**
CrossLang NV, Franklin Rooseveltlaan 348/8, 9000 Gent, Belgium
firstname.lastname@crosslang.com

Eirini Kaldeli
National Technical University of Athens, Greece
ekaldeli@image.ntua.gr

Abstract

The AI4Culture project, funded by the European Commission (2023-2025), developed a platform (https://ai4culture.eu) to educate cultural heritage (CH) professionals in AI technologies. Acting as an online capacity building hub, the platform describes openly labeled data sets and deployable and reusable tools applying AI technologies in tasks relevant to the CH sector. It also offers tutorials for tools and *recipes* for the combination of tools. In addition, the platform allows users to contribute their own resources. The resources described by project partners involve applications for optical or handwritten character recognition (OCR, HTR), generation and validation of subtitles, machine translation, image analysis, and semantic linking. The partners customized various tools to enhance the usability of interfaces and components. Here, we zoom in on the use case of correcting OCR/HTR output using various means (such as an unstructured manual transcription) to facilitate multilingual accessibility and create structured ground truth (text lines with image coordinates).

1 Introduction

The AI4Culture project, which was funded by the DIGITAL program of the European Commission (EC) and took place from April 2023 until March 2025, developed an online capacity building hub for AI technologies in the sector of cultural heritage (CH). The platform makes CH data and tools involving varying modalities more accessible, understandable, and multilingual, supports heritage preservation, and contributes to making the common European data space for CH[1] more interoperable with AI technologies. The project coordinator is the AILS Laboratory of the National Technical University of Athens (NTUA). Partners in the CH sector include the Europeana Foundation, the European Fashion Heritage Association, the DigitGLAM unit at the University of Leuven, and the Institute for Sound and Vision. The technical partners include CrossLang, Datable, Datoptron, Pangeanic, and Translated, as well as the Machine Translation (MT) Research Unit at Fondazione Bruno Kessler (FBK) and the Digital Safety and Security Center of the Austrian Institute of Technology (AIT).

During the project, the partners focused on four types of technologies: (1) optical or handwritten character recognition (OCR, HTR) of scanned documents and MT of the transcriptions; (2) generation and validation of subtitles; (3) MT of documents and metadata; and (4) enrichment of metadata through image analysis and semantic linking. The partners customized tools to enhance the performance and usability of interfaces and components and organized workshops and a hackathon to involve stakeholders. The latter can also enrich the platform by contributing their own resources.

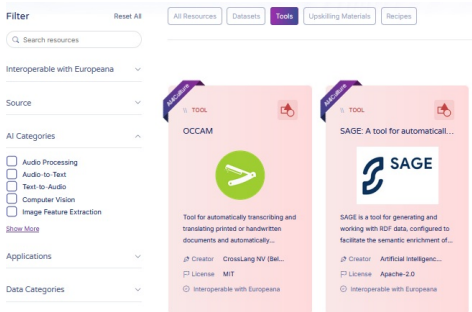

Figure 1: Exploring tools on the AI4Culture platform

2 Platform

The platform https://ai4culture.eu, launched in October 2024 and shown in Figure 1, offers a wide variety of AI-related resources: (1) descriptions of openly labeled data sets for training, testing, and evaluating models; (2) descriptions of de-

© 2025 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

[1] https://www.dataspace-culturalheritage.eu/en

Proceedings of Machine Translation Summit XX Volume 2, pages 99–100, June 23-27, 2025

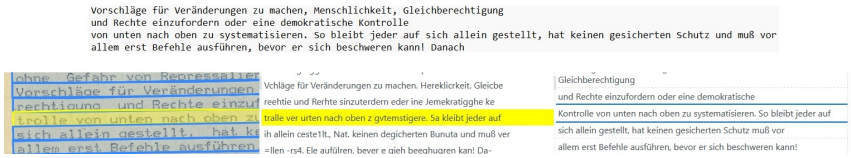

Figure 2: Unstructured manual transcription supporting the correction of OCR output

ployable and reusable tools for applying AI technologies in CH tasks; (3) tutorials (upskilling material) on such tasks; and (4) *recipes* illustrating the combination of tools for complex tasks. Target users include CH professionals and students, data providers, researchers, and AI developers.

The project partners registered descriptions of the tools they customized and other relevant tools. They focused on open source tools and the possibility to run tools locally. The tools provided by the partners can interact with `https://www.europeana.eu` and other CH data space components. The partners created data sets during the project, which they describe on the platform (for instance, PageXML files containing OCR/HTR transcriptions). After registering, platform users can contribute by uploading their own resources, thus raising awareness of their work. The platform allows for looking up resources based on criteria such as the AI technologies used, application types, etc. A resource description links to the repository where the actual resource is stored.

The project hosted a series of capacity building activities to provide professionals with hands-on experience. These include workshops on various technologies, the recordings of which are available on the platform, and a hackathon at the University of Leuven. A series of interviews with technical partners is also available on the platform.

3 Customization

Existing software has been customized in various ways: (1) FBK and Translated set up an open-source automatic subtitling system (Gaido et al., 2024); (2) Pangeanic combined computer-aided translation functionalities with CH-oriented MT engines; (3) NTUA and Datoptron extended their tools for semantic enrichment and validation of metadata and integrated them with the CH data space (Kaldeli et al., 2024); (4) Datable provided an object and color detection tool; (5) CrossLang added an open source HTR tool to its OCCAM transcription and translation environment[2] and functionality for automatic correction of OCR/HTR output and thus improved MT of transcriptions (Vanallemeersch et al., 2024), and (6) AIT interlinked Transcribathon[3] with the OCCAM services.

Regarding OCR/HTR, the two correction approaches reported by Vanallemeersch et al. (2024) (using a lexicon and language models) were extended towards the project end with a method matching output to an existing unstructured manual transcription (i.e. flat text). A final manual validation of the result of the approaches (structured, i.e. text lines with image coordinates) leads to ground truth useful for training new models or fine-tuning existing ones. While the first two approaches show variable performance (especially for low initial output quality), the third one shows clear results, as illustrated in Figure 2, where the (approximate) correspondence between lines in the image and a very long line in the manual transcription is detected. This "recycling" technique allows for reaching a minimally low CER score (character error rate), even for very poor original output quality.

Acknowledgments

AI4Culture is funded by the EC's DIGITAL program (project 101100683).

References

Marco Gaido, Sara Papi, Mauro Cettolo, Roldano Cattoni, Andrea Piergentili, Matteo Negri, and Luisa Bentivogli. 2024. Automatic subtitling and subtitle compression: FBK at the IWSLT 2024 subtitling track. In *Proceedings of IWSLT 2024*, pages 134–144.

Eirini Kaldeli, Alexandros Chortaras, Vassilis Lyberatos, Jason Liartis, Spyridon Kantarelis, and Giorgos Stamou. 2024. Combining automatic annotation with human validation for the semantic enrichment of cultural heritage metadata. In *Proceedings of CHR 2024*, pages 353–368.

Tom Vanallemeersch, Sara Szoc, and Laurens Meeus. 2024. AI4Culture: Towards multilingual access for cultural heritage data. In *Proceedings of EAMT 2024*, pages 59–60.

[2]https://ai4culture.crosslang.dev/ui

[3]https://europeana.transcribathon.eu