An introduction to Transkribus: how to use Handwritten Text Recognition in research and teaching

Mansutti, Sara
s.mansutti@readcoop.eu
READ COOP, Austria

Brief description

The tutorial provides a practical introduction to Transkribus, a comprehensive platform for the digitisation, AI-powered text recognition, transcription and searching of historical documents. The participants will learn how to run the layout analysis (also for tables and complex layouts) and train and apply Text Recognition AI-models to automatically generate machine-readable transcriptions of images of text. In the end, they will have a general understanding of the potential of Handwritten Text Recognition to open new research questions and how Transkribus can be used to collaborate, involve volunteers and teach in DH and History classes.

Instructor

Sara Mansutti works as Education Manager at READ-COOP, the European cooperative society maintaining and further developing the Transkribus platform. She holds a M.A. in Italian Philology and Literature from the University of Udine and is now pursuing a PhD in Digital Humanities at University College Cork. Her research interests include early modern history, archival studies, and Digital Humanities, in particular, combining Handwritten Text Recognition and crowdsourcing.

Target audience

The tutorial is open to researchers, teachers and students working with handwritten and printed primary sources, with interest in understanding if Handwritten Text Recognition and Transkribus can help them to produce reliable automatic transcriptions, enrich them with tags, search and extract data, and answer new research questions.

Requirements

No previous knowledge of the platform is required. The tutorial will be combined with hands-on sessions: to fully enjoy them, the participants are advised to bring their own laptop and prepare in advance a folder with about ten images of handwritten/printed documents on which they would like to apply Text Recognition.

Outline

The study of our written past has often been limited by the difficulty of reading handwritten documents and the considerable amount of time required to transcribe them manually. These factors have often prevented the study of large collections of manuscripts and archival documents as well as limited the discoverability only to the published ones. Over the last 30 years, libraries and archives have massively digitized their collections and made them available online, providing access to a large number of written items, regardless of users' location and time. However, the way of exploring and studying them has not changed: instead of reading the manuscript in the library's reading room, the user can now read a digitized image of it on their screen. What Handwritten Text Recognition is making possible is to automatically make these images of text searchable and readable, not only for a human but also for a machine. The potential is enormous both for analysing these texts with digital techniques and for opening new research questions.

In recent years, Transkribus has emerged as a comprehensible and user-oriented platform that allows everyone to apply Handwritten Text Recognition and train customised AI models. It was developed within the Horizon 2020 "READ" EU project by a consortium of leading research groups from all over Europe, headed by the University of Innsbruck, and is now sustained and further developed by the READ-COOP European Cooperative Society.

Besides the technology, however, the core of Transkribus has been its user community, to such an extent that Transkribus has been defined as a community of practices. Many users collaborate with the team to test, improve and suggest new features, lead workshops, support new users, share their positive and negative experiences with the platform, and make their AI models public so that other users can apply them to their documents. Collaboration is at the heart of Transkribus, and this tutorial aims to show not only the many features available but also how the platform can be used as a collaborative tool.

The workshop will begin with an introduction to Handwritten Text Recognition. Afterwards, the main features of the Transkribus platform will be presented and explained: manually and automatically running the layout analysis of the pages; training baselines models to detect complex layouts; training and applying AI models to automatically transcribe handwritten and printed texts; evaluating and computing the accuracy of a model; working with tables; enriching transcriptions with textual and structural tags; exporting the documents in various formats.

Explanations will be alternated with hands-on sessions where participants will work on their own documents within Transkribus.

Finally, the tutorial will present how Trankribus can be used in the classroom as a teaching tool to explain how Text Recognition works, teach palaeography, engage students in reading historical primary sources, and show them how digital tools can be integrated into historical research. Some teaching projects already using Transkribus will be mentioned, and the recently redesigned Transkribus Learn platform will be explained and tested with the participants.

Learning outcomes

By the end of the tutorial, participants will:
- Develop an understanding of what Automatic Text Recognition is and how it works;

- Practise with Transkribus and learn how to: run the layout and textual recognition; train baseline and text recognition models; enrich transcriptions with structural and textual tags; export the documents in different formats (including TEI);
- Understand the opportunities that Automatic Text Recognition can offer to answer new types of research questions;
- Explore the possibility of using Transkribus to engage volunteers and students.

Bibliography

Muehlberger, Guenter / **Seaward, Louise** / **Terras, Melissa** / **Oliveira, Sofia Ares** / **Bosch, Vicente** / **Bryan, Maximilian** / **Colutto, Sebastian** / **Déjean, Hervé** / **Diem, Markus** / **Fiel, Stefan** / **Gatos, Basilis** / **Greinoecker, Albert** / **Grüning, Tobias** / **Hackl, Guenter** / **Haukkovaara. Vili** / **Heyer, Gerhard** / **Hirvonen Lauri** / **Hodel, Tobias** / **Jokinen, Matti** / **Kahle, Philip** / **Kallio, Mario** / **Kaplan, Frederic** / **Kleber, Florian** / **Labahn, Roger** / **Lang, Eva Maria** / **Laube, Sören** / **Leifert, Gundram** / **Louloudis, Georgios** / **McNicholl, Rory** / **Meunier, Jean-Luc** / **Michael, Johannes,** / **Mühlbauer, Elena** / **Philipp, Nathanael** / **Pratikakis, Ioannis** / **Puigcerver Pérez, Joan** / **Putz, Hannelore** / **Retsinas, George** / **Romero, Verónica** / **Sablatnig, Robert** / **Sánchez, Joan Andreu** / **Schofield, Philip** / **Sfikas, Giorgos** / **Sieber, Christian** / **Stamatopoulos, Nikolaos** / **Strauß, Tobias** / **Terbul, Tamara** / **Toselli, Alejandro Héctor** / **Ulreich, Berthold** / **Villegas, Mauricio** / **Vidal, Enrique** / **Walcher, Johanna** / **Weidemann, Max** / **Wurster, Herbert** / **Zagoris, Konstantinos** (2019): "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study" in *Journal of Documentation* 75, 5: 954-976. DOI: 10.1108/JD-07-2018-0114.

Romein, C. Annemieke / **Kemman, Max** / **Birkholz, Julie M.** / **Baker, James Baker** / **De Gruijter, Michel** / **Meroño-Peñuela, Albert** / **Ries, Thorsten** / **Ros, Ruben** / **Scagliola, Stefania** (2020): " State of the Field: Digital History" in *History* , 105: 291-312. DOI: 10.1111/1468-229X.12969.

Terras, Melissa (2022): "The role of the library when computers can read: Critically adopting Handwritten Text Recognition (HTR) technologies to support research", in Hervieux, Sandy / Wheatley, Amanda (eds): *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries*. Atlanta: ACRL - Association of College & Research Libraries 137-148.