Workshop: Transkribus Lite

Format: In person (and if desired: hybrid)

Audience: All interested in the use of Transkribus Lite (whether first time users, or those using eXpert initially)

Organisers: Dr. C.A. Romein (Huygens Institute for History and Culture of the Netherlands, Amsterdam);

Joost Oosterhuis (Universiteit van Amsterdam);

Bram Jacobs (READ COOP)

Length: workshop basic (3 hours) + workshop advanced (3 hours),

Transkribus eXpert to Lite

Recognising handwritten texts through palaeography is a time-consuming task. Historical Handwritten texts can be a challenge to read for untrained eyes. Even when reading handwritten texts frequently, it takes a lot of time. With the EU funded tranScriptorium-project (2013-2015) and the READ-project (Recognition and Enrichment of Archival Documents – 2016–2019) a consortium worked on the creation of a tool - the later Transkribus - to automatically recognise handwritten texts. 'Making digitised and digital sources available is increasingly becoming a core element in many research projects.' (Romein et al. 2020, p. 294) Artificial Intelligence (AI) played and plays a significant role in 'unlocking the past' as Transkribus' slogan now reads.

Transkribus is one of the tools that can be trained to do handwritten text recognition (HTR). It is one of the tools, that, thanks to its 100,000+ users has gathered a massive amount of training material on their servers which allows the AI to 'learn' and 'adapt' to challenging handwriting and has decreased the amount of training data needed to create a model drastically. While in 2018 one needed around 60-80 pages of training material; the number has dropped to 30-50 pages (10.000 words) by the end of 2022. With the emergence of more basemodels the minimum of words can even be reduced to 6.000; expectations are that with the TransformerHTR-models the number might be much lower again.

Transkribus eXpert – the original desktop version of the tool – was used in 2015 for the first time. New features were integrated over the course of the years and allowed users to use various tools within the eXpert version: e.g. OCR (ABBYY), Text2Image, various HTR engines (HTR+ and Pylaia), and compare the results. Time has changed and most tools are now web-based; hence the transformation of Transkribus eXpert to Transkribus Lite.While some 'veterans' may have a preference for eXpert (desktop version) they too, will see the benefits of using Lite when they e.g. need to work with others on a project.

Transkribus Lite is – for most users – more intuitive and straightforward and – very important – much easier, as it does not require Java installation. It is therefore much easier to use within classes/ workshops and it is much easier to explain to students and volunteers alike. Furthermore, the number of buttons has been drastically reduced. Transkribus Lite will be *the* version to work with, as the eXpert-version is being phased out/ will not receive any feature updates anymore.

About the trainers

Two of the trainers are active users of Transkribus as researchers(-to-be), while the third organiser is affiliated with Transkribus (as an account manager Benelux) *and* uses Transkribus for a family project. During the training, the use of examples from various projects where they are involved in will give practical insights and understanding of the use within daily research-life. Transkribus will be presented from the users' point of view to ease the use; the presence of an account manager is meant to answer particular questions about the practical and financial implication of Transkribus in your (future) projects.

Workshop DHBenelux

This workshop consists of two parts. You can attend both parts, but you can also participate in just one of them, please have a thorough look at the program to fit your needs. The workshop will be a hands-on, data-driven workshop. The organisers warmly welcome the participants to bring their own images/ scans/ photos of archival or library materials; this is not limited to handwritten material as printed texts are welcomed too. However, if you do not have source material available or are not (yet) at liberty to upload them to the servers in Innsbruck, the organisers can provide you with some material to practice on.

Part 1. *Basic workshop* (3 hours)

No prior knowledge is needed. Do bring a laptop (or tablet) please.
We do assume that the participants in this workshop/ these workshops will create a user account at https://lite.transkribus.eu/ and bring (remember) their username and password.

00:01–00:20 hour	- Welcome and introductions
	  - what is your discipline
	  - what type of sources do you want to work with
	  - what period are you focussing on
	  - particular challenges in your sources

00:20-00:35 hour	READ COOP introduction
00:35-00:45 hour	Upload, servers and sensitivities - What happens when you upload your files (where are they?)

	- How do you do an upload? - File structure - Sensitive sources… what to do? - Changing names of your files?
00:45-01:00 hour	Practise uploading files + short break

01:00-01:15 hour	Lay-out recognition: the importance of LA
	- Textual structure and the computer
	  - Regions
	  - Baselines
	  - Polygones
	- How to do manual lay-out recognition?
	- How to do automatic lay-out recognition?
	- Training your own LA-model? (Advanced)

01:15-01:30 hour	Lay-out recognition: practise
01:30-01:45 hour	Lay-out: structure tags - Why structure your layout? - How to do it - Best practises
01:45-02:00 hour	Practice structure tags + break
02:00-02:15 hour	Transcribing your texts - Where to transcribe - Best practices with challenges (abbreviations etc.)
02:15-02:30 hour	Transcribing your texts - Adding your (manual) transcriptions
02:30-02:40 hour	Applying an existing model - Where do I find public models? - Will it work on my text (transkribus.ai)?
02:40-02:55 hour	Creating your own model - What is a base model? - What do I need to build my own model? - How do I ‘get it going’?
02:55-03.00 hour	Wrap up

Part 2. *Advanced workshop* (3 hours)

Prior knowledge of working with Transkribus is preferred for this part of the session, this could be with eXpert, or with Lite (the basic workshop provided in advance is sufficient). Do bring a laptop (or tablet) please.

03:00–03:10 hour	Outline of this workshop session/ introductions when needed
03:10-03:35 hour	Full text search/ Smart search
03:35-04:15 hour	Text to Image - If you have existing transcriptions (and images) but they are not yet combined within Transkribus. - Please bring your own if you have! Practise is possible.
04:15-04:30 hour	Break
04:30-04:45 hour	Textual tagging - E.g. entities. - Adding preferred entities - Practice
04:45-05:00 hour	Tabels - How to work with tables?
05:00-05:25 hour	Training your own LA-model? - Baselines - P2PaLa
05.25:05.30 hour	Short break
Annemieke: 05:30-05:55 hour	Working with a team or volunteers - How to add people to your collection? - Best practices with running volunteer projects
Joost: 05:30-05:55 hour	Read & Search (platform feature) - Showing some examples - Explaining benefits for projects
Bram: 05:30-05:55 hour	Questions about pricing and your projects (Q&A) - Fellowships - Projects and finances
05:55-06:00 hour	Wrap-up

On display	Scantent - What can you do with it? - Which apps can you use it with? (Testing it during breaks.)

Bibliography

ROMEIN, C.A., KEMMAN, M., BIRKHOLZ, J.M., BAKER, J., DE GRUIJTER, M., MEROÑO-PEÑUELA, A., RIES, T., ROS, R. and SCAGLIOLA, S. (2020), State of the Field: Digital History. History, 105: 291-312. https://doi.org/10.1111/1468-229X.12969