Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D1RP93

Title: Training HTR-Models for a Bilingual Digital Edition of the Ethica Complementoria

Creator: Annika Rockenberger - **ORCID:** 0000-0001-9515-8262

Affiliation: University of Oslo (uio.no)

Principal Investigator: Annika Rockenberger

Data Manager: Håvard Loeng, Tuva Kongshaug, Annika Rockenberger

Project Administrator: Annika Rockenberger

Contributor: Håvard Loeng

Funder: University Of Oslo, Teksthub+digital Humanities

Template: Digital Curation Centre

Project abstract:

We aim to create a dataset to be used as the basis for a bilingual (Danish/German) digital scholarly edition of one of the most popular books on ‘etiquette’ in early modern Germany and Northern Europe: the Ethica Complementoria.

Originally written in German, the book made its way to the Nordic region through the Danish translation from 1678. This first Danish print will be published in parallel with the German version used for the translation.

The transcription project is part of a larger project on the book and revision history of the Ethica Complementoria, led by Annika Rockenberger and will be conducted by Håvard Loeng. An overview of all editions has been published digitally at the Herzog August library: http://diglib.hab.de/ebooks/ed000738/start.htm.

Manual transcription of two 300+ page texts is not feasible anymore. However, traditional Optical Character Recognition (OCR) yields inferior results for older printed books. Therefore, we aim to test, evaluate, improve, and build upon the NorFraktur model from the National Library of Norway. NorFraktur is a public Handwritten Text Recognition (HTR) model in Transkribus. It was trained on the HTR algorithm developed by READ Coop to recognise manuscripts and older prints automatically.

The development project contributes to both a digital scholarly edition with open access (planned as part of the publications by the Norwegian Language and Literature Society at bokselskap.no) and to the improvement and expansion of an open HTR model that the scholarly community can reuse for early modern prints in Norwegian (including Danish and German).

Start date: 06-26-2023

End date: 10-31-2023

Last modified: 07-08-2024

Copyright information:

Training HTR-Models for a Bilingual Digital Edition of the Ethica Complementoria

Data Collection

What data will you collect or create?

An automatic transcription of the Danish print of the Ethica Complementoria from 1678.

The resulting data will be available in plain text format (txt) and semi-structured XML.

How will the data be collected or created?

The transcription uses the Read Coop Transkribus HTR model NorFraktur_1600_PyLaia created by the National Library of Norway. The scans of the physical print are uploaded into the Transkribus application. First, a layout recognition module is applied and manually corrected using the default baseline model from the app. Then, text recognition is performed for selecting 25 pages using the NorFraktur_1600_PyLaia model. The results are quality checked by the research assistant, and corrections are applied to the text within the app. The corrected text is then used to improve the HTR model using the training module in the Transkribus app. Then, the improved model will be applied to a different selection of 20 pages and checked for error rate. If the error rate could be improved significantly, the automatic transcription of the remaining 250+ pages is performed. If the error rate is too high, another round of manually correcting transcription errors and training the model for improvement is performed.

Documentation and Metadata

What documentation and metadata will accompany the data?

A human-readable README file will accompany the data set.

The XML file will be in a standardised dialect, Prima Page Content XML. The schema can be found at http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15 http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15/pagecontent.xsd.

Ethics and Legal Compliance

How will you manage any ethical issues?

There are no ethical issues known at this point.

The source data is published CC0; the HTR model is released publicly for re-use.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

There are no copyright or other IPR issues known. The source text is from 1678.

The created dataset will be licensed under CC0.

Storage and Backup

How will the data be stored and backed up during the research?

Read Coop Transkribus stores and hosts the scans of the source material after upload. Read Coop ensures backup of data in the app.

The collection is shared with the project's research assistant.

For the final transcription, a public GitHub repository will be created where the XML and the plain text files incl. documentation, will be stored and version controlled.

How will you manage access and security?

The project lead and the research assistant can access the collection on Read Coop's Transkribus app containing the source material and the transcription. The project lead is the collection owner; the assistant temporarily has the role of editor.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

The goal of the project is to make the dataset available long-term. This will be achieved by publishing the improved HTR model on Transkribus, archiving the XML and plain text files on Dataverse.no and preparing the XML file for further archiving in a different metadata standard (TEI P5) to be archived and made accessible by bokselskap.no.

The training data will not be kept; Read Coop owns them.

Researchers in the fields of literary studies, history, cultural studies, linguistics and digital humanities will be able to use and re-use the datasets.

What is the long-term preservation plan for the dataset?

Archiving on dataverse.no and publication as a digital scholarly edition using XML/TEI P5 on bokselskap.no. This should ensure the availability of the data for at least five years, most likely more.

Dataverse.no is free of charge for datasets created by researchers affiliated with the University of Oslo.

Bokselskap.no has agreed to publish the digital edition without costs to the project.

The data will also be available on GitHub in a public repository archived regularly on Zenodo.org.

Data Sharing

How will you share the data?

Openly, freely, and publicly on repositories like Dataverse.no, Bokselskap.no and Zenodo.org.

Sharing conditions are CC0 for the transcriptions and CC BY 4.0 for all commentary and additional texts by the project lead.

Are any restrictions on data sharing required?

No.

Responsibilities and Resources

Who will be responsible for data management?

The project lead is responsible for data management. The University of Oslo owns the project data and its outputs. Read Coop owns the machine learning algorithms for training HTR and baseline models.

What resources will you require to deliver your plan?

Access to Transkribus app for individual users: project lead and research assistant. (OK)

VSCode incl. extensions for XML and XSLT: project lead (OK), research assistant (to be done).

Preferably: Oxygen XML editor academic license. (Pending)

Shareable credits for Transkribus. (Pending)

Planned Research Outputs

Text - "Ethica Complementoria Digital Scholarly Edition – Redux. Series of blogposts"

Series of public blog posts on the website of the parent project "Georg Greflinger - Digitale Edition". These accompany the project: https://greflinger.hypotheses.org/.

In addition, microblogging will be done on the social network platform Mastodon via the project lead's personal account.

Dataset - "Ethica Complementoria. Danish print 1678"

XML file of the automatic transcription of the 1678 print. Transcribed with Read Coop's Transkribus, using the NorFraktur_1600_PyLaia public HTR model. Transcription quality-checked and corrected by Håvard Loeng, research assistant.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Ethica Complementoria Digital Scholarly Edition – ...	Text	2023-10-30	Open	None specified	1 MB	Creative Commons Attribution 4.0 International	None specified	No	No
Ethica Complementoria. Danish print 1678	Dataset	2023-10-30	Open	Dataverse.no Bokselskap.no	1 MB	Creative Commons Zero v1.0 Universal	None specified	No	No

Related Works

Articles

- https://greflinger.hypotheses.org/721
- hloeng. 2023. “Transcribing the 1st Danish Translation of the Ethica Complementoria and the Tranchierbuch from 1678.” [Article]. https://doi.org/10.58079/p4zm.
- Annika Rockenberger. 2023. “Automated Text Recognition with ChatGPT 4.” [Article]. https://doi.org/10.58079/p4zn.