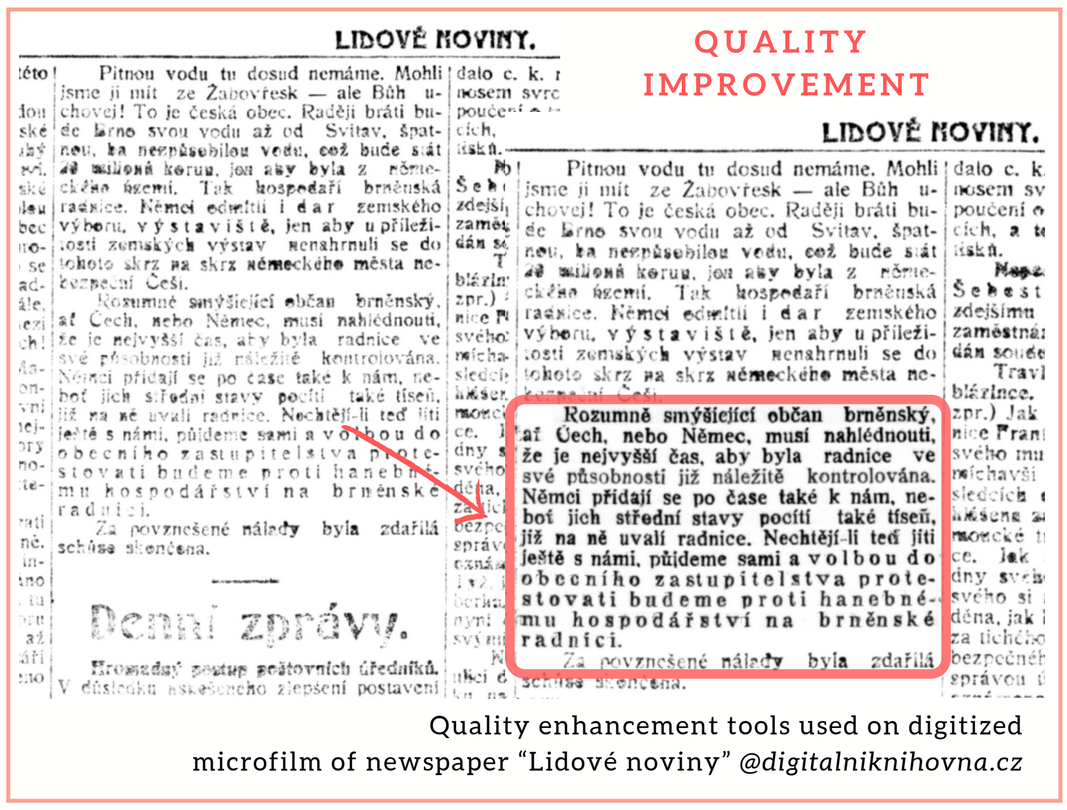
MORAVIAN LIBRARY

The Moravian Library is the second largest library in Czech Republic and a research organisation carrying out basic, applied and experimental research or software development in the field of digital libraries and cultural heritage **<labs.mzk.cz>**

LIMITS OF AUTHENTICITY OF DIGITIZED OBJECTS

Alžbeta Zavřelová - Petr Žabička, Moravian Library <Alzbeta.Zavrelova, Petr.Zabicka>@mzk.cz

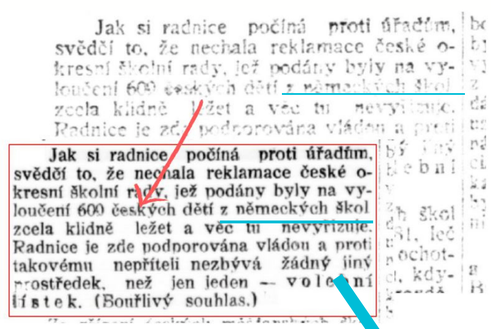


QUALITY IMPROVEMENT

Quality enhancement tools used on digitized microfilm of newspaper "Lidové noviny" *@digitalniknihovna.cz*

This poster discusses the limits of authenticity of digital objects and current possibilities of modifying digital historical documents using machine learning methods.
Our results provide us with tools that make us reconsider the importance of digital interpretation and processing of digitized objects.

Tools for automatic classification or full text indexing of the OCR often work with language models that may highly affect the content of the resulting text.



These pictures show the visual quality improvement by OCR of old newspapers and the possibility to make changes in the text (see below). *@digitalniknihovna.cz*

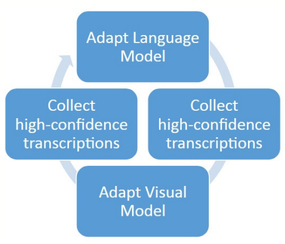
CRITICAL APPROACH TO THE DIGITAL COPY

Each collection is a subjective interpretation of the creator's view or ideological attitude. Digitized historical collections are believed to be credible by its nature. Intentional document forgery has always been around because of financial profit, privilege, power or influence gain. Even a small change in the text may produce the desired results for a certain group.

DIGITAL COPY = INTERPRETATION OF OBJECT

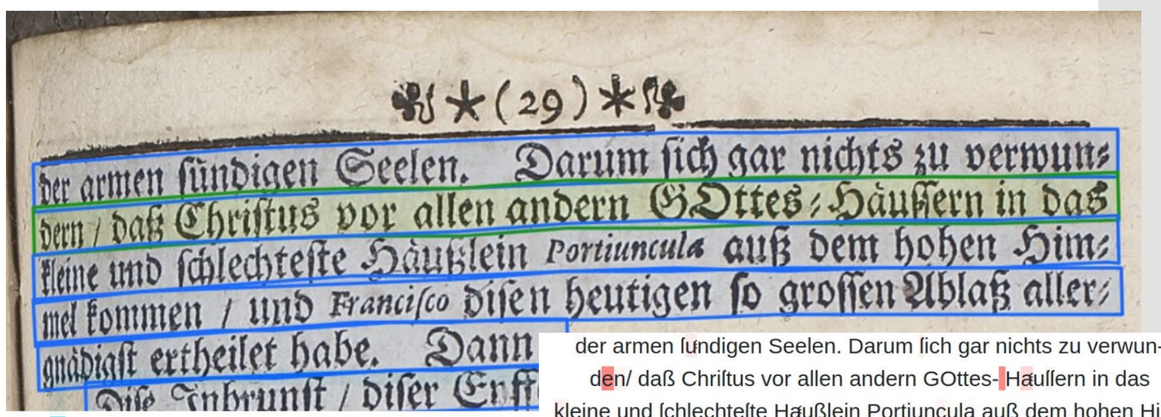
DIGITAL COPY ≠ ORIGINAL OBJECT

Many objects need to be modified for better accessibility or readability during digitization, it may include:



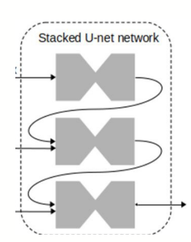
- special collections: curved pages flattening, unfolding text lines
- old medium copies: audio cleaning, image noise reduction
- quality enhancement
- OCR/HTR
- text reconstructions

How to modify historical documents?



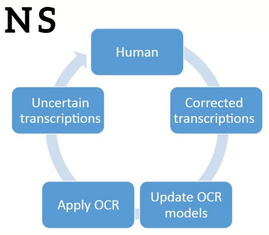
Web application used for text of early prints (-1860) *@PERO*

The **PERO - Advanced content extraction and recognition for printed and handwritten documents for better accessibility and usability** project aims to create tools to improve accessibility of digitized historic documents based on methods of machine learning (neural networks), computer vision and language modelling **<https://pero.fit.vutbr.cz/>** The results of the project will be integrated in large research infrastructure **Lindat/Clariah-CZ** - Digital Research Infrastructure for Language Technologies, Arts and Humanities.

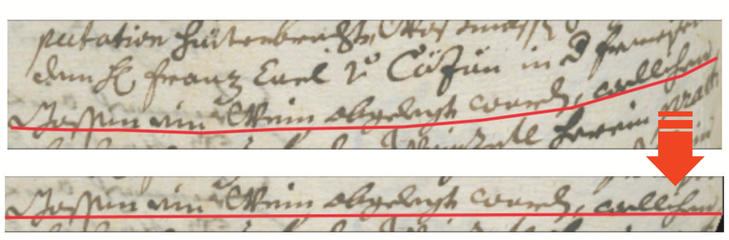


WEB APPLICATION FOR MANUAL CORRECTIONS

- text line extraction: detection tool (baseline and height)
- align lines with existed transcriptions
- text lines + transcriptions = dataset/model
- manual error corrects - OCR model adaptations

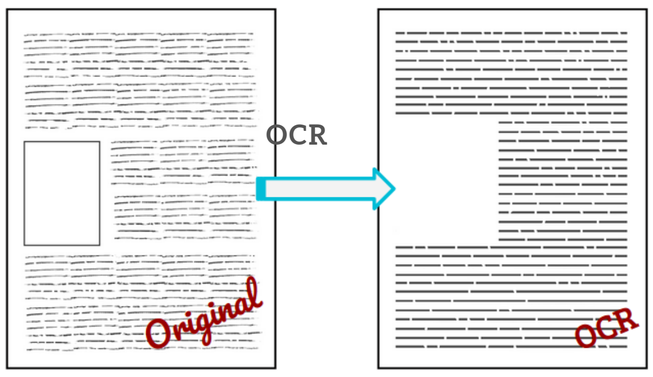


BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

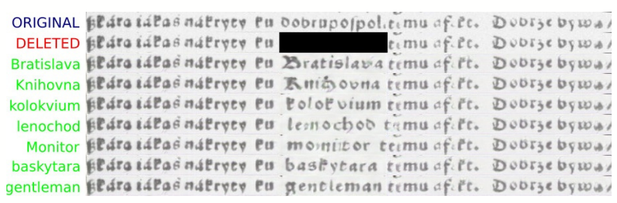


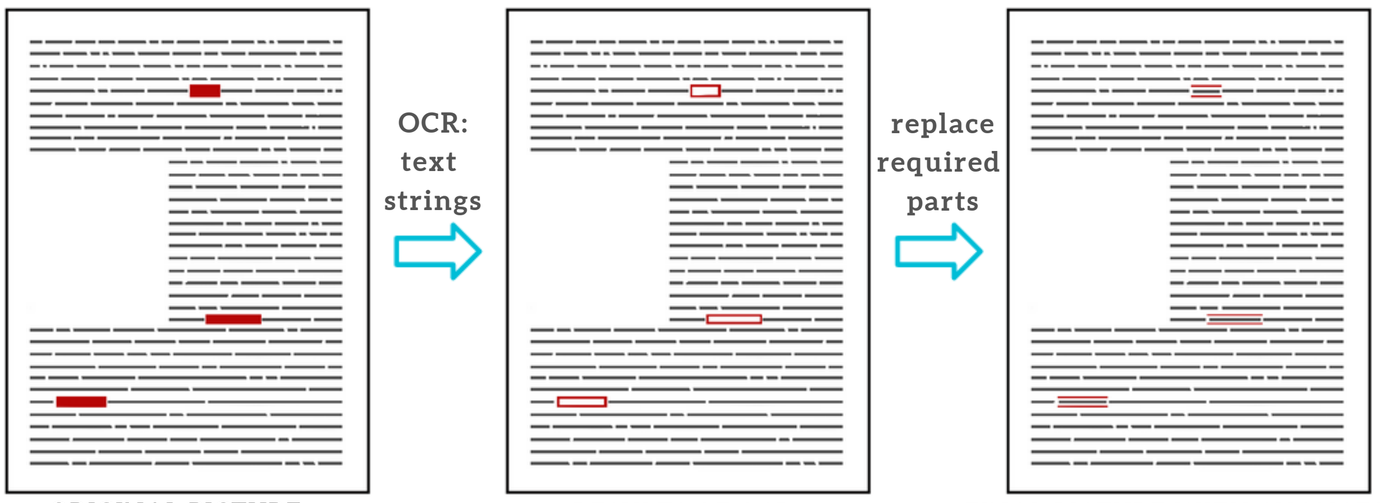
Automated textline unfolding of crooked hand-written text @PERO

THE METHOD OF TEXT MANIPULATION BASED ON GANS @PERO



An example of text manipulation based on Generative Adversarial Networks @PERO





ORIGINAL PICTURE with manual deletion

RESULTS OF OCR text with empty spaces

FINAL PICTURE WITH CHANGES

ACKNOWLEDGEMENTS

This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic (LINDAT/CLARIAH-CZ: programme of Large Infrastructure for Research, Experimental Development and Innovation, LM2018101, 2019-2022) and by the Ministry of Culture of Czech Republic (PERO:NAKI II, DG18P02OVV05, 2018-2022)

DOWNLOAD

You can easily download all applications and tools from the PERO GitHub: <https://github.com/DCGM> **INTERESTED TO GET INVOLVED?** We are still looking for new partner institutions that can provide us their digitized documents with transcriptions we can use to improve our learning models.