Digital humanities in libraries

Jiří Dufka – Alžbeta Zavřelová

Moravská zemská knihovna v Brně

Content

- What do metadata in the libraries look like?
- How to read them?
- How useful are they?
- Some tips for advanced search in digital libraries
- How can digitized texts be recognized?
- Are you able to use such tools?

Metadaten der Bücher(ausgaben)

Deskriptive Metadaten in den Bibliotheken

- Struktur der alten Zettelkatalogen als Grundlage

- ermöglichen das Suchen in Katalogen
- ermöglichen das Suchen und Präsentation in digitalen Bibliotheken
- unterstützen eine Reihe spezieller Datenbanken
- bilden einen der größten Korpusse von den Open Data überhaupt

Suchen der Digitalisate

Katalog = Suchdienste,

Digitale Bibliothek = Blättern, Nutzung der Bilddaten

im Katalog Suchen und in der digitaler Bibliothek (leider immer nur) blättern

Suchen:

- im Verbundkatalog
- in Online Bibliographien
- in der digitalen Bibliothek

Verbundkataloge

- Auschnitte der Aufnahmen aus den lokalen Katalogen.
- ohne lokalspezifischen Angaben (z. B. Provenienz usw.)
- Karlsruher Virtueler Katalog – effizienter Verbundkatalog der Verbundkataloge (zeigt auch die digitale Ergebnisse)
- Souborný katalog ČR – klassischer Verbundkatalog der großen tschechischen Bibliotheken
- knihovny.cz – tschechischer Facetten- und Verbundkatalog, kann auch die Digitalen Bibliotheken zusammen mit den Bibliothekskatalogen durchsuchen

Suchen der Digitalisate in Online Bibliographien

- beste Aufnahmen für die Auflage, angeknüpfte Digitalisate
- enthalten keine exemplar-spezifischen Daten (Provenienz, Einband usw.)
- deutsche Drucke
  - Bücher vor 1800 aus dem deutschen Sprachraum: VD16, VD17, VD18,
  - deutschsprachige Liederdrucke: VD-Lied,
  - Zeitschriftendatenbank
- tschechische Drucke
  - auf tschechisch gedruckte Bücher vor 1800: Knihopis,
  - nicht tschechischsprachigen Bohemica: BCBT,
  - knihoveda.cz – Facettenkatalog KPS+BCBT,
  - ČNB: Tschechische Nationalbibliographie nach 1801 (in Tschechien gedruckte Bücher)
- Incunabula: ISTC (alle Sprachen zusammen)

Suche der Digitalisate in Tschechien

digitalniknihovna.cz

- gemeinsame Oberfläche für verschiedene digitale Bibliotheken, deren Inhalt ist nicht identisch
- es gibt keine gemeinsame Schnittstelle für den Zugang zu den digitalisierten Büchern (vom Urheberrecht geschützte Bücher können nur der Bibliothek veröffentlicht, die sie im Bestand hat)
- bessere Zugänglichkeit der Texte für angemeldete LeserInnen
- Registrdigitalizace.cz – hilft den Bibliotheken, dupplizite Digitalisation zu vermeiden, kann aber als Wegweiser auf die Digitalisate in verschidenen Bibliotheken dienen (auch für Periodika)

Handschirften und Drucke vor 1800

- Manuscriptorium.cz - Aggregator der Digitalisate und Aufnahmen mit einer Schnittstelle für die personalisierte Bearbeitung der ausgewählten Bilder (nad der Anmeldung). Ursprünglich nur für Handschriften bestimmt

Spezialisierte Datenbanken

- Kalliope - Nachlässe, Autographen, Verlagsarchive
- Provenio - Datenbais zur Provenienzforschung

Kartographische Visualisation

- https://mapa.knihoveda.cz/ - tschechische Druckproduktion bis 1800
- Kramarsketisky.mzk.cz - tschechische Lieddrucke (in Bau)

Kartensuche

- Oldmapsonline.org

Text recognition technology at the Moravian Library

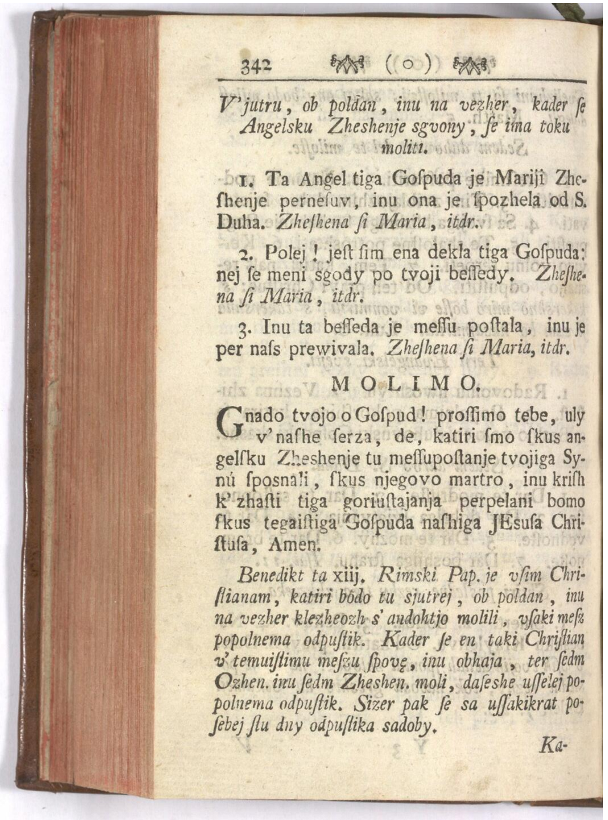
Digitisation

Mass digitisation of library collections has become a necessity in memory institutions all around the world.

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

- simple interventions
  - brightness, contrast, image cropping or stitching, noise reduction, ..
- innovative digital tools (using machine learning methods)
  - quality improvement, OCR,
  - flatten curved pages and unfolding text lines, scan text in narrow bookbinding,
  - OCR/HTR text reconstructions, ...

What is OCR/HTR?



342 (o)

V' jutru, ob poldan, inu na vezher, kader ſe Angelsku Zheshenje sgvony, ſe ima toku moliti.

1. Ta Angel tiga Goſpuda je Mariji Zheſhenje perneſuv, inu ona je ſpozhela od S. Duha. *Zheſhena ſi Maria, itdr.*

2. Polej! jeſt ſim ena dekla tiga Goſpuda: nej ſe meni sgody po tvoji beſſedy. *Zheſhena ſi Maria, itdr.*

3. Inu ta beſſeda je meſſu poſtala, inu je per naſs prewivala. *Zheſhena ſi Maria, itdr.*

MOLIMO.

Gnado tvojo o Goſpud! proſſimo tebe, uly v' naſhe ſerza, de, katiri ſmo ſkus angelſku Zheshenje tu meſſupoſtanje tvojiga Synú ſposnali, ſkus njegovo martro, inu kriſh k' zhaſti tiga goriuſtajanja perpelani bomo ſkus tega iſtiga Goſpuda naſhiga JEsuſa Chriſtuſa, Amen.

Benedikt ta xiij. Rimski Pap. je vſim Chriſtianam, katiri bódo tu sjutrej, ob poldan, inu na vezher klezheozh s' andohtjo molili, vſaki meſz popolnema odpuſtik. Kader ſe en taki Chriſtian v' temuiſtimu meſzu ſpovę, inu obhaja, ter ſedm Ozhen. inu ſedm Zheshen. moli, daſeshe uſſelej popolnema odpuſtik. Sizer pak ſe sa uſſakikrat poſebej ſlu dny odpuſtika sadoby.

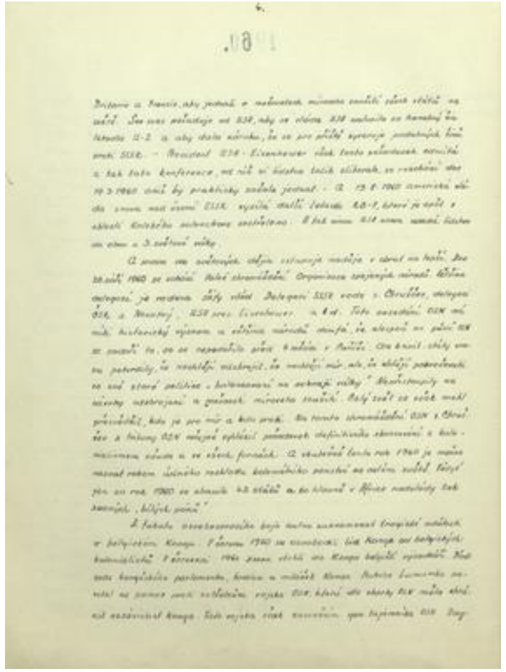
Ka-

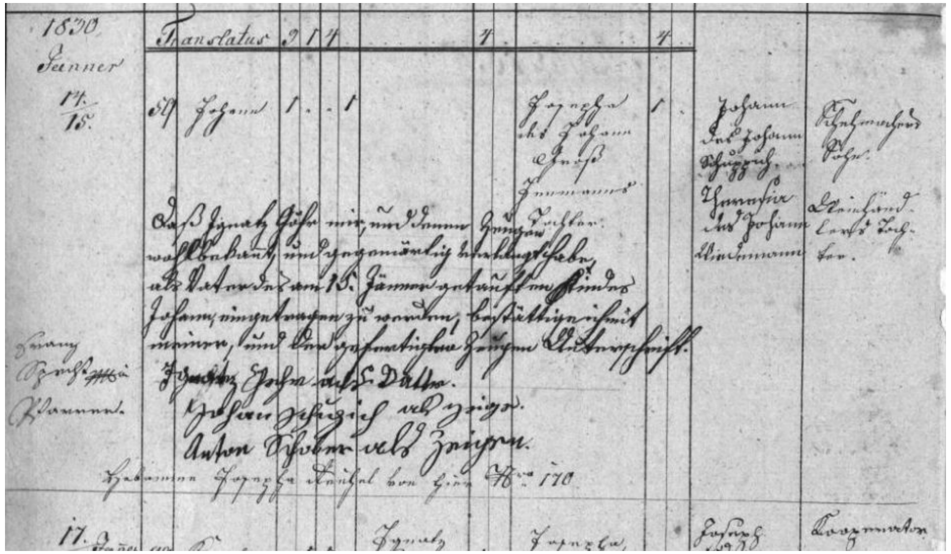
LIDOVE NOVINY

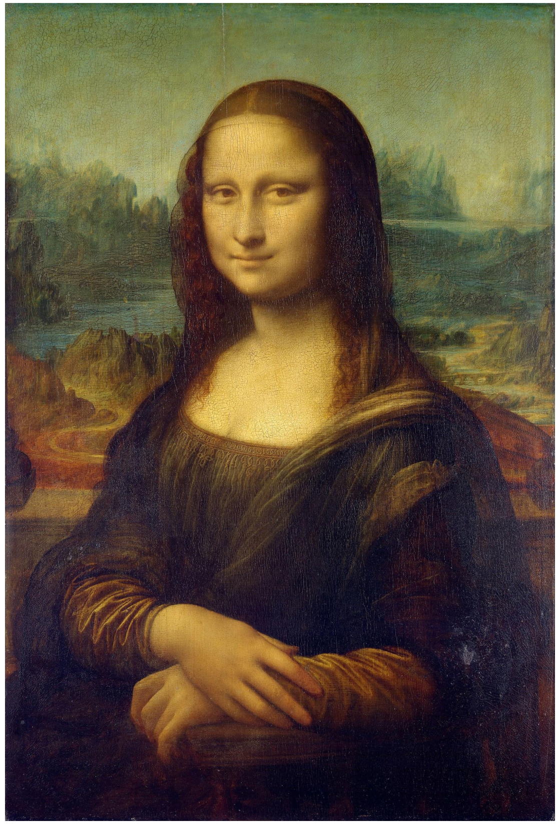
(Dříve „Moravské Listy" a „Pozor".)

...je o 6. stranách.

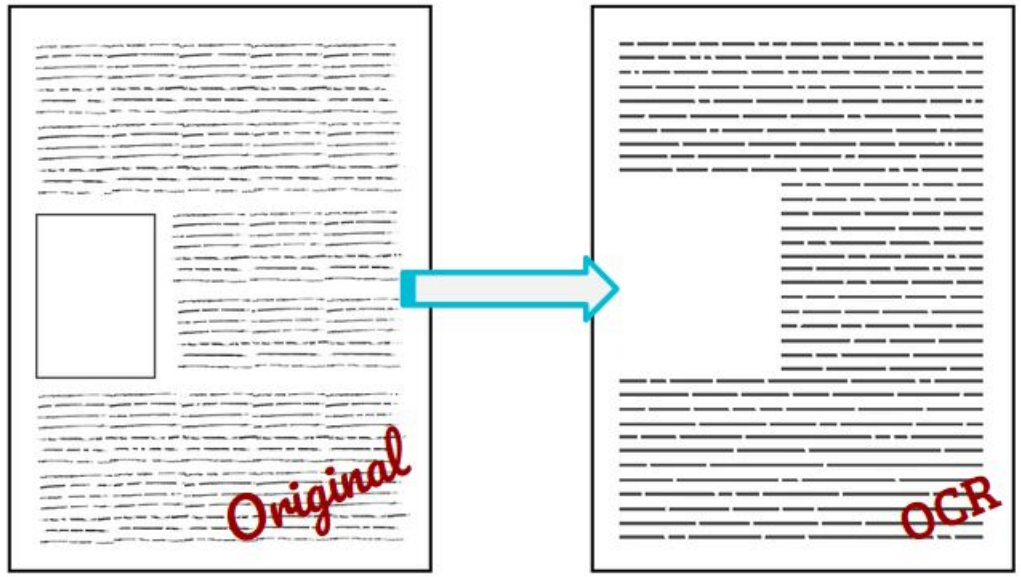
...romluvil.







What is OCR/HTR?



OCR = Optical Character Recognition

HTR = Handwritten Text Recognition

- conversion of image data into machine-readable text

OCR: scans of **printed** documents

- modern books, magazines, prints of digital documents etc.
- old newspaper, early printed books, ...

HTR: scans of **handwritten** documents

- modern handwriting
- rare books and manuscripts - historical libraries and archival collections

Limits of OCR

bad quality scans

- digitized microfilms of several newspaper titles (unique)

historical collections

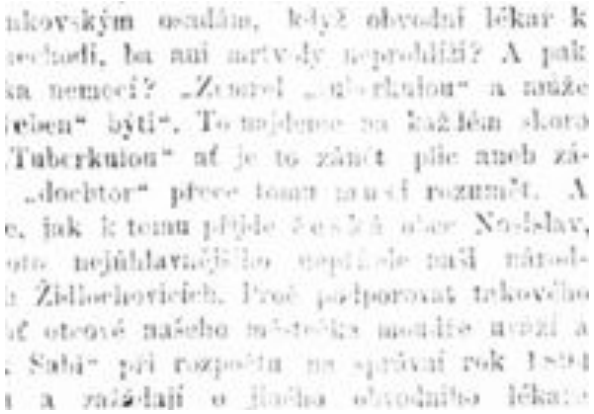
- early prints, incunabula, manuscripts, old maps
- special scripts/font

cultural specifics

- Czech language and diacritical mark

better tools → better OCR

good known OCR - [ABBYY]
good known HTR - [Transkribus]



PERO project

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**. (2018-2022)

inspiration - Adam Matthew: Colonial America

- aims:
  - automatic quality enhancement of damaged documents
  - automatic OCR for older prints (inc. early printed books)
  - semi-automatic HTR (handwritten documents, inc. historical manuscripts)

BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

MORAVIAN LIBRARY

PERO project

Collected datasets:

- ML digital library - 608 books up to y1820 (69.315 pages)
- czech prints - low quality periodicals (microfilms)
- IMPACT - european early prints (9 languages, 27.373 pages)
- Deutsches Textarchiv (6.000 prints)
- Bentham manuscripts (21.403 pages)
- Czech correspondence from the 20th century (2.000 letters)
- other digital libraries and archives

The results of the project will be integrated in large research infrastructure Lindat/Clariah-CZ - Digital Research Infrastructure for Language Technologies, Arts and Humanities.

Language models

Language models highly affect the content of OCR results - improve visual recognition by strong language knowledge

Unicorns from France likes to visit libraries.

vs

Unicorns from France like to visit libraries.

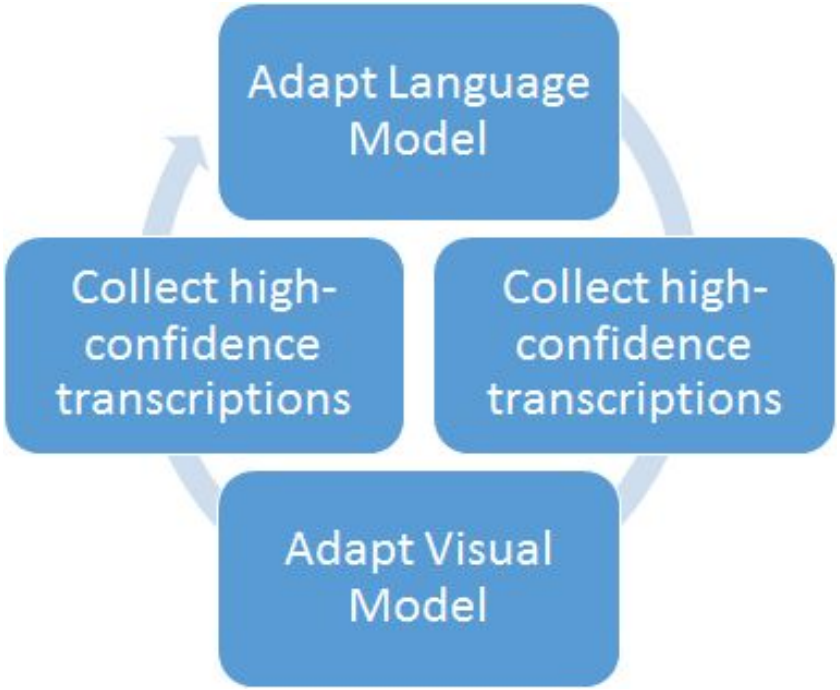
What is the next likely word?

People like to go to …… . (library, cinema, pubs, restaurants, work?)

Language models

Adapt visual and language models to a specific document:

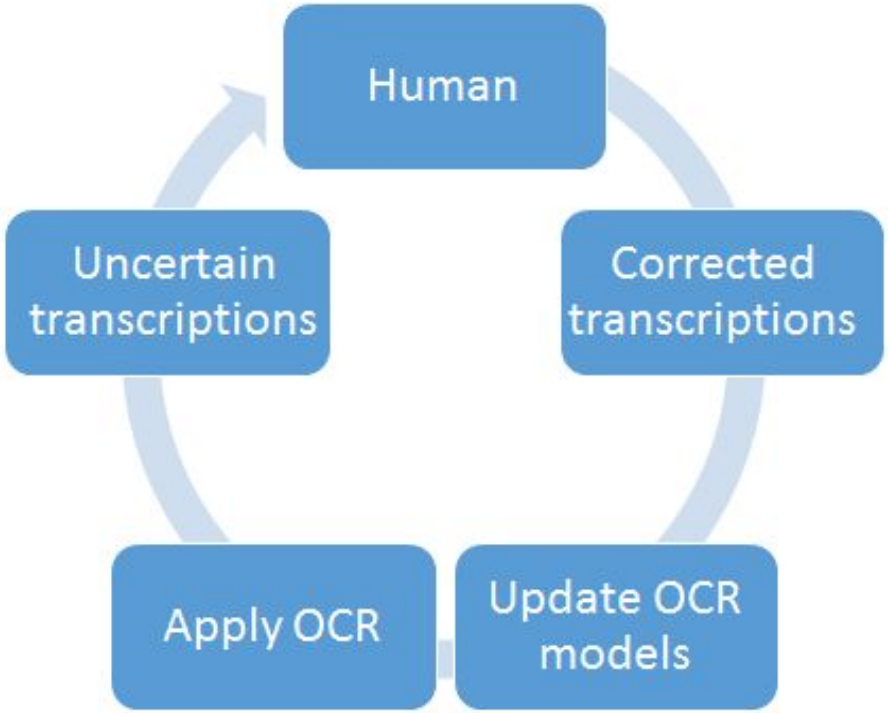
- Start with general models trained on large set of documents
- Adapt models to new document/writer/language
- Automatically adapt visual model by information from language model
- Automatically adapt language model by information from the visual model



Language models

Humans can help with adaptation :

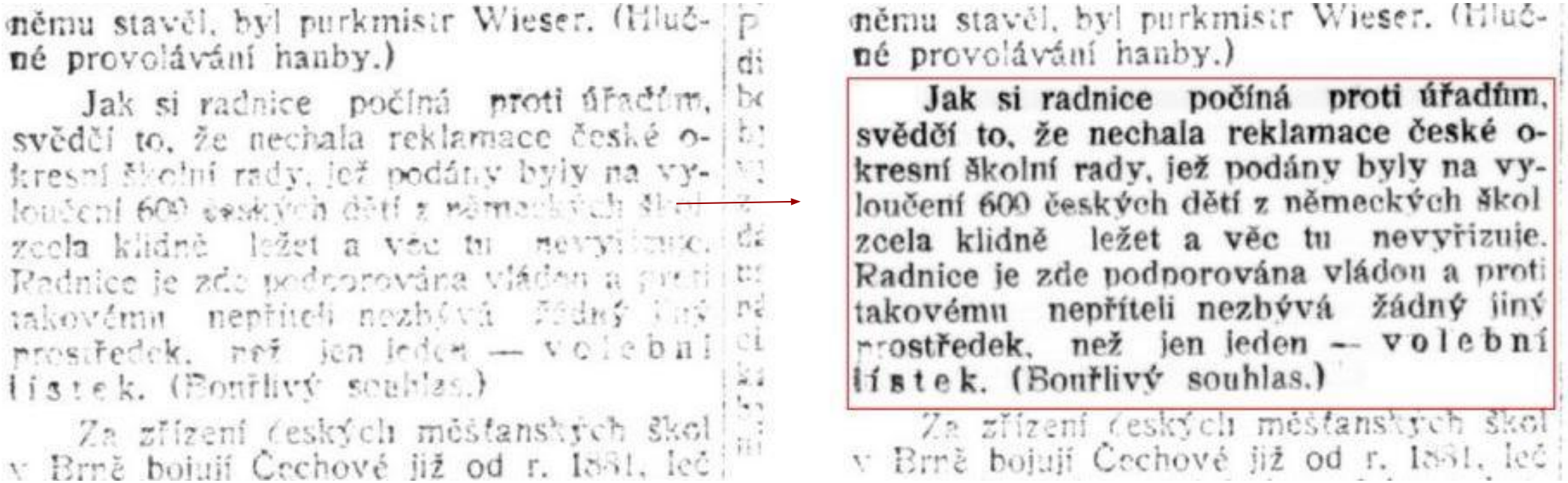
- For out-of domain documents or high accuracy
- System processes new documents
- Uncertain or poor quality text lines are selected
- Uncertain characters are highlighted
- Human annotators correct transcription mistakes
- Corrected transcriptions are used to adapt the system and improve recognition



PERO-OCR

1) **automatic quality enhancement of damaged documents**

- quality improvement of digitized documents (eg. on microfilms)



němu stavěl, byl purkmistr Wieser. (Hluč-
né provolávání hanby.)

Jak si radnice počíná proti úřadům,
svědčí to, že nechala reklamace české o-
kresní školní rady, jež podány byly na vy-
loučení 600 českých dětí z německých škol
zcela klidně ležet a věc tu nevyřizuje.
Radnice je zde podporována vládou a proti
takovému nepříteli nezbývá žádný jiný
prostředek, než jen jeden — volební
lístek. (Bouřlivý souhlas.)

Za zřízení českých měšťanských škol
v Brně bojují Čechové již od r. 1881, leč

before

němu stavěl, byl purkmistr Wieser. (Hluč-
né provolávání hanby.)

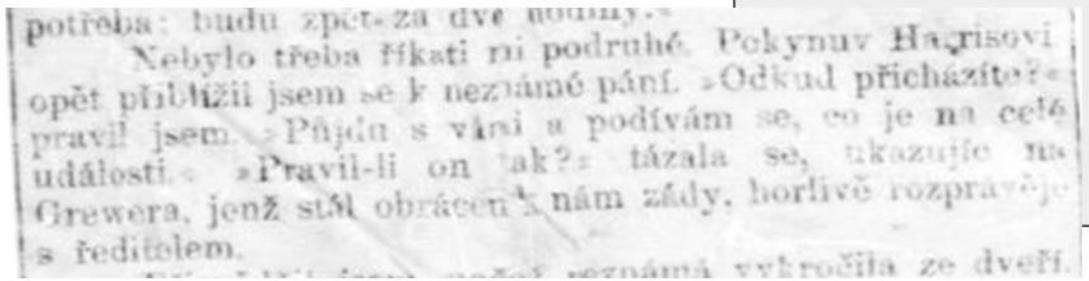
Jak si radnice počíná proti úřadům,
svědčí to, že nechala reklamace české o-
kresní školní rady, jež podány byly na vy-
loučení 600 českých dětí z německých škol
zcela klidně ležet a věc tu nevyřizuje.
Radnice je zde podporována vládou a proti
takovému nepříteli nezbývá žádný jiný
prostředek, než jen jeden — volební
lístek. (Bouřlivý souhlas.)

Za zřízení českých měšťanských škol
v Brně bojují Čechové již od r. 1881, leč

after

PERO-OCR

- *low-quality print example*



ABBYY	-livlo i ludi i |H>d nl.é l'- V r. Haj-isov; přichazít-: ■i i- ii« ceh dál osti i i Pra vil- nž stál , nera l
PERO	Nebylo třeba říkati ni podruhé. Pokynuv Hagrisovi opět přiblížil jsem se k neztámé pání. »Odkud přicházíte?« pravil jsem. »Půjdu s váni a podívám se, co je na celé události.« »Pravil-li on ak? tázala se, ukazujíc na Grewera, jenž stál obrácenk nám zády, horlivě rozprávěje s ředitelem.

PERO-OCR online application

PERO OCR demonstration application

This application demonstrates capabilities of pero-ocr python package developed in project PERO at Brno University of Technology.

You can watch videos demonstrating document management, page layout editting and text recognition, correction and review.

The application allows users to automaticaly transcribe several types of printed and handwritten documents. The provided OCR engines are able to transcribe even very low-quality printed documents in most european languages including Latin, old documents in Fraktur and similar scripts in German and Czech, and handwritten documents mainly in Czech language.

The application provides efficient interface for text corrections and several formats of transcriptions for download (ALTO, PAGE XML, plain text). Be aware that the images and corrections you provide may be used for further training of our systems.

make your account on:

pero-ocr.fit.vutbr.cz

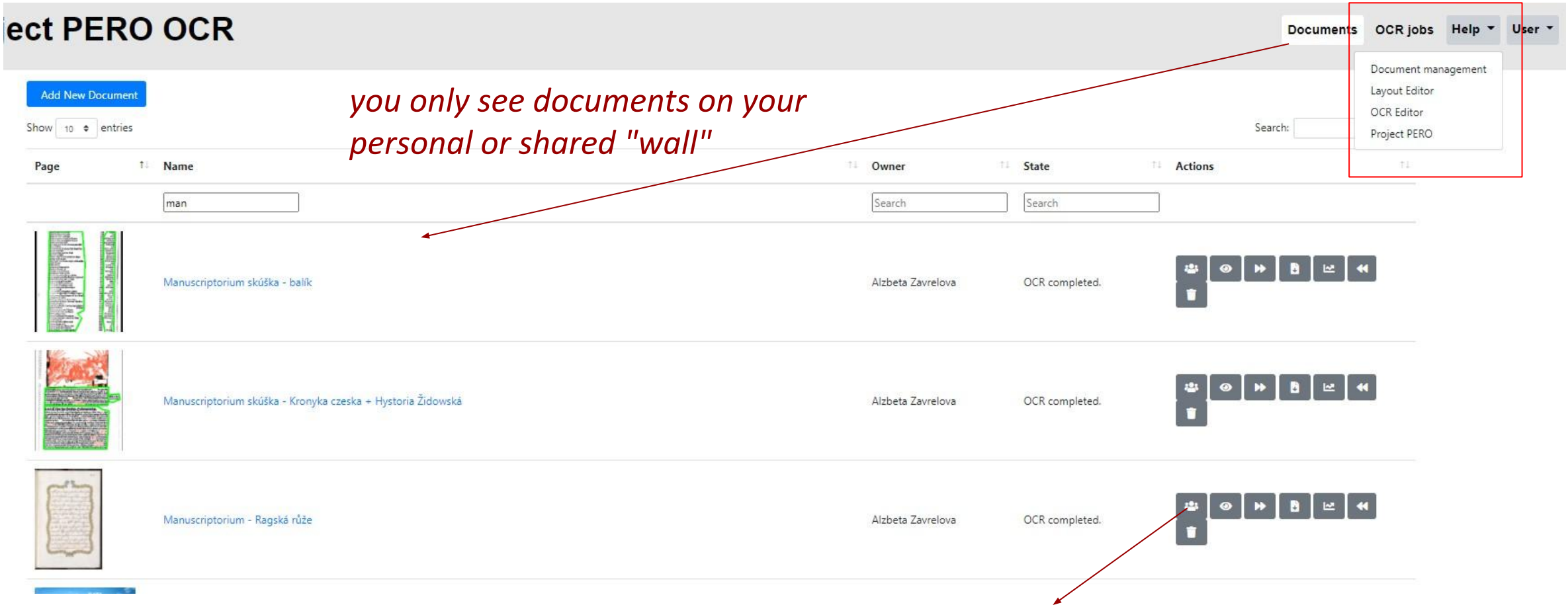
MORAVSKÁ ZEMSKÁ KNIHOVNA

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

MINISTERSTVO KULTURY

PERO-OCR online application

how to use the application step by step



you only see documents on your personal or shared "wall"

"Edit collaborators" you can share your documents with other registered users

PERO-OCR

2) automatic OCR for older prints (inc. early printed books)

- old newspaper (mixed fonts)
- text recognition of Fraktur or antikva fonts
- PERO for old prints

3) semi-automatic HTR = OCR for handwritten documents, inc. historical manuscripts (european)

- modern handwritten docs (try yours!)
- documents from the 20th century (chronicles, ..)
- historical manuscripts - Kurent script (experimental - charters, manuscripts, ..)

article: Projekt PERO – OCR pro historické texty

How to modify historical documents?

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.

90s - attention to possibilities of open digital images misuse, which has expanded significantly with technology development

Critical approach to the digital copy - each collection is a subjective interpretation of the creator's view or ideological attitude (digital copy = interpretation of an object!)

Digitized historical collections are **believed to be credible by its nature**

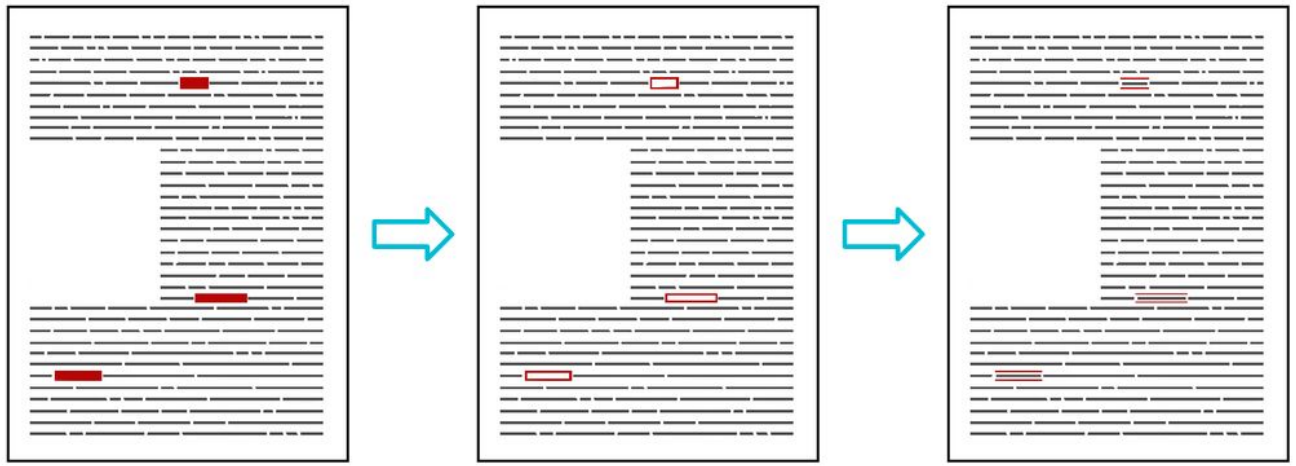
→ tools to make us reconsider the importance of digital interpretation and processing of digitized objects

How to modify historical documents?

The method of text manipulation based on Generative Adversarial Networks @PERO-OCR

find problematic part → text reconstruction → FIX the image using reconstructed text

- changes are (almost) not visible on the final picture



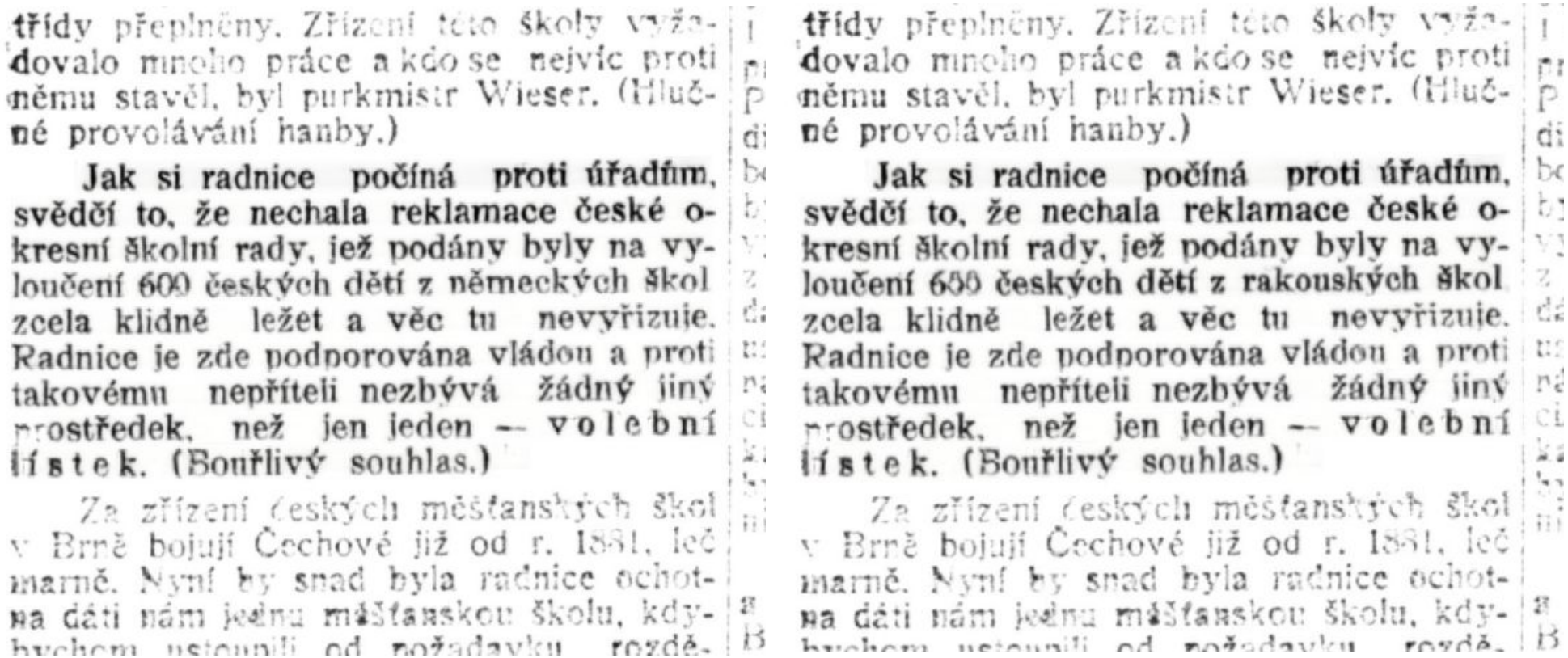
original picture
with manual deletion

results of OCR
text with empty spaces

final picture
with changes

How to modify historical documents?

Can you find change in the text?



třídy přeplněny. Zřízení této školy vyžadovalo mnoho práce a kdo se nejvíc proti němu stavěl, byl purkmistr Wieser. (Hlučné provolávání hanby.)

Jak si radnice počíná proti úřadům, svědčí to, že nechala reklamace české okresní školní rady, jež podány byly na vyloučení 600 českých dětí z německých škol zcela klidně ležet a věc tu nevyřizuje. Radnice je zde podporována vládou a proti takovému nepříteli nezbývá žádný jiný prostředek, než jen jeden — volební lístek. (Bouřlivý souhlas.)

Za zřízení českých měšťanských škol v Brně bojují Čechové již od r. 1881, leč marně. Nyní by snad byla radnice ochotna dáti nám jednu měšťanskou školu, kdy-

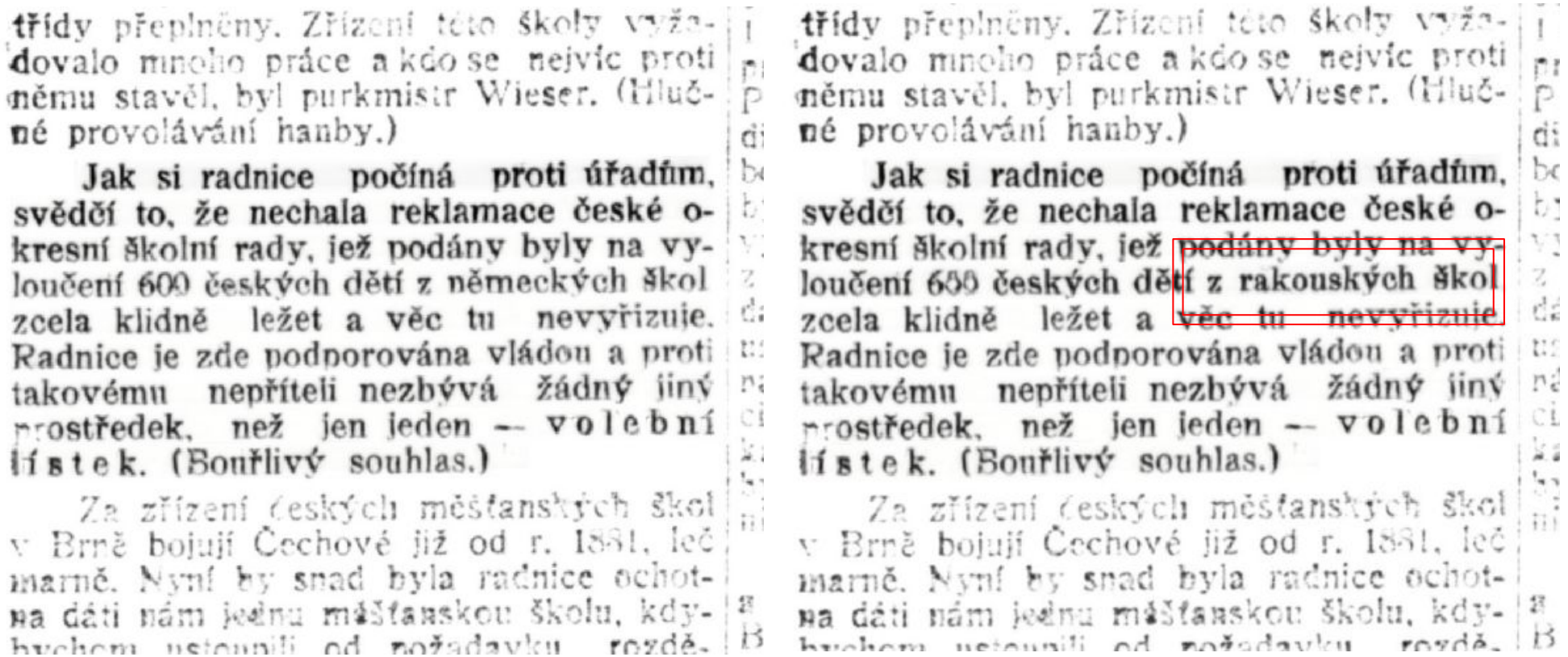
třídy přeplněny. Zřízení této školy vyžadovalo mnoho práce a kdo se nejvíc proti němu stavěl, byl purkmistr Wieser. (Hlučné provolávání hanby.)

Jak si radnice počíná proti úřadům, svědčí to, že nechala reklamace české okresní školní rady, jež podány byly na vyloučení 600 českých dětí z rakouských škol zcela klidně ležet a věc tu nevyřizuje. Radnice je zde podporována vládou a proti takovému nepříteli nezbývá žádný jiný prostředek, než jen jeden — volební lístek. (Bouřlivý souhlas.)

Za zřízení českých měšťanských škol v Brně bojují Čechové již od r. 1881, leč marně. Nyní by snad byla radnice ochotna dáti nám jednu měšťanskou školu, kdy-

How to modify historical documents?

Can you find change in the text?



třídy přeplněny. Zřízení této školy vyžadovalo mnoho práce a kdo se nejvíc proti němu stavěl, byl purkmistr Wieser. (Hlučné provolávání hanby.)

Jak si radnice počíná proti úřadům, svědčí to, že nechala reklamace české okresní školní rady, jež podány byly na vyloučení 600 českých dětí z německých škol zcela klidně ležet a věc tu nevyřizuje. Radnice je zde podporována vládou a proti takovému nepříteli nezbývá žádný jiný prostředek, než jen jeden — volební lístek. (Bouřlivý souhlas.)

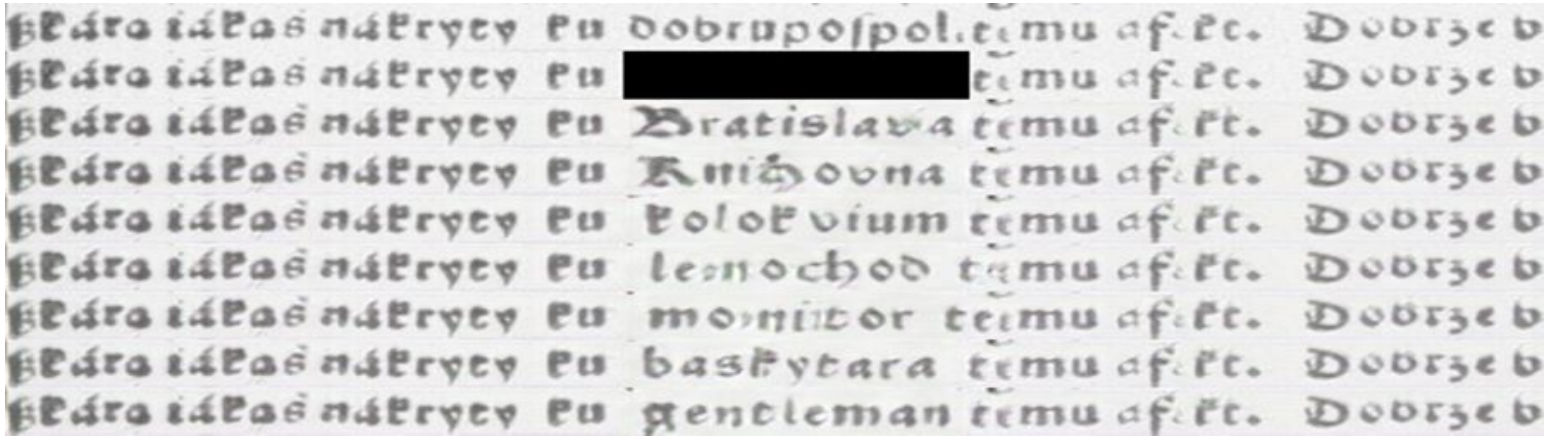
Za zřízení českých měšťanských škol v Brně bojují Čechové již od r. 1881, leč marně. Nyní by snad byla radnice ochotna dáti nám jednu měšťanskou školu, kdy-

třídy přeplněny. Zřízení této školy vyžadovalo mnoho práce a kdo se nejvíc proti němu stavěl, byl purkmistr Wieser. (Hlučné provolávání hanby.)

Jak si radnice počíná proti úřadům, svědčí to, že nechala reklamace české okresní školní rady, jež podány byly na vyloučení 600 českých dětí z rakouských škol zcela klidně ležet a věc tu nevyřizuje. Radnice je zde podporována vládou a proti takovému nepříteli nezbývá žádný jiný prostředek, než jen jeden — volební lístek. (Bouřlivý souhlas.)

Za zřízení českých měšťanských škol v Brně bojují Čechové již od r. 1881, leč marně. Nyní by snad byla radnice ochotna dáti nám jednu měšťanskou školu, kdy-

How to modify historical documents?



New technologies OPEN new questions we need to solve..

- How can we prove the authenticity of our documents?
- How can we guarantee our digital collections are credible for the future?

contact

Moravian Library in Brno: www.mzk.cz/en

Jiří Dufka

head of department

Manuscripts and early printed books

Jiri.Dufka@mzk.cz

Alžbeta Zavřelová

research assistant - PERO project

Innovation unit (research and development)

Alzbeta.Zavrelova@mzk.cz