

# 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 Digitalate

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 Virtueller 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 verschiedenen Bibliotheken dienen (auch für Periodika)

## Handschriften und Drucke vor 1800

- Manuscriptorium.cz - Aggregator der Digitalisate und Aufnahmen mit einer Schnittstelle für die personalisierte Bearbeitung der ausgewählten Bilder (nach 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](https://kramarsketisky.mzk.cz/) - tschechische Lieddrucke (in Bau)

## Kartensuche

- [Oldmapsonline.org](https://oldmapsonline.org/)



# Text recognition technology at the Moravian Library



# Digitisation

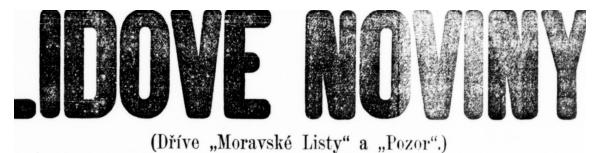
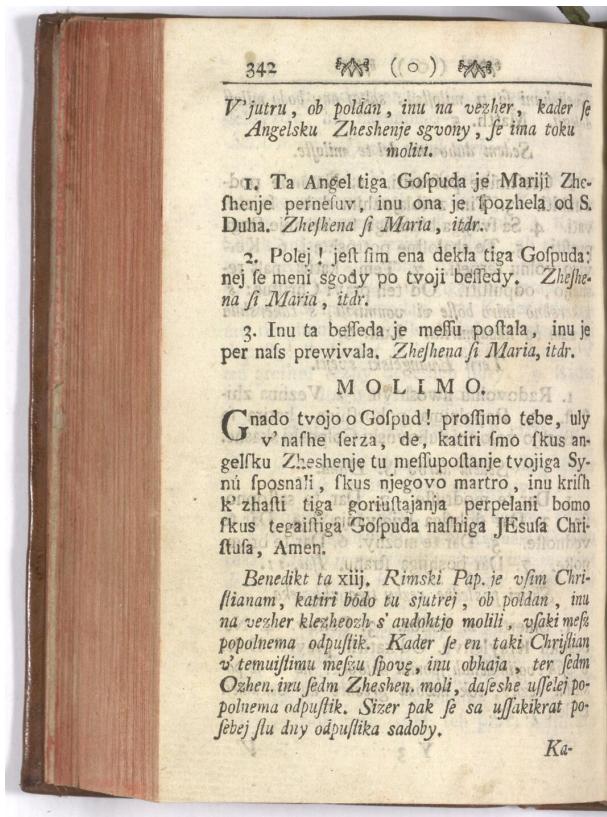
Mass digitisation of library collections has become a necessity in memory institutions all around the world.

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?



(Dříve „Moravské Listy“ a „Pozor“.)

in je o 6. stranách

*V'jutru, ob pol'dan, inu na vezher, kader se  
Angelsku Zheshenje sgvony, se ima toku  
valim si molit.*

— Ta Angel tiga Gospuda je Mariji Zhe-  
shenje perneluv, inu ona je spozhela od S.  
Duha. *Zhešena si Maria*, itdr.

2. Polej! jest ſim ena dekla tiga Gospuda:  
nej ſe mení ſgody po tvoji běſedy. *Zheſte na ſi Maria*, itdr.

# M O L I O, O M I L O,

**G**nado twojo o Gospud! prossimo tebe, uly v' nafshe serza, de, katiri smo skus angelsku Zhesherenje tu meflupostanje twojiga Synospalni, skus njegovo marto, inu krish k' zhasiti tiga gorušljajanja perpelani bomo skus tegatiliga Golpuda nafshiga JEsuša Christusa, Amen.

Benediktia xiiij. Rimski Pap. je všim Christianam, kاتri bodo tu sjurej, ob poldan, inu na vezher klezheozb z audihoz motili, uskati mesz popolnemu odpustit. Kader je tu takz Christian z temujsimu meszu sproe, inu obhaja, ter sedm Ozhen, inu sedm Zheshen, mojh, daleske usselej popolnemu odpustit. Sizer pak se sa usfakirat posjebej slu dny odpustka sadoby.

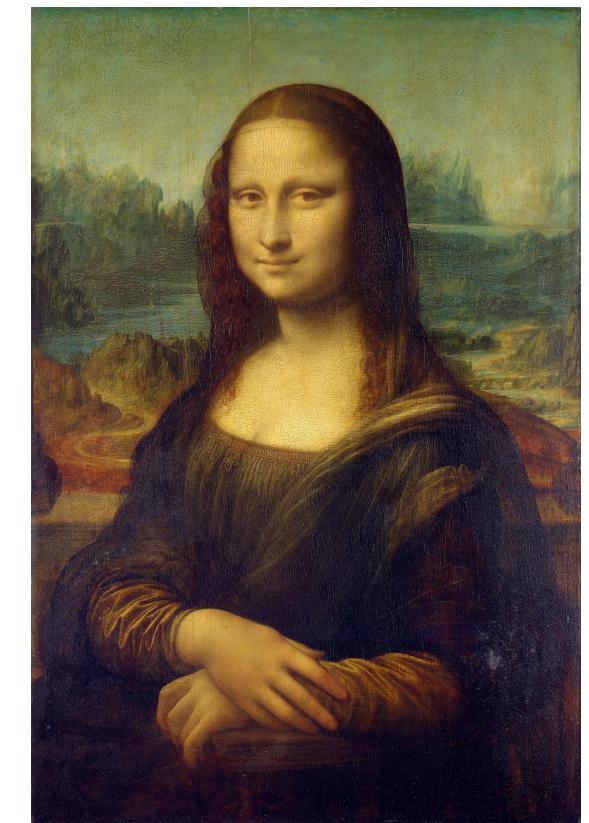
K<sub>a</sub>-

1820.	Translatus	214	7	4	
Ganner					
17. 18.	etij. Jezuist. 1...1		Jezuist. Jesuistum Jesuistis Jesuistis	Jezuist. Jesuistum Jesuistis Jesuistis	Polytheist. Polytheist. Polytheist. Polytheist.
	Das Blaue Gold mit dem kleinen Grünem Grünem Gold. wollt blau und grünemöndig herleitungsgebend als das andere am 15. Januar gebünt und gefündet Jefaminius und Antonius und Petrus und Paulus minor, inn dem unerträglichen Jezuistengewicht. Jezuist. Epoca vel Dulta. Jezuist. Jezuist. vel Jezuist. Antonius vel Zmigrod.				
	Zmigrod von Jan 170.				
17.9.		Lyneth	Fazula	Ezra	Laevanion

Doprava a finanční aktivity mimořádného významu zahrnuje plánované a užité. Sice jsou výdaje o 33,6 mil. za období 1958-1960 určeny na hradby a hrany 2,2, ale zdejší nároky, když se jedná o plánované výdaje, jsou jenom 33,6. - Prezentace 238 - Lisenovská vlna tedy všechny výdaje, z nichž tato konference, než nás v hledání také vloží, se nachází v rozmezí 19,7-20,6 mil. byly pořízeny údajně jednotkou. - G. Z. 19.6.1960 zaznamenává zprávu nad rukou ELLA, vyznačuje jednotce 40-7, které je potřeba aby vložila kontaktní adresy jednotlivých občanů. A pak nás ELLA zase naznačuje, že obec je 3. nejdůležitější.

A mons uti videretur, adhuc vixitq; mense q; vixit in hoc  
tempore, 1950 et anno 1951. Tunc vides, quod organicae res ipsae, vixit.  
adversus, q; restans. Tunc vides, deinde 1951 vixit. Et Cicerius, inq;  
151, a Novitate, 1952 vixit. Autem, q; dicit. Tunc vixitq; 1951  
mense, hisceterumq; vixit in ultima vixitq; mense, q; vixit in plu-  
re vixit. Ita q; ad responsum pateris & filiorum & fratris. Quia tunc vixit,  
et paternus, & fratres, vixerunt, ita vixit, q; vixit. Ita dicitur paternus  
et ad eum patriles, heredemque ex auctoritate, q; vixit. Hereditateq;  
vixit, vixerunt, q; patriles, heredemque ex auctoritate, q; vixit. Ita q; vixit  
paternus, ita q; vixit pater, & filii pateri. Quia tunc vixitq; 1950 & 1951  
tunc & huiusq; tunc vixitq; spesq; patriles, heredemque ex auctoritate  
vixit, vixerunt, q; patriles, heredemque ex auctoritate, q; vixit.

*2. februári önkormányzati bíróság meghallgatott ülésére* működött a teljesítőkörű Körüljárásban (működésben), mely Körüljárás teljesítőkörű önkormányzati bíróságként működött. 1994. január 20-án a Körüljárás teljesítőkörű önkormányzati bíróságának ülésén a működési előírások szerint a bíróság elnöke a következőképpen nevezte ki a bírói ügyeket:



# What is OCR/HTR?

**OCR = Optical Character Recognition**

**HTR = Handwritten Text Recognition**

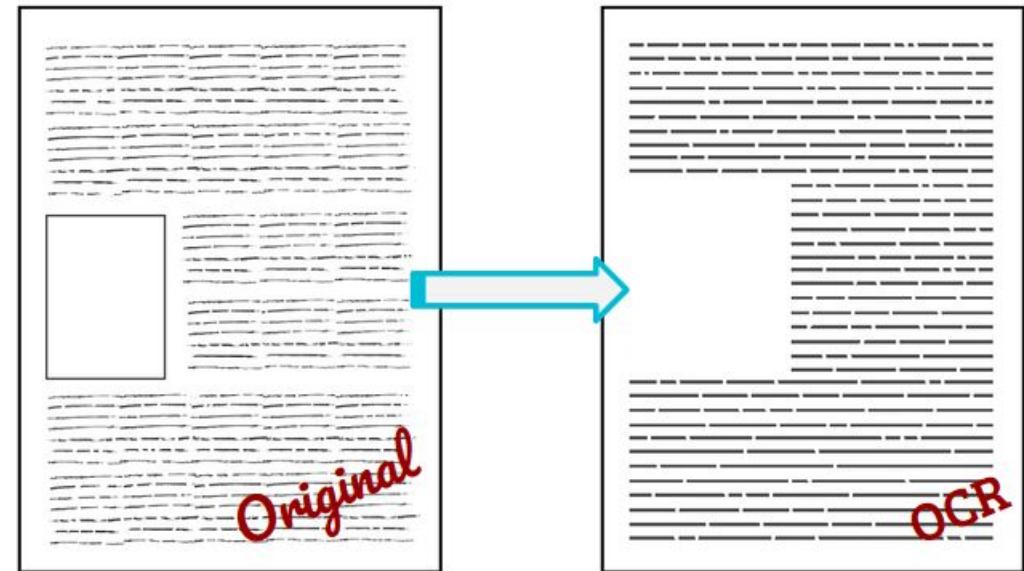
- conversion of image data into  
machine-readable text

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

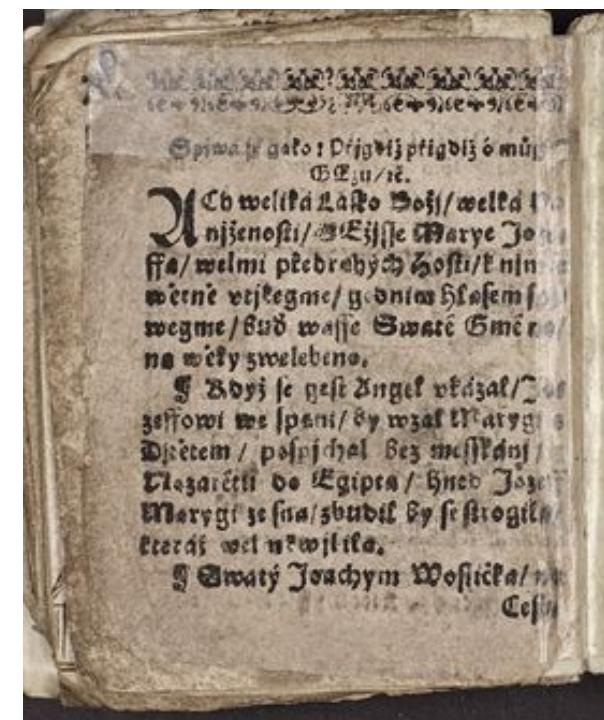
- Czech language and diacritical mark

*better tools → better OCR*

*good known OCR - [ABBYY](#)*

*good known HTR - [Transkribus](#)*

ukovským osudům, když obyvatelé k sebe hodi, ba ani smrt vydě neprodrží? A pak ka nemoci? „Zemřel „glaukom“ a zároveň „lešen“ být“. To můžeme na každém skoro „Tuberkulon“ ně je to zácnit plíce aneb zá „doctior“ přece jenom naučit rozumět. A e, jak k tomu přijde český autor Nodlavl, aby nejúhodnějšího dopisůla napsal národní Židlochovičich. Proč podporovat tekověho a otovové našeho místního mužstva učinil a „Sabin“ při rozpuštění na správní rok 1894 i k založení o židloch obecnosti lékaře



# 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)



# 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?

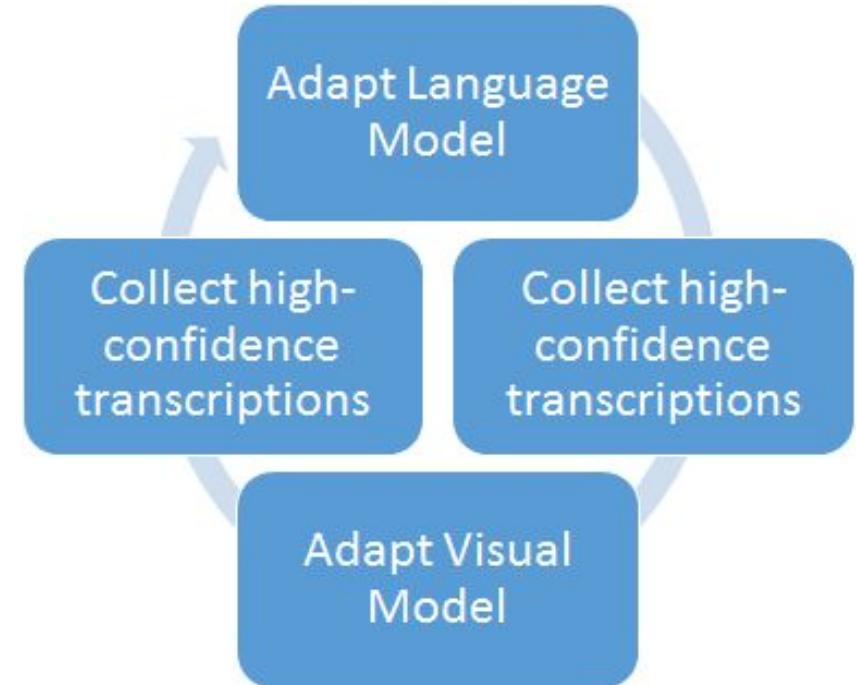
*People like to go to ..... (library, cinema, pubs, restaurants, work?)*



# Language models

Adapt visual and language models to a specific document:

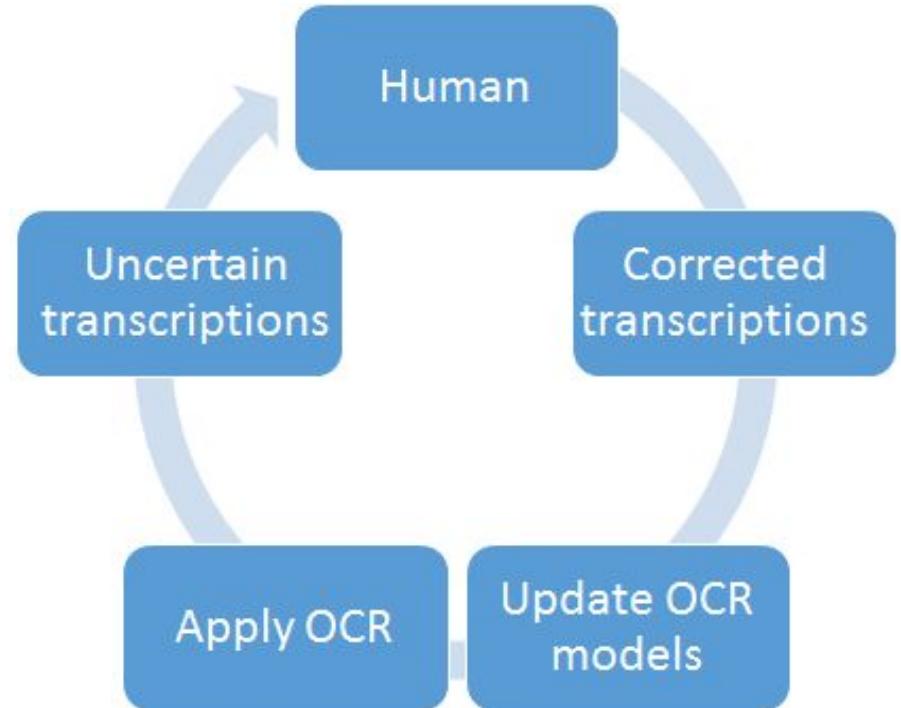
- Start with general models trained on large set of documents
- Adapt models to new document/writer/language
- Automatically adapt visual model by information from language model
- Automatically adapt language model by information from the visual model



# 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. (Hlučné provolávání hanby.)

Jak si radnice počiná 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 vše tu nevyřizuje. Radnice je zde podporována vládou a proti takovému nepříteli nezbývá žádný jiný prostředek, než jen jeden — volební lístek. (Bončlivý souhlas.)

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

*before*

P  
di  
b  
v  
z  
d  
t  
b  
r  
c  
k  
t  
n

němu stavěl, byl purkmistr Wieser. (Hlučné provolávání hanby.)

Jak si radnice počiná 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 vše tu nevyřizuje. Radnice je zde podporována vládou a proti takovému nepříteli nezbývá žádný jiný prostředek, než jen jeden — volební lístek. (Bončlivý souhlas.)

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

*after*



# PERO-OCR

- *low-quality print example*

potřeba: budu zpět za dve noumy.  
Nebylo třeba říkat ni podruhé. Pokynuv Hagrisovi  
opět přiblížil jsem se k neznámé páni. »Odkud přicházíte?«  
pravil jsem. »Půjdu s váni a podívám se, co je na celé  
události.« »Pravil-li on tak?« tázala se, ukazujíc na  
Grewera, jenž stál obrácen k nám zády, horlivě rozprávěje  
s ředitelem.

ceh dál osti i i Pra vil- nž stál , nera I

-livlo i ludi i |H>d nl.é l'- V r. Haj-isov; přichazít-: ■i i- ii«

ABBYY

Nebylo třeba říkat ni podruhé. Pokynuv Hagrisovi  
opět přiblížil jsem se k neznámé páni. »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 obrácen k nám zády, horlivě rozprávěje  
s ředitelem.

PERO



# PERO-OCR online application



## Project PERO OCR

Email:  Password:

## 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 editing](#) and [text recognition, correction and review](#).

The application allows users to automatically 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](http://pero-ocr.fit.vutbr.cz)



# PERO-OCR online application

*how to use the application  
step by step*

ect PERO OCR

**Add New Document**

Show 10 entries

Page	Name	Owner	State	Actions
	man	Search	Search	
	Manuscriptorium skúška - balík	Alzbeta Zavrelova	OCR completed.	
	Manuscriptorium skúška - Kronika česka + Hystoria Židowskā	Alzbeta Zavrelova	OCR completed.	
	Manuscriptorium - Ragská růže	Alzbeta Zavrelova	OCR completed.	

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

**Documents** **OCR jobs** **Help** **User**

Document management  
Layout Editor  
OCR Editor  
Project PERO

**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

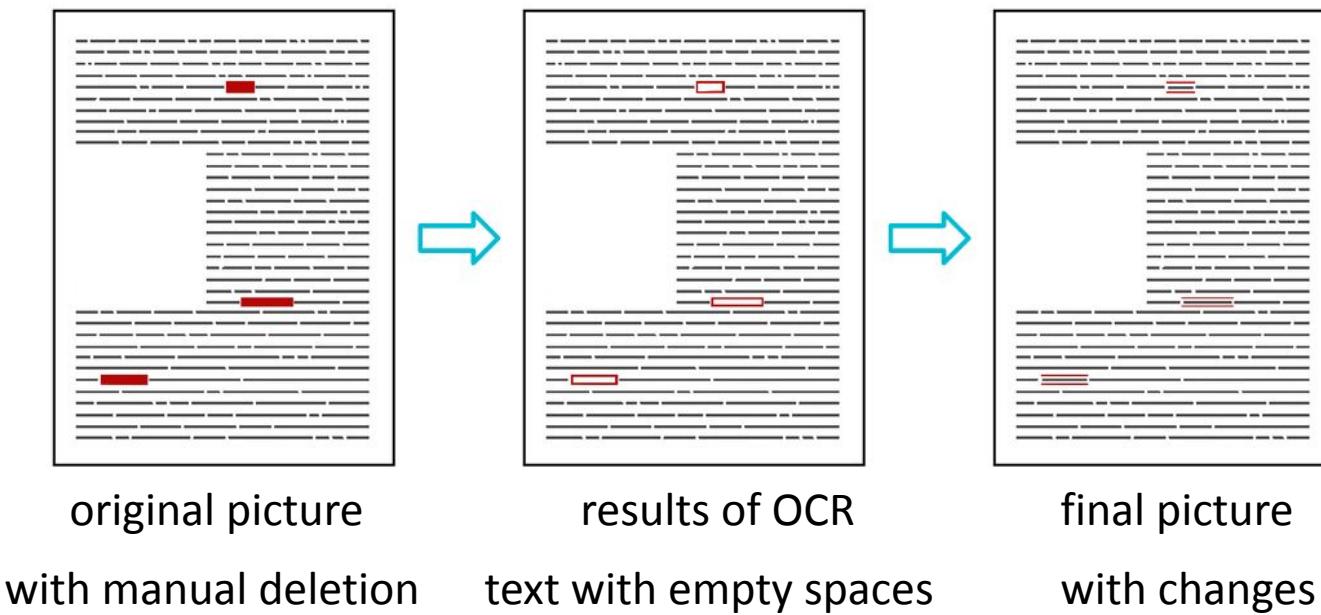


# How to modify historical documents?

**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



# How to modify historical documents?

*Can you find change in the text?*

třídy přeplněny. Zřízení této školy vyžadovalo mnoho práce a kdo se nejvíce 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.** (Bonžlivý souhlas.)

Za zřízení českých měšťanských škol v Brně bojují Čechové již od r. 1881, leč marně. Nyní by snad byla radnice ochotna dátě nám jednu měšťanskou školu, když budešem ustanovili od pořadavku rozdě-

třídy přeplněny. Zřízení této školy vyžadovalo mnoho práce a kdo se nejvíce 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.** (Bonžlivý souhlas.)

Za zřízení českých měšťanských škol v Brně bojují Čechové již od r. 1881, leč marně. Nyní by snad byla radnice ochotna dátě nám jednu měšťanskou školu, když budešem ustanovili od pořadavku rozdě-



jí  
pří  
pří  
di  
be  
bí  
ví  
z  
dá  
tu  
rě  
ci  
ka  
bí  
ni  
a  
B

jí  
pří  
pří  
di  
be  
bí  
ví  
z  
dá  
tu  
rě  
ci  
ka  
bí  
ni  
a  
B

# How to modify historical documents?

*Can you find change in the text?*

třídy přeplněny. Zřízení této školy vyžadovalo mnoho práce a kdo se nejvíce 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.** (Bonžlivý souhlas.)

Za zřízení českých měšťanských škol v Brně bojují Čechové již od r. 1881, leč marně. Nyní by snad byla radnice ochotna dátě nám jednu měšťanskou školu, když budešem ustanovili od pořadavku rozdě-

třídy přeplněny. Zřízení této školy vyžadovalo mnoho práce a kdo se nejvíce 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.** (Bonžlivý souhlas.)

Za zřízení českých měšťanských škol v Brně bojují Čechové již od r. 1881, leč marně. Nyní by snad byla radnice ochotna dátě nám jednu měšťanskou školu, když budešem ustanovili od pořadavku rozdě-



# How to modify historical documents?

Bratislavské národné muzeum v Bratislavě  
Bratislavské národné muzeum v Bratislavě

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](http://www.mzk.cz/en)

**Jiří Dufka**

head of department

Manuscripts and early printed books

[Jiri.Dufka@mzk.cz](mailto:Jiri.Dufka@mzk.cz)

**Alžbeta Zavřelová**

research assistant - PERO project

Innovation unit (research and development)

[Alzbeta.Zavrelova@mzk.cz](mailto:Alzbeta.Zavrelova@mzk.cz)

