Transkribus & Pero OCR

Miniúvod k HTR a OCR technologiím

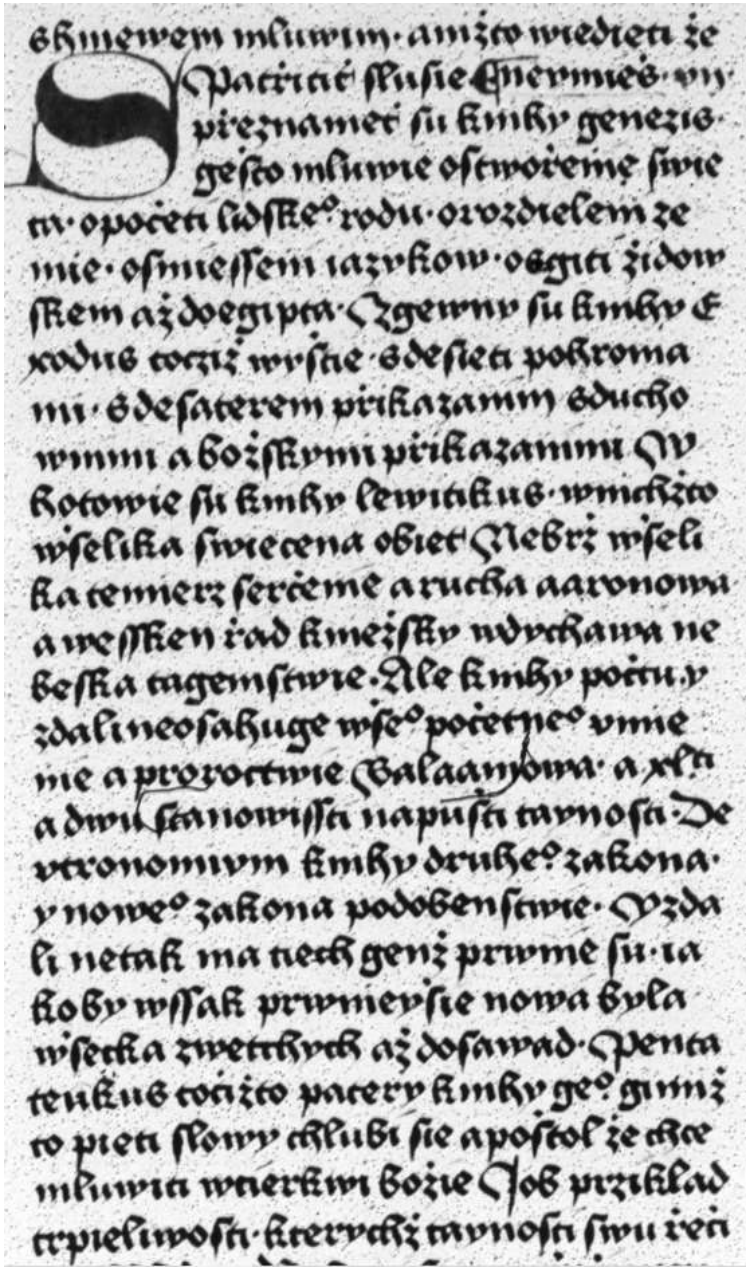
Anna Michalcová
ÚČJTK, FF UK
12. 10. 2022

Tento projekt vznikl v rámci zastřešujícího projektu Specifického vysokoškolského výzkumu 2022 – 260 554, řešeného na Filozofické fakultě Univerzity Karlovy.

OCR vs. HTR

- **OCR** (Optical Character Recognition)
  - technologie, která umí rozpoznávat jednotlivé znaky (může být problematické v případě starého textu)
  - pokud OCR umožňuje vytvořit trénovací data, zvýší se i úspěšnost jeho použití na konkrétní dokument
- **HTR** (Handwritten Text Recognition)
  - pracuje s celými řádky a dekóduje řádek jako celek
- Hlavní rozdíl z pohledu uživatele spočívá v tom, že fáze analýzy rozložení/segmentace je integrována do mechanismu OCR, zatímco v případě HTR se jedná o samostatný krok v pracovním postupu

Pero	Transkribus
s hiewem mluwim, a nižto wiedieti že	shniewem mluwím.anižto wiedieti že¬
Patřitiť slusie snevmieš vi	Patřitít ſluſíe Gnevmíeš.vii.
přeznameť su knihy genezis	přeznaňíeť ſu knihý genezíš.
gesto mluwie oſtwořeňie swie	geſto mluwíe oſtwořeňíe ſwie¬
ta, opočeti lidske° rodu, orozdielení ze	ta.opočeti lidſkeo rodu.orozdieleni ze¬
mie, oſmiesſeni iazvkow, ogiti židow	nie.oſmíeſſeňí iazyków.osgití židow
skem až dogipta. Zgewny su knihy s	ſkém aždoegipta.Zgewny ſu knihy k¬
rodus toczіž wystie sdesieti pohroma	rodus toczіž wyſtie. sdeſíeti pohroma¬
mi, sdesaterem přikazanim sducho	mi.sdeſaterem přikazaňím sducho
wnimi a božskymi přikazanimi Q	wňími a božſkymi přikazaňími.W¬
hotowie su knihy lewitikus, wnichžto	hotowie ſu knihý lewitikus.wnichžto
wſelika swiecena obiet Nebrž wseli	wſelika ſwiecená obiet Nebrž wſeli
ka temierz serčenie a rucha aaronowa	ka temierz ſerčeňíe a rucha aaronowa¬
a wesſken řad k nieżsky udychawa ne	a weſſken řad knieżſky wdychawa ne¬
beska tagemstwie. Ale knihy počtu. y	beſka tagemſtwíe.Ale knihý počtu.y
zdali neosahuge wſe° početne° vmie	zdali neoſáhuge wſeo vmie¬
nie a proročtwie Balaamowa a rlti	ňíe a proroctwíe Balaamowa.a. xlti.
a dwu stanowiſſti napusti tavnosti. De	a dwu ſtanowiſſti napúſti taynoſti.De¬
vtronomivm knihy druhe° zakona.	vtronomiyňí knihy druheo zakona¬
y nowe° zakona podobenstwie. yzda	y noweo zakona podobenſtwíe.Yzda¬
li netak ma tiech genž prwnie su ia	li netak ma tiech genž prwňíe ſu.ia
koby wsak prwnievsie nowa byla	koby wſſak prwnieyſíe nowa byla
wsecka zwetchych až dofawad. penta	wſecka zwetchych až doſawad.Penta¬
tenkus totižto pacery knihy ge° gimiž	teukus točižto patery knihy geo gími ž
to pieti slowy chlubi sie a postol že chce	to píetí ſlowý chlubi ſie a poſtol že chce¬
mluwiti wtierkwi božie Job prziklad	mluwiti wcíerkwi božíe Iob príklad
trpieliwosti, kterychž taynosti swu řeči	trpieliwoſti.kterýchž taynoſtí ſwú řéči



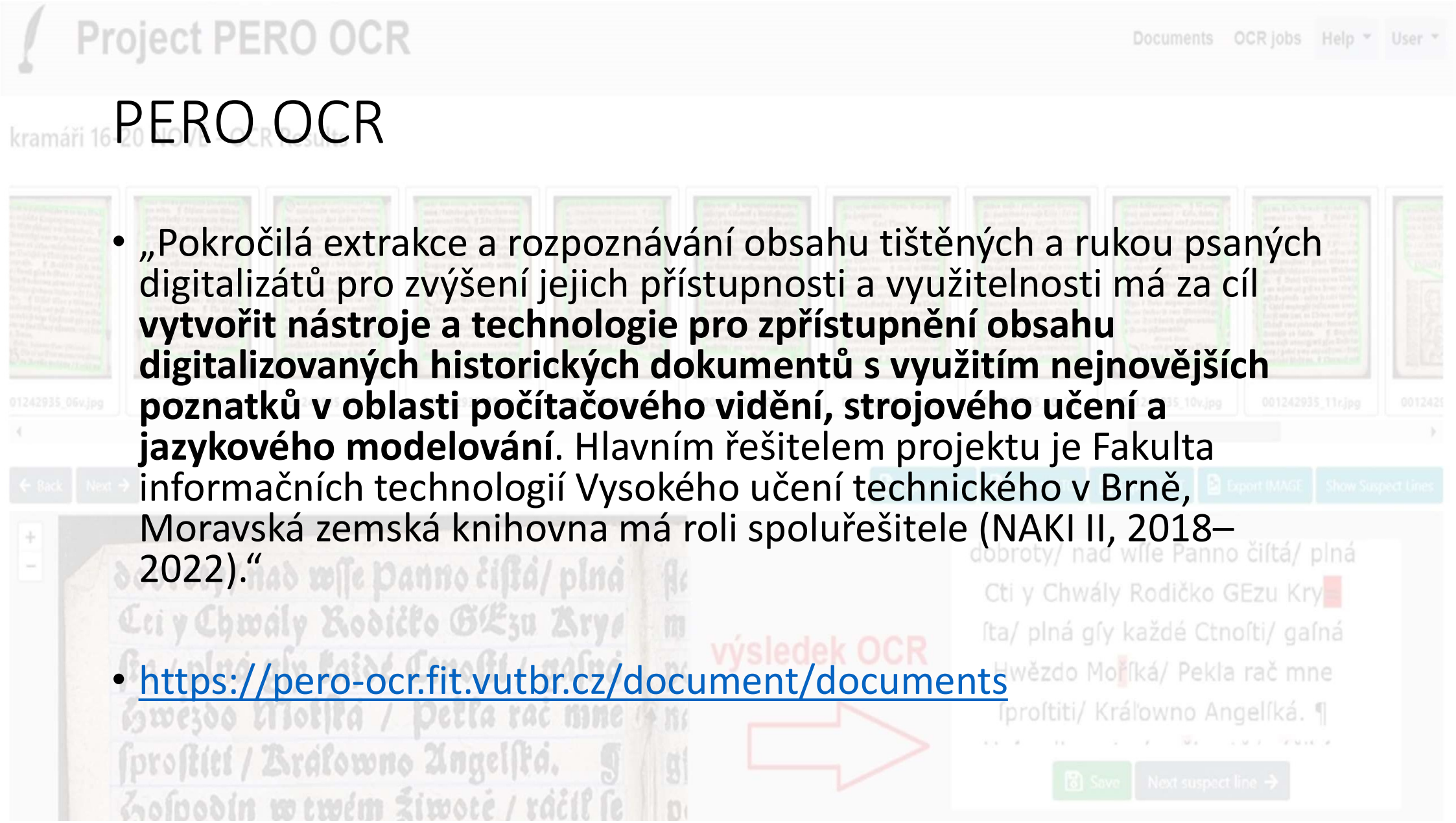
Transliterace vs. transkripce

- PERO OCR i Transkribus využívají transliteraci
- oproti paleografickému přepisu nejde o nápodobu rukopisu (např. rozlišování různých grafických podob písmen s výjimkou rozlišování dlouhého a kulatého *s*), ale o přepis znak za znak
- „**Transliterace** je pohodlné řešení otázek, vlastně pohodlné přesunutí všech sporných otázek z vydavatele na činitele. **Transkripce** je odnětí těch svízelů zhodnocení, co je (z hlediska vyjadřovacího) obsah a co forma, uživatelům edic a podání již jistého názoru na text a jeho obsah. Tedy jistá interpretace.“
- DAŇHELKA, Jiří (2013). Dopis Jiřího Daňhelky Marii Skalické, in: *Textologie a starší česká literatura*. Ed. J. Sichálek. Praha: Ústav pro českou literaturu Akademie věd České republiky v. v. i., s. 213–214.

PERO OCR

- „Pokročilá extrakce a rozpoznávání obsahu tištěných a rukou psaných digitalizátů pro zvýšení jejich přístupnosti a využitelnosti má za cíl **vytvořit nástroje a technologie pro zpřístupnění obsahu digitalizovaných historických dokumentů s využitím nejnovějších poznatků v oblasti počítačového vidění, strojového učení a jazykového modelování**. Hlavním řešitelem projektu je Fakulta informačních technologií Vysokého učení technického v Brně, Moravská zemská knihovna má roli spoluřešitele (NAKI II, 2018–2022).“

- https://pero-ocr.fit.vutbr.cz/document/documents



Transkribus

- Transkribus je platforma pro **rozpoznávání textu, analýzu obrazu a rozpoznávání struktury historických dokumentů**. Platforma byla vytvořena v rámci dvou projektů EU tranScriptorium a READ. Byl vyvinut univerzitou v Innsbrucku. Od 1. července 2019 platformu řídí a dále rozvíjí **READ-COOP**.
- nahrání dokumentů, transkripční nástroje zdarma × automatický přepis placený (kreditový systém)

Transkribus

- **kdo všechno Transkribus využívá?**
- archivy: Berlínský státní archiv, Zeit.punkt (digitalizace novin ze Severního Porýní-Vestfálska z let 1801–1945), Archiv města Amsterdam, Národní archiv Finska, Arolsen Archives (International Center on Nazi Persecution) – archivy nyní prohledatelné
- univerzity, akademie: k vytvoření korpusu textů, umožňuje pracovat na transkripci z různých míst
- Kati: University of Innsbruck / Federal State of Tirol
- EnrichEuropean+ (podpora hromadné práce pro evropské kulturní dědictví)
- Österreichischer Bibelübersetzer (Berlin-Brandenburgischen Akademie der Wissenschaften)
- Rakouská akademie věd + READ-COOP (optické rozpoznávání hudby)

Transkribus

APPLICATION

Please, send a **short CV** and **cover letter** to Jan Odstrcilik | jan.odstrcilik@oeaw.ac.at **before 11th September 2022.**
As a subject, use „**HTR School 2022**".

Please don't forget to indicate the team/ module you would like to take (i.e. Carolingian Latin, late medieval Latin, medieval German, medieval Czech).
If you have any questions, don't hesitate to ask.

Michael Berger, *University of Vienna* | michael_berger@univie.ac.at
Tim Geelhaar, *Bielefeld University* | tim.geelhaar@uni-bielefeld.de
Tobias Hodel, *University of Bern* | tobias.hodel@unibe.ch
Sarah Hutterer, *Unviersität Wien* | sarah.hutterer@univie.ac.at
Anna Michalcová, *Czech Academy of Sciences* | a.michalcova@ujc.cas.cz
Stephan Müller, *University of Vienna* | stephan.mueller@univie.ac.at
Jan Odstrčilík, *Austrian Academy of Sciences* | jan.odstrcilik@oeaw.ac.at
Steffen Patzold, *University of Tübingen* | steffen.patzold@uni-tuebingen.de
Leon Pürstinger, *Austrian Academy of Sciences* | leon.puerstinger@oeaw.ac.at

Helmut Reimitz, *Princeton University* | hreimitz@princeton.edu
Dennis Wegener, *University of Vienna* | dennis.wegener@univie.ac.at
Vicent Bosch , *Transkriptorium, Universitat Politècnica de València* | vbosch@transkriptorium.com
Gerda Heydemann, *Freie Universität Berlin* | gerda.rummel-heydemann@fu-berlin.de
Daniela Mairhofer, *Princeton University* | daniela.mairhofer@princeton.edu
Joan Andreu Sanchez, *Transkriptorium, Universitat Politècnica de València* | jandreu@prhlt.upv.es
Alejandro Toselli, *Transkriptorium, Universitat Politècnica de València* | ahector@prhlt.upv.es
Enrique Vidal, *Transkriptorium, Universitat Politècnica de València* | evidal@prhlt.upv.es

• • •

WWW.OEAW.AC.AT

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

CALL FOR PARTICIPATION
WINTER SCHOOL | OCTOBER - DECEMBER 2022

INTRODUCTION INTO HTR
HANDWRITTEN TEXT RECOGNITION
TECHNOLOGIES OF MEDIEVAL MANUSCRIPTS
LATIN|GERMAN|CZECH

Organisation:
Institute for Medieval Research of the Austrian Academy of Sciences
Manuscript, Rare Books and Archival Studies Initiative (MARBAS), Princeton University

In cooperation with:
CRC 1288 Practices of Comparing, University Bielefeld
Department of German Studies, University of Vienna
Digital Humanities, Walter Benjamin Kolleg, Universität Bern
History Department, University of Tübingen
tranSkritporium, Valencia
tranSkriptorium.com

Contact:
Dr. Jan Odstrčilík
Austrian Academy of Sciences
Jan.Odstrcilik@oeaw.ac.at

4 Zoom online sessions | Oct 21, Nov 4 and 25, Dec 9
3-day-workshop in person at Vienna | December 19-21

INSTITUT FÜR MITTELALTERFORSCHUNG

Transkribus

Polish – General Model

Free Public AI Model for Handwritten Text Recognition with Transkribus

Transkribus > Public Models > Polish – General Model

« Back to all public models

A general model for historical and modern Polish handwriting.

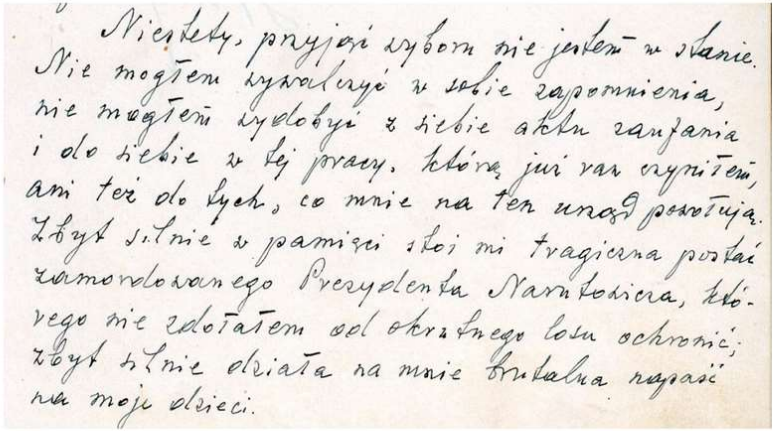


Image Source: Wikipedia

Model Overview

Name:
Transkribus Polish M2

Creator:
Transkribus Team

Model ID:
44976

Century:
n/a

Languages:
Polish

Script:
Latin alphabet

Engine:
PyLaia

Material:
Handwritten

Štátna vedecká knižnica v Banskej Bystrici
a
Filozofická fakulta Univerzity Mateja Bela v Banskej Bystrici

Vás pozývajú
na vedeckú konferenciu s medzinárodnou účasťou

Digital humanities
Nástroje sprístupňovania historického dedičstva

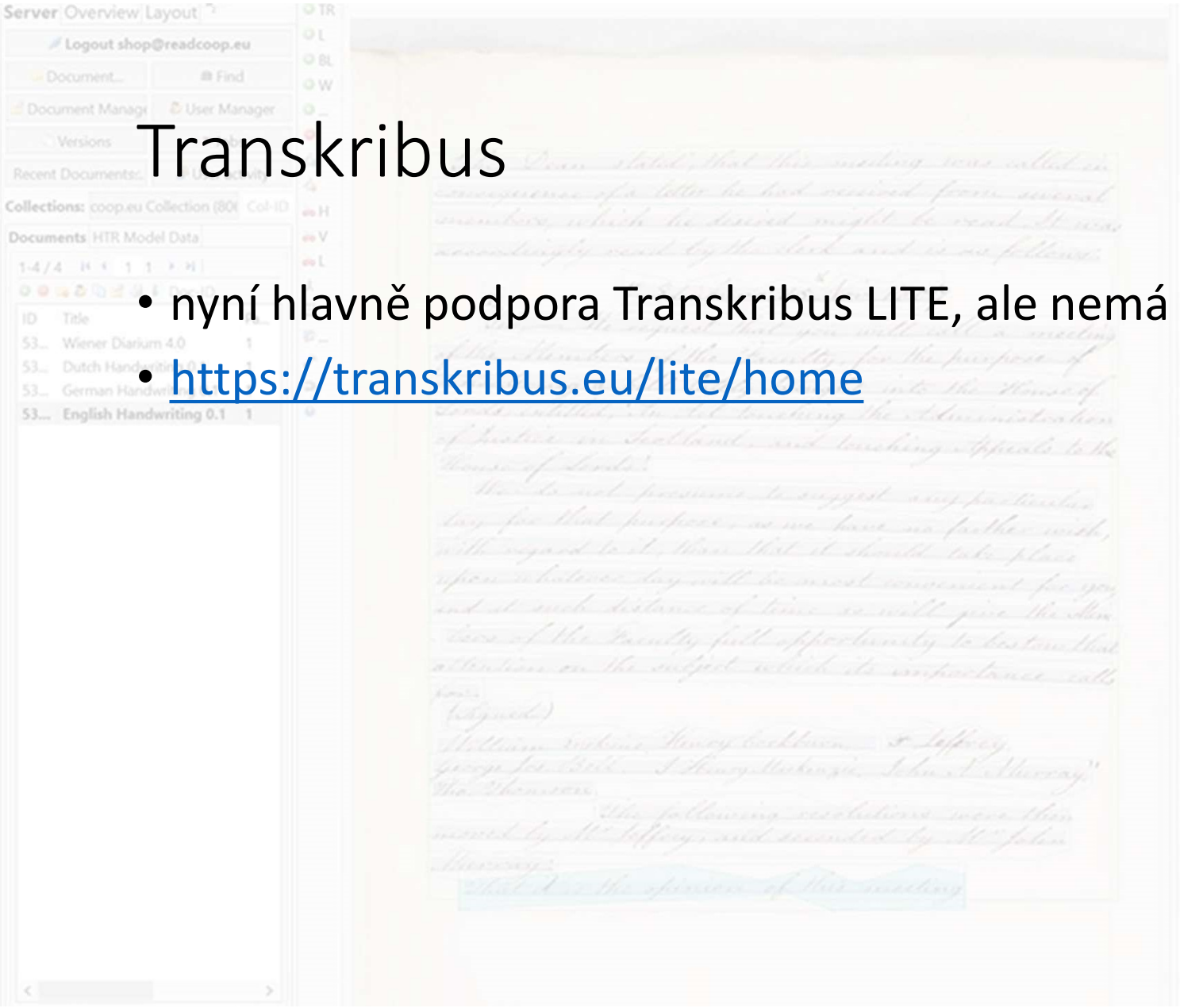
usporiadanú v dňoch 12. – 13. októbra 2022

Štátna vedecká knižnica v Banskej Bystrici
Lazovná 9

AGENTÚRA NA PODPORU VÝSKUMU A VÝVOJA

Transkribus

- nyní hlavně podpora Transkribus LITE, ale nemá zatím všechny funkce
- https://transkribus.eu/lite/home



Transkribus v1.10.1 (06_03_2020_10:18), Loaded doc: PRÓBKA, ID: 1105318, Page 1, file: próbka 1.jpg* [Image Meta Info: (Resolution:96.0, w*h: 2601 * 925)] [current line: w*h: 1850 * 85]

Search current document... 1 /1 In Progress

Server | Overview | Layout | Metadata | **Tools**

Layout Analysis

Method: CITlab Advanced Configure...

Current page Pages (1): 1-1

Find Text Regions Only use un

Find Lines in Text Regions

Run

Text Recognition

Method: HTR (CITlab)

Models... Train...

Run...

Compute Accuracy

Other Tools

P2PaLA...

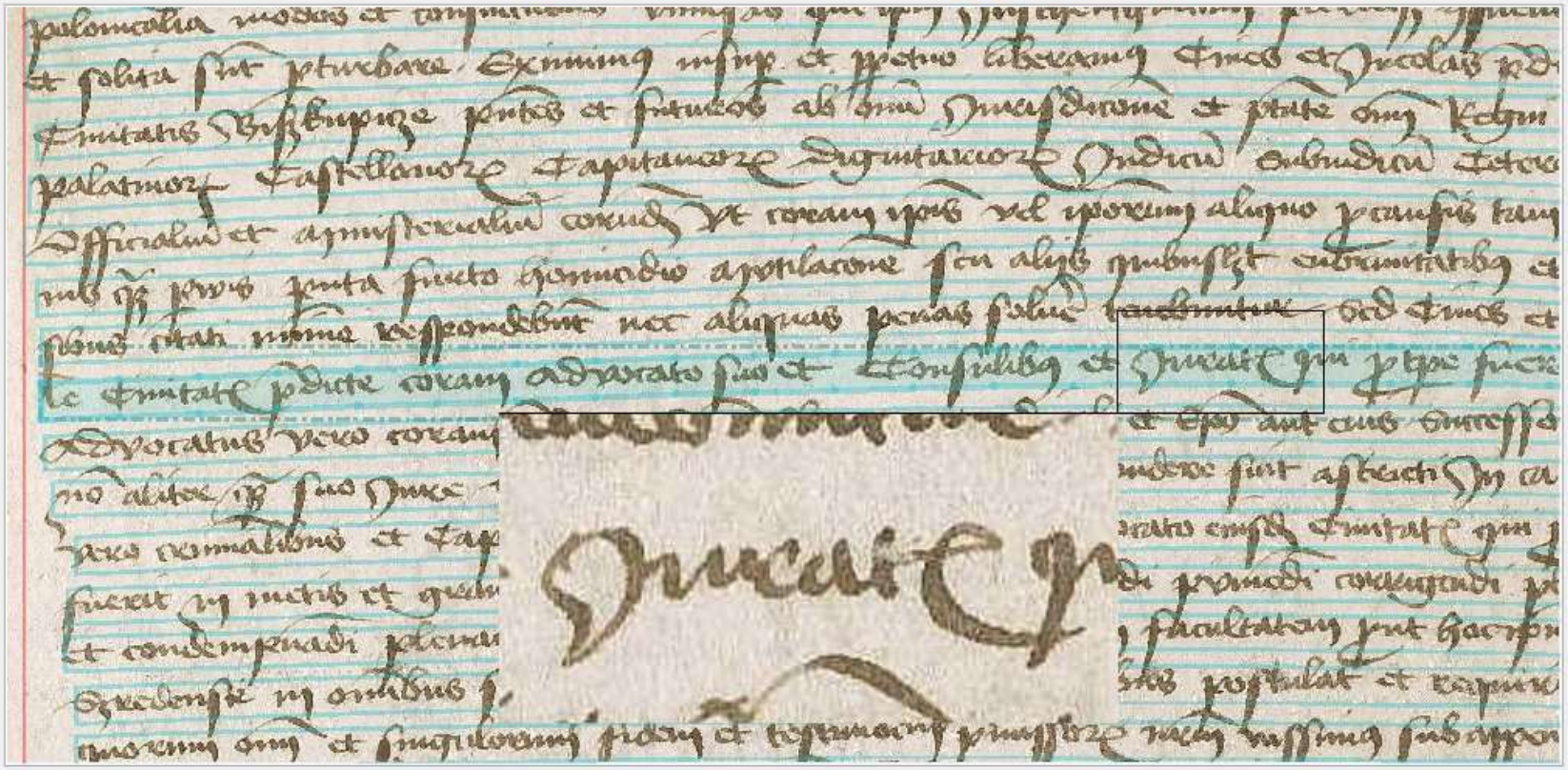
Text2Image...

Current page Pages (1): 1-1

Add Baselines to Polygons

Add Polygons to Baselines

TR L BL W ... H V L ...



1-10 sibus· citati· minime· respondebunt· nec· aliquas· penas· solvere· tenebuntur· sed· cives· et· Inco

1-11 le· civitatis· predicte· coram· advocato· suo· et· consulibus· et·

Děkuji za pozornost!