

# Nejnovější modely transkripce

## Západoslovanské historické rukopisy a tisky

**Prof. PhDr. Dušan Katuščák, PhD. - ŠVK Banská Bystrica a Slezská univerzita v Opavě**

*„Táto práca bola podporená Agentúrou výskumu a vývoja na základe Zmluvy č. APVV-22-0049“ Projekt DIGINARCH 2025-2029*

# Osnova

- 1) Internetový manifest IFLA
- 2) Transkripce historických dokumentů
- 3) Supermodely transkripce
- 4) Projekt APVV DIGINARCH-AI



# AI a paměťové instituce

<https://ipk.nkp.cz/docs/IFLA-AI-Entry-Point-to-Libraries-and-AI-cs.pdf>

**IFLA Knihovny a AI – úvod do tématu Internetový manifest- Květen 2025**

**Globální názor na AI z pozice knihoven (informačních institucí)**

- 1. Úloha knihovny vzhledem na technologie AI**
- 2. Možné přínosy umělé inteligence**
- 3. Možné aplikace umělé inteligence**
- 4. Možná rizika umělé inteligence**
- 5. Scénáře pro diskusi o AI v rámci profesního rozvoje**

# Úlohy knihoven

## Vliv knihoven na AI

Knihovny budou spolupracovat na definici odpovědného používání AI.

Svůj vliv mohou uplatňovat na šesti různých úrovních:

1. Vývoj nebo licencování AI pro služby knihoven
2. Vývoj a školení AI (ve veřejném zájmu).
3. Poradenství uživatelům při službách AI.
4. AI gramotnost
5. Poradenství o AI v širší komunitě.
6. Prosazování odpovědného využívání umělé inteligence ve společnosti a vhodné regulace.

# Význam informačních institucí pro AI

- AI považují za **metasystém**, který je závislý na zdrojích, vytvořených tvořivými lidmi
- AI je bytostně závislá na **digitálních** informačních zdrojích
- Digitální **repozitáře** jsou pro AI zásadní
- Knihovny, archivy a jiné instituce jsou **klíčovými tvůrci** digitálních zdrojů pro AI
- Knihovny patří do **jádra zdrojů** pro AI
- Bez digitálních zdrojů informačních institucí některé **AI nemůžou existovat**
- Je nevyhnutně potřebné, **aby AI uváděly využívané zdroje**

# Transkripce historických dokumentů

## Proč je to důležité?

- Textové písemné dědictví (rukopisy, staré tisky, strojopisy...) nejde celkem dobře (nebo vůbec) transkribovat klasickými nástroji OCR (ABBYY FineReader)
- OCR 87%-93% tištěných dokumentů
- **Obrazové galerie** - Digitalizáty (digitální faksimile) v depozitářích nejsou digitálními knihovnami (Manuskriptorium)
- Skenování bez transkripce OCR/HTR - je jenom poloviční digitalizace
- K dispozici jsou nástroje a platformy transkripce využívající (AI) strojové učení
- Modely transkripce pro západoslovanské jazyky si musíme vytvořit sami

# OCR Assistant Bezplatný nástroj OCR

Asistent OCR Využijte bezplatný nástroj OCR k rozpoznávání textu a tabulek, převodu PDF do formátu WORD a zlepšení efektivity práce a života

## Rozpoznávání slov



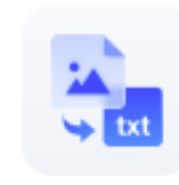
### Univerzální rozpoznávání znaků

Převádějte obrázky do upravitelných formátů slov, txt



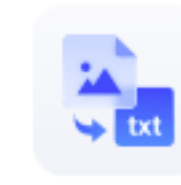
### Tisk rozpoznávání textu

Převádějte obrázky do upravitelných formátů slov, txt



### Rozpoznávání rukopisu

Převádějte obrázky do upravitelných formátů slov, txt



### Identifikace starověkých knih

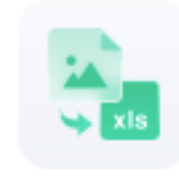
Převádějte obrázky do upravitelných formátů slov, txt

## Identifikace tabulky



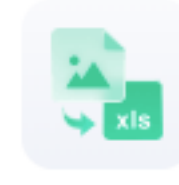
### Univerzální identifikace tabulek

Převod tabulek v obrázcích na upravitelné soubory aplikace Excel



### Rozpoznávání ohraničené tabulky

Převod tabulek v obrázcích na upravitelné soubory aplikace Excel



### Rozpoznávání tabulek bez okrajů

Převod tabulek v obrázcích na upravitelné soubory aplikace Excel

## PDF do souboru



### PDF do Wordu

Převod souborů PDF do formátu Word



### PDF dělají Markdown

Převod souborů PDF do formátu Markdown



### PDF do Excelu

Převod souborů PDF do formátu Excel



### PDF až JPG

Převod souborů PDF do formátu JPG



### PDF dělají PPT

Převod souborů PDF na dokumenty PowerPoint

## Další konverze

Microsoft Teams

Slovo do PDF

Excel do PDF

Obrázek do PDF

Word do JPG



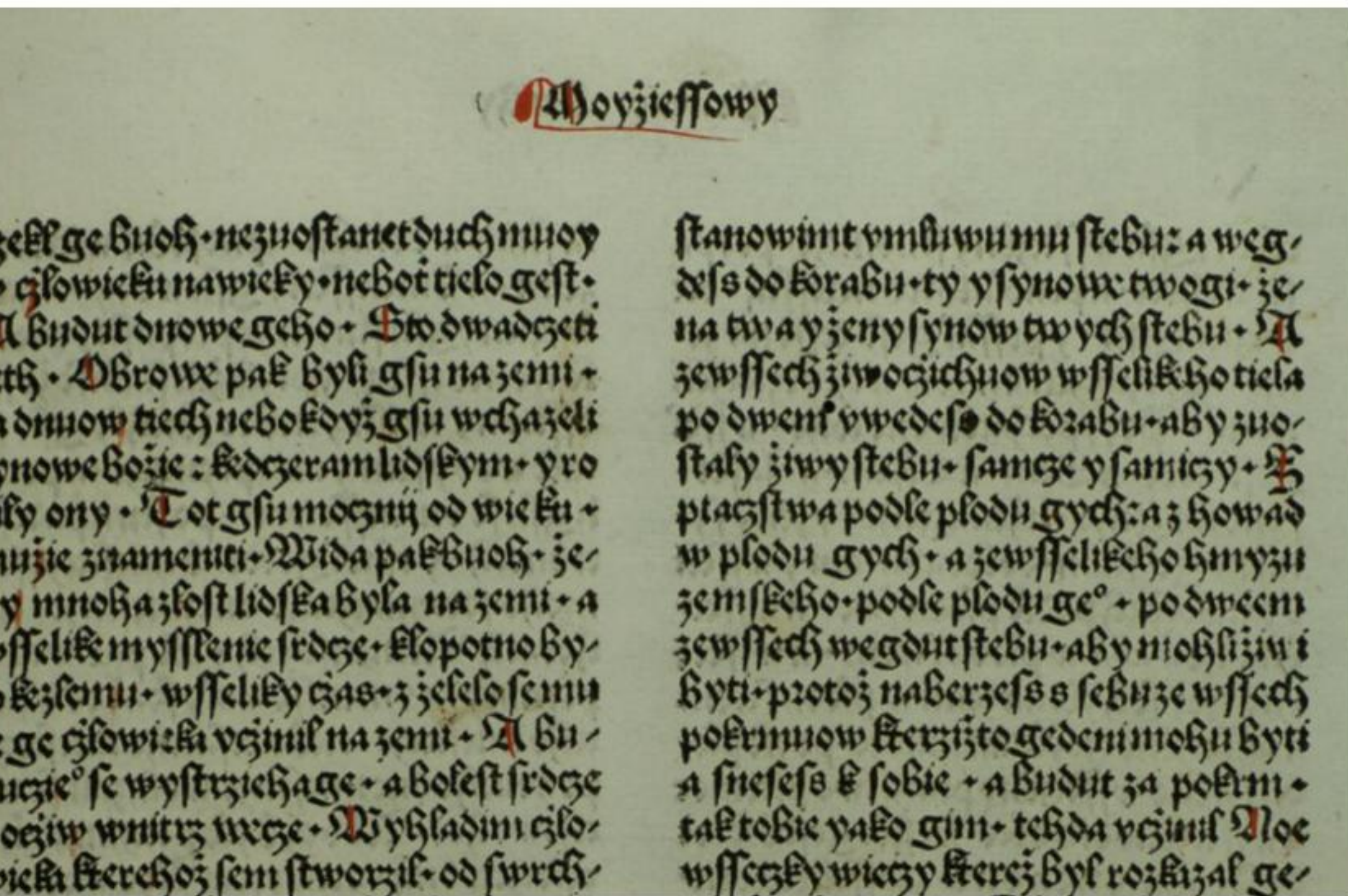
# Nástroj na rozpoznávanie rukou písaného textu | Inteligentný rukopis na text pomocou umelej inteligencie

网页版仅供在线体验，完整功能请下载客户端！

客户端下载

各VIP版本功能预览

Result



Roicomy  
 meefgeBuog-neuoftanetsudjno9  
 fanowinntomfwumu(teBu:ameg.  
 N  
 wolowilitnamicEyoneBotticio gift.  
 sfssokrabu ty yfynoutogt-c  
 梦  
 ΣsutsooSmac  
 ия бАујснУ[уноу тpyд(теБи+  
 34G  
 a  
 ;emfcdjjiocicguomwfccliRgoticl  
 重春  
 og  
 on:agggo  
 g0f.2  
 Tynowe Boite:RocamlisEym+yro  
 falyjytebu-amgcyfamigy+3  
 开4  
 sifyony. Cotglumognų oo wicEu-  
 enmoltn:jrusold leodaflanid  
 g  
 g  
 ofw  
 mplosn gyc.aielEo5my3t  
 tgonm g

Microsoft Teams

Opätovné nahranie

复制



[Home](#)[Desk](#)[Models](#)[Public models](#)[My models](#)[Training lab](#)[Sites](#)[All \(635\)](#)[Favorites \(3\)](#)[Public \(344\)](#)

MODEL TITLE

LANGUAGES

WORDS

LABELS

CER

The Text Titan I ter

Super Model

Danish, Dutch; Flemish

+9

31 127 219

Featured

The Text Titan I

Super Model

German, Dutch; Flemish

+4

0

2.95%

Featured

Dutch Dean

Super Model

Featured

Dutch; Flemish

24 721 830

3.50%

Dansk Dokumentalist

Super Model

Danish

12 849 906

3.70%

Featured

German Genius

Super Model

German

21 100 000

4.50%

Featured

# Supermodel 1

## The text Titan

Produkt týmu Transkribus.

Nejlepší verze modelu Text Titan. 11. 06. 2025

Není trénován na češtinu a slovenčinu

Trénovaný na obrovském masívu **31 127 219** slovech.

Obecný model rozpoznávání bez nutnosti vlastního tréninku.

Dobře funguje jak s tištěnými, tak ručně psanými texty

Text Titan I ter lepší CER téměř o 3 procenta oproti Text Titan (2023)

# Supermodely Transkribus

## Veřejné a privátní

- Můj cíl (od 2019...)
- Vytvořit Univerzální supermodel transkripce pro západoslovanské jazyky + němčina, latina, maďarčina
- Koordinace úsilí odborníků a nadšenců (Česko, Polsko, Maďarsko a Transkribus)
- Snad společný projekt aplikovaného výzkumu?
- Integrovat náš supermodel do supermodelu Titan

# 1 WEST SLAVIC Supermodel 1.0

ID437773

Model transkripce historických dokumentů v západoslovanských jazycích

Integruje částečné modely:

a) Slovenský výzkumný projekt UMB a ŠVK Banská Bystrica Skriptor na Slovensku 2020-2024

b) Studentská grantové soutěž (SGS) na Slezské univerzitě v Opavě (3 diplomky, 4 modely)

c) HTR Winter School 2022, 2023, 2024. (Viedli: Dr. Jan Odstrčilík a Anna Michalcová)

Výpočet trval ca 60 hodin. 250 cyklů,  
**399 806 slov; 71 129 řádků**

Tréningový set 1666 stran CER 2.98%  
Validačný set 166 stran CER 12.66%

**Přesnost 87,34% je nedostatečná! (Min 95%)**



Name: 1666 stran Train Set

Language: Slovak; Czech; Sorbian; Polish

WEST SLAVIC Supermodel 1.0

Creator: 166 stran Valid Set

dusankatuscak@gmail.com

Description:

WEST SLAVIC Supermodel 1.0  
The model of transcription of historical documents in West Slavic languages (Czech, Slovak, Polish, Sorbian) integrates partial models that were created in a) the Slovak research project Skriptor in Slovakia, b) in the framework of the Student Grant Competition (SGC) at the University of Silesia in Opava

Parameters:

Max epochs 250  
Early stopping 20  
Epochs trained 250

Document Type: Handwritten

Nr. of Words: 399806

Nr. of Lines: 71129



Save

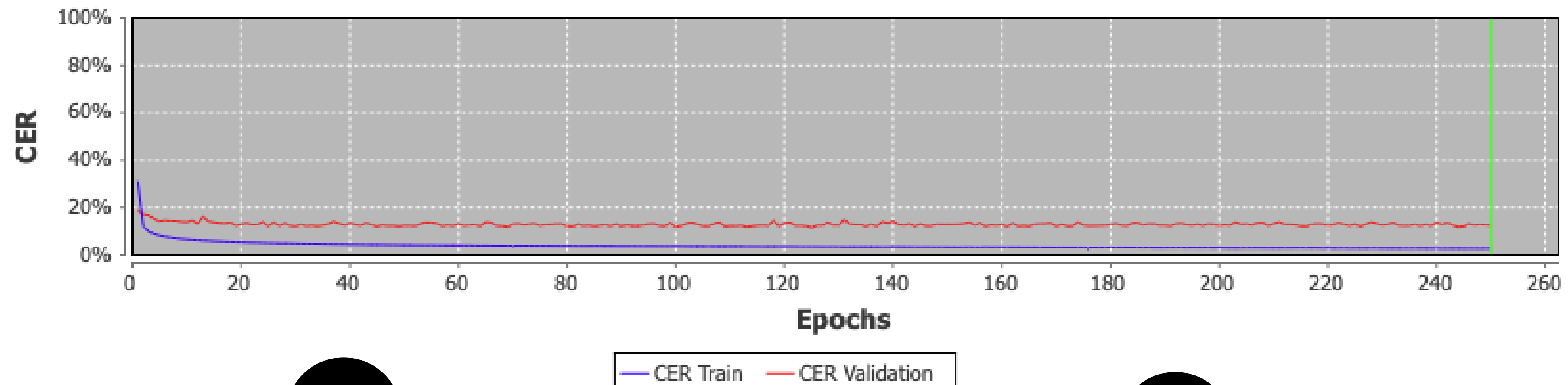
Train Set

Validation Set

Characters

Show advanced parameters...

### Learning Curve



CER on Train Set: 2.98%

CER on Validation Set: 12.66%

# 2. ID 63569

## Slovenský supermodel M1

Slouží ke transkripci **rukopisů** v slovenštině, latině, maďarštině, češtině, slovakizované češtině.

Data z projektu ID: APVV-19-0456 SKRIPTOR.

1359 stran v kvalitě Ground Truth (GT)

56713 řádků a 333 777 slov

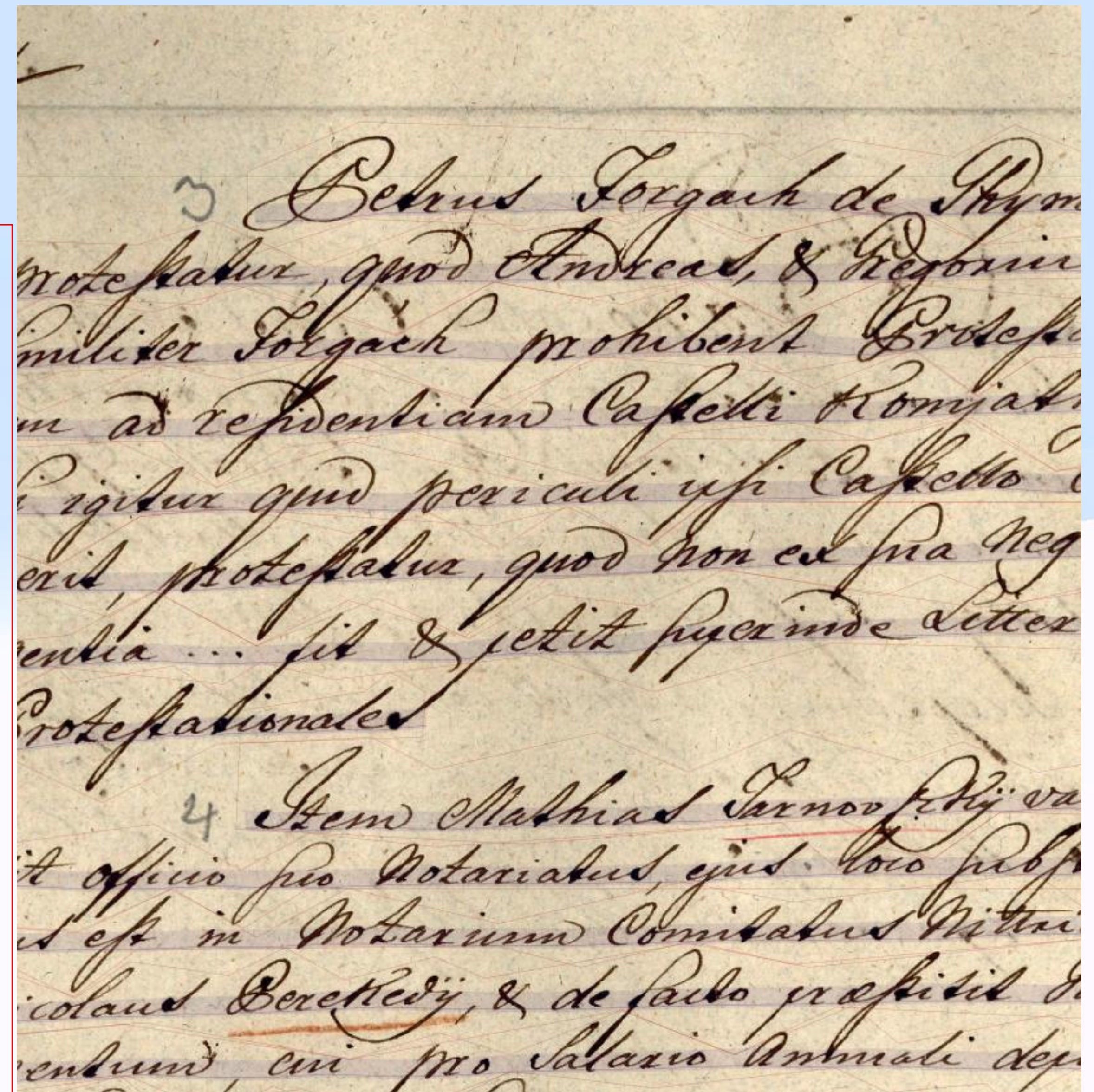
1224 stran Train Set; CER 4,90 %

135 stran Valid Set CER 5,30 %

**Přesnost 94,7 % = Dobrá**

Katuščák, D., Nagy, I., Maliniak, P., Kurhajcová, A., Tomeček, O., Kunec, P., & Bôbová, M. (2024). Slovak Supermodel M1 (SSM1) : Matej Bel University SKRIPTOR project datasets (First version (20240503)) [Data set].

Zenodo. <https://doi.org/10.5281/zenodo.11109087>.



# 3. Český supermodel SGS

## ID220865

Agregovaný model transkripce českých historických rukopisů (novověk)

Studenti Filozofické fakulty Slezské univerzity v Opavě v rámci projektu Studentské grantové soutěže v roce 2024

Základem supermodelu je:

1. model, jehož autorem je student Lukáš Němec.
2. model, jehož autorem je student Vojtěch Říha.

+

15 stran rukou psané zápisy v GT - Matice slezská

947 pages Train et 2.98%  
93 pages Valid set 5.83%.

Přesnost 94,17% (Dobrá)



# 5. Old Czech Handwriting

## ID341425

Model vychází z přepisů v kvalitě GT ze čtyř staročeských zdrojů:

1. Padeřovská bible (1432–1435)
2. Cisteriánská bible (1456)
3. Nový testament Martina Lupáče (1440)
4. Pražská bible (1488)

Polodiplomatické pravidla přepisu stanovených Pero OCR a "Směrnice pro vydávání starších českých textů" od Jiřího Daňhelky

Lze jej použít na staročeské rukopisy psané bastard, především ze 14. a 15. století.

Michalcová, A., Bazelides, K., Hajič, J., Pěnkavová, E., Maniaková, L., Kreisingerová, H., Filipová, J., Chi-hung Lu, & Dvořáková, M. (2022). Padeřov-Bible-handwriting-ground-truth: Initial release [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7467034>; 2) Cistercian Bible (New York, The Morgan Library & Museum, shelfmark MS M.752, 1456); 3) New Testament of Martin Lupáč (Vienna, Austrian National Library, shelfmark Cod. 1175, 1440), Zenodo: <https://doi.org/10.5281/zenodo.10619017>; 4) Prague Bible (Vienna, Österreichische Nationalbibliothek, shelfmark Ink 13.C.5, 1488), Zenodo: <https://doi.org/10.5281/zenodo.14524858>





## 6. Slovenský supermodel pro tisk a psací stroj (SSPT1)

### ID78289

Základem jsou slovenské, české, lužickosrbské srbské tištěné historické dokumenty a strojopisné dokumenty

Přibližně 60-80% strojopisných dokumentov, najmä z rokov 1850-1990

1. Projekt Skriptor (UMB a SVK v Baňské Bystrici),
2. Přepisy připravené studenty Slezské univerzity v Opavě v rámci SGS a vzdělávací aktivity.

Stránky v kvalitě Ground Truth

37 897 řádků a 200 697 slov.

489 stran Train Set CER 1,48 %

59 stran pro Valid Set CER 1,45 %.

Přesnost transkripce 98,55% Excelentný model

Katuščák, D., Nižníková, L., Mikušková, M., Halfarová, N., Gajdošová, T., Málková, L., Taufrová, N., Nagy, I., Kováčová-Pohlová, K., Šmida, M., & Kociánová, N. (2024). Slovenský Supermodel P&T1 (SSPT1) : Datasets projektu SKRIPTOR Univerzity Mateja Bela (First version (20240520)) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.11218527>



**Sert a Pravda.**

So gest:

welmi pěkně smyslené, vtěšsené

**Hystorie a Rozprávky**

w několik stech,

Ěterěz se pro wyraženi myšli a pro

**Wydawatel k wptimným Čechům.**

**G**estliže kdy lidé rádi hystorie čítali, gisté je za našeho věku nehradí; a protož také nyní husťši na swětlo wyšly. My suce Čechové dosawád melifom nich počtem chlubiti se nemůžeme; wšak křeč w těchto několik letech na swětlo wysly, vědce gsau na důkaz, dšlem je Čechům písny spisowatelé nescházeti, a dšlem je Čechš jazyl není tak nechludš, gak mnozš před wšle, wšle lcau čubili, abn se w njem mšle.

**Katechysmu.**

Česty wplojeno býti má. A tak těmito slovy Zbor. Cyrlew/gedna a tás wěc se mjni, neb rozum).

**14. Kádli ty chodíš do Zboru/ to gest, do společného shromáždění wěřjých?**

**W prawdě rad.**

**15. Proč?**

Proto, abyš darů a milostj spasytědných, strze přiluhowanj přitomně nabýwal. Take abyš se cet a k chwále Bojž. Pšničky duchownj s wěřjými zpsjwal / a w nich srdce své, prawdu Bojž naplněné, Bohu obětowal / Modlitby společné s nimi činil / Slowa Bojžho poslauchal / Swátosti Těla a Krwe Dáně časem přijwal / a tak se strze ty wěcy w swěte

byla splněna na 140 procent. K těmto úspěchům p  
rostlinné výroby přispěle svou produkcí i nové  
středisko okrasných dřevia, které JZD převzalo.  
Dobrych výsledků v rostlinné výrobě dosáhlo dru  
díky tomu, že byla realizována opatření, která  
na pro konsolidaci družstva a zaměřena na záro  
zabezpečení výšivy rostlin s cílem dosáhnout se  
nosí.

Navzdor úspěšné byla také zemědělská produk  
křivočinné výroby. Zvláště příznivá situace se  
robě mláka, kterou JZD svýžilo proti plánu o 1  
29

**LUŽICA.**

Měsáčník za zabawu a powučěnja.  
Škrumásky česowskeho a delnjeluzatskeho Serbow.  
W powětu tšch serbow křowu Měsáčník w Čechowch  
Wydawatel: Měsáčník w Čechowch  
Wydawatel: Měsáčník w Čechowch

Wydawatel: Měsáčník w Čechowch  
Wydawatel: Měsáčník w Čechowch

# Nový projekt DIGINARCH-AI APVV-24-0049

Digitalizácia a inteligentná obsahová analýza historických dokumentov pre odbornú archívnu činnosť a historickú vedu (2025 - 2029, 263 000 EUR = ca 6.5 mil. CZ)

- Univerzita Mateja Bela v Banskej Bystrici
- Štátna vedecká knižnica v Banskej Bystrici
- Štátny archív v Košiciach
- Štátny archív v Banskej Bystrici
- Archívne fondy Osídľovací úrad pre Slovensko v Bratislave a Fond národnej obnovy 1945-1955
- Ca 200 000 listín
- Digitalizácia, výskum, konzervovanie
- Transkripcia strojopisných a rukopisných dokumentov
- Tvorba nového supermodelu transkripcie



~~3~~

~~Krátko po oslobodení Československa 19.5.1945 vydal prezident republiky E. Beneš dekrét o národnej správe majetkov Nemcov, Maďarov, zradcov a kolaborantov.~~

~~Zákonný článok 27 Zbierky zákonov a nariadení zo dňa 17.7.1945 potom tento tvrdý kurz voči Nemcom a Maďarom ešte viac skonkrétnil tým, že priamo výtýčil spôsob riešenia rodnostnej otázky odsunom Nemcov a Maďarov z územia republiky a reemigráciou zahraničných Slovákov a Čechov ako aj vnútornou kolonizáciou, ktorú mal riadiť na Slovensku novovytvorený Osídľovací úrad pre Slovensko so sídlom v Bratislave a v Čechách Osídľovací úrad pre Čechy a Moravu so sídlom v Prahe.~~

~~Oficiálnym cieľom tejto politiky bolo teda to, aby~~

Region1

1 -3-

2 Krátko po oslobodení Československa 19.5.1945 vydal  
3 prezident republiky E. Beneš dekrét o národnej správe majet-  
4 kov Nemcov, Maďarov, zradcov a kolaborantov.  
5 Zákonný článok 27 Zbierky zákonov a nariadení zo dňa-  
6 17.7.1945 potom tento tvrdý kurz voči Nemcom a Maďarom ešte  
7 viac skonkrétnil tým, že priamo výtýčil spôsob riešenia ná-  
8 rodnostnej otázky odsunom Nemcov a Maďarov z územia republi-  
9 ky a reemigráciou zahraničných Slovákov a Čechov ako aj vnú-  
10 tornou kolonizáciou, ktorú mal riadiť na Slovensku novovyt-  
11 vorený Osídľovací úrad pre Slovensko so sídlom v Bratisla-  
12 ve a v Čechách Osídľovací úrad pre Čechy a Moravu so sídlom  
13 v Prahe.  
14 Oficiálnym cieľom tejto politiky bolo teda to, aby sa  
15 dosiahlo navrátenie všetkých oblastí Československa pôvod-  
16 nému slovanskému obyvateľstvu. K tomu je treba dodať, že  
17 Slováci tak v Maďarsku ako aj v Juhoslávii nejavili veľký  
18 záujem o reemigráciu do vlasti a celá akcia sa skončila



## DIGIARCH\_SAKE\_00Ú a FNO Košice

I. OOÚ a FNO v Košiciach

A. Obežníky, vyhlášky, nariadenia OÚ, FNO, povereníctiev a ministerstiev.

Do tejto spisovej skupiny bola zaradená prevažne časť veľkého množstva cyklostylovaného materiálu, ktorý sa našiel medzi zachovanými archíviami. Obežníky bez dátumu sú obvyčajne zaradené na konci ročníka.

B. Rozhodnutia českosl... [ Show more ]

Hľadať



Pokročilé hľadanie

Zobraziť metadáta

Zoradiť od Dátum vytvorenia

Zobrazenie: Tabuľka

Vidieť ako...

Filtre



Zatiaľ žiadny filter.

Miniatúra (thumbnail)

Názov

Číselné

Dátum



SAKE\_00Ú a FNO Košice.xlsx



Združený inventár. Oblastné úradovne osídľovacích úradov na východnom Slovensku 1946 - ...