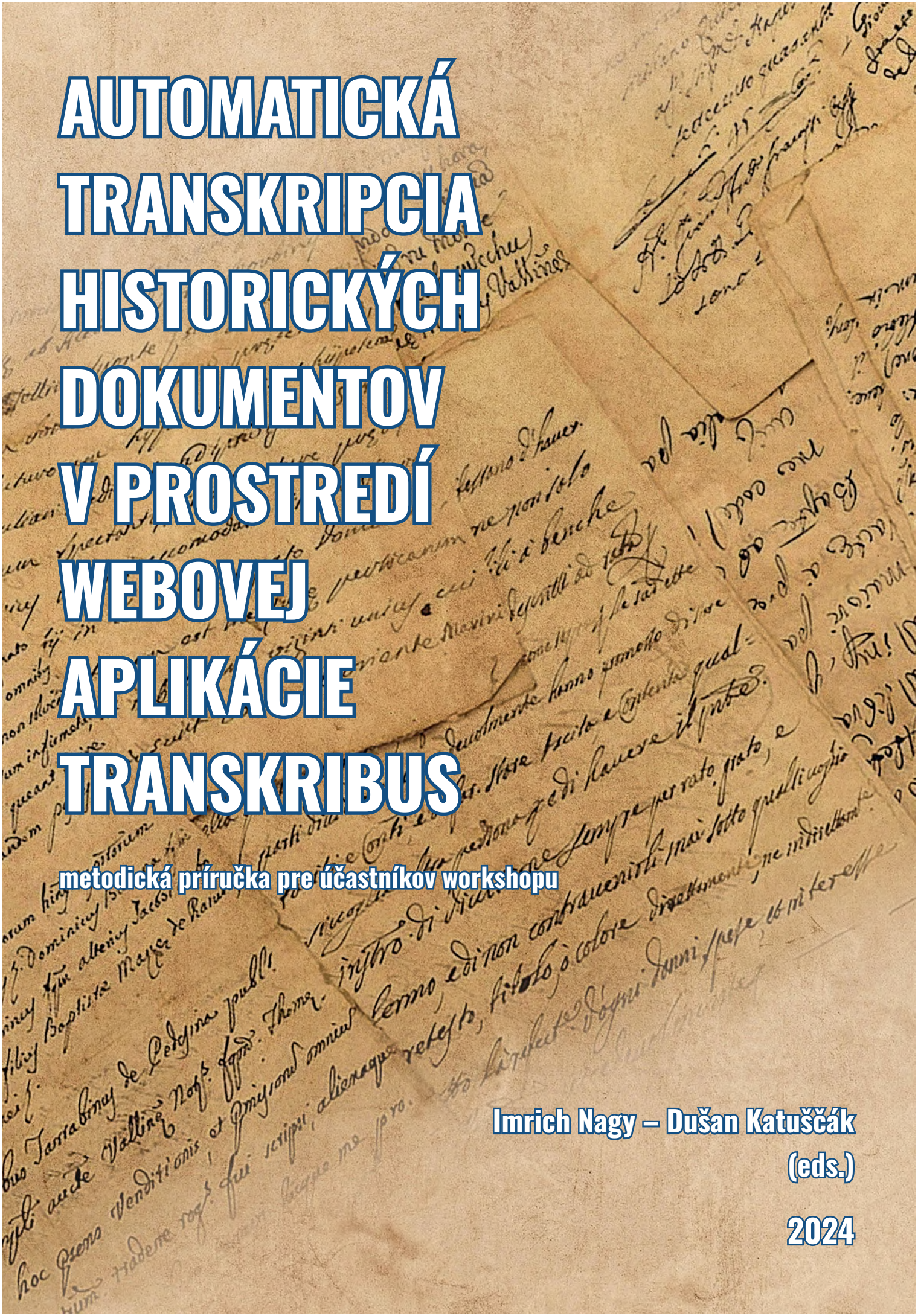


AUTOMATICKÁ TRANSKRIPCIA HISTORICKÝCH DOKUMENTOV V PROSTREDÍ WEBOVEJ APLIKÁCIE TRANSKRIBUS

metodická príručka pre účastníkov workshopu

Imrich Nagy – Dušan Katuščák
(eds.)

2024

Automatická transkripcia historických dokumentov v prostredí webovej aplikácie Transkribus

metodická príručka pre účastníkov workshopu

Mária Bôbová, Dušan Katuščák, Alica Kurhajcová, Pavol Maliniak, Michaela Mikušková, Imrich Nagy, Lucia Nižníková, Oto Tomeček

Elektronická metodická príručka je výstupom z riešenia projektu APVV-19-0456 SKRIPTOR – Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov (2020 – 2024).

Editori: Imrich Nagy, Dušan Katuščák

Autori: Mária Bôbová, Dušan Katuščák, Alica Kurhajcová, Pavol Maliniak, Michaela Mikušková, Imrich Nagy, Lucia Nižníková, Oto Tomeček

Jazyková korektúra: Lucia Nižníková

Grafická úprava: Miroslav Chladný

Verzia Transkribus app. 3.1.0.103

Táto práca bola podporená Agentúrou na podporu výskumu a vývoja na základe zmluvy č. *APVV-19-0456 SKRIPTOR – Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov*.

V metodike boli využité poznatky a výsledky z výskumov podporených v rámci Studentské grantové soutěže na Slezské univerzitě v Opavě (SGS 2022, SGS 2023, SGS 2024) a datasetov študentov Filozoficko-přírodovědecké fakulty pod vedením Dušana Katuščáka.

© BELIANUM. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici 2024

DOI: https://doi.org/10.24040/2024.9788055721439



Táto publikácia je šírená pod licenciou Creative Commons Attribution 4.0 International Licence CC BY (uvedenie autora).

ISBN 978-80-557-2143-9

Obsah

Slovo na úvod 6

1 Webová platforma Transkribus app 8
1.1 Pripojenie na internet 8
1.2 Vstup cez webový prehliadač 8
1.3 Čo Transkribus umožňuje 10

2 Otvorenie aplikácie Transkribus app 12
2.1 Registrácia 12
2.2 Hlavná plocha 13
2.2.1 Záložka *Desk* 13
2.2.2 Záložka *Models* 14
2.2.3 Záložka *Sites* 15
2.2.4 Záložka *Jobs* 18
2.2.5 Záložka *User* 18

3 Zbierka 19
3.1 Vytvorenie zbierky 19
3.1.1 Nová zbierka 22
3.1.2 Názvy zbierok 23
3.1.3 Kontrola kvality pred importom 23
3.1.4 Zálohovanie. Archivovanie 23
3.2 Správa používateľov zbierky 24

4 Príprava dokumentu na automatickú transkripciu 26
4.1 Kritériá výberu digitalizátov 26
4.2 Popis fondov, zbierok a dokumentov 27
4.3 ScanTent a DocScan pre archívy a knižnice 28
4.3.1 ScanTent 29
4.3.2 Aplikácia DocScan 31
4.3.3 Bezpečnosť údajov v aplikácii DocScan 33
4.3.4 Snímanie pomocou zariadenia ScanTent a aplikácieq DocScan 33
4.3.5 Práca s aplikáciou DocScan 34
4.3.5.1 Odoslanie dokumentu na platformu Transkribus 36
4.3.5.2 Nastavenia 36
4.3.5.3 Automatické orezávanie, otáčanie a mazanie 37
4.3.5.4 Manuálne orezanie 38

4.4 Snímanie – zhrnutie 39
4.4.1 Snímanie 39
4.4.2 Snímanie v praxi 39
4.4.3 Formáty obrázkov: formát JPG, JPEG 39
4.4.4 Pixel – základ pre ukladanie digitálneho obrazu 39
4.4.5 Príklad štruktúry konvolučnej siete 40
4.5 Import digitalizátov do webovej aplikácie platformy Transkribus 41
4.5.1 Privátnosť zbierok a dokumentov 45

5 Segmentácia dokumentov v Transkribuse 46
5.1 Spôsoby segmentácie 46
5.1.1 Výber strán 48
5.1.2 Automatická segmentácia 50
5.1.3 Manuálna segmentácia 51
5.1.4 Výber modelu 57
5.1.5 Pokročilé nástroje na nastavenie automatickej segmentácie textu 57
5.1.6 Automatická segmentácia a rozpoznávanie textu 60
5.2 Opravy po automatickej a manuálnej segmentácii 63
5.2.1 Korekcia textových rámcov (*Text Regions*) 63
5.2.2 Korekcie riadkov (*Lines*) 67
5.2.3 Kontrola a úprava poradia čítania textových rámcov a riadkov 72
5.2.3.1 *Visibility* (viditeľnosť objektov segmentácie) 73
5.2.3.2 *Layout* 75
5.2.3.3 Práca so stĺpcami 77
5.2.3.4 Vkladanie medziriadkov 78
5.3 Segmentácia tabuliek 82

6 Tvorba modelu automatickej transkripcie 85
6.1 Prepis dokumentu (príprava vzorky *Ground Truth*) 85
6.2 Spustenie trénovania modelu 91
6.3 Úspešnosť modelu a jeho zdokonaľovanie 96
6.3.1 Vyhodnotenie úspešnosti modelu 96
6.3.2 Zdokonaľovanie modelu 97
6.4 Supermodely 100
6.4.1 Výhody používania supermodelov 100
6.4.2 Slovenský supermodel M1 pre rukopisy (SSM1) 100

7 Priebeh automatickej transkripcie v aplikácii Transkribus 102
7.1 Výber dokumentu na automatickú transkripciu 102
7.2 Výber snímok na automatickú transkripciu 103

7.3 Výber nastavení na automatickú transkripciu 104
7.4 Výber modelu na automatickú transkripciu 105
7.5 Spustenie automatickej transkripcie 108
7.6 Výsledok automatickej transkripcie 109
7.7 Kontrola kreditov a systém spoplatnenia automatickej transkripcie 109

8 Možnosti práce s textom po automatickej transkripcii 111
8.1 Textové tagy 111
8.1.1 Priradenie textového tagu 112
8.1.2 Ostatné textové tagy 114
8.1.3 Vytvorenie textového tagu 114
8.2 Štrukturálne tagy 115
8.2.1 Zviditeľnenie štrukturálnych tagov 115
8.2.2 Správa štrukturálnych tagov 116
8.2.3 Priradenie štrukturálneho tagu 117
8.2 Export výstupov 118
8.3.1 Štandardné možnosti exportu 118
8.3.2 Prístup k exportovaným súborom 120

Slovník pojmov 121

Použité zdroje 129

Slovo na úvod

Technologický pokrok vo využívaní nástrojov strojového učenia (*Machine Learning*) a umelej inteligencie AI (*Artificial Intelligence*) sa postupne stáva súčasťou našej každodennosti a vedome či nevedome sme s ním konfrontovaní aj pri rôznych špecifických odborných činnostiach, pri ktorých bolo nahradenie vedomostí a zručností človeka strojom donedávna nepredstaviteľné. Neznamená to však, že by sa ľudská odbornosť stávala zbytočnou. Sú to práve invenčnosť, zručnosť a um človeka, ktoré využili a adaptovali existujúce i novo sa vyvíjajúce technológie na zvládanie presne definovateľných, rutinných a opakujúcich sa algoritmov. Takým procesom je aj zostavovanie textov na ľubovoľné témy prostredníctvom chatovacích robotov, ktoré je v súčasnosti až emblematickým symbolom pokroku vo využiteľnosti AI.

Tak trochu v tieni týchto populárnych nástrojov zostávajú dlhodobo vyvíjané a v praxi overované nástroje AI schopné vskutku mimoriadnym spôsobom zmeniť a v podstate nanovo zadefinovať vysoko odborné činnosti v jednotlivých profesiách. Pozoruhodným príkladom toho je aj platforma Transkribus vyvinutá vďaka multilaterálnej spolupráci významných európskych vedeckých inštitúcií v rámci projektov *tranScriptorium* (2013 – 2015) a *READ* (2016 – 2019) financovaných z programov EÚ. Lídrom tejto spolupráce je Univerzita v Innsbrucku a vedúcou postavou Dr. Günter Mühlberger, ktorí výsledky predchádzajúcich projektov pretavili do veľmi dynamicky sa vyvíjajúceho a účinného nástroja na automatickú transkripciu dokumentov v rukopisnej aj tlačenej podobe ľubovoľnej geografickej, historickej či jazykovej proveniencie. Vďaka tomu je v súčasnosti Transkribus verejne dostupný komerčný produkt, ktorý prostredníctvom združenia READ-COOP European Cooperative Society ponúka všetkým individuálnym a inštitucionálnym záujemcom riešenie pre vskutku efektívnu digitalizáciu historických dokumentov s plnotextovými digitálnymi výstupmi v najrozličnejších formátoch podľa požiadavky zadávateľa. Pre pamäťové inštitúcie a ich používateľov z radov laickej i odbornej verejnosti je to doslova revolučná zmena, ktorá zásadným spôsobom do budúcnosti zmení ich prácu a má potenciál priniesť mimoriadne výsledky v poznaní a sprístupňovaní našej histórie a nehmotného kultúrneho dedičstva.

Univerzita Mateja Bela v Banskej Bystrici sa v spolupráci so Štátnou vedeckou knižnicou v Banskej Bystrici v rámci projektu *SKRIPTOR – Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov podporeného Agentúrou pre podporu vedy a výskumu (APVV-19-0456)* podujala aplikovať nástroj Transkribus na slovacikálne historické rukopisné a tlačené dokumenty a overiť jeho využiteľnosť v podmienkach slovenských pamäťových inštitúcií (archívov). Na Slovensku ide o ojedinelý pilotný projekt, ktorý môže otvoriť cestu k žiaducemu nasadeniu moderných technológií pri digitalizačných projektoch našich pamäťových inštitúcií v záujme širokého sprístupnenia informácií z digitalizovaných dokumentov a ich ďalšieho odborného využitia.

V rámci riešenia projektu sa už podarilo vytvoriť desiatku funkčných modelov na automatickú transkripciu slovacikálnych rukopisov zo 16. – 20. storočia a tiež aj historických tlačí. Významnou pridanou hodnotou je nadobudnutie know-how na prácu s platformou Transkribus. Jeho sprostredkovanie záujemcom a zástupcom pamäťových inštitúcií zo Slovenska formou workshopov považujeme za zmysluplné zavŕšenie nášho snaženia. Za týmto účelom sme zostavili metodickú príručku, ktorá v postupných krokoch predstavuje jednotlivé fázy digitalizácie a automatickej transkripcie dokumentu na platforme Transkribus.

Pokrok vo vývoji aplikácie Transkribus v rokoch 2023 a 2024 je poznamenaný postupným ukončovaním podpory desktopovej aplikácie – Transkribus expert klienta. V ekosystéme Transkribus aktuálne prebieha podstatná a zásadná zmena. Namiesto expert klienta, v ktorom sme

v rámci výskumu SKRIPTOR získavali prvé skúsenosti, a v ktorom sme tvorili prvé modely transkripcie, sa začíname zoznamovať s webovou aplikáciou Transkribus. Novú aplikáciu už nie je potrebné inštalovať na lokálne počítače. Práca s ňou prebieha výlučne cez webový prehliadač prostredníctvom internetu.

Preto sme sa rozhodli pôvodnú metodickú príručku doplniť a upraviť na prácu s webovou aplikáciou. Metodika predstavuje jednotlivé fázy digitalizácie a automatickej transkripcie dokumentu na platforme Transkribus.

Na tomto mieste je dôležité upozorniť, že platforma Transkribus sa stále vyvíja. Na stránkach https://readcoop.eu/transkribus/ sa nachádzajú manuály a videá na prácu s Transkribusom. Niektoré inštrukcie a názorné ukážky, ktoré boli aktuálne v minulosti a podľa ktorých boli inštrukcie pripravené, nemusia odrážať vlastnosti a funkcie nových verzií.

Veríme, že využitie možností AI vo forme práce s Transkribusom prinesie novú dynamiku do digitalizácie, uchovávania a sprostredkovania nehmotného kultúrneho dedičstva na Slovensku.

Dušan Katuščák – Imrich Nagy

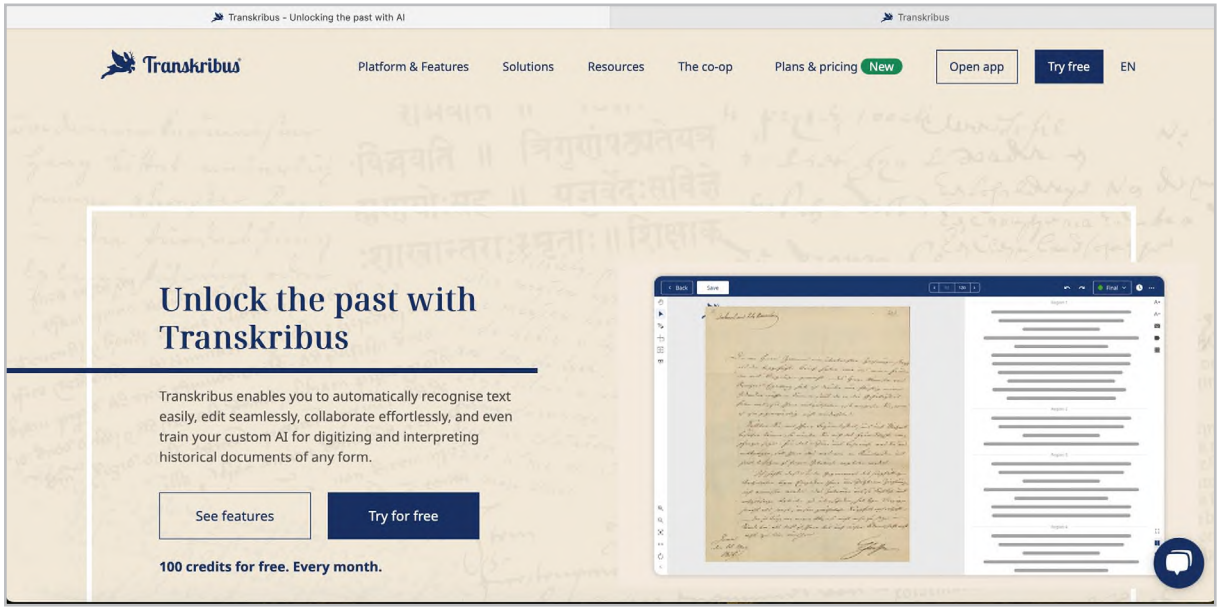
1 Webová platforma Transkribus app

1.1 Pripojenie na internet

Na prácu na platforme Transkribus je potrebné mať k dispozícii nepretržité vysokorýchlostné pripojenie na internet. Všetky vaše úkony sa budú robiť v režime vzdialeného prístupu na serveroch platformy Transkribus. Súbory a verzie strán, s ktorými pracujete, sa ukladajú na serveroch platformy. Výhodou je, že k nim máte prístup z ktoréhokoľvek miesta a ktoréhokoľvek počítača (cez webový prehliadač alebo prostredníctvom nainštalovanej aplikácie).

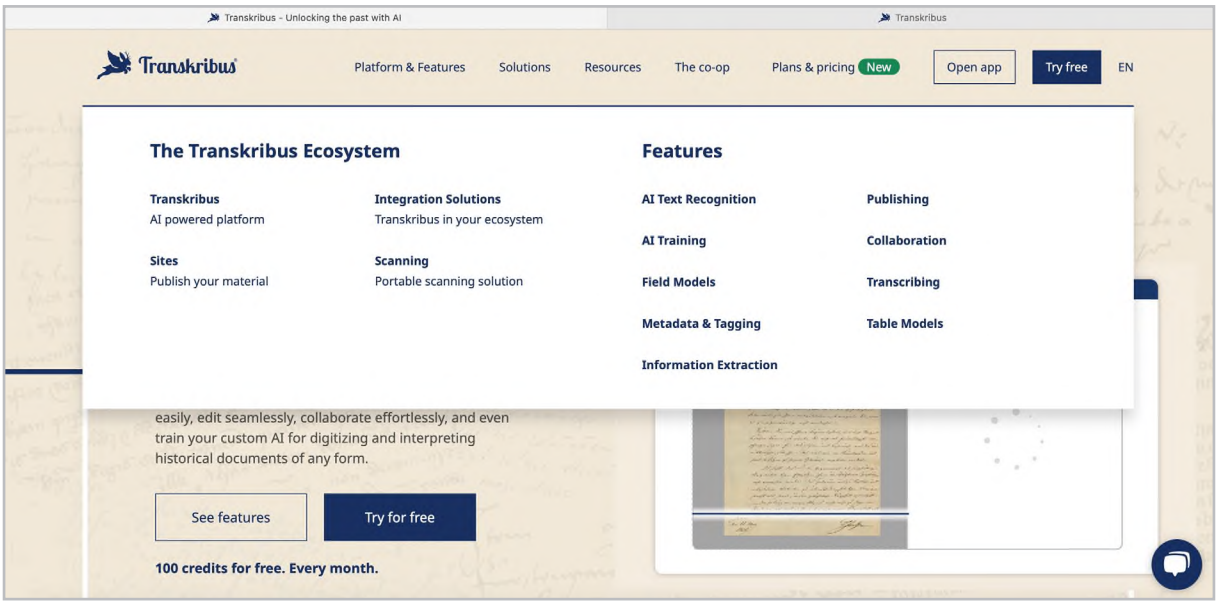
1.2 Vstup cez webový prehliadač

Do webovej platformy Transkribus app vstúpime cez webový prehliadač (napr. Chrome, Edge, Safari; odporúčame Chrome) zadaním adresy https://www.transkribus.org/. Transkribus app nie je potrebné inštalovať.



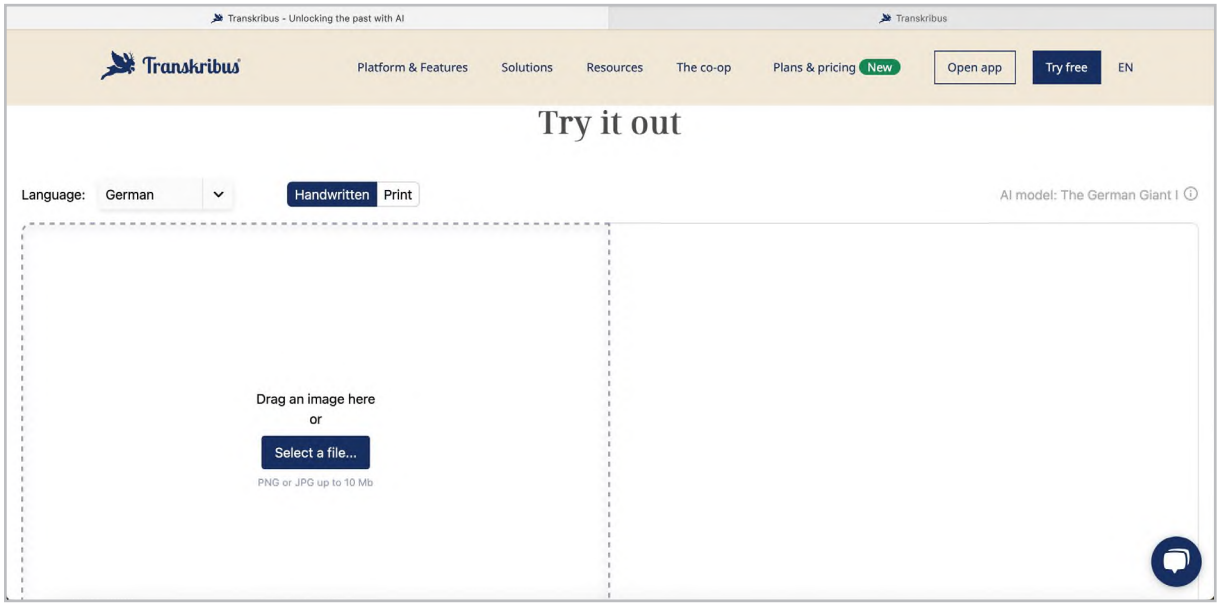
Obrázok 1 Úvodná propagačná stránka https://www.transkribus.org/

Na hornej lište obrazovky sa nachádzajú voľby, cez ktoré získate *základné informácie o e*kosystéme Transkribus a to: Platforma a funkcie (*Platform & Features*), Riešenia (*Solutions*), Zdroje (*Resources*), Obchodné združenie (*The co-op*), Plány a ceny (*Plans & pricing*). V okienku vpravo hore sú umiestnené voľby Otvoriť aplikáciu (*Open app*), Vyskúšať bezplatne (*Try for free*) a jazyk (*EN*) pre možnosť vrátiť sa k angličtine alebo zvoliť iný jazyk rozhrania. Odporúčame postupne sa aspoň zbežne oboznámiť s obsahom týchto položiek. V dolnej časti sa nachádza osobitná voľba Pozrieť funkcie (*See features*) a *Vyskúšať* bezplatne (*Try for free*).



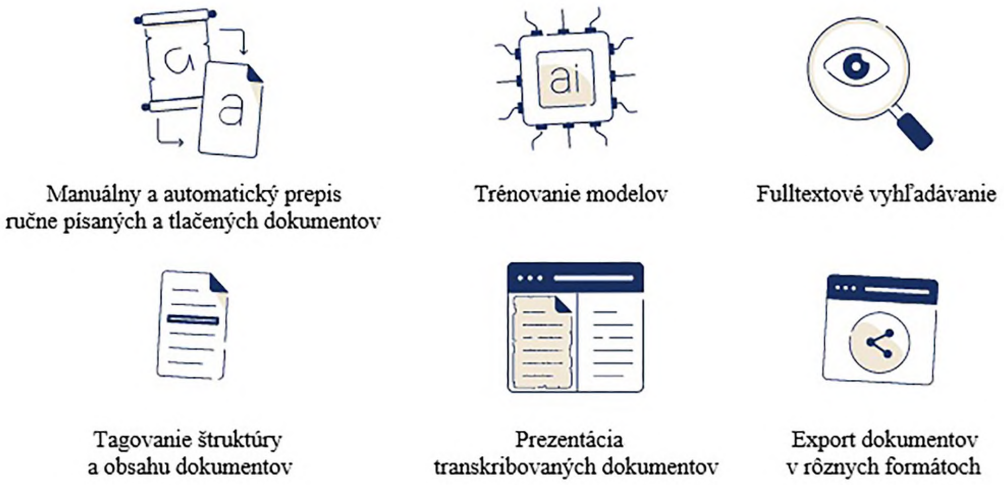
Obrázok 2 Prístup k jednotlivým častiam ekosystému Transkribus v prostredí webovej aplikácie

Rolovaním stránky sa dostanete k možnosti vyskúšať zdarma *rýchlu transkripciu s použitím niektorého verejného modelu, a to tak, že do okna vložít*e zo schránky alebo zo svojho počítača obrázok – digitalizát (digitálne faksimile) vo formáte JPG alebo PNG *s veľkosťou* maximálne 10 MB. M*ôžete* zvoliť jazyk dokumentu a vybrať typ dokumentu, teda či ide o rukopis (*Handwritten*) alebo tlač (*Print*). Pre niektoré jazyky, pre ktoré ešte neexistujú verejné modely transkripcie, je voľba typu dokumentu nedostupná.



Obrázok 3 Možnosť bezplatne vyskúšať transkripciu

1.3 Čo Transkribus umožňuje



Obrázok 4 Možnosti platformy Transkribus (základný prehľad)

Funkcie a návody aplikácie Transkribus

Prvé kroky v Transkribuse https://help.transkribus.org/getting-started

1. Registrácia a prehľad používateľského rozhrania
2. Vytvorenie zbierky
3. Nahrávanie súborov
4. Použitie kreditu

Rozpoznávanie rozloženia https://help.transkribus.org/layout-recognition

1. Automatické rozpoznanie rozloženia
2. Rozšírené nastavenia konfigurácie rozloženia
3. Manuálna úprava rozloženia
4. Modely základných čiar
5. Modely polí
6. Tabuľkové modely
7. Noviny
8. P2PaLA

Rozpoznávanie textu https://help.transkribus.org/text-recognition

1. Automatická transkripcia dokumentov
2. Výber modelu
3. Verejné modely
4. Supermodely

Trénovanie modelov https://help.transkribus.org/training-models

1. Trénovanie modelov rozpoznávania textu
2. Príprava dát
3. Nastavenie modelu a trénovanie
4. Miera chybovosti znakov a krivka učenia
5. Presnosť výpočtu chybovosti
6. Pretrénovanie pomocou technológie PyLaia

Tagovanie https://help.transkribus.org/tagging

1. Tagovanie
2. Štrukturálne tagy
3. Textové tagy

Vyhľadávanie https://help.transkribus.org/searching

1. Možnosti vyhľadávania
2. Fulltextové vyhľadávanie
3. Fuzzy vyhľadávanie
4. Inteligentné vyhľadávanie

Správa dát https://help.transkribus.org/management

1. Správa zbierok
2. Správa dokumentov
3. Správa používateľov – roly a povolenia na prístup
4. Správa modelov
5. Virtuálna klávesnica

Sťahovanie (export dát) https://help.transkribus.org/downloading

Sites (nástroj na prezentáciu výsledkov) https://help.transkribus.org/sites

1. Nastavenie stránky Sites
2. Sites: O stránke
3. Sites: Preskúmať
4. Sites: Viacjazyčná podpora

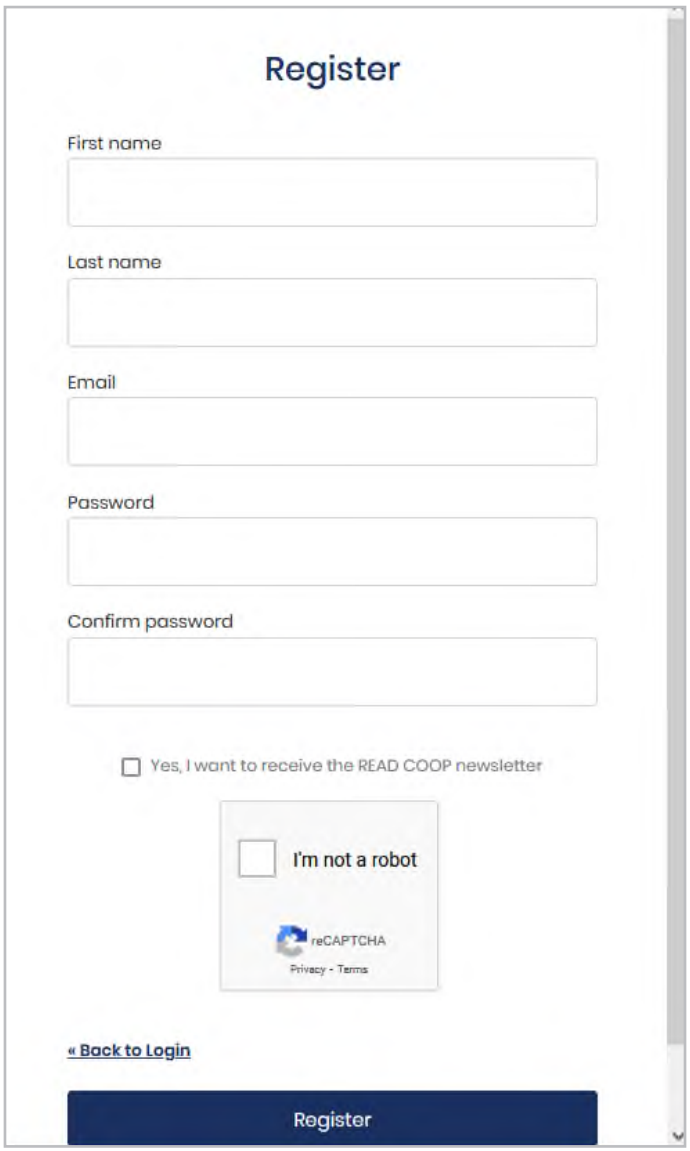
2 Otvorenie aplikácie Transkribus app

Profesionálnu prácu s aplikáciou Transkribus začnite kliknutím na voľbu Otvoriť aplikáciu (*Open app*).

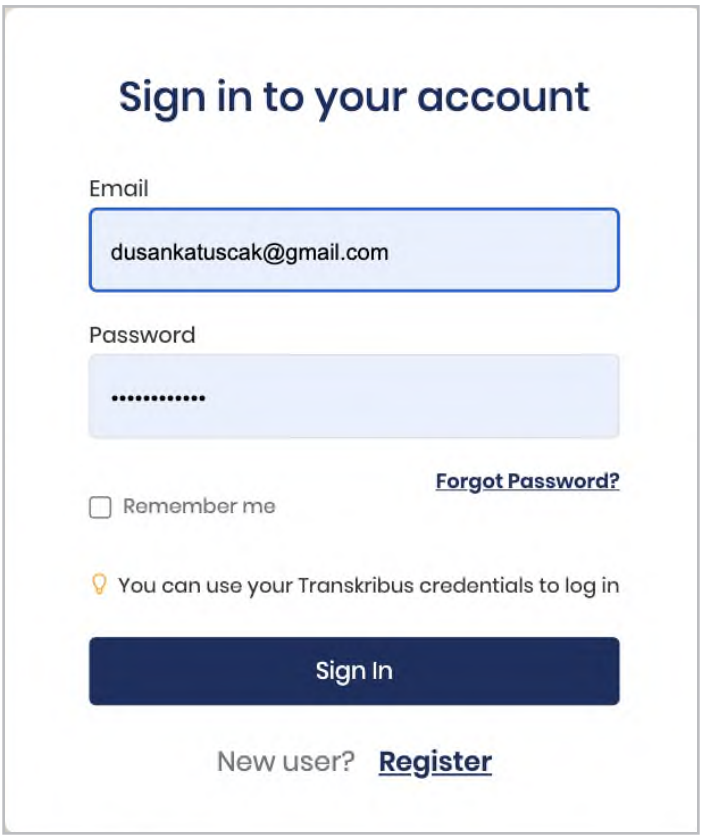
2.1 Registrácia

Na prácu s platformou Transkribus sa treba registrovať. Registrácia účtu Transkribus je jednoduchá, stačí sa bezplatne zaregistrovať. Ak ste sa zaregistrovali, máte vytvorený vlastný účet a s ním máte prístup:

1. na platformu Transkribus,
2. do aplikácie Transkribus vo webovom prehliadači.



Obrázok 5 Registrácia. Okno na zápis registračných údajov používateľa na platformu Transkribus app



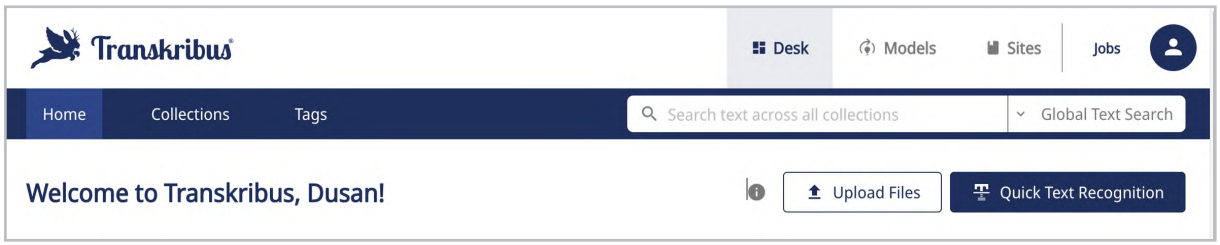
Obrázok 6 Prihlásenie do účtu Transkribus

2.2 Hlavná plocha

Po prihlásení sa do aplikácie Transkribus sa na adrese https://app.transkribus.org/home zobrazí hlavná plocha aplikácie – stránka *Desk*.

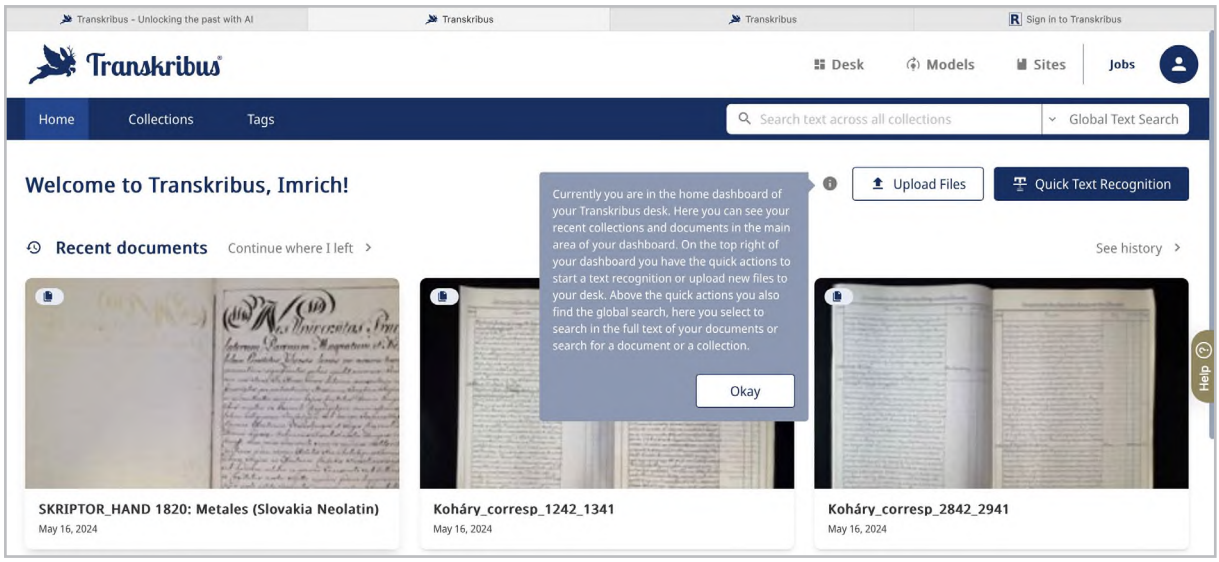
Vedľa nej sa vpravo hore nachádzajú záložky Modely transkripcie (*Models*), nástroj *Sites*, Úlohy (*Jobs*) a silueta používateľa.

2.2.1 Záložka *Desk*



Obrázok 7 Hlavná stránka záložky Desk

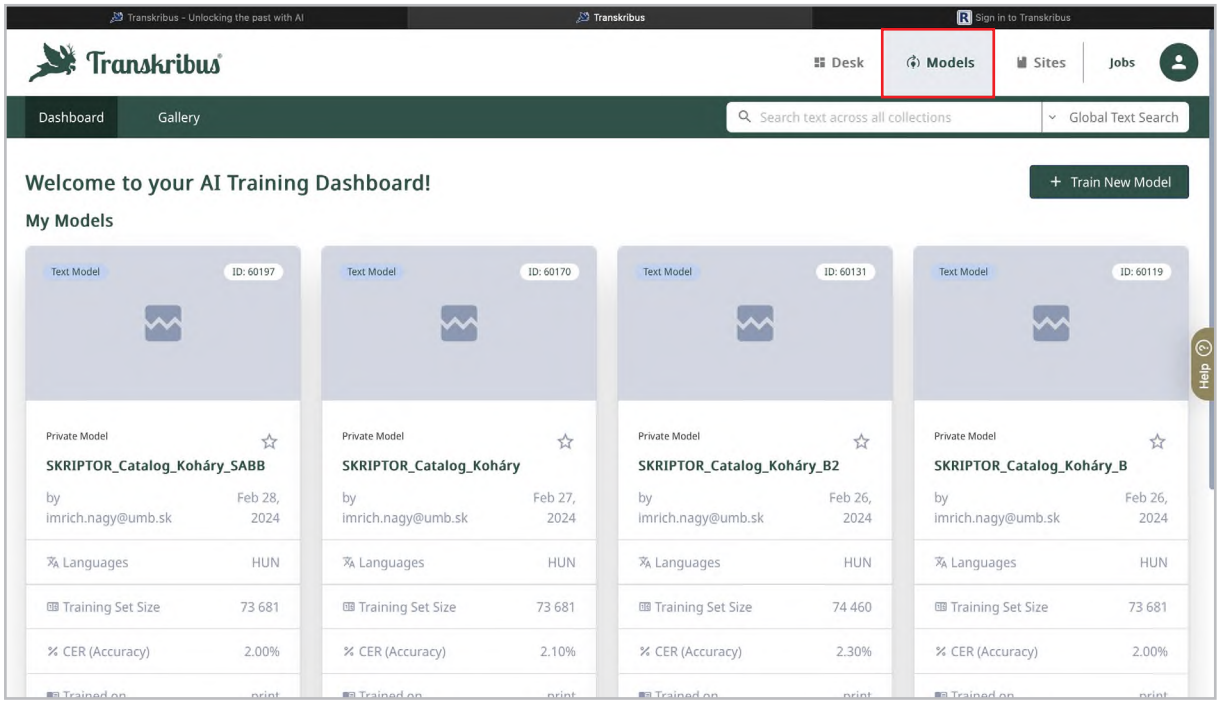
Keď začínate pracovať s aplikáciou Transkribus a po prihlásení máte otvorenú hlavnú plochu s funkciou *Nahrať súbory*, súbory nenahrávajte. Najprv si vytvorte zbierku. Súbory, dokumenty, potom nahráte do zbierky.



Obrázok 8 Záložka Desk a najnovšie dokumenty, s ktorými ste pracovali

Na tmavomodrej lište sa nachádzajú voľby Domov (*Home*), Zbierky (*Collections*), Značky/ Tagy (*Tags*) a okná na vyhľadávanie vo všetkých vašich zbierkach a tiež globálne vyhľadávanie zbierok alebo dokumentov podľa názvov.

2.2.2 Záložka *Models*

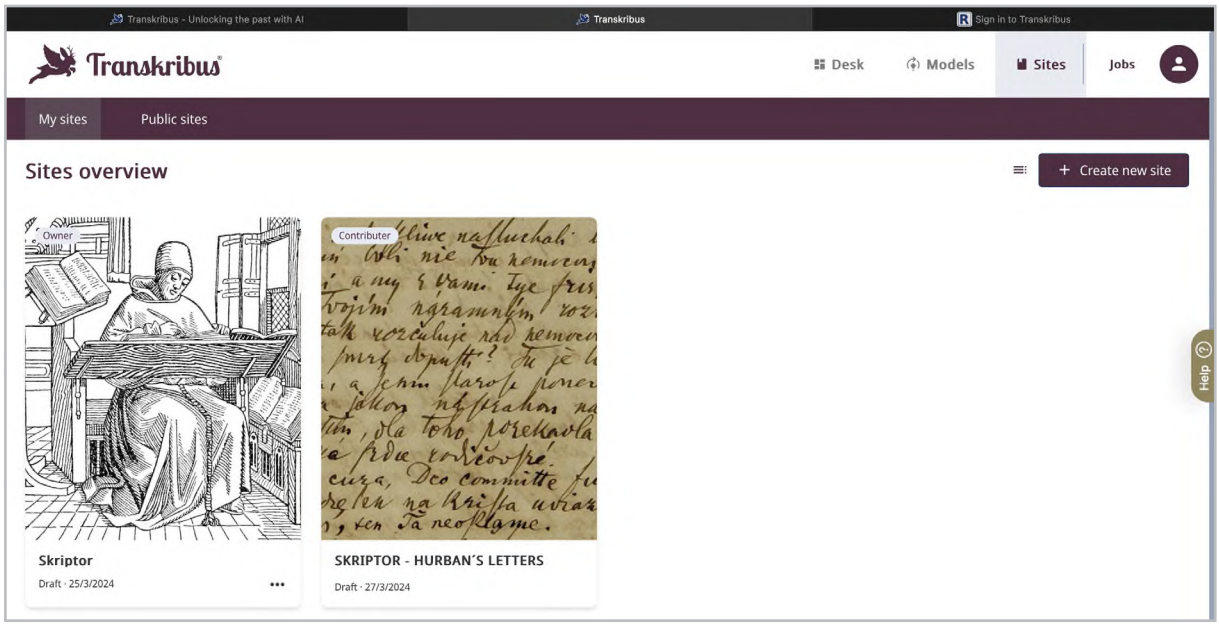


Obrázok 9 Záložka Models

Na tejto záložke sa zobrazia sa všetky vaše alebo s vami zdieľané modely transkripcie. Na stránke je dostupné *identifikačné číslo modelu (ID)*, základný popis modelu a možnosť *úpravy metadát* popisujúcich model a možnosť upraviť *popis modelu s ukážkami digitalizátov*.

Na tejto záložke sa nachádza aj voľba +Trénovať nový model (*+Train New Model)*. Túto možnosť vyberte v prípade, že máte pripravené dáta na trénovanie nového modelu, ideálne v kvalite prepisu *Ground Truth*.

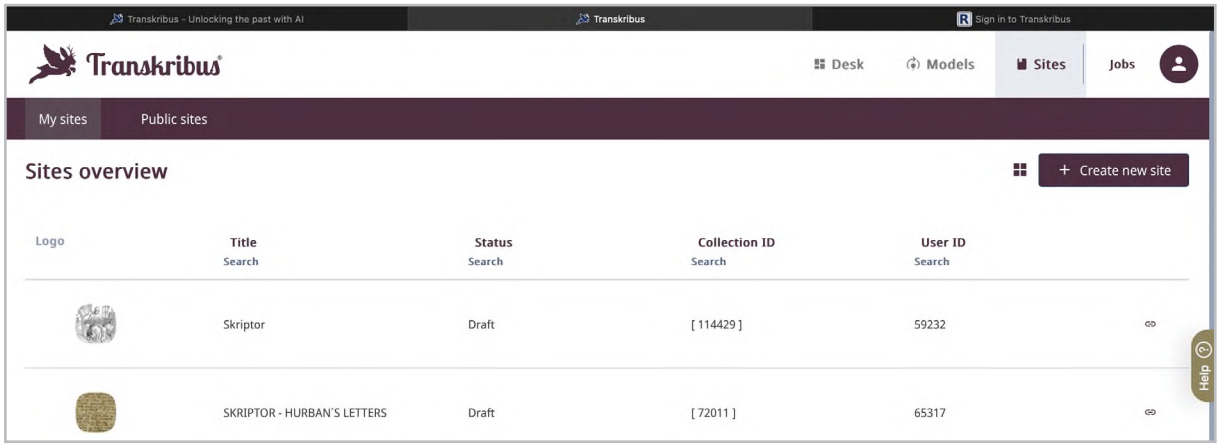
2.2.3 Záložka *Sites*



Obrázok 10 Záložka Sites

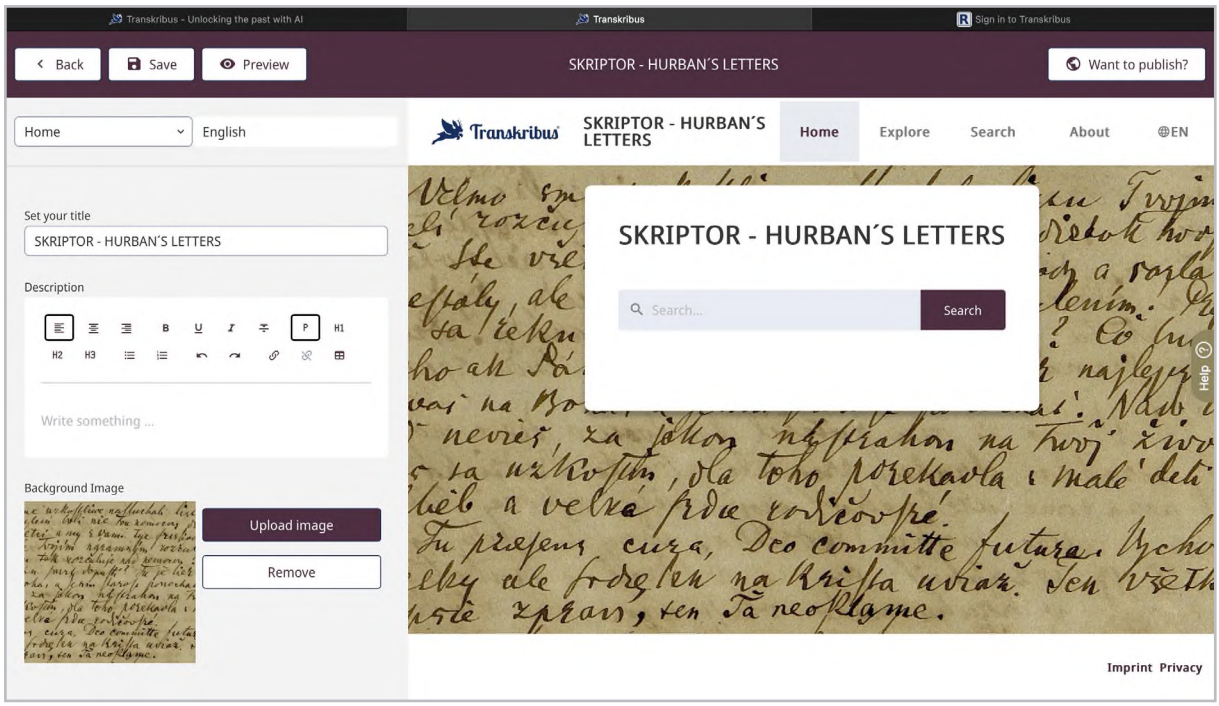
Táto záložka umožňuje prístup do vašich osobných webových stránok, ktoré slúžia na sprístupnenie vašich zbierok a dokumentov pre širokú verejnosť na internete. Webové stránky sú buď privátne alebo verejné.

Cez funkciu vytvorenia novej stránky (*+Create new site*) sa dostanete k popisu zvolenej stránky, v danom príklade k popisu stránky A. Kurhajcovej SKRIPTOR – Hurbanove listy.



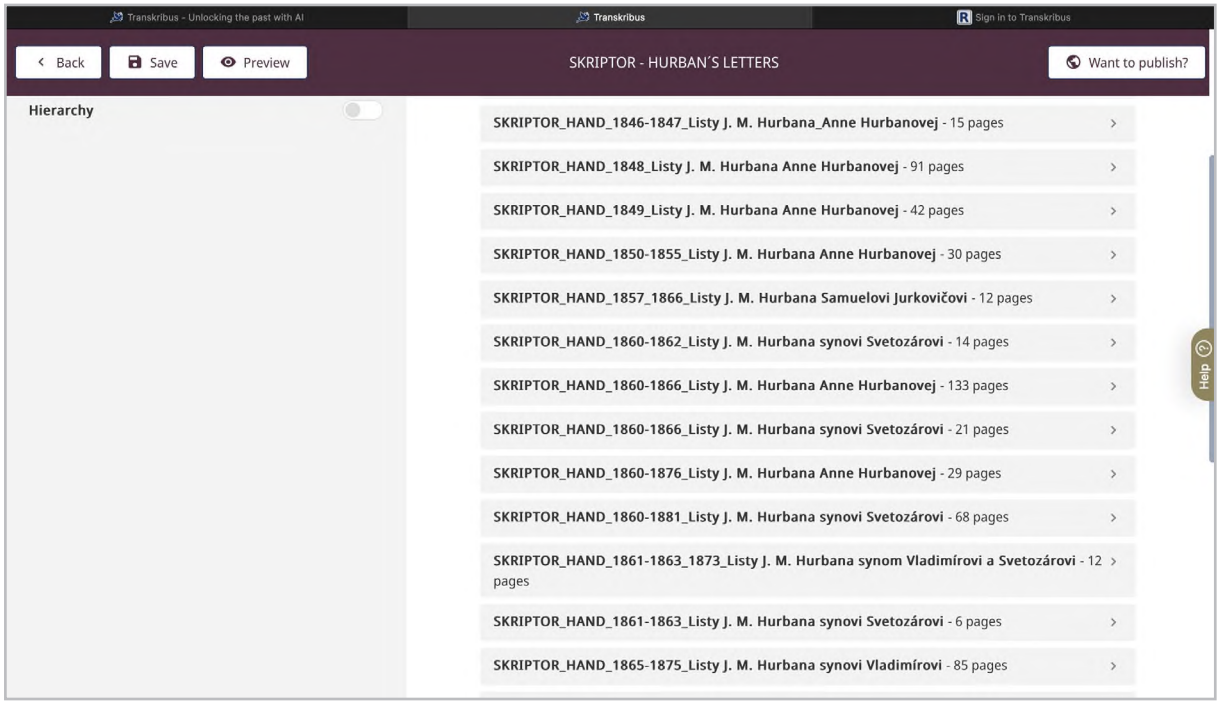
Obrázok 11 Prehľad vašich stránok na Sites sprístupňujúcich zbierky

Prehľad obsahuje zoznam vašich stránok, logo/logá, názvy, status, ID zbierky, ID používateľa. V poslednom stĺpci sa nachádza symbol hyperlinku. Po kliknutí na tento symbol sa otvorí stránka tak, ako ju môže vidieť verejnosť na internete, ak sa vlastník zbierky rozhodol stránku zverejniť.



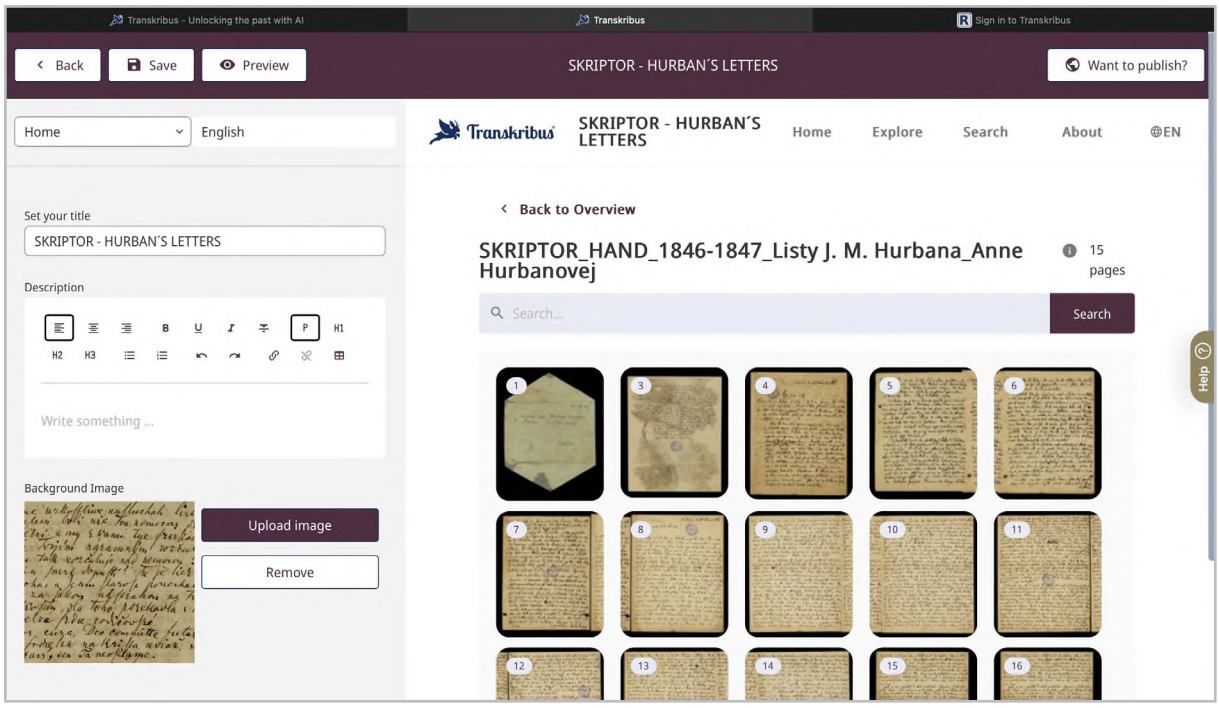
Obrázok 12 Ukážka možností tvorby vlastnej stránky v Sites na prezentáciu zbierky na internete

Cez voľbu Náhľad (*Preview*) a následne Preskúmať (*Explore*) sa dostanete k samotnému zoznamu jednotlivých dokumentov patriacich do zbierky, ktorú zverejňujete.

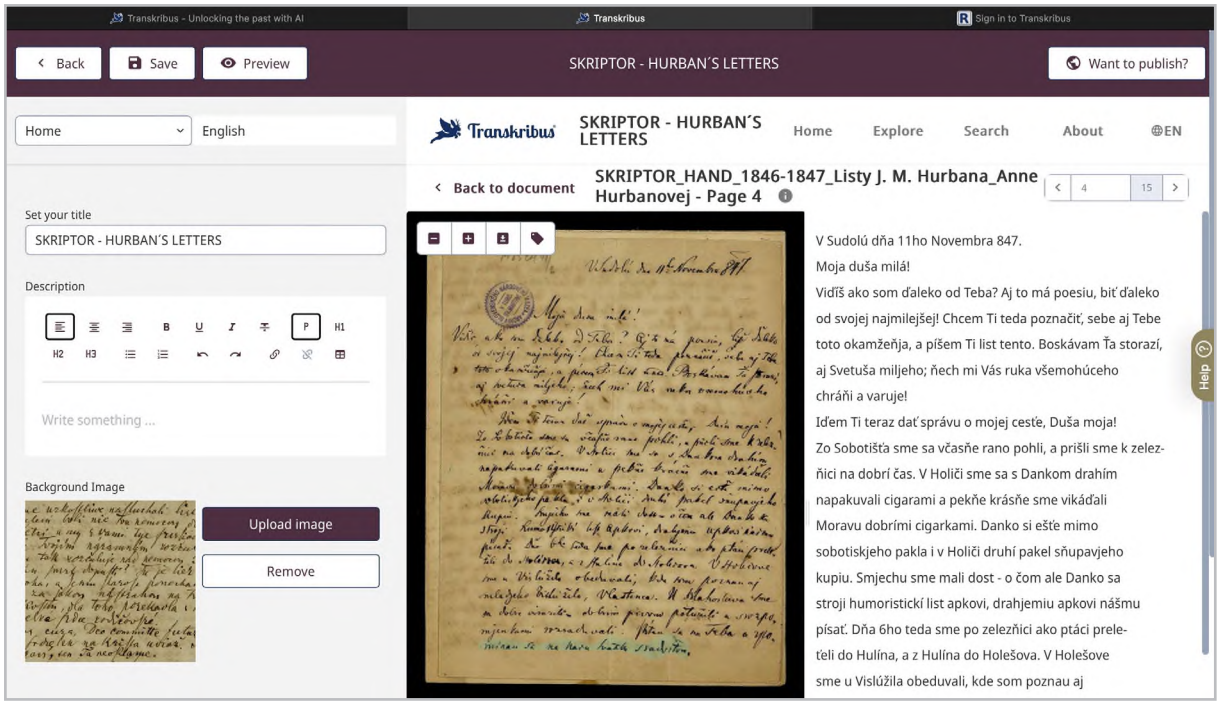


Obrázok 13 Zoznam dokumentov v Sites určených na zverejnenie na internete

Kliknutím na šípku vpravo od názvu dokumentu sa zobrazia strany dokumentu ako digitalizáty.

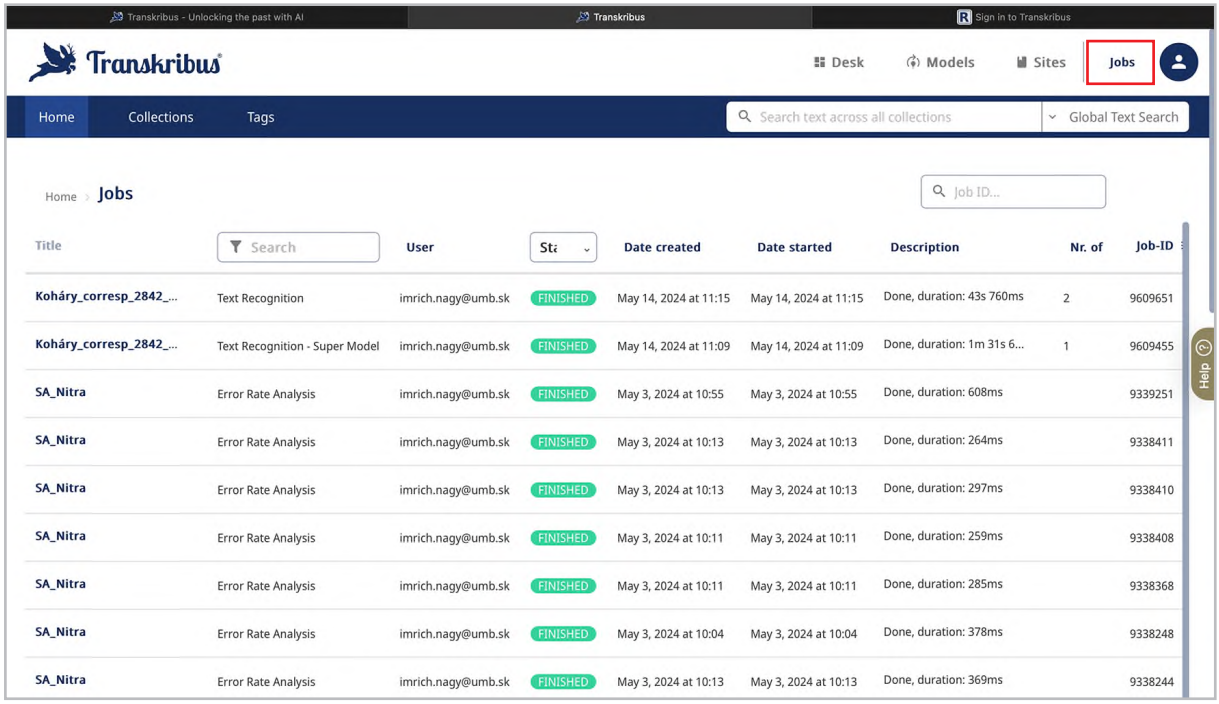


Obrázok 14 Zobrazenie strán dokumentu



Obrázok 15 Vľavo zobrazenie strany ako digitalizátu a vpravo jej transkripcia

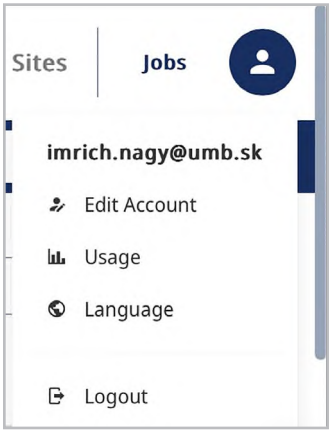
2.2.4 Záložka *Jobs*



Obrázok 16 Ukážka s prehľadom úloh (Jobs), ktoré používateľ robil na serveri

2.2.5 Záložka *User*

Kliknutím na symbol siluety získate prístup k informáciám o svojom účte. Môžete upraviť svoj účet, informácie o použití a o jazyku. Tu sa nachádza aj možnosť odhlásiť sa z aplikácie.

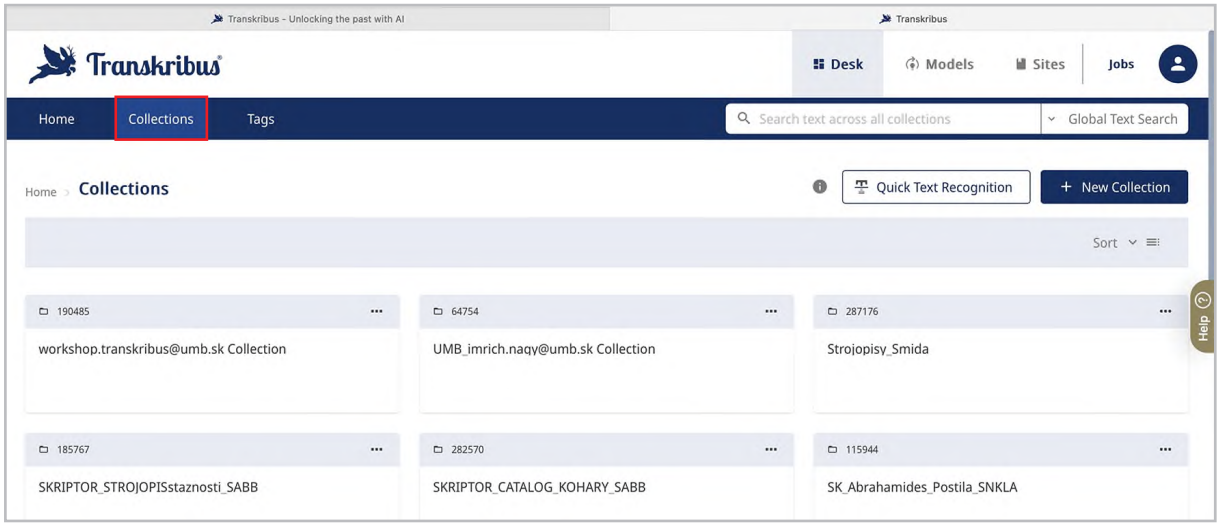


Obrázok 17 Silueta User

3 Zbierka

3.1 Vytvorenie zbierky

Skôr ako začnete pracovať s aplikáciou Transkribus, vytvorte si svoju zbierku, prípadne niekoľko zbierok. Na tmavomodrej lište kliknite na voľbu Zbierky (*Collections*).



Obrázok 18 Zobrazenie zbierok po kliknutí na voľbu Collections

Otvorí sa zoznam všetkých vašich a s vami zdieľaných zbierok. Zoznam sa zobrazí buď ako mozaika – položky vedľa seba, alebo kliknutím na ikonu troch čiarok vpravo funkciou +Nová zbierka (*+New Collection*) zobrazíte zbierky ako jednoduchý zoznam zbierok s počtom dokumentov v nich.

Zbierku môže vymazať oprávnený používateľ, ktorým je vlastník zbierky (*owner*) cez ikonu troch bodiek vedľa ID zbierky.

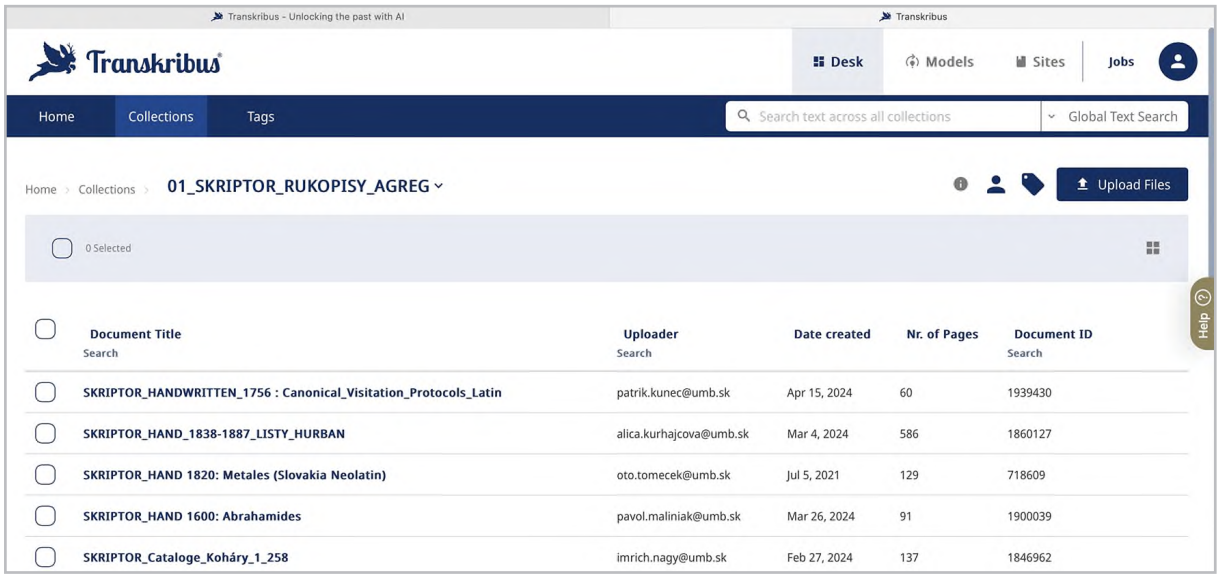
Dôležité je pochopiť, ako sú zbierky a dokumenty štruktúrované. Na obrázku nižšie vidieť logickú štruktúru zbierok v aplikácii Transkribus.

Zbierka A													
Dokument 1													
Strana 1	Strana 2	Strana 3	Strana 4	Strana 5	Strana 6	Strana 7	Strana 8	Strana 9	Strana 10	Strana 11	Strana 12	Strana 13	Strana ...n
Dokument 2													
Strana 1	Strana 2	Strana 3	Strana 4	Strana 5	Strana 6	Strana 7	Strana 8	Strana 9	Strana 10	Strana 11	Strana 12	Strana 13	Strana ...n
Dokument 3													
Strana 1	Strana 2	Strana 3	Strana 4	Strana 5	Strana 6	Strana 7	Strana 8	Strana 9	Strana 10	Strana 11	Strana 12	Strana 13	Strana ...n
Zbierka B													
Dokument 1													
Strana 1	Strana 2	Strana 3	Strana 4	Strana 5	Strana 6	Strana 7	Strana 8	Strana 9	Strana 10	Strana 11	Strana 12	Strana 13	Strana ...n
Dokument 2													
Strana 1	Strana 2	Strana 3	Strana 4	Strana 5	Strana 6	Strana 7	Strana 8	Strana 9	Strana 10	Strana 11	Strana 12	Strana 13	Strana ...n
Dokument 3													
Strana 1	Strana 2	Strana 3	Strana 4	Strana 5	Strana 6	Strana 7	Strana 8	Strana 9	Strana 10	Strana 11	Strana 12	Strana 13	Strana ...n

Obrázok 19 Štruktúra zbierok na platforme Transkribus

Zbierku možno chápať ako *priečinok* obsahujúci dokumenty. Zbierky sa zvyčajne používajú na projektovom základe. Napríklad všetky dokumenty patriace do jedného projektu sú usporiadané v jednej zbierke. Príklad: zbierku na účely projektu *Transkripcia korešpondencie J. M. Hurbana* môžeme nazvať *Hurban_listy*.

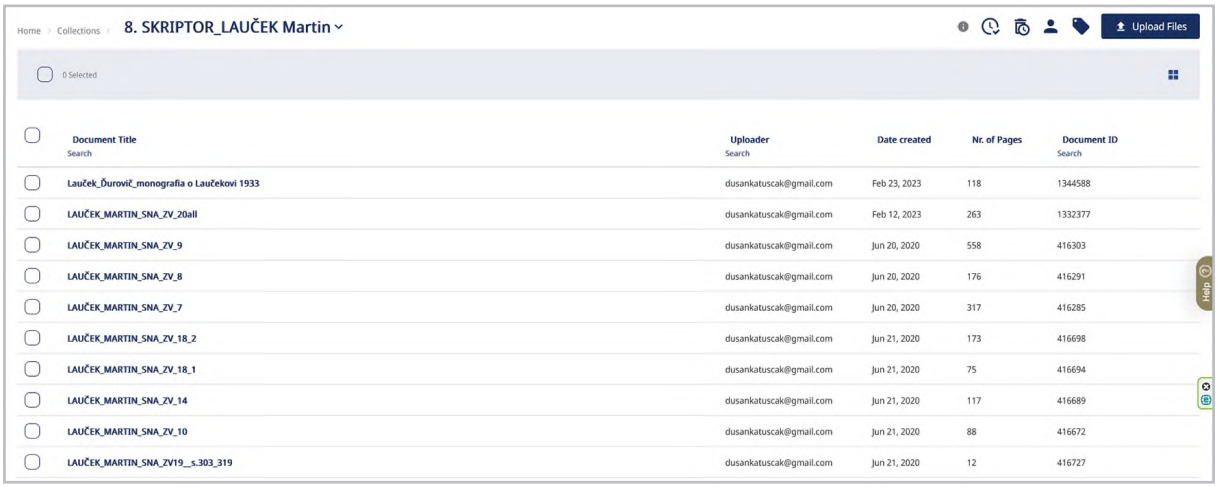
V jednej zbierke môže byť uložených viac *dokumentov.* Dokumenty pozostávajú z jednej alebo viacerých strán. Napríklad v zbierke *Hurban_listy* sú ako dokumenty jednotlivé listy. V projekte Skriptor má každý riešiteľ vytvorený vlastný účet – vlastný projekt a vlastné zbierky.



Obrázok 20 Zbierky v aplikácii Transkribus

Zbierky vytvárajte tak, aby zodpovedali organizácii fondov, zbierok a dokumentov v inštitúcii. Napríklad v archíve SNM v Martine je zbierka rukopisnej korešpondencie Andreja Kmeťa deponovaná v piatich škatuliach.

Dokumenty môžete ako vlastník zbierky (*owner*) usporiadať tak, že uložíte *všetky listy Andreja Kmeťa do zbierky Andrej Kmeť* written letters. V nej má napríklad dokument List Kmeťa adresátke Balkovej a 3 strany.

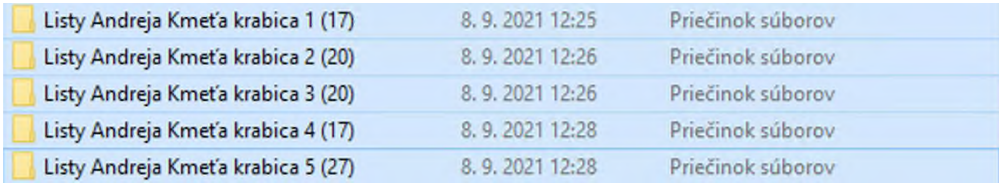


Obrázok 21 Príklad zbierky s dokumentmi a stranami

Ak chcete presne dodržať spôsob, akým sú dokumenty uložené v archíve, môžete pre každú z piatich škatúľ vytvoriť samostatnú zbierku s dokumentmi v tejto škatuli. Takto môžete mať

napríklad vo svojej zložke päť zbierok a v každej zbierke dokumenty, listy podľa uloženia v škatuliach.

Poznámka: To, ako chcete mať usporiadané zbierky, by ste mali mať premyslené už pri snímaní dokumentov (skenovaní, fotografovaní).



Obrázok 22 Usporiadanie zbierok v osobnom počítači podľa škatúľ v archíve

Ak chcete postupovať takto, vytvorte najprv vo svojej zložke päť zbierok a pomenujte ich. Napríklad *Kmeť 1*, *Kmeť 2*, *Kmeť 3*, *Kmeť 4*, *Kmeť 5*.

Do takto vytvorených zbierok nahrajte jednotlivé dokumenty (do *Kmeť 1* dokumenty zo škatule č. 1, atď.)

Ak chcete dodržať spôsob uloženia archívnych dokumentov (päť škatúľ), vytvorte 5 zbierok. V škatuli č. 1 je 17 zložiek (obalov s listami podľa adresátov). Do zbierky *Listy Andreja Kmeťa* škatuľa č. 1 nahrajte všetky dokumenty (listy) podľa adresátov. Vytvárať 17 zbierok pre každú zložku by nemalo praktický zmysel. V prípade korešpondencie Andreja Kmeťa ide o homogénny fond: 5 škatúľ s listami usporiadanými podľa adresátov.

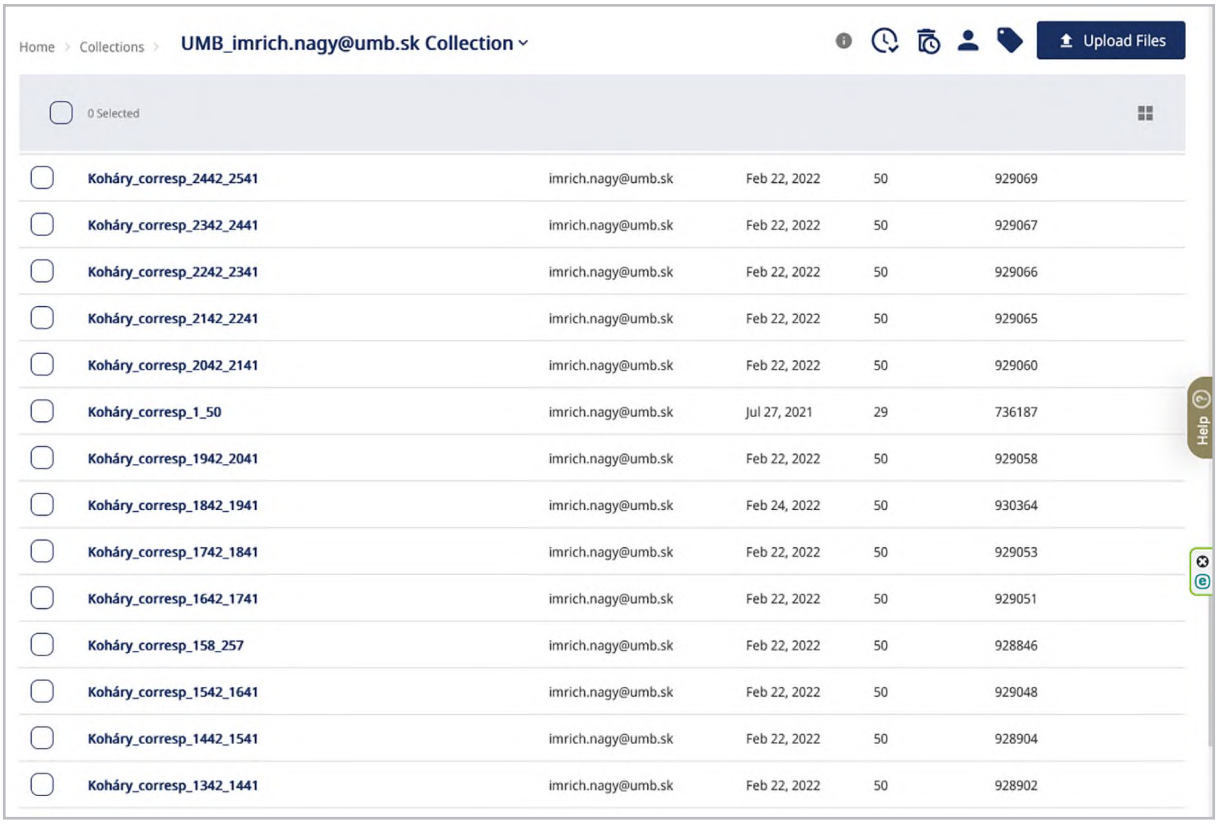
Všetky listy si už v osobnom počítači vopred pripravte ako súbory vo formáte PDF na import na platformu bez ohľadu na uloženie v archíve. Usporiadanie všetkých listov podľa adresátov v abecednom poradí je pre používateľa prijateľnejšie.

Alternatívne je možné do zbierky nahrať samostatne jednotlivé fyzické zväzky. Napríklad v zbierke Martina Laučeka *Collectanea* je základné rozdelenie podľa označovania zväzkov z archívov podľa miesta uloženia: SNK, z archívu SNM a z OsZK a potom podľa označenia v jednotlivých archívoch. Niektoré rozsiahlejšie zväzky sú rozdelené na menšie dokumenty kvôli experimentom v projekte, čo však nie je potrebné. Napríklad zväzok 13 je rozdelený na 5 častí.



Obrázok 23 Usporiadanie zväzkov podľa miesta uloženia

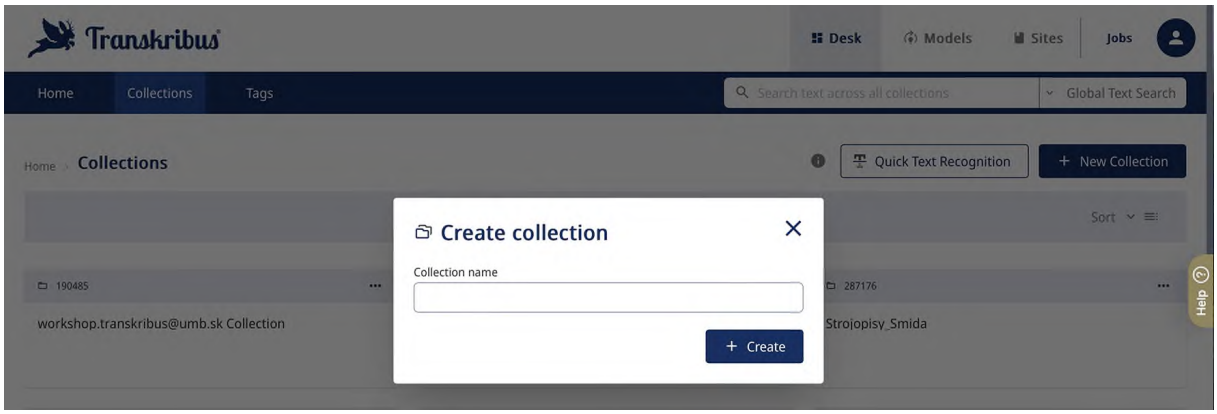
Niekedy je na účely experimentovania vhodné rozdeliť jednotlivé zväzky podľa počtu strán (napríklad 50 s.)



Obrázok 24 Usporiadanie rozdelených zväzkov v zbierke Koháry

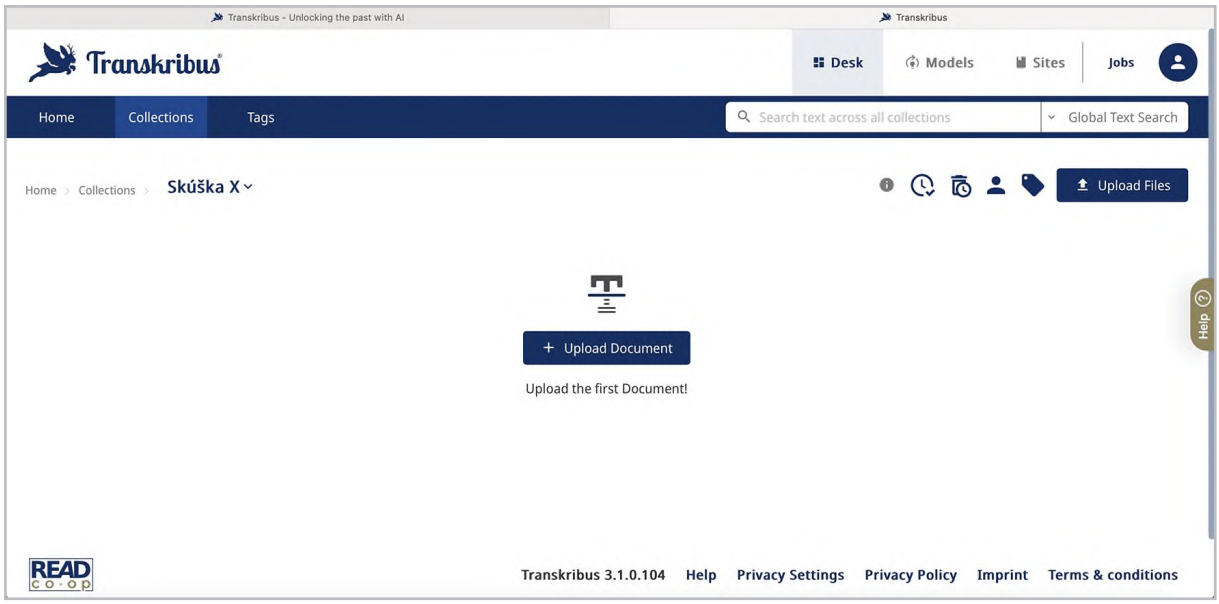
3.1.1 Nová zbierka

Novú zbierku vytvoríte kliknutím na tlačidlo zbierky na tmavomodrej lište a potom na tlačidlo vpravo hore *+New Collection*.



Obrázok 25 Vytvorenie zbierky

Otvorí sa okno Vytvoriť zbierku (*Create collection*). Doň napíšte Názov zbierky (*Collection name*) a kliknite na tlačidlo +Vytvoriť (*+Create*). Do zbierky nahrajte súbory (dokumenty) – digitalizáty (digitálne faksimile) dokumentu/dokumentov, ktoré chcete transkribovať.



Obrázok 26 Vytvorená zbierka pripravená na nahrávanie dokumentov/súborov

3.1.2 Názvy zbierok

Pomenovanie zbierok a dokumentov si pripravte vopred už v procese snímania (skenovania). Digitálne dokumenty označujte spôsobom, ktorý je určený pre archívnu prax. Inštitúcie, ktoré sa rozhodnú pre transkripciu pomocou platformy Transkribus, môžu pomenovať zbierky a dokumenty tak, ako ich majú vo svojich fondoch.

3.1.3 Kontrola kvality pred importom

Kontrola kvality pred nahrávaním dokumentov do aplikácie Transkribus je mimoriadne dôležitá, pretože umožňuje udržiavať poriadok na platforme, organizovať dokumenty a pripravovať ich na editovanie a sprístupnenie na internete cez nástroj *Sites*.

Skontrolujte:

- úplnosť dokumentu,
- kvalitu orezania strán,
- kvalitu nasnímania (ostrosť, kontrast, farebnosť, úplnosť snímanej plochy – strany, presvity),
- orientáciu strán,
- poradie strán,
- formát nasnímania,
- odstráňte duplicitné strany,
- vložte chýbajúce strany.

3.1.4 Zálohovanie. Archivovanie

Po nasnímaní je z dôvodu archivácie, zálohovania a ďalšej manipulácie potrebné vytvoriť:

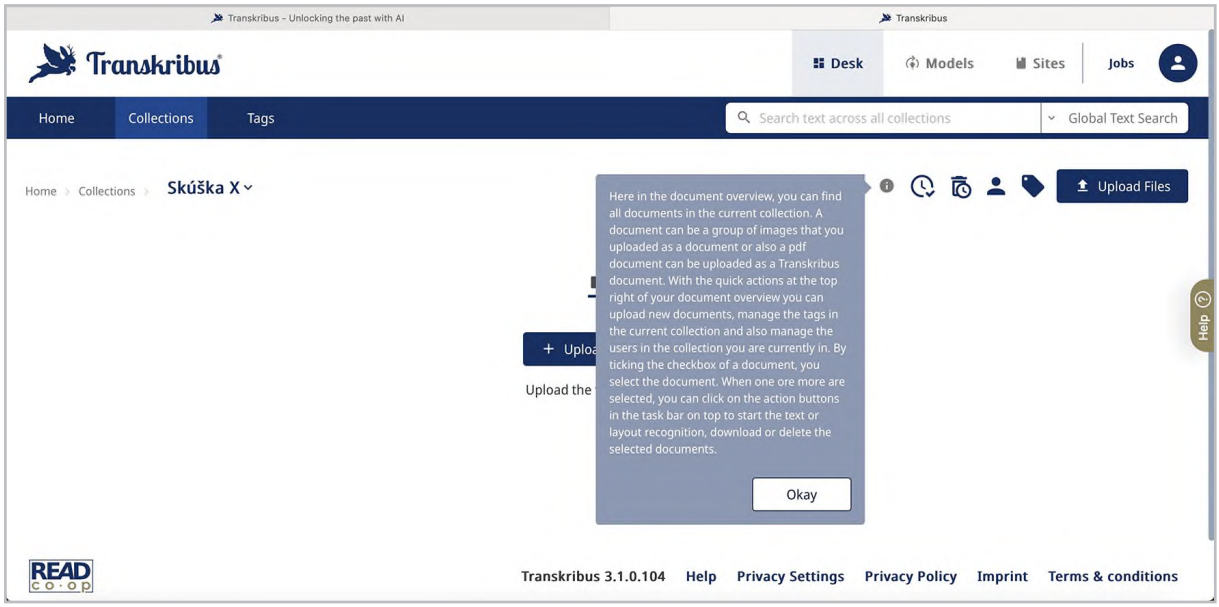
1) *archívnu kópiu* na úložisku alebo na externých nosičoch. Ide o adresár, do ktorého uložíte obrázky v „surovom“, needitovanom formáte, v akom boli nasnímané (fotografované, skenované) JPG/TIFF/RAW/PNG,

2) *derivovanú kópiu*. Adresár, v ktorom budú už „surové“ obrázky po postprocesingu, teda po následnej úprave a kontrole kvality upravené, opravené, orezané, úplné so správnou orientáciou a v dobrej kvalite. Kópiu tohto adresára v najkvalitnejšom dohodnutom formáte nahrajte na nosič (CD, SD, USB, externý disk) alebo na digitálne úložisko a:
   a) poskytnite ho inštitúcii ako vlastníkovi alebo správcovi zbierky,
   b) uložte ho v inštitúcii na účely neverejného prístupu nedostupného cez internet. Určte miesto uloženia a zodpovednú osobu (systémový knihovník, kurátor a pod.),
3) *pracovnú kópiu* v adresári na svojom počítači s dokumentmi v derivovanom formáte PDF, z ktorého nahráte zbierky a dokumenty na platformu Transkribus,
4) *transkribovanú kópiu* v adresári alebo adresároch s exportovanými súbormi, ktoré sú výsledkom transkripcie,
5) *datasety* so súbormi *Ground Truth* a vlastnými modelmi. Exportované datasety z aplikácie Transkribus sa odporúča uložiť na ďalšie použitie podľa interných inštrukcií v digitálnom repozitári inštitúcie alebo napríklad v aplikácii *Zotero*. Datasety vytvorené v aplikácii sú autorskými dielami a citujú sa podobne ako iné dokumenty podľa normy ISO 2709 (2021). Niektoré platformy ako *GitHub* pohodlne umožňujú využiť úložisko čiastočne automaticky prepojené s *európskym repozitárom Zenodo*. *GitHub* sa potom postará o tvorbu verzií (a tvorbu vydaní). *Zenodo zároveň* prideľuje uloženým objektom DOI.

Dokumenty nahrávate (importujete) na platformu Transkribus preto, aby ste ich automaticky transkribovali a následne sprístupnili odbornej a širšej verejnosti.

Pomocou aplikácie Transkribus má zmysel transkribovať väčšie fondy, zbierky a dokumenty, teda stovky až tisícky strán.

3.2 Správa používateľov zbierky



Obrázok 27 Vysvetlivky k nahrávaniu dokumentov do zbierky

Vpravo hore máte k dispozícii tieto ikony: hodiny (*Activity*), odpadový kôš (*Recycle Bin*), silueta používateľa (*User Manager*) a záložka na správu tagov (*Manage Tags*).

Z týchto funkcií má praktický význam najmä funkcia Správa používateľov (*User Manager*) pod ikonou siluety používateľa. Prostredníctvom nej môžeme pridávať používateľov pracujúcich na tej istej zbierke a dokumentoch.

Používateľ/používateľa pridávate kliknutím na symbol siluety a potom na voľbu +Pridať používateľa (*+Add User*). Zadáte jeho e-mailovú adresu, ktorú používa na prihlásenie na platformu a do aplikácie Transkribus. Ak zadáte inú adresu, ktorá sa nenachádza medzi registrovanými používateľmi, používateľa nie je možné pridať.

4 Príprava dokumentu na automatickú transkripciu

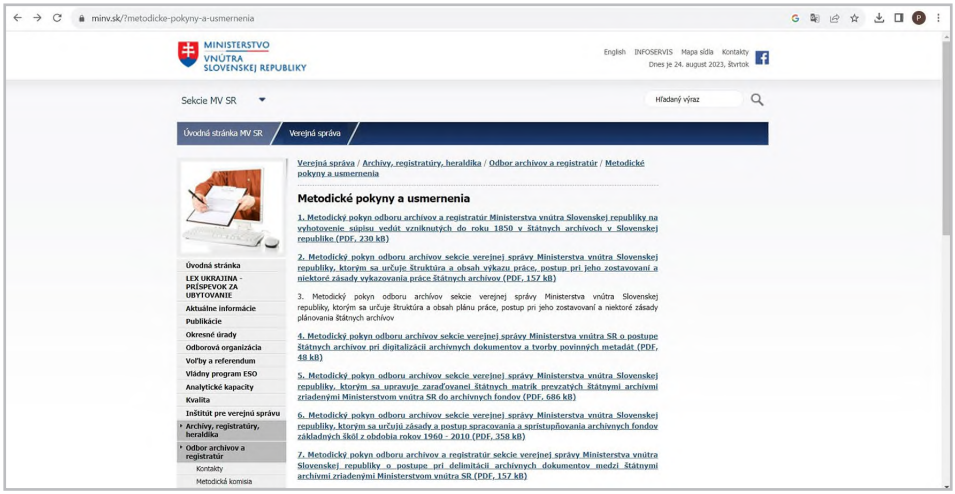
Prípravná fáza na prácu s dokumentom na platforme Transkribus zahŕňa kritériá, ktoré by mali digitalizáty spĺňať, ich správny popis, samotné vyhotovenie kvalitných digitalizátov a následný import do aplikácie Transkribus.

4.1 Kritériá výberu digitalizátov

V podmienkach slovenskej archívnej praxe upravuje výber archívnych dokumentov určených na digitalizáciu Metodický pokyn odboru archívov sekcie verejnej správy Ministerstva vnútra SR o postupe štátnych archívov pri digitalizácii archívnych dokumentov a tvorby povinných metadát č. SVS-OA-2011/23406-001 z roku 2011. Pokyn definuje prednostný výber archívnych fondov a archívnych zbierok ako aj technické parametre systematickej digitalizácie. Odporúča na digitalizáciu dokumenty sprístupnené archívnou pomôckou, chronologicky spred roka 1526, často využívané bádateľmi a fyzicky najohrozenejšie. Z hľadiska technických parametrov stanovuje vyhotovenie digitálnych kópií z originálnych dokumentov alebo mikrofilmov. Každý záber digitálnej kópie archívneho dokumentu má byť uložený ako samostatný súbor s popisným reťazcom pozostávajúcim zo šiestich, resp. siedmich častí v tvare:

SK_aaaa_fffff_iiiii_ssss_x.ext

Štruktúru reťazca tvorí kód krajiny (SK), štvormiestne číslo archívu (z číselníka štátnych archívov SR, v príklade aaaa), päťmiestne číslo archívneho súboru (z aplikačného programového vybavenia AFondy, v príklade fffff), päťmiestne označenie inventárneho čísla alebo signatúry archívneho dokumentu (v príklade iiiii), štvormiestne poradové číslo snímky v rámci inventárneho čísla alebo signatúry (v príklade ssss), znak x označuje písmeno alebo číslicu rozlišujúce digitálnu kópiu (konzervačnú, pre interné potreby alebo na študijné účely), prípona .ext označuje grafický formát (JPEG alebo TIFF).



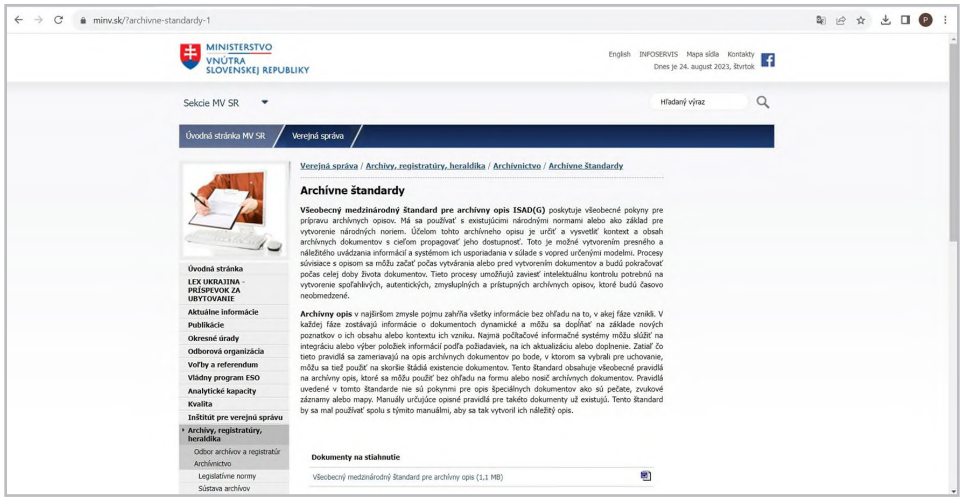
Obrázok 28 Metodický pokyn MV SR č. SVS-OA-2011/23406-001

Voľba dokumentu pre aplikáciu Transkribus sa oproti tomu vyznačuje niektorými špecifikami. Na rozdiel od bežnej digitalizácie v pamäťových inštitúciách nie sú prioritou poškodené a ďalšou manipuláciou ohrozené archívne dokumenty. Vzhľadom na ich stav zachovania (porušenosť, fragmentárnosť) nie sú príliš vhodné na segmentáciu, transkripciu a nadväzujúce postupy. Naopak vhodnejšie sú intaktne zachované archívne dokumenty. Na prácu v Transkribuse sú najefektívnejšie rozsiahle rukopisy vyhotovené jednou pisárskou rukou najlepšie v krátkom

časovom úseku. Z hľadiska potrieb bádateľov a vedeckého výskumu sem možno zaradiť napríklad matričnú agendu, kanonické vizitácie, sčítacie operáty, parcelné protokoly a pod. S prihliadnutím na charakter platformy je vhodné vyberať dokumenty odrážajúce špecifiká slovenského kultúrneho okruhu, ktoré sú atraktívne aj pre zahraničných užívateľov.

4.2 Popis fondov, zbierok a dokumentov

Význam presného popisu je osobitne dôležitý pre digitálne objekty nadobúdajúce podobu elektronického informačného zdroja. Na rozdiel od fyzického vyhotovenia vznikajú digitálne dokumenty iba vďaka softvéru. Stráca sa tým jedinečnosť a proveniencia fyzicky zachovaných archívnych dokumentov. Elektronické dokumenty preto nadobúdajú zvýšené požiadavky na overovanie faktov (*fact-checking*) s dôrazom na dôveryhodnosť, spoľahlivosť, ale aj pôvodnú provenienciu a hierarchiu. Popis a citovanie má umožňovať dohľadanie fyzicky zachovaných zdrojov. Vzhľadom na medzinárodný obsah a rozšírenie platformy Transkribus je pre digitalizované dokumenty (materiál textovej povahy) vhodné využívať štandardizované medzinárodné popisy a normy.



Obrázok 29 Všeobecný medzinárodný štandard pre popis archívnej jednotky

Medzinárodná rada archívov schválila a zverejnila niekoľko štandardov zjednocujúcich popis archívnych dokumentov vrátane digitalizátov. Všeobecný medzinárodný štandard pre popis archívnej jednotky (*General International Standard for Archival Description – ISAD(G)*) vychádza z provenienčného princípu a definuje dvadsaťšesť položiek popisu. Odbor archívov a registratúr Ministerstva vnútra SR sprístupnil slovenský preklad druhého vydania štandardu z roku 1999 aj s príkladmi viacúrovňových popisov pre sieť štátnych archívov na Slovensku. Keďže archívne dokumenty uchovávajú aj iné subjekty, napr. kultúrne inštitúcie a súkromní vlastníci, vznikli ďalšie normy. Medzinárodný štandard pre archívne autoritné záznamy právnických osôb, fyzických osôb a rodín (*International Standard of Archival Authority Record for Corporate Bodies, Persons and Families – ISAAR(CPF)*) je rozšírený najmä v mimoeurópskom priestore. Ďalšie úpravy obsahuje Všeobecný medzinárodný štandard pre popis inštitúcií s archívnymi dokumentmi (*International Standard for Describing Institutions with Archival Holdings – ISDIAH*).

Možnosti pre jednotný popis digitalizovaných dokumentov zo štátnych archívov, ale aj cirkevných archívov, knižníc, múzeí, galérií, pamiatkových úradov, vedeckých ústavov a pod. poskytuje štruktúra popisného reťazca pre platformu Transkribus – projekt *Skriptor*. Na rozdiel

od archívnej terminológie metodika vychádza z určenia pre digitálny repozitár. Obsahuje fixné názvy zbierok a súborov určených na automatickú transkripciu s dôrazom na prehľadnosť a zrozumiteľnosť. Štruktúru reťazca tvorí názov zbierky, názov podzbierky a zdroj/vlastník. Za fixnými časťami nasledujú premenlivé hodnoty dopĺňané podľa konkrétnej situácie a podľa skenovaných objektov.

Tieto hodnoty sú najmä:

1) označenie (číslo) zväzku (signatúra),
2) počet listov,
3) rok(y) RRRR alebo RRRR-RRRR.

Celý názov entity určený na nahratie do Transkribusu môže mať napríklad takúto štruktúru:

LAUČEK_MARTIN_SNA_ZV_13_5

Skriptor_Hurban listy_SNKLA_2A3_1875_Pauliny-Tóth Viliam

Visitatio canonica_CV18_DABB

Ak je snímaný objekt určený pre digitálny repozitár, vloží sa na začiatok reťazca referenčný kód a názov jednotky popisu – v prípade štátnych archívov podľa ISAD(G) kód krajiny, archívu, fondu alebo zbierky.

4.3 ScanTent a DocScan pre archívy a knižnice

DocScan a *ScanTent* sú nástroje, ktoré pomáhajú snímať historické dokumenty na účely transkripcie v dobrej kvalite. Informácie o nástrojoch sú dostupné z hlavnej stránky združenia READ-COOP na https://readcoop.eu/transkribus/?sc=Transkribus.

V bádateľniach archívov bádatelia používajú na snímanie vlastné zariadenia, fotoaparáty, mobilné telefóny, tablety a podobne. ScanTent a DocScan sú prijateľnou alternatívou k bežným zariadeniam na snímanie dokumentov v archívoch a knižniciach a výborným riešením pre inštitúcie, ktoré nemajú pre používateľov k dispozícii kvalitnejšie stolové skenery alebo ktoré svoje dokumenty ešte nemajú zdigitalizované a prístupné pre používateľov.

Obrázky snímané týmto spôsobom je možné poskytnúť inštitúcii na dohodnutom nosiči alebo na uloženie do inštitucionálneho digitálneho repozitára, archívu alebo knižnice. Ak však máte možnosť rozhodnúť sa medzi zariadeniami ScanTent a DocScan a profesionálnym skenerom dokumentov, uprednostnite profesionálny skener.

Pre digitalizáciu platí zásada, že snímanie – skenovanie sa robí v najvyššej možnej kvalite, na najvyššej dosiahnuteľnej úrovni. Kvalita snímaných obrázkov je kľúčová pre efektívnu transkripciu. Skúsenosti ukazujú, že kvalita snímania by mala byť okolo 600 DPI. Historické rukopisy predstavujú *de facto* špecifickú grafiku, pre ktorú sa niekedy odporúča snímanie v kvalite 900 až 1200 DPI. Práca s vysokokvalitnými obrázkami však môže vyžadovať postprocesing, čiže následné spracovanie pomocou špeciálnych softvérov na úpravu obrazu.

Pri práci s nástrojmi DocScan a ScanTent trvá naskenovanie knihy, teda fyzického zväzku s 300 stranami, približne 12 –15 minút. To je 150 obrázkov, pretože v zariadení sa snímajú naraz obidve strany otvoreného zväzku prakticky až do veľkosti A3. Spravidla teda môžete nasnímať **viac ako 500 obrázkov za hodinu.**

4.3.1 ScanTent

ScanTent je možné zakúpiť priamo z hlavnej stránky platformy Transkribus kliknutím na voľbu ScanTent.

Je optimálnym riešením na snímanie voľných alebo zviazaných dokumentov v bádateľniach – pre nízkonákladové a vysokokvalitné snímanie (skenovanie). Cena ScanTentu je aktuálne 239,00 € vrátane 20% DPH plus poštovné.

Niektoré inštitúcie majú pre bádateľov a čitateľov v študovniach a bádateľniach desiatky zariadení ScanTent. Napríklad Francúzska národná knižnica ich mala v roku 2023 celkovo 40. Na získanie ďalších informácií kontaktujte scantent@caa.tuwien.ac.at.

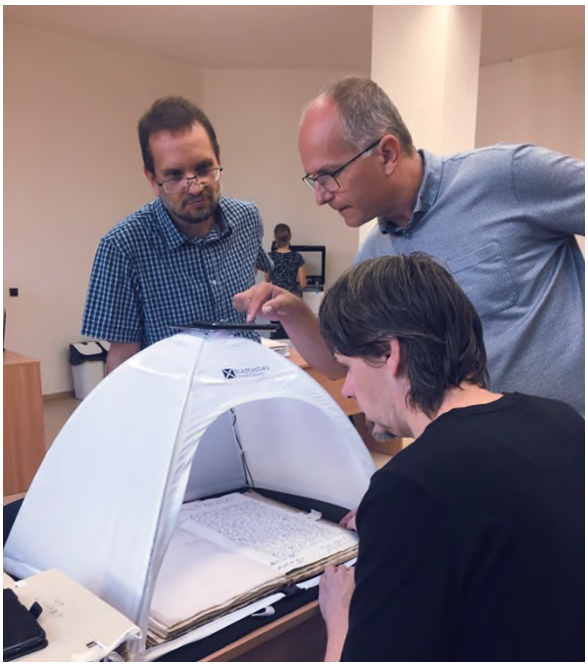
Na účely obstarania a evidencie majetku je to v katalógu tovarov a služieb tovar v rámci skupiny „statívy na fotoaparáty“ vedený pod číslom 90 391.

Charakteristika zariadenia ScanTent:

- **profesionálne fotografické prostredie** na snímanie vysokokvalitných obrázkov bez dodatočného svetla. Stan je z nylonovej hodvábnej látky s vnútornými príchytkami na led osvetlenie,
- **LED osvetlenie s USB napájaním** na nepriame osvetlenie dokumentov – pripojenie na notebook alebo iný zdroj (napríklad powerbank),
- **tmavá plstená látka** na základni ako optimálny podklad,
- **veľká základná plocha**, takže používatelia môžu vložiť ruky do zariadenia a držať zviazané dokumenty otvorené oboma rukami,
- skenovanie dokumentov **veľkosti približne A3 alebo aj o niečo väčších**,
- ľahký (500 gramov) a **skladateľný**, zmestí sa do malého puzdra.



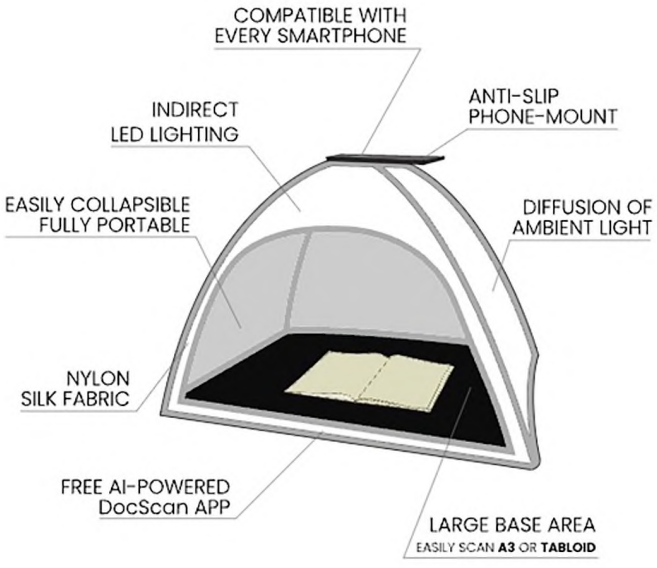
Obrázok 30 Prototyp ScanTentu použitý na snímanie zväzkov Martina Laučeka v SNA v Bratislave a v SNM v Martine (2018)



Obrázok 31 Novší model ScanTentu použitý na snímanie v Diecéznom archíve Banskobystrického biskupstva v Badíne v rámci projektu Skriptor (10.09.2020)

Popis častí ScanTentu:

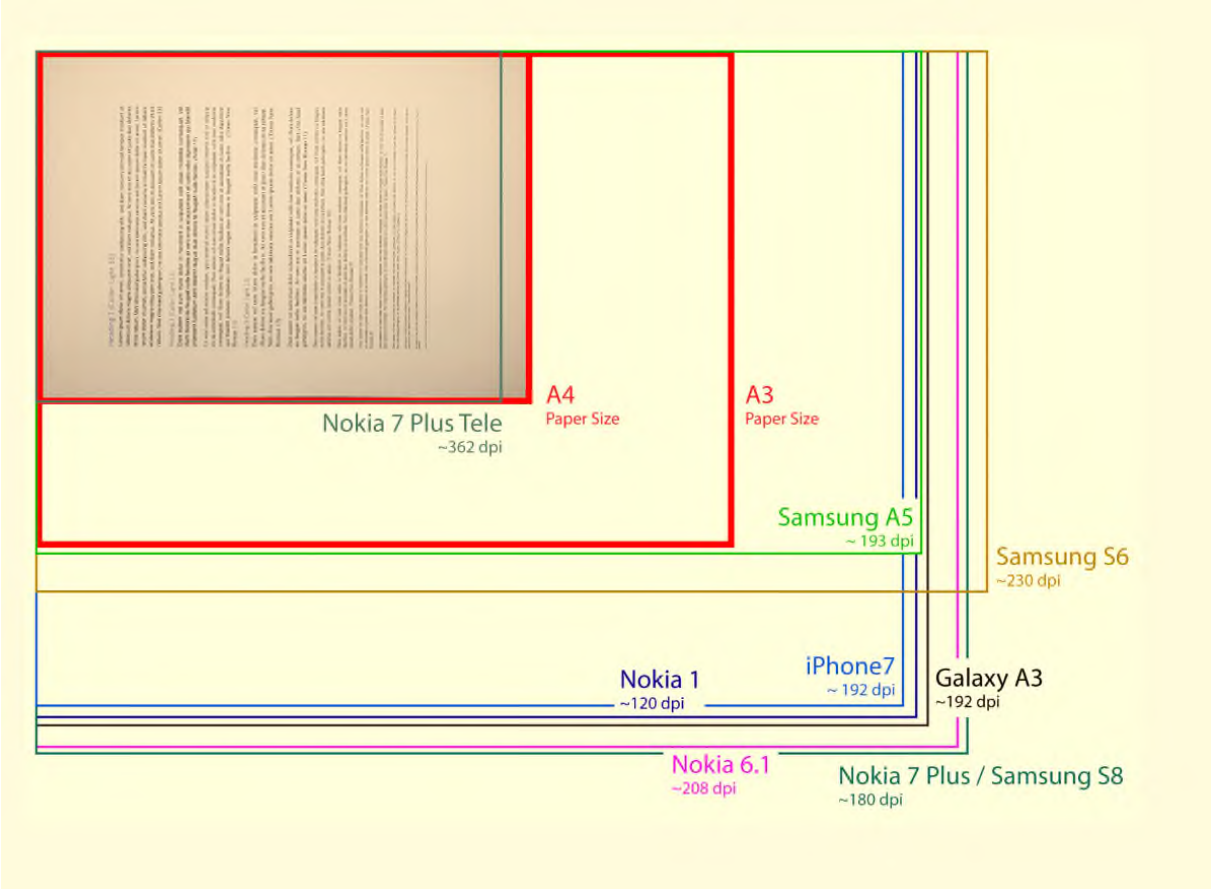
1. kompatibilita s každým smartfónom (*Compatible with every smartphone*)
2. nepriame LED osvetlenie (*Indirect LED Lighting*)
3. protišmyková plocha pod smartfón (*Anti-slip phone-mount*)
4. difúzne osvetlenie (*Diffusion of ambient light*)
5. nylonová hodvábna látka (*Nylon silk fabric*)
6. voľne dostupná aplikácia DocScan s umelou inteligenciou (*Free AI-Powered DocScan App*)
7. veľká základná plocha (A3 alebo časopisecký formát tabloid 280 mm × 430 mm) (*Large base area*)



Obrázok 32 Komponenty ScanTentu

Rýchly prehľad montáže ScanTentu nájdete vo videu https://youtu.be/iL2WNNi5VEI

Po nastavení zariadenia ScanTent môžete začať so snímaním – fotografovaním. Ak okolité svetlo nie je dostatočné (čo sa stáva veľmi zriedka), zapnite svetlá do prenosného počítača alebo USB zásuvky.

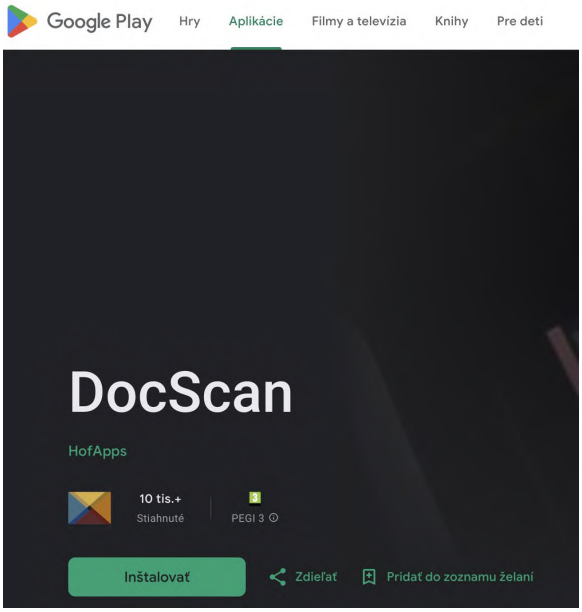


Obrázok 33 Experimenty vývojového tímu s rôznymi smartfónmi

Vývojový tím platformy Transkribus testoval osem rôznych smartfónov a meral ich rozlíšenie. Zorné pole a rozlíšenie jednotlivých smartfónov môžete vidieť na obrázku č. 33. DocScan nepodporuje telefóny Nokia 7 Plus Tele a iPhone 7.

4.3.2 Aplikácia DocScan

Aplikácia DocScan sa používa so zariadením ScanTent. Je to softvér, ktorý vyvinula Technická univerzita vo Viedni v rámci európskeho projektu READ. Aplikáciu DocScan si môžete bezplatne stiahnuť z obchodu Google Play na adrese
https://play.google.com/store/apps/details?id=at.ac.tuwien.caa.docscan



Obrázok 34 Stránka Google Play na stiahnutie a inštaláciu aplikácie DocScan

Aplikácia DocScan bola vyvinutá špeciálne na digitalizáciu kníh a archívnych dokumentov pomocou smartfónu. V súčasnosti je k dispozícii prednostne pre telefóny so systémom Android.

DocScan je určený na skenovanie historických dokumentov v kombinácii so ScanTentom. Zobrazuje strany v živom náhľade a robí skeny v kvalite dostatočnej pre platformu Transkribus. V automatickom režime *Series* sníma obrázok po otočení stránky po pripojení k zariadeniu ScanTent. Umožňuje teda rýchlo skenovať knihy alebo dokumenty bez interakcie s vaším mobilným telefónom.

Charakteristika aplikácie DocScan:

- rýchla a spoľahlivá detekcia stránok dokumentu,
- jednoduchý režim (*Single*) na manuálne snímanie jednotlivých obrázkov,
- sériový režim (*Series*) na automatické snímanie obrázkov (pohyb je detekovaný automaticky). Po otočení automaticky sníma ďalší obraz dvojstrany,
- schopnosť otáčať a orezávať stránky,
- priame nahrávanie dokumentov na server Transkribus.

Výhody:

- vysoká kvalita obrazu – moderné inteligentné telefóny poskytujú vynikajúcu kvalitu obrazu s vysokým rozlíšením,
- nákladovo efektívne – pre koncového používateľa aj pre knižnicu/archív,
- žiadne licenčné poplatky,
- nie je potrebná žiadna používateľská podpora z archívu alebo knižnice – používatelia sa s aplikáciou DocScan rýchlo zoznámia sami,
- priateľské k autorským právam – používatelia snímajú a ukladajú obrázky na svojom vlastnom zariadení, nie na tých, ktoré vlastní knižnica alebo archív,
- DocScan ponúka možnosť „masového skenovania“, kde je možné obrázky vytvorené používateľmi pridať do digitálnych fondov knižnice alebo archívu.

4.3.3 Bezpečnosť údajov v aplikácii DocScan

Bezpečnosť sa začína pochopením toho, ako vývojový tím zhromažďuje a zdieľa vaše údaje. Postupy ochrany osobných údajov a zabezpečenia sa môžu líšiť v závislosti od vášho používania, regiónu a veku. Vývojový tím aktuálne poskytuje nasledujúce informácie a môže ich časom aktualizovať.

- Táto aplikácia môže zdieľať tieto typy údajov s tretími stranami: miesto a osobné údaje.
- Táto aplikácia môže zhromažďovať tieto typy údajov: osobné informácie, fotografie a videá, súbory a dokumenty.
- Dáta sú pri prenose šifrované.
- Údaje nie je možné vymazať.

4.3.4 Snímanie pomocou zariadenia ScanTent a aplikácie DocScan

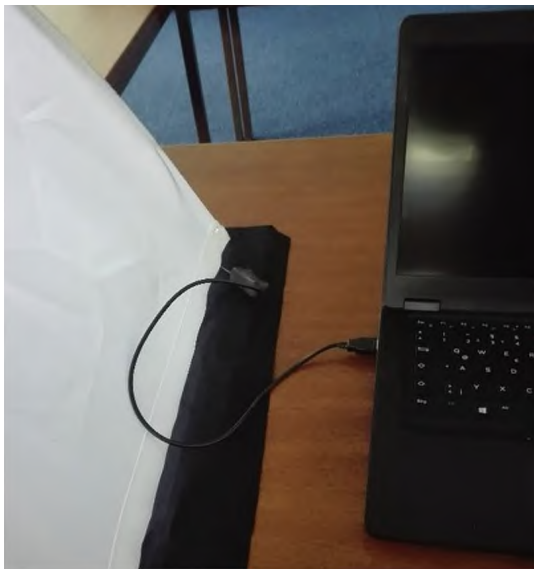
Položte smartfón na podložku v hornej časti ScanTent tak, aby šošovka fotoaparátu smerovala nadol. Šošovka by mala byť zarovnaná s otvorom v hornej časti zariadenia.

Polohu smartfónu je vhodné nastaviť paralelne vzhľadom na snímanú plochu a orientáciu strany. Poloha smartfónu by mala zostať počas snímania dokumentu v stabilnej a rovnakej polohe vo vzťahu k snímanému dokumentu, aby dodatočne nebolo potrebné korigovať orientáciu strán alebo opakovane snímať nesprávne snímanú plochu strany. Smer snímania ukazuje obrázok písmena „T“.

Dôležité: ScanTent umiestnite vyššie alebo nižšie podľa toho, či chcete pri snímaní sedieť alebo stáť.

Displej smartfónu musí byť rovnobežný so smerom dokumentu. Ak stojíte pred ScanTentom, musíte vidieť na displej a vedieť čítať správy DocScanu na smartfóne. Mobilný telefón by mal byť orientovaný rovnako ako strana.

Pre pohodlie snímania je možné obrazovku aplikácie DocScan zrkadliť na ďalšom počítači, takže ju môžete vidieť a ovládať cez počítač nielen cez smartfón položený na ScanTente.



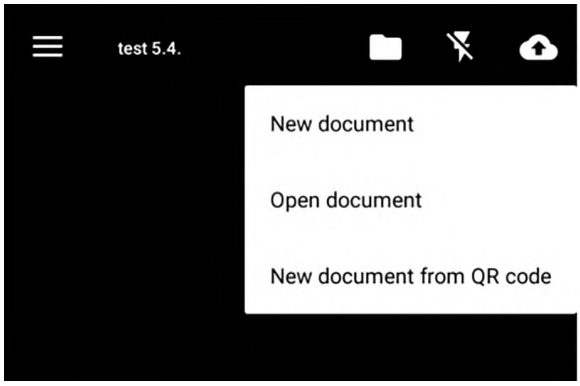
Obrázok 35 Pripojenie osvetlenia LED k notebooku



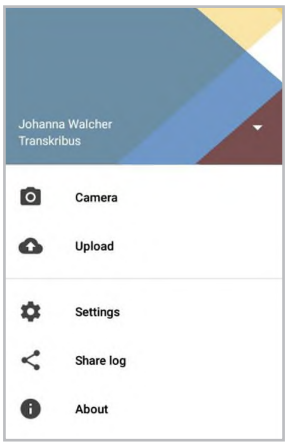
Obrázok 36 ScanTent pripravený na snímanie smartfónom

4.3.5 Práca s aplikáciou DocScan

Otvorte aplikáciu kliknutím na ikonu DocScan v telefóne.



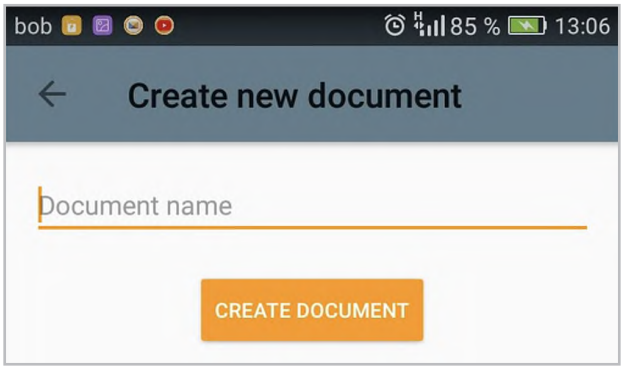
Obrázok 37 Spojenie DocScanu s aplikáciou Transkribus za účelom prenosu údajov z DocScanu



Obrázok 38 Plocha aplikácie DocScan prihláseného používateľa

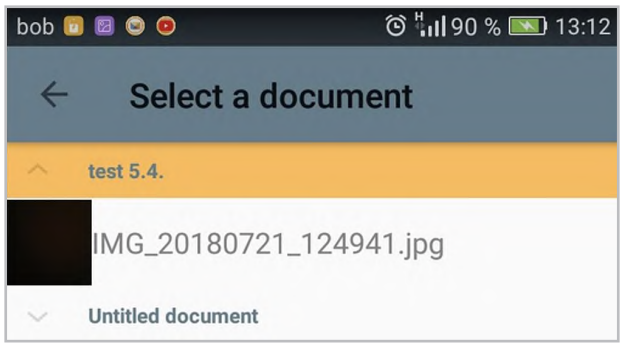
Kliknite na záložku Dokumenty (*Documents*). Priraďte svojmu dokumentu názov.

Vyberte možnosť Vytvoriť dokument (*Create document*). Všetky obrázky, ktoré následne nasnímate, budú uložené pod týmto menom vo vašom telefóne a zostanú v ňom, aj keď ich nahráte do Transkribusu.



Obrázok 39 Vytvoriť a popísať nový snímaný dokument

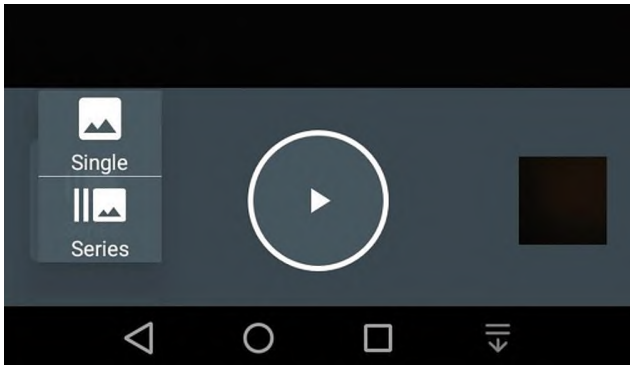
K existujúcim dokumentom môžete pridať nový dokument kliknutím na ikonu „+“. Kliknite na ikonu priečinka v pravej hornej časti aplikácie. Potom vyberte možnosť Otvoriť dokument. Vyberte názov existujúceho dokumentu a zvoľte možnosť Použiť vybratý dokument.



Obrázok 40 Pridanie nového dokumentu k existujúcim dokumentom

Po popísaní dokumentu môžete začať skenovať. Umiestnite telefón na vrchnú časť zariadenia ScanTent.

Na hlavnej stránke kliknite na záložku Kamera (*Camera*). Môžete si vybrať, či chcete nasnímať jednotlivé obrázky manuálne alebo nastaviť aplikáciu tak, aby automaticky zachytávala obrázok pri každom otočení stránky. Môžete si vybrať z možností Manuálne/Jednotlivo (*Manual/Single*) alebo Automaticky/Hromadne (*Automatic/Series*) v ľavej dolnej časti aplikácie.



Obrázok 41 Režimy snímania: Manual/Single režim, Automatic/Series režim

Snímanie spustite kliknutím na ikonu fotoaparátu v krúžku.

Na telefóne zapnite zvuk. Ten upozorní na otočenie strany. Otočenie a zosnímanie indikuje aj svetelný signál, ak ho máte zapnutý.

Strany otáčajte opatrne, neponáhľajte sa, aby DocScan stačil zaostriť, a aby správne snímal celú plochu. Unáhlené pohyby môžu spôsobiť nedostatočné zaostrenie a rozmazanie snímaného obrazu.

Po snímaní dokumentu je potrebná kontrola kvality snímania alebo postprocesing, čiže následné spracovanie obrazov v dokumente. Zamerajte sa na úplnosť, možné duplicity, orientáciu strán a pod.

Proces snímania je možné vrátiť cez ikonu troch vodorovných čiarok. Ku kamere sa dostanete cez tú istú ikonu.

4.3.5.1 Odoslanie dokumentu na platformu Transkribus

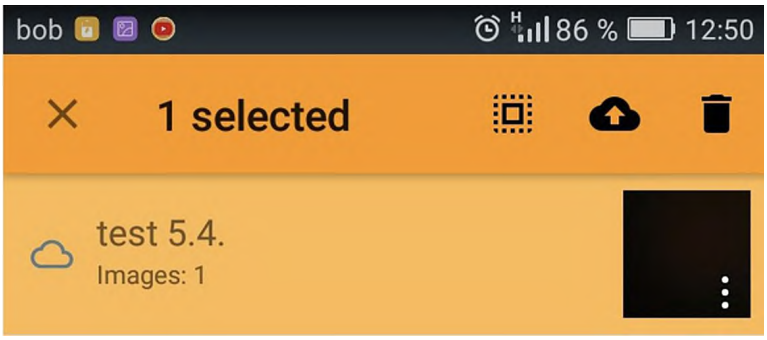


Obrázok 42 Nahrávanie do Transkribusu cez ikonu Cloud

Stlačte ikonu *Cloud* v pravej hornej časti aplikácie. V prípade potreby sa prihláste do svojho účtu v Transkribuse.

Vyberte dokument, ktorý chcete nahrať na platformu. Znova kliknite na ikonu *Cloud*.

Otvorte Transkribus vo svojom počítači. Svoje nahraté dokumenty nájdete v zbierke s názvom *DocScan – Uploads*.



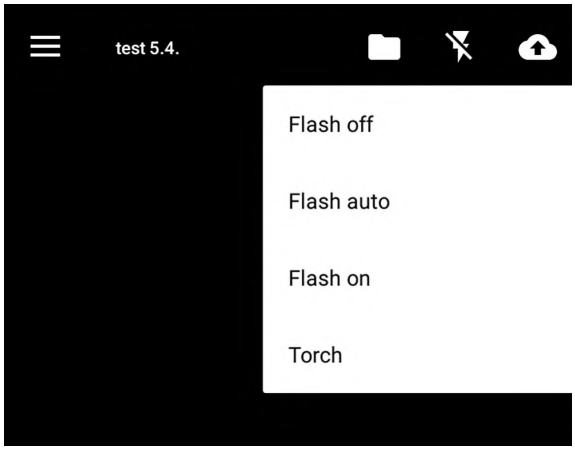
Obrázok 43 Výber súboru na nahratie do Transkribusu

Nahrávanie na platformu Transkribus je zvyčajne hotové o niekoľko minút. Ak prenášate veľké množstvo snímok, môže to trvať o niečo dlhšie.

4.3.5.2 Nastavenia

Ďalšie nastavenia nájdete a upravíte kliknutím na ikonu troch vodorovných čiarok vľavo hore a výberom voľby Nastavenia (*Settings*).

Blesk nastavíte stlačením ikony blesku v pravom hornom rohu aplikácie. Na výber sú štyri možnosti: vypnutý blesk (*Flash off*), automatický blesk (*Flash auto*), zapnutý blesk (*Flash on*) a svetlo (*Torch*).

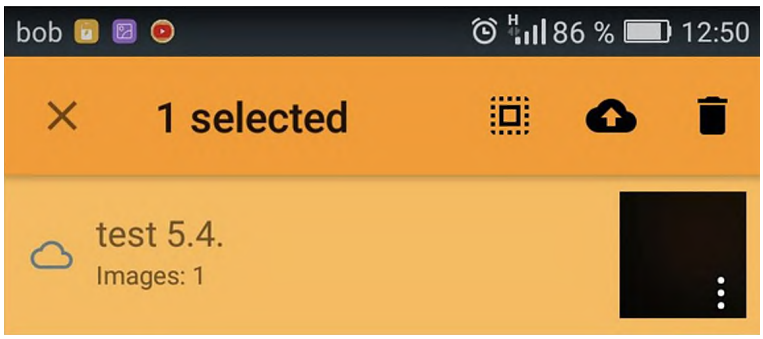


Obrázok 44 Nastavenie blesku

4.3.5.3 Automatické orezávanie, otáčanie a mazanie

Na orezanie a otočenie obrázkov podľa potreby môžete použiť aplikáciu DocScan.

1. Po nasnímaní obrázka stlačením miniatúry v pravom dolnom rohu aplikácie otvorte nastavenia úprav.

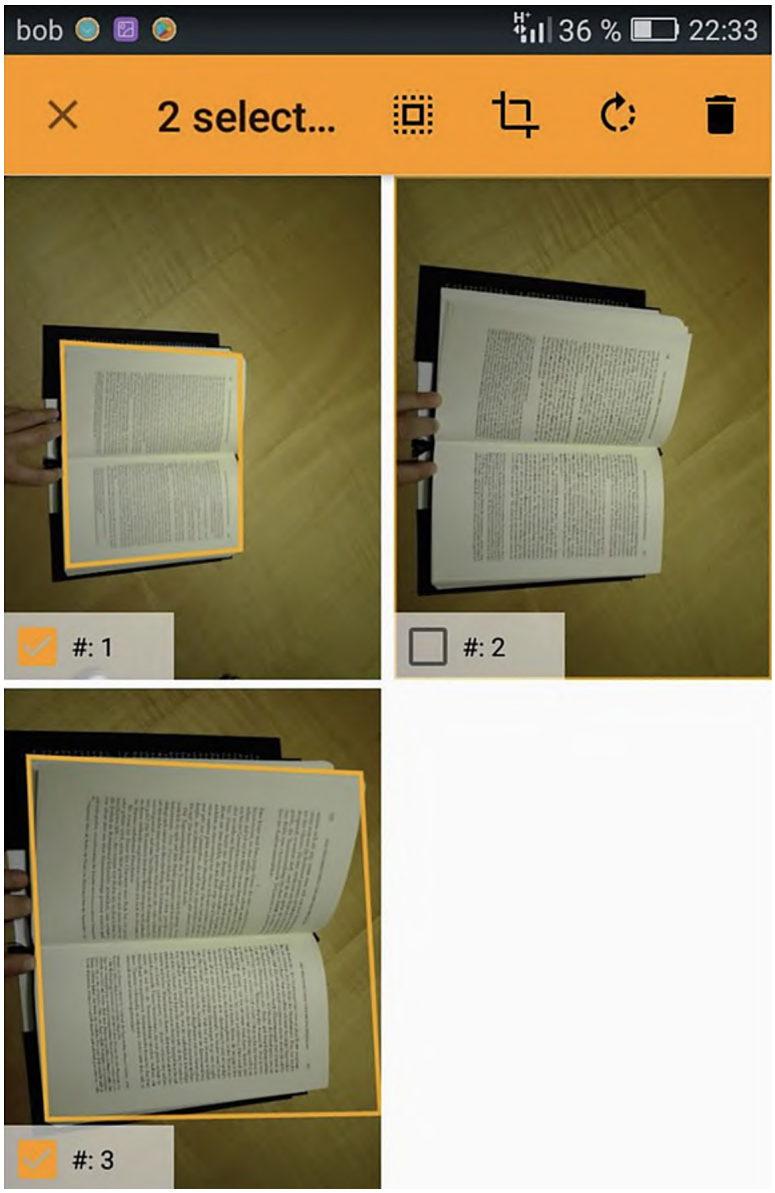


Obrázok 45 Výber strán dokumentu na orezanie cez miniatúru

2. Všetky strany sa zobrazia v žltých rámoch.

Poznámka: Keď je aktivované orezanie, do žltého rámu sa pridá niekoľko pixelov, takže na obrázku sa zobrazí celá strana.

3. Označte súbory, ktoré chcete orezať.



Obrázok 46 Výber strán na orezanie

Vďaka funkcii automatického orezania nemusíte presúvať rámy do správnej polohy, aplikácia to za vás urobí automaticky.

4.3.5.4 Manuálne orezanie

1. Kliknite na ikonu orezania v spodnej časti obrazovky.
2. Potiahnite rohy obrázka do požadovanej polohy.
3. Kliknite na ikonu orezania v pravom hornom rohu obrazovky a uložte orezaný obrázok.
4. Na ďalšej obrazovke kliknite na ikonu Uložiť (*Save*).

Obrázky môžete otáčať, zdieľať alebo odstrániť (zahodiť do koša) tak, že vyberiete potrebné strany a kliknete na príslušnú ikonu.



Obrázok 47 Funkcie Otočiť, Zdieľať, Orezať a Zahodiť do koša

4.4 Snímanie – zhrnutie

4.4.1 Snímanie

- snímanie je jeden z procesov digitalizácie,
- vykonáva sa pomocou technického zariadenia vhodného na digitalizáciu, konkrétne zariadenia na zachytenie digitálneho obrazu ako:
  - digitálne fotoaparáty,
  - kamery,
  - knižné skenery,
  - iné skenery.
- v súvislosti s digitalizáciou hovoríme o výsledkoch snímania – obrazoch:
  - digitalizáty
  - digitálne faksimile.

4.4.2 Snímanie v praxi

Snímanie vo verejných informačných inštitúciách by sa malo realizovať v súlade s právnymi predpismi. V archíve je možné so súhlasom vedenia archívu snímanie archívnych dokumentov klasickou kamerou a digitálnou kamerou a fotografickou technikou pri dodržaní týchto zásad:

- na účely automatickej transkripcie, pokiaľ je to možné, použiť dokumenty nasnímané profesionálnymi skenermi a obrazmi v najvyššej dosiahnuteľnej kvalite,
- minimálna kvalita skenovania by mala byť 300 DPI,
- nakoľko pri historických rukopisoch ide de facto o grafiku, je vhodné skenovať vo vyššej kvalite,
- pre platformu Transkribus je možné snímať dokumenty do formátu veľkosti A3 pomocou zariadenia ScanTent a softvérom DocScan.

4.4.3 Formáty obrázkov: formát JPG, JPEG

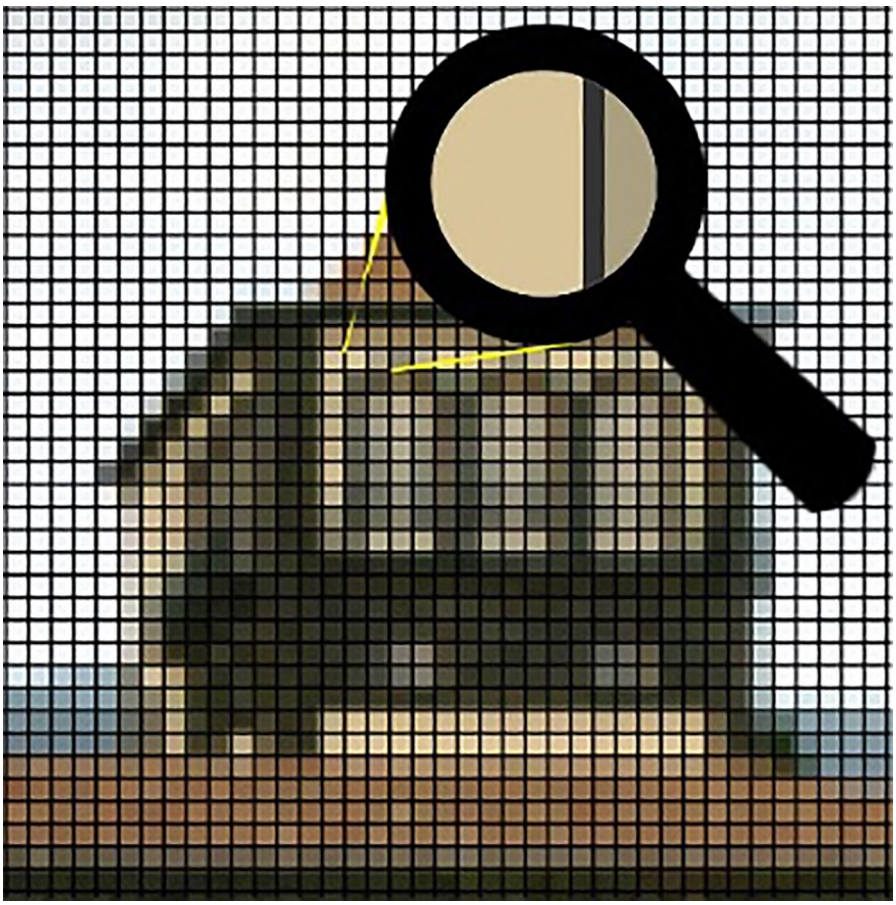
Najrozšírenejšie sú formáty s príponou .jpg, .jpeg alebo .JPG, .JPEG – medzi nimi nie je žiadny rozdiel. V týchto formátoch ukladajú súbory všetky fotoaparáty aj mobilné zariadenia, ak používame napríklad aplikáciu DocScan. Formát JPEG (JPG) je de facto široko používaným štandardom na ukladanie digitálnych snímok.

4.4.4 Pixel – základ pre ukladanie digitálneho obrazu

Čo je pixel:

- najmenšia jednotka obrazových informácií,
- skratka pre prvok obrázka,
- skratka „pix“ znamená jednu plnofarebnú bodku obrázka,

- nemá predpísaný tvar – môže byť štvorcový, kruhový alebo ľubovoľný – možno si ho predstaviť ako obdĺžnik vytvorený rozrezaním obrazu na určitý počet vertikálnych a horizontálnych segmentov.

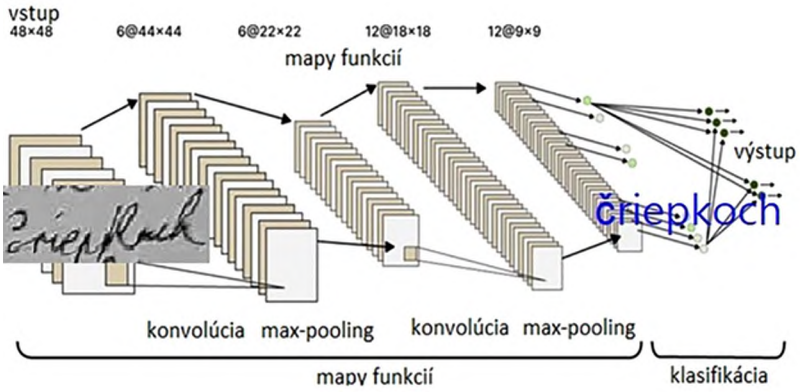


Obrázok 48 Pixely obrázka.
Zdroj: https://www.digimanie.cz/formaty-pro-ukladani-fotografii-1dil-zaklady/1962

Aby počítač mohol pracovať s obrázkom, musia sa pixely previesť do bitov (nuly a jednotky). Pre *čiernobiele obrázky* platí, že každý pixel sa zvyčajne skladá z 8 bitov (1 bajt). Pre farebné obrázky platí, že ak sa používa farebná schéma RGB (červená, zelená, modrá), tak pre každú farbu sa použije jeden bajt, čiže (3 x 8), t. j. 24 bitov (3 bajty) na pixel – v tomto prípade hovoríme o farebnej hĺbke 8 bitov na kanál alebo 3 x 8 = 24 bitov na pixel (24 bpp = bit na pixel). V odbornej praxi to však často nestačí, preto sa používa vyššia farebná hĺbka, t. j. 16 bitov (2 bajty) na kanál, t. j. 3 x 16 = 48 bitov na pixel (bpp).

4.4.5 Príklad štruktúry konvolučnej siete

Obrázok ilustruje proces fungovania konvolučnej siete.



Obrázok 49 Pixely v transkripcii - konvolúcia

Historické staré a vzácne tlače, strojopisy a hlavne rukopisy spravidla nie je možné uspokojivo transkribovať – vtedy prichádza na pomoc umelá inteligencia. V snahách sprístupniť historické písomné dedičstvo sa pozornosť výskumníkov koncentruje na transkripciu a strojové učenie s použitím konvolučných neurónových sietí. Ide o proces, v ktorom sa nasnímaný obrázok mení na text. Aby mohol byť obrázok spracovaný počítačom napr. pri transkripcii, musia sa obrazové informácie, teda pixely previesť do číselnej formy. Na vstupe (Input) procesu rozpoznávania nejakého predmetu, napríklad písma, tváre, zvieraťa, auta sú pixely obrázka. Vstupný obraz má napríklad pixel 48 x 48. Potom sa postupne použijú množiny filtrov (Mapy funkcií – *Feature Maps*) na extrahovanie lokálnych obrazových príznakov (tvar, farba) prostredníctvom operácie konvolúcia (convolution), čo je matematická operácia. Filtre sú v podstate masky, ktoré sú „prehodené“ cez obrázok, aby sa zistilo, či im niečo vyhovuje. Konečný súbor funkcií sa potom vloží do husto pripojenej siete, odkiaľ pochádza skutočná univerzálna predikčná sila tohto algoritmu (*classification*). Takáto sieť sa môže naučiť aproximovať akúkoľvek primerane dobre vycvičenú funkciu s ľubovoľnou presnosťou, pokiaľ je sieť dostatočne veľká. V prípade transkripcie rukopisov to prakticky znamená, že na cvičenie modelu je potrebný veľký súbor cvičných dát. Otázka je, aký veľký by ten súbor mal byť, aby výsledky transkripcie boli čo najpresnejšie. Na základe rozdielu medzi predpoveďou modelu a *Ground Truth* sa parametre vo vnútri siete aktualizujú iteratívne.

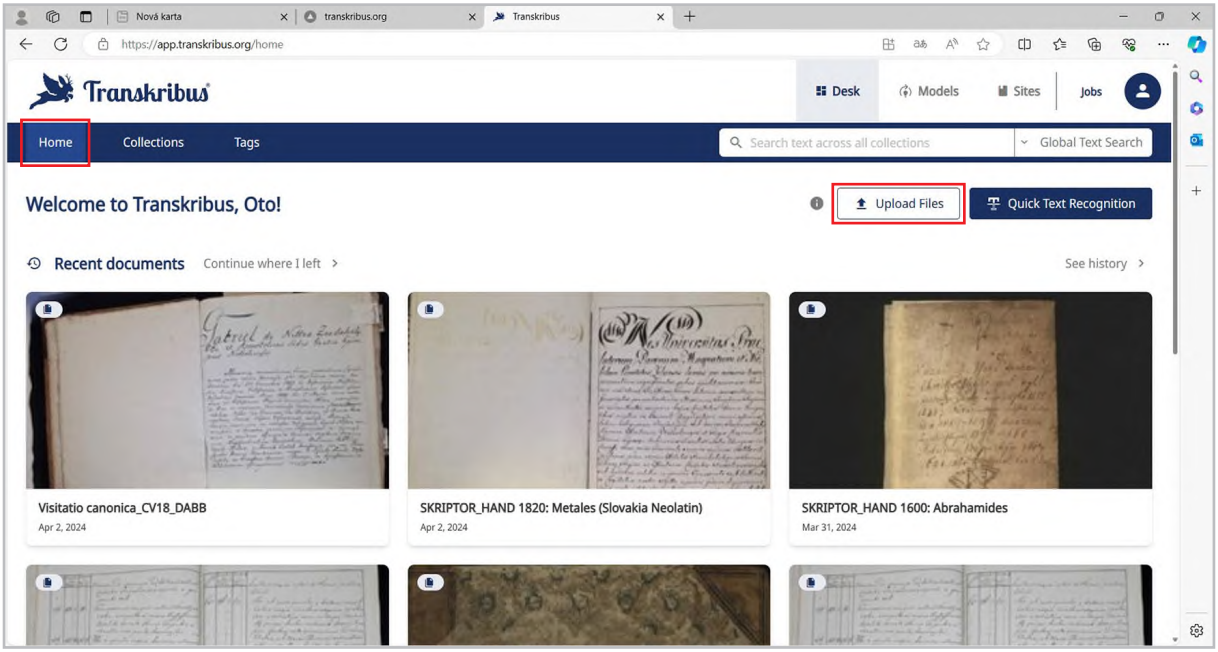
4.5 Import digitalizátov do webovej aplikácie platformy Transkribus

Importovanie digitalizovaných dokumentov (digitalizátov) na webovú platformu Transkribus je možné priamo prostredníctvom aplikácie DocScan (kapitola *4.3.5.1 Odoslanie dokumentu na platformu Transkribus*) alebo ako samostatné vopred pripravené (naskenované alebo digitálne nafotené) dokumenty z osobného počítača. Druhá možnosť je výhodná v tom prípade, ak si potrebujete digitalizáty pred ich importovaním vopred pripraviť. Ich eventuálna úprava môže zahŕňať vyselektovanie najkvalitnejších verzií digitalizátov prípadne dodatočnú úpravu ich rozmerov a kvality (napr. nastavenie jasu, kontrastu, farebnej sýtosti a pod.). Po takejto úprave digitalizátov si ich finálne verzie uložte v samostatnom adresári/priečinku osobného počítača, z ktorého ich budete neskôr importovať.

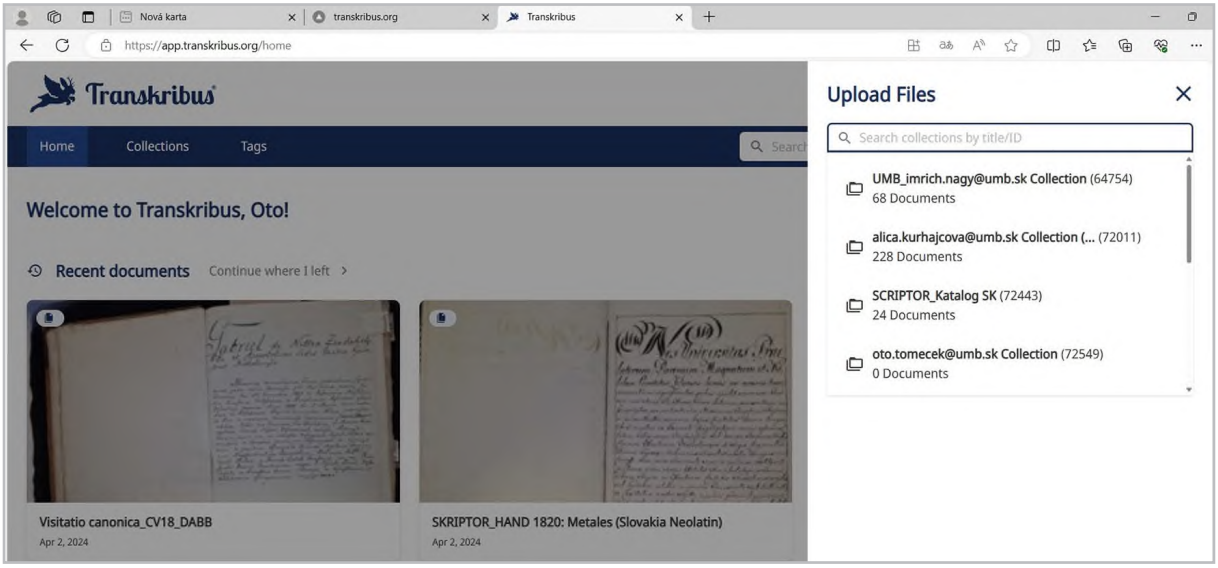
V prípade akejkoľvek práce v prostredí webovej aplikácie platformy Transkribus vrátane importovania digitalizátov je nevyhnutné byť prihlásený. Prvým krokom pri importovaní digitalizátov na server platformy je teda prihlásenie do svojho účtu prostredníctvom e-mailovej adresy a hesla. Po prihlásení si ako ďalší krok vyberte spôsob vloženia pripraveného digitalizátu na webovú platformu podľa toho či už máte alebo nemáte vytvorenú vlastnú zbierku *(Collection*).

Založenie vlastnej zbierky je nevyhnutným predpokladom a podmienkou pri vkladaní digitalizátov. Zbierka predstavuje prostredie, do ktorého vkladáte jednotlivé digitalizáty (bližšie kapitolu *3 Zbierka*).

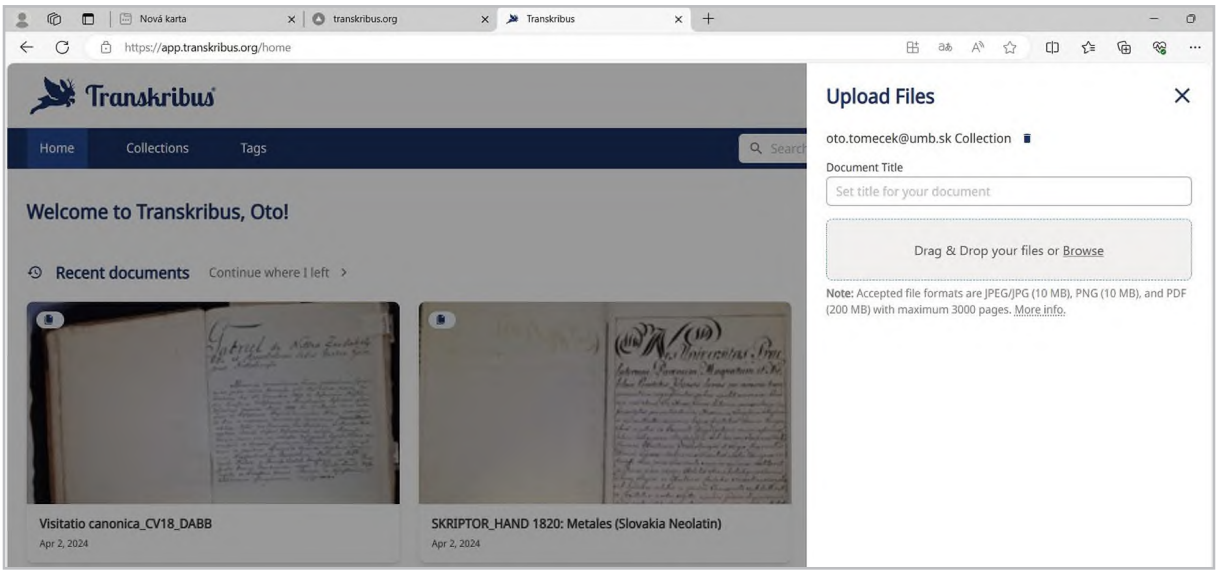
V prípade, že v prostredí platformy máte k dispozícii vlastnú zbierku, môžete pri nahrávaní digitalizátov postupovať dvoma spôsobmi. Pri prvom spôsobe kliknite na záložke *Home* na tlačidlo Nahrať súbory *(Upload Files)*. Po otvorení nového okna na pravej strane obrazovky vyberte prostredníctvom voľby prehľadávania zbierok *Search collections by title/ID* príslušnú zbierku zo zoznamu. Po tomto výbere napíšte do okna *Document Title* názov dokumentu, pod ktorým bude figurovať v tejto zbierke. Pomocou funkcie *Upload Files* pretiahnite súbor/priečinok, ktorý chcete nahrať na platformu. Druhou možnosťou je, že kliknete na funkciu nahrávania súborov *(Upload Files)* a súbor vyhľadáte podľa jeho umiestnenia vo svojom osobnom počítači.



Obrázok 50 Nahratie dokumentu cez záložku Home, krok 1

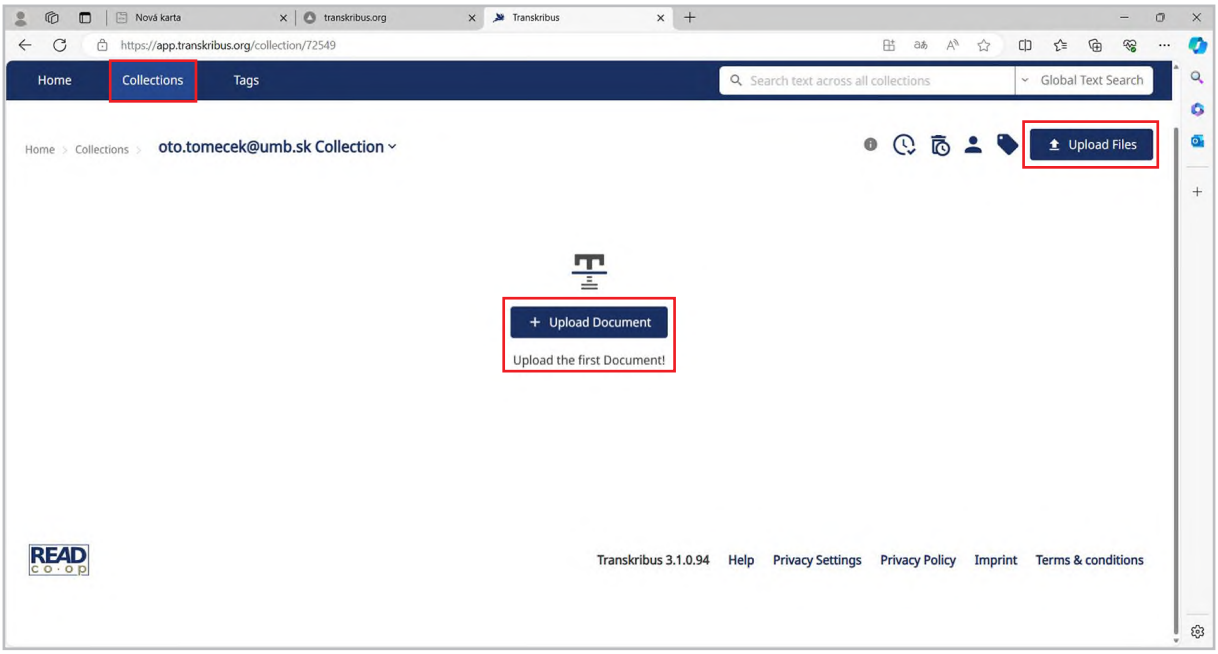


Obrázok 51 Nahratie dokumentu cez záložku Home, krok 2 – výber zbierky



Obrázok 52 Nahranie dokumentu cez záložku Home, krok 3 – zadanie názvu dokumentu a jeho vloženie

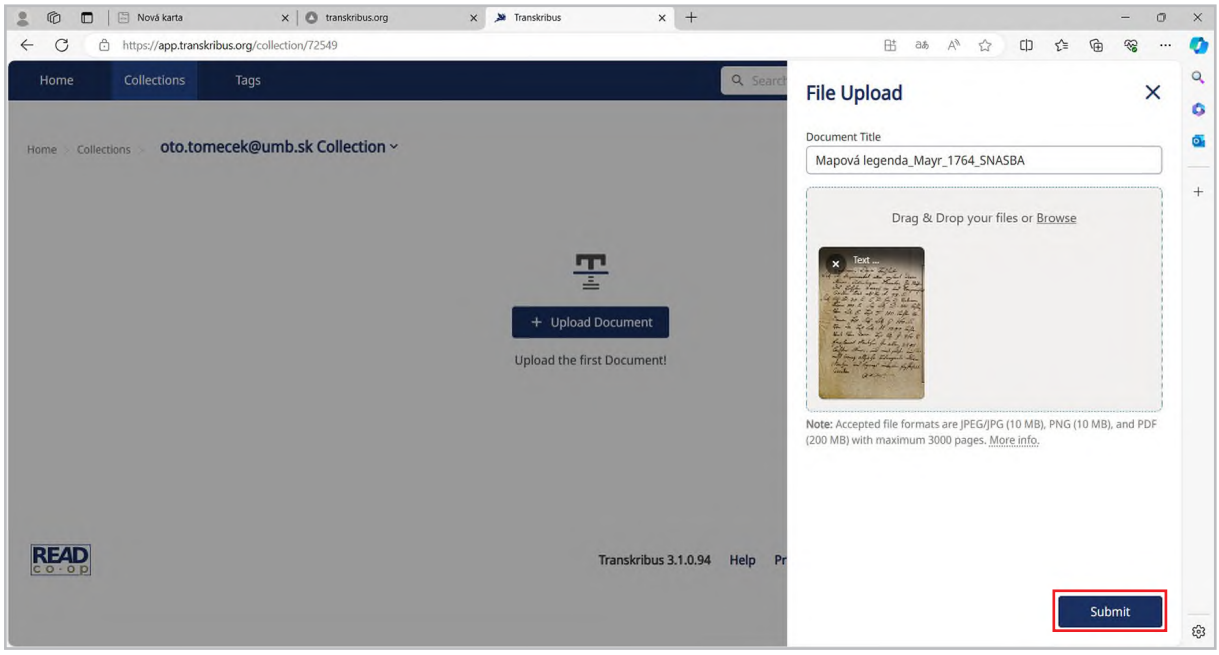
Pri druhom spôsobe sa prekliknite – podobne ako v prípade, keď ešte nemáte vlastnú zbierku – zo záložky *Home* na záložku *Collections*. Následne sa na obrazovke zobrazí prehľad zbierok prístupných vo vašom konte. Vyberte príslušnú zbierku. V prípade, že zbierka ešte neobsahuje žiadne dokumenty, uprostred obrazovky sa zobrazí funkcia +Nahrať dokument (*+Upload Document*). V pravej časti obrazovky sa otvorí nové okno s názvom *File Upload*. Ak zvolená zbierka už obsahuje nejaké dokumenty, kliknite na modro zvýraznenú funkciu *Upload Files* v pravom hornom rohu obrazovky. Rovnako ako v predošlom prípade sa v pravej časti obrazovky otvorí nové okno *File Upload*.



Obrázok 53 Nahranie dokumentu zo stránky Collections. Tlačidlo v strede slúži na nahratie prvého dokumentu, tlačidlo vpravo hore na nahratie ďalších dokumentov.

Ďalší postup je v oboch prípadoch rovnaký. Do horného riadku okna *File Upload* označeného *Document Title* napíšte názov dokumentu, pod ktorým bude figurovať v príslušnej zbierke. Do okna pod ním pretiahnite súbor/priečinok, ktorý chcete nahrať na platformu. Druhou možnos-

ťou je, že kliknete priamo do okna a súbor vyhľadáte podľa jeho umiestnenia vo svojom osobnom počítači. Keď sa súbor natiahne do uvedeného okna, jeho prenos na platformu Transkribus spustíte kliknutím na tlačidlo *Submit* v pravej dolnej časti obrazovky.



Obrázok 54 Prenos digitalizátu alebo priečinka s digitalizátmi na platformu Transkribus

Po kliknutí na položku *Submit* sa okno *File Upload* zmení na okno *Jobs*, v ktorom môžete sledovať priebeh procesu importovania dokumentov. Dĺžka prenosu dokumentov na server môže závisieť od aktuálnej vyťaženosti samotného servera, predovšetkým však od veľkosti prenášaných dokumentov (digitalizátov). Po ukončení nahrávania digitalizátov je potrebné prekliknúť sa na záložku *Home* a následne sa znova vrátiť na záložku *Collections*. Tým by mali byť importované digitalizáty viditeľné v rámci predmetnej zbierky. Všetky obrázky vybraté naraz sa nahrajú ako jeden dokument.

Aktuálne je možné digitalizáty vkladať len v podporovaných formátoch JPEG/JPG, PNG a PDF. Jednotlivé digitalizované obrázky vo formátoch JPEG a PNG by nemali presahovať veľkosť 10 MB. Rozlíšenie obrázkov by malo ideálne dosahovať hodnotu 300 dpi. Vyššie rozlíšenie nie je potrebné, keďže nijako neprispieva k zlepšeniu rozpoznania a automatickej transkripcie textu. Vložený súbor alebo priečinok s viacerými obrázkami sa po naimportovaní na platformu považuje za samostatný dokument. Každý obrázok v rámci súboru/priečinka predstavuje jednu stranu dokumentu. V prípade vkladania dokumentu vo formáte PDF by nemala jeho veľkosť presahovať 200 MB. Maximálny počet vložených strán v rámci jedného súboru vo formáte PDF by nemal presahovať hodnotu 3000.

Okrem toho v okne odovzdávania získate informácie o obmedzeniach nahrávania týkajúcich sa typov súborov, veľkostí a počtu strán. Ak súbory nespĺňajú tieto kritériá, dostanete upozornenie o chybe.

Iné spôsoby vkladania digitalizátov aktuálne nie sú prístupné. V rámci dnes už nepodporovanej platformy Transkribus expert klient je však možné využiť aj iné cesty vkladania digitalizátov. Prvým spôsobom je cesta *Private FTP*, ktorá umožňuje nahrať viac súborov/priečinkov naraz. Tento spôsob si vyžaduje inštaláciu klienta *FTP*. Pri druhom spôsobe nahrávania dokumentov, cestou *IIIF manifest*, sa dokumenty vkladajú priamo z webovej stránky po skopírovaní a vlože-

ní URL adresy. Tento spôsob je možné použiť vtedy, ak inštitúcia (archív, knižnica) poskytuje online prístup k svojim zdigitalizovaným dokumentom prostredníctvom štandardu *IIIF*. Aj pri treťom spôsobe nahrania dokumentov, cestou *DFG Viewer METS*, sa dokumenty vkladajú na platformu priamo z webu jednoduchým vložením URL adresy do určeného poľa. Do budúcnosti sa očakáva, že všetky tieto tri, aktuálne nedostupné, cesty nahrávania dokumentov budú neskôr dostupné aj na webovej aplikácii platformy Transkribus.

4.5.1 Privátnosť zbierok a dokumentov

Všetky dokumenty nahraté na platformu Transkribus sú štandardne súkromné. Sú uložené na serveroch spoločnosti READ-COOP SCE, ktorá softvér vyvíja a spravuje. Všetky servery sú umiestnené v rakúskom Innsbrucku v súlade s GDPR a údaje môžu byť spracovávané podľa podmienok na webovej stránke READ-COOP SCE.

Ak chcete zdieľať svoju zbierku s ostatnými používateľmi Trankribusu, aby ste mohli spolupracovať s viacerými používateľmi, postupujte podľa kapitoly *3.2 Správa používateľov zbierky*.

5 Segmentácia dokumentov v Transkribuse

Keď máte dokument nahratý v aplikácii Transkribus, môžete začať s rozpoznávaním rozloženia (*Layout Recognition*). Výsledkom analýzy je segmentácia snímok dokumentu, t. j. identifikácia jednotlivých prvkov, rozlíšenie štruktúry, horizontálnej orientácie textu a určenie poradia čítania textu.

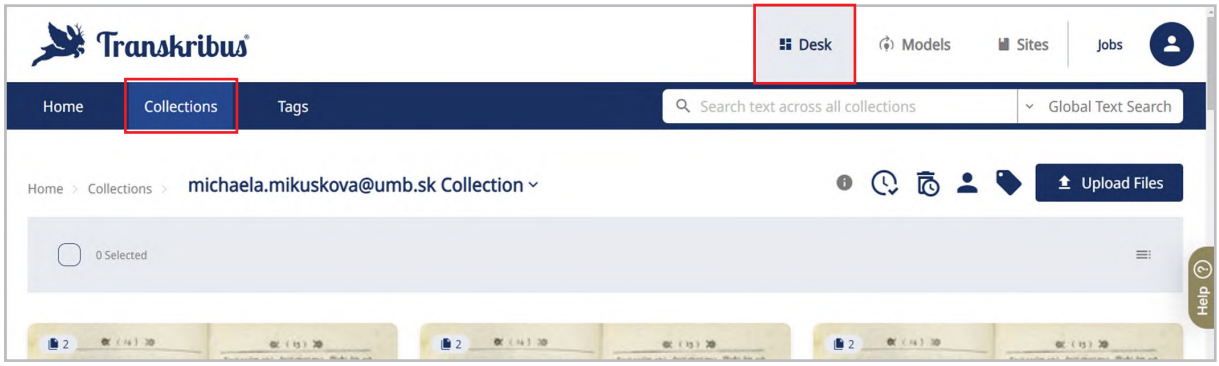
Pri segmentácii sa uplatňuje metóda analýzy obrazu a textovej analýzy, ktorých výsledkom je členenie textu na časti, resp. objekty. Tie sa následne budú prepájať s textom, ktorý je výsledkom transkripcie.

Každý objekt segmentácie určuje, kde sa na nachádzajú:

- **textové rámce** (*Text Regions*) – vymedzujú oblasti s textom, môže ísť o hlavný text dokumentu, čísla strán, marginálie, tabuľky a i., označené sú zeleným obrysom v tvare obdĺžnika,
- **oblasti čiar** (*Lines*) – vymedzujú čiaru, ktorá sa tiahne pozdĺž spodnej časti textového riadku. Ide o najdôležitejší referenčný bod na rozpoznávanie textu, na základe ktorého sa v ďalšom kroku softvér učí čítať jednotlivé znaky; v rámci textových rámcov majú čiary modrú farby, po kliknutí na čiaru sa oblasť s textom nad ňou vysvieti oranžovým obrysom,
- okrajové a nadbytočné časti dokumentu, ktoré nie sú dôležité pre proces transkripcie a trénovania modelu.

Súradnice objektov sa v procese segmentácie ukladajú do súboru príslušnej stránky dokumentu. Správna segmentácia textu výrazne ovplyvňuje prepis dokumentu, kvalitu vytrénovaného modelu, korekciu transkripcie a proces spracovania prepísaného textu.

Pre spustenie segmentácie si v hlavnej ponuke v hornej časti aplikácie otvorte záložku *Desk* a v tmavomodrej lište zvoľte záložku *Collections*.



Obrázok 55 Nastavenie aplikácie na vykonanie segmentácie

5.1 Spôsoby segmentácie

Rozpoznanie rozloženia (*Layout Recognition*) sa v aplikácii Transkribus vykonáva:

- **automaticky spustením úlohy rozpoznávania textu** – analýza rozloženia objektov na snímkach dokumentu a prepis textu v jednom kroku (viac v kapitole *5.1.6 Automatická segmentácia a rozpoznávanie textu*),
- **segmentáciou objektov na snímkach v samostatnom kroku**.

Voľba spôsobu segmentácie závisí od typu dokumentu, s ktorým pracujete a očakávaných výsledkov. Ak chcete text prepisovať ručne, plánujete vytrénovať nový model na prepis dokumen-

tu, resp. plánujete vytrénovať vlastný model na rozpoznanie rozloženia objektov na snímkach dokumentu, odporúčame proces segmentácie a prepisu dokumentu oddeliť.

Aj samotnú segmentáciu prvkov na snímkach môžete urobiť dvomi spôsobmi:

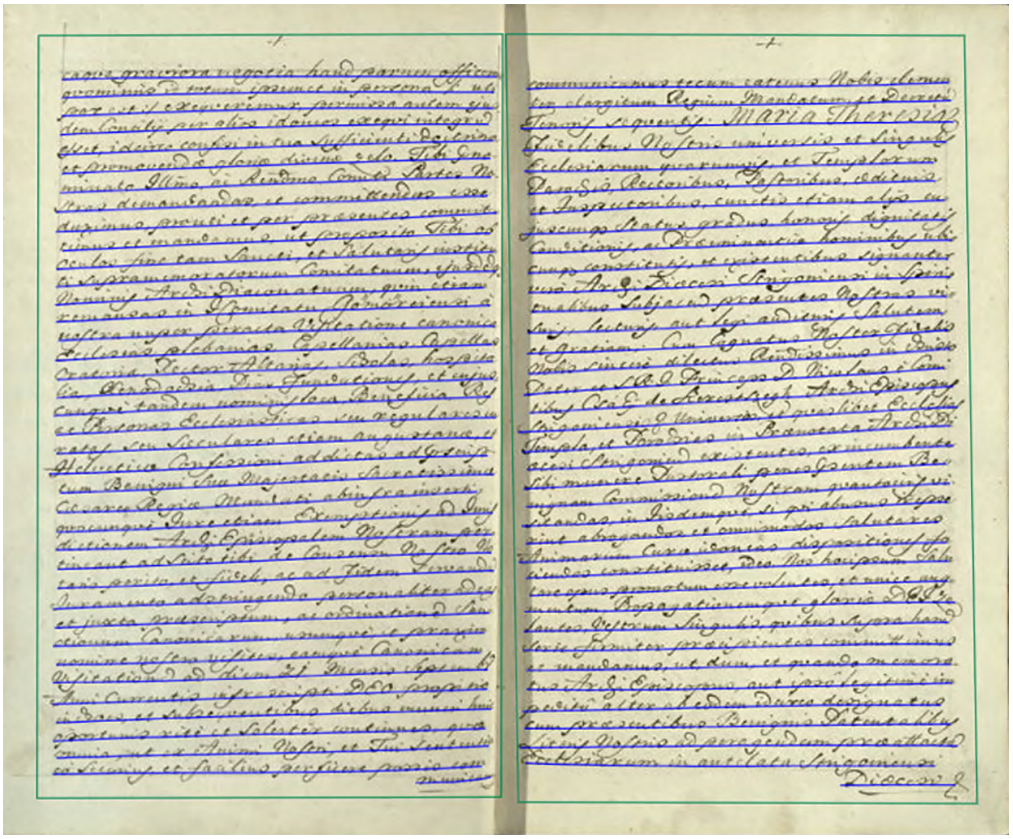
- **automaticky** – označenie textových rámcov a riadkov necháte urobiť výlučne softvér,
- **manuálne** – spočíva v manuálnom vytvorení textových rámcov a automatickej segmentácii riadkov.

Objekty segmentácie, textové rámce a riadky môžete označiť aj výlučne manuálnym spôsobom, t. j. bez použitia funkcií automatickej segmentácie. Ide však o veľmi prácny a časovo náročný proces. Nástroje na tvorbu objektov sú popísané v kapitole *5.2 Opravy po automatickej a manuálnej segmentácii*.

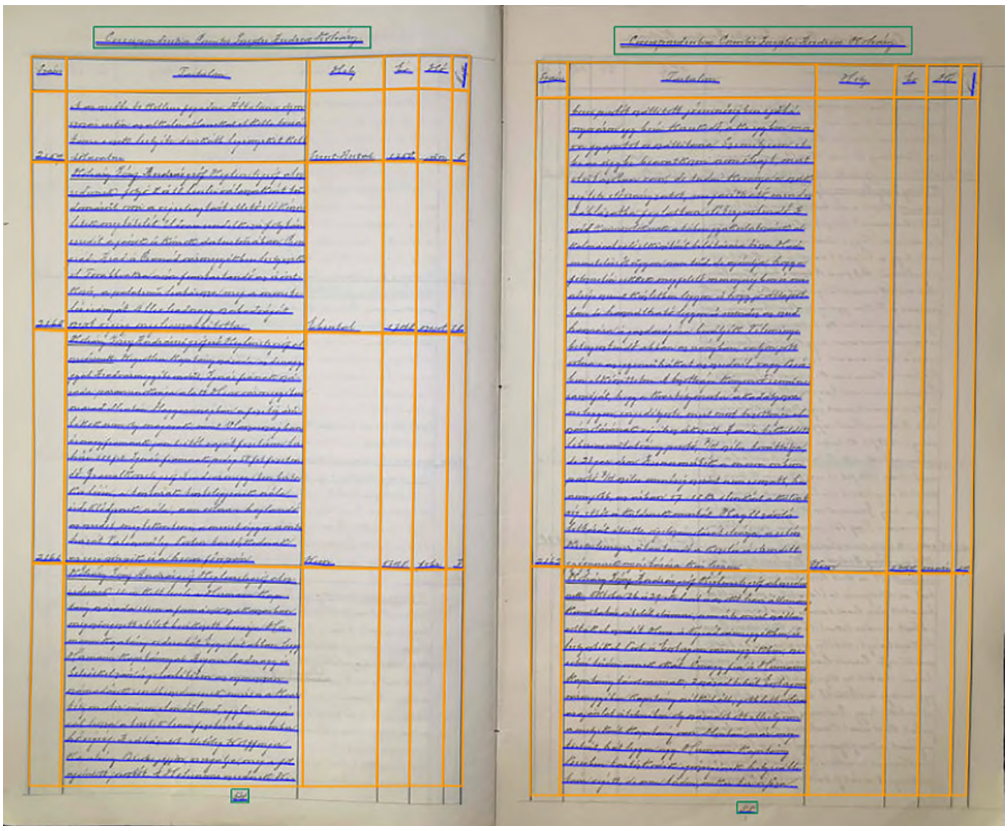
Výber spôsobu segmentácie závisí od štruktúry a obsahu dokumentu, s ktorým pracujete. Nesprávne zvolený typ segmentácie môže viesť k časovo náročným opravám. Automatická segmentácia rozpozná, kde sa text na snímke dokumentu graficky nachádza, rozozná základné textové rámce a riadky v nich, ale nerozlišuje typ obsahu. Text vo vytvorených rámcoch zoradí podľa súradníc objektov na snímke, spravidla od ľavého horného rohu smerom nadol. Automatickú segmentáciu je preto vhodné použiť na dokumenty s jednoduchou štruktúrou a jasným poradím riadkov.

Pri komplikovanom rozložení textu je však potrebné zadefinovať viac textových rámcov. Manuálnu segmentáciu je vhodné použiť pri členitom obsahu a zložitejšej štruktúre textu dokumentu, napr. ak text obsahuje poznámky pod čiarkou, stĺpce, tabuľky, alebo sa v dokumente vyskytujú marginálie a i.

Na obrázkoch nižšie vidieť príklady dokumentov vhodných na automatickú a manuálnu segmentáciu.



Obrázok 56 Príklad dokumentu s jasnou štruktúrou poradia textových rámcov a riadkov vhodného na automatickú segmentáciu



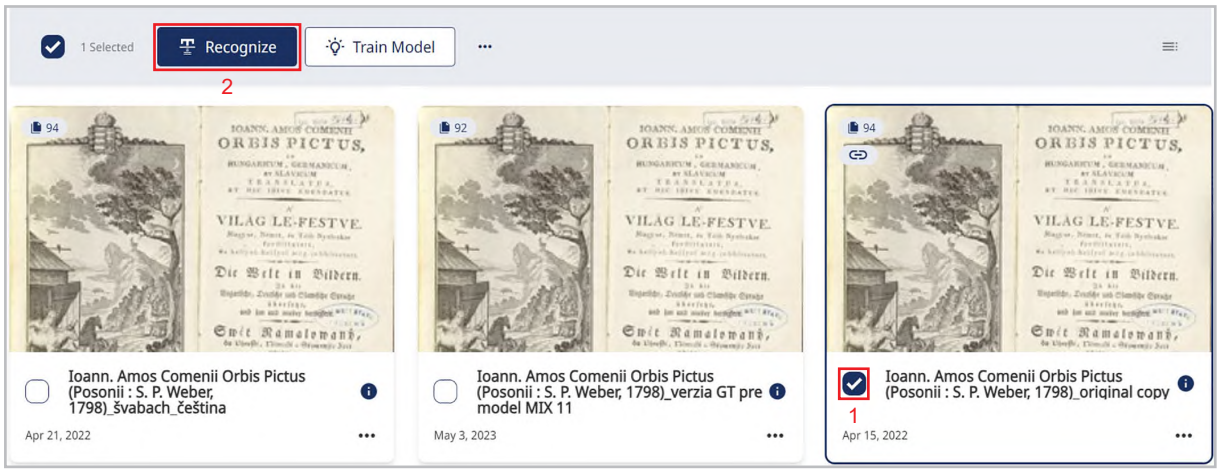
Obrázok 57 Príklad dokumentu so zložitou štruktúrou textu vhodného na manuálnu segmentáciu

5.1.1 Výber strán

Automatickú a manuálnu segmentáciu môžete spustiť na celom dokumente alebo na vybraných stranách. Na začiatok odporúčame spustiť segmentáciu na jednej strane, aby ste si overili, či je vybraný spôsob segmentácie pre váš typ dokumentu vyhovujúci.

Nastavenie rozsahu segmentácie:

- **celý dokument** – zakliknite prázdne okienko na úrovni zoznamu dokumentov vo vlastnej zbierke a následne kliknite na tlačidlo *Recognize*, ktoré sa zobrazí v lište nad zoznamom dokumentov,



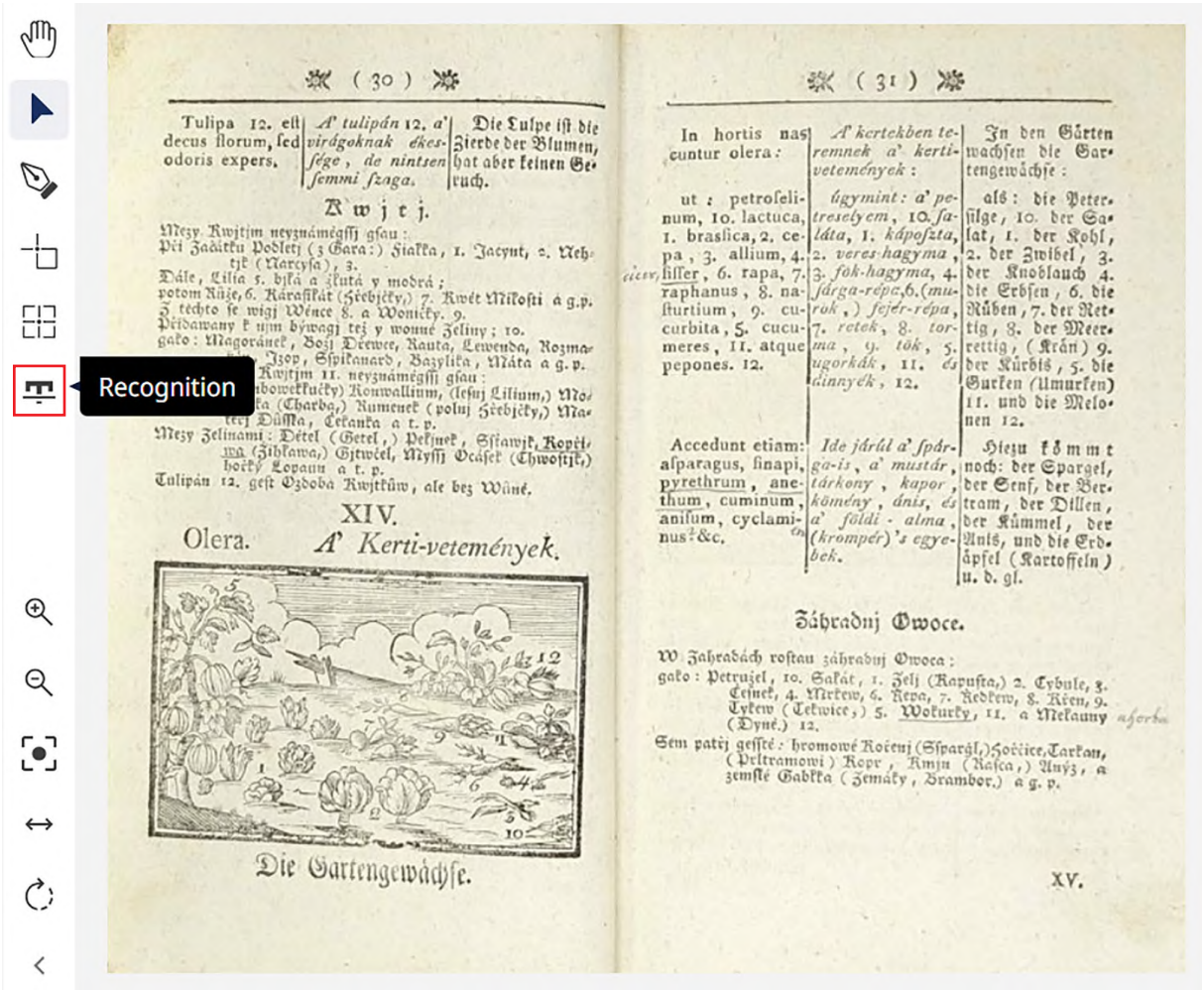
Obrázok 58 Výber segmentácie na úrovni celého dokumentu

- **na vybraných snímkach dokumentu** – zakliknite prázdne okienko na úrovni zoznamu snímok vybraného dokumentu a následne kliknite na tlačidlo *Recognize,* ktoré sa zobrazí v lište nad zoznamom,



Obrázok 59 Výber segmentácie na úrovni snímok/stránok dokumentu

- **na jednej snímke dokumentu** – kliknite na ikonku *(Recognition)*, ktorá sa nachádza vo vertikálnom paneli nástrojov v ľavej časti okna editora.



Obrázok 60 Výber segmentácie na úrovni editora dokumentu

5.1.2 Automatická segmentácia

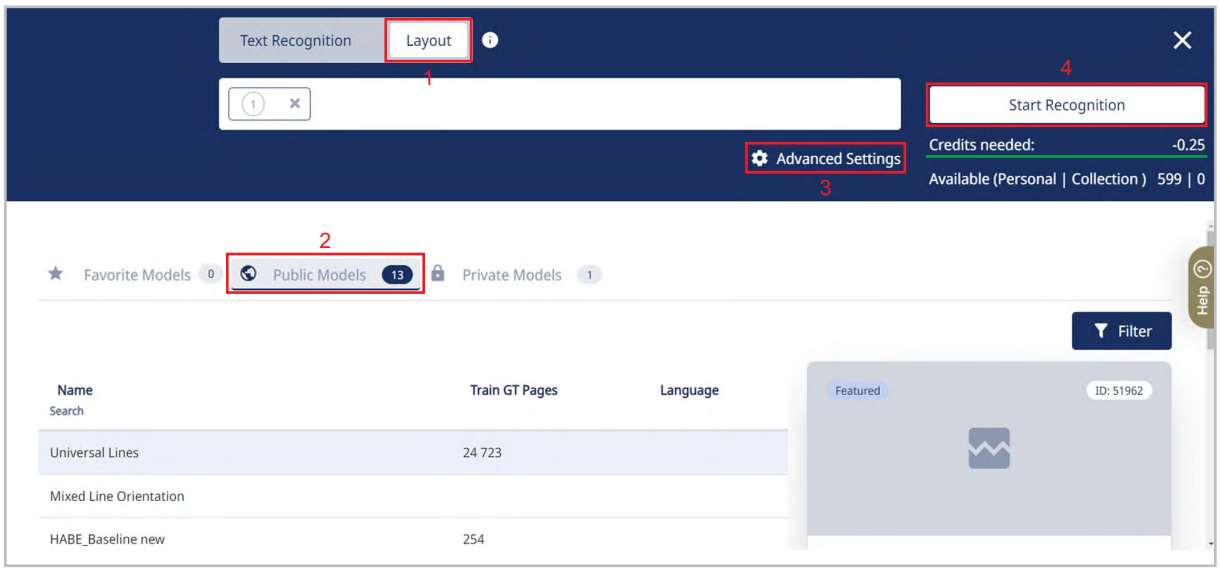
Pri automatickej segmentácii softvér na snímke dokumentu sám vyznačí textové rámce a riadky, zároveň určí aj poradie ich čítania.

Nastavenie a spustenie automatickej segmentácie

Pred spustením segmentácie treba:

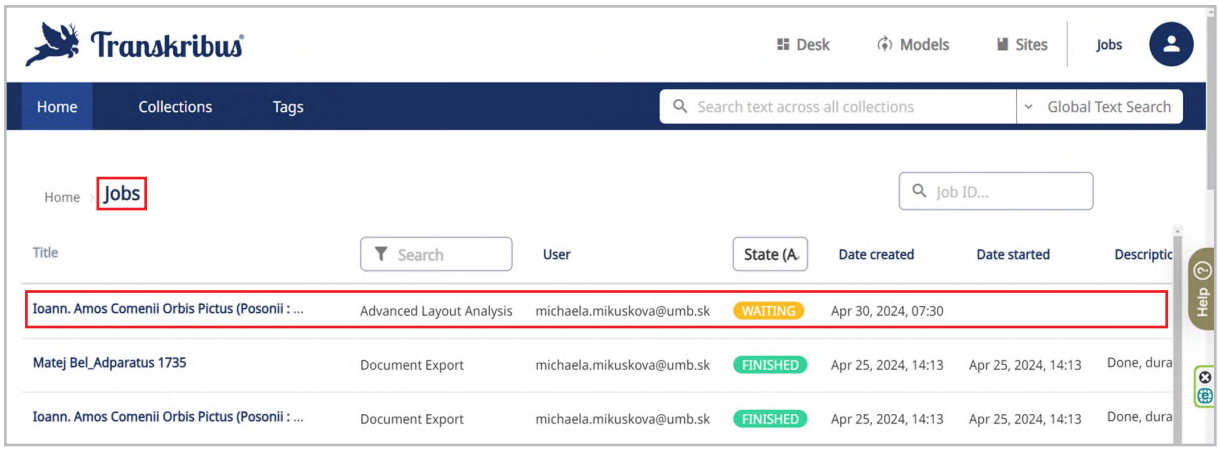
1. vybrať snímky (dokument), na ktorých chcete segmentáciu vykonať (viac v kapitole *5.1.1 Výber strán*),
2. prejsť na funkciu rozpoznávania cez ikonu alebo na tlačidlo *Recognize*,
3. v otvorenom okne s nástrojmi rozpoznávania v hornej časti sa prepnúť do sekcie Rozloženie (*Layout*),
4. vybrať najvhodnejší z dostupných modelov na rozpoznanie rozloženia (viac v kapitole *5.1.4 Výber modelu*),
5. ak je to potrebné, kliknúť na rozšírené nastavenia (*Advanced Settings*) a v otvorenom dialógovom okne upraviť parametre segmentácie (viac v kapitole *5.1.5 Pokročilé nástroje segmentácie*).

Segmentáciu spustíte kliknutím na tlačidlo *Start Recognition*.



Obrázok 61 Dôležité prvky nastavenia a spustenia automatickej segmentácie v dialógovom okne Layout. Zelenou farbou je podčiarknutý počet kreditov potrebných na tento úkon.

Po spustení segmentácie sa zobrazí záložka *Jobs*, kde vidíte prehľad úloh, ktoré ste na serveroch Transkribusu zadávali.



Obrázok 62 Zobrazenie zoznamu úloh a stavu ich riešenia na záložke Jobs. V červenom rámčeku je označená úloha automatickej segmentácie.

Záložku *Jobs* môžete opustiť a pracovať na iných úkonoch, prípadne aj ukončiť prácu s aplikáciou Transkribus. Stav riešenia zadanej úlohy si môžete kedykoľvek skontrolovať kliknutím na záložku *Jobs* ktorá sa nachádza sa v hlavnej ponuke v hornej časti aplikácie pri prihlásení. Po ukončení procesu si zobrazte príslušnú snímku a skontrolujte výsledok segmentácie.

5.1.3 Manuálna segmentácia

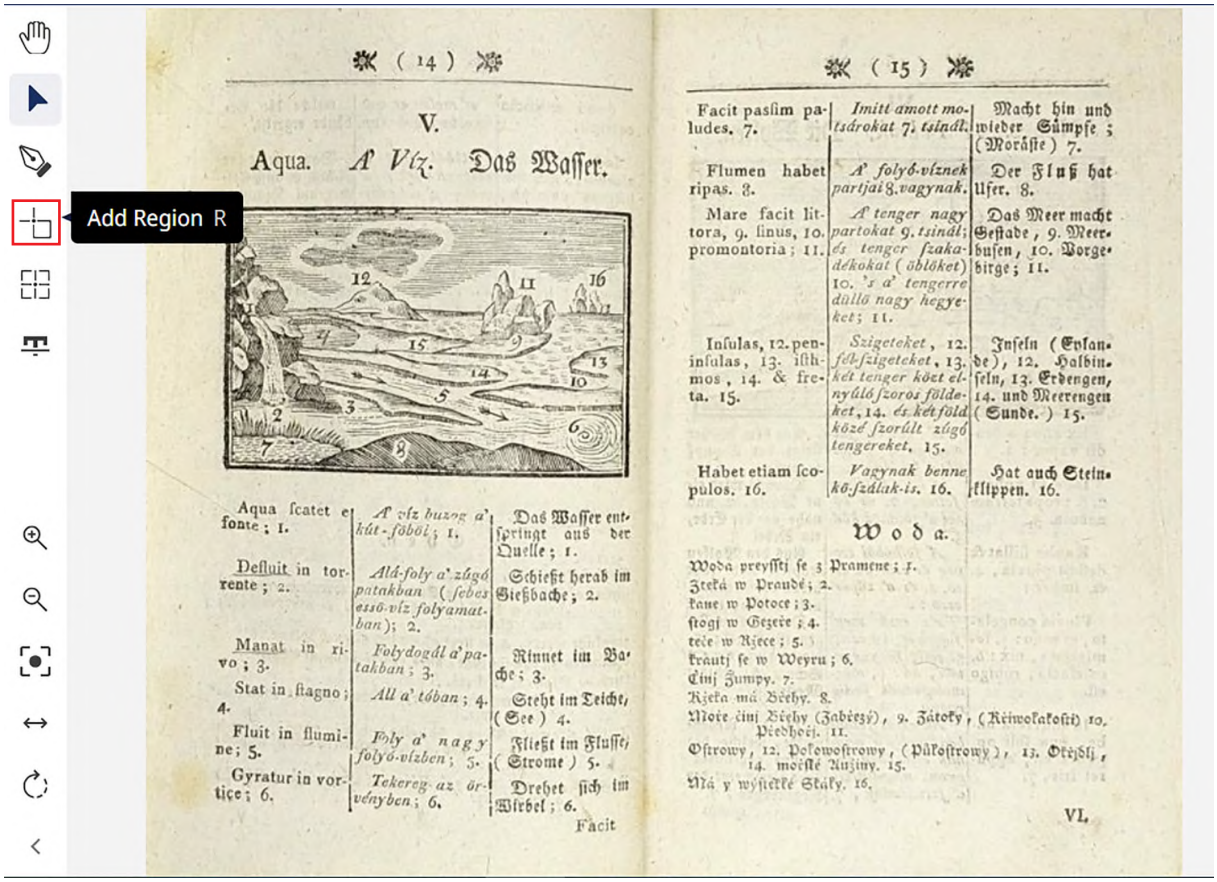
Pri manuálnej segmentácii je proces segmentácie rozdelený do dvoch krokov:

1. manuálne označenie textových rámcov,
2. spustenie automatickej segmentácie riadkov v manuálne označených textových rámcoch.

Treba starostlivo zvážiť členenie textu a jeho rozdelenie na textové rámce. Počet rámcov na jednej strane dokumentu závisí od jeho štruktúry a obsahu.

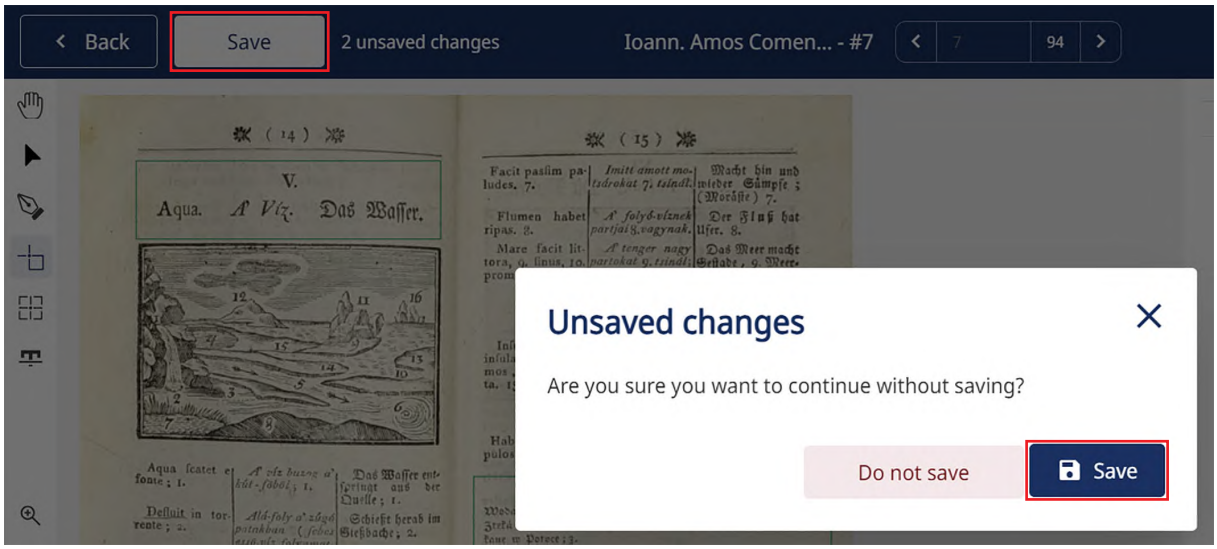
Manuálne označenie textových rámcov

Nástroje na vytváranie textových rámcov a iné úpravy segmentácie nájdete vo vertikálnom paneli nástrojov v ľavej časti okna editora. Na vytvorenie rámca kliknite na ikonku Pridať textový rámec (*Add region*) alebo na klávesnici stlačte písmeno R.



Obrázok 63 Pridanie textového rámca

Textové rámce majú zvyčajne tvar štvoruholníka (štvorca alebo obdĺžnika). Na snímke kurzorom kliknite na miesto, z ktorého chcete textový rámec začať vytvárať, t. j. na oblasť, kde sa bude nachádzať jeden z vrcholov štvoruholníka. Z tohto miesta postupným ťahaním po snímke označte priestor, ktorý bude vymedzovať príslušný textový rámec. Označenie textového rámca ukončíte kliknutím. Hranice vytvoreného rámca označujú zelené čiary. Tie môžete ďalej upravovať a posúvať. Nesprávne vytvorený rámec môžete vymazať (viac v kapitole *5.2 Opravy po automatickej a manuálnej segmentácii*). Pri prechode na inú stránku dokumentu zmeny uložte kliknutím na ikonu Uložiť (*Save*) v hornej časti editora alebo v dialógovom okne, ktoré sa otvorí vždy, keď opúšťate snímku a vykonané zmeny na nej ste predtým neuložili.



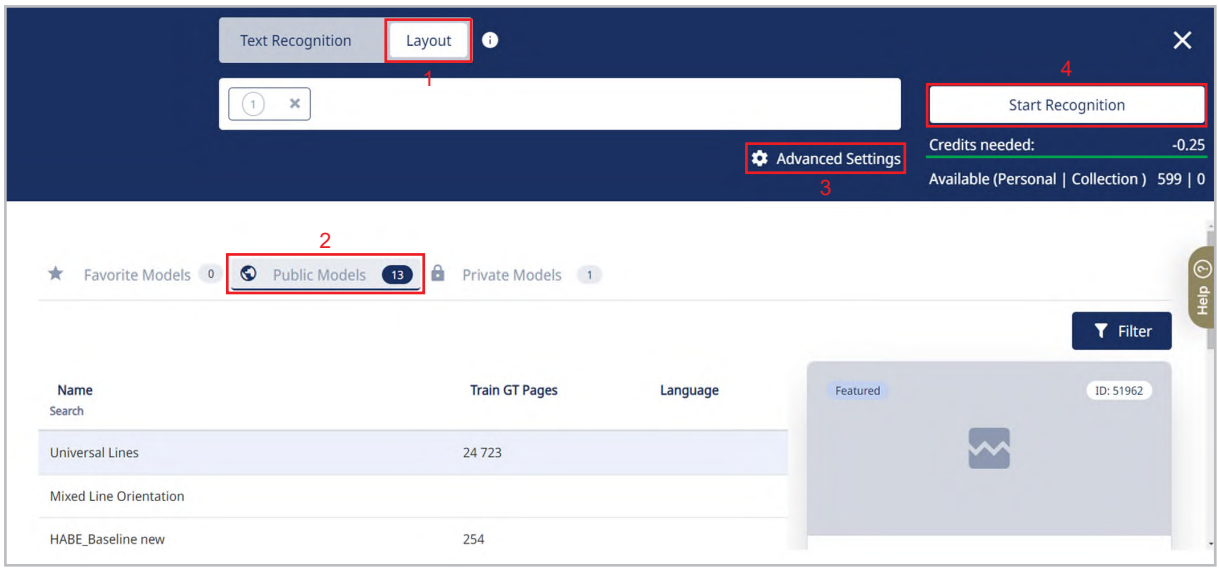
Obrázok 64 Ikony na uloženie vykonaných zmien

Spustenie automatickej segmentácie riadkov

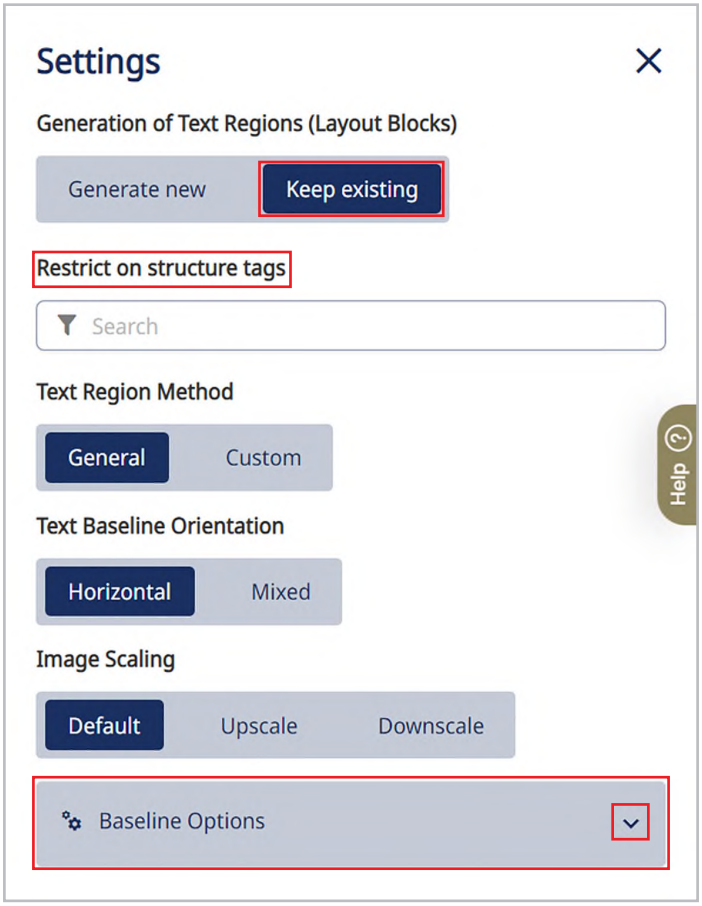
Keď máte vymedzené textové rámce, môžete pristúpiť k automatickej segmentácii riadok. Tento proces môžete spustiť na jednej alebo viacerých snímkach dokumentu. Pred spustením automatickej segmentácie riadkov treba:

1. vybrať snímky (dokument), na ktorých chcete segmentáciu vykonať (viac v kapitole *5.1.1 Výber strán*),
2. prejsť na funkciu rozpoznávania cez ikonu alebo na tlačidlo *Recognize*,
3. v otvorenom okne s nástrojmi rozpoznávania v hornej časti sa prepnúť do sekcie Rozloženie (*Layout*),
4. vybrať najvhodnejší z dostupných modelov na rozpoznanie rozloženia (viac v kapitole *4.1.4 Výber modelu*),
5. kliknúť na Rozšírené nastavenia (*Advanced Settings*) – v otvorenom dialógovom okne pri nastavení generovania textových rámcov *Generation of Text region (Layout Blocks)* prepnite na voľbu *Keep existing* – softvér pri rozpoznávaní rozloženia ponechá segmentáciu vami zadaných textových rámcov.

Segmentáciu spustíte kliknutím na tlačidlo *Start Recognition*.



Obrázok 65 Dôležité prvky nastavenia a spustenia manuálnej segmentácie v dialógovom okne Layout. Zelenou farbou je podčiarknutý počet kreditov potrebných na tento úkon.



Obrázok 66 Nastavenie ponechania už zadefinovaných textových rámcov

Segmentáciu je možné obmedziť len na určitý typ štruktúry obsahu pomocou funkcie *Restrict on structure tags* (viac v kapitole *8 Možnosti práce s textom po automatickej transkripcii*). V nastaveniach je možné navoliť viac štrukturálnych metadát súčasne. Pri vyhľadávaní systém automaticky ponúka dostupné metadáta. Metadáta však musia byť na stránkach vopred označe-

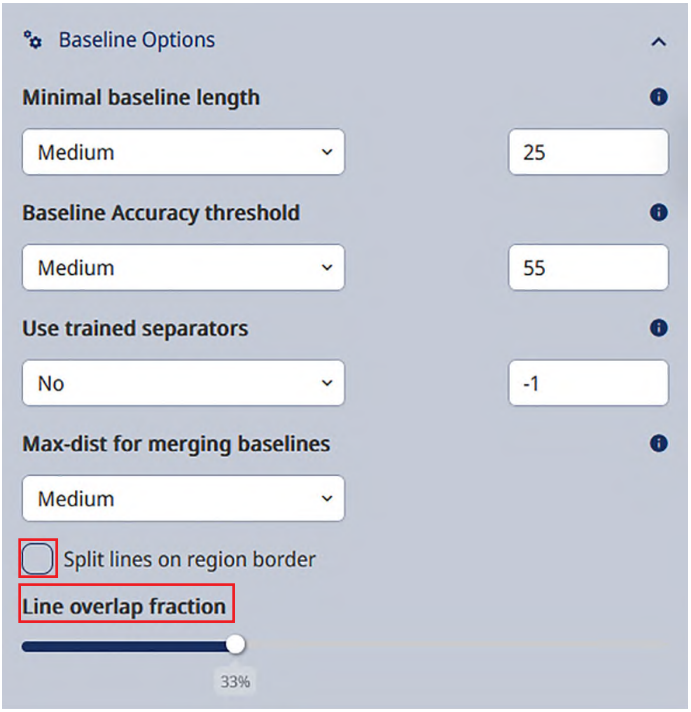
né. Tieto nastavenia je vhodné použiť v prípade, že trénujete model na automatickú segmentáciu dokumentu, s ktorým pracujete.



Obrázok 67 Výber štrukturálnych metadát pre segmentáciu

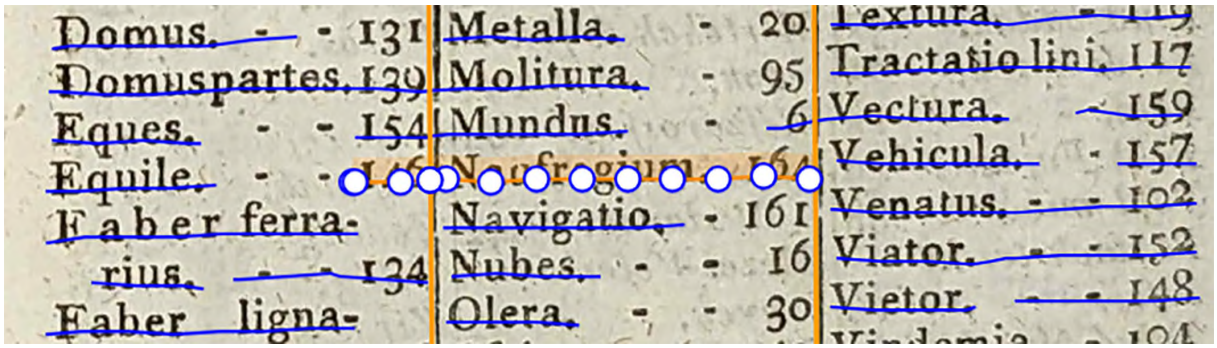
Ak ste v nastavení pre generovanie textových rámcov *Generation of Text region (Layout Blocks)* prepli na voľbu Ponechať existujúce rámce (*Keep existing*), rozbaľte ponuku ďalších nastavení kliknutím na voľbu *Baseline Options*. Zakliknutie funkcie delenia riadkov na hranici textových rámcov (*Split lines on region border*) zabezpečí, že sa riadky striktne riadia hranicou textového rámca. Toto nastavenie je veľmi užitočné pri segmentácii dokumentov, v ktorých sa textové rámce nachádzajú tesne vedľa seba a text v nich sa končí takmer na hranici rámca. Zabránite tým spojeniu čiar do jednej dlhej čiary (príklady na obrázkoch nižšie).

Nastaviť môžete podiel prekrytia čiar a textových rámcov nastavením hodnoty *Line overlap fraction*. Ak chcete, aby čiary prekrývali hranice textových rámcov treba hodnotu zvýšiť. V prípade, že ste zaklikli funkciu *Split lines on region border*, odporúčame ponechať toto nastavenie na najnižšej hodnote.

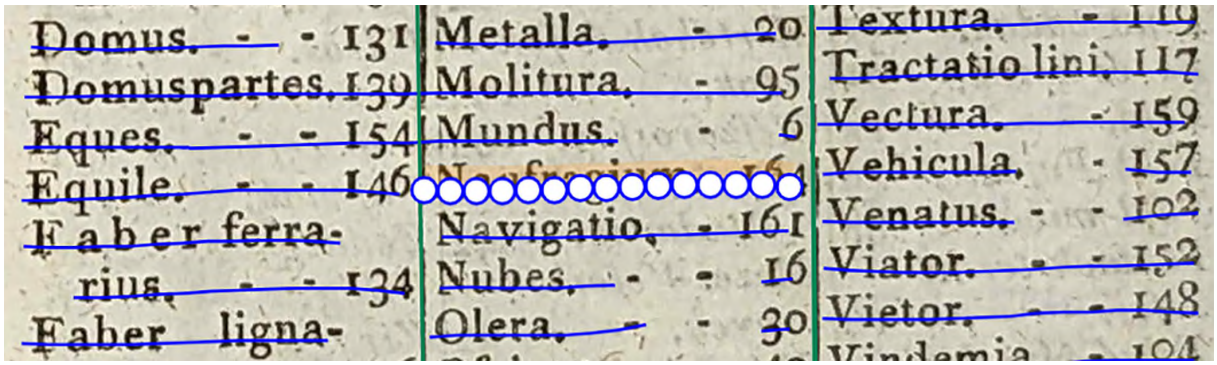


Obrázok 68 Výber funkcie delenia riadkov na hranici textových rámcov a nastavenie hodnoty prekrytia

Segmentáciu spustíte kliknutím na tlačidlo *Start Recognition*.

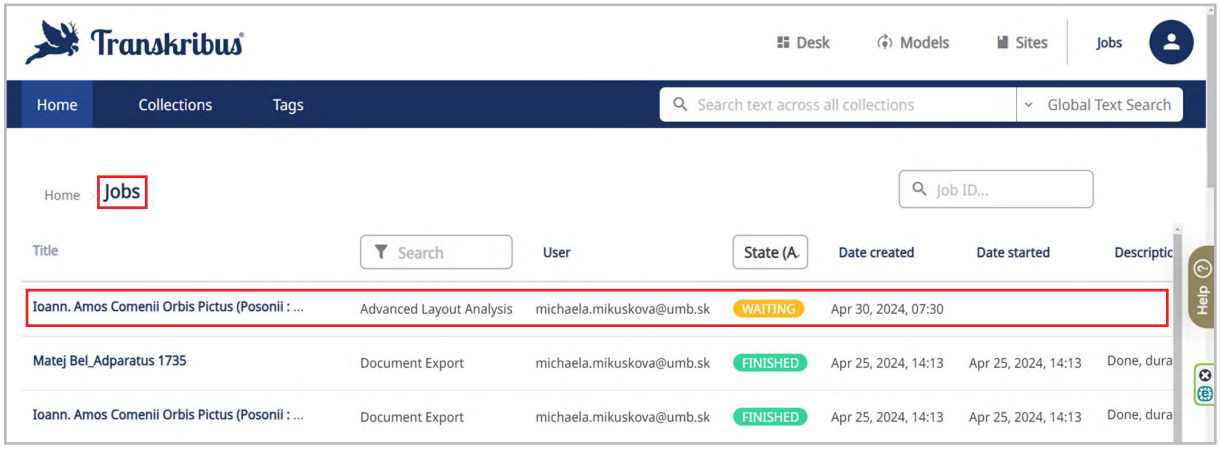


Obrázok 69 Výsledok automatickej segmentácie bez použitia funkcie delenia riadkov na hranici textových rámcov – riadky v dvoch susediacich textových rámcoch sú spojené



Obrázok 70 Výsledok automatickej segmentácie s použitím funkcie delenia riadkov na hranici textových rámcov – riadky v dvoch susediacich textových rámcoch sú oddelené

Po spustení segmentácie sa zobrazí záložka *Jobs*, v ktorej vidíte prehľad úloh, ktoré ste na serveroch Transkribusu zadávali.



Obrázok 71 Zobrazenie zoznamu úloh a stavu ich riešenia na záložke Jobs. V červenom rámčeku je označená úloha automatickej segmentácie.

Záložku *Jobs* môžete opustiť a pracovať na iných úkonoch, prípadne aj ukončiť prácu s aplikáciou Transkribus. Stav riešenia zadanej úlohy si môžete kedykoľvek overiť kliknutím na záložku *Jobs*, ktorá sa nachádza sa v hlavnej ponuke v hornej časti aplikácie po prihlásení. Po ukončení procesu si zobrazte príslušnú snímku a skontrolujte výsledok segmentácie.

5.1.4 Výber modelu

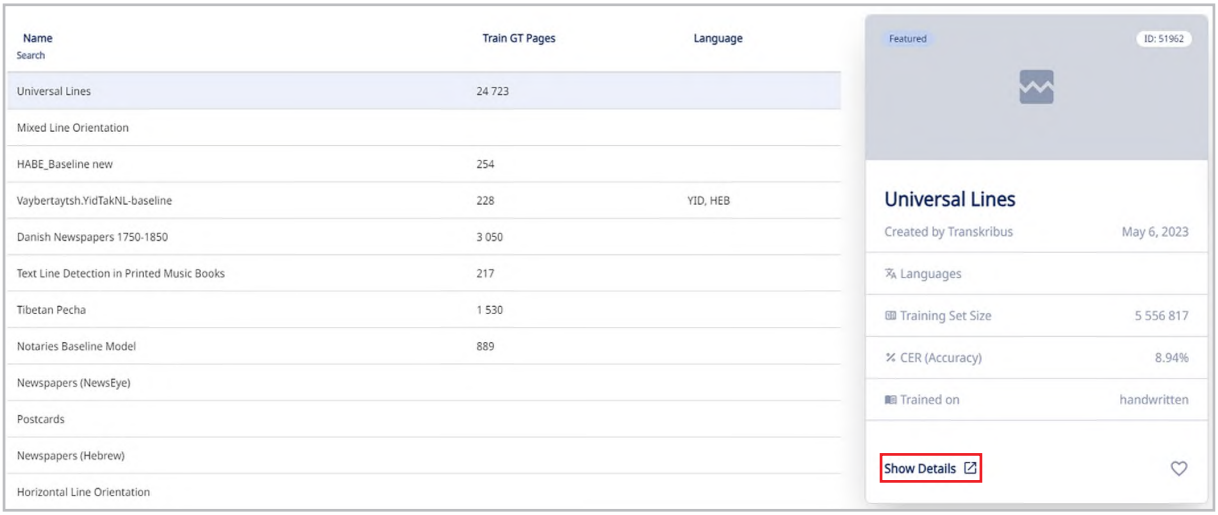
Pred spustením automatickej segmentácie si môžete vybrať z niekoľkých verejných modelov, ktoré slúžia na rozpoznávanie štruktúry dokumentov s rôznym usporiadaním textu.

K dispozícii je niekoľko (v čase písania tejto príručky ich bolo trinásť) vytrénovaných modelov rozloženia obsahu, ktoré vytrénovali vývojári softvéru Transkribus alebo používateľská komunita:

- *Universal Lines* – najvšeobecnejší model, ktorý je na platforme aktuálne k dispozícii. Tento model odporúčame použiť, ak si nie ste istí výberom optimálneho modelu, ktorý bude vyhovovať vlastnostiam dokumentu, s ktorým pracujete,
- *Mixed Line Orientation* – model pre rôznorodé rozloženie textu na snímkach, t. j. text je napísaný viacerými smermi,
- *Horizontal Line Orientation* – model pre dokumenty s homogénnym rozložením textu, t. j. len horizontálne alebo vertikálne čiary.

Na výber sú aj modely zohľadňujúce štruktúru novín, pohľadníc a modely vytrénované pre špecifickú typológiu dokumentu. Použiť môžete aj vlastný model, ktoré ste si vycvičili na základe štruktúry objektov špecifických pre typ dokumentu, s ktorým pracujete.

Základné informácie o každom modeli sa zobrazia v okienku vedľa zoznamu. Podrobnejšie informácie získate kliknutím na voľbu detailu *Show Details*. V základných nastaveniach je predvolený verejný model (*Public model*) Universal Line.



Obrázok 72 Ponuka modelov na segmentáciu dokumentu

5.1.5 Pokročilé nástroje na nastavenie automatickej segmentácie textu

Proces automatickej segmentácie je defaultne nastavený a nemusí vyhovovať každému dokumentu. Používatelia aplikácie Transkribus majú k dispozícii nástroje na úpravu predvolených parametrov. Dialógové okno s rozšírenou ponukou na nastavenie konfigurácie sa otvorí po kliknutí na Rozšírené nastavenia (*Advanced Settings*) v hornej časti okna názvom vybraného dokumentu.



Obrázok 73 Otvorenie nástrojov konfigurácie automatickej segmentácie

Nastavenie pozostáva z dvoch krokov:

1. úprava parametrov segmentácie textových rámcov a obrázkov,
2. úprava parametrov segmentácie riadkov.

Parametre na nastavenie generovania textových rámcov (Text Regions) a obrázkov (Image)

Po analýze výskytu riadkov dochádza k ich zoskupeniu do blokov.

Základné nastavenie tvorby textových rámcov súvisí s voľbou spôsobu segmentácie:

- generovať nové textové rámce (*Generate new*) – systém vytvorí nové textové rámce pri použití automatickej segmentácie,
- ponechanie existujúcich textových rámcov (*Keep existing*) – systém ponechá textové rámce vytvorené počas manuálnej segmentácie. Okrem toho môžete obmedziť rozpoznávanie textových oblastí na typ označených štrukturálnych metadát pomocou *Restrict on structure tags* (viac v kapitole *8 Možnosti práce s textom po automatickej transkripcii*). V nastaveniach je možné navoliť viac štrukturálnych metadát súčasne. Pri vyhľadávaní systém automaticky ponúka dostupné metadáta. Tieto metadáta však musia byť na stránkach vopred označené. Tieto nastavenia je vhodné použiť v prípade, že trénujete model na automatickú segmentáciu dokumentu, s ktorým pracujete.

Zhlukovanie textových rámcov (*Text Region Method*):

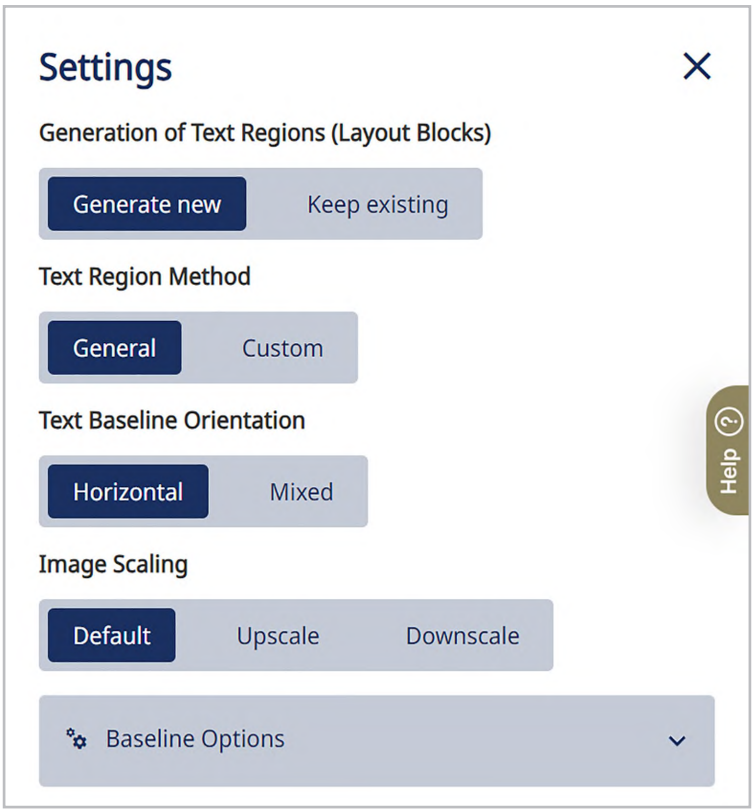
- všeobecné (*General*) – zhlukuje riadky zľava doprava. S nastavením tejto hodnoty súvisí aj nastavenie orientácie riadkov (*Text Baseline orientation*). Nastavte hodnotu *Horizontal*, ak sa v dokumente nachádzajú len horizontálne orientované riadky, alebo hodnotu *Mixed*, ak sú v dokumente aj riadky otočené o 0, 90, 180 a 270 stupňov;
- vlastné (*Custom*) – ide o jednoduché aglomeratívne zhlukovanie založené na najľavejšom bode každého riadku. Zhlukuje čiary na základe ich vzdialenosti. Môžete nastaviť, či na snímke má byť jedna oblasť textu (*One*), niekoľko (*Few*), stredne veľa (*Medium*), veľa (*Many*) alebo ich počet voliteľne prispôsobte (*Custom*).

Orientácia textu (*Text Baseline Orientation*):

- horizontálna (*Horizontal*) – ponechajte nastavenie pri prevládajúcej horizontálnej orientácii textu na snímkach,
- zmiešaná (*Mixed*) – použite v prípade, že na snímkach sa nachádza text napísaný v horizontálnom i vertikálnom smere.

Škálovanie obrázkov (*Image scaling*):

- môžete sa rozhodnúť, či chcete zvýšiť škálovanie obrázkov s nízkym rozlíšením alebo znížiť škálovanie obrázkov s vysokým rozlíšením. Túto funkciu odporúčame vyskúšať len vtedy, keď segmentácia s predvolenými nastaveniami nefunguje, napr. deteguje žiadne/málo základných čiar.



Obrázok 74 Nastavenie parametrov pokročilej segmentácie textových rámcov

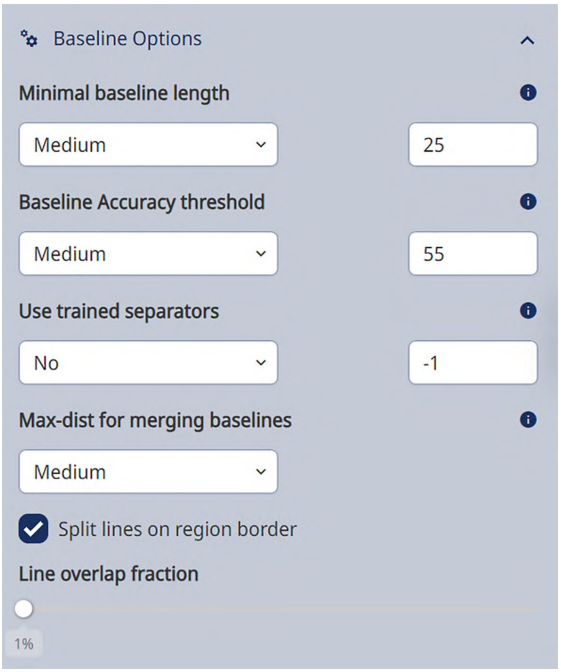
***Parametre na rozpoznanie rozloženia riadkov* (*Lines*)**

Úpravu prednastavených hodnôt odporúčame, ak pri segmentácii bolo rozpoznaných príliš málo/veľa základných čiar alebo ak boli nesprávne spojené/oddelené. Pre každý parameter môžete vybrať jednu z troch navrhovaných hodnôt – nízka (*Low*), stredná (*Medium*), vysoká (*High*) alebo si hodnotu prispôsobte (*Custom*):

- minimálna dĺžka základnej čiary (*Minimal baseline length*) sa udáva v pixeloch. Ak algoritmus v procese segmentácie deteguje základné čiary pod nastavenou dĺžkou, vynechá ich;
- prahová hodnota presnosti základnej čiary (*Baseline accuracy threshold*) – v prvej fáze rozpoznávania rozloženia sa každý pixel označí ako základná čiara, oddeľovač alebo iné. Prah presnosti základnej čiary sa pohybuje v rozmedzí od 0 do 255. Vyššie hodnoty sa prejavia vo väčšej presnosti rozpoznaných základných čiar. Pri obrázkoch s nižším rozlíšením sa pri neúspešnej detekcii základných čiar odporúča hodnoty znížiť;
- použitie vytrénovaných oddeľovačov (*Use trained separators*) – oddeľovače sú malé zvislé čiary nakreslené vedľa každej základnej čiary, označujú jej začiatok a koniec. Rozpoznávajú sa v prvej fáze analýzy rozloženia. Prahová hodnota oddeľovača sa pohybuje v rozmedzí od 0 do 255. 0 znamená, že oddeľovače sa vôbec nepoužívajú. Zvyčajne aj nižšie hodnoty zabránia spájaniu základných čiar. Použite napr. hodnotu 1, ak chcete informácie o oddeľovačoch používať niekedy (*Sometimes*) a vyššie hodnoty, ak ich chcete používať stále (*Always*);
- maximálna vzdialenosť na zlučovanie (*Max-dist for merging baselines*) – v druhej fáze sa softvér pokúša zlúčiť blízke základné čiary za predpokladu, že je ich vzdialenosť menšia ako nastavená hodnota. Použite hodnotu *Low* na zlúčenie čiar, ktoré sa na dokumente nachádzajú bližšie ako 0,5 % šírky obrazu, *Medium* na zlúčenie čiar, ktoré sú

bližšie ako 1 % šírky obrazu alebo *High* na zlúčenie čiar, ktoré sú od seba vzdialené viac ako 1 %, ale bližšie ako 5 % šírky obrazu. Vo väčšine prípadov by mala dobre fungovať voľba *Medium*;

- rozdelenie čiar na hranici textových rámcov (*Split lines on region border*) zabezpečí, že sa riadky striktne riadia hranicou zadefinovaných textových rámcov. Toto nastavenie je veľmi užitočné pri segmentácií dokumentov, v ktorých sa textové rámce nachádzajú tesne vedľa seba a text v nich sa končí takmer na hranici rámca. Zabráni sa tým spojeniu čiar do jednej dlhej čiary;
- podiel prekrývania čiar (*Line overlap fraction*) – môžete zvýšiť minimálny podiel prekrytia medzi zistenou čiarou a existujúcim textovým rámcom v prípade, že chcete, aby čiara mierne presahovala hranicu textového rámca. Ak sa čiara prekrýva s viacerými textovými rámcami, vyberie sa oblasť s najväčším prekrytím.



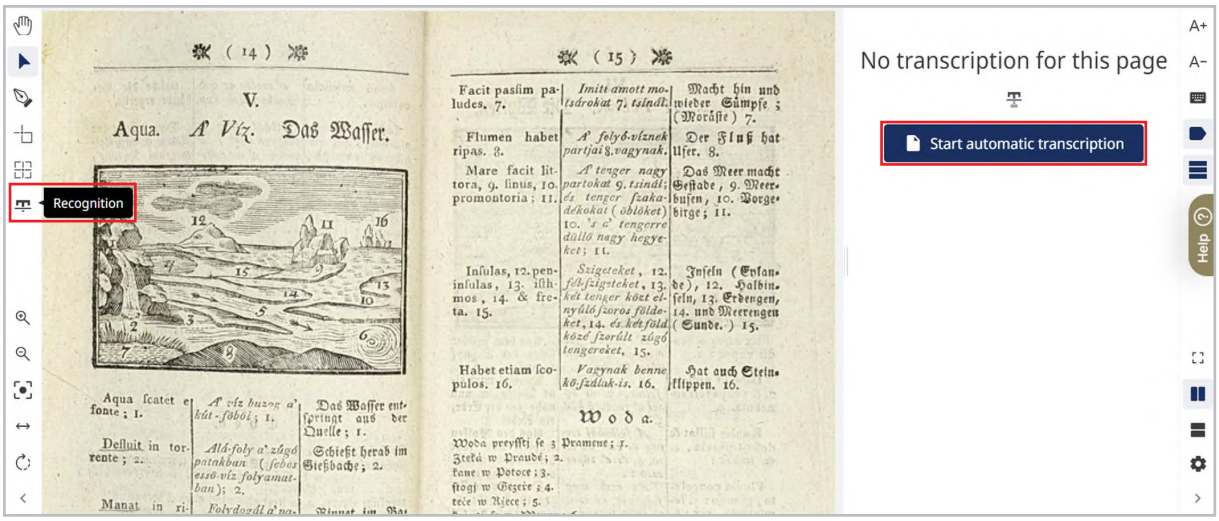
Obrázok 75 Nastavenie parametrov segmentácie

5.1.6 Automatická segmentácia a rozpoznávanie textu

Automatické rozpoznanie rozloženia (*Layout recognition*) a transkripciu dokumentu môžete vykonať v jednom kroku. Softvér automaticky na snímkach vyznačí textové rámce a riadky, určí poradie ich čítania a zároveň prepíše text identifikovaný na čiarach. Pri transkripcii dokumentu týmto spôsobom treba aplikovať niektorý z vytrénovaných textových modelov.

Textový model je algoritmus umelej inteligencie vytrénovaný na určitom počte údajov (obrázkov a prepisov), ktorý dokáže zistiť najpravdepodobnejšiu postupnosť znakov pre každý segmentovaný riadok textu. Všeobecný model pre všetky rukopisy zatiaľ neexistuje, preto musíte vybrať čo najvhodnejší model pre písmo a jazyk dokumentu, s ktorým pracujete.

V rámci programu Transkribus je k dispozícii niekoľko verejných modelov, ktoré sprístupnila komunita a tím vývojárov programu Transkribus, aj súkromné modely, ktoré vytrénovali samotní používatelia.



Obrázok 76 Prechod do nastavení automatickej segmentácie a transkripcie

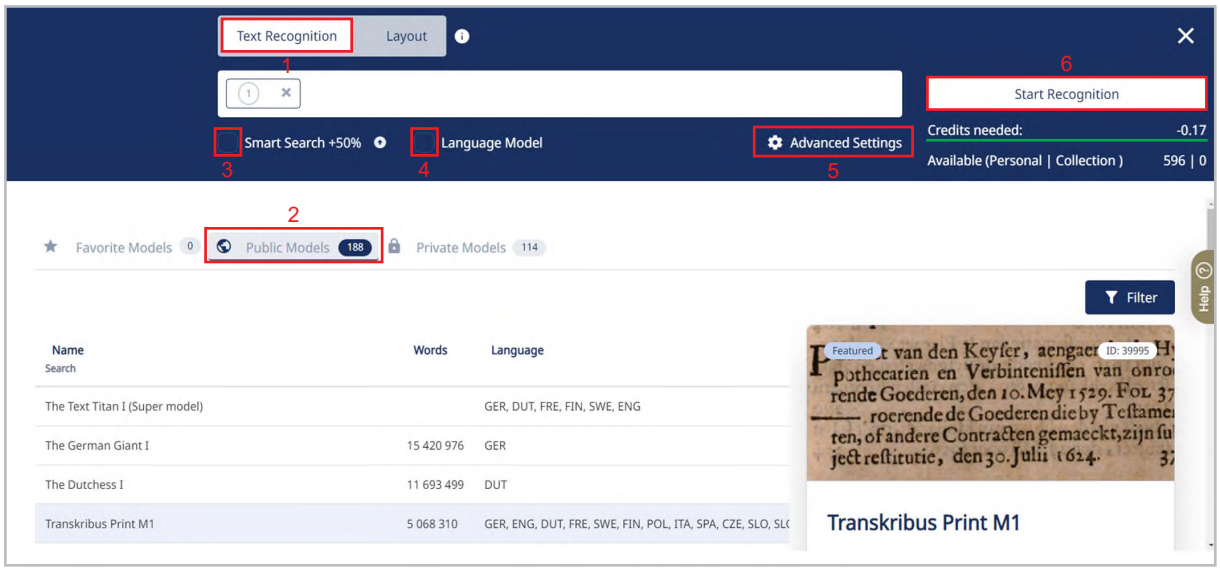
Nastavenie a spustenie automatickej segmentácie a prepisu

Pred spustením segmentácie treba:

1. vybrať snímky (dokument), na ktorých chcete segmentáciu a transkripciu vykonať (viac v kapitole *5.1.1 Výber strán*),
2. prejsť na funkciu rozpoznávania cez ikonu alebo tlačidlom *Recognize*,
3. v otvorenom okne s nástrojmi rozpoznávania v hornej časti prepnúť do sekcie *Text Recognition*,
4. vybrať najvhodnejší z dostupných modelov na transkripciu – modely môžete prehľadávať aj cez ikonu *Filter* podľa názvu, jazyka, typu dokumentu, obdobia a i.,
5. ak je to potrebné, kliknúť na rozšírené nastavenia (*Advanced Settings*) a v otvorenom dialógovom okne upraviť parametre segmentácie,
6. zakliknúť funkciu *Smart Search* (nie je dostupná pre každý model) – s jej použitím sa v procese prepisu hľadá aj niekoľko alternatív pre každé rozpoznané slovo automatickej transkripcie, v procese spracovania transkribovaného dokumentu umožňuje v prepise nájsť slová aj v prípade, že ich model rozpoznávania textu prepísal nesprávne,
7. zakliknúť funkciu *Language Model* – použitie jazykových modelov vo väčšine prípadov zvyšuje presnosť rozpoznávania.

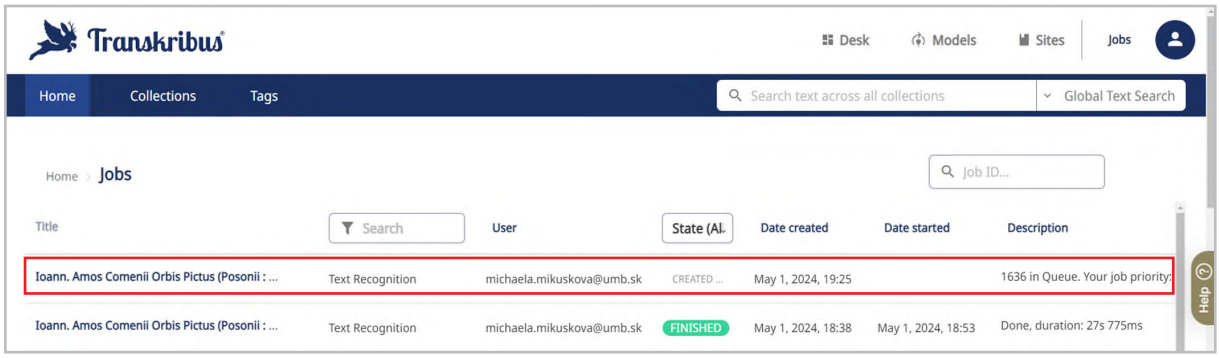
Segmentáciu spustíte kliknutím na tlačidlo *Start Recognition*.

Odporúčame vyhnúť sa používaniu veľkých modelov, ktoré vytrénovala komunita Transkribus s možnosťou jazykového modelu alebo *Smart Search*, pretože z dôvodu neštandardnej veľkosti nemusia správne fungovať. Okrem toho sa neodporúča používať takéto veľké modely ani ako základné modely (*Base Models*) pri trénovaní vlastných modelov.



Obrázok 77 Dôležité prvky nastavenia a spustenia automatickej segmentácie v dialógovom okne Text Recognition. Zelenou farbou je podčiarknutý počet kreditov, potrebných na tento úkon.

Po spustení rozpoznávania textu môžete skontrolovať stav riešenia zadanej úlohy kliknutím na záložku *Jobs*, ktorá sa nachádza v hlavnej ponuke v hornej časti aplikácie pri prihlásení. Tu vidíte prehľad úloh, ktoré ste na serveroch Transkribusu zadávali. Po ukončení procesu si zobrazte príslušnú snímku a na nej skontrolujte výsledok rozpoznávania.



Obrázok 78 Zobrazenie zoznamu úloh a stavu ich riešenia na záložke Jobs. V červenom rámčeku je označená úloha automatickej segmentácie.

Generovanie textu v dokumentoch so zložitou štruktúrou

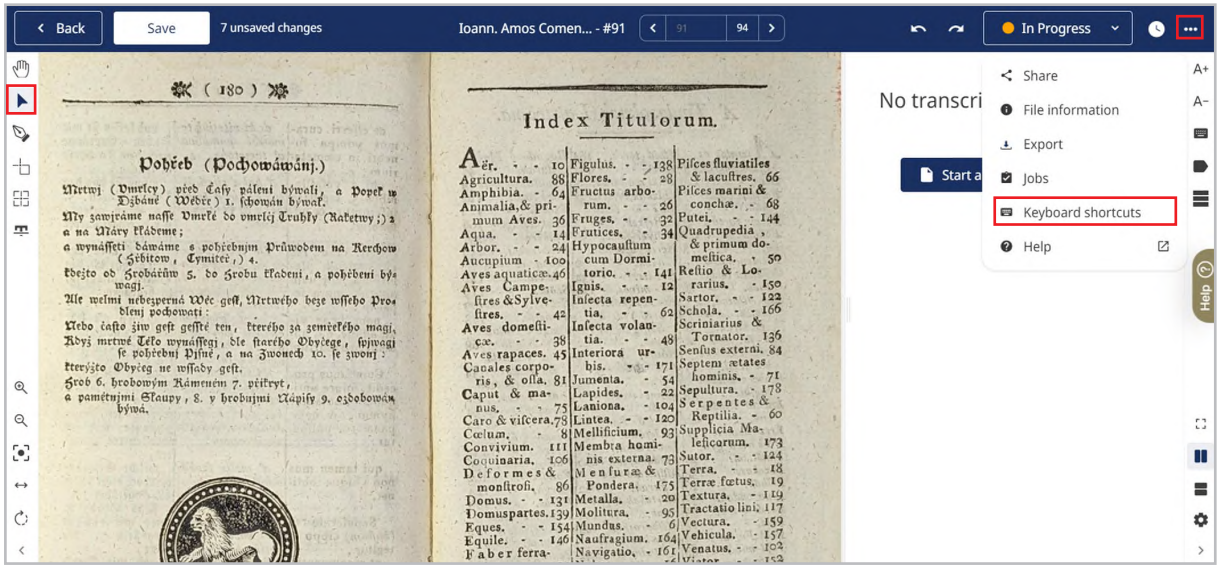
V prípade, že pracujete s dokumentom so zložitejšou štruktúrou, napr. text obsahuje tabuľky, marginálie, viacero stĺpcov, odporúčame:

1. vykonať rozpoznanie rozloženia (*Layout recognition*) objektov segmentácie – textové rámce a riadky,
2. skontrolovať poradie čítania objektov segmentácie (viac v kapitole *5.2 Opravy po automatickej a manuálnej segmentácii*),
3. spustiť funkciu rozpoznávania textu (*Text recognition*).

5.2 Opravy po automatickej a manuálnej segmentácii

Výsledky manuálnej a automatickej segmentácie nie sú vždy ideálne a vo väčšine prípadov je potrebné urobiť ďalšie opravy. Časová náročnosť úprav závisí od štruktúry dokumentu a zvoleného typu rozpoznania rozloženia textu, t. j. segmentácie.

Nástroje na korekciu segmentácie sú dostupné vo vertikálnom paneli okna editora, ktorý sa nachádza naľavo od snímky dokumentu, s ktorým pracujete. Jednotlivé funkcie sú zastúpené ikonkami. Keď prejdete kurzorom na ikonku, zobrazí sa popis funkcie. Zvolený nástroj sa deaktivuje vykonaním požadovaného úkonu alebo kliknutím na ikonku  (*Selection mode*) alebo stlačením klávesnice ESC. Klávesové skratky na úpravu segmentácie nájdete v sekcii *Keyboard shortcuts* kliknutím na tri bodky v pravom hornom rohu okna editora.



Obrázok 79 Ikonky nástrojov editora segmentácie

5.2.1 Korekcia textových rámcov (*Text Regions*)

Textové rámce, ktoré sa pri automatickej i manuálnej segmentácii vytvárajú, majú tvar štvorca alebo obdĺžnika v závislosti od textu, ktorý označujú. Mali by obklopovať celý text, ktorý sa nachádza na snímke dokumentu a má byť predmetom transkripcie. Počet a štruktúra textových rámcov závisí od štruktúry a obsahu dokumentu. Pri manuálnej segmentácii je niekedy potrebné vytvoriť špecifické typy a tvary textových rámcov. Aj pri automatickej segmentácii textových rámcov môžu nastať prípady, keď je nutné urobiť čiastočné korekcie. Editor poskytuje niekoľko nástrojov na prácu s textovými rámcami.

Prispôsobenie textového rámca

Štandardne sú hranice textového rámca na seba kolmé a definované štyrmi kontrolnými bodmi, ktoré vymedzujú vrcholy rámca. Textové rámce je možné prispôsobovať posúvaním kontrolných bodov prípadne posúvaním čiar označujúcich hranice rámcov.

Pri manuálnom vytváraní textových rámcov môžu nastať prípady, že sa rámce prekrývajú alebo text z jedného rámca čiastočne prechádza do druhého. Rámce je možné upravovať pridávaním kontrolných bodov, čím sa vytvorí polygón.

Na úpravu hraníc textových rámcov:

- kliknite na textový rámec, ktorý chcete upraviť,
- prejdite kurzorom na zelenú čiaru označujúcu hranice textového rámca a na požadovanom mieste kliknutím pridajte ďalšie kontrolné body,
- textový rámec pomocou pridaných bodov upravte na požadovaný tvar.

Rozdelenie textového rámca

Textové rámce niekedy treba rozdeliť, lebo text, ktorý rámec označuje, spolu nesúvisí, napr. hlavný text a marginálne poznámky.

Pre rozdelenie jedného textového rámca na dva rámce označte kurzorom. Podľa toho, ako potrebujete rámec rozdeliť, stlačte na klávesnici príslušné písmeno:

- horizontálne rozdelenie – kláves H (*Split element horizontally*),
- vertikálne rozdelenie – kláves V (*Split element vertically*),
- prispôsobiteľné rozdelenie – kláves C (*Custom split*).

Všetky tri možnosti s označením *Vertical split, Horizontal split* a *Custom split* sa zobrazia aj v prípade, že na oblasť textového rámca kliknete pravým tlačidlom myši.

Stlačením klávesu na vykonanie požadovanej funkcie alebo vybraním požadovaného úkonu z ponuky sa v editore snímky zobrazí modrá čiara. V označenom textovom rámci preneste túto čiaru kurzorom na miesto, kde ho potrebujete rozdeliť a kliknite.

Spojenie textových rámcov

Automatickou segmentáciou môžu vzniknúť dva textové rámce, ktoré treba spojiť do jedného. Na spojenie viacerých rámcov:

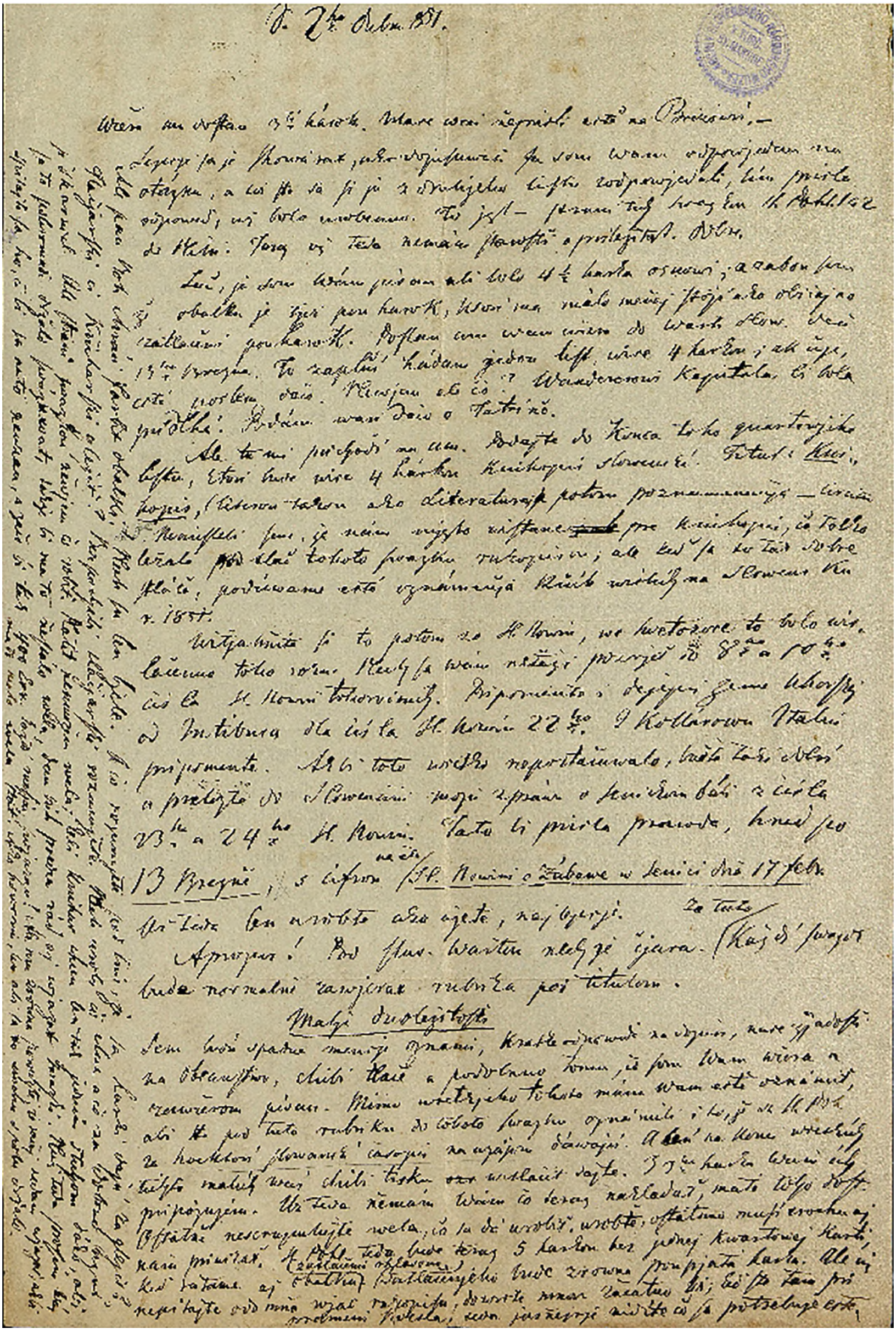
- na klávesnici stlačte kláves CTRL a kurzorom kliknite na rámce, ktoré chcete spojiť,
- na klávesnici stlačte kláves M (*Merge two elements*).

Odstránenie textového rámca

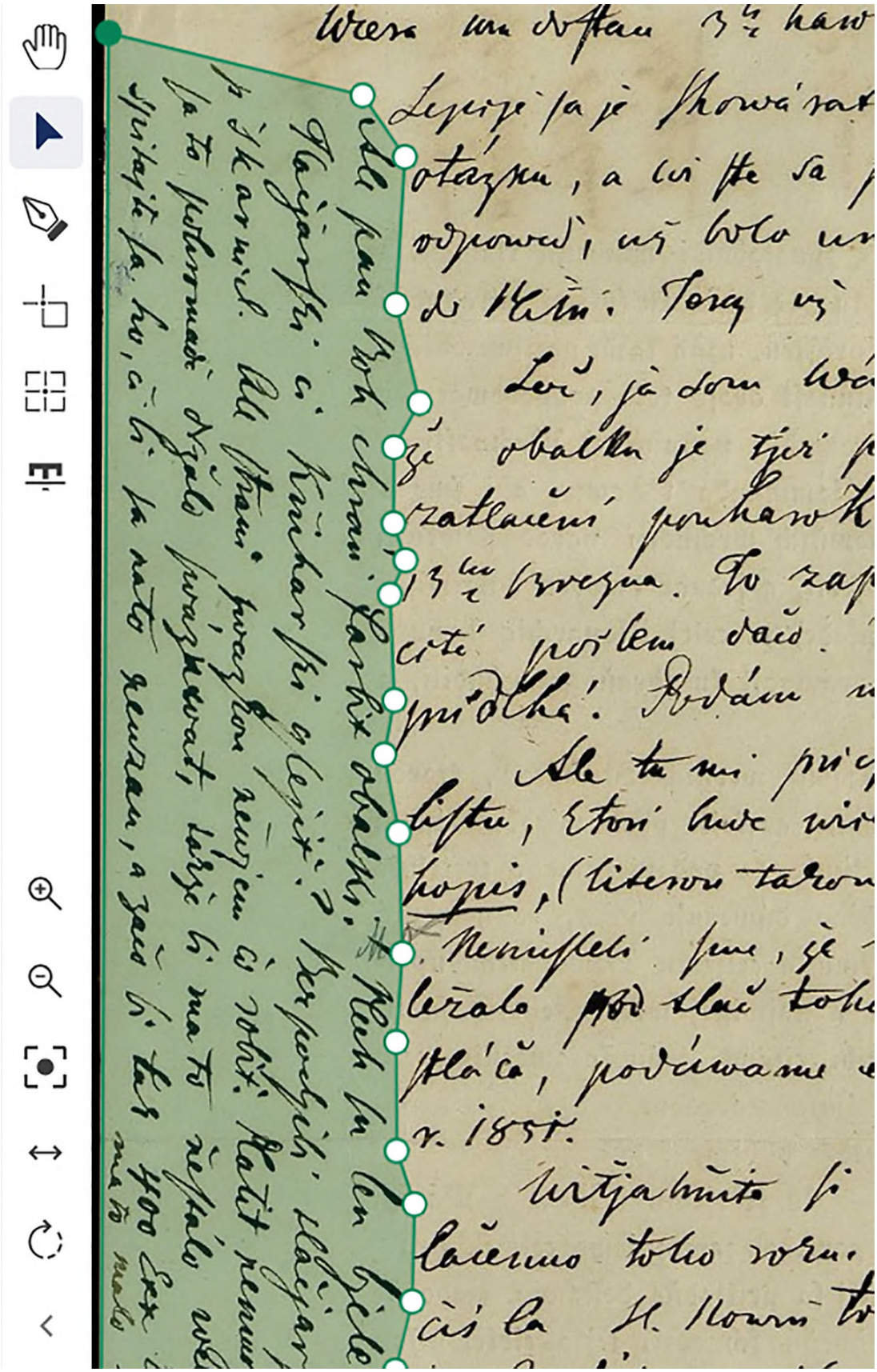
Pri automatickej segmentácii môže vzniknúť nežiaduci textový rámec na mieste, kde sa nachádzajú rôzne šmuhy, text presvitajúci z inej strany a pod. Vyskytnúť sa môžu aj prípady, že v jednom textovom rámci vzniknú dva rámce. Tieto treba odstrániť, aby nenarúšali štruktúru dokumentu, prípadne neoznačovali nežiaduce riadky, ktoré by mohli znižovať kvalitu vytrénovaného modelu. Na odstránenie rámca:

- kurzorom označte rámec, ktorý chcete vymazať,
- na klávesnici stlačte klávesu DELETE/DEL (*Delete selected element*).

Ak odstraňujete rámec, v ktorom sú označené riadky, odstránia sa aj tie.



Obrázok 80 Kombinované horizontálno-vertikálne rozloženie textu, ktoré si vyžaduje vytvorenie špecifických textových rámcov.



Obrázok 81 Detail manuálnej úpravy textových rámcov (polygónov) pridávaním nových kontrolných bodov a ich presúvaním na čiare vytvoreného textového rámca. Vyžaduje tvorbu dvoch samostatných rámcov a samostatné vykreslenie komplikovaného rozdelenia rámcov.

5.2.2 Korekcie riadkov (*Lines*)

S chybami sa stretnete aj pri automatickej segmentácii riadkov. Riadky vymedzujú oblasti čiary a sú základným referenčným bodom na rozpoznávanie textu. Ich úprave je preto potrebné venovať zvýšenú pozornosť.

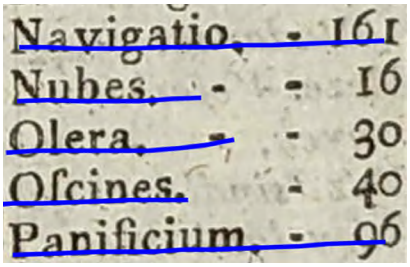
Čiary popisujú polyčiaru tiahnucu sa pozdĺž spodnej časti riadku písaného alebo tlačeného textu. Na snímke dokumentu sú zobrazené tenkou čiarou modrej farby. Po kliknutí na čiaru príslušného riadku sa farba zmení na oranžovú. V tomto zobrazení vidieť, že čiara spája body, ktorých počet závisí od dĺžky textu nachádzajúceho sa v príslušnom riadku.



Obrázok 82 Spôsob označenia základnej čiary

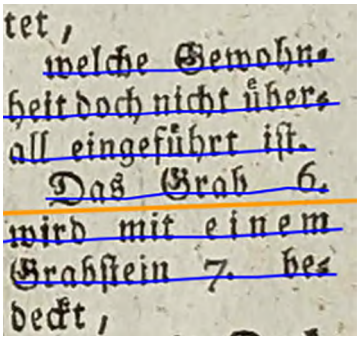
Najčastejšie sa môžete pri analýze textu stretnúť s týmito chybami:

- čiara nekopíruje celý text v príslušnom riadku,



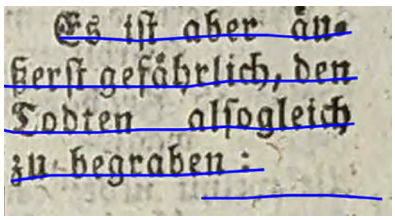
Obrázok 83 Nedotiahnutá čiara

- čiara sa nevytvorí tam, kde sa nachádza text dokumentu,



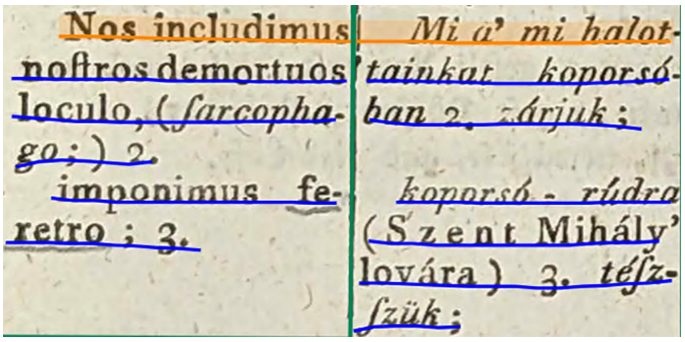
Obrázok 84 Nevytvorená čiara

- čiara sa vytvorí tam, kde sa text dokumentu nenachádza (napr. šmuha na papieri, text presvitajúci z druhej strany listu a pod.),



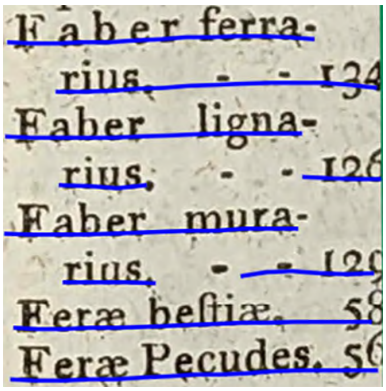
Obrázok 85 Čiara vytvorená na mieste, kde sa nevyskytuje text

- vytvorí sa jedna čiara cez susediace textové rámce,



Obrázok 86 Čiara prechádzajúca cez dva textové rámce

- vytvorí sa viac základných čiar namiesto jednej.



Obrázok 87 Prerušovaná čiara v jednom riadku

Podobne ako na korekciu textových rámcov má editor dostupné nástroje aj na korekciu čiar.

Úprava označenia riadku

Keď na farebné označenie riadku kliknete, zistíte, že čiaru tvorí niekoľko pospájaných kontrolných bodov. Začiatok a koniec čiary nemusí presne zodpovedať textu. Prax ukázala, že nie je nevyhnutné začiatok a koniec označenia doťahovať. Dôležité je, aby čiara správne kopírovala spodok riadku a písmená na nej sedeli. Niekedy je však potrebné čiaru upraviť, prípadne predĺžiť. Môžete tak urobiť dvomi spôsobmi:

1. natiahnutím okrajov čiary v požadovanom smere:
   - kliknite na posledný bod základnej čiary,
   - posuňte ho do požadovanej strany.
2. pridaním nových bodov na okrajoch čiary:
   - na klávesnici stlačte kláves A,
   - kurzorom vyberte miesto, na ktorý chcete kontrolný bod pridať a čiaru predĺžte.

Pridanie čiary

Ak sa pri manuálnej segmentácii nevytvoril riadok tam, kde sa nachádza text:

- v editore vyberte ikonku alebo na klávesnici stlačte kláves B,

- postupným klikaním kurzorom na spodnej línii písmen pridajte kontrolné body po celej dĺžke riadku textu,
- tvorbu čiary ukončíte dvojitým kliknutím alebo stlačením klávesu ENTER v poslednom bode.

Čiaru odporúčame označovať viacerými klikmi pozdĺž celého riadku tak, aby kopírovala písmená aj v prípade, že riadok nie je napísaný rovno.

Odstránenie čiary

Na odstránenie prebytočného riadku:

- kurzorom označte čiaru, ktorú chcete odstrániť,
- na klávesnici stlačte kláves DELETE/DEL (*Delete selected element*).

Rozdelenie čiary

Ak potrebujete rozdeliť riadok, ktorý prechádza do viacerých textových rámcov:

- kurzorom označte čiaru, ktorú chcete rozdeliť,
- na klávesnici stlačte kláves H (*Split element horizontally*),
- kurzorom kliknite na miesto, kde treba čiaru rozdeliť.

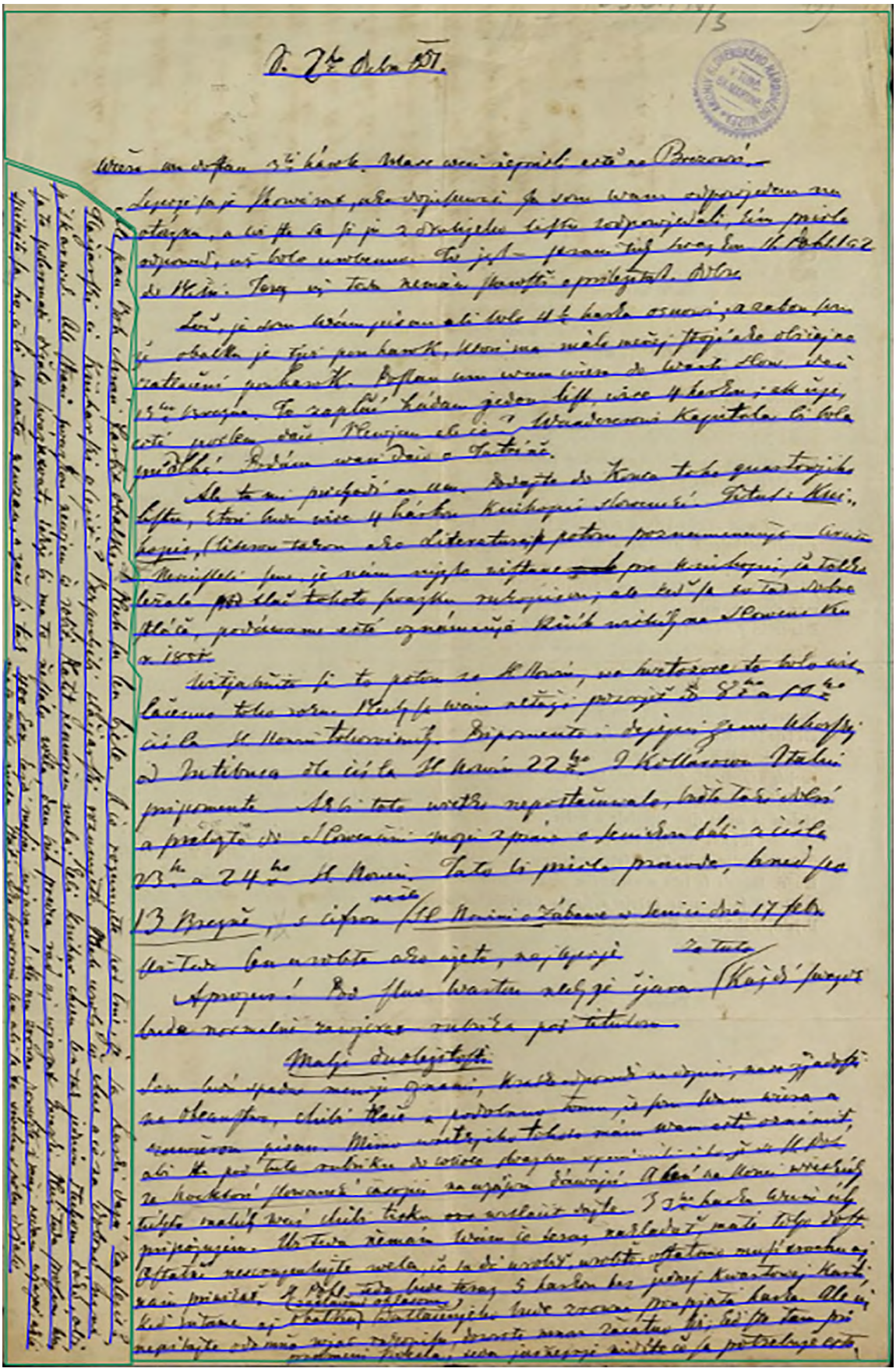
Spojenie čiar

Algoritmus niekedy nerozpozná štruktúru riadku a namiesto jedného riadku vytvorí dva, resp. aj viac. Pre spojenie čiar:

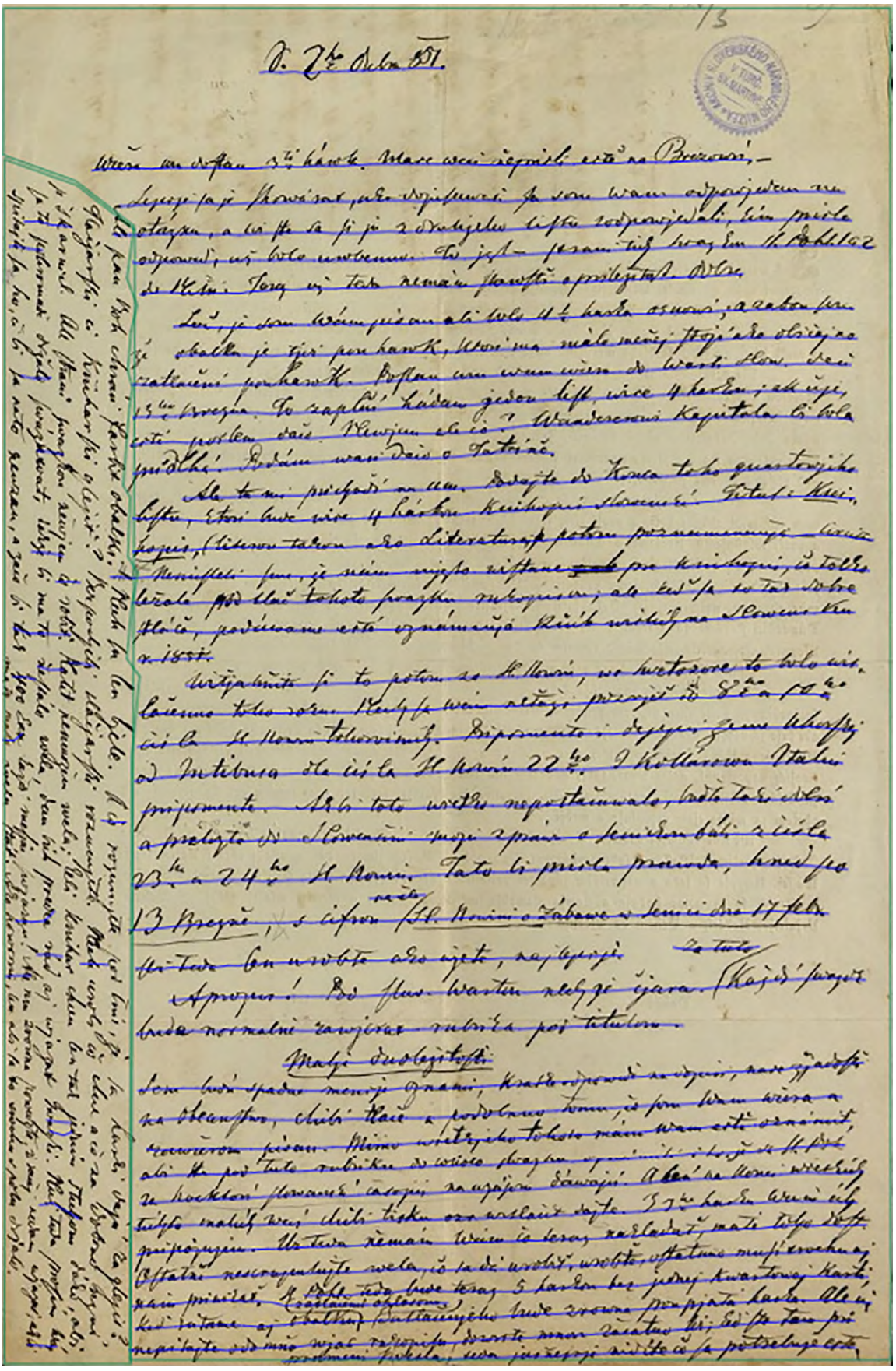
- na klávesnici stlačte kláves CTRL a kurzorom kliknite na čiary, ktoré chcete spojiť,
- na klávesnici stlačte kláves M (*Merge two elements*).

Čiary je možné zadefinovať aj vertikálne a kombinovať rôzne smery čiar na jednej stránke dokumentu (napr. pri pohľadniciach alebo ako uvádza príklad nižšie).

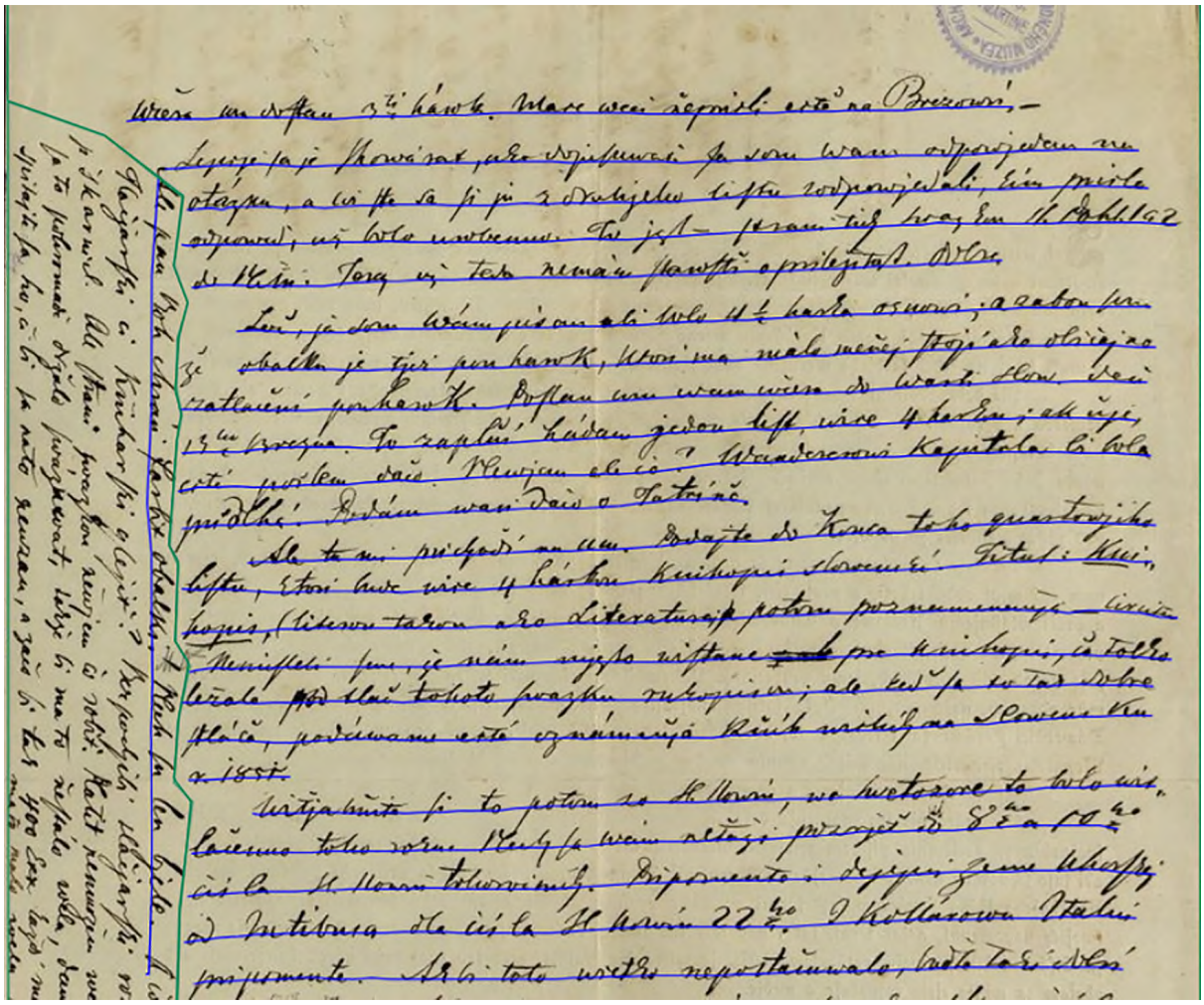
Segmentáciu dokumentu, v ktorom sa nachádza text napísaný rôznymi smermi, môžete vykonať použitím vhodného modelu na rozpoznanie rozloženia (*Mixed Line Orientation*). Ak je takto napísaných riadkov v dokumente málo, môžete ich dať nasegmentovať aj samostatne použitím vhodného modelu na rozpoznanie rozloženia (*Horizontal Line Orientation*) alebo ich označiť manuálne.



Obrázok 88 Výsledok segmentácie použitím modelu rozloženia Horizontal Line Orientation v samostatnom kroku rozpoznania rozloženia



Obrázok 89 Neuspokojivé výsledky automatickej segmentácie horizontálno-vertikálneho členenia textu



Obrázok 90 Manuálne doplňanie riadkov pri horizontálno-vertikálnom členení textu dokumentu

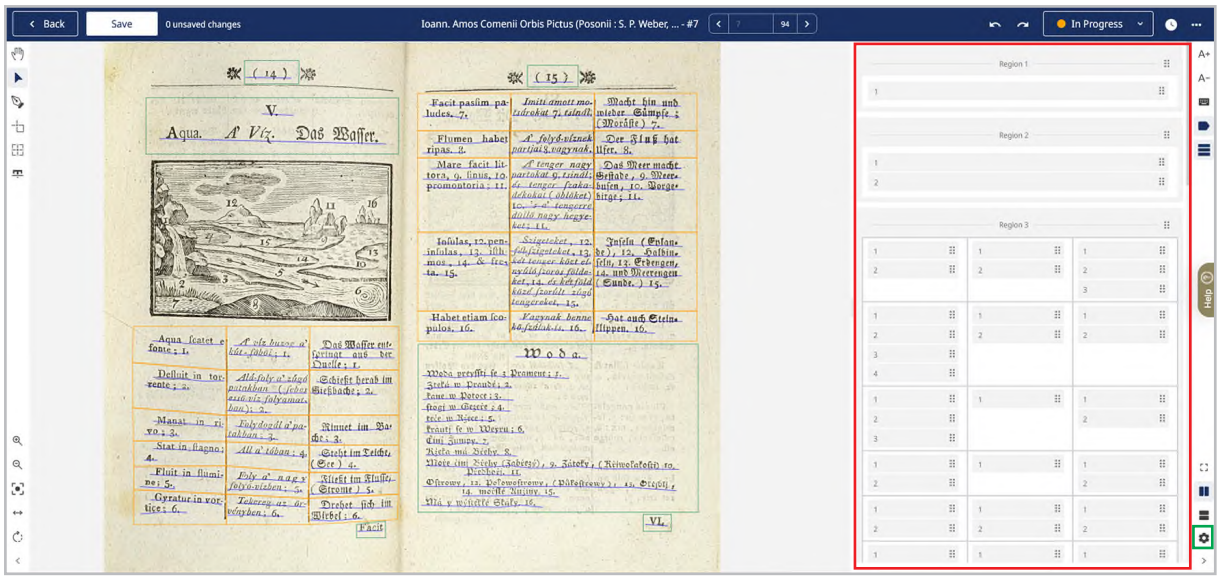
5.2.3 Kontrola a úprava poradia čítania textových rámcov a riadkov

Mnohé dokumenty obsahujú nielen hlavný text, ale aj poznámky pod čiarou, marginálie, ktoré pridali iní používatelia dokumentu, prípadne je obsah dokumentu veľmi štruktúrovaný, napr. je zapísaný v stĺpcoch, obsahuje tabuľky a pod. Algoritmus pri rozpoznaní rozloženia usporadúva textové rámce a čiary podľa ich grafického výskytu a automaticky ich čísluje podľa súradníc na snímke dokumentu, pričom postupuje od ľavého horného rohu smerom nadol.

Pre trénovanie modelu nie je dôležité striktné poradie čítania textových rámcov a riadkov v nich. Toto poradie je však dôležité, ak chcete s textom následne pracovať, zverejniť ho pre iných používateľov alebo ho vydať v tlačenej podobe. Na to, aby bol text s náročným rozložením pre čitateľa usporiadaný zrozumiteľne, má aplikácia Transkribus k dispozícii nástroje, vďaka ktorým môžete zmeniť poradie čítania textových rámcov a riadkov a usporiadať ich do logického sledu.

Nástroje na úpravu poradia čítania textových rámcov a riadkov:

1. nástroj *Visibility* na zobrazenie číslovania textových rámcov a čiar,
2. štruktúra *Layout.*



Obrázok 91 Umiestnenie nástrojov na úpravu textových rámcov a čiar. Nástroj Layout, na obrázku označený červeným obrysom, sa zobrazí po vyznačení textových rámcov. Nástroj Visibility sa zobrazí po kliknutí na ozubené koliesko nastavení (Settings), na obrázku označené zeleným obrysom.

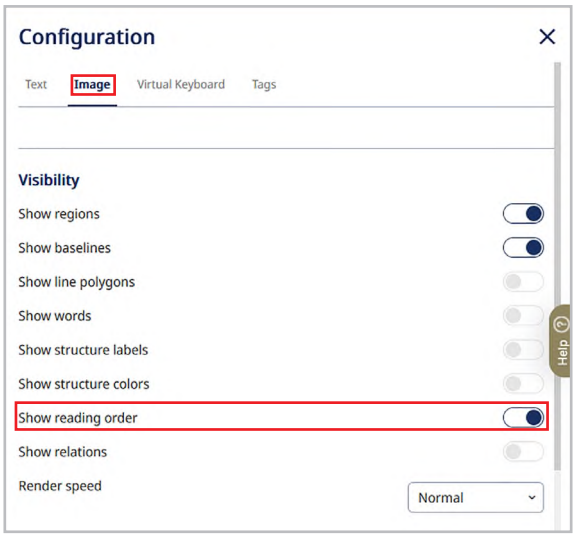
Pri dokumentoch so zložitým usporiadaním textu, kde sa poradie riadkov neriadi bežnými pravidlami, a pri dokumentoch, v ktorých ste vykonali viaceré manuálne opravy segmentácie, je možné oba nástroje kombinovať.

5.2.3.1 *Visibility* (zobrazenie objektov segmentácie)

Táto funkcia slúži na zobrazenie poradia čítania objektov segmentácie.

Po kliknutí na ikonku nastavení ⚙ (*Settings*) sa otvorí okno *Configuration*, ktoré v záložke *Image* ponúka možnosti na zobrazenie jednotlivých objektov segmentácie:

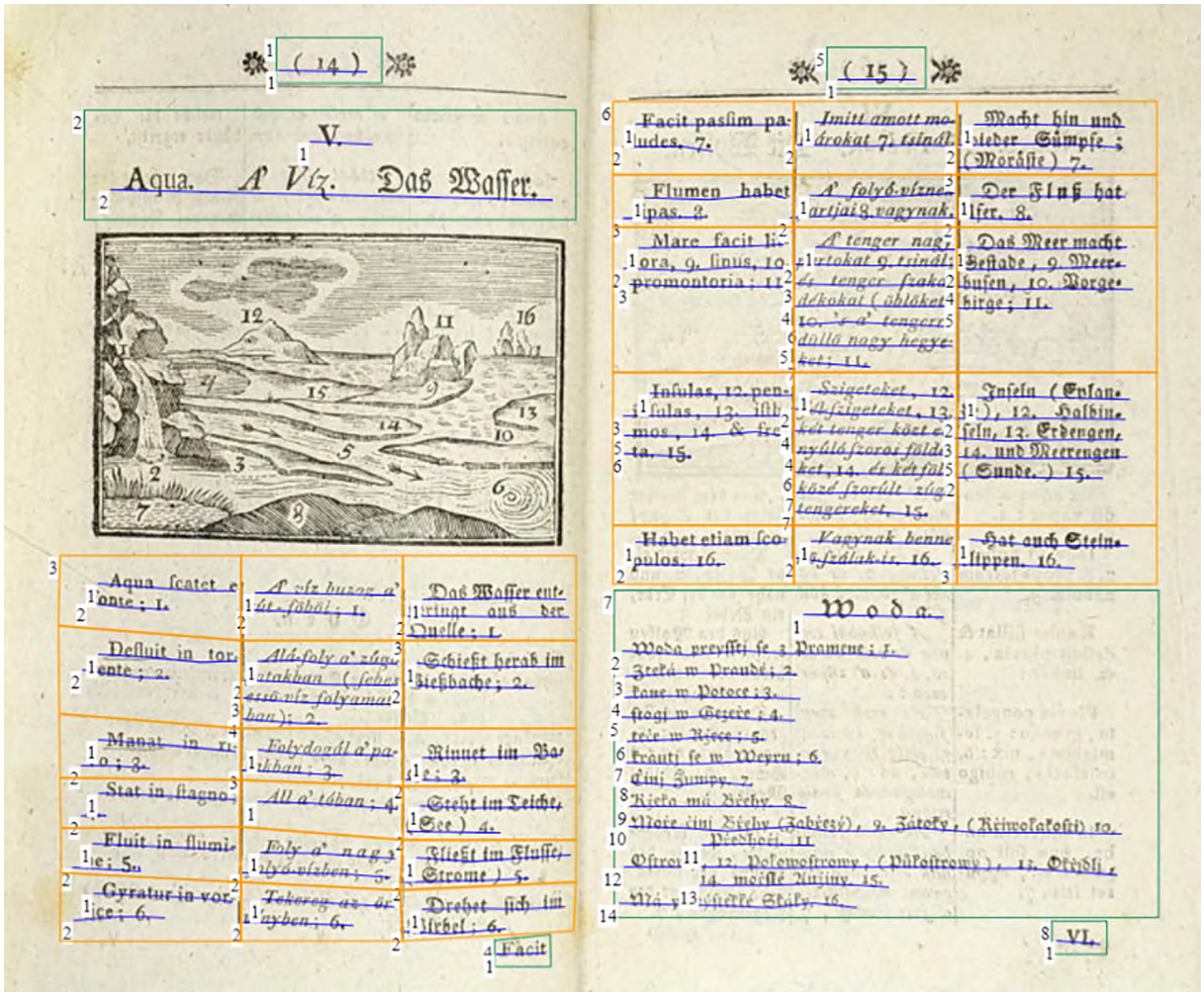
- textových rámcov (*Show regions*),
- čiar (*Show lines*),
- polygónov čiar (*Show lines polygons*),
- slov (*Show words*),
- textového označenia štrukturálnych tagov (*Show structure labels*),
- farebného rozlíšenia štrukturálnych tagov (*Show structure colors*),
- číselného označenia poradia čítania textových rámcov a riadkov (*Show reading order*),
- vzťahov (*Show relations*) a i.



Obrázok 92 Ponuka zobrazenia objektov segmentácie v konfiguračných nastaveniach obrázka

Automaticky býva nastavené zobrazenie textových rámcov a riadkov. Ostatné možnosti si vyberáte podľa toho, ktorý objekt potrebujete zobraziť.

Každý textový rámec má svoje vlastné číslovanie čiar, t. j. prvá čiara textového rámca má byť označená číslicou jeden.



Obrázok 93 Zobrazenie správneho poradia čítania textových rámcov a riadkov štruktúrovaného textu

5.2.3.2 *Layout*

Táto funkcia slúži na zobrazenie poradia čítania objektov segmentácie, ale ako textový editor, do ktorého sa prepisuje text zodpovedajúci príslušnému textovému rámcu a čiare v procese transkripcie.

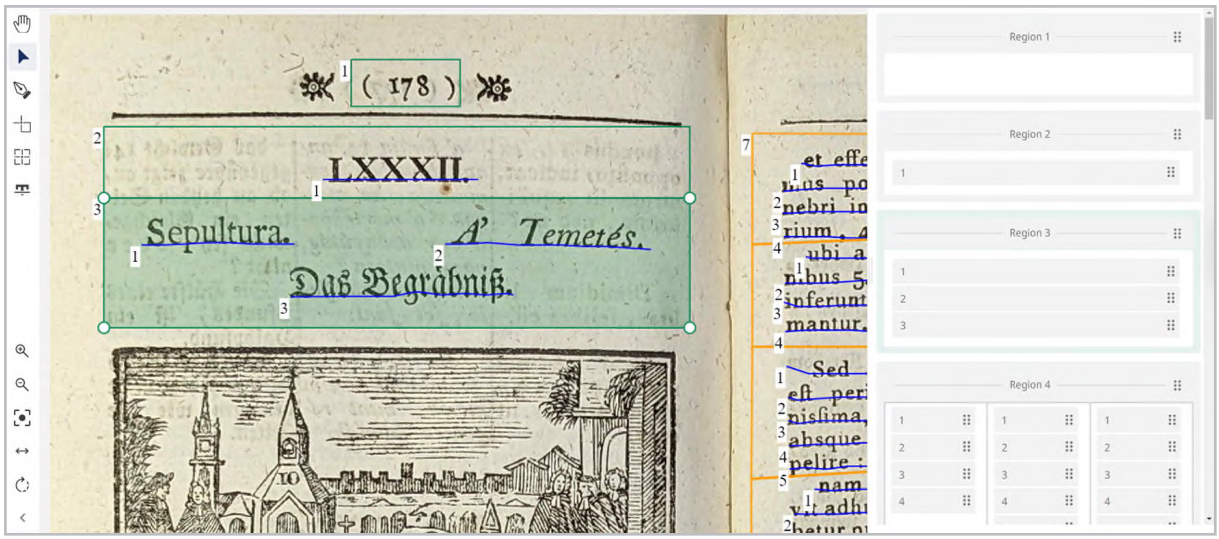
Štruktúra *Layout* sa zobrazí automaticky po vytvorení prvých textových rámcov na snímke dokumentu. Zobrazuje čísla textových rámcov (*Regions*) a čísla čiar prislúchajúcich danému riadku bez ohľadu na to, či je zapnuté zobrazenie číslovania čiar na snímke cez funkciu *Visibility*.



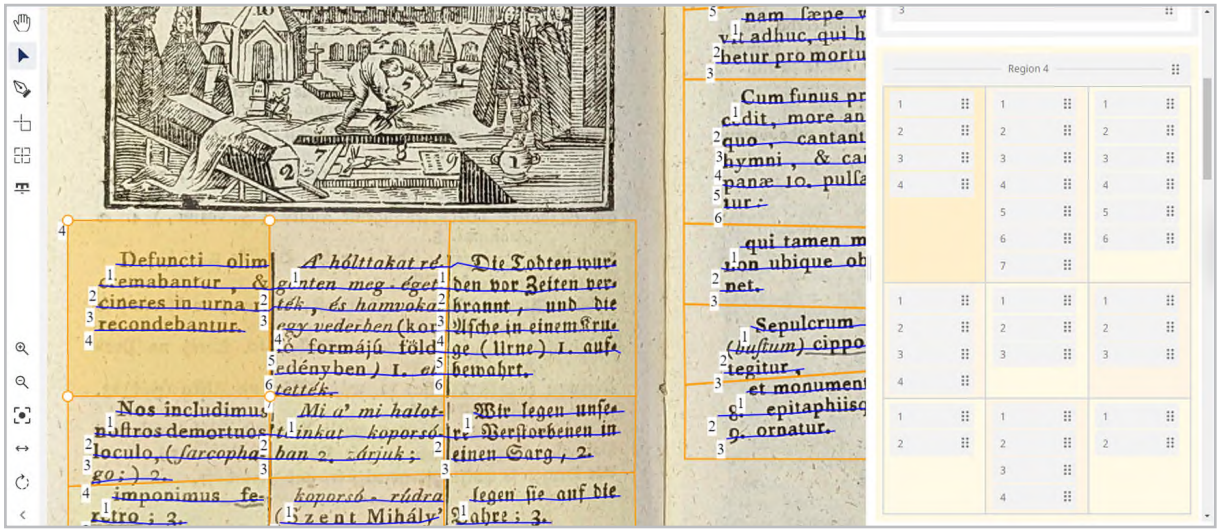
Obrázok 94 Náhľad na štruktúru objektov v zobrazení Layout

Kontrola a úprava poradia čítania objektov segmentácie v zobrazení *Layout*:

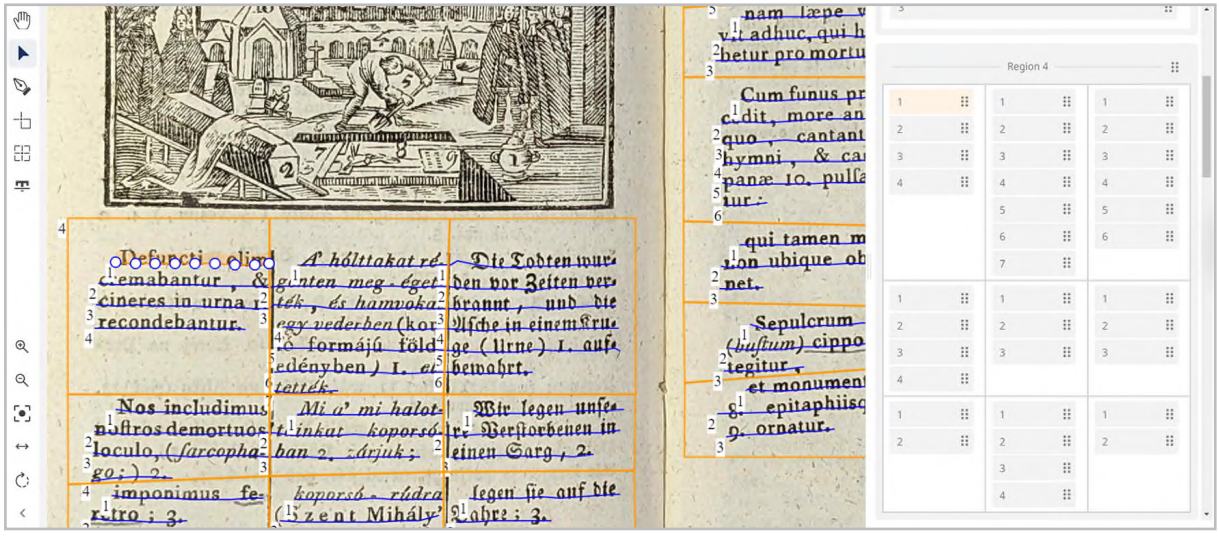
- cez funkciu *Visibility* si zapnite funkciu *Show reading order*, t. j. zobrazovanie poradia čítania objektov segmentácie (textových rámcov a riadkov),
- kliknite na objekt, ktorý je zoradený nesprávne,
- objekt sa zvýrazní na snímke dokumentu aj v *Layoute* (farebné podsvietenie zodpovedajúce typu objektu),
- zmeňte poradie objektu potiahnutím na požadované miesto (podobne ako presúvate označený text v programe Word).



Obrázok 95 Príklad zobrazenia textového rámca v Layoute



Obrázok 96 Príklad zobrazenia bunky tabuľky v Layoute



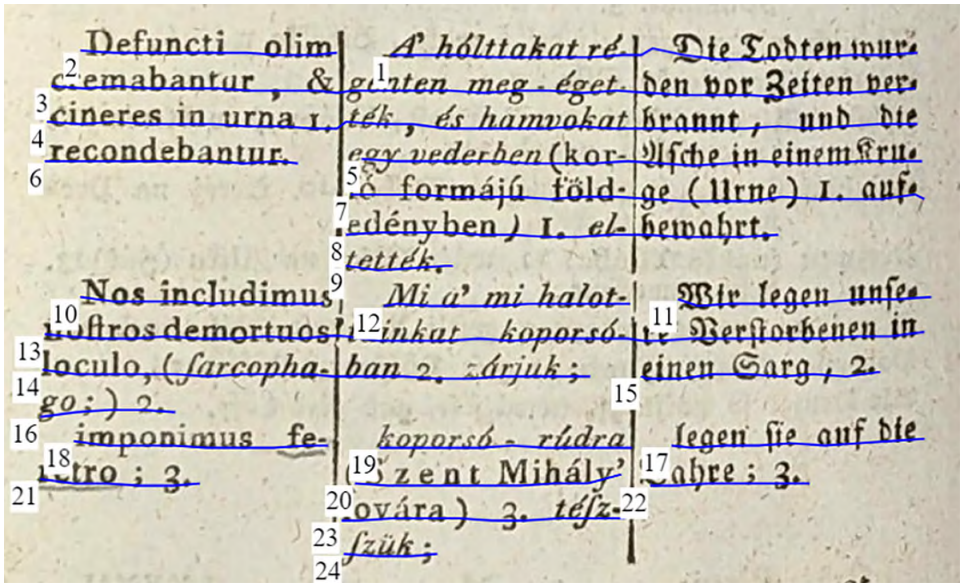
Obrázok 97 Príklad zobrazenia riadku v Layout

5.2.3.3 Práca so stĺpcami

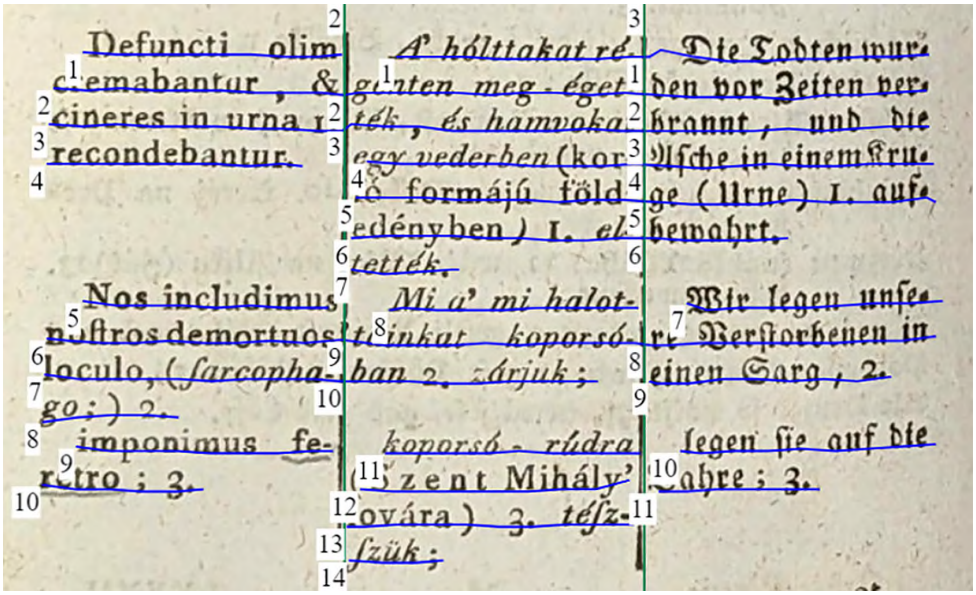
Dodatočne možno upraviť aj poradie čítania objektov, ktoré majú iné usporiadanie, napr. stĺpce. Program automaticky priraďuje poradie čítania na základe horizontálneho usporiadania riadkov na stránke namiesto toho, aby riadky zoradil podľa stĺpcov. Čiastočne tento problém odstránite nasledovne:

- stlačte kláves V (*Split element vertically*) a rozdeľte textový rámec podľa usporiadania stĺpcov na snímke,
- keď je každý stĺpec vyčlenený v samostatnom textovom rámci, poradie čítania riadkov sa automaticky aktualizuje.

Na príkladoch nižšie vidieť, že vertikálnym rozdelením stĺpcov došlo aj k rozdeleniu čiar, ktoré prechádzali cez viacero stĺpcov (napr. riadky č. 3, 4, 13 a 14 v prvom stĺpci). Tento krok vo väčšine prípadov vyžaduje následnú kontrolu a korekciu poradia čítania riadkov.



Obrázok 98 Poradie čítania riadkov v stĺpcoch pred vertikálnym rozdelením textového rámca

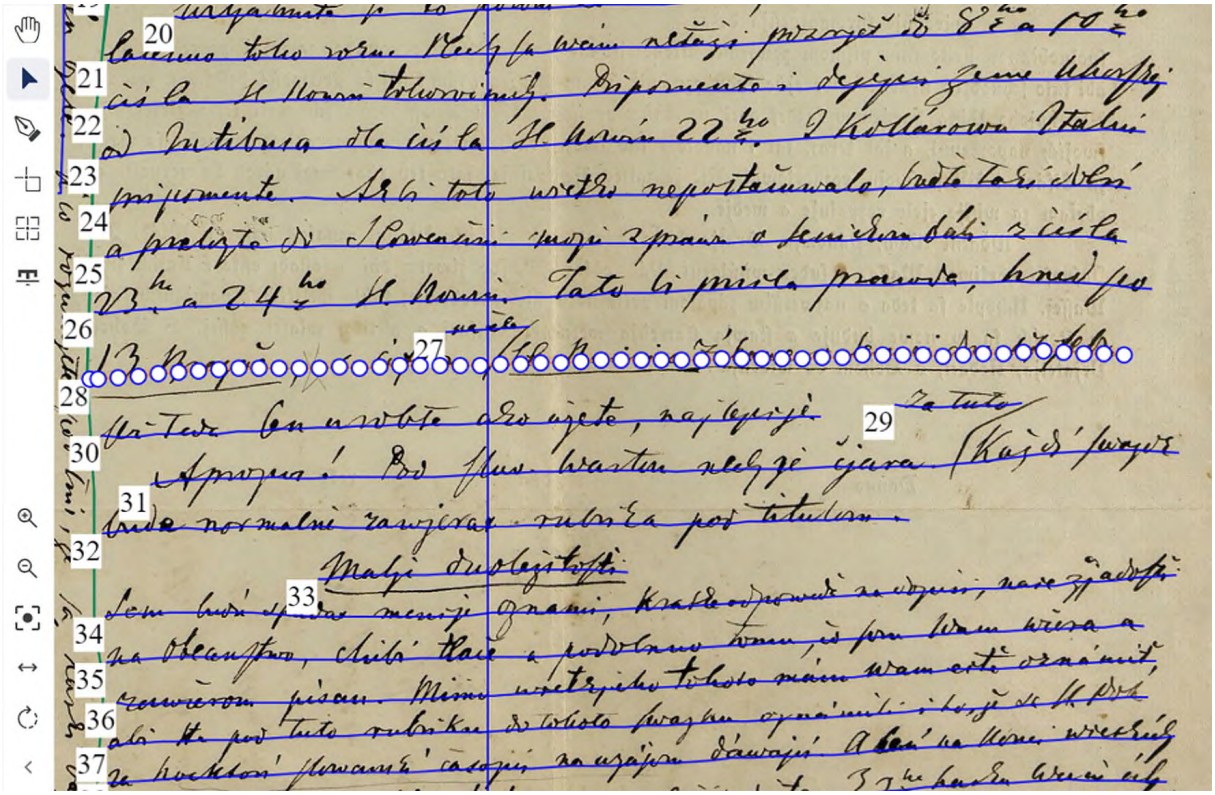


Obrázok 99 Poradie čítania riadkov v stĺpcoch po vertikálnom rozdelení

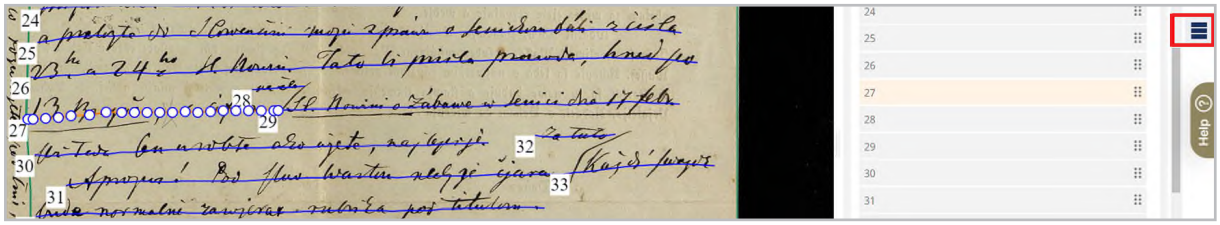
5.2.3.4 Vkladanie medziriadkov

Pri rukopisných textoch sa môžete stretnúť s vloženým textom (vsuvkou), ktorou autor do pôvodného textu vkladá nový obsah. Vložený text vytvára medziriadok, ktorý treba správne včleniť do štruktúry a obsahu dokumentu tak, aby text logicky nasledoval. Na vygenerovanie správneho poradia čítania treba urobiť manuálne úpravy:

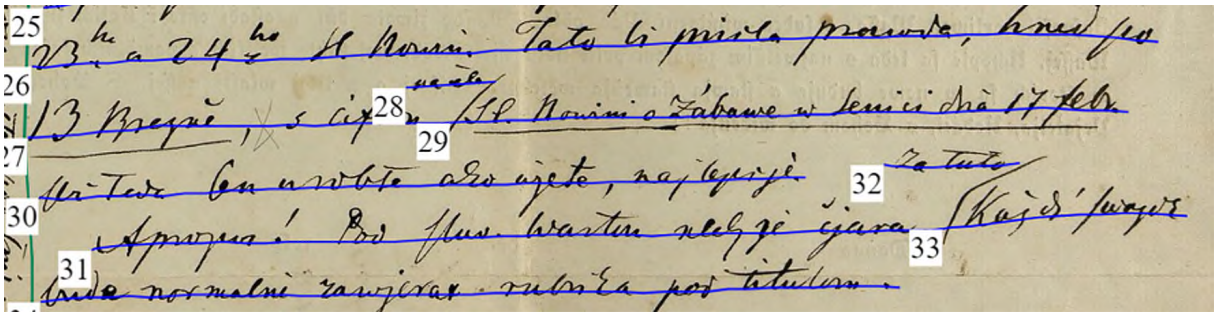
- cez funkciu *Visibility* si nastavte zobrazenie číslovania riadkov (*Show lines reading order*),
- kliknutím označte riadok nachádzajúci sa pod vloženým textom,
- na klávesnici stlačte kláves V (*Split element vertically*) a rozdeľte čiaru na mieste, kde vložený text obsahovo patrí,
- opravte číslovanie poradia čiar.



Obrázok 100 Rozdelenie riadku, do ktorého treba vložiť vsunutý text

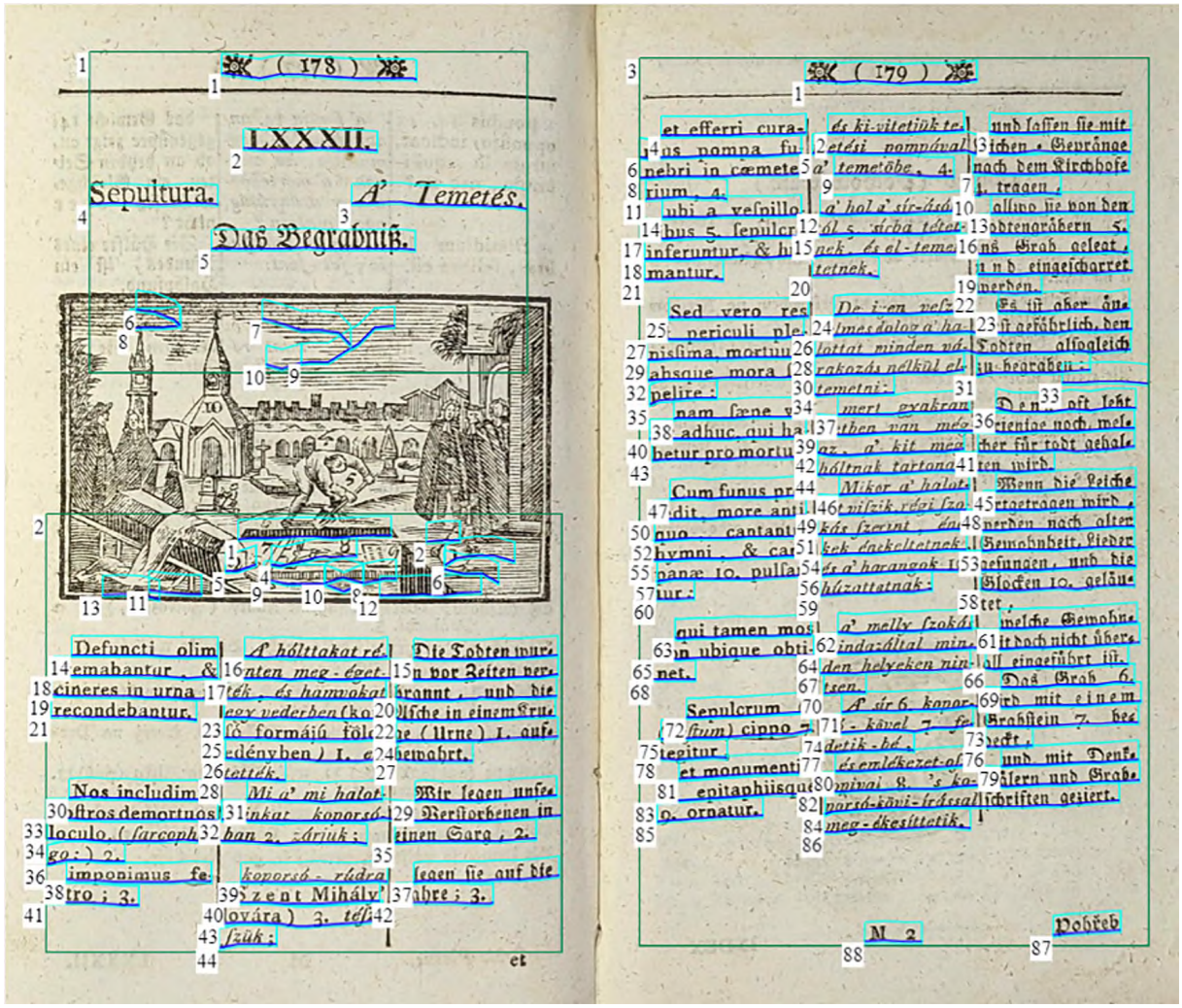


Obrázok 101 Úprava poradia číslovania: riadok 28 bude prečíslovaný na 27, riadok 27 bude mať číslo 28, poradie čítania riadku 29 je správne



Obrázok 102 Správne poradie číslovania riadkov s vloženým textom po manuálnych úpravách. Na tomto príklade vidieť dva vložené texty s upraveným poradím čítania (riadky 28 a 32)

Na obrázkoch nižšie uvádzame príklady úpravy stránky po automatickej a manuálnej segmentácii štruktúrovaného textu.



Obrázok 103 Neuspokojivé výsledky automatickej segmentácie textových rámcov a čiar

Algoritmus na obrázku 103 automaticky identifikoval tri textové rámce, pričom do dvoch zahrnul aj ilustrácie. Pre ľahšiu identifikáciu riadkov by bolo vhodnejšie oddeliť text napísaný v stĺpcoch do samostatných textových rámcov. To môžete urobiť tromi spôsobmi:

1. vytvoríte samostatné textové rámce pre každý stĺpec zvlášť,
2. použijete nástroj na prácu s tabuľkami (viac v kapitole *5.3 Segmentácia tabuliek*),

3. dodatočne rozdelíte stĺpce použitím klávesu V (*Split elements vertically*) (viac v kapitole *5.2.3.3 Práca so stĺpcami*).

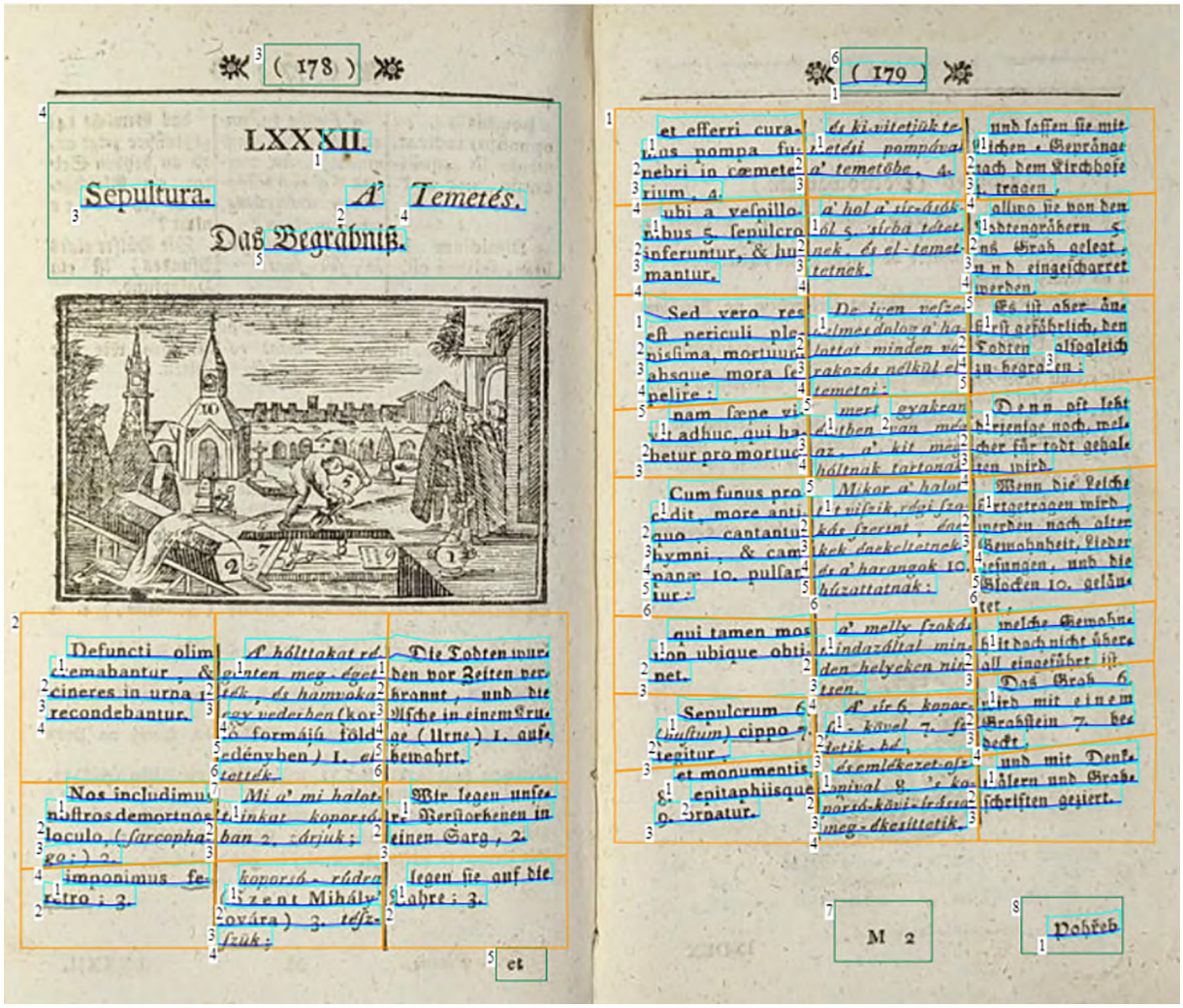
Tým, že neboli správne segmentované textové rámce, nie je správne ani poradie čítania riadkov. Navyše algoritmus detegoval aj riadky v ilustrácii, ktorá zasahuje do dvoch textových rámcov.

Napriek tomu, že text na obrázku 103 nachádzajúci sa v stĺpcoch nebol pri segmentácii rozdelený do samostatných textových rámcov, softvér automaticky segmentoval text do stĺpcov. Viditeľných je niekoľko chýb, kde text, ktorý sa nachádza v rôznych stĺpcoch, je nasegmentovaný ako jeden riadok, napr. do dvoch stĺpcov prechádzajú riadok 19 na strane 178, riadky 18, 26, 28 na strane 179. Text na snímke je mierne naklonený (smeruje zdola nahor), čo má za následok nesprávne poradie čítania riadkov, pretože riadok nachádzajúci sa vyššie má z hľadiska nastavenia algoritmu vyššiu prioritu. Preto sú na strane 179 takmer všetky riadky číslované sprava doľava.

Manuálna korekcia poradia čítania riadkov takto segmentovaného dokumentu by bola časovo náročná (minimálne 10 – 15 min. na jednu snímku). Vyžiadala by si:

- vymazanie nesprávne identifikovaných čiar,
- doplnenie chýbajúcich čiar,
- úpravu nesprávne vymedzených čiar, napr. riadok 9 v strednom stĺpci na strane 179,
- rozdelenie spojených čiar, ktoré sa majú nachádzať v rôznych textových rámcoch (stĺpcoch),
- rozdelenie textových rámcov do stĺpcov s použitím funkcie *Split element vertically* (kláves V), prípadne aj oddelenie ostatných častí textu, napr. paginácia, kustódy s použitím funkcie *Split element horizontally* (kláves H),
- usporiadanie textových rámcov do správneho poradia,
- kontrolu poradia čítania riadkov v textových rámcoch a presun nesprávneho poradia riadkov v *Layout.*

Nesprávne usporiadanie textu nemá vplyv na trénovanie modelu, pretože softvér sa učí čítať znaky bez ohľadu na logické usporiadanie textu. Taktiež nemá vplyv na následnú transkripciu dokumentu. Sťažuje však transkripciu nevyhnutného počtu strán potrebných na trénovanie modelu a zároveň komplikuje percepciu prepísaného dokumentu.



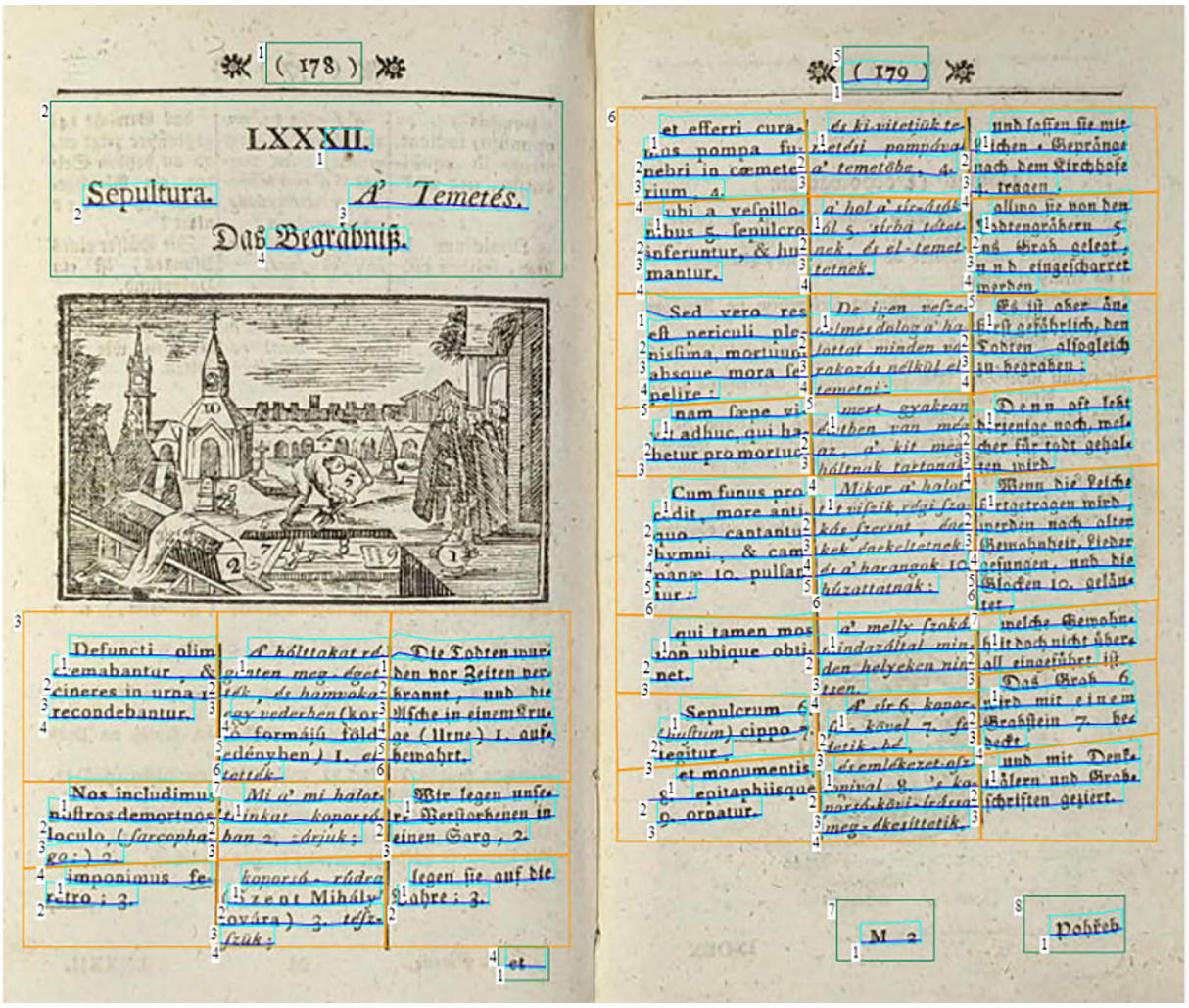
Obrázok 104 Uspokojivé výsledky manuálnej segmentácie textových rámcov a automatickej segmentácie riadkov

Textové rámce na obrázku 104 sú manuálne rozčlenené do viacerých textových rámcov, ktoré označujú jednotlivé časti textu, nie sú však logicky správne usporiadané. Na oddelenie textu v stĺpcoch bol použitý nástroj na segmentáciu tabuliek (viac v kapitole *5.3 Segmentácia tabuliek*).

Vďaka použitiu funkcie *Split lines on region border* pri nastavovaní automatickej segmentácie riadkov nedošlo k spojeniu čiar prechádzajúcich medzi jednotlivých stĺpcami textu.

V segmentácii riadkov je viditeľných niekoľko chýb, napr. neoznačená kustóda na strane 178, neoznačené číslovanie paginácie na strane 179, nedotiahnutá čiara riadku 3 v šiestej bunke pravého stĺpca tabuľky na strane 179, neidentifikovaný riadok v piatej bunke pravého stĺpca tabuľky na strane 179 a i.

Manuálna korekcia chýb takto segmentovaného dokumentu zaberie približne dve minúty na jednu snímku.



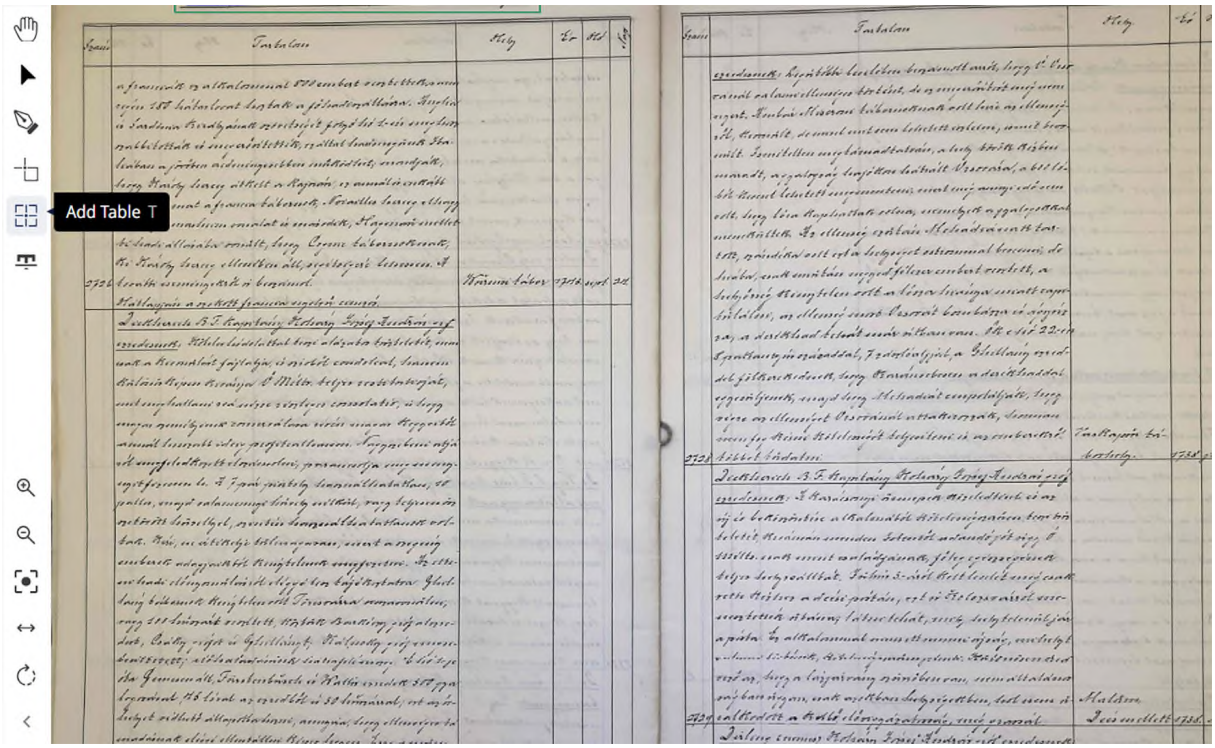
Obrázok 105 Upravené číslovanie poradia čítania objektov segmentácie

5.3 Segmentácia tabuliek

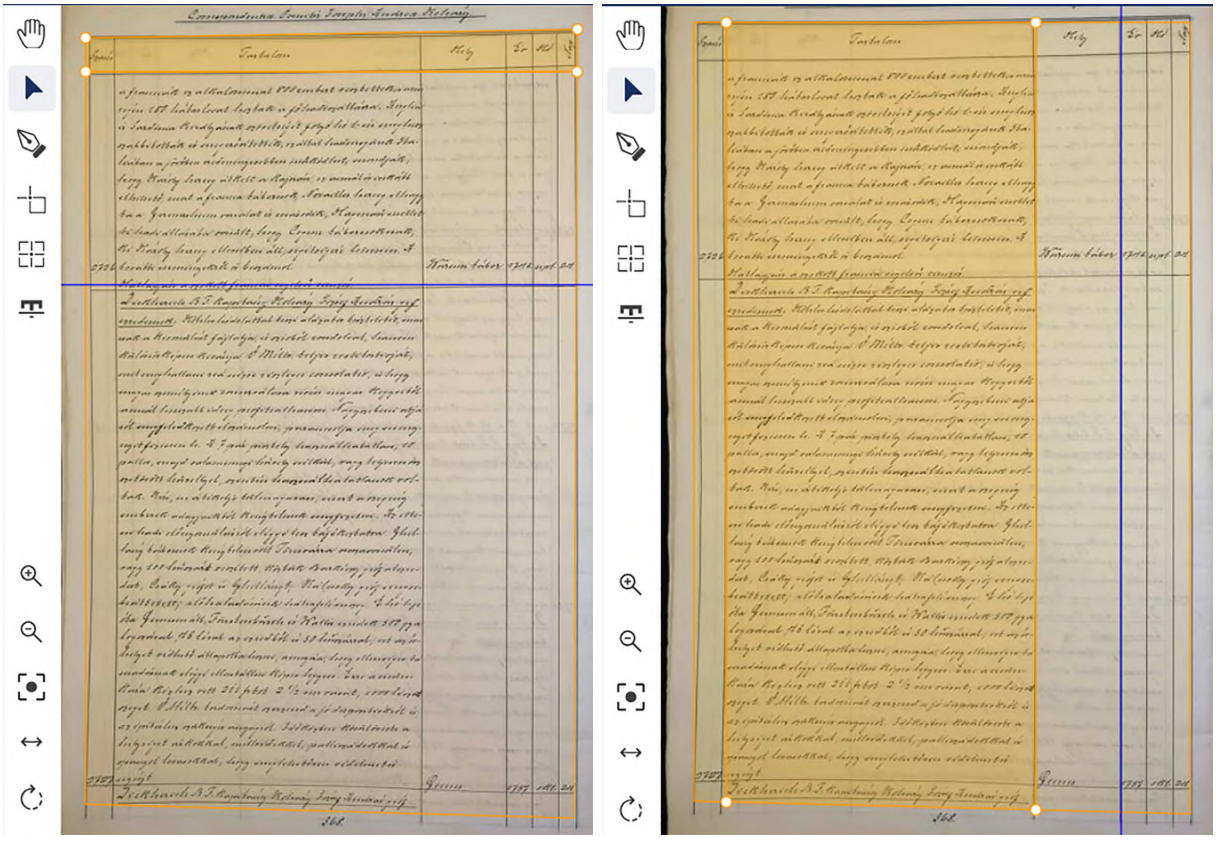
Segmentácia tabuliek v aplikácii Transkribus je poloautomatický proces. Najskôr treba vytvoriť štruktúru tabuľky a následne spustiť automatickú segmentáciu riadkov. Segmentovanie tlačených a ručne kreslených tabuliek umožňuje editor tabuliek. Vďaka nemu si manuálne vytvoríte vonkajšie hranice tabuľky. Takto zadefinovaná oblasť tabuľky následne zjednodušuje a zefektívňuje využívanie ostatných funkcií editora na tvorbu vnútornej štruktúry, t. j. rozdelenie textu do stĺpcov a riadkov.

Pri manuálnom vytváraní tabuľky postupujte nasledovne:

- kliknite na ikonku Pridať tabuľku (*Add table*), ktorá sa nachádza vo vertikálnom paneli nástrojov v ľavej časti okna editora alebo na klávesnici stlačte kláves T,
- na snímke dokumentu označte celú oblasť tabuľky,
- pomocou funkcie *Split element horizontally* (kláves H) rozdeľte tabuľku na riadky – kliknite na všetky čiary, ktoré definujú spodné línie buniek tabuľky,
- pomocou funkcie *Split element vertically* (kláves V) vytvorte v tabuľke stĺpce – kliknite na všetky čiary, ktoré definujú bočné línie buniek tabuľky.



Obrázok 106 Výber funkcie na segmentáciu tabuliek

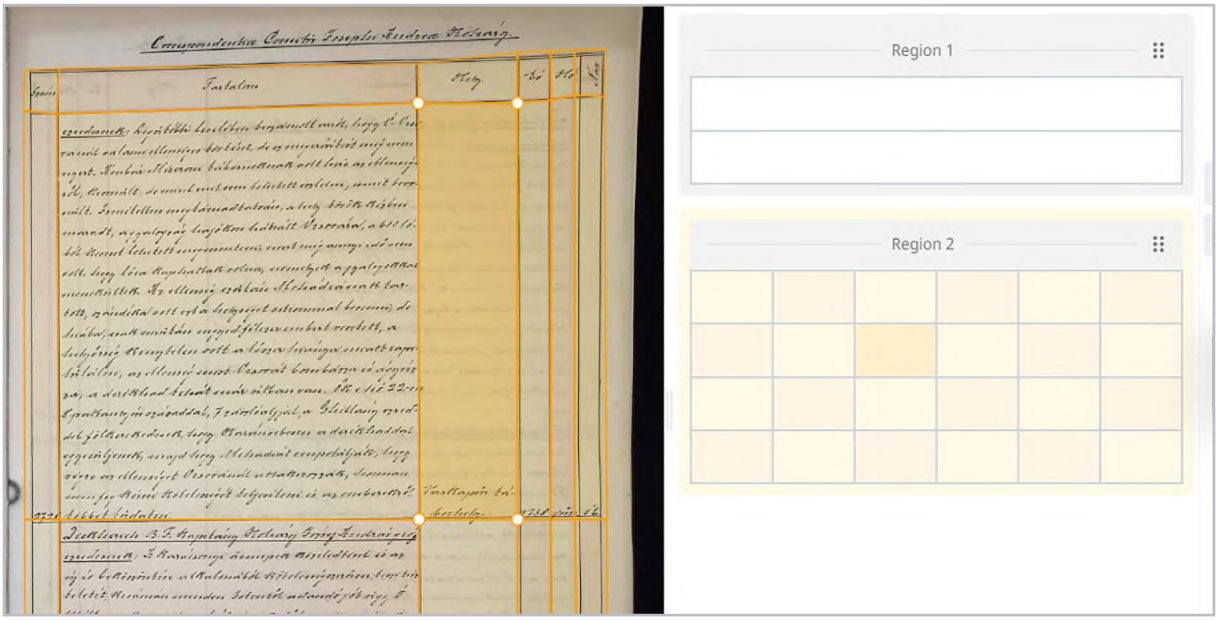


Obrázok 107 Horizontálne a vertikálne členenie tabuľky

Po rozdelení tabuľky na bunky je zvyčajne potrebné manuálne korigovať čiary, ktoré tvoria tvar jednotlivých buniek. Môžete tak urobiť posúvaním bodov/vrcholov oblasti bunky alebo presúvaním okrajov bunky. Existuje aj možnosť pridávať ďalšie body, ktoré by umožnili vytvorenie

špecifických polygónov na presnejšie kopírovanie textu umiestneného v bunke prípadne textu presahujúceho do vedľajšej bunky, vo webovej aplikácii Transkribus však táto funkcia momentálne nie je dostupná.

Segmentovaná tabuľka predstavuje jeden textový rámec. V záložke *Layout* môžete skontrolovať poradie čítania jednotlivých buniek. Poradie čítania jednotlivých buniek je automaticky nastavené po riadkoch od ľavého horného rohu k pravému spodnému rohu tabuľky.



Obrázok 108 Označenie textových rámcov a bunky v záložke Layout

Ak je rozloženie tabuľky na viacerých stránkach podobné, môžete štruktúru tabuľky skopírovať a vložiť z jednej stránky na ostatné:

- vyberte tabuľku a stlačte klávesovú skratku CTRL+C,
- prejdite na ďalšiu stránku a stlačte klávesovú skratku CTRL+V,
- upravte tabuľku tak, aby zodpovedala predlohe na snímke.

Túto úlohu odporúčame vykonať po nakreslení stĺpcov, ale ešte pred nakreslením riadkov, pretože riadky majú tendenciu sa na jednotlivých stranách viac líšiť a ich ručné prispôsobenie môže zabrať viac času.

Po ukončení segmentácie tabuľky môžete pristúpiť k automatickej alebo manuálnej segmentácii čiar (pozri kapitolu *5.1 Spôsoby segmentácie*) a kontrole poradia čítania riadkov (pozri kapitolu *5.2.3 Kontrola a úprava poradia čítania textových rámcov a riadkov*).

6 Tvorba modelu automatickej transkripcie

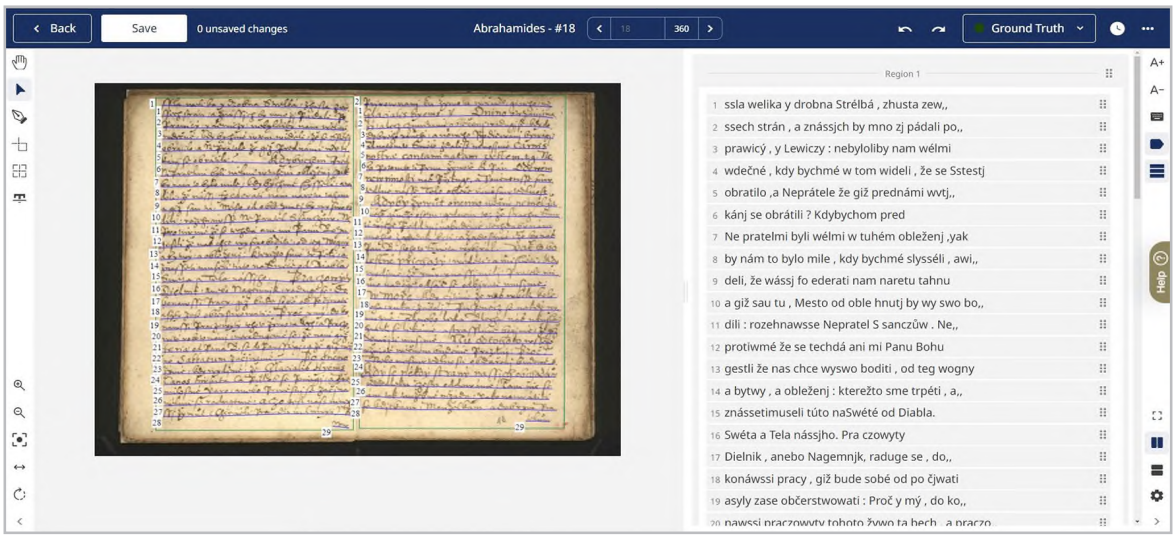
Transkribus umožňuje trénovaním vytvoriť vlastný model na rozpoznávanie rukopisných alebo tlačených textov, ktorý sa použije na automatickú transkripciu celej zbierky dokumentov. Využíva podoblasť umelej inteligencie – nástroj strojového učenia PyLaia. Proces tvorby modelu na jeden alebo viacero rukopisov zahŕňa prepis dokumentu (prípravu vzorky *Ground Truth*), spustenie trénovania modelu, vyhodnotenie úspešnosti modelu a jeho zdokonaľovanie.

6.1 Prepis dokumentu (príprava vzorky *Ground Truth*)

Prepisy v aplikácii Transkribus môžu byť použité na trénovanie modelu *PyLaia* a tiež ako základ pre vytvorenie knižnej alebo digitálnej pramennej edície. Na trénovanie modelu postačuje jednoduchý prepis. Účinnosť modelu závisí od kvality trénovaného materiálu (manuálnej transkripcie), kvality digitalizátov a ich paleografickej náročnosti. Existujú aj pokročilé možnosti prepisu na prípravu digitálnej edície. Obsahujú napr. na úpravu poradia textu, použitie historických znakov, pridávanie značiek (tagov), metadát a rozpisovanie skratiek.

Jednoduchý prepis na trénovanie modelu

Po segmentácii dokumentu kliknutím na vybranú stranu v zbierke môžete prepisovať rukopis. Pole textového editora je možné nastaviť na bočnom paneli nástrojov vpravo dole buď ako vertikálne stĺpcové zobrazenie (*Column View*) alebo horizontálne riadkové zobrazenie (*Row View*). Pre každý (základný) riadok na obrázku nájdete zodpovedajúci riadok v textovom editore. Zachovajte rovnaké poradie (číslovanie) riadkov v textovom editore aj v segmentovanom dokumente. Prepíšte text podľa zdrojového dokumentu. Dokument môže prepisovať viac používateľov, ale nemali by spracovávať rovnakú stranu dokumentu súčasne.



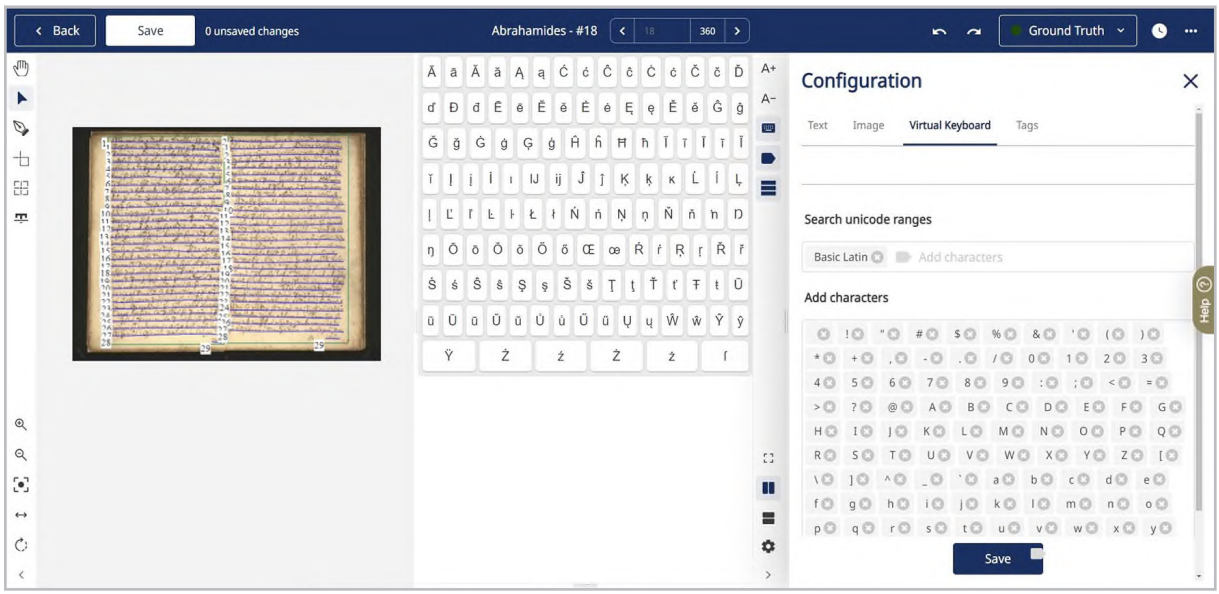
Obrázok 109 Prepis segmentovaného dokumentu. Stĺpcové zobrazenie textového editora

Transkripcia a virtuálna klávesnica

Prepis, ktorý bude slúžiť ako základ pre vedeckú edíciu, by mal používateľovi poskytnúť viac kontextových údajov ako jednoduchý prepis. V tomto prípade zohráva dôležitú úlohu nielen strojové čítanie, ale aj vlastné čítanie používateľa.

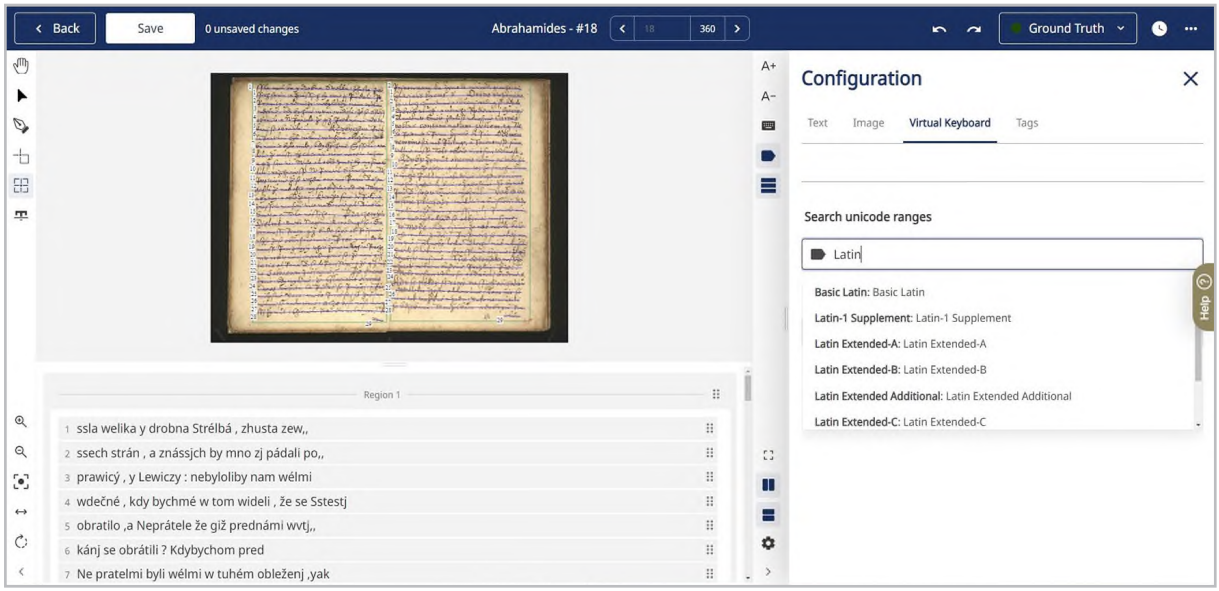
Počas prepisu môžete pridať špeciálne znaky a symboly zo štandardizovanej sady *Unicode* použitím virtuálnych klávesníc (*Virtual keyboards*) v poli textového editora. V prípade súbežnej

práce viacerých osôb je potrebné zabezpečiť jednotné používanie znakov z konkrétnej klávesnice. Znaky z rozdielnych klávesníc sa totiž môžu prejaviť vo zvýšenej chybovosti počas následného trénovania modelov. Kliknutím na ikonku nastavenia (*Settings*) na bočnom paneli nástrojov vpravo sa v okne konfigurácie (*Configuration*) zobrazí ponuka virtuálnych klávesníc. Predvolená je klávesnica *Basic Latin*. Funkcia pridávania znakov (*Add characters*) v konfigurácii zobrazuje znaky ku klávesnici, ktoré je možné dopĺňať a odoberať po uložení zmien (*Save*).



Obrázok 110 Predvolená virtuálna klávesnica Basic Latin v okne konfigurácie

Ak znaky zo základnej klávesnice *Basic Latin* nepostačujú, pole v okne konfigurácie umožňuje hľadať v sadách *Unicode* (*Search unicode ranges*). Vpísaním slova *Latin* do vyhľadávacieho poľa sa rozšíri ponuka klávesníc s najčastejšie zastúpeným písmom – latinkou. Pre grafémy používané počas prepisu historických rukopisov aj tlačí je zvlášť vhodná napr. klávesnica *Latin-1 Supplement*. Zmenu klávesnice je vždy potrebné potvrdiť kliknutím na ikonku Uložiť (*Save*). Vpísaním do vyhľadávacieho poľa môžete zmeniť a pridať ďalšie písma: *Cyrillic*, *Greek*, *Hebrew* a pod.



Obrázok 111 Ponuka klávesníc so znakmi z latinského písma v okne konfigurácie. Textový editor v riadkovom zobrazení

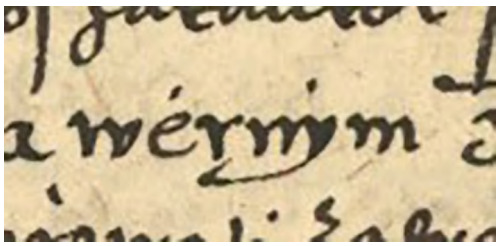
Diakritika a ligatúry

Existujú dve možnosti spracovania správneho prepisu znakov:

Možnosť 1 **Mierna normalizácia** podľa slovníka

Hlavné pravidlo, ktoré sa tu uplatňuje: ak jasne vidíte základný znak glyfu (grafémy) a ak sa základný znak zároveň používa v slovníku na vyjadrenie tohto glyfu, zachovajte základný znak.

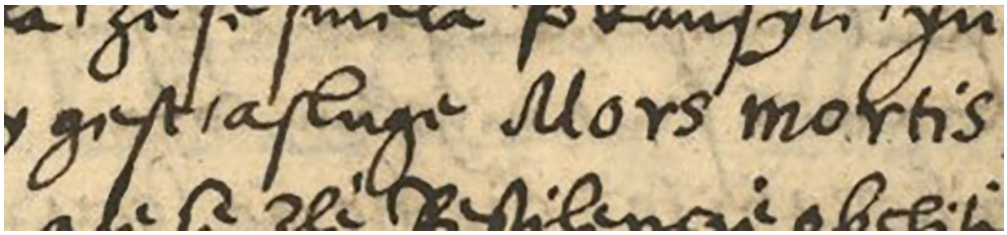
> *Príklad 1:* litery *e* a *y* v latinskej minuskule sa v mnohých dokumentoch objavujú s diakritickými znamienkami odlišnými od súčasného úzu (bodka, dvojbodka).



Obrázok 112 Slovo wernym zapísané novogotickou kurzívou. Litery sú prepísané bez diakritiky

V jednoduchom prepise ich môžete prepísať ako latinskú minuskulnú literu *e* a *y*, keďže základný znak je stále jasne viditeľný.

> *Príklad 2:* Latinská minuskulná litera *s* sa vo väčšine európskych historických písiem vyjadruje dvoma grafémami. Nachádzame preto jasný rozdiel medzi okrúhlym *s* a kurzívnym dlhým *ſ*.

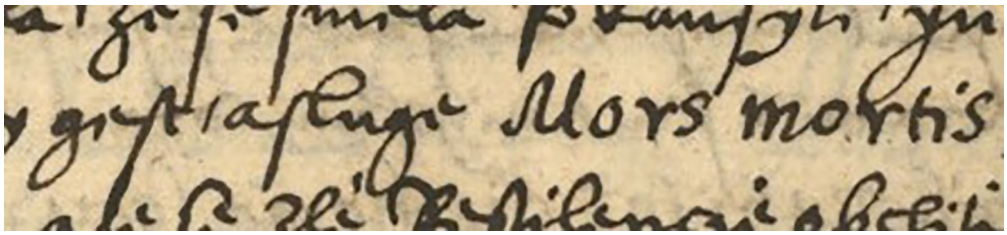


Obrázok 113 Jednoduchý prepis dlhého a okrúhleho s v slovách gest, a sluge Mors mortis

Aj keď existuje jasný rozdiel, jednoduchý prepis by použil okrúhle *s* v obidvoch prípadoch.

Možnosť 2 **Paleografický prepis (transliterácia)**

Ortograficky vernému prepisu zodpovedá označenie transliterácia. Na platforme Transkribus sa pre všetky druhy prepisu konvenčne používa pojem transkripcia.



Obrázok 114 Paleografický prepis dlhého a okrúhleho s v slovách geſt, a ſluge Mors mortis

V **tlačených textoch** môže zohrať rolu prepisovanie ligatúr. Znovu možno použiť rovnaké pravidlo: hoci sa špecifické kombinácie písmen ako napríklad *ft* alebo *ft*, keď sa spájajú dve grafémy, dajú vyjadriť aj špecifickými znakmi *Unicode*, odporúčame ich prepisovať bez ligatúr podľa slovníka.

Interpunkčné znamienka

Interpunkčné znamienka sa prepisujú rovnakým spôsobom ako ostatné znaky. Použite príslušný znak na klávesnici. Na rozdiel od transkripčných pravidiel, ktoré interpunkčné znamienka pridávajú alebo vynechávajú podľa dnešného ponímania, odporúčame v prepise zachovať pôvodné znamienka. Napr. dvojbodky v historických textoch sa často používajú na značenie skracovania slov. Mali by sa prepisovať ako dvojbodky.

Prepis novovekých dokumentov by mal odrážať predlohu, aj keď sa interpunkčné znamienko použilo spôsobom, ktorý nezodpovedá súčasnému úzu.

Prepis stredovekých dokumentov by nemal používať modernú interpunkciu. Vhodnejšie je vynechať všetky interpunkčné znamienka alebo použiť špecifické symboly zo sady *Unicode*.

Zásady prepisu v Transkribuse

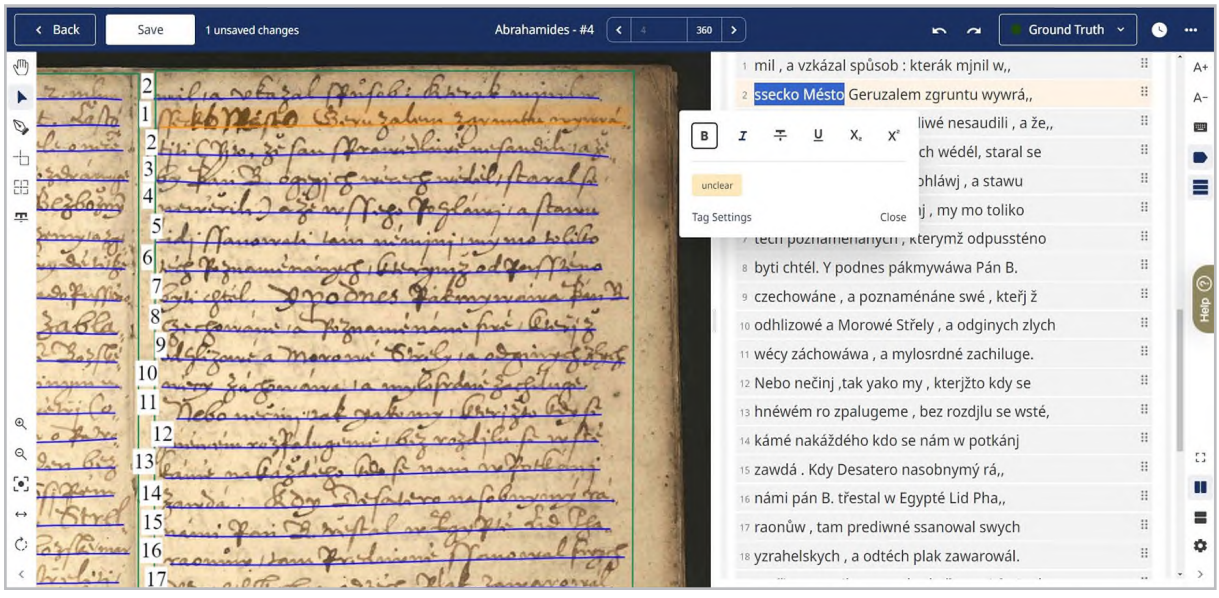
Je dôležité, aby bol jednoduchý prepis určený na trénovanie modelu konzistentný, vypracovaný jednotnou metodikou.

Prepis má zodpovedať rukopisnej predlohe vrátane pisárskych chýb. Slová by sa mali oddeľovať alebo spájať podľa predlohy.

Podľa rukopisu rozlišujte minuskulu a majuskulu. Ak literu nie je možné jasne rozlíšiť, rozhodnutie závisí od vás.

Slová s rozdeľovníkom na konci alebo uprostred riadku majú byť prepísané a rozdelené podľa pôvodného textu.

Po zvýraznení vybraného prepisu kurzorom v riadku textového editora sa automaticky zobrazí okno s ponukou odlíšiť zvýraznený prepis boldom, kurzívou, prečiarknutím, podčiarknutím, dolným indexom a horným indexom. Tieto funkcionality môžete využiť kliknutím na príslušnú ikonku.



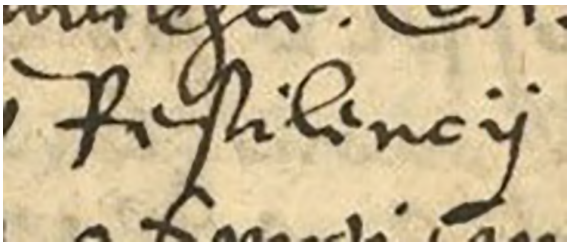
Obrázok 115 Okno s ponukou odlíšenia prepisovaného textu boldom, kurzívou, prečiarknutím, podčiarknutím, dolným alebo horným indexom

Zásady používania špeciálnych znakov

Odlišné typy a druhy písma (napr. gotické a humanistické) nie sú osobitne značené. Skratky prepisujte podľa predlohy – nerozpisujte ich. Platí to pre historické spôsoby skracovania slov aj pre skratky používané v súčasnosti.

Diakritické znaky môžete vynechať (v prípade jednoduchého prepisu) alebo ich použiť podľa predlohy (v prípade transliterácie).

Častým prípadom je zamieňanie hlások *i* a *j*, ktoré je v rukopisoch náročné rozlíšiť najmä v majuskule. Zdvojenie znakov *ii* alebo *ij* sa prejavilo v používaní grafémy *ÿ*. Vkladajte ju pomocou virtuálnej klávesnice. Platí pritom odporúčanie, aby sa každý znak pre dostatočné osvojenie strojového učenia vyskytol v prepísanej vzorke aspoň päťdesiatkrát.



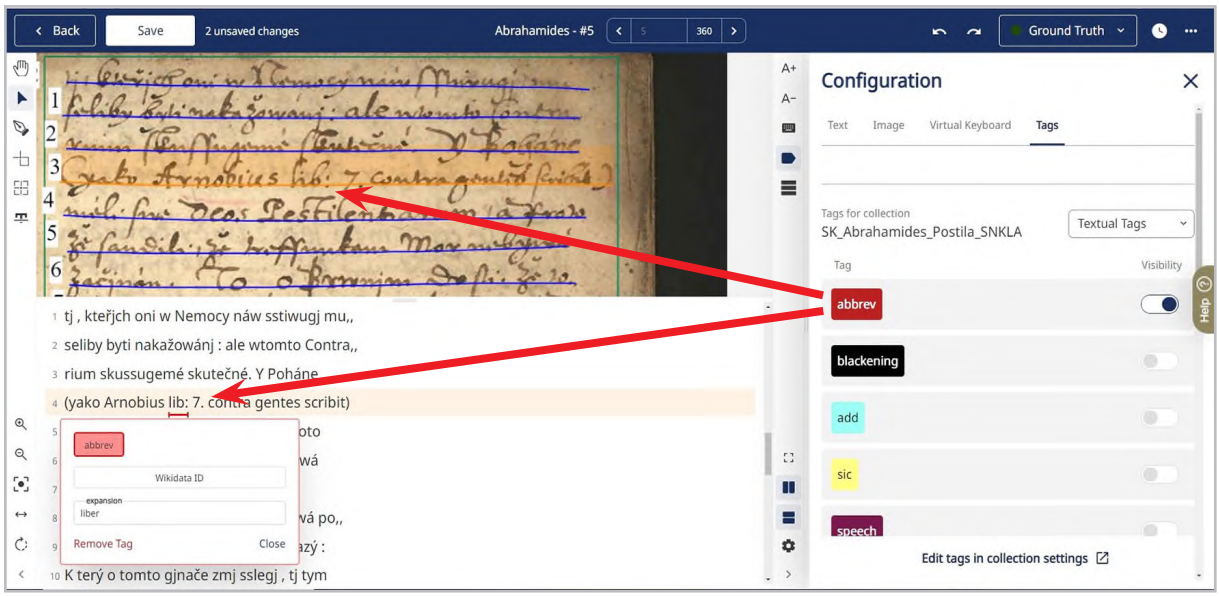
Obrázok 116 Dvojhláska v podobe samostatnej grafémy v zápise slova pestilencÿ

Ligatúry môžete rozpisovať, nie je potrebné používať pritom osobitné znaky ako pri skratkách. Ak sa rozhodnete ponechať ligatúru (napr. *æ*), mala by sa v prepísanej vzorke vyskytnúť v odporúčanom počte.

K formám zápisu hlásky *s*, okrem už spomínaného okrúhleho a dlhého *s*, patrí aj dvojité *s* často v podobe ligatúry *ß*. Ostré *s* môžete prepisovať ako *ss* alebo použiť znak *ß*, ak sa v prepise vyskytuje päťdesiatkrát.

Tagovanie skratiek

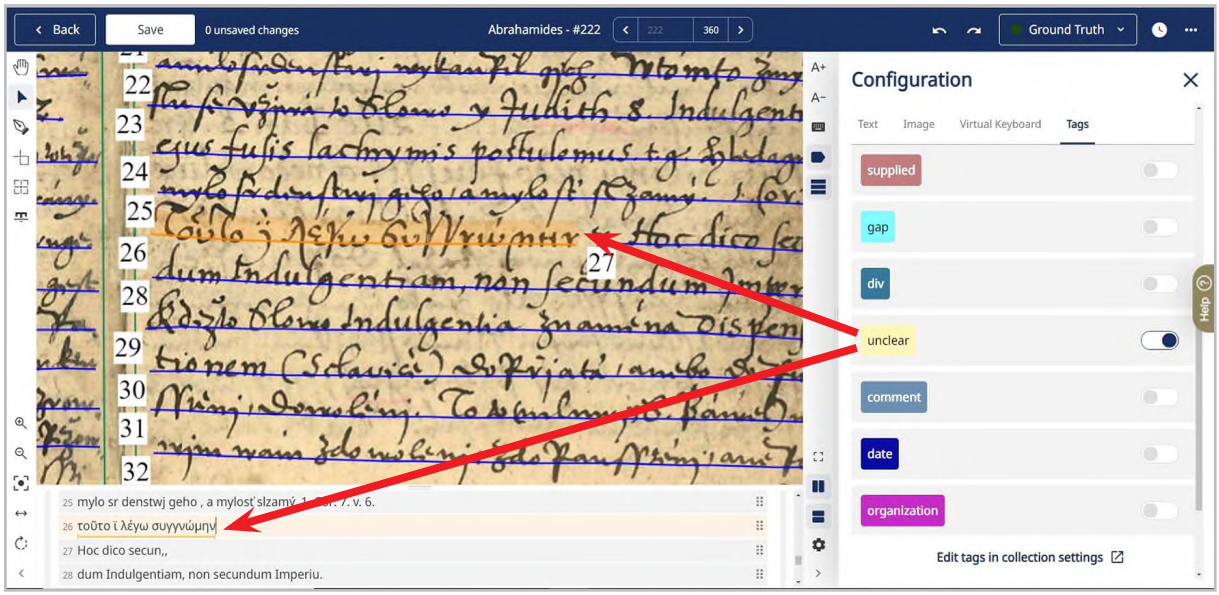
Prepis dokumentu v textovom editore je možné doplniť tagmi. Kliknutím na bočný panel nástrojov vpravo dolu na ikonku nastavenia sa v okne konfigurácia (*Configuration*) zobrazí ponuka tagov. Vyberte z nich textové tagy (*Textual Tags*). V zozname sa prvé zobrazia skratky (*abbrev*), pričom prekliknutím voľby v riadku vpravo označíte ich viditeľnosť. V prípade opakujúcich sa skratiek môžete skrátené slovo alebo jeho časť s výskytom skratky označiť v textovom editore kurzorom – zostane podčiarknuté červenou čiarou. V automaticky otvorenom okne sa po kliknutí na funkciu *abbrev* ukáže riadok s označením rozpísanie (*expansion*), v ktorom môžete skratku vysvetliť.



Obrázok 117 Označenie skratky lib: v textovom editore a jej rozpísanie pomocou tagu

Nečitateľné miesta

V prípade nečitateľnej pasáže alebo pasáže zapísanej iným druhom písma môžete v ponuke textových tagov zvoliť označenie Nejasné (*unclear*). Po zvýraznení vybranej pasáže kurzorom v textovom editore sa automaticky zobrazí okno s funkcionalitou *unclear*. Ak váš prepis umožňuje čítať pasáž aj iným spôsobom, môžete po zakliknutí tagu v riadkoch doplniť alternatívne čítanie a vysvetliť dôvody čítania. V textovom editore pasáž zostane podčiarknutá žltou čiarou. Takéto pasáže nemusia byť zahrnuté do trénovania modelu.



Obrázok 118 Označenie pasáže zapísanej gréčtinou tagom unclear

Stavy dokumentu a verzia *Ground Truth*

Editované alebo dokončené strany označte zodpovedajúcim príznakom (odlíšeným farebne) na hornej lište vpravo:

- prebiehajúci (*In Progress*) je označenie pre stranu, ktorú je stále možné prepisovať,

- hotový (*Done*) sa používa na označenie strany, ktorá je už prepísaná, ale ešte vyžaduje kontrolu,
- finálna verzia (*Final*) označuje prepísanú a skontrolovanú stranu,
- základná pravda (*Ground Truth*) je výsledná verzia prepisu strany vhodnej pre tvorbu modelu.



Obrázok 119 Označenie stavu transkribovanej strany v paneli nástrojov

6.2 Spustenie trénovania modelu

Pred spustením trénovania modelu je potrebné pripraviť si vzorku *Ground Truth* (viac v kapitole *6.1 Prepis dokumentu*), t. j. k originálu čo najpresnejší prepis (manuálny alebo automaticko-manuálny), ktorý sa umelá inteligencia naučí „čítať“. V závislosti od typu prepisovaného dokumentu (tlač, rukopis) a počtu rúk (resp. meniaceho sa štýlu rukopisu autora) sa odporúča trénovať model na 5 000 až 15 000 slovách, čo zodpovedá prepisu približne 25 až 75 strán:

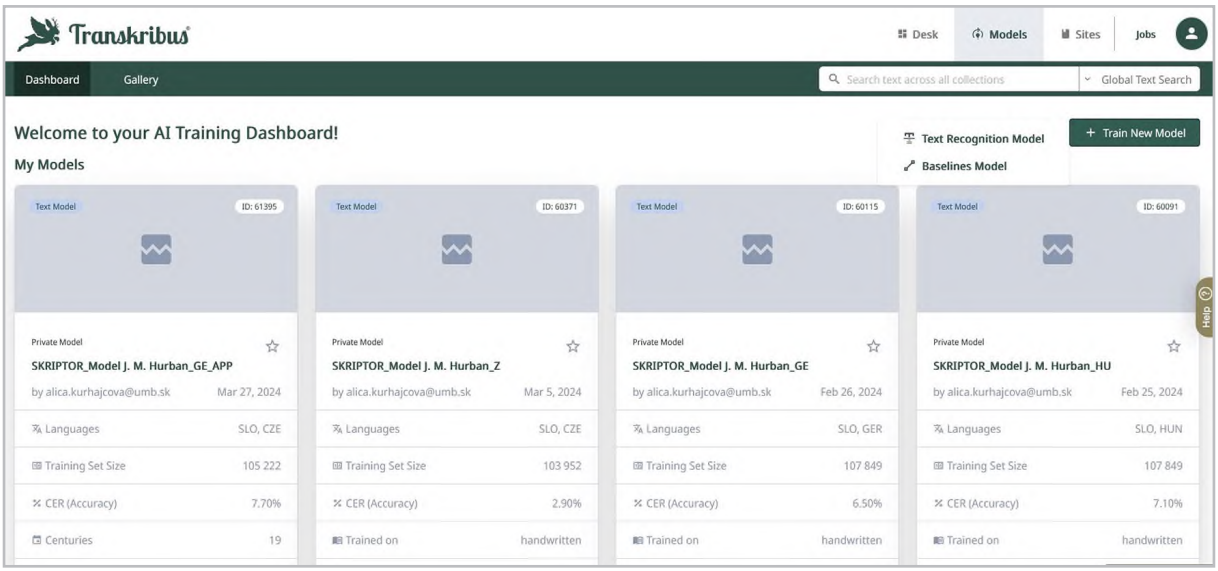
- v prípade tlačeného textu na približne 5 000 slovách,
- v prípade rukopisného textu na aspoň 10 000 slovách pre každú ruku.

Ak chcete trénovať model na rozpoznanie troch rôznych „rúk“, mali by ste prepísať aspoň 30 000 slov, 10 000 slov pre každú ruku. Platí to aj v prípade jedného autora, ak sa jeho rukopis v priebehu života menil. Veľký model trénovaný na viac ako 100 000 slovách, ktorý obsahuje rôzne ruky z rovnakého obdobia a regiónu, by mal byť schopný rozpoznať aj rukopis, ktorý sa do trénovania nedostal (aj keď výsledky jeho prepisu môžu byť v porovnaní s trénovanými stranami o niečo horšie).

Je dôležité, aby strany vo vzorke *Ground Truth* boli **reprezentatívne**, t. j. aby obsahovali varianty všetkých typov písiem (resp. aj jazykov, abecied, no aj štýlov písania), ktoré má váš model rozpoznať (čiže prepísať) súčasne. Strany zahrnuté do vzorky *Ground Truth* majú vplyv na kvalitu modelu.

Po príprave vzorky *Ground Truth* nasleduje spustenie trénovania nového modelu (*+Train New Model*). Funkciu nájdete na záložke Modely (*Models*) vpravo hore. Po kliknutí na tlačidlo trénovania modelu máte možnosť trénovať:

- model na rozpoznávanie textu, resp. textový model (*Text Recognition Model*),
- model na základné čiary (*Baselines Model*).



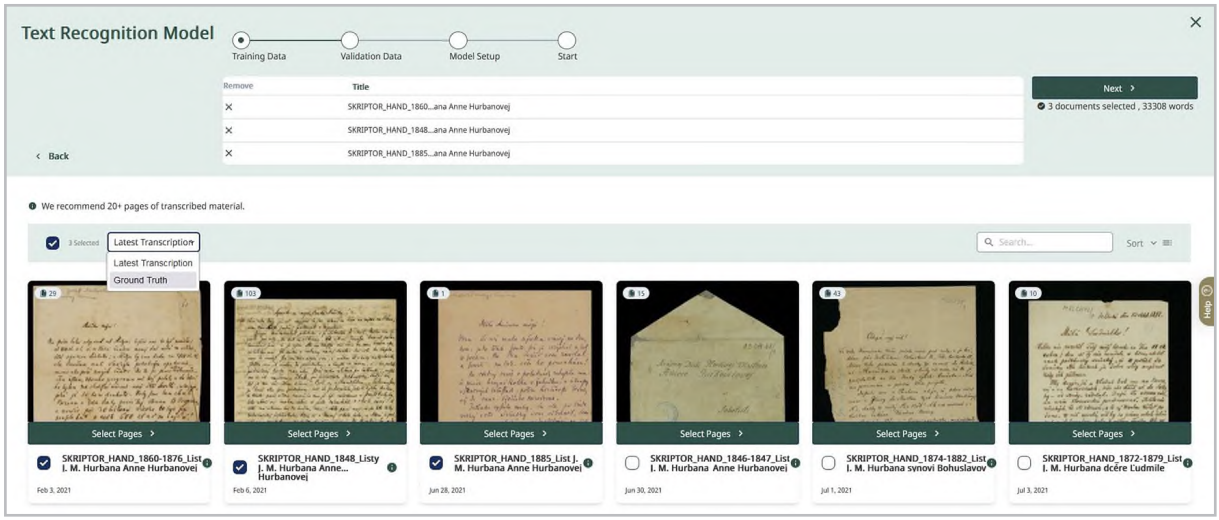
Obrázok 120 Záložka s funkciou spustenia trénovania nového modelu

Postup trénovania oboch modelov je podobný (odlišné sú napríklad nastaviteľné parametre) a pozostáva zo štyroch krokov:

Krok 1 Cvičné dáta (*Training Data*)

V prvom okne vyberáte z príslušnej zbierky dokumentov cvičné dáta – buď ako celé súbory alebo konkrétne strany (*Select Pages*), na ktorých sa model môže vytrénovať. Zmeny pri výbere dát priebežne ukladajte (*Save and go back*). Na cvičných dátach sa stroj „učí“, t. j. pri každom cykle „prečíta“ rovnakú stranu, pričom chybne prečítané znaky pri každom nasledujúcom cykle vyradí. Pod vyobrazenými súbormi sa nachádza možnosť vybrať verziu prepisu – poslednú (*Latest Transkription*) alebo bezchybnú (*Ground Truth*). Pri výbere druhej možnosti máte istotu, že sa do cvičných dát nedostanú strany s iným, t. j. neželaným príznakom (napr. *In Progress*).

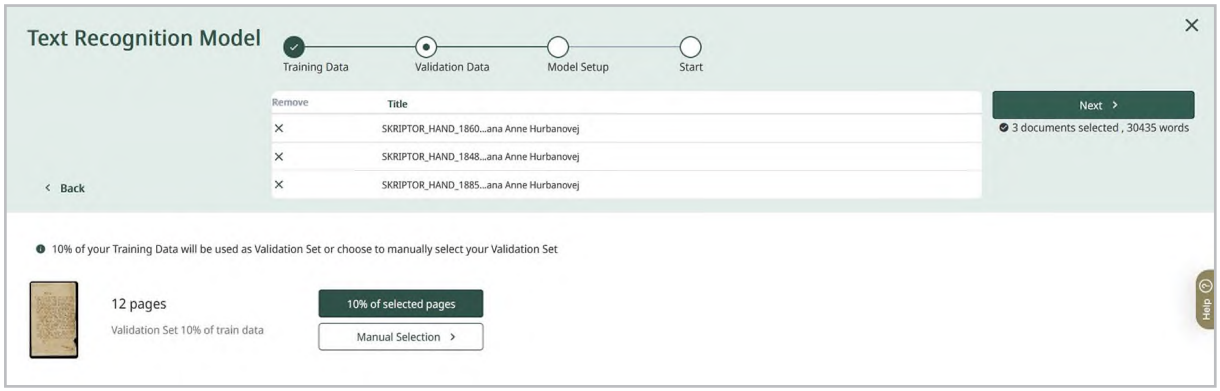
Pod tlačidlom *Next*, ktorým po každom výbere prejdete k ďalšiemu kroku (oknu), sa zobrazí počet vybraných dokumentov, a pri trénovaní textových modelov aj počet slov.



Obrázok 121 Náhľad prvého kroku: Výber cvičných dát

Krok 2 Overovacie dáta (*Validation Data*)

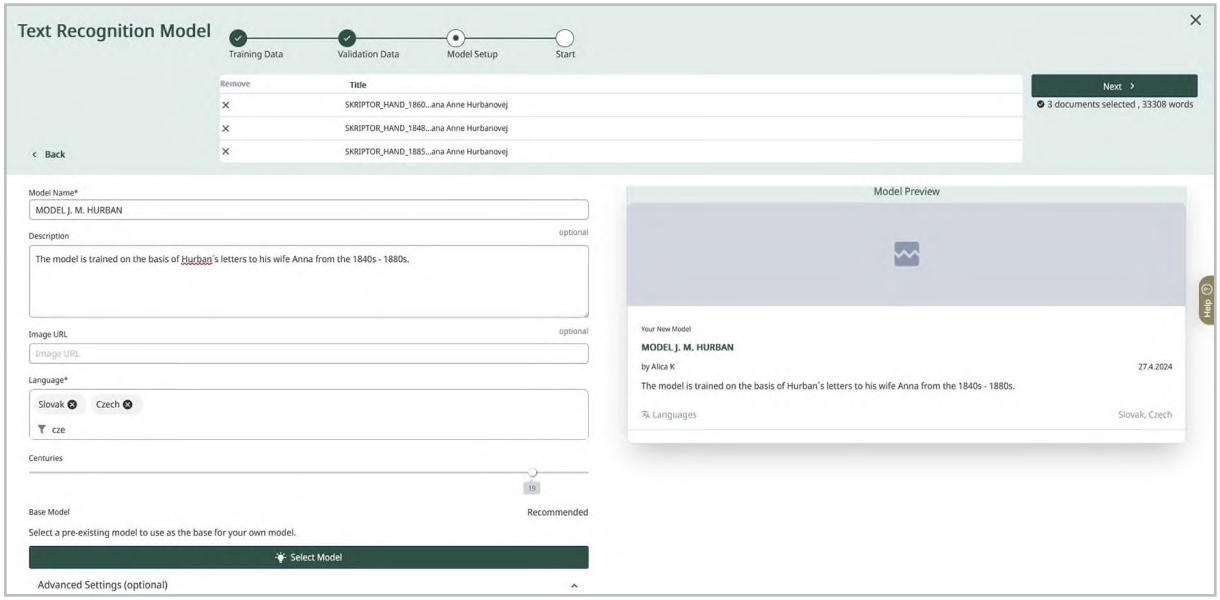
Otvorí sa okno, v ktorom si vyberáte overovacie dáta, na ktorých sa presnosť vytrénovaného modelu automaticky odskúša (overí). V porovnaní s cvičnými dátami sú overovacie dáta menšie, spravidla 10 % (prípadne 5 % či len 2 %) z celkovej vzorky *Ground Truth*. Systém umožňuje aj manuálny výber strán (*Manual Selection*). Overovacie dáta by však mali byť reprezentatívne, t. j. mali by obsiahnuť príklady všetkých písmen, jazykov a iných atribútov zahrnutých v cvičných dátach. Pokiaľ sú príliš homogénne, výkon modelu môže byť nízky, prípadne skreslený.



Obrázok 122 Náhľad druhého kroku: Výber overovacích dát

Krok 3 Nastavenie modelu (*Model Setup*)

V ľavej časti nového okna si nastavte vstupné údaje ako aj ďalšie parametre, ktorými môžete zvýšiť funkčnosť a efektívnosť trénovaného modelu. Ako prvé uveďte **povinné údaje** – názov modelu (*Model Name*) a jazyk/y dokumentu/ov (*Language*) a následne **voliteľné údaje** – popis modelu (*Description*), URL adresu obrázka (*Image URL*) a na pohyblivej osi príslušné storočia (*Centuries*), z ktorých dokumenty pochádzajú. Vložené údaje sa zobrazia v pravej časti okna v Náhľade modelu (*Model Preview*).

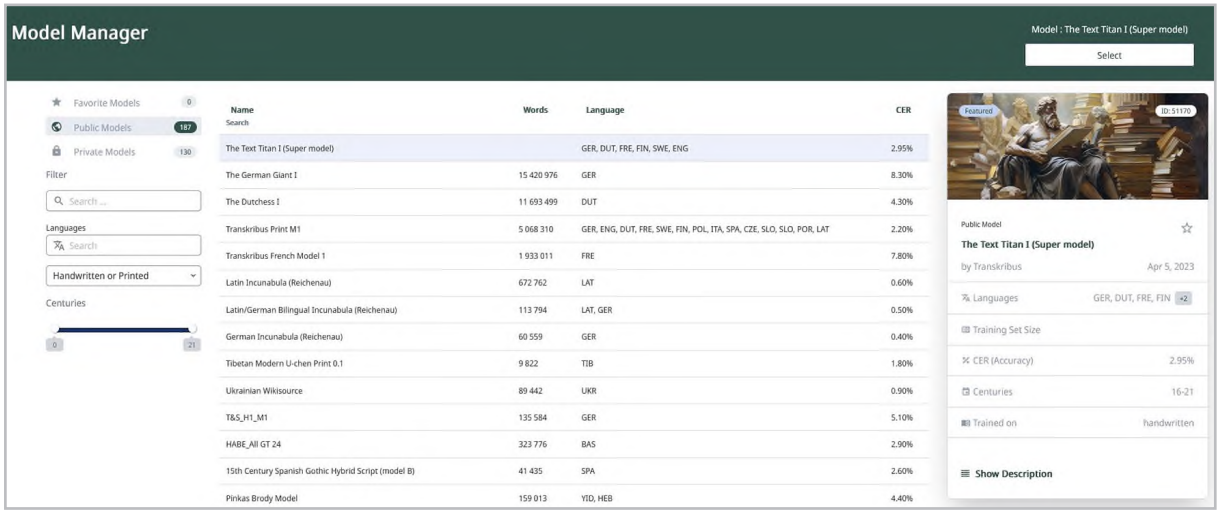


Obrázok 123 Náhľad tretieho kroku: Nastavenie a popis modelu

Ak chcete zefektívniť strojové učenie pri trénovaní modelu na rozpoznávanie textu (*Text Recognition Model*), máte možnosť vybrať si **základný model** (*Base Model*). Keď do vami tré-

novaného modelu pridáte dáta základného modelu, umožní vám to začať trénovanie s menšou vzorkou a za istých okolností aj zlepšiť vami vytrénovaný model (viac o zdokonaľovaní modelu v kapitole *6.3 Vyhodnotenie úspešnosti modelu a jeho zdokonaľovanie*).

Prehľad základných modelov sa otvorí stlačením tlačidla Vybrať model (*Select model*). V ľavom stĺpci vyberáte z ponuky vlastných, ešte nezverejnených modelov (*Private Models*) alebo z verejne dostupných modelov (*Public Models*) iných používateľov, a to za predpokladu, že majú podobné vlastnosti ako vaše cvičné dáta. Na vyhľadanie základného modelu s podobnými vlastnosťami slúži filter s možnosťou vyhľadávať model podľa kľúčových slov (*Search...*), jazykov, storočí, či typov písma, a to podľa toho, či pracujete s písaným textom, tlačeným alebo s oboma typmi (*Handwritten or Printed*). Zhrnutie charakteristík každého modelu (vrátane percentuálneho vyjadrenia chybovosti znakov *% CER Accuracy*) sa zobrazí po kliknutí na príslušný model v pravej časti okna. Detailnejší popis modelu s grafickým znázornením procesu trénovania a miery chybovosti nájdete pod ikonkou Zobraziť popis (*Show Description*). Základný model pridáte kliknutím na príslušný model a stlačením tlačidla Vybrať (*Select*) v pravom hornom rohu.



Obrázok 124 Okno s ponukou verejných a súkromných základných (base) modelov

Následne podľa typu dokumentu a skúseností, aké nadobudnete pri práci s nástrojom PyLaia, môžete vyplniť pokročilé parametre (*Advanced Settings – optional*) a podľa potreby meniť predvolené nastavenia. Rozlišujeme:

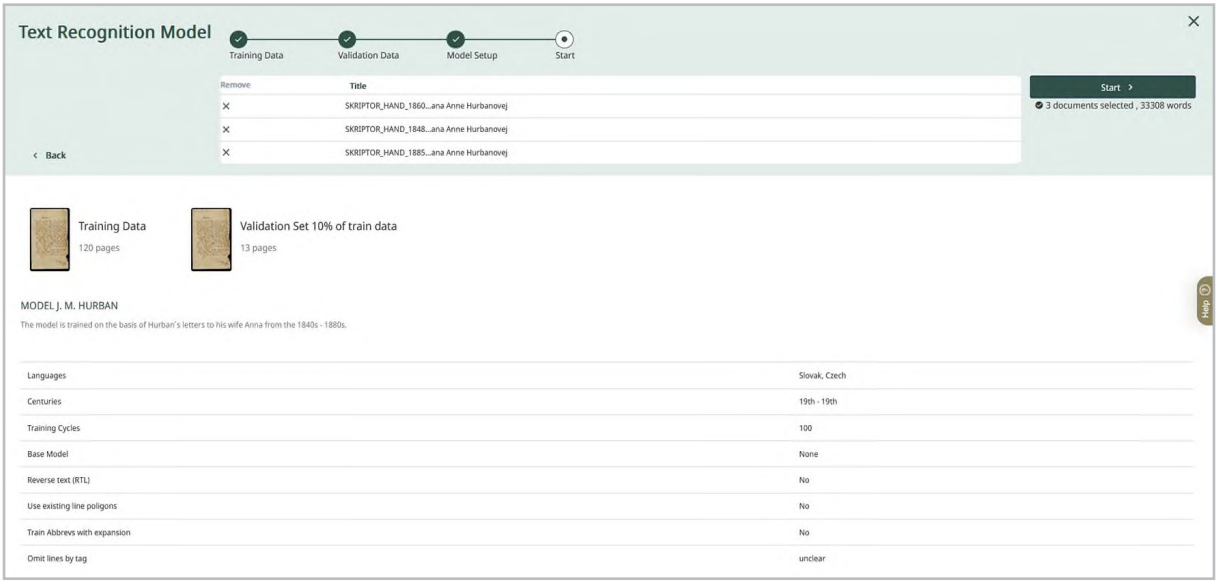
Parametre pri trénovaní modelu na rozpoznávanie textu:

- **cykly trénovania** (*Training Cycle*) – zadávate celkový počet opakovaní trénovania. Stroj sa opakovane „učí“ čítať cvičné dáta, t. j. pri každom cykle prečíta tú istú stranu a vyhodnotí ju. Na začiatok sa odporúča ponechať predvolené nastavenie (100 cyklov). Treba mať na pamäti, že zvyšovaním počtu cyklov sa aj proces trénovania predlžuje a naopak znižovaním zasa skracuje. Zvyšovanie počtu cyklov nemusí mať vplyv na výslednú úspešnosť modelu,
- **predčasné zastavenie** (*Early Stopping*) – zadávate minimálny počet opakovaní tréningu. Pre väčšinu modelov postačuje predvolené nastavenie (20 cyklov). Znamená to že, ak sa hodnoty modelu zlepšujú, trénovanie bude aj po dosiahnutí 20 cyklov pokračovať. Ak však už hodnoty nebudú vykazovať zlepšenie, trénovanie sa automaticky zastaví a vyhodnotí,

- **obrátený text** (*Reverse Text*) – túto možnosť označte, ak je smer písania na obrázku opačný ako pri prepise (napr. originál bol napísaný sprava doľava a prepísaný text zľava doprava)
- **na trénovanie použite existujúci polygónový ťah** (*Use existing line polygons for training*) – túto možnosť označte, ak chcete počas trénovania zohľadniť existujúci, nie predvolený polygónový ťah,
- **trénovať skratky** (*Train Abbrevs with expansion*) – túto možnosť použite, ak chcete dosiahnuť lepšie výsledky pri rozpoznávaní skratiek,
- **vynechať riadky označené tagom** (*Omit lines by tag*) – túto možnosť označte, ak chcete z procesu trénovania vynechať riadky obsahujúce slová označené tagmi Nejasný (*unclear)* alebo Medzera (*gap*). Vynecháte tak nielen označené slovo, ale aj celý riadok, keďže trénovanie prebieha na úrovni riadkov.

Parametre pri trénovaní modelu na základné čiary:

- **cykly trénovania** (*Training Cycle*) – zadávate len celkový počet opakovaní tréningu,
- **rýchlosť učenia** (*Learning Rate*) – predvolená hodnota 0,001, ktorú odporúčame ponechať, definuje, ako rýchlo bude učenie pri prechode od jedného cyklu k druhému prebiehať,
- **nastavenie mierky** (*Scaling Adjustment*) – upravujete veľkosť obrázku, t. j. minimálne a maximálne hodnoty rozsahu škálovania (*min/max*) a mieru overenia (*validation*),
- **sklon a posúvanie** (*Tilt & Shift*) – upravujete obraz naklonením alebo posunutím jeho častí, čo môže byť užitočné pri trénovaní modelu s rôznymi polohami a uhlami písania,
- **efekt kolísania** (*Wobble Effect*) – túto možnosť označte, ak chcete do obrázka pridať „efekt zvlnenia“, ktorý kopíruje prirodzené nezrovnalosti, ktoré vznikli pri písaní alebo tlači. Hodnoty „X“ a „Y“ regulujú intenzitu tohto efektu v horizontálnom a vo vertikálnom smere,
- **otáčanie obrazu** (*Image Rotation*) – touto funkciou simulujete smerovanie obrázku. Je na vás, či si vyberiete možnosť: bez otáčania (*None*), náhodné otočenie o 90, 180 alebo 270 stupňov (*Quarter Turns*) alebo náhodné uhly (*Random*),
- **tenké alebo zhustené riadky** (*Thin or Thicken Lines*) – upravujete hrúbku čiar na obrázku, čo môže byť užitočné na rozpoznanie tenších alebo hrubších ťahov rukopisu. Predvolená hodnota je 1.



Obrázok 125 Náhľad štvrtého kroku: Zhrnutie údajov a nastavení pred spustením trénovania modelu

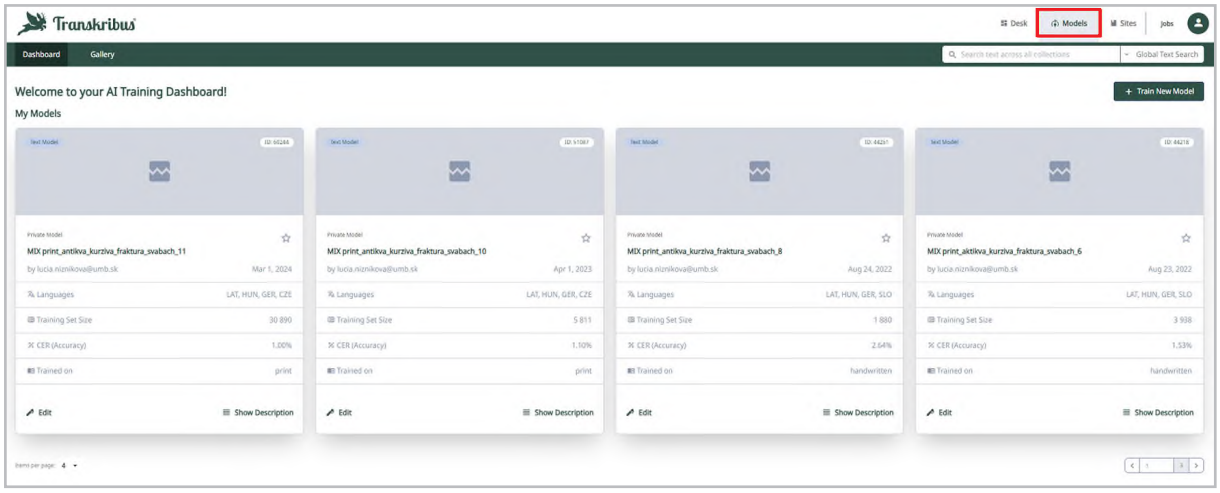
Krok 4 **Spustiť trénovanie** (*Start*)

V poslednom okne sa zobrazí zhrnutie pomeru cvičných a overovacích dát, vložených základných a voliteľných údajov a nastavení. Ak sú pre vás tieto dáta postačujúce (napr. počet strán v cvičných alebo overovacích dátach), stlačte tlačidlo *Start*, čím spustíte trénovanie modelu. Ak chcete niektoré strany z cvičných alebo overovacích dát odobrať, použite tlačidlo *x Remove*. Proces spracovania údajov a priebeh trénovania si môžete skontrolovať v zobrazení hlavného menu vpravo pod záložkou *Jobs*, podrobnejšie v *Open Full Jobs Table*.

6.3 Úspešnosť modelu a jeho zdokonaľovanie

6.3.1 Vyhodnotenie úspešnosti modelu

Po vytrénovaní modelu platforma ponúkne výsledok v podobe grafu a percentuálneho vyjadrenia chybovosti znakov v automaticky prepísanom texte. Dostupný je na záložke *Models* v ľavej časti základnej lišty, na ktorej sa zobrazuje zoznam dostupných modelov: vlastných a/alebo modelov, ktoré vám sprístupnili iní používatelia.

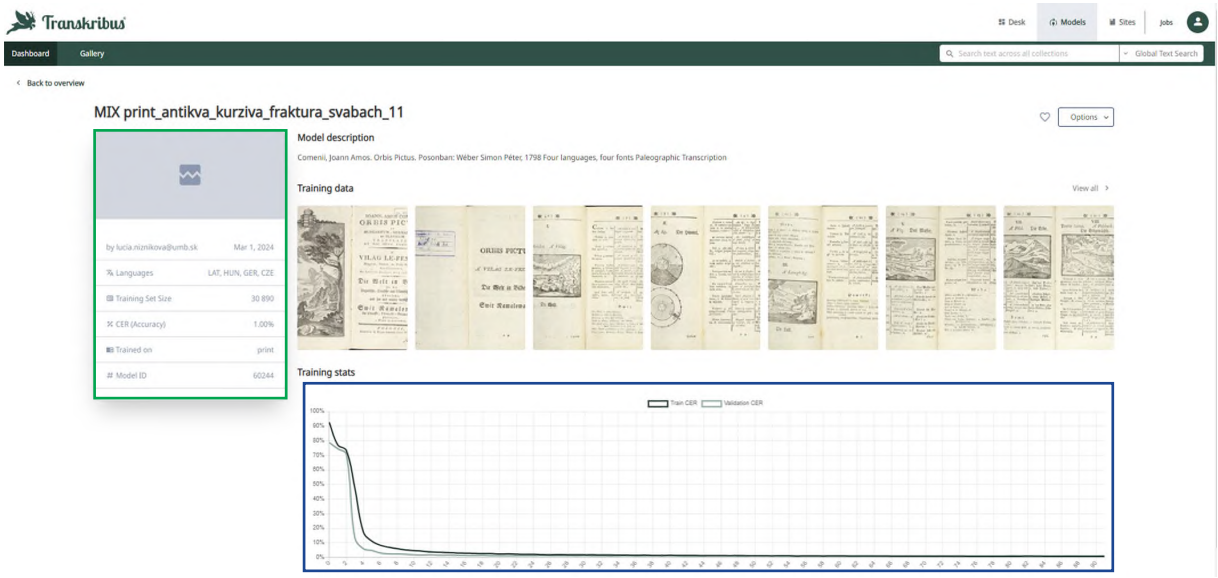


Obrázok 126 Náhľad stránky s dostupnými modelmi

Po kliknutí na možnosť *Show Description* príslušného modelu sa zobrazí vyhodnotenie, ktoré má dve časti.

Prvá časť obsahuje údaje o transkribovanom dokumente, ktoré ste zadali pred spustením modelu, a údaje, ktoré sa generujú automatizovane:

- názov a popis modelu (*Model description*),
- meno vlastníka a dátum vytvorenia modelu,
- jazyk/y dokumentu (*Languages*),
- počet slov v cvičnej vzorke, ktorý softvér prepočíta sám (*Training Set Size*)
- údaj o miere chybovosti znakov vyjadrený v percentách (*% CER Accuracy*)
- informáciu o tom, či bol model trénovaný na tlači alebo rukopisnom dokumente (*Trained on*),
- jedinečné identifikačné číslo modelu na platforme Transkribus (*Model ID*).



Obrázok 127 Údaje o modeli a jeho charakteristikách

Druhá časť vyhodnotenia (*Training Stats*) graficky zobrazuje proces trénovania, t. j. krivku učenia sa (tzv. *Learning Curve*). Čierna čiara predstavuje priebeh trénovania, zelená čiara predstavuje priebeh vyhodnocovania naučeného na overovacích dátach.

Os y zobrazuje **mieru chybovosti znakov CER** (*Character Error Rate*). Krivka sa bez použitia base modelu spravidla začína na 100 % a tým, ako sa model trénuje a zlepšuje, postupne klesá. Miera chybovosti znakov porovnáva celkový počet znakov (n) vrátane medzier s minimálnym počtom vložení (i), nahradení (s) a vymazaní (d) znakov potrebných na dosiahnutie rovnakého výsledku ako vo vzorke *Ground Truth*.

Vzorec na výpočet miery chybovosti znakov: CER = [(i + s + d) / n]*100

Každá malá chyba pri prepise je štatisticky plnohodnotnou chybou. To znamená, že chýbajúca čiarka, „u“ namiesto „v“, „á“ namiesto „ä“, medzera navyše alebo veľké písmeno namiesto malého sa počítajú ako chyby.

Os x predstavuje cykly, t. j. priebeh tréningu. Počas procesu trénovania vykonáva Transkribus vyhodnotenie po každom cykle. Pri nastavovaní parametrov na trénovanie modelu si môžete sami určiť počet cyklov. Čím je ich viac, tým dlhšie trénovanie trvá. Model na Obr 127 bol vytrénovaný pri počte 92 cyklov. V tomto prípade bol maximálny počet cyklov nastavený na 250, ale trénovanie sa automaticky zastavilo skôr, pretože model sa už nezlepšoval.

Informácie o vytrénovanom modeli dopĺňa prehľad strán originálneho dokumentu, ktoré boli zaradené do trénovania (*Training Data*).

6.3.2 Zdokonaľovanie modelu

Ak výsledné hodnoty chybovosti modelu nevyšli podľa vašich predstáv a zdá sa vám, že model sa dá vylepšiť, môžete využiť niektorý z týchto postupov.

Postup 1 Oprava manuálnej transkripcie

Pri podrobnej analýze automatickej transkripcie po vytrénovaní modelu možno zistíte, že softvér upozornil aj na vaše vlastné chyby pri manuálnej transkripcii. V tejto fáze môžete manuál-

ny prepis strán opätovne skontrolovať a opraviť prípadné nezrovnalosti. Vždy dbajte na to, aby strany vo vzorke *Ground Truth* boli prepísané presne a bezchybne, v opačnom prípade môže každý nesprávne prepísaný znak či akékoľvek odchýlenie od originálu negatívne ovplyvniť výsledok trénovania modelu.

Postup 2 Zvýšenie počtu slov v cvičných dátach

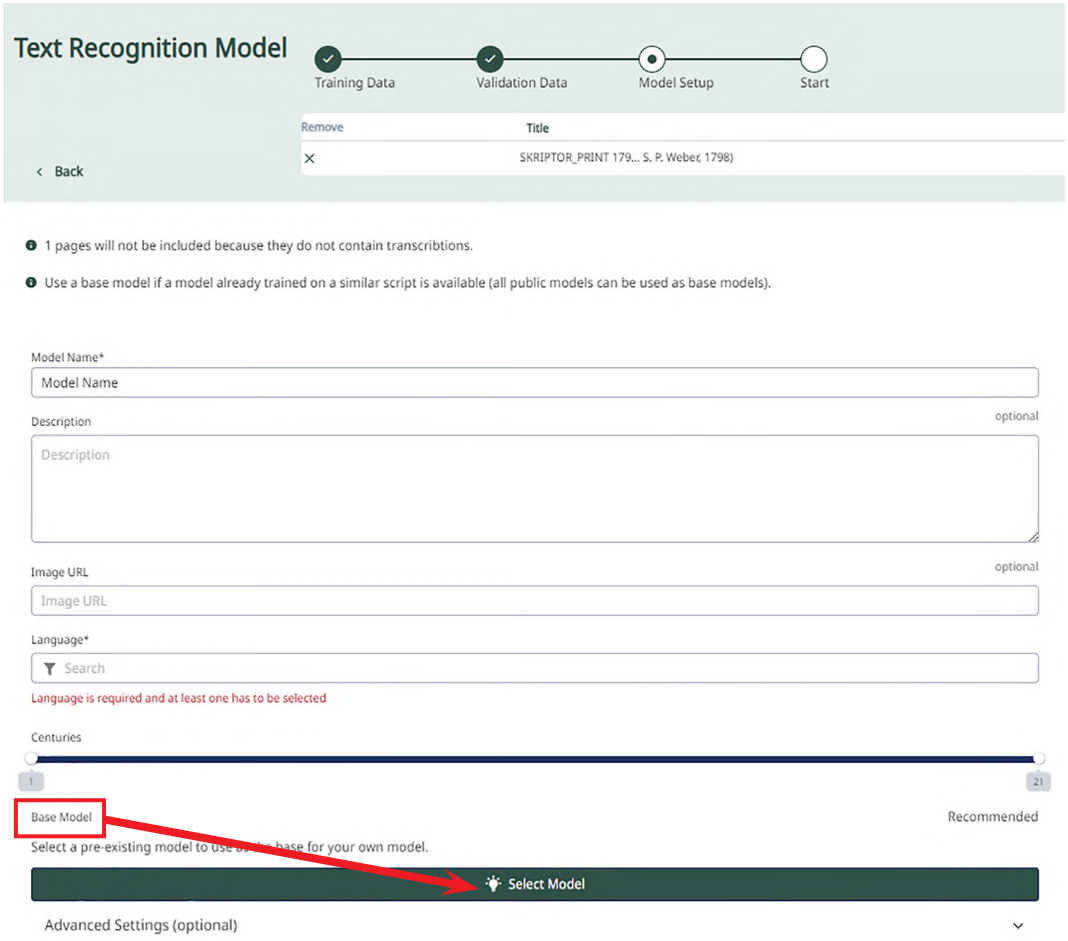
Odporúčaný minimálny počet slov zahrnutých do trénovania modelu je pre rukopisné texty 10 000 – 15 000 slov, v prípade tlačených textov približne 5 000 slov. V niektorých prípadoch toto odporúčané minimum nie je dostačujúce, preto je potrebné rozsah zväčšiť. Ide o metódu priebežného zvyšovania strán v cvičných dátach (*Training Data*) o automaticky transkribované strany (t. j. na základe aktuálnej verzie modelu). Automaticky prepísané strany sa následne opravia na úroveň *Ground Truth* a pripoja k predošlým, „bezchybne" pripraveným stranám, aby sa mohlo spustiť trénovanie nového, väčšieho súboru. Keďže vo väčšine prípadov platí pravidlo, že čím vyšší je výskyt znaku v cvičných dátach, tým lepšie sa ho stroj naučí rozpoznávať, týmto postupom môžete dosiahnuť zníženie miery chybovosti na úrovni znakov v overovacích dátach aj o niekoľko percentuálnych bodov.

Postup 3 Použitie základného modelu (*Base Model*)

Z doterajšej praxe sa ako najefektívnejší spôsob zdokonaľovania modelu javí využitie tzv. základného modelu (*Base Model*). Základný model si pamätá, čo sa naučil. Preto každé nové trénovanie (teoreticky) zlepšuje jeho kvalitu. Nový model sa učí od svojho predchodcu, a tak sa stáva lepším. Trénovanie pomocou základného modelu je preto mimoriadne vhodné pre veľké všeobecné modely, ktoré sa priebežne vyvíjajú počas dlhého obdobia. Zároveň umožňuje trénovať nový model s menšou vzorkou cvičných dát. Ak sa rozhodnete využiť základný model, máte na výber dve možnosti:

- použiť najlepšiu verziu vlastného modelu,
- použiť model iného používateľa, ktorý je verejne dostupný.

Ak chcete spustiť trénovanie nového modelu pomocou základného modelu, na záložke *Text Recognition Model* vo fáze nastavovania parametrov modelu (*Model Setup*) zvoľte v poli *Base Model* možnosť *Select Model*. Z ponuky vyberte základný model, ktorý bol trénovaný na dokumente s podobnými charakteristikami, ako je váš. Upozorňujeme, že výberu základného modelu od iného používateľa musí predchádzať dôkladná analýza postupov a metód, ktoré boli na jeho vytrénovanie použité. Ak vybraný základný model nebol trénovaný na podobnom type rukopisného alebo tlačeného fontu a nemá špecifické znaky/grafémy, ktoré obsahuje váš dokument, na trénovanie nového modelu, resp. zdokonaľovanie vášho pôvodného modelu je nevhodný.



Obrázok 128 Nastavovanie parametrov trénovania – výber základného modelu

Na tomto mieste je dôležité podotknúť, že na výsledné hodnoty vytrénovaného modelu vplýva viacero faktorov:

Faktory, ktoré môžete ovplyvniť:

- kvalita digitalizátu – preexponované alebo inak nekvalitné snímky nahraďte lepšími zábermi,
- charakter textov, ktoré sa počas trénovania rozhodnete vložiť do overovacích dát – či už ide o mieru ich reprezentatívnosti, kvalitu alebo počet znakov na príslušnej strane (napr. v poslednom prípade platí, že čím menej znakov na strane, tým väčšie percento chybovosti).

Faktory, ktoré nedokážete ovplyvniť:

- kvalita originálnej tlače – ak je originálna tlač nekvalitná, písmo nevýrazné (málo sýte), text obsahuje zásahy perom/ceruzkou (podčiarknuté riadky, škrtance, nadpísané slová a pod.), machule a iné nečistoty na papieri,
- pri rukopisných textoch platí dvojnásobne, že odchýlky v rukopise (napr. zmena štýlu písania, častý výskyt autorských korektúr, hromadné uvádzanie číselných údajov) môžu negatívne ovplyvniť výslednú úspešnosť modelu.

Preto je dôležité pri príprave vzorky *Ground Truth* zvoliť čo najtypickejšie, reprezentatívne a nepoškodené strany z rukopisu/tlače.

6.4 Supermodely

6.4.1 Výhody používania supermodelov

Supermodel je jeden veľký, veľmi všeobecný model so schopnosťou súčasne rozpoznať ručne písaný aj tlačený text. To môže byť užitočné najmä pri práci so zmiešanými materiálmi. Niektoré archívne fondy alebo zbierky rukopisov môžu mať rôzne typy písma, tlačené aj ručne písané dokumenty, strojopisy, predtlačené formuláre vyplnené ručne, kartotéky atď. So supermodelmi môžete použiť model na oba typy textu, čo znamená, že pri práci s ručne písanými aj tlačenými dokumentmi nepotrebujete rôzne modely ani nemusíte neustále meniť nastavenia.

Super Model Text Titan I. verejne dostupný v aplikácii Transkribus je pozoruhodne vhodný na spracovanie širokej škály materiálov a písomností. Aj keď dostupné supermodely nie sú momentálne jemne laditeľné alebo trénovateľné užívateľmi, poskytujú vynikajúci okamžitý výkon naprieč mnohými heterogénnymi typmi materiálov, čo používateľom pomôže rýchlo vytvoriť *Ground Truth* pre tréning prispôsobených modelov *PyLaia*.

Komunita Transkribus sa usiluje o tvorbu podobných supermodelov ako *Text Titan I.*, ktoré by boli čo najlepšie prispôsobiteľné pre špecifické potreby.

Špecializovaný model PyLaia trénovaný pre dobre definovaný materiál však môže stále prinášať lepšie výsledky, ale vytvorenie trénovacích údajov pre takýto špecializovaný model možno značne urýchliť tým, že časť materiálu najskôr spracujete pomocou supermodelu a opravíte ho manuálne.

Na Slovensku sme sa v rámci výskumu APVV v projekte *Skriptor* pokúsili o vytvorenie agregovaného modelu supermodelu zatiaľ samostatne pre rukopisy a osobitne pre tlače a strojopisy pre slovacikálne, resp. západoslovanské jazyky.

6.4.2 Slovenský supermodel M1 pre rukopisy (SSM1)

Dňa 24.04.2024 o 9,09 hod sme spustili v platforme Transkribus Expert verzia 1.27.0 vytvorenie nového agregovaného slovenského supermodelu. Základ pre tvorbu supermodelu pre určité slovacikálne rukopisy tvorili parciálne modely riešiteľov úloh v projekte *Skriptor*

Model má označenie *ID63569 Slovak Supermodel M1 (SSM1)*. Na tvorbu modelu sme použili 2583 strán v kvalite *Ground Truth* (GT) v počte 56713 riadkov a 333 777 slov. Z toho 1224 strán v cvičnom súbore (*Train set*) a 135 strán na overenie nového modelu (*Validation set*). Dokumenty na tento model sú v slovenčine, latinčine, maďarčine a češtine, resp. slovakizovanej češtine. Vstupné rukopisy mali rôznu kvalitu, pokiaľ ide o digitalizáty. Niektoré digitalizáty boli použité z digitálnych repozitárov spravidla v dobrej kvalite 600 dpi a niektoré boli výsledkom snímania pomocou zariadenia *ScanTent* a softvéru *DocScan* v dostatočnej kvalite 300 dpi.

Samotné ukážky považujeme pre ďalšie používanie a zdokonaľovanie modelu za veľmi dôležité, pretože ďalší výskumníci získajú predstavu o podobnom alebo odlišnom písme vlastných dokumentov, ktoré chcú transkribovať.

Pre supermodel boli dosiahnuté hodnoty *CER Train set*: 4,90 % na cvičnom súbore a *CER Validation set*: 5,30 % na overovacom súbore.

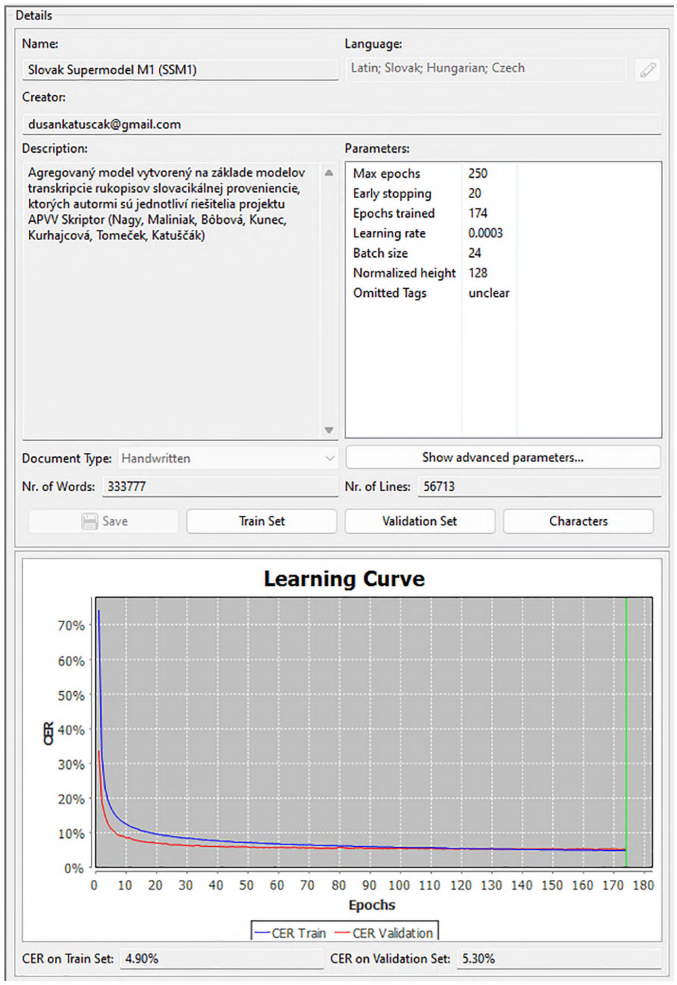
Tvorba modelu SSM1 na serveri Transkribus trvala dva dni, 5 hodín a 16 sekúnd, teda 53 hodín a 58 minút. Proces trénovania sa skončil po 174 cykloch.

Model SSM1 je na Slovensku a v Čechách prvým pokusom o tvorbu nástroja, prostredníctvom ktorého by bolo možné automaticky sprístupniť určité typy rukopisných dokumentov, ktoré sú podobné písmam použitým na jeho tvorbu.

V žiadnom prípade nemožno *ID63569 Slovak Supermodel M1 (SSM1)* považovať za definitívny univerzálny model transkripcie historických rukopisov slovacikálnej proveniencie všetkých typov a období. Varieta písiem a rukopisných štýlov je nekonečná a tvorba optimálneho agregovaného modelu predstavuje výzvu pre ďalších výskumníkov a entuziastov v nasledujúcich rokoch. Domnievame sa však, že náš prvý agregovaný model SSM1 môže uľahčiť automatickú transkripciu ďalších analogických rukopisov.

Výskumný tím plánuje sprístupniť datasety v rámci udržateľnosti projektu v rokoch 2024 – 2028 prednostne na výskumné a vzdelávacie účely pre inštitúcie a výskumníkov, ktorí budú chcieť prispieť k tvorbe modelu historických rukopisov v okruhu západoslovanských jazykov, resp. jazykov slovacikálnej a bohemikálnej proveniencie.

Takáto automatická transkripcia prirodzene neprinesie okamžité uspokojivé výsledky. Môže však uľahčiť „hrubú“ postupnú automatickú transkripciu ďalších strán, ich manuálnu opravu do stavu *Ground Truth* a následné použitie väčších datasetov na zdokonalenie nového modelu na báze nášho SSM1. Po vytvorení ďalších stoviek a tisícov strán bude možné pristupovať k tvorbe ďalších generácií nových modelov na základe SSM1. Vývoj by mohol pokračovať pre rukopisy modelmi nových generácií SSM2, SSM3 a pod. respektíve pre tlače a strojopisy (P-Print) ako supermodely SSP1, SSP2, SSP3 atď.



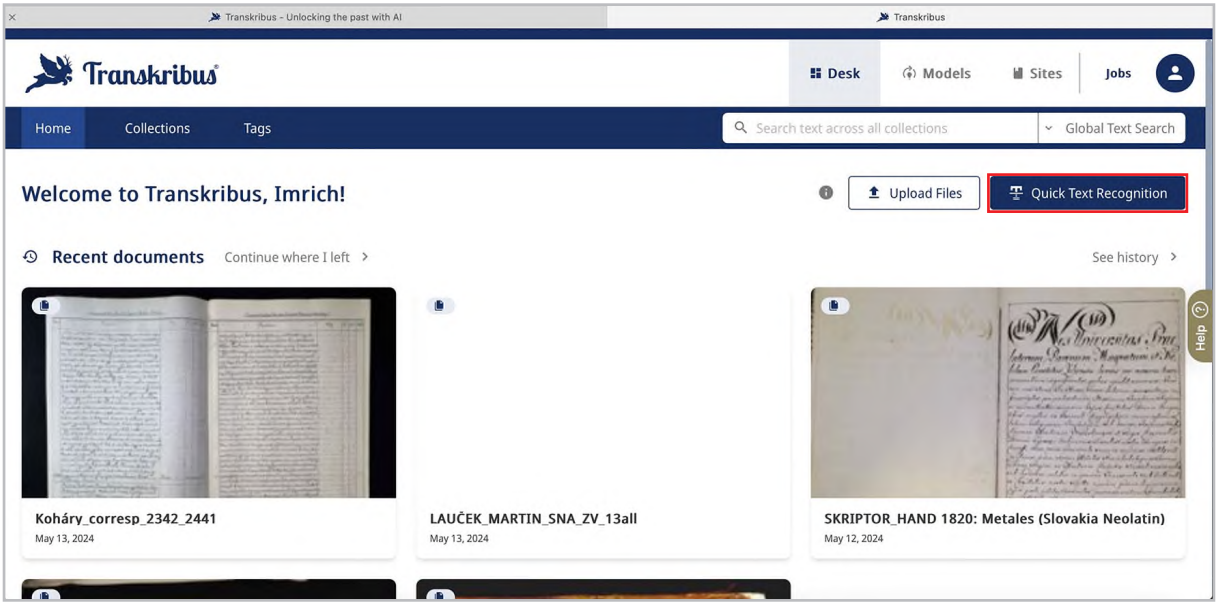
Obrázok 129 Charakteristiky supermodelu Slovak Supermodel M1 (SSM1)

7 Priebeh automatickej transkripcie v aplikácii Transkribus

Automatická transkripcia dokumentu je završením práce v aplikácii Transkribus, od ktorého sa očakáva výstup v podobe zrozumiteľného a všestranne použiteľného digitálneho prepisu. Pred samotnou realizáciou automatickej transkripcie si ešte raz skontrolujte, či ste vykonali všetky prípravné kroky:

- dokument máte digitalizovaný,
- digitalizáty ste importovali na platformu Transkribus,
- vykonali ste segmentáciu textu,
- máte model na automatickú transkripciu svojho dokumentu (vytrénovali ste si vlastný model, resp. chcete použiť adekvátny model z portfólia voľne dostupných modelov na platforme Transkribus).

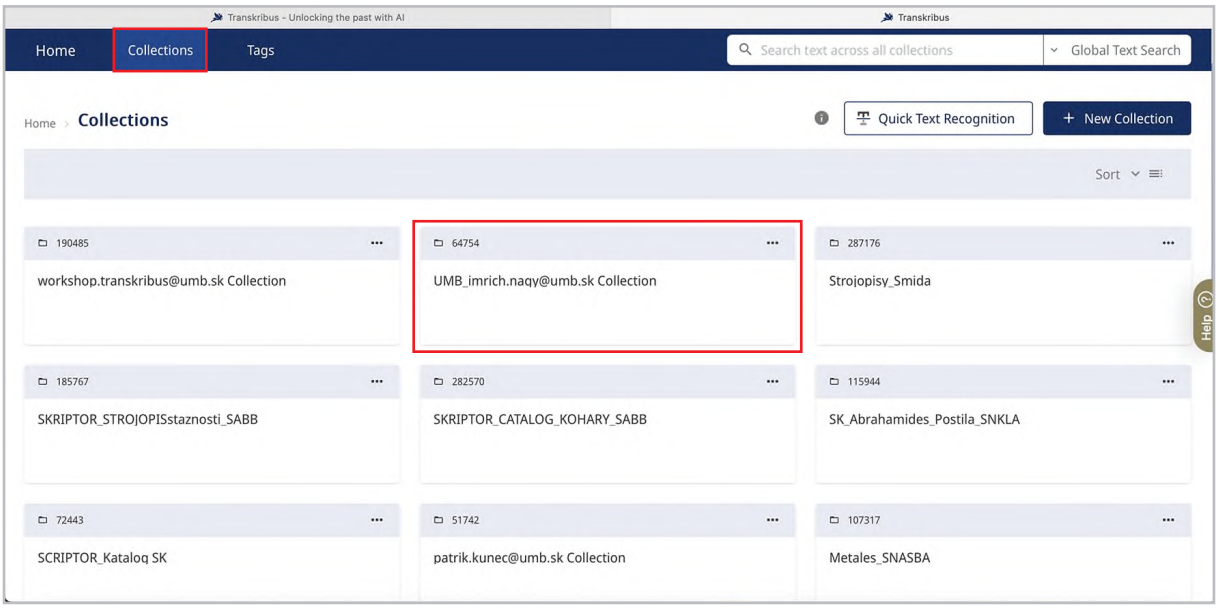
Poznámka: Transkribus umožňuje aj vykonanie tzv. rýchlej automatickej transkripcie hneď z úvodnej obrazovky (záložka Home, resp. Desk) kliknutím na tlačidlo rýchleho rozpoznávania (Quick Text Recognition) v pravej hornej časti obrazovky s preddefinovaným verejným modelom pre konkrétny jazyk. To znamená, že všetky kroky budú automatizované. Momentálne je však táto možnosť limitovaná obmedzenou dostupnosťou verejných modelov pre konkrétny jazyky a typy dokumentu.



Obrázok 130 Tlačidlo na prístup k rýchlej automatickej transkripcii na úvodnej stránke aplikácie Transkribus

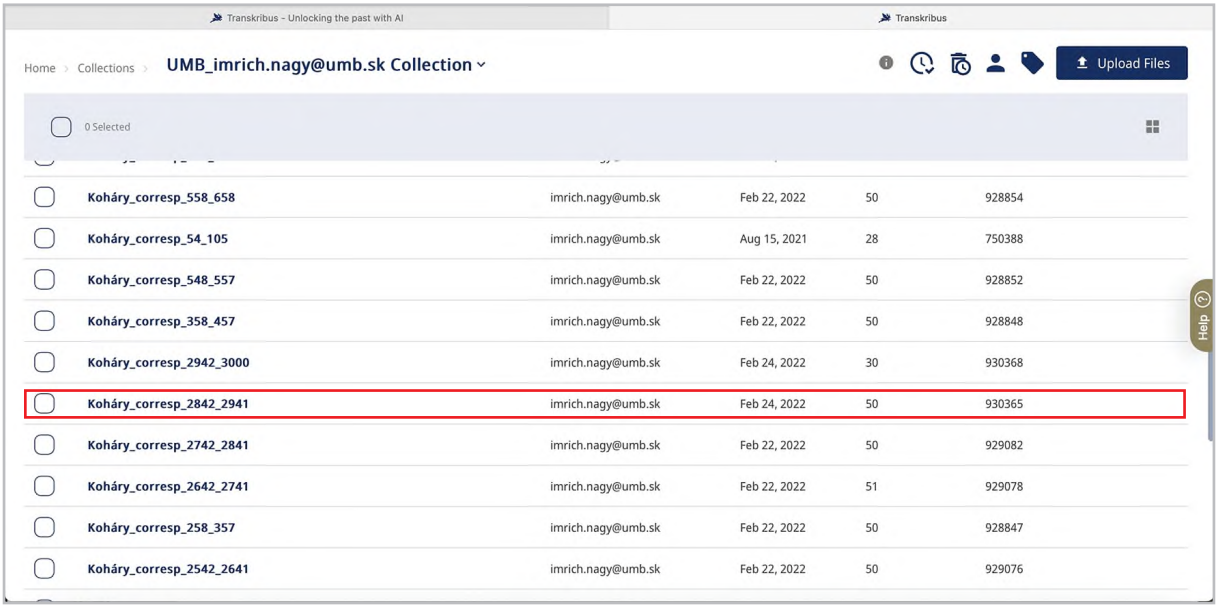
7.1 Výber dokumentu na automatickú transkripciu

Na hlavnej lište domovskej stránke otvorte záložku Zbierky (*Collections*) a zvoľte si zbierku, v ktorej sa nachádza dokument, ktorý chcete transkribovať.



Obrázok 131 Na záložke Zbierky (Collections) si vyberte zbierku, v ktorej máte uložený dokument na automatickú transkripciu

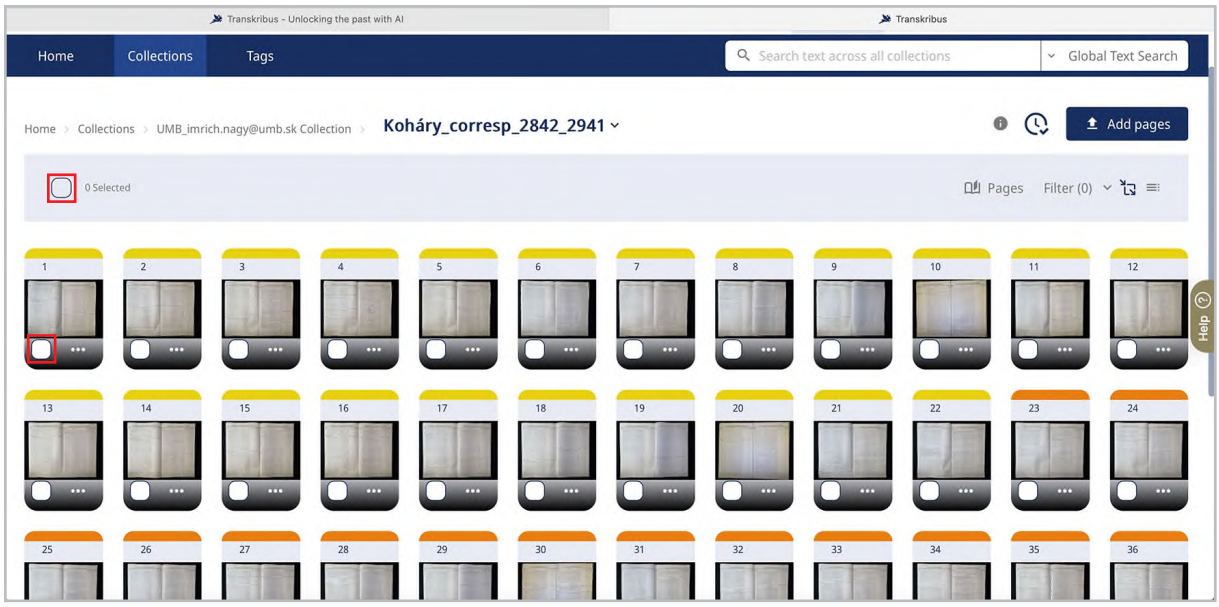
Zo zoznamu dokumentov v zbierke si zvoľte kliknutím dokument, ktorý už obsahuje strany s dokončenou segmentáciou.



Obrázok 132 Výber dokumentu na automatickú transkripciu zo zvolenej zbierky

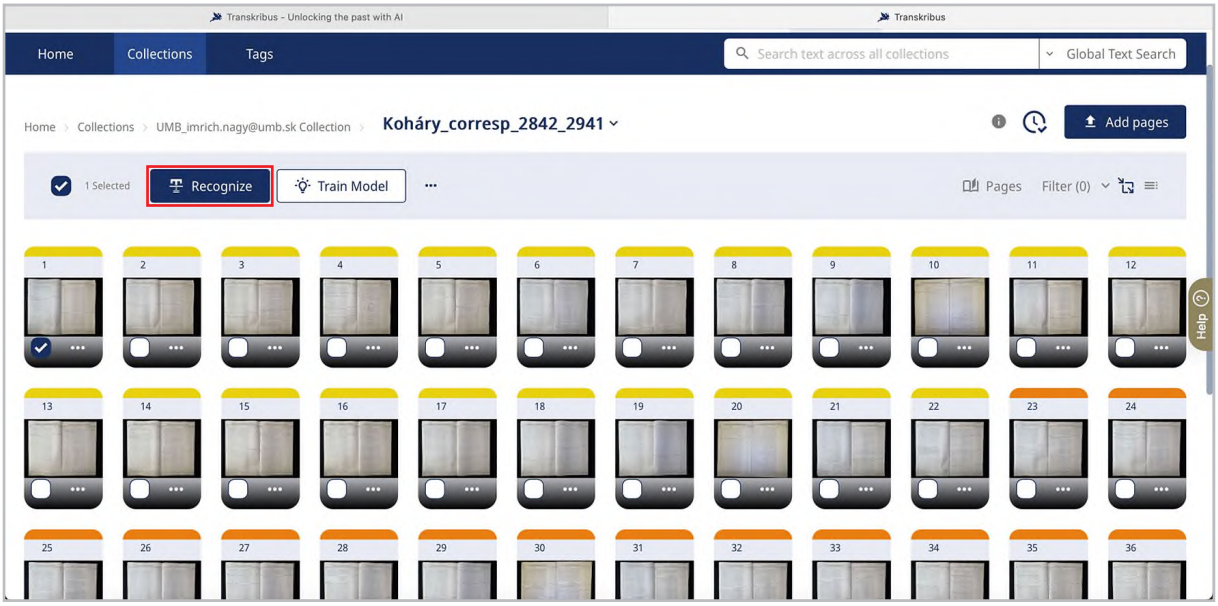
7.2 Výber snímok na automatickú transkripciu

Z ponuky snímok v dokumente si zvoľte konkrétne strany pripravené na automatickú transkripciu ich individuálnou voľbou (zakliknutím štvorčeka pod miniatúrou strany), resp. voľbou všetkých strán v dokumente (zakliknutím štvorčeka v lište nad miniatúrami). V orientácii vám môže pomôcť farebné zvýraznenie strán podľa stavu strany, napríklad ak stranu s ukončenou segmentáciou označíte stavom Hotovo (*Done*), je zvýraznená žltou farbou na rozdiel od strany, kde ste len začali robiť úpravy (*In Progress*), ktorá je zvýraznená oranžovou farbou.



Obrázok 133 Výber snímok (strán) dokumentu na automatickú transkripciu

Po označení strany (strán), ktoré chcete transkribovať, sa vám v hornej lište sprístupnia tlačidlá *Recognize*, *Train Model*, resp. tri bodky sprístupňujúce ďalšie voľby.



Obrázok 134 Zvolená strana na automatickú transkripciu a sprístupnené tlačidlo Recognize

7.3 Výber nastavení na automatickú transkripciu

Po kliknutí na tlačidlo *Recognize* sa vám otvorí nová stránka *Text Recognition*, ktorá ponúka nastavenia pre automatickú transkripciu.

V hornej lište máte sumarizáciu strán, ktoré ste zvolili na automatickú transkripciu. Pod ňou sa nachádza štvorček s možnosťou zakliknutia funkcie tzv. šikovného vyhľadávania (*Smart Search*), ktorou sa do vyhľadávania v prepise indexujú všetky možné variácie slov, t. j. aj tie, ktoré sa v prepise nenachádzajú. Pri základnom plnotextovom vyhľadávaní (*Fulltext search*) môžete vyhľadať len presné podoby slov, ktoré sa v prepise nachádzajú. Vďaka funkcionalite *Smart Search* teda môžete vyhľadať aj slovo, ktoré bolo napríklad prepísané nesprávne.

Táto voľba však nie je dostupná pri aplikácii tzv. supermodelov (modelov vytrénovaných na mimoriadne rozsiahlej a z rôznych aspektov univerzálnej cvičnej sade).

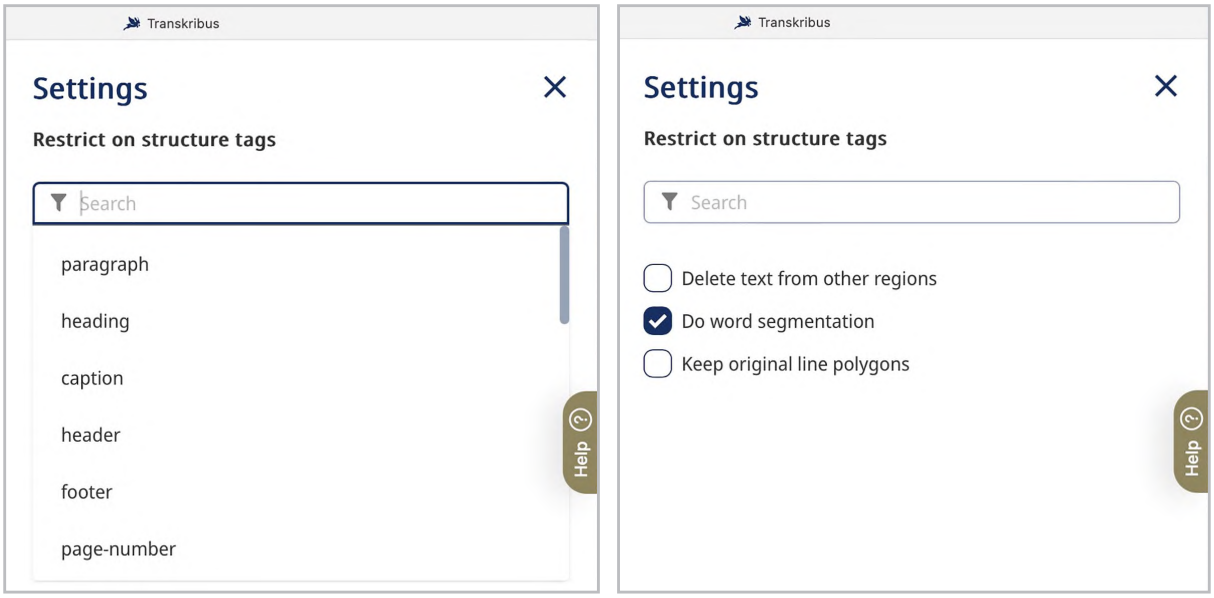
Okrem toho si môžete zvoliť aj aplikáciu jazykového modelu pri automatickej transkripcii. Voľba tejto možnosti sa odporúča najmä vtedy, ak máte dostupný špecifický model pre váš dokument.



Obrázok 135 Horná lišta na stránke Text Recognition s voľbami šikovné vyhľadávanie (Smart Search), jazykový model (Language Model) a pokročilé nastavenia (Advanced Settings)

Napravo od predchádzajúcej voľby máte možnosť upraviť pokročilé nastavenia (*Advanced Settings*). Aktuálne ponúkajú iba 4 možnosti:

- obmedziť transkripciu na vybrané časti strany, ktoré ste označili štrukturálnymi značkami/tagmi (*Restrict on structure tags*), ktorú dopĺňa možnosť vymazať text z iných rámcov (*Delete text from other regions*),
- vykonať segmentáciu slov (*Do word segmentation*) – táto voľba je predvolená a po vykonaní automatickej transkripcie doplní do snímky začiatky a konce slov. Umožňuje to spätnú kontrolu aj ďalšiu analýzu (napr. úspešnosti aplikácie jazykového modelu),
- zachovanie pôvodnej čiary polygónov (*Keep original line polygons*) je vhodné zvoliť, ak ste pri segmentácii robili jej manuálne úpravy.



Obrázok 136 Pokročilé nastavenia pre automatickú transkripciu. Vpravo: ponuka štrukturálnych značiek (tagov) pre voľbu obmedzenia transkripcie na vybrané časti strany

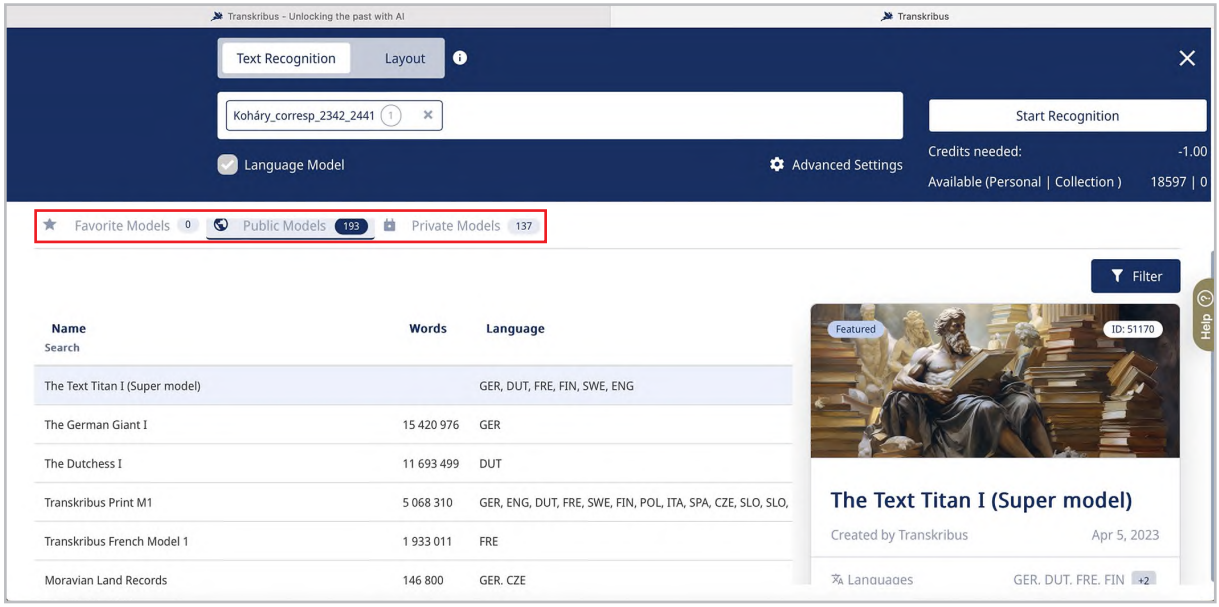
7.4 Výber modelu na automatickú transkripciu

Najdôležitejším krokom je výber vhodného modelu na automatickú transkripciu. Prehľad dostupných modelov je v dolnej časti stránky pre rozpoznávanie textu (*Text Recognition*), ktorá má osobitnú lištu na triedenie modelov:

Obľúbené modely (*Favorite Models*), ktoré ste si pre rýchlejšie vyhľadanie takto označili zakliknutím srdiečka na karte modelu.

Verejné modely (*Public Models*) sprístupnené ich autormi na využívanie celou komunitou registrovanou na platforme Transkribus.

Súkromné modely (*Private Models*) sú modely, ku ktorým máte prístup v rámci vášho konta (vaše vlastné modely, resp. modely vlastníkov iných zbierok, ku ktorým máte prístup).

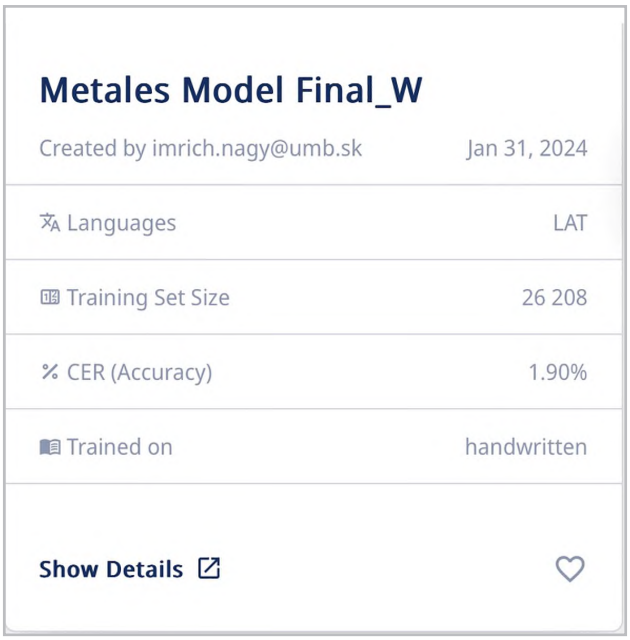


Obrázok 137 Lišta s ponukou modelov – obľúbené (Favorite Models), verejné (Public Models) a súkromné (Private Models)

Po zakliknutí modelu sa v pravej časti stránky zvýrazní jeho karta so základnými charakteristikami:

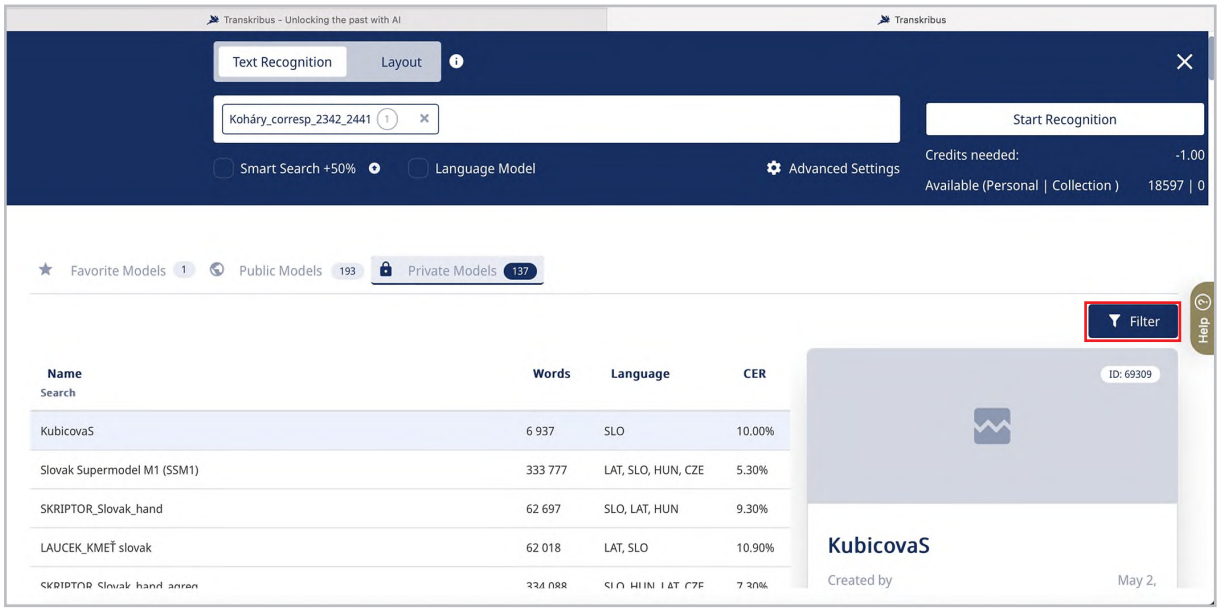
- autor a dátum vytvorenia modelu (*Created*),
- jazyk(y) modelu (*Languages*),
- rozsah cvičného súboru (*Training Set Size*),
- úspešnosť modelu (*% CER (Accuracy)*),
- typ dokumentu (*handwritten / print*).

V spodnej časti karty modelu ešte možno otvoriť jeho opis s grafom (*Show Details*) a napokon je v pravom dolnom roku voľba srdca pre zaradenie modelu medzi obľúbené (*Favorites*).



Obrázok 138 Karta modelu s jeho základnými charakteristikami

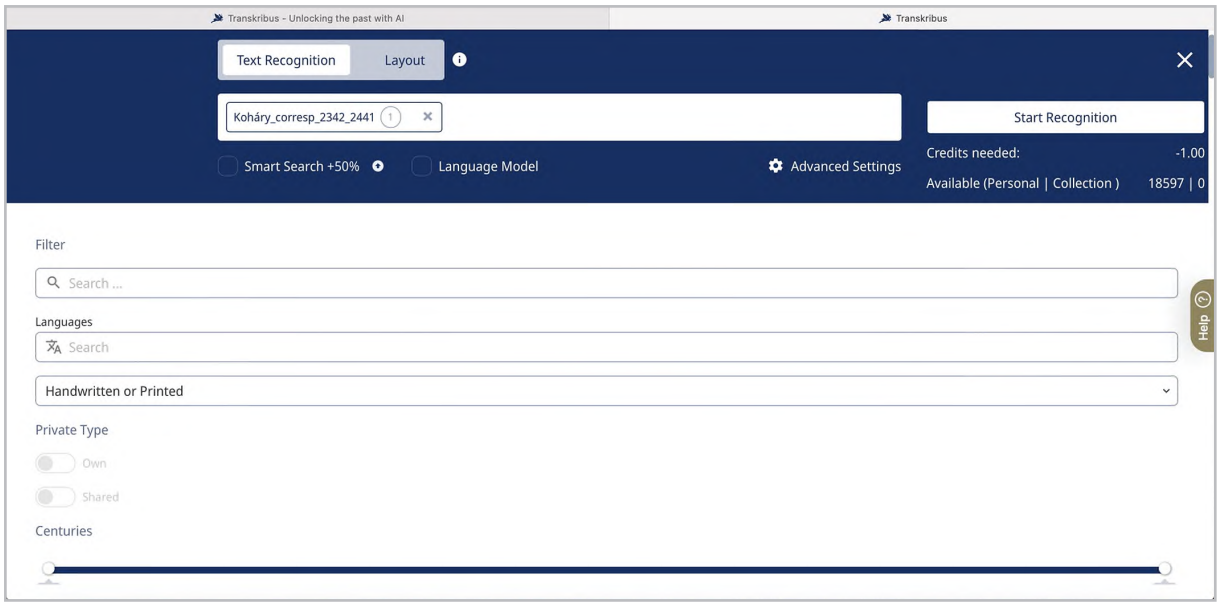
Pre uľahčenie výberu vhodného modelu (jeho rýchlejšie vyhľadanie) stránka ponúka aj filter, ktorý aktivujete kliknutím na tlačidlo (*Filter*) nachádzajúce sa nad kartou aktívneho modelu.



Obrázok 139 Tlačidlo na filtrovanie dostupných modelov

Filter ponúka možnosť vyhľadávať podľa:

- kľúčového slova,
- jazyka modelu,
- typu dokumentu (Handwritten / Print),
- typu súkromného modelu – vlastný (*Own*), resp. zdieľaný (*Shared*),
- storočia (*Centuries*), ktorým sa dá zúžiť časové obdobie zhodujúce sa so vznikom vlastného dokumentu.

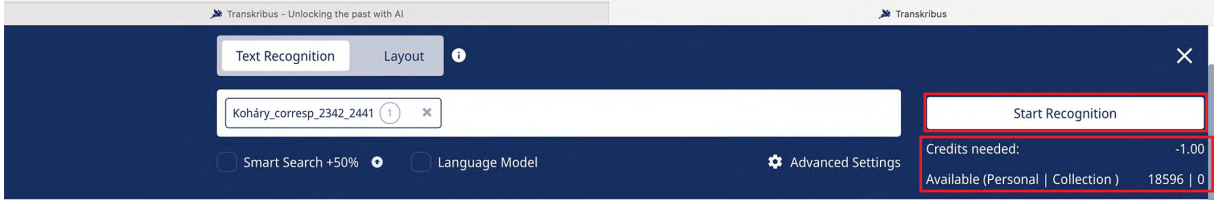


Obrázok 140 Voľby v ponuke na filtrovanie dostupných modelov

Voľba modelu rozhoduje o úspešnosti automatickej transkripcie. Preto je vhodné vytrénovať si vlastný model. Ak ho nemáte, môžete aplikovať jeden z univerzálnych, tzv. supermodelov (napr. *The Text Titan I.*) Pamätajte však na to, že model môžete aplikovať len na dokument, ktorý jazykovo podporuje.

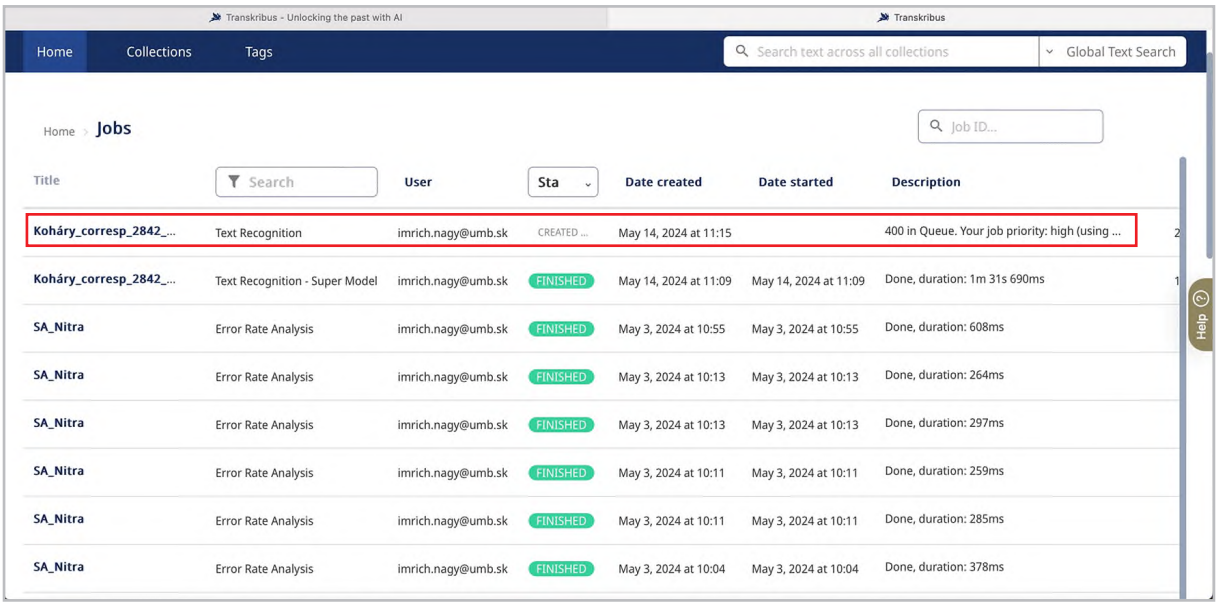
7.5 Spustenie automatickej transkripcie

Po voľbe modelu a ostatných nastaveniach môžete spustiť samotný proces automatickej transkripcie kliknutím na tlačidlo *Start Recognition* v pravej časti hornej lišty. Ide o **spoplatnenú akciu**, na čo vás upozorňuje **cena v kreditoch** (*Credits needed*) uvedená pod tlačidlom. Zároveň máte prehľad aj o stave konta vašich kreditov – osobných aj tých priradených k vašej zbierke *(Available (Personal / Collection)*).



Obrázok 141 Tlačidlo na spustenie automatickej transkripcie (Start Recognition), cena za transkripciu a aktuálny stav kreditov na konte

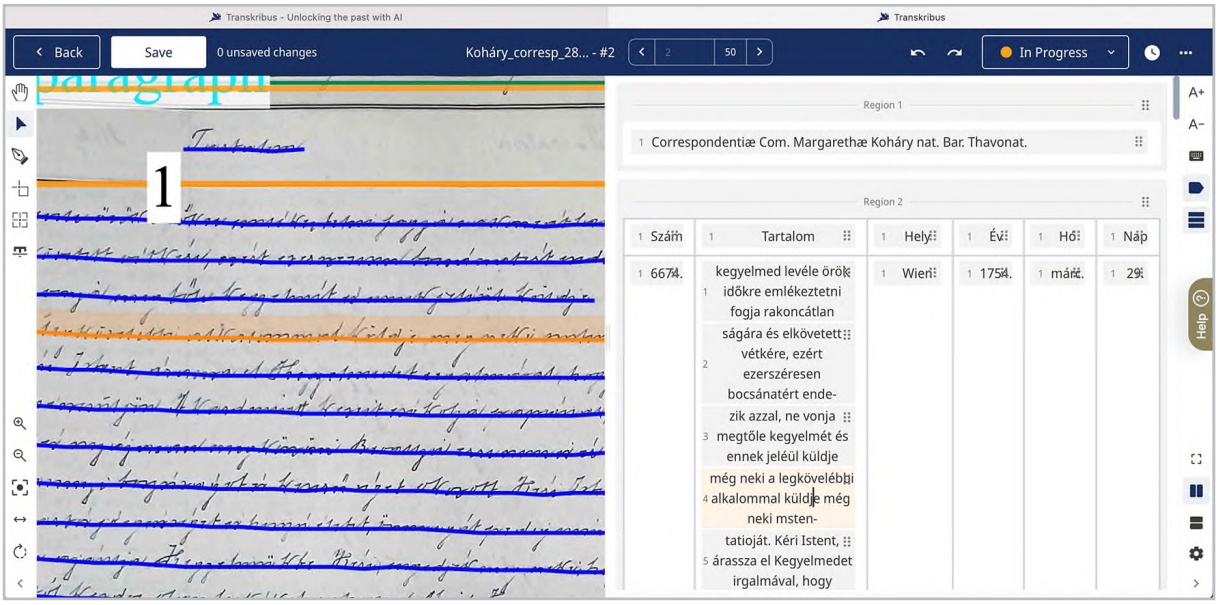
Po spustení akcie sa vaša obrazovka prepne na prehľad úloh, kde vidíte ich aktuálny stav. Skontrolovať si to môžete aj neskôr z domovskej stránky (*Home*) cez záložku *Jobs*. Operácia automatickej transkripcie, ktorú ste spustili, je uvedená v prvom riadku s príslušným statusom: *CREATED / PENDING / RUNNING / FINISHED* a základným popisom. Vaša požiadavka na automatickú transkripciu sa zaradí do poradia podľa aktuálne spracovávaných požiadaviek na serveroch platformy Transkribus. Čakacia doba na výsledok závisí od poradia a náročnosti jednotlivých operácií. Samotný proces automatickej transkripcie netrvá dlho – približne minútu na jednu snímku (v závislosti od dĺžky prepisovaného textu).



Obrázok 142 Prehľad úloh – v prvom riadku zaradená požiadavka na automatickú transkripciu

7.6 Výsledok automatickej transkripcie

Po skončení úlohy si môžete výsledok pozrieť a skontrolovať vyhľadaním príslušnej strany vo vašom dokumente (zbierke).

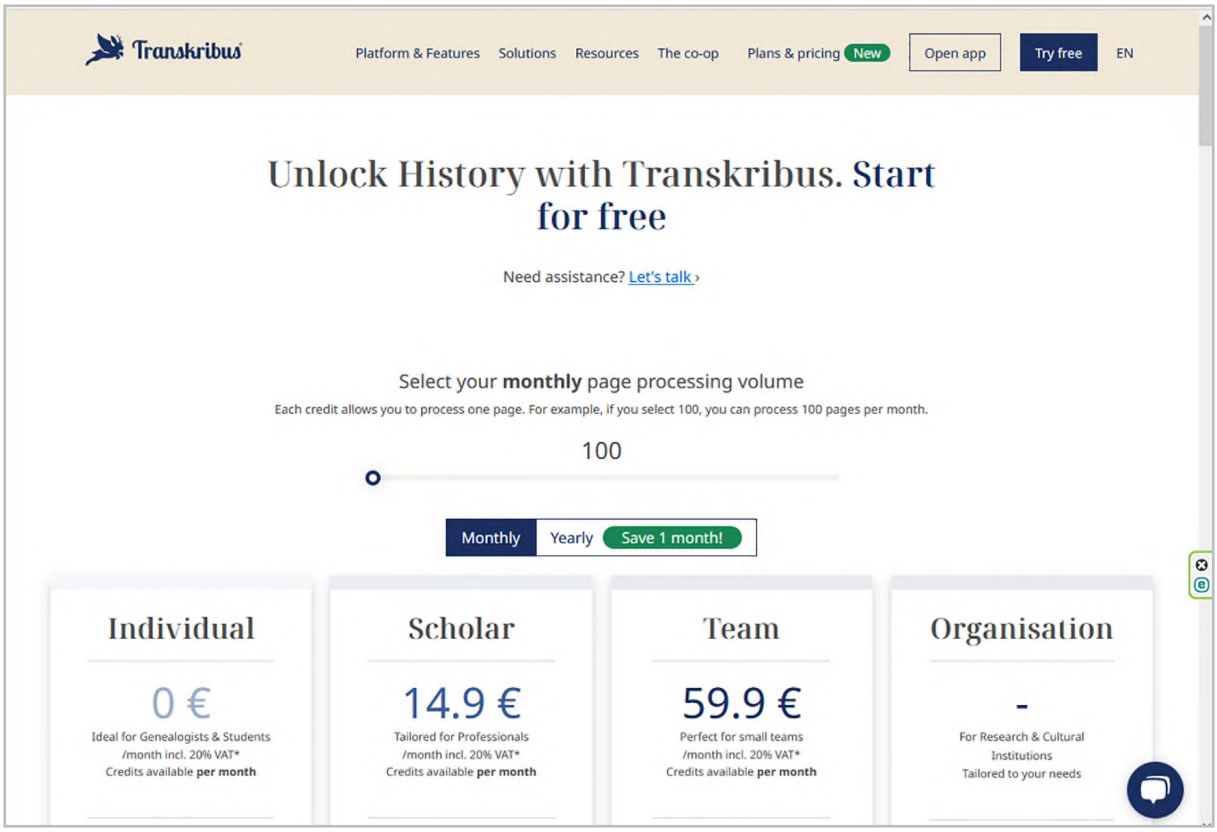


Obrázok 143 Výsledok automatickej transkripcie so zvýraznenými riadkami

7.7 Kontrola kreditov a systém spoplatnenia automatickej transkripcie

Ako sa uvádza vyššie, automatická transkripcia je spoplatnená formou kreditov, pričom cenová politika sa môže časom meniť. V súčasnosti je kreditmi spoplatnená segmentácia (0,25 kreditu / strana) aj transkripcia (0,5 – 1 kredit / strana). Pri registrácii si volíte jednu zo štyroch úrovní predplatného – individuálne predplatné je zdarma, máte však obmedzené počty úloh za mesiac a nemôžete využívať všetky funkcionality platformy (napr. verejné supermodely). V cene predplatného je zahrnutých 100 kreditov na mesiac. Ak potrebujete viac kreditov, musíte si ich

dokúpiť (predávajú sa v balíkoch od 1000 kreditov vyššie (cena za 1 kredit vychádza na 0,24 €). Bližšie informácie získate pri registrácii, resp. na webovej stránke: https://www.transkribus.org/plans.



Obrázok 144 Prehľad úrovní predplatného na platforme Transkribus

8 Možnosti práce s textom po automatickej transkripcii

Kapitola uvádza možnosti práce s prepísaným textom, ktorá ho zmení na dátovú základňu a export požadovaných obrazových či textových informácií, s ktorým chcete pracovať už v inom formáte alebo inom programe.

Text získaný po automatickej transkripcii a jej kontrole môžete obohatiť o dodatočné informácie. Spočíva vo vyčlenení významných údajov v rámci textu podľa nastavených kritérií. Uskutočňuje sa jeho označením zodpovedajúcimi tagmi (značkami).

Rozlišujeme dva základné typy tagov:

Textové tagy, ktoré definujú pojmy a frázy v texte a slúžia na označenie na úrovni oblasti, riadku, slova alebo aj jednotlivých znakov. Úpravy urobíte ľavým kliknutím na označený prepísaný text v textovom editore.

Štrukturálne tagy, ktoré definujú štruktúru dokumentu a sú založené na oblastiach textu a riadkov. Úpravy urobíte pravým kliknutím na digitalizovaný snímok v obrazovej časti pracovnej plochy.

S tagmi sa môžete stretnúť aj v predošlých krokoch práce na platforme, môžu pomôcť pri segmentácii, pri prepise i trénovaní modelu. Preto by ste mali zvážiť ich možné využitie už v úvode práce s textom a aké máte očakávania po získaní jeho prepisu.



Obrázok 145 Záložka na správu tagov

8.1 Textové tagy

Prepísané texty môžete obohatiť o textové tagy, ktoré podrobnejšie charakterizujú zvolený výraz alebo pomáhajú pri bližšej identifikácii neistých výrazov.

Na to slúžia atribúty tagov. Predstavujú vlastnosti tagu, poskytujú podrobnejšie informácie o obsahu tagu a teda o konkrétnom označenom výraze. Možno ich použiť na vyčlenenie a spracovanie údajov z prepisu (napr. dátum uvedený v dokumente obohatený o storočie, osobné meno rozšírené o dátumy narodenia a úmrtia osoby). Nie je potrebné vytvárať atribúty pre každý tag, závisí to od konkrétnych potrieb výskumu.

Platforma ponúka preddefinované tagy, s ktorými môžete pracovať ihneď alebo si môžete podľa potreby vytvoriť vlastné tagy. Práca s tagmi je možná až po priradení konkrétneho tagu k požadovanému výrazu.

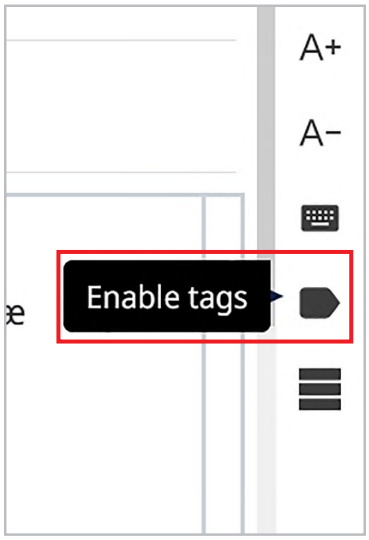
Rozlišujeme:

1. autoritatívne tagy (napr. osobné meno, geografické miesto, dátum, inštitúcia, abstraktná identita),
2. ostatné textové tagy (napr. skratky, nečitateľné výrazy, vymazaný text),
3. vlastné tagy.

8.1.1 Priradenie textového tagu

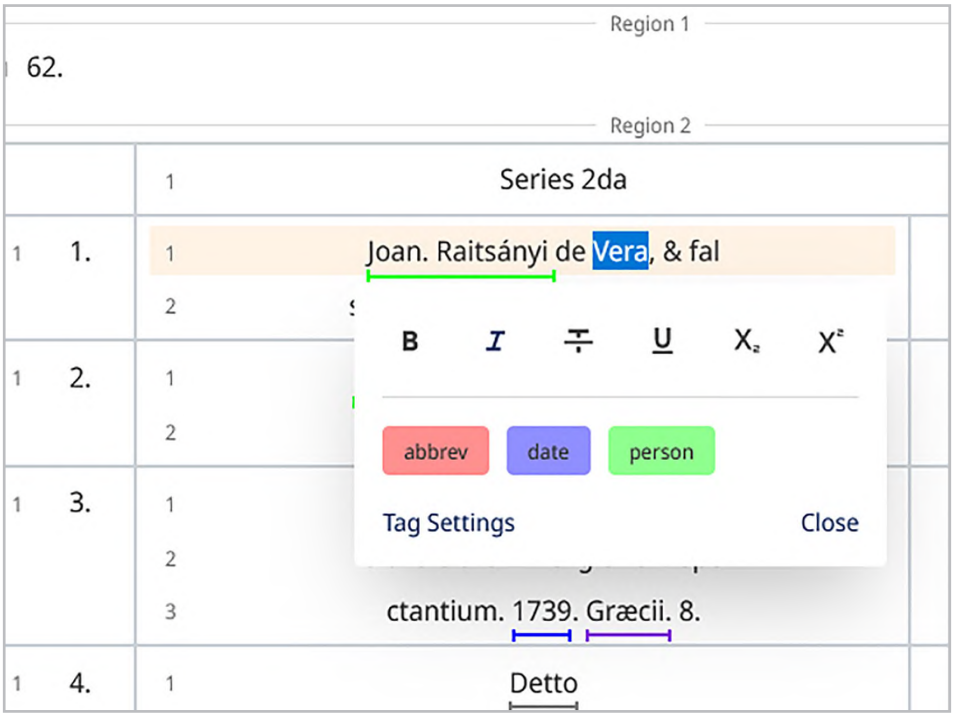
Textové tagy môžete použiť pre výrazy na úrovni oblasti, riadku, slova alebo aj jednotlivých znakov. Označujte však len nevyhnutné časti textu, ktoré majú byť vyhľadateľné. Každý tag sa používa samostatne na zvolený výraz, ale v prípade potreby je možné k rovnakému výrazu priradiť aj viacero tagov. Priradiť textový tag môžete už počas manuálneho prepisu dokumentu alebo po jeho automatickom prepise.

Na prácu s textovými tagmi je najprv potrebné povoliť zobrazovanie označovacieho okna. Kliknite na ikonu Povoliť tagy (*Enable tags*) umiestnenú v hornej časti pravej bočnej lišty.



Obrázok 146 Ikona na povolenie práce s textovými tagmi

Následne v textovom editore zvýraznite požadovaný výraz. Otvorí sa okno, v ktorom si buď len vyberiete vhodný tag alebo výraz v prípade potreby rozšírite aj o jeho atribúty.

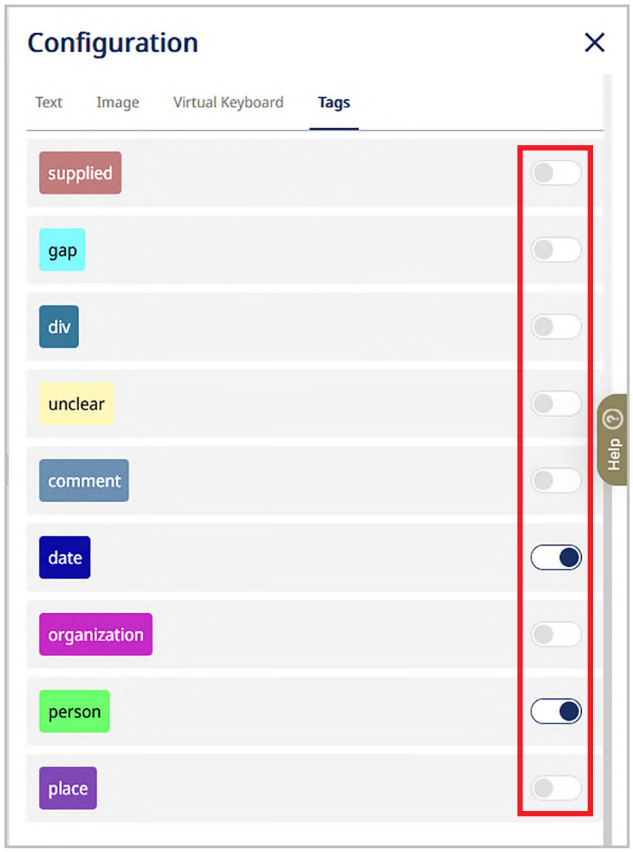


Obrázok 147 Možnosti výberu v zobrazovacom okne



Obrázok 148 Označený výraz rozšírený o atribúty

Môžete si vybrať, ktoré prednastavené tagy budete v okne vidieť. Viditeľnosť tagov nastavíte cez otvorené okno zakliknutím voľby Nastavenie tagov (*Tag Settings*) alebo cez Nastavenia (*Settings*) a konfiguráciu tagov (*Configuration Tags*) v spodnej časti pravej bočnej lišty. Vykonané zmeny sa aktualizujú po opätovnom načítaní webovej stránky.



Obrázok 149 Okno nastavenia viditeľnosti tagov

8.1.2 Ostatné textové tagy

Patria do skupiny preddefinovaných tagov, ktoré majú stanovené pravidlá použitia. Využívajú sa pri označovaní úprav textu daných dokumentom alebo dodatočných. V prameňoch sa z rôznych dôvodov vyskytuje aj nečitateľný text, ktorý sa nedá presne a dôveryhodne prepísať alebo čitateľný nemôže byť.

Príklad 1 Ak je pôvodný text preškrtnutý, ale stále čitateľný, prepíšte ho čo najvernejšie a dodatočne ho označte ako prečiarknutý pomocou tlačidla Prečiarknutie (*Strikethrough*) umiestneného v okne označovania. Použitie tagu je viditeľné jednak v zmene vzhľadu výrazu, jednak na pozadí k nemu platforma pridá textový tag.

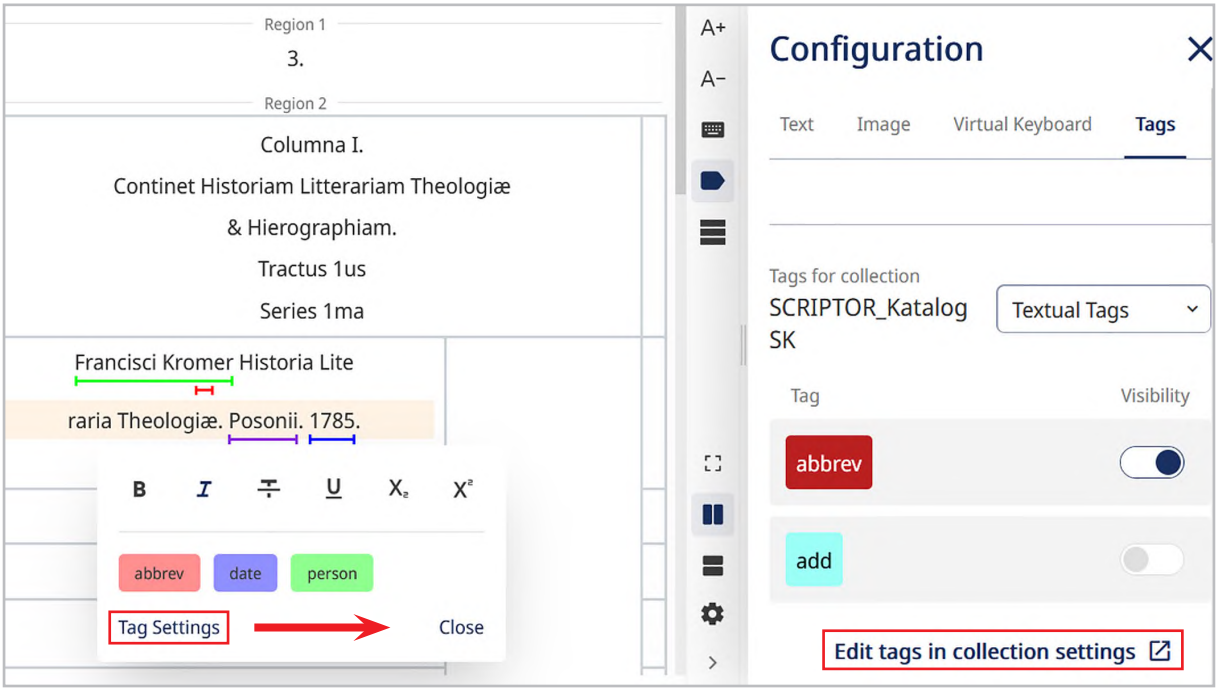
Príklad 2 Ak si nie ste istí správnosťou prepisu, prepísaný text označte tagom Nejasný (*Unclear*), aby ste sa ním mohli zaoberať neskôr. Môžete pridať aj alternatívy a návrhy pre nečitateľný výraz ako atribút tagu. Riadky s takto označeným výrazom nie sú zahrnuté do trénovania modelu.

Príklad 3 Ak je text úplne nečitateľný, také miesto označte tagom Medzera (*Gap*).

Príklad 4 V niektorých prípadoch sa dá nečitateľný znak alebo znaky uhádnuť a tak sa dajú jednoducho prepísať. Namiesto pridania obvykle používaných hranatých zátvoriek doplnený text označte tagom Nahradené (*Supplied*).

8.1.3 Vytvorenie textového tagu

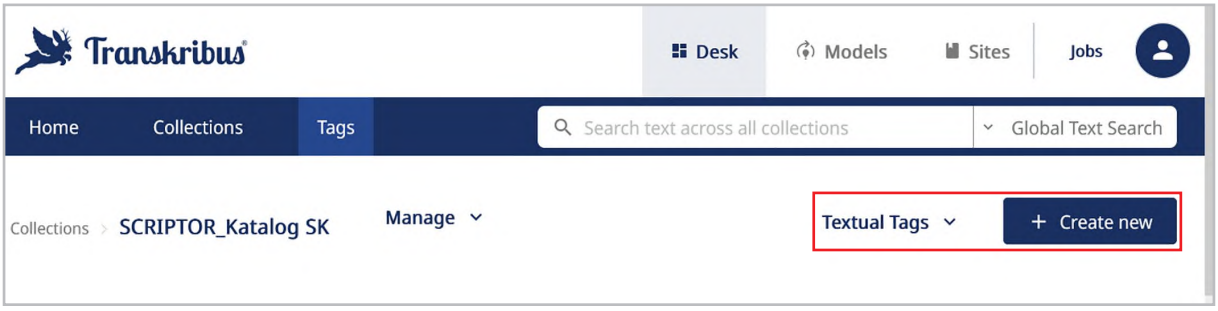
Okrem preddefinovaných, ihneď dostupných tagov môžete používať aj ľubovoľné vlastné tagy. Vlastný tag vytvoríte kliknutím na voľbu Upraviť tagy v nastaveniach zbierky (*Edit tags in collection settings*). Je umiestnená v Konfigurácii tagov (*Configuration Tags*), ktorá sa nachádza v ikone Nastavenia (*Settings*) v spodnej časti pravej bočnej lišty. Následne budete presmerovaní na samostatnú stránku záložky Tagy (*Tags*) v hornej základnej záložke *Desk*, cez ktorú spravujete tagy celej zbierky.



Obrázok 150 Presmerovanie k úprave tagov

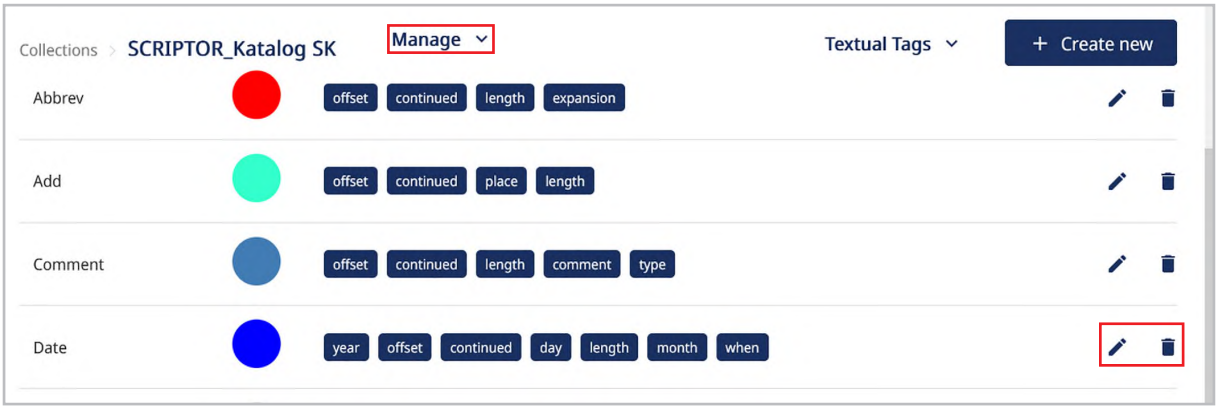
Kliknite na voľbu Vytvoriť nový tag (*Create new*), zadajte názov, vyberte farbu a možné atribúty a tlačidlom Vytvoriť (*Create*) uložte zmeny.

Zmeny sú v textovom editore viditeľné až po aktualizácii webovej stránky.



Obrázok 151 Tlačidlo na vytváranie nových tagov

Zakliknutím možnosti Spravovať (*Manage*) môžete robiť ďalšie úpravy tagov a ich atribútov, či už cez ikonu Upraviť (*Edit*) alebo cez ikonu Vymazať (*Delete*). Odstrániť tag môžete aj cez tlačidlo Odstrániť tag (*Remove tag*).



Obrázok 152 Ikony na úpravu tagov

8.2 Štrukturálne tagy

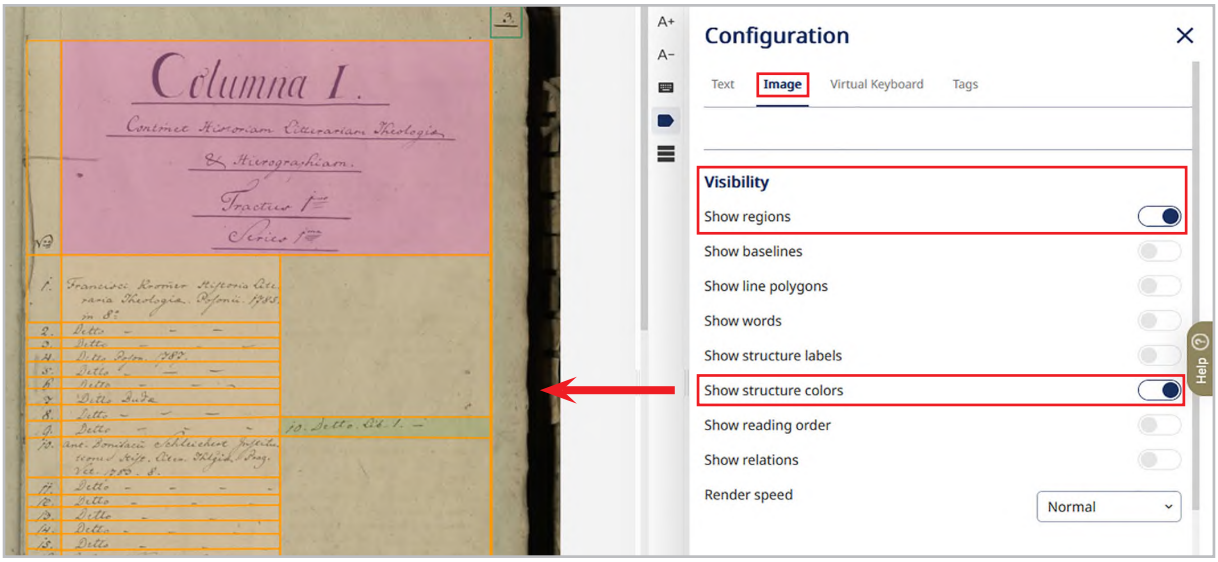
Prepísané texty sa dajú obohatiť aj o štrukturálne tagy (napr. odsek, nadpis, marginálie), ktoré umožňujú definovať štruktúru dokumentov. Je to doplnková možnosť, ktorú môžete využiť na označenie sekcií, ktoré vás zaujímajú (napr. vyčlenenie rôznych typov rukopisu v dokumente) alebo ak chcete obmedziť rozpoznávanie textu na určité typy štruktúry namiesto rozpoznávania celej stránky. Nie je potrebné označovať každý prvok dokumentu.

Rovnako ako textové aj štrukturálne tagy sú centrálne spravované na úrovni zbierky.

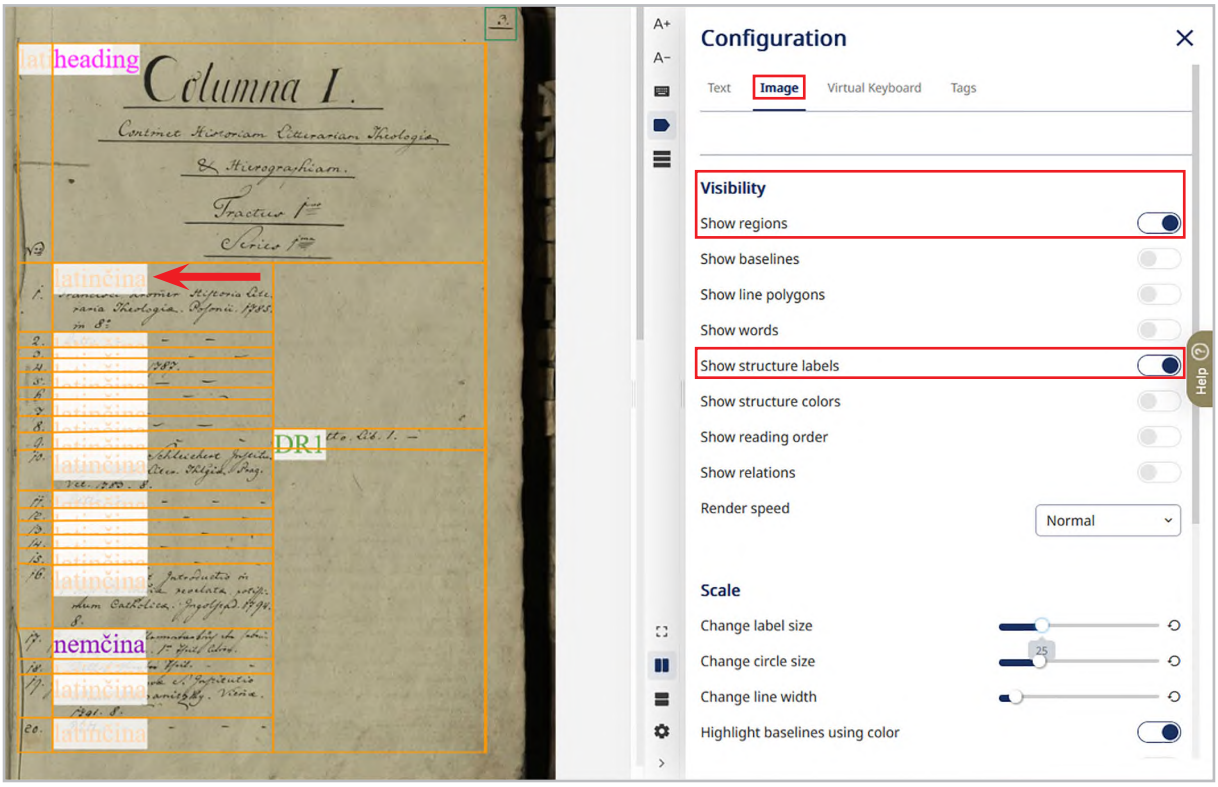
8.2.1 Zviditeľnenie štrukturálnych tagov

Na prácu so štrukturálnymi tagmi musíte povoliť zviditeľnenie ich označovania. Cez ikonu Nastavenia (*Settings*) umiestnenú v spodnej časti pravej bočnej lišty vyberte možnosť Obrázok (*Image*). Tu sa nastavujú pomocné prvky, ktoré sú viditeľné v obrázkovej sekcii pracovnej plochy. V časti Viditeľnosť (*Visibility*) najprv zakliknite Zobraziť oblasti (*Show regions*). Potom si vyberte spôsob zviditeľnenia. Prvá možnosť je cez textové označenie zakliknutím tlačidla Zobraziť označenia štruktúry (*Show structure labels*). Veľkosť označenia tagu si prispôsobíte

v časti Mierka (*Scale*). Druhou možnosťou je farebné rozlíšenie štrukturálnych tagov cez tlačidlo Zobraziť farby štruktúry (*Show structure colors*).



Obrázok 153 Označenie farebného rozlíšenia štrukturálnych tagov



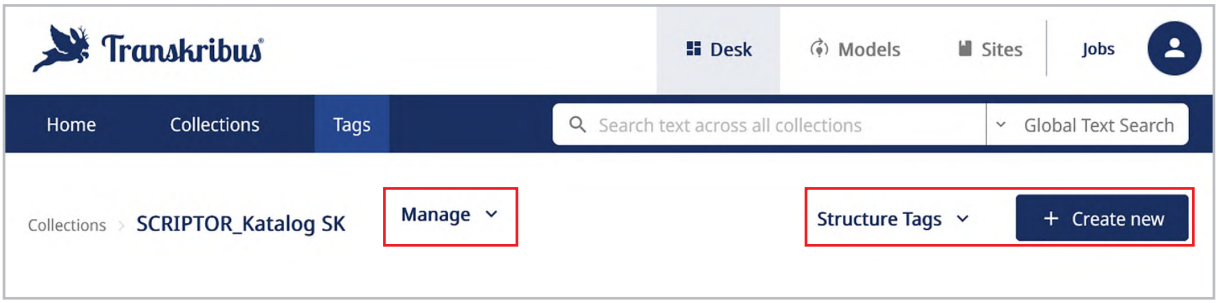
Obrázok 154 Označenie textového rozlíšenia štrukturálnych tagov

8.2.2 Správa štrukturálnych tagov

Ikona Nastavenia (*Settings*) a jej záložka Tagy (*Tags*) umiestnená v spodnej časti pravej bočnej lišty slúži len na možnosť výberu zviditeľnenia konkrétnych štrukturálnych tagov. Ďalšie úpravy sa uskutočňujú cez tlačidlo Upraviť tagy v nastaveniach zbierky (*Edit tags in collection settings*) umiestnené v spodnej časti záložky. Presmeruje vás na samostatnú stránku záložky Tagy (*Tags*) v hornej základnej záložke *Desk*, cez ktorú spravujete tagy celej zbierky.

Po prepnutí na Štrukturálne tagy (*Structural tags*) môžete pridať, upraviť alebo odstrániť vlastné štrukturálne tagy cez možnosť Spravovať (*Manage*). Predvolené tagy nie je možné vymazať ani upraviť, dá sa upraviť len ich viditeľnosť.

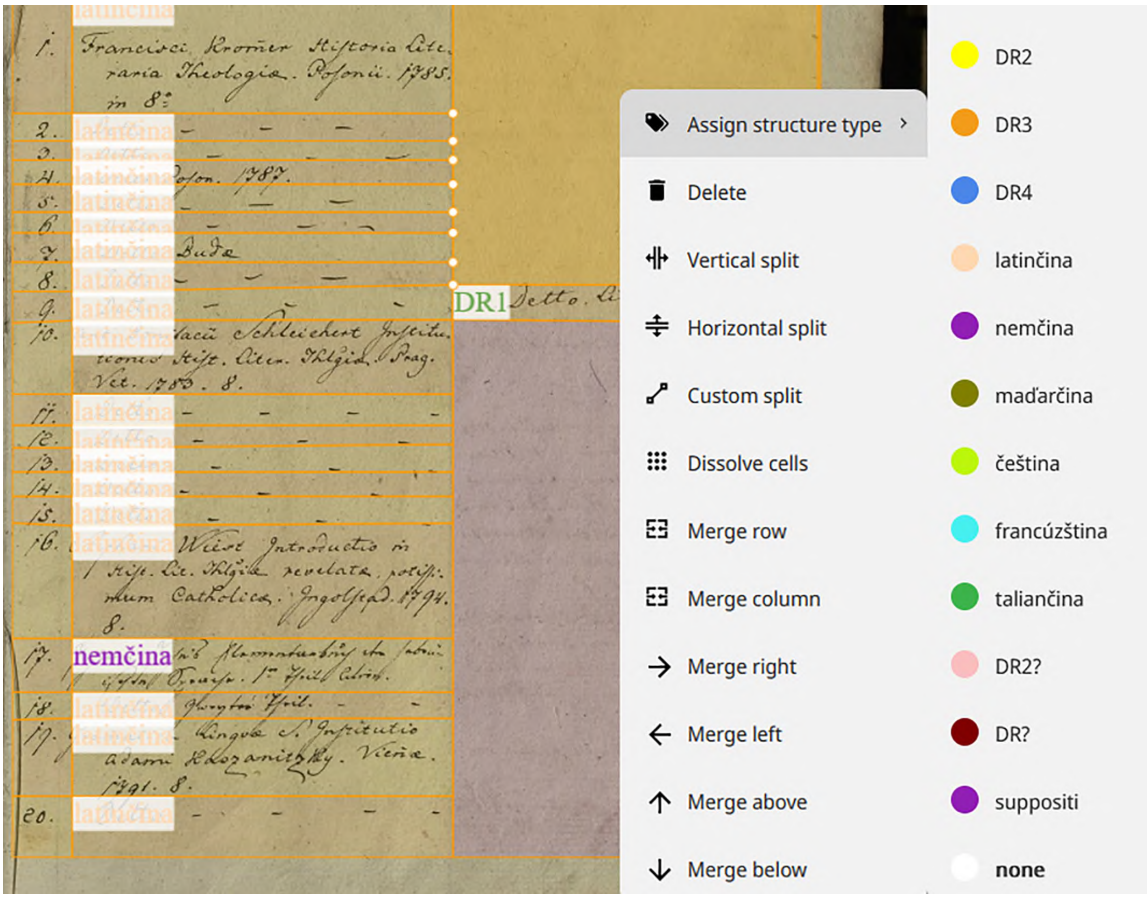
Nový štrukturálny tag vytvoríte kliknutím na tlačidlo Vytvoriť nový (*Create new*). Zadajte názov, vyberte farbu a zmeny uložte kliknutím na tlačidlo Vytvoriť (*Create*). Zmeny týkajúce sa názvu alebo farebného označenia môžete robiť aj v už existujúcom vlastnom štrukturálnom tagu cez ponúkané ikony Upraviť tag (*Edit tag*) a Odstrániť tag (*Remove tag*).



Obrázok 155 Vytvorenie nového štrukturálneho tagu

8.2.3 Priradenie štrukturálneho tagu

Štrukturálne tagy sa priraďujú k oblastiam textu a oblastiam riadkov. Vyberte požadovanú oblasť a kliknite na ňu pravým tlačidlom myši. Cez prvú položku nového okna Priradiť typ štruktúry (*Assign structure type*) si vyberte z vyrolovaných možností vhodný štrukturálny tag. Viditeľné sú iba štrukturálne tagy, ktoré ste tak označili v Nastaveniach (*Settings*).



Obrázok 156 Možnosti priradenia štrukturálneho tagu

Postup je rovnaký aj pri odstraňovaní štrukturálneho tagu z vybranej oblasti, ibaže z ponuky vyrolovaných možností vyberiete voľbu Žiadny (*None*).

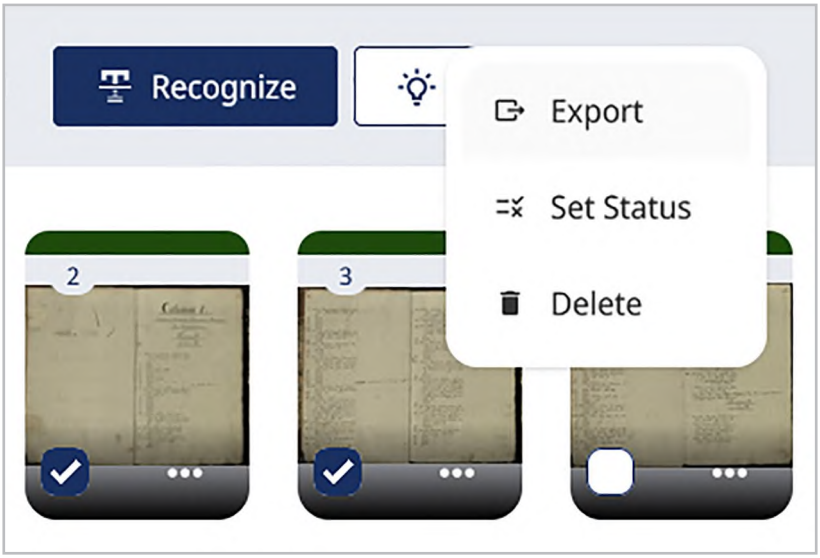
Rovnaký štrukturálny tag môžete naraz priradiť viacerým oblastiam. Najprv označte zvolené oblasti podržaním tlačidla CTRL na klávesnici a následne kliknite pravým tlačidlom myši a vyberte vhodný štrukturálny tag.

8.3 Export výstupov

So svojimi snímkami a prepismi môžete pracovať aj mimo platformy. Slúži na to export výstupov, ktorý umožňuje uloženie, publikovanie alebo ďalšiu analýzu prepísaných alebo pridaných údajov. Rôzne funkcie vám umožnia prispôsobiť výstup podľa formátu súboru a možností, ktoré uprednostňujete.

8.3.1 Štandardné možnosti exportu

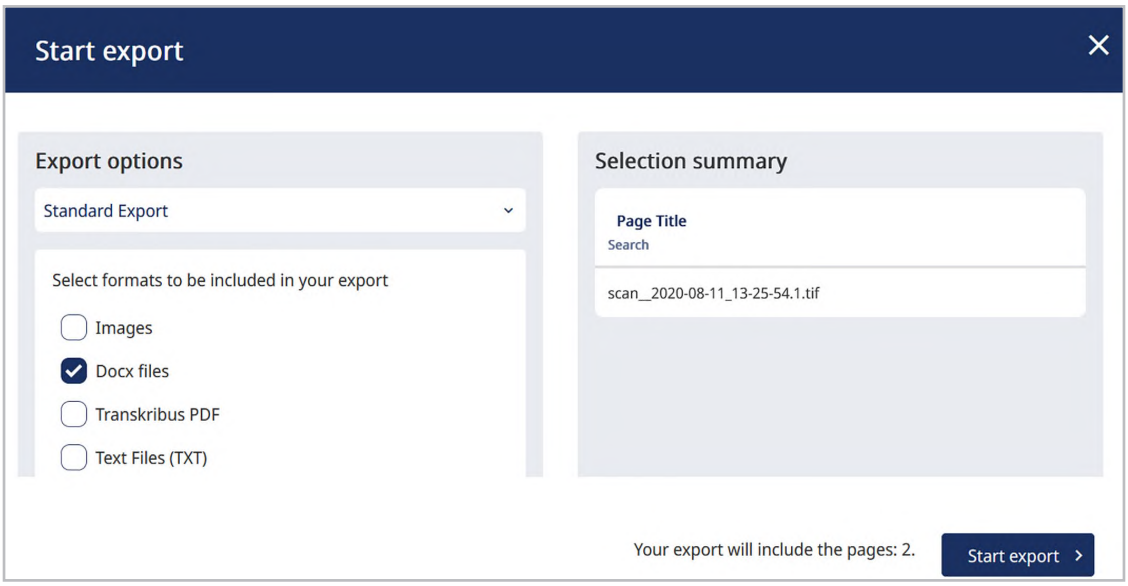
Označte dokumenty alebo snímky určené na export v prázdnom okienku stránky a zakliknite možnosť Export (*Export*) v trojbodkovej ponuke na hornom paneli s ponukami.



Obrázok 157 Okno pre export výstupov

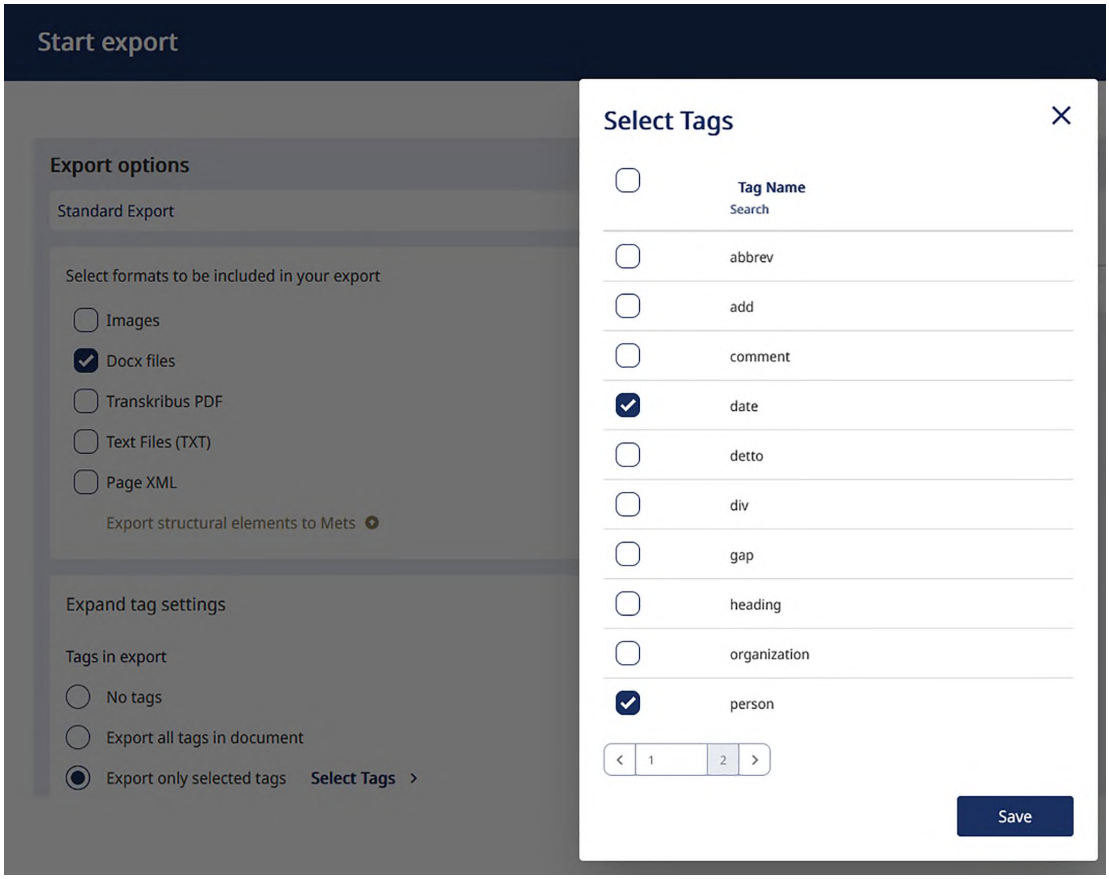
Následne v okne Spustiť export (*Start export*) vyberte Štandardný export (*Standard Export*) v Možnostiach exportu (*Export options*).

Potom vyberte z viacerých možností formáty, v ktorých si želáte výstup exportovať: obrázky/snímky (*Images*), súbory docx (*docx*), Transkribus PDF, textové súbory (txt) (*Text Files*) alebo XML stránky (*Page XML*).



Obrázok 158 Možnosti okna pre Spustiť export

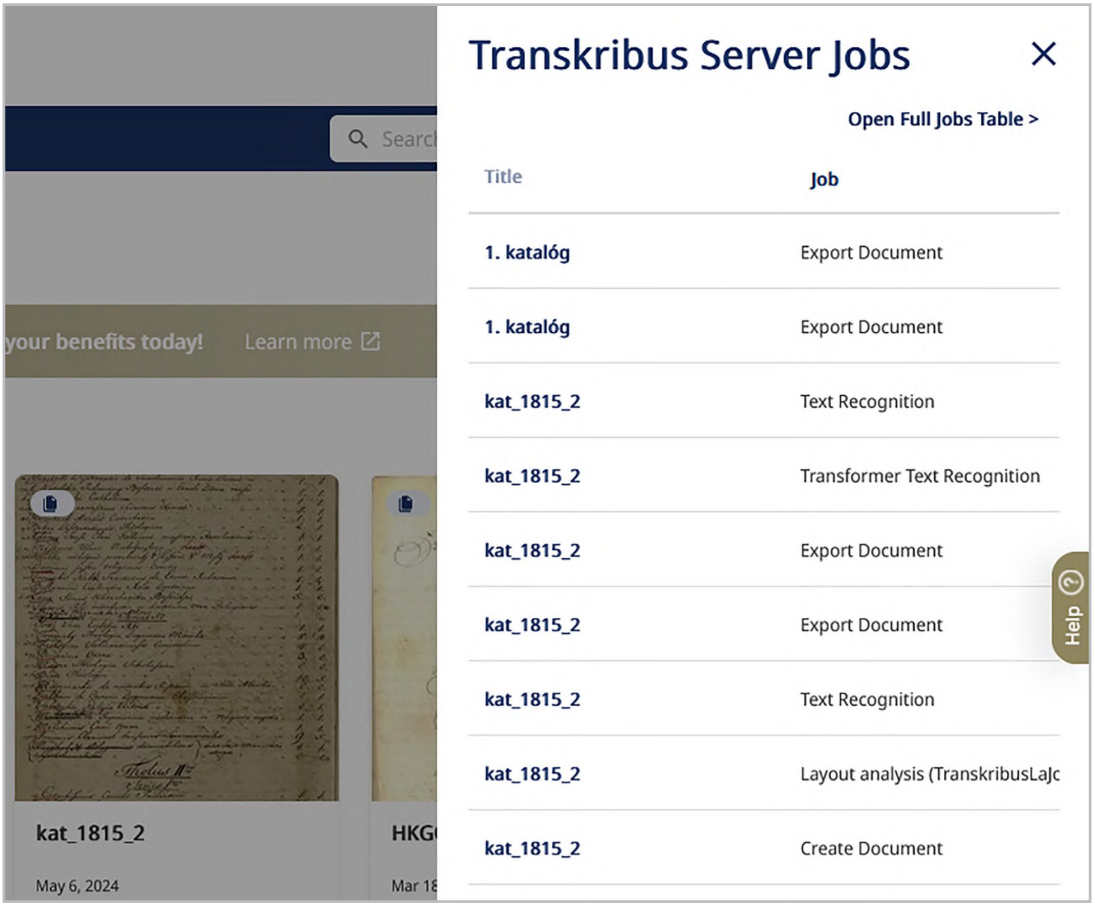
- Prostredníctvom Transkribus PDF exportujete dokumenty ako súbory PDF s vloženým textom.
- Prostredníctvom XML stránky exportujete súbory na ďalšie technické použitie alebo analýzu.
- V prípade výberu možnosti súborov docx a Transkribus PDF sa ponuka rozšíri aj o možnosť exportovať tagy, či už všetky alebo len konkrétne. Výber formátov je možné kombinovať.



Obrázok 159 Výber možnosti exportovať tagy

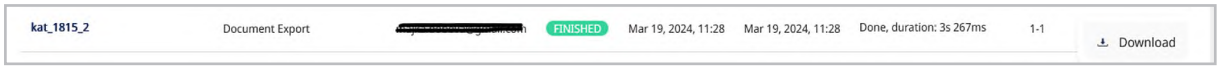
8.3.2 Prístup k exportovaným súborom

Export vybraných súborov sa spracováva na serveri platformy. Priebeh plnenia úlohy môžete sledovať na záložke Úlohy (*Jobs*) v hornom paneli.



Obrázok 160 Okno plnenia úloh

Po ukončení sťahovania súborov sa cez tlačidlo Otvoriť tabuľku úloh (*Open Full Jobs Table*) môžete prekliknúť na celú tabuľku úloh a stiahnuť si požadované súbory zakliknutím možnosti Stiahnuť (*Download*) v poslednom stĺpci Akcia (*Action*) vpravo. Zároveň dostanete e-mail s odkazom na stiahnutie súborov, ktorý je platný dva týždne.



Obrázok 161 Sťahovanie súborov

Slovník pojmov

Archívne fondy a zbierky. Historické rukopisné, prípadne strojopisné dokumenty na transkripciu sa nachádzajú prevažne v archívoch. Historické tlačené dokumenty sa nachádzajú najmä v knižniciach, ale aj u iných právnických alebo fyzických osôb. Na usporiadanie archívnych fondov sa u nás používa *Klasifikačná schéma archívnych fondov a zbierok štátnych archívov na Slovensku*. Na najvyššej úrovni majú archívy spravidla svoje zoznamy archívnych fondov a zbierok. Tieto zoznamy obsahujú všeobecné atribúty fondu a zbierky: názov archívneho fondu/zbierky, časové rozpätie, rozsah veľkosti archívneho fondu/zbierky v bežných metroch, prístupnosť a typ archívnej pomôcky. Výber konkrétnych dokumentov na transkripciu a výskum záleží na erudícii výskumníka, pretože rozsah a hĺbka spracovania fondov a zbierok sú rôzne.

CER (*Character Error Rate*). Miera chybovosti znakov porovnáva pre danú stranu celkový počet znakov (n) vrátane medzier s minimálnym počtom vložení (i), nahradenia (s) a vymazania (d) znakov, ktoré sú potrebné na získanie výsledku *Ground Truth*. Ide teda o chyby v porovnaní s presným, referenčným textom. Vzorec na výpočet CER: CER = [(i + s + d)/n]*100. Každá malá chyba v prepise je štatisticky plnohodnotná chyba. To znamená, že každá chýbajúca čiarka, „u“ namiesto „v“, dodatočná medzera alebo dokonca veľké písmeno namiesto malého písmena sú zahrnuté v CER ako chyba. Považuje sa za potvrdené a overené konštatovanie, že: a) ak je hodnota chybovosti znakov CER nižšia ako 10 %, čo je 10 a menej chýb na sto znakov, tak výsledok transkripcie je dobrý, čitateľný, a ak je to účelné, je možné ďalšie editovanie výstupu; b) ak je chybovosť znakov CER ≤ 5 %, tak výsledok transkripcie je veľmi dobrý; c) ak je chybovosť znakov CER pod 3 %, potom je možné považovať výsledky transkripcie za výborné a chybovosť znakov CER pod 2,5 % za excelentné.

Cvičné dáta (*Training Data*). Pozostávajú zo strán, na ktorých sa model trénuje. Na cvičnom súbore sa stroj „učí“, pri každom cykle „prečíta“ rovnakú stranu, pričom chybne prečítané znaky pri každom nasledujúcom cykle vyradí.

DocScan. Open source aplikácia pre Android navrhnutá pre ScanTent. Identifikuje strany dokumentu v živom náhľade a robí snímky v kvalite dostatočnej na transkripciu. V automatickom režime nasníma obrázok po otočení strany dokumentu. Umožňuje rýchlo snímať knihy alebo dokumenty bez interakcie s mobilným telefónom. Obrazovku smartfónu je možné zdieľať na obrazovke počítača a vzdialene ovládať smartfón napríklad cez TeamViever. Vďaka spoločnosti ifunplay a aplikácii DocScan možno ScanTent používať aj s operačným systémom iOS v iPhonoch. Držiak na vrchnej časti zariadenia ScanTent umožňuje umiestnenie smartfónu, optimálny pozorovací uhol a konštantnú vzdialenosť. Ak denné svetlo nestačí, biele LED pásiky poskytujú rovnomerné osvetlenie, ktoré maximalizuje kvalitu obrazu.

Dokument (*Document*). V štruktúre systému Transkribus je dokument zvyčajne zaradený do zbierky. Dokument môže byť presunutý do inej existujúcej zbierky. Základné metadáta k dokumentu sú: jedinečný číselný identifikátor, názov dokumentu, meno osoby, ktorá nahrala dokument do zbierky v Transkribuse, dátum a čas nahratia do zbierky, názov zbierky, do ktorej dokument patrí. Dokument je možné zobraziť vo forme Prehľad (*Overview*) s jednotlivými stranami a grafickým rozlíšením stavu stránky (napr. *In progress*, *Done*, *Final, Ground Truth*). V záložke Rozloženie (*Layout*) sú viditeľné texty transkripcie strán, riadky textu, poradie čítania riadkov strojom, identifikátor riadka a koordináty umiestnenia elementov v riadku.

Export. Ak chceme pracovať s obrázkami a prepismi mimo platformy Transkribus, môžeme svoje dokumenty exportovať do bežnejších formátov, ako sú docx, PDF, xls, PageXML, TEI-XML alebo txt. Možnosti zahŕňajú export celých strán, obrázkov, textu alebo štrukturálnych

prvkov. Exportovať je možné do adresára v lokálnom počítači alebo exportovať na server Transkribusu, z ktorého dostanete oznámenie po skončení exportu.

Formát JPG, JPEG. Najrozšírenejší je formát, ktorý sa vyskytuje s príponou .jpg, .jpeg. V tomto formáte ukladajú súbory všetky fotoaparáty aj mobilné zariadenia, ak používate napríklad DocScan. V niektorých fotoaparátoch je možné voliť jeden formát alebo snímanie v dvoch formátoch JPG a RAW (ARW). Výhodou formátu JPG je, že obrázok sa dá zobraziť prakticky v každom zariadení – v mobilnom telefóne, televízore alebo vo webovom prehliadači. Zaberá málo miesta na disku, je úsporný, pretože ide o kompresiu so stratou. Nevýhodou tohto formátu je, že obrázok každou úpravou stráca kvalitu pri každom uložení. V projektoch transkripcie používame na snímanie mobilnými zariadeniami formát JPG na archivovanie a v transkripcii spravidla pracujeme s derivovaným formátom PDF.

Formát PNG. Skratka v preklade znamená prenosná sieťová grafika (*Portable Network Graphics*), čiže ide o bezstratový kompresný formát pre obrázky a fotografie využívaný najmä na internete.

Formát RAW. znamená, že nasnímaný súbor je „surový", nespracovaný a dáta nie sú komprimované. Dáta v tomto formáte sú veľmi veľké a na ich spracovanie je potrebný špeciálny softvér, napríklad komerčný Zoner Photo Studio alebo open source FastStone Image Viewer. Výsledné obrázky majú vysokú kvalitu a po úprave sú vhodné na kvalitné editovanie.

Formát TIFF. Vyskytuje sa s príponami .tiff, tif. Pri ukladaní do tohto formátu spravidla nedochádza ku kompresii dát. Ak áno, tak ide o bezstratovú kompresiu aj pri opakovanom ukladaní. Súbor zachováva maximum informácií z formátu RAW pri editácii. Nevýhodou je veľkosť súborov vo formátoch TIFF. V profesionálnych projektoch digitalizácie je formát TIFF najvhodnejší na dlhodobé archivovanie.

Formáty obrázkov. Snímky je možné tvoriť, ukladať a upravovať v rôznych formátoch. Najčastejšie ide o súbory vo formátoch RAW a JPG. Z hľadiska úprav fotografií je dôležitý formát TIFF.

Gotické písmo malo niekoľko druhov. Napríklad francúzska textúra s veľmi ostrým lomom a štíhlou stavbou, talianska širšia a okrúhlejšia rotunda s miernejším lomením oblúkov, zmiešané písmo – bastarda, v Nemecku švabach – písmo širších, oválnejších tvarov a fraktúra – písmo užších a špicatejších tvarov s ozdobnými úponkami. Vynálezom kníhtlače (v roku 1450 Johannom Gutenbergom) sa tento druh písma veľmi rozšíril najmä v krajinách hovoriacich po nemecky.

Ground Truth (základná pravda). Vzorka manuálne prepísaných a dôsledne skontrolovaných a korigovaných strán dokumentu, ktorá sa používa pri trénovaní modelu automatickej transkripcie.

Import dokumentov (*Upload*). Po vytvorení zbierky v Transkribuse je potrebné nahrať dokumenty. Potom je možné spustiť nástroje, ako sú analýza rozloženia (segmentácia) alebo rozpoznávanie textu (transkripcia). Údaje v Transkribuse sú vždy súkromné a prístupné iba jednotlivým používateľom. Vlastník zbierky (*Owner*) môže umožniť prácu aj iným používateľom (*Users*) s oprávneniami, ktoré im pridelí (*Owner, Editor, Transcriber, Reader*).

ISAD(G) (*General International Standard Archival Description*). Medzinárodný štandard, ktorý definuje zoznam prvkov a pravidiel na popis archívov a popisuje druhy informácií, ktoré musia a mali by byť zahrnuté v takýchto opisoch. Vytvára hierarchiu popisu, ktorá určuje, aké informácie by mali byť zahrnuté na akej úrovni. V súvislosti s výskumom a experimentmi s transkripciou archívnych dokumentov považujeme za vhodné, aby boli transkribované fondy,

zbierky a dokumenty popísané na štandardnej úrovni. Tento štandard poskytuje rámec pre spoločný prístup a nie rigidný formát.

Model. Na platforme Transkribus je model entita, ktorá je výsledkom použitia softvéru strojového učenia a umelej inteligencie a hlbokých neurónových sietí na rozpoznávanie historických rukopisných a tlačených textov. Platforma Transkribus umožňuje používateľom trénovať model rozpoznávania textu rukou (PyLaia) na automatické spracovanie zbierky dokumentov. Model je potrebné trénovať tak, aby rozpoznal určitý štýl písania zobrazovaním obrázkov dokumentov a umožnil ich presný prepis. Podľa typu textu môžu používatelia na transkripciu použiť verejne dostupný model alebo vytvoriť vlastný model, prípadne trénovať vlastný model s použitím základného modelu.

OCR (*Optical Character Recognition*). Optické rozpoznávanie znakov alebo optická čítačka znakov je elektronická alebo mechanická konverzia obrázkov ručne písaného alebo vytlačeného textu na strojovo kódovaný text či už z naskenovaného dokumentu alebo fotografie.

Overovacie dáta (*Validation Data*). Pozostávajú zo strán dokumentu, na ktorých sa presnosť vytrénovaného modelu automaticky overí (odskúša). V porovnaní s cvičnými dátami sú preto menšie, spravidla 10 % z celkovej vzorky *Ground Truth*. Na druhej strane overovacie dáta by mali byť reprezentatívne, t. j. mali by obsiahnuť príklady všetkých písmen, jazykov a iných atribútov zahrnutých v cvičnom súbore. V opačnom prípade, čiže ak sú overovacie dáta príliš homogénne, výkon modelu môže byť nízky, prípadne skreslený.

Polygóny (*Polygons*). Historické dokumenty majú niekedy zložité usporiadanie a pozostávajú z rôznych rozložení, čo môže viesť k problémom s poradím čítania prvkov textu. Pri komplikovaných rozloženiach si rýchlo všimneme, že ručne nakreslené textové oblasti sa môžu prekrývať. Tento problém sa dá ľahko vyriešiť úpravou pravouhlých oblastí textu pridaním bodov a tým vytvorením polygónov.

Poradie čítania. V systéme Transkribus poradie čítania zobrazuje na segmentovanej stránke to poradie, v ktorom bude stroj transkripcie čítať riadky textu na obrázku stránky. Toto poradie čítania sa vytvára automaticky počas segmentácie, ale možno ho neskôr zmeniť aj manuálne. Pri automatickej analýze rozloženia je poradie čítania určené súradnicami riadkov na obrázku: horný riadok, ktorý je najviac vľavo, je číslo jedna, atď. Dôležité je vedieť, že poradie čítania nie je relevantné pre samotné trénovanie, ale môže sťažovať čítanie transkribovanej strany. Ak sa má prepis exportovať a ďalej použiť na vydanie, tak poradie čítania je potrebné zadať správnym spôsobom, aby bol text v správnom poradí. Dá sa to urobiť jednoducho zapnutím poradia čítania funkciou Viditeľnosť tvaru (*Shape visibility*).

Presnosť modelu. Presnosť modelu sa vypočíta automaticky pri trénovaní modelu ako pomer chybne prečítaných znakov – udáva sa v % CER.

Princípy popisu ISAD(G) sa riadia štyrmi všeobecnými zásadami: 1) Opis od všeobecného ku konkrétnemu – viacúrovňový opis sa začína od všeobecnej úrovne opisu, ktorá je zvyčajne fondmi, a pokračuje do podrobnejších úrovní, ako sú podfondy, séria, súbor, položka, atď. Táto hierarchická štruktúra musí byť reprezentovaná a správne definovaná v archívnom opise. 2) Informácie relevantné pre úroveň opisu – informácie na každej úrovni opisu sa musia týkať len archívnej jednotky opísanej na tejto úrovni. 3) Prepojenie opisov – každá archívna jednotka musí byť prepojená so svojou nadradenou úrovňou v rámci hierarchie a jej úroveň musí byť explicitná. 4) Neopakovanie informácií – aby sa zabránilo opakovaniu, všeobecné informácie spoločné pre skupinu sa musia deklarovať na najvyššej možnej úrovni. Podúrovne musia zase obsahovať spoločné informácie, ktoré sa vzťahujú na jej nižšie úrovne.

PyLaia. Nástroj na rozpoznávanie rukopisného textu, ktorý umožňuje používateľovi nastaviť si jednotlivé parametre transkripcie. Zmeniť sa dá aj sieťová štruktúra PyLaia, čo je príležitosť pre ľudí, ktorí poznajú strojové učenie. Úpravy neurónovej siete je možné vykonať prostredníctvom úložiska GitHub. Dokumenty, ktoré boli transkribované pomocou modelu PyLaia, je možné prehľadávať pomocou plnotextového vyhľadávania (*Solr*) v Transkribuse.

READ (*Recognition and Enrichment of Archival Documents*). Projekt, ktorého riešenie prebiehalo v rokoch 2016 – 2019 v rámci európskeho programu Horizon 2020. Výskum bol predtým financovaný ako súčasť projektu *tranScriptorium*. Tento projekt získal finančné prostriedky zo 7. rámcového programu Európskej únie pre výskum a technologický rozvoj podľa dohody o grante č. 600707. Viac o projekte https://cordis.europa.eu/project/id/674943.

READ-COOP. Združenie na udržateľnosť a vývoj platformy Transkribus. V októbri 2022 malo združenie 113 členov z 27 krajín. Jedinou členskou krajinou zo strednej a východnej Európy bolo v tom čase Slovensko. V READ-COOP sa kupujú kredity. Nejde o zisk združenia, ale o príjem, ktorý sa používa na výskum, vývoj a infraštruktúru platformy.

Riadkové polygóny (*Line polygons*). Oblasti, ktoré sa nachádzajú v textových rámcoch a možno ich opísať ako mnohouholníky, v ktorých je všetok ručne písaný/tlačený text v riadku. Keďže nemajú pre proces transkripcie bezprostredný význam, riadkové polygóny by sa nemali opravovať. Ak sa niečo má zmeniť v rozložení riadkov dokumentu, vždy to treba urobiť na úrovni základnej čiary (*Baseline*). Základná čiara by mala prebiehať pozdĺž spodnej časti textového riadku, písmená by na nej mali sedieť a zostupne smerovať nižšie. Riadkové polygóny sa prispôsobia automaticky, keď niečo zmeníte na základnej úrovni. Zobrazí sa vyskakovacie okno s otázkou, či chcete zmeniť aj nadradený riadok, čo treba potvrdiť.

Segmentácia (*Segmentation*). Uplatnenie metódy analýzy obrazu a textovej analýzy, pričom výsledkom tejto analýzy je určenie členenia stránky textu na časti – analýzou sa vyznačujú najmä textové rámce, horizontálne členenie textu, podstatné, prípadne okrajové, nadbytočné časti obrazu a riadky. Jednotlivé nahraté dokumenty v zbierke majú v aplikácii Transkribus formu obrázkov, ktoré vznikli v procese snímania (skenovania). Sú to snímky stránok dokumentov nahratých na platformu Transkribus vo formáte PDF, JPG/JPEG a PNG. Snímky je potrebné segmentovať, identifikovať jednotlivé prvky obrázkov. Na účely transkripcie dokumentu je najprv potrebné obrázok rozdeliť na textové rámce a riadky (*Text Regions* a *Lines*). Segmentáciu je možné vykonať niekoľkými kliknutiami a vo väčšine prípadov si úkon nevyžaduje manuálne opravy. To závisí od zložitosti štruktúry vstupného dokumentu. V aplikácii Transkribus sa segmentácia spustí automaticky, keď sa spustí úloha rozpoznávania textu. Automatická pokročilá analýza rozloženia vo svojom štandardnom nastavení zvyčajne rozpozná jeden textový rámec na obrázku so zodpovedajúcimi základnými čiarami. Existujú však aj rozloženia, pri ktorých sa odporúča použitie viacerých textových rámcov. Ide o situácie, keď existujú poznámky na okraji alebo poznámky pod čiarou a podobné opakujúce sa prvky. Ak sú tieto textové oblasti, ktoré sa líšia obsahom a štruktúrou, obsiahnuté v jednej textovej oblasti, analýza rozloženia jednoducho počíta riadky zhora nadol. Toto poradie čítania nezohľadňuje, kam text skutočne patrí z hľadiska obsahu, ale len to, kde sa na stránke graficky nachádza. Oprava automaticky vygenerovaného, ale neuspokojivého poradia čítania môže byť časovo náročná. Problému možno ľahko predísť vytvorením niekoľkých textových rámcov.

Sites. Webový nástroj Transkribusu, ktorý online formou sprístupňuje dokumenty zo zbierky vytvorenej v aplikácii Transkribus. Webové rozhranie bohaté na funkcie je ideálne na sprístupnenie historických dokumentov a vyhľadávanie na webe.

SKRIPTOR. Projekt APVV-19-0456 (2020 – 2024). Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov (*Innovative disclosure of written heritage of Slovakia through the automatic transcription of historical manuscripts*). Riešiteľské organizácie: Univerzita Mateja Bela v Banskej Bystrici (zodpovedný riešiteľ doc. Mgr. Imrich Nagy, PhD.), Štátna vedecká knižnica v Banskej Bystrici – partner (garant prof. PhDr. Dušan Katuščák, PhD.).

Snímanie. Jeden z procesov digitalizácie. Vykonáva sa pomocou vhodného technického zariadenia na digitalizáciu, akým sú zariadenia na zachytenie digitálneho obrazu (digitálne fotoaparáty a kamery, skenery na knihy, dokumenty alebo mikrofilmy, audio- a videohardvér) pripojené na vhodnú počítačovú platformu. Je možné rozlíšiť dve rôzne metódy snímania: skenovanie a fotografovanie, používanie digitálnych kamier/fotoaparátov, mobilných telefónov. Na účely automatickej transkripcie, ak je to možné, sa používajú dokumenty nasnímané profesionálnymi skenermi a obrazmi v najvyššej dosiahnuteľnej kvalite. Minimálna kvalita skenovania by mala byť 300 DPI. Keďže pri historických rukopisoch ide de facto o grafiku, je vhodné skenovať vo vyššej kvalite. Pre platformu Transkribus je možné snímať dokumenty do formátu veľkosti A3 zariadením ScanTent so softvérom DocScan.

Stav dokumentu. Rôzne stavy spracovania strany: *New* (nový – stav pre nové nahraté dokumenty), *In Progress* (prebiehajúci – automatická zmena stavu po úprave strany), *Done* (hotový – stránka je prepísaná, ale vyžaduje ďalšiu kontrolu), *Final* (finálna verzia – stránka je prepísaná a skontrolovaná), *Ground Truth* (základná pravda – 100 % správne prepísaná strana). Znamená to, že sa zaznamenáva práca s každou jednotlivou stranou a verzii strany sa môžu priradiť rôzne stavy v závislosti od toho, aký pokrok sa na nich dosiahol.

Štrukturálne metadáta – tagy (*Structural metadata – tags*). V štruktúre systému Transkribus je možné pomocou funkcie štrukturálneho značkovania vo funkcionalite metadáta označiť, „značkovať“ (*Mark-up*) prvky štruktúry dokumentov. Okrem toho je možné trénovať modely tak, aby automaticky rozpoznali štruktúru dokumentov. Pridaním tagov, teda štrukturálnych značiek sa vytvoria cvičné dáta pre tento proces. Nie je potrebné označovať každý prvok dokumentu, stačí sa zamerať na označenie sekcií, ktoré vás zaujímajú. Rozhranie štrukturálneho označovania v Transkribuse umožňuje rozdeliť dokumenty do štruktúrnych sekcií, ako sú odseky, nadpisy alebo čísla strán, pridať prispôsobené kategórie značiek pre vaše individuálne potreby a v budúcnosti použiť tieto štrukturálne informácie na trénovanie modelu.

Tabuľky. Tlačené a ručne kreslené tabuľky sú bežné v historických dokumentoch všetkých typov. V súčasnosti sa tabuľky musia v Transkribuse kresliť ručne pomocou editora tabuliek. Technológia, ktorá umožní automatické rozpoznávanie tabuliek, je vo vývoji. Momentálne ide v práci s tabuľkami o poloautomatický proces. Na účely transkripcie je najprv potrebné manuálne vytvorenie štruktúry tabuľky v Transkribuse a prepis textu, ktorý tabuľka obsahuje. Ak majú tabuľky v dokumente rovnakú štruktúru na viacerých stranách, je možné schému pripravenej štruktúry tabuľky použiť na dávkové rozpoznanie ďalších strán s tabuľkami. Ak teda majú viaceré strany rovnakú štruktúru tabuľky alebo šablónu tabuľky, pripraví sa manuálne tabuľka len pri prvom výskyte a potom sa distribuuje na ďalšie strany pomocou súpravy nástrojov *nomacs*. Na transkripciu tabuliek sa najprv vytvoria textové rámce (*Text Region*) pre všetky informácie, ktoré nepatria do tabuľky. Týka sa to informácií v hornej časti, spodnej časti alebo po stranách stránky, ktoré evidentne nie sú súčasťou tabuľky ako napríklad čísla strán, čísla riadkov, termíny, akékoľvek iné označenia alebo anotácie. Následne sa vytvoria textové rámce pre jednotlivé bunky tabuľky, horizontálne a vertikálne čiary, a koriguje sa text v bunkách tabuľky na strane. Grafickú schému tabuľky, ohraničenie tabuľky a buniek je možné použiť

na ďalšie rovnaké strany s tabuľkami. Bunky sa ohraničujú pomocou nástroja Ohraničovanie buniek (*Cell borders*).

Textový rámec (*Text region*). Ak chcete vygenerovať automatický prepis pomocou platformy Transkribus, musíte dokumenty rozdeliť na textové rámce, v nich vymedziť riadkové polygóny a základné čiary. V predvolenom nastavení je oblasť textu obdĺžnik, ktorý obklopuje všetok ručne písaný alebo tlačený text obsiahnutý v obrázku. Textový rámec možno podľa všeobecného rozloženia upraviť aj pridaním kontrolných bodov, čím sa vytvorí polygón.

Transkribus expert klient (*Transkribus Expert Client*). Pôvodná klientska verzia Transkribusu so všetkými funkcionalitami, ktoré sa postupne aplikovali pri vývoji systému. Posledná je verzia 1.27.0. V súčasnosti sa už tento nástroj nepodporuje – jediným podporovaným nástrojom pre prístup a prácu v systéme Transkribus je webová aplikácia, do ktorej sa postupne implementujú všetky pôvodné funkcionality expert klienta.

Transkribus web app (pôvodne označovaný ako *Transkribus Lite*). Automaticky transkribuje a umožňuje pohodlnú úpravu historických dokumentov. V súčasnosti už má aplikovanú väčšinu funkcionalít Transkribus expert klienta. V Transkribus web app je teda možné realizovať všetky fázy potrebné na automatickú transkripciu: import dokumentu, segmentáciu, trénovanie modelu, automatickú transkripciu a export transkripcie vo zvolenom formáte.

Transkribus. Komplexná platforma na digitalizáciu dokumentov, na rozpoznávanie textu podporované umelou inteligenciou, ako aj na prepis a prehľadávanie historických dokumentov – z akéhokoľvek miesta, kedykoľvek a v akomkoľvek jazyku. Platforma integruje nástroje vyvinuté výskumnými tímami v celej Európe vrátane tímu na rozpoznávanie vzorov a technológie ľudského jazyka Technickej univerzity vo Valencii a skupiny CITlab University Rostock. V októbri 2023 mal Transkribus viac ako 100 000 registrovaných používateľov a viac ako 40 miliónov rozpoznaných strán. Platforma bola vytvorená v kontexte dvoch projektov EÚ tranScriptorium (2013 – 2015) a READ (2016 – 2019).

Transkripcia (prepis). Na platforme Transkribus sa používa termín transkripcia vo význame prepisu rukopisného alebo tlačeného historického textu v určitom jazyku a automatický prepis textu v tom istom jazyku. Napríklad rukopis v maďarčine sa prepisuje pomocou znakovej sady tlačenej latinky. Nejde teda o prepis medzi jazykmi, ale o prepis v rámci jedného jazyka.

Transliterácia. Ortograficky vernému prepisu zodpovedá označenie transliterácia. Na platforme Transkribus sa pre všetky druhy prepisu konvenčne používa pojem transkripcia.

Trénovanie modelu. Pomocou nástroja Transkribus možno trénovať model rozpoznávania rukopisného textu, aby bolo možné automaticky transkribovať zbierky dokumentov. Model je výsledkom trénovania, preto je pri jeho tvorbe potrebné trénovať tak, aby stroj rozpoznal určitý štýl písania v zobrazovaných obrázkoch dokumentov a poskytol ich viac-menej presný prepis. Na trénovanie modelu je potrebných 5 000 až 15 000 slov (približne 25 – 75 strán) prepísaného materiálu. Prepis sa získa manuálnym prepisom riadok po riadku presne podľa predlohy. Prepis si možno uľahčiť použitím už prepísaných a dostupných dokumentov alebo postupovať pri príprave cvičných dát s použitím základného modelu. Pri práci s tlačeným textom sa zvyčajne vyžaduje menšie množstvo cvičných dát ako pri rukopisoch. Použitím základného modelu je možné znížiť množstvo požadovaných cvičných dát. Ako základný model sa môže použiť buď niektorý z verejne dostupných modelov PyLaia v Transkribuse, ktorý by mohol byť vhodný pre vaše dokumenty, alebo jeden z vlastných modelov, ktoré sme už predtým vytrénovali.

Verejné modely transkripcie (*Public Models*). Modely Transkribusu, ktoré je možné použiť na podobné dokumenty. Pre každý model je k dispozícii krátky opis cvičného materiálu, pre

ktoré jazyky môže byť model užitočný a kto ho vytvoril a vytrénoval. Cieľom je sprístupniť používateľom Transkribusu čoraz viac modelov, aby mohli ťažiť z kooperácie a sieťového efektu a tým šetriť prácu a čas. V súčasnosti je dostupných viac ako 100 verejných modelov napríklad: nemecký kurent, noviny, časopisy, rôzne tlače a rukopisy; viacjazyčný model pre tlače v rôznych jazykoch (holandčina, angličtina, fínčina, francúzština, nemčina, švédčina); všeobecný model pre francúzske rukopisy, nemecká bastarda 15. stor.; dánska fraktúra a historické rukopisy a strojopisy; holandské rukopisy a tlače; estónske rukopisy; fínske noviny a rukopisy; francúzske rukopisy a tlače; hlaholika; latinčina; neolatinčina; ruština; španielske rukopisy a tlače a i.

Verzie. Pri práci s dokumentom v aplikácii Transkribus sa pri každom spustení úlohy alebo uložení dokumentu vytvorí nová verzia dokumentu. Výhodou je, že sa vždy môžete vrátiť k starším verziám a pokračovať v práci na nich, čo zabraňuje strate údajov v Transkribuse Prehľad uložených verzií dokumentu si otvoríte kliknutím na ikonu hodiniek (*Version History*). Pri verziách jednotlivých stránok je vždy informácia o stave strany (*Page status*), používateľovi, dátume zmeny, nástroji zmeny a identifikátoroch.

Virtuálna klávesnica. Editačný nástroj aplikácie Transkribus, ktorý umožňuje pridávať znaky sady Unicode (ISO 10646) a špeciálne znaky, ktoré nie sú dostupné na bežnej klávesnici. Funkcia Pridávanie znakov (*Add characters*) v konfigurácii zobrazuje znaky ku klávesnici, ktoré je možné dopĺňať a odoberať po uložení zmeny (*Save*).

WER (*Word Error Rate*). Miera chybovosti slov v transkripcii.

Základná čiara (*Line*) (*Baseline, BL*). Najdôležitejší referenčný bod na rozpoznávanie textu. Popisuje polyčiaru, ktorá sa tiahne pozdĺž spodnej časti rukou písaného/tlačeného textového riadku. Segmentáciu textu na riadkové polygóny a základné čiary je možné vykonať automaticky. Pri zložitých rozloženiach a v závislosti od konkrétneho písma v rukopisoch/tlačiach sa však môžu vyskytnúť prípady, keď je potrebné vykonať manuálne opravy. Základná čiara by mala prebiehať pozdĺž spodnej časti textového riadku, písmená by na nej mali sedieť a zostupne smerovať nižšie. Základná čiara pozostáva z jednotlivých bodov, ktoré je možné nastaviť pri manuálnej úprave segmentácie.

Základný model (*Base model*). Ak tvoríme vlastné, generické modely na rozpoznávanie textu, tak nepracujeme so základnými modelmi. Pri trénovaní so základnými modelmi je však každé trénovanie založené na existujúcom modeli, t. j. na základnom modeli. Toto je spravidla posledný model na rozpoznávanie textu, ktorý bol vytrénovaný v nejakom projekte. Základné modely si „pamätajú“ to, čo sa už „naučili“. Preto každé nové trénovanie teoreticky zlepšuje kvalitu nového modelu. Nový model sa učí od svojho predchodcu a stáva sa tak lepším. Preto je trénovanie so základnými modelmi obzvlášť vhodné aj pre veľké generické modely, ktoré sa neustále vyvíjajú počas dlhého časového obdobia. Ak chcete vykonať trénovanie so základným modelom, jednoducho si v cvičnom nástroji okrem obvyklých nastavení vyberte konkrétny základný model. Potom na karte údaje modelu pre rozpoznávanie textu (*Model data*) vložte cvičné a overovacie dáta základného modelu ako aj nové cvičné a overovacie dáta. Okrem toho môžete pridať ďalšie nové strany *Ground Truth* a začať s trénovaním.

Zálohovanie a archivovanie. V procesoch snímania je nevyhnutné zvoliť metódu zálohovania a archivovania zdrojových obrázkov a ich derivátov. Základné pravidlo o zálohovaní vyžaduje urobiť najmenej tri kópie na dva rôzne nosiče a jednu – archívnu zálohu mať na vzdialenom mieste. Každá snímka by mala mať aspoň dve kópie, a to na dvoch rôznych úložiskách, napríklad na SD karte, disku, externom disku, v digitálnom repozitári.

Zbierka (*Collection*). V štruktúre systému Transkribus sú dva kľúčové prvky: zbierky a dokumenty. Zbierka je nadradená dokumentu. Dokumenty sú usporiadané do zbierok. Zbierky možno chápať ako priečinky obsahujúce dokumenty. Zbierky sa zvyčajne tvoria podľa konkrétneho projektu. Napríklad všetky dokumenty patriace k jednému projektu sú usporia- dané do jednej zbierky. Dokumenty pozostávajú z jednej alebo viacerých strán dokumentu. Každá zbierka v Transkribuse má jedinečný identifikátor (ID). Každý dokument v zbierke má jedinečný číselný identifikátor, názov dokumentu, počet strán dokumentu, meno osoby, ktorá nahrala dokument do Transkribusu, dátum a čas nahratia, meno vlastníka zbierky. V zbierke je možné manažovať – tvoriť, vymazať, upravovať, pridávať a upravovať oprávnenia používateľom zbierky so súhlasom a rozhodnutím vlastníka zbierky, pracovať s kreditmi k zbierke. Ku každému dokumentu je možné popísať všeobecné metadáta a metadáta k jednotlivým stranám, ako aj štrukturálne a textové metadáta a komentáre. Používateľ môže mať niekoľko zbierok s rôznymi dokumentmi. Na účely prezentačnej vrstvy *Sites* je potrebné vytvoriť jednu spoločnú zbierku. Všetky zbierky a dokumenty v Transkribuse sú súkromné.

Použité zdroje

DRAŠKABA, Peter a Jozef HANUS, prekl. Všeobecná medzinárodná norma pre opis archívnej jednotky. *Slovenská archivistika* [online]. 2000, roč. 35, č. 1, s. 197 – 215 [cit. 2023-08-17]. ISSN 2730-0323. Dostupné na: https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/archivnictvo/slovenska_archivistika/Slovenska%20archivistika_1-2020.pdf

KATUŠČÁK, Dušan: Metodológia a metodika transkripcie historických textov. In: KATUŠČÁK, Dušan a Imrich NAGY, eds. *Automatická transkripcia slovacikálnych historických dokumentov* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela, 2022, s. 18 – 47 [cit. 2023-08-29]. ISBN 978-80-557-2020-3. Dostupné na: https://doi.org/10.24040/2022.9788055720203

KERESTEŠ, Peter. Archívny dokument a jeho definícia. *Slovenská archivistika* [online]. 2022, roč. 52, č. 1, s. 137 – 147 [cit. 2023-18-17]. ISSN 2730-0323. Dostupné na: https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/archivnictvo/slovenska_archivistika/SA%201-2022,%20roc.%2052.pdf

KÖRMENDY, Lajos. Štandardizovanie opisu archívnej jednotky: odborný nástroj v kontexte národnej a regionálnej tradície. *Slovenská archivistika* [online]. 2000, roč. 35, č. 2, s. 222 – 235 [cit. 2023-08-17]. ISSN 2730-0323. Dostupné na: https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/archivnictvo/slovenska_archivistika/Slovenska%20archivistika_2-2020.pdf

KURHAJCOVÁ, Alica. Keď sa stroj učí čítať Hurbanove listy. In: *Automatická transkripcia slovacikálnych historických dokumentov* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela, 2022, s. 124 – 145 [cit. 2023-10-09]. ISBN 978-80-557-2020-3. Dostupné na: https://doi.org/10.24040/2022.9788055720203

Metodický pokyn odboru archívov sekcie verejnej správy Ministerstva vnútra SR o postupe štátnych archívov pri digitalizácii archívnych dokumentov a tvorby povinných metadát č. SVS-OA-2011/23406-001 [online]. Bratislava, 2011 [cit. 2023-08-18]. Dostupné na: https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/odbor_archivov_a_registratur/MP_digitalizaciaAD_metadata.pdf

NAGY, Imrich. Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. *Slovenská archivistika* [online]. 2021, roč. 51, č. 2, s. 53 – 67. [cit. 2023-08-17]. ISSN 2730-0323. Dostupné na: https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/archivnictvo/slovenska_archivistika/SA%202-2021,%20roc.%2051.pdf

NAGY, Imrich. Sprístupnenie Csákósovho katalógu korešpondencie Koháryovcov pomocou automatickej transkripcie. In: KATUŠČÁK, Dušan a Imrich NAGY, eds. *Automatická transkripcia slovacikálnych historických dokumentov* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela, 2022, s. 66 – 83 [cit. 2023-08-15]. ISBN 978-80-557-2020-3. Dostupné na: https://doi.org/10.24040/2022.9788055720203

PÉKOVÁ, Monika. Od analógového archívneho dokumentu k jeho digitálnej kópii. In: GRESCHOVÁ, Eva a František CHUDJÁK, eds. *Zborník Spoločnosti slovenských archivárov 2015.* Bratislava: Spoločnosť slovenských archivárov, Slovenské múzeum ochrany prírody a jaskyniarstva, 2016, s. 78 – 81. ISBN 978-80- 971356-2-1.

PIHAN, Roman. Formáty pro ukládání fotografií - 1. díl: základy. *DIGIMANIE* [online]. oXyShop, 31.10.2007 [cit. 2024-05-14]. ISSN 1214-2190. Dostupné na: https://www.digimanie.cz/formaty-pro-ukladani-fotografii-1dil-zaklady/1962

Resource Center. In: *READ-COOP* [online]. Innsbruck: READ-COOP SCE, last update 2023 [cit. 2023-08-28]. Dostupné na: https://readcoop.eu/transkribus/resources/

ŠEDIVÝ, Juraj a Hana PÁTKOVÁ, eds. *Vocabularium parvum scripturae latinae* [online]. Bratislava – Praha, 2008 [cit. 2023-08-25]. Dostupné na: https://manuscripta.at/Ma-zu-Bu/dateien/Vocabularium_parvum_scripturae_Latinae_2008.pdf

Transkribus: help center [online]. [cit. 2023-08-28]. Dostupné na: https://help.transkribus.org/

Všeobecný medzinárodný štandard pre archívny opis ISAD(G) [online]. 2. vyd. Bratislava, 2015 [cit. 2023-08-18]. Dostupné na: https://www.minv.sk/?archivne-standardy-1



© BELIANUM. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici 2024

DOI: https://doi.org/10.24040/2024.9788055721439

Táto publikácia je šírená pod licenciou Creative Commons Attribution 4.0 International Licence CC BY (uvedenie autora).

Zdroj fotografie na obálke: https://transkribus.org

Zdroj fotografie v zadnej tiráži: https://www.facebook.com/photo/?fbid=510677900861981&set=ecnf.100057593253225

ISBN 978-80-557-2143-9