1992 - 2022
Slovenská
asociácia
knižníc

KNIŽNICE 2022

Priestor pre všetkých

Zborník z celoslovenskej konferencie

ISBN
978-80-972135-4-1

www.sakba.sk

Knižnice 2022 – Priestor pre všetkých

Príspevky z celoslovenskej konferencie konanej 4. – 5. októbra 2022 v Liptovskom Jáne

Autori príspevkov:

Iveta Babjaková, Katarína Cigánová, Peter Černek, Jitka Dobbersteinová, Soňa Jakešová, Zora Kalmanová, Dušan Katuščák, Jitka Kmeťová, Radomíra Kodetová, Alena Kulíková, Ondrej Látka, Lenka Malovcová, Eva Medviďová, Michaela Mikušková, Jakub Pavčík, Zuzana Prachárová, Andrea Sivaničová, Slavka Polgárová, Soňa Šóky, Kristína Šteiningerová, Jaroslav Šušol, Pavol Tomašovič, Milota Torňošová, Valéria Závadská

Zostavovateľ: Ondrej Látka

Rok: 2022

Rozsah: 153 strán

ISBN: 978–80–972135–4–1 (online)

DOI: https://doi.org/10.33542/CKK22–4–1

Recezenti: Mgr. Beáta Bellérová, PhD., Mgr. Jana Ilavská, PhD.

Vydavateľ: Slovenská asociácia knižníc, Bratislava

Za odbornú a jazykovú stránku príspevkov zodpovedajú autori.

Rukopis neprešiel redakčnou ani jazykovou úpravou.

Vychádza iba elektronicky.

OBSAH

ROBOT S UMELOU INTELIGENCIOU SPRÍSTUPŇUJE PÍSOMNÉ DEDIČSTVO

Dušan Katuščák

Štátna vedecká knižnica v Banskej Bystrici; Slezská univerzita, Ústav bohemistiky a knihovníctva, Opava
dusan.katuscak@fpf.slu.cz

Abstrakt

Príspevok je zameraný na transkripciu a sprístupnenie písomného dedičstva v kontexte globálnych trendov. Poukazuje na význam poznania globálnych trendov pre knižnice pre riadenie a vzdelávanie. Uvádza informácie o nástrojoch rozpoznávania historického tlačeného a rukopisného písma. Popisuje ciele a obsah projektu SKRIPTOR zameraného na inovatívne sprístupnenie dokumentov a poznatkov o vedeckej komunikácii a intelektuálnom dedičstve novoveku a modernej doby širokej verejnosti a odbornej komunite. Výskumnou témou a predmetom záujmu projektu SKRIPTOR sú historické dokumenty, písomné dedičstvo, jeho prieskum, výskum, digitalizácia, automatická transkripcia, uchovávanie a sprístupnenie. Výskumným problémom projektu SKRIPTOR je tvorba čo najlepších modelov automatického rozpoznávania textov a ilustruje niektoré výsledky transkripcie.

Abstract

The contribution is focused on the transcription and making available of the written heritage in the context of global trends. It points out the importance of knowledge of global trends for libraries for management and education. Provides information on historical print and manuscript recognition tools. It describes the goals and content of the SKRIPTOR project aimed at innovatively making documents and knowledge about scientific communication and the intellectual heritage of the early modern and modern times available to the general public and the professional community. The research topic and object of interest of the SKRIPTOR project are historical documents, written heritage, its exploration, research, digitization, automatic transcription, preservation and access. The research problem of the SKRIPTOR project is the creation of the best possible models of automatic text recognition and illustrates some transcription results**.**

Kľúčové slová

globálne trendy, knižnice, písomné dedičstvo, projekt SKRIPTOR, rukopisy, transkribus, transkripcia, umelá inteligencia

Key words

artificial intelligence, global trends, libraries, manuscripts, SCRIPTOR project, transcription, transcription, written heritage

Globálne trendy v knihovníctve

Trend vo všeobecnosti znamená prevládajúcu tendenciu, novú orientáciu nejakého štýlu (napríklad v móde), prevládajúci prúd, smer alebo rozvoj niečoho určitým smerom. Trend je to, čo je alebo bude práve populárne jednoducho to, čo je „trendy". v jednotlivých sektoroch, oblastiach a odboroch je trend určitý špecifický smer rozvoja.

Poznanie trendov je dôležité najmä v riadení a v sektore vzdelávania. Pokiaľ ide o význam trendov pre riadenie, je zrejmé, že poznanie trendov je nevyhnutnou podmienkou pre riadenie. Poznanie sektorových, odborových, a všeobecných trendov rozvoja je rozhodujúce pre tvorbu politík, stratégií, programov a projektov. Bez poznania trendov nie je možné vytvoriť politiky a strategické rámce pre rozvoj akejkoľvek oblasti. v riadení musí mať hlavný manažér jednak skúsenosti, tvorivú intuíciu, víziu a jednak musí poznať trendy. Subjekt, ktorý štartuje nejaký trend alebo pôsobí ako kľúčový nositeľ trendu sa volá trendsetter alebo trend–setter. Známe sú osobnosti a spoločnosti, ktoré udávajú trendy

v móde, v automobilovom priemysle a pod. Existuje množstvo zdrojov, ktoré sa zaoberajú trendami v knihovníctve a informačných inštitúciách.

Niektoré trendy vo vzdelávaní knihovníkov

Poznanie globálnych trendov v knihovníctve má vo vzdelávaní knihovníkov dôležité miesto. Súčasní študenti knihovníctva, informačných alebo mediálnych štúdií nastúpia do praxe o tri, päť a viac rokov. na jednej strane si musia osvojiť poznatky o súčasnej práci informačných inštitúcií. o niekoľko rokov začnú pracovať v nových podmienkach, ktoré budú výsledkom sociálnych, ekonomických, technologických a iných zmien. Jednoducho, naši absolventi musia byť pripravení nie na to, čo je teraz, ale na to, čo s určitosťou bude a nastane v budúcnosti.

Semestrálna práca predmetu Úvod do štúdia

V mojom predmete Úvod do štúdia zoznamujem študentov 1. ročníka (BC) s globálnymi trendmi v knihovníctve. Vychádzam pritom z materiálov Americkej knihovníckej asociácie (ALA, American Library Association)[1]. ALA sa systematicky venuje trendom, ktoré majú vplyv na verejné, akademické a iné knižnice a informačné inštitúcie. ALA je nositeľom trendov (trendsetter) v odbore knihovníctva na svete. Usilujem sa zoznámiť študentov s strendmi, ktoré skôr alebo neskôr ovplyvnia alebo už ovplyvňujú knihovníctvo u nás.

Študenti 1. ročníka v predmete *Úvod do štúdia* boli povinní vypracovať semestrálnu prácu v rozsahu 5 – 10 strán. Bez odovzdania semestrálnej práce nebolo možné sa prihlásiť na skúšku. Cieľom semestrálnych prác bolo predložiť semestrálnu prácu so 4 kapitolami: *1) Podstata trendu; 2) Preklad popisu trendu; 3) Prečo na tom záleží; 4) Môj názor na trend.* Úlohou študentov bolo preložiť z angličtiny do češtiny popisy, charakteristiky trendov, ktoré už sú v súčasnosti a v budúcnosti budú dôležité pre smerovanie knižníc a ich transformáciu. Každý jeden študent má pridelený jeden trend. k názvu pridelia pri konečnej redakcii seminárnej práce bodovú hodnotu trendu od 1 do 10. Na hrubý, základný preklad z angličtiny do slovenčiny je možné použiť strojový prekladač (napr. Google translator). Preklad bolo potrebné redigovať a upraviť tak, aby bol opis trendu zrozumiteľný a jasný.

Osobitne cenná bola samostatná kapitola *Môj názor na trend* v ktorej študenti ukázali, že ich téma trendov zaujala, boli motivovaní a popísali trendy vlastnými slovami – ako danému trendu rozumejú, vysvetlili, či považujú daný trend za aktuálny aj v našom prostredí pre knižnice. Potom podľa vlastnej úvahy pridelili trendu bodovú hodnotu dôležitosti trendu v škále od 1 do 10 (10 je najdôležitejší trend). Podobnú úlohu majú aj študenti 1. ročnika v zimnom semestri roku 2022. Výsledky študentských prác sú uložené na stránke *Opavský knihovník* a sú dostupné kliknutím na meno a lebo názov trendu z obrázku (Obrázok 1). na účely prezentácie sme použili farebné rozlíšenie trendov podľa konceptu ALA. Farby označujú trendy: týkajúce sa *spoločnosti, technológií, vzdelávania, ekológie, vládnutia, ekonómie a demografie*.

Poznanie trendov môže informačných špecialistov (knihovníkov, archivárov, múzejníkov, dokumentaristov) inšpirovať k tomu, aby sa v predstihu kompetentne a včas zamerali na stratégie, plány a projekty, ktoré im umožnia dosiahnutie hlavného cieľa. a tým hlavným cieľom sú kvalitné univerzálne a špeciálne diverzifikované *služby* špecialistov pre svoju komunitu. v akademickom sektore môžu byť trendy témami záverečných a kvalifikačných prác a vedeckých projektov.

Technologické trendy

Medzi technologické trendy patria napríklad: *rozpoznávanie tvárí, umelá inteligencia, roboti, internet vecí, haptické (dotykové) technológie, odpojenie, samoriaditeľné (autonómne) vozidlá, „blockchain“, dáta všade.* Niektoré technologické trendy si už postupne osvojujeme. Pritom ide prevažne o osvojovanie si poznatkov o skúsenostiach iných, prípadne o vlastné overovacie experimenty.

[1] https://www.ala.org/tools/future/trends

Praktické skúsenosti už majú knižnice napríklad s robotmi, napríklad v službách knižníc, v digitalizácii, ako aj s umelou inteligenciou, internetom vecí, rozpoznávaním textov (OCR).

Obrázok 1 Globálne trendy v knihovníctve. Semestrálne práce študentov

Digital humanities a trend rozpoznávania textov

Najvýznamnejší pokrok vo výskume, vývoji a aplikáciách v digitalizácii v spoločenských a humanitných odboroch, čiže v *digital humanities* (POOLE, 2017) nastal najmä v posledných desiatich rokoch. Predmetom odborného záujmu je automatické optické rozlišovanie *písma* (OCR). *Digital humanities* považujeme *za* spoločné pomenovanie a prierezovú metodológiu pre všetky aplikácie informačných a komunikačných technológií (IKT) v spoločenských a humanitných vedách, odboroch a disciplínach a im zodpovedajúcej praxi. Táto metodológia sa komplexne uplatnila v projekte READ, ktorý

sa realizoval v rámci programu Horizon 2020[2] (MÜHLBERGER, G., 2016). Projekt READ má všetky atribúty metodológie *digital humanities*. k týmto atribútom patrí najmä: a) kooperácia bádateľov; b) scientizácia v spoločenských a humanitných odboroch; c) interdisciplinarita; d) tímovosť (medziinštitučná, medzištátna, univerzity, knižnice, archívy, galérie, múzeá); e) výrazné zapojenie informatikov do výskumu, vzdelávania a sprístupňovania poznatkov; f) umelá inteligencia (Hidden Markov Model (HMM)) .

Rozpoznávanie tlačeného a rukopisného písma

Kým OCR bežných *tlačených* dokumentov je už dávnejšie dostatočne zvládnuté pomocou kvalitných nástrojov OCR, tak náročnejšej problematike OCR historických *rukopisov* a tlačí s využitím umelej inteligencie sa venujú desiatky výskumníkov a experimentátorov len v posledných rokoch. Pokrok nastal realizáciou projektu READ[3], ktorý ako vedecký projekt základného výskumu podliehal priamo Európskej komisii a bol ročne hodnotený nezávislými hodnotiteľmi[4]. Hlavným výstupom projektu je použiteľná platforma a nástroj *Transkribus*, ktorá predstavuje svetovú inováciu zameranú na transkripciu historických rukopisov a dokumentov. v strednej a východnej Európe je Slovensko zatiaľ jedinou krajinou v združení READ–COOP, ktorá sa usiluje rozpracovať podnety Európskeho základného výskumu READ v projekte aplikovaného výskumu SKRIPTOR.

Od polovice 20. storočia sa rozpoznávanie znakov tlačených a rukopisných dokumentov rozvíjalo spoločne s OCR. Najprv sa naskenované obrázky tlačeného textu konvertovali na strojový kód a porovnávali sa shotovými šablónami písma. Tlačené dokumenty obsahujú znaky z vopred definovanými, hotovými znakovými súbormi, a preto je porovnávanie jednoduchšie. Avšak, aj softvéry OCR pre tlačené znaky sú schopné ďalšieho „doučovania“.

Rukopisné texty však predstavujú odlišný problém kvôli množstvu odlišností rukopisov, rúk, zmien rukopisov v čase. Rukopisy sa stali novou výzvou pre informatikov. Najprv, v 80. rokoch, sa výskum a vývoj rozpoznávania rukopisov rozvíjal s používaním štatistických metód. v 90. rokoch nasledoval výskum a vývoj rozpoznávania vzorov v kombinácii s umelou inteligenciou a vývoj hlbokých neurónových sietí v rokoch 2000 a 2010. Išlo aj o obdobie významného rozvoja a zvyšovania kapacít informačných a komunikačných technológií.

READ–COOP

Projekt READ skončil 30. júna 2019. Následne vzniklo medzinárodné združenie READ–COOP SCE (Societas Cooperativa Europeae – SCE), a to 1. júla 2019. Jeho cieľom je udržať a ďalej rozvíjať platformu *Transkribus*. Odborníci a inštitúcie majú záujem o pokračovanie a vývoj služby Transkribus. V súčasnosti má READ–COOP 113 členov z 27 krajín sveta. Desiatky tisíc používateľov *Transkribus* pracujú s touto platformou.

Platforma Transkribus

Výsledkom projektu READ je hlavne platforma *Transkribus. v* tejto platforme sú implementované výsledky základného výskumu projektu READ. Vytvorenie výskumnej platformy *Transkribus* bolo okrem základného výskumu jedným z hlavných cieľov projektu READ. Približne 2,5 milióna eur z 8,2 milióna eur sa investovalo do rozvoja tejto výskumnej infraštruktúry. Teraz vznikajú nadväzujúce projekty,

[2] Výskum bol predtým financovaný ako súčasť projektu *tranScriptorium.* Tento projekt získal finančné prostriedky zo siedmeho rámcového programu Európskej únie pre výskum, technologický rozvoj podľa dohody o grante č. 600707.

[3] READ Recognition and Enrichment of Archival Documents, ktorého riešenie prebiehalo v rokoch 2016 – 2019 v rámci programu Horizon2020. [cit 2.10.2021]. Dostupné z https://cordis.europa.eu/project/id/674943

[4] Dušan Katuščák bol jedným z troch hodnotiteľov projektu READ pre Európsku komisiu.

v ktorých pokračuje základný aj aplikovaný výskum. Osvojovanie si platformy *Transkribus* môže mať aj významné ekonomické efekty.

Podľa údajov z internej dokumentácie projektu READ sa trhové ceny manuálneho prepisu historických rukopisov pohybujú od 10 EUR až do 30 EUR alebo viac za jednoduchú angličtinu, nemčinu, latinčinu za konkrétny rukopis. Ak predpokladáme 15 EUR za stranu ako priemerné náklady, tak v projekte READ operátori vygenerovali peňažnú hodnotu 4 – 6 miliónov EUR. Tieto údaje sú pridanou hodnotou a potenciálnym zdrojom rozvoja novozaloženého združenia READ–COOP a presvedčivým potvrdením základnej koncepcie výskumu smerujúcej k novým poznatkom a súčasne ku komerčnému využitiu nástrojov, ktoré sú výsledkami aplikácie nových poznatkov.

Kľúčový inovatívny nástroj pre transkripciu historických rukopisných dokumentov je *Transkribus expert*. Je to komplexná platforma na digitalizáciu, rozpoznávanie textu podporované umelou inteligenciou, ako aj na prepis a vyhľadávanie historických dokumentov – z akéhokoľvek miesta, kedykoľvek a v akomkoľvek jazyku.

S *Transkribus Lite* je možné použiť *Transkribus* v prehliadači osobných počítačov a smartfónov. Mnohé z funkcií klienta *Transkribus Expert* môžu byť použité aj v *Transkribus Lite*. Výsledky transkripcie sú dostupné cez portál *Read&Search*. Platforma *Transkribus* integruje nástroje vyvinuté výskumnými skupinami v celej Európe, vrátane *Skupiny pre rozpoznávanie vzorov a technológie ľudského jazyka* Technickej univerzity vo Valencii a skupiny *CITlab University* v Rostocku.

Spoločnou víziou vedcov, expertov a iných používateľov z oblasti písomného dedičstva je, aby sa verejne dostupné modely transkripcie postupne stali užitočným spoločným nástrojom pre automatickú transkripciu historických dokumentov. Je potrebné dosiahnuť takú úroveň, aby už nebolo potrebné tvoriť pre každú zbierku rukopisov a tlačí samostatný model. pre používateľov by malo ísť o akúsi „čiernu skrinku“ (black box), v ktorej robot, umelá inteligencia sama vyberie z integrovaných modelov najvhodnejší model transkripcie historických tlačí, rukopisov, strojopisov a iných dokumentov, ktoré používateľ chce študovať alebo sprístupniť. k tomuto cieľu však vedie dlhá cesta a je nevyhnutné trpezlivo tvoriť množstvo parciálnych modelov. Optimálne by bolo, keby sme mali jeden univerzálny model pre slovenské rukopisy.

Projekt SKRIPTOR

Na Slovensku sme začali pracovať s platformou *Transkribus* v roku 2017 a informovali sme verejnosť o našich prvých modeloch transkripcie rukopisov. Spočiatku išlo o individuálnu iniciatívu, ktorá vďaka osvietenej ústretovosti ľudí z Univerzity Mateja Bela prerástla do inštitucionálneho výskumu v projekte SKRIPTOR.

V platforme *Transkribus* používame stroj umelej inteligencie *HTR+* (Handwritten Text Recognition) a *PyLaia*. Tieto stroje zatiaľ nemôžu okamžite automaticky transkribovať rôzne historické rukopisy. Najprv musí byť stroj vyškolený na konkrétny typ písma a rukopisu. Hlavným cieľom praktických experimentov v projekte SKRIPTOR je v súčasnosti tvorba *modelov* transkripcie.

Zmyslom projektu SKRIPTOR je, aby súčasťou spoločného medzinárodného úsilia boli aj naši odborníci, a aby budúca „čierna skrinka“ bola pripravená poskytnúť pomoc všetkým pri transkripcii slovacikálnych historických zbierok a dokumentov (slovenčina, čeština, latinčina, maďarčina, poľština a iné). V súčasnej fáze vývoja je dôležité zamerať pozornosť na tvorbu modelov transkripcie na základe väčších zbierok, ktoré obsahujú stovky a tisíce strán.

Výskumníci zapojení do projektu SKRIPTOR po počiatočnej nedôvere k možnostiam transkripcie sa postupne stávajú expertmi. Darí sa im vytvárať veľmi dobré až excelentné modely transkripcie archívnych dokumentov a starých tlačí.

Tvoríme modely, ktoré umožnia transkripciu písomného dedičstva z našej kultúrnej a jazykovej oblasti, pre ktorú sú charakteristické určité druhy písma, jazyky, znaky, štýly, diakritika a pod.

Jednotliví výskumníci projektu SKRIPTOR informujú o stave výskumu a tvorbe modelov pre zvolené archívne zbierky na konferencii k projektu SKRIPTOR v októbri 2022. D. Katuščák uvažuje o mieste transkripcie v *digital humanities a o* svojich výsledkoch transkripcie. P. Maliniak vysvetľuje postup a skúsenosti s transkripciou rukopisných kázní Izáka Abrahamidesa. Študentka K. Kováčová popisuje možnosti transkripcie nemeckej rukopisnej kuchárskej knihy z roku 1667. P. Kunec a jeho študent M. Katreniak prinášajú poznatky o transkripcii kanonických vizitácií. M. Mikušková a L. Nižníková popisujú prístup k transkripcii historickej tlače. O. Tomeček popisuje prístup a výsledky transkripcie novolatinského rukopisu reambulačného protokolu. M. Bôbová uvažuje o možnostiach využitia digitalizácie vo výskume dejín knižnej kultúry. A. Kurhajcová vysvetľuje postupy a skúsenosti s transkripciou rukopisu J. M. Hurbana. I. Nagy sa venuje postupu a výsledkom transkripcie Csákósovho katalógu korešpondencie Kohárovcov.

Dosiahnuté výsledky, know–how a skúsenosti nás viedli k úsiliu zaviesť revolučnú a inovatívnu platformu *Transkribus* na Slovensku a podnietiť výskum aj v Čechách[6]. Usilujeme sa rozvinúť medzinárodné kontakty a zaviesť poznatky jednak do systému vzdelávania a jednak do praxe pamäťových a fondových inštitúcií prostredníctvom projektov výskumu a vývoja.

Rezervovaný prístup

Predstavitelia *digital humanities* na Slovensku majú k tejto iniciatíve (umelej inteligencii), ako k podozrivej novote, rozličné postoje. od nadšených prejavov súhlasu a obdivu až po veľmi rezervované až odmietavé postoje (typu „to nie je nič pre nás“, „máme iné starosti“, „umelá inteligencia nenahradí nás expertov“). Známy jav zo zápasov tradicionalistov a novátorov. Často ide o reakcie, ktoré na jednej strane síce verbálne deklarujú záujem o „digitalizáciu“ a „umelú inteligenciu“, no na druhej strane svedčia o nedostatočných vedomostiach o problematike a možnostiach digitalizácie a využitia umelej inteligencie. Problémom je zrejme aj fakt, že *Transkribus* nie je hotový nástroj, „policový softvér“ hotový na „klikanie“, ale nástroj, ktorý sa kolektívne stále tvorí a zdokonaľuje. Postoje niektorých svedčia skôr o uprednostnení tradičných paradigiem práce a výskumu, než o reálnej snahe hľadať inovatívne nástroje sprístupnenia a interpretácie nášho obrovského historického písomného dedičstva ako súčasti európskeho kultúrneho dedičstva.

Ako je to s chybovosťou transkripcie

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*.

[6] HITEXT. Slezská univerzita v Opave pripravila v r. 2022 návrh projektu aplikovaného výskumu s akronymom HITEXT v programe NAKI III. Rozpočet ca 13 mil. korún. Projekt sa v r. 2022 posudzuje. Mimo toho sa problematika rieši v rámci vzdelávania a v projekte študentskej grantovej súťaže v r. 2022. Vyriešiť a implementovat v Českej republike najnovšie poznatky z evrópského základného výskumu automatickej transkripce textov historických dokumentov (READ) so špecifickým zameraním na pramene v moravskoslezskom regióne.

Alternatívne systémy transkripcie

Vo viacerých vyspelých krajinách sa realizovali projekty masovej digitalizácie a vznikli mohutné digitálne repozitáre a archívy tlačených a rukopisných dokumentov[7]. po masovej digitalizácii nastal čas aj na využívanie digitálneho obsahu získaného digitalizáciou rukopisov. Ak sa má z naskenovaných obrazov rukopisných dokumentov získať použiteľný, editovateľný text, je potrebné použiť pokročilú technológiu rozpoznávania HTR *Transkribus*, prípadne rovnaký, ale komerčný *Quartex (*Adam Matthew Digital *2018*).

Výskumníci projektu SKRIPTOR sa *prednostne* venujú platforme *Transkribus* a transkripcii rukopisných zbierok a okrajovo aj transkripcii tlačí.

Existuje celý rad iných nástrojov transkripcie: *OCR4all*, ktorý bol vyvinutý na digitalizáciu starých tlačí. Aplikácia *eScript* slúži na transkripciu rukopisov a tlačí. Nástroj *Rescribe* je určený pre stolné počítače na vykonávanie rozpoznávania OCR na obrazových súboroch, súboroch PDF a knihách Google. Jedným z použiteľných nástrojov transkripcie je aj *Pero.cz* (MARTÍNEK, 2020). Systém *ABBYY Cloud OCR SDK* je veľmi kvalitná aplikácia v cloude prostredníctvom webového rozhrania API. Aj ku *ABBYY Cloud OCR SDK* existuje viac ako 10 alternatív. Najlepšou alternatívou je *Online OCR*, ktoré je zadarmo. Ďalšie skvelé stránky a aplikácie podobné *ABBYY Cloud OCR SDK* sú okrem *Transkribusu* aj *Kofax Omnipage, Geekersoft OCR Word Recognition a i2OCR.* Mnoho aplikácií používa ako základ systém *Tesserakt*.[8] Pred výskumníkmi v budúcnosti stojí úloha vypracovať kritériá hodnotenia funkcionality a kvality nástrojov, aplikácií a platforiem transkripcie.

SKRIPTOR – niektoré výsledky transkripcie

Experimenty – tvorba modelov (staré a vzácne tlače)

Z digitálnych zbierok ŠVKBB sme v projekte APVV vykonali experimenty s týmito dokumentmi:

Časopis "Jitřenka" (Schwabacher) JPG formát, 1840, 128 s. Pozri: JITRENKA JP – Schwabacher (Transkribus.eu)).

Publikácia "Jánošík" (Schwabacher) JPG formát, 1862, 32 s. Pozri: (Jánošík – Schwabacher (Transkribus.eu)).

Noviny"Obzor" (Antikva and Schwabacher), PDF formát, 1866, 293 s. (Noviny OBZOR_1866.pdf – Schwabacher (Transkribus.eu))

[7] Na Slovensku išlo o mimoriadny a v európskom kontexte bezprecedentný národný projekt masovej digitalizácie a konzervovania v gescii Slovenskej národnej knižnice (SNK) v Martine s názvom *Digitálna knižnica a digitálny archív* 2012-2015. Jeho iniciátorom a autorom bol Dušan Katuščák. (Katuščák et al. 2008,2011a,2011b,2011c, 2021 ai). Projekt sa čiastočne realizoval na základe zmluvy medzi SNK a Úradom vlády SR zo 7. marca 2012 o poskytnutí nenávratného finančného príspevku vo výške vyše 49 miliónov eur. Vybudovaná je unikátna infraštruktúra: 20 skenerov, z toho 10 digitalizačných robotov a poloautomatov, archív na dlhodobú ochranu digitálneho obsahu, platforma Slovakiana na sprístupňovanie digitálnych dokumentov, vytvorených je 73 nových pracovných miest. Cieľom bolo digitalizovať ca 3 milióny dokumentov a fakticky celý slovacikálny knižničný fond, knihy, noviny, časopisy, zborníky ai. Unikátnosť projektu spočívala v integrácii masovej priemyselnej digitalizácie a priemyselného konzervovania degradujúceho kyslého papiera. Po podstatných zmenách manažmentu v roku 2012 sa do roku 2021 sa digitalizovalo len ca 10% z plánovaného objemu a celkove sa použilo v SNK ca 60 miliónov eur. Masová deacidifikácia papiera sa nerealizuje, takže papier, ako nosič ďalej nevratne degraduje (nevratný termodynamický dej). Digitálne dokumenty nie sú dostupné online. Stav digitalizácie je čiastočne kriticky popísaný v analýzach Ministerstva kultúry Slovenskej republiky (MKSR, 2019 a MKSR, 2020.

[8] Tesseract je nástroj na rozpoznávanie textu s otvoreným zdrojovým kódom (OCR) dostupný pod licenciou Apache 2.0. Pozri: https://tesseract-ocr.github.io/tessdoc/. Motor Tesseract bol pôvodne vyvinutý ako proprietárny softvér v laboratóriách Hewlett Packard v Bristole v Anglicku a Greeley v Colorade v rokoch 1985 až 1994, s ďalšími zmenami vykonanými v roku 1996.

Monografia "Církev of Ewanjelicko–Lutheránská" (Schwabacher) JPG formát, 1861, 375 s. (Author J. M. Hurban) (Hurban_Cirkev ev – Schwabacher (Transkribus.eu))

Noviny "Moravské noviny" (Antikva and Schwabacher), PNG formát, 1849, 20 s. (Moravské noviny – Seite 1 – Schwabacher (Transkribus.eu)). Source: Ústav pro českou literaturu AV ČR, v. v. i. | Digitalizovaný archiv časopisů (cas.cz).

Noviny „Opavský Besedník“ (Antikva a Schwabacher), formát PNG, 1861, 9 s. Transkribus Read&Search. (https://Transkribus.eu/r/slovakia–state/#/documents/763182) Zdroj: Digital Repository Kramerius 5, State research library, Ostrava. Dostupné: Moravskoslezská vědecká knihovna v Ostravě | Digitální knihovna Kramerius (digitalniknihovna.cz).

Tabuľka 1 Výsledky tréningu modelov fraktúry a antikvy

Transkripcia fraktúry (Schwabacher)								
Date	OCR method	Train file		Valid file		Accuracy CER		ID model
		pages	lines	pages	lines	Train	Valid	
20210824	OCR base 29418	7	8092	1	888	0,20%	0,91%	36160
20210905	OCR base 29418	9	11231	4	1179	0,18%	1,07%	36358
20210912	OCR base 29418	17	20805	5	2252	0,39%	0,44%	36550
20210913	OCR base 36550	7	2462	3	276	0,03%	1,78	36607

Zbierka rukopisnej korešpondencie Andreja Kmeťa

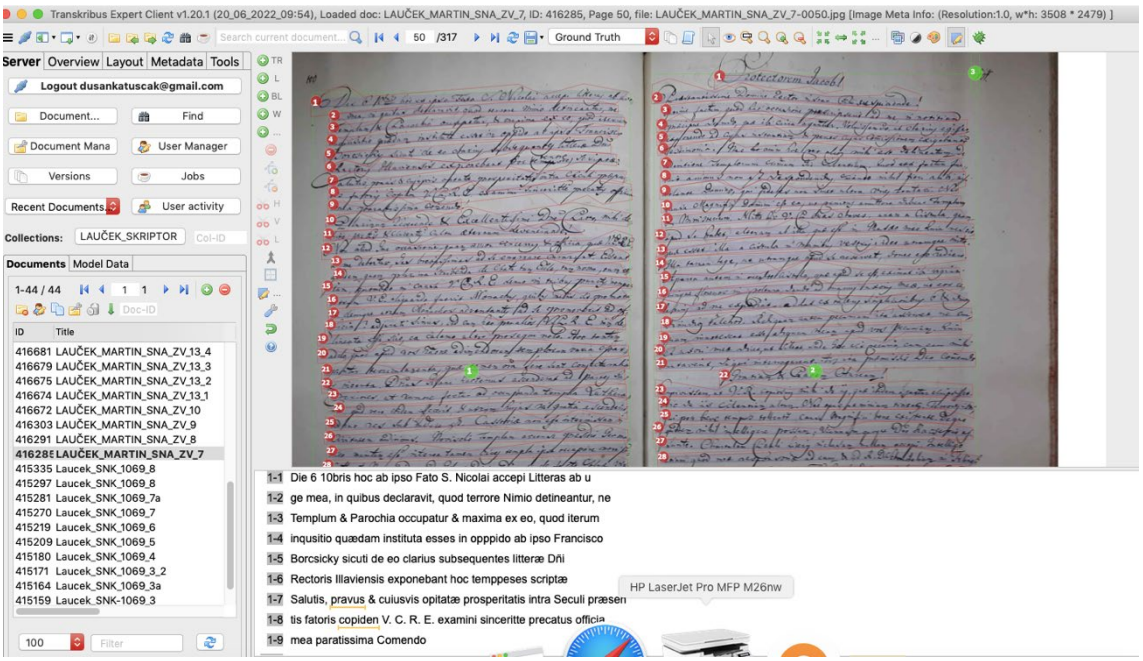
Andrej Kmeť: Príklad transkripcie – list A. Kmeťa – Búlovskému: https://Transkribus.eu/r/slovakia–state/#/documents/777242

Tabuľka 2 Modely transkripcie rukopisnej korešpondencieAndreja Kmeťa

MODELY TRANSKRIPCIE RUKOPISNEJ KOREŠPONDENCIE ANDREJA KMEŤA MODELS OF TRANSCRIPTION OF ANDREJ KMEŤ'S HANDWRITTEN DOCUMENTS											
Date	Method	Model	Training set		Validation set		CER accuracy set		Number of cycles	CER/WER	
		ID	pages	lines	pages	lines	training	validation	(epochs)	characters	words
20190125	CITlabHT+	10135	125	22549	26	3497	1.15%	3.37%	200	5.97%	21.60%
20190201	CITlabHT+	10410	152	29905	46	4499	1.27%	2.97%	200	6.19%	22.13%
20190205	CITlabHT+	10548	166	29411	46	4573	1.37%	1.84%	200	5.91%	21.87%
20201012	CITlabHT+	26809	111	18071	98	2921	0.44%	7.25%	500	6.08%	21.87%
20210410	CITlabHT+	31888	119	19291	13	3126	1.15%	5.16%	200	3.77%	12.27%
20210821	CITlabHT+	36009	185	28672	26%	4703	1,8%	5.79%	200	2.48%	7.73%

Zbierka Collectanea Martina Laučeka

Ide o pokračujúcu virtualizáciu a transkripciu 24 zväzkov rukopisov zo zbierky *Collectanea* (prevažne latinčina, 18. st.). Tisíce strán.



Obrázok 2 Príklad transkripcie latinského textu zo zbierky Collectanea Martina Laučeka zv. 7

Transkripcia hornolužickej a dolnolužickej srbčiny

Časopis Lužica 1909

Na automatickú transkripciu sme vytvorili nový, doteraz neexistujúci model transkripcie lužickej srbčiny v platforme Transkribus expertný klient. Presnosť transkripcie v najlepšom komerčnom systéme *FineReader OCR* sa pohybuje od 87% do 92% (DROBAC, S., 2020)

Naša presnosť transkripcie je 99,2%. pri danom dokumente. Všetky ďalšie ročníky by bolo možné transkribovať odteraz podobne ako tento ročník.

Rozpracované zbierky

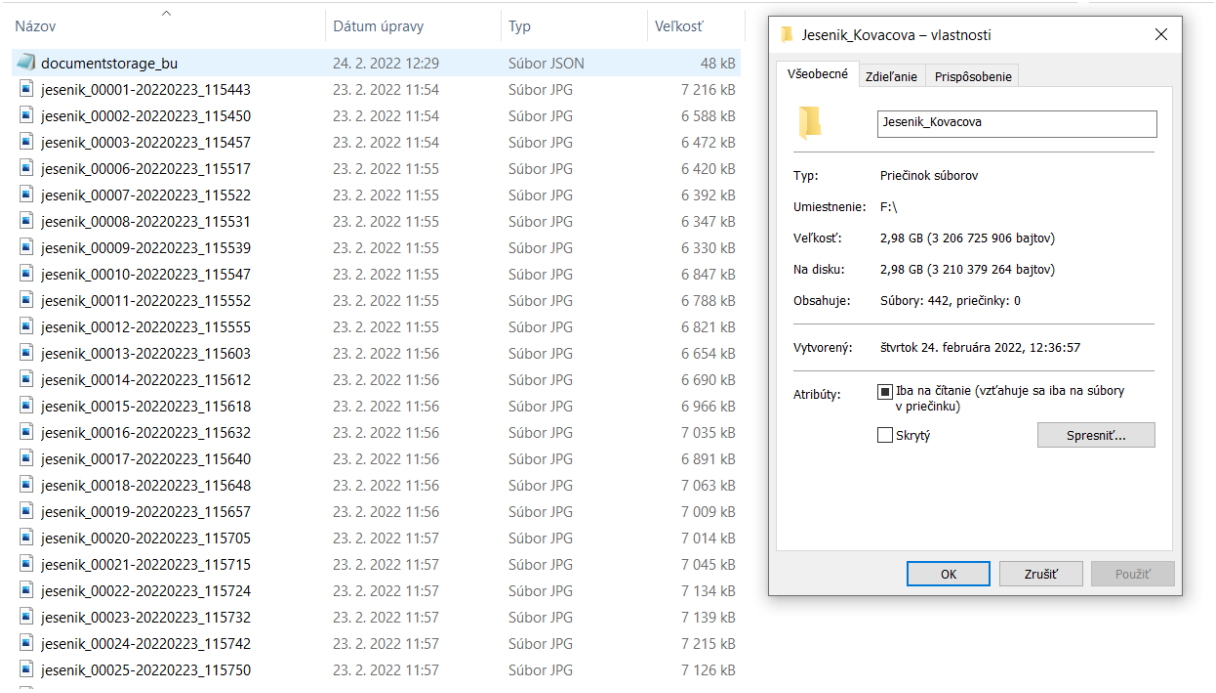
Archív Jeseník – rukopisná kuchárska kniha z roku 1667. 2,98 GB, 442 obrázkov (vyše 800 s). Práce popísané v bakalárskej práci. Pokračovanie v diplomovej práci Kláry Kováčovej. Zatiaľ transkripcia bez učenia. Použitý existujúci model transkripcie nemeckého kurentu.

Archív Rimavská Sobota

Vďaka pomoci štátneho archívu v Banskej Bystrici (p. Moniky Nagyovej) sme s dr. M. Bôbovou mali možnosť snímať dva dokumenty v archíve v Lučenci. v Rimavskej Sobote sme snímali s použitím zariadenia ScanTent a softvéru DocScan dokumenty:

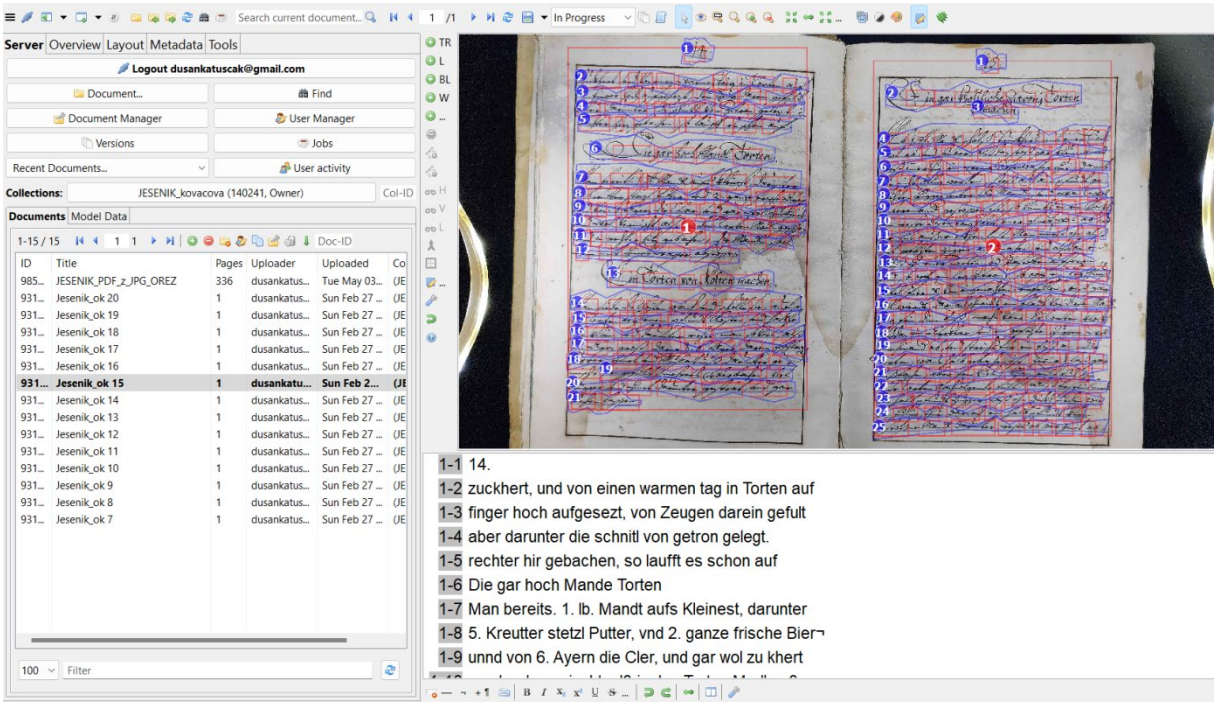
Mestečko Tisovec. Kurentálny protokol. Derivovaný súbor PDF, 500 MB, ca 500 obr., 175 dvojstrán. Rukopis, slovenčina, cca 1780 a n., potrebný je dôkladný popis dokumentu, tréning modelu, transkripcia, sprístupnenie.

Školská zápisnica, rukopis, slovenčina, 1836, cca 110 MB (tiež záznam o Dobšinskom...). Potrebný je dôkladný popis dokumentu, tréning modelu, transkripcia, sprístupnenie.



Názov	Dátum úpravy	Typ	Veľkosť
documentstorage_bu	24. 2. 2022 12:29	Súbor JSON	48 kB
jesenik_00001-20220223_115443	23. 2. 2022 11:54	Súbor JPG	7 216 kB
jesenik_00002-20220223_115450	23. 2. 2022 11:54	Súbor JPG	6 588 kB
jesenik_00003-20220223_115457	23. 2. 2022 11:54	Súbor JPG	6 472 kB
jesenik_00006-20220223_115517	23. 2. 2022 11:55	Súbor JPG	6 420 kB
jesenik_00007-20220223_115522	23. 2. 2022 11:55	Súbor JPG	6 392 kB
jesenik_00008-20220223_115531	23. 2. 2022 11:55	Súbor JPG	6 347 kB
jesenik_00009-20220223_115539	23. 2. 2022 11:55	Súbor JPG	6 330 kB
jesenik_00010-20220223_115547	23. 2. 2022 11:55	Súbor JPG	6 847 kB
jesenik_00011-20220223_115552	23. 2. 2022 11:55	Súbor JPG	6 788 kB
jesenik_00012-20220223_115555	23. 2. 2022 11:55	Súbor JPG	6 821 kB
jesenik_00013-20220223_115603	23. 2. 2022 11:56	Súbor JPG	6 654 kB
jesenik_00014-20220223_115612	23. 2. 2022 11:56	Súbor JPG	6 690 kB
jesenik_00015-20220223_115618	23. 2. 2022 11:56	Súbor JPG	6 966 kB
jesenik_00016-20220223_115632	23. 2. 2022 11:56	Súbor JPG	7 035 kB
jesenik_00017-20220223_115640	23. 2. 2022 11:56	Súbor JPG	6 891 kB
jesenik_00018-20220223_115648	23. 2. 2022 11:56	Súbor JPG	7 063 kB
jesenik_00019-20220223_115657	23. 2. 2022 11:56	Súbor JPG	7 009 kB
jesenik_00020-20220223_115705	23. 2. 2022 11:57	Súbor JPG	7 014 kB
jesenik_00021-20220223_115715	23. 2. 2022 11:57	Súbor JPG	7 045 kB
jesenik_00022-20220223_115724	23. 2. 2022 11:57	Súbor JPG	7 134 kB
jesenik_00023-20220223_115732	23. 2. 2022 11:57	Súbor JPG	7 139 kB
jesenik_00024-20220223_115742	23. 2. 2022 11:57	Súbor JPG	7 215 kB
jesenik_00025-20220223_115750	23. 2. 2022 11:57	Súbor JPG	7 126 kB

Obrázok 3 Zoznam súborov a technické metadáta k rukopisu kuchárskej knihy z archívu v Jeseníku



Obrázok 4 Automatická transkripcia strany rukopisu kuchárskej knihy z roku 1667

Bibliografické odkazy

[1] ALA Trends., 2022. Dostupné [26.9.2022] https://www.ala.org/tools/future/trends

[2] ADAM MATTHEW DIGITAL, 2018. "Handwritten text recognition: artificial intelligence transforms discoverability of handwritten manuscripts", [cit. 2.10.2021]. Dostupné z www.amdigital.co.uk/products/handwritten–text–recognition

[3] DROBAC, S. 2020. OCR and post–correction of historical newspapers and journals (Doctoral dissertation). Helsinky : University of Helsinki, 2020. ISBN 978–951–51–6511–4 (paperback) ISBN 978–951–51–6512–1 (PDF), [cit. 10.6.2022]. Available: OCR and post–correction of historical newspapers and journals (helsinki.fi)

[4] KATUŠČÁK, D., 2008. Súčasný stav formovania stratégie digitalizácie na Slovensku. In: *Kolokvium knihovních a informačních pracovníků zemí V4+*. 6. – 8. července 2008, Brno, ČR. Elektronický zborník, s. 30–46.

[5] KATUŠČÁK, D., 2021. Pochybná hodnota za veľa peňazí? In: *Kultúrny kyslík*. 2021, č. 2, s. 14–17. ISSN 1339–6919. [cit. 3. 10. 2021]. Dostupné z kulturny_kyslik_2021_2.pdf (ikp.sk)

[6] KATUŠČÁK, D. – KATUŠČÁK, M., 2011c. Základná koncepcia národného projektu digitálna knižnica. In: *Knižnica*, roč. 12, 2011, č. 2, s. 6–10. [cit. 2.10.2021] Dostupné z Knižnica 2_2011.indd (snk.sk)

[7] KATUŠČÁK, D. ET. Al., 2011a. *Digitálna knižnica a digitálny archív.* Národný projekt. Operačný program informatizácie spoločnosti OPIS2. Implementácia 2010–2015. Martin: Slovenská národná knižnica, 2011. [Kompletný projekt k žiadosti o nenávratný finančný príspevok zo štrukturálnych fondov Európskej únie ca 4000 s.]

[8] KATUŠČÁK, D., 2011b. Národný projekt digitálna knižnica a digitálny archív. In. *Bulletin Slovenskej asociácie knižníc*. Bratislava : SAK, 2011. 38 s. [Opis projektu] Dostupné na http://dusan.katuscak.net/2011/12/02/digitalna–kniznica–a–digitalny–archiv–opis2/

[9] KATUŠČÁK, D., 2011d. Situační zpráva o národním projektu SNK Digitální knihovna a digitální archiv. In: *12. konference Archivy, knihovny, muzea v digitálním světě 2011.* Praha : SKIP, 30. listopadu a 1. prosince 2011 v konferenčním sále Národního archivu v Praze, Archivní 4, Praha 4 – Chodovec. [cit. 2.10.2021] Dostupné z http://old.skipcr.cz/dokumenty/akm–2011/Katuscak.pdf

[10] MARTÍNEK, J. – LENC, L. – KRÁL, P., 2020. *Building an efficient OCR system for historical documents with little training data.* Neural Comput & Applic 32, 17209–17227 (2020). [cit. 2.10.2021] https://doi.org/10.1007/s00521–020–04910–x

[11] MÜHLBERGER, G., 2016. READ (Recognition and Enrichment of Archival Documents) – 2016–2019. [Projektová štúdia]. [cit 6.10.2021.] Dostupné z https://www.academia.edu/22653102/H2020_Project_READ_Recognition_and_Enrichment_of_Archival_Documents_–_2016–2019

[12] MÜHLBERGER, G. et al., 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5), 954–976. DOI: https://doi.org/10.1108/JD–07–2018–0114

[13] ZELGER, J. – SAGMEISTER, D., 2014. User–driven correction of OCR errors: combining crowdsourcing and information retrieval technology. In: Antonacopoulos, A. & Schulz, K. U. (Eds.), DATeCH'14: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, Madrid, Spain, 19–20 May 2014 (pp. 53–56). New York, NY: Association for Computing Machinery. DOI: https://doi.org/10.1145/2595188.2595212

[14] MÜHLBERGER, G. – COLUTTO, S. – KAHLE, P., [2016 Preprint] *Handwritten Text Recognition (HTR) of Historical Documents as a Shared Task for Archivists, Computer Scientists and Humanities Scholars. The Model of a Transcription & Recognition Platform (TRP).* [cit. 5.10.2021]. Dostupné z Günter Mühlberger | Univerzita v Innsbrucku – Academia.edu

[15] POOLE, ALEX H., 2017. The Conceptual Ecology of Digital Humanities. In: *Journal of Documentation*, roč. 73, 2017, č. 1, s. 91 – 122. [cit. 3. 10. 2021]: Dostupné z https://www.academia.edu/27862789/The_Conceptual_Ecology_of_Digital_Humanities

[16] STROBEL, P.B. – CLEMATIDE, S. – VOLK, M., 2020. How Much Data do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR. In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020),* pages 3551–3559 Marseille, 11–16 May 2020 c European Language Resources Association (ELRA)

[17] MKSR, 2019. Revízia výdavkov na kultúru. Priebežná správa. Október 2019 Kap. 4.4 Projekt digitalizácie, s. 75–78. [cit. 2.10.2021] Dostupné 28.9.2021: Revizia_vydavkov_na_kulturu_priebezna_sprava_compressed.pdf (gov.sk)

[18] MKSR, 2020. Revízia výdavkov na kultúru. Záverečná správa. Júl 2020. Kap. 4.9 Digitalizácia kultúrneho dedičstva, 132–139. [cit. 2.10.2021] Dostupné 28.9.2021: Revizia_vydavkov_na_kulturu_–_zaverecna_sprava_compressed.pdf (gov.sk)

Príspevok vznikol vďaka podpore Projektu APVV–19–NEWPROJECT–17816 (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].