Univerzita Hradec Králové
Filozofická fakulta

Diplomová práce

2021 Bc. Ondřej Tomiška

Univerzita Hradec Králové

Filozofická fakulta

Katedra pomocných věd historických a archivnictví

Automatizovaný přepis rukopisných historických dokumentů a jejich využití pomocí moderních IT

Diplomová práce

Autor: Bc. Ondřej Tomiška

Studijní program: Historické vědy

Forma studia: prezenční

Studijní obor: Moderní systémy archivnictví

Vedoucí práce: doc. RNDr. Štěpán Hubálovský, Ph.D.

Hradec Králové, 2021

Univerzita Hradec Králové
Filozofická fakulta

Zadání diplomové práce

Autor: **Ondřej Tomiška**

Studium: F19NP0031

Studijní program: N7105 Historické vědy

Studijní obor: Archivnictví

Název diplomové práce: **Automatizovaný přepis rukopisných historických dokumentů a jejich využití pomocí moderních IT**

Název diplomové práce AJ: Automatized transcription of handwritten historical records and their use with modern IT

Cíl, metody, literatura, předpoklady:

Cílem práce je analyzovat možnosti využití moderních technologií strojového učení, HTR (Transkribus, Google Vision API) a OCR (Tesseract, Microsoft OCR Azure, Amazon Textract), které umožňují automatizované rozpoznávání a převod rukopisného (ale i tištěného) archivního materiálu do textové digitální podoby a tím zpřístupnit velké množství archiválií. Toto má veliký význam nejen pro badatele, ale i pro další zpracování moderními postupy práce s textovými daty jako jsou web scraping, nástroje textové analýzy (Python Natural Language Toolkit), a další metody spadající pod Natural Language Processing (NLP) a Data Mining.

Každá zmíněná technologie i její konkrétní implementace jsou popsány a komparovány, jsou pro ně uvedeny nejběžnější zástupci, je provedena literární rešerše pro zjištění současného stavu vědeckého bádání. Zmíněny jsou také praktické příklady reálného využití těchto technologií a postupů fungující jako určitá ukázka jejich využití a jejich aplikace do prostředí českého archivnictví. Technologie jsou přímo porovnávány se současnými běžnými řešeními digitalizačních procesů v prostředí českých archivů, zanalyzovány jejich nedostatky a napak klady a možný nevyužitý potenciál v kontextu zmíněných technologií.

Services. READ Recognition and Enrichment of Archival Documents [online]. 2018 [cit. 2019-11-12]. Dostupne z: https://read.transkribus.eu/services/

Natural Language Toolkit. NLTK 3.4.5 documentation [online]. 2019 [cit. 2019-11-13]. Dostupne z: https://www.nltk.org/

Reiter, Brian-Patrick, How to Analyse Non-Digital Historical Archives of Large Organizations ? A Text-Mining Case Study (September 26, 2019). WST Working Paper Series. Available at SSRN: https://ssrn.com/abstract=3460121

About us. Impact digitisation.eu: centre of competence [online]. 2019 [cit. 2019-11-12]. Dostupne z: https://www.digitisation.eu/about/

A set of benchmarks for Handwritten Text Recognition on historical documents. Pattern Recognition [online]. Universitat Politecnica de Valencia Camino de Vera, Spain, 2018, 51(94), 122-134 [cit. 2019-11-12]. Dostupne z: https://www.sciencedirect.com/science/article/pii/S003132031930200

ALCTS Preservation and Reformatting Section, Working Group on Defining Digital Preservation, ? Definitions of Digital Preservation,? [online]. [cit. 2019-10-29]. Association for Library Collections {\& Technical Services, Dostupne z: http://www.ala.org/alcts/resources/preserv/defdigpres0408.

M. CORRADO, Edward a Heather MOULAISON SANDY. Digital Preservation for Libraries, Archives, and Museums [online]. 2. Rowman {\& Littlefield, 2017, 402 s. [cit. 2019-10-28]. ISBN 1442278722.

WALLACE, Niamh; FEENEY, Mary. An Introduction to Text Mining. Qualitative and Quantitative Methods in Libraries, [S.l.], v. 7, n. 1, p. 23-30, feb. 2019. ISSN 2241-1925. Available at: http://qqml-journal.net/index.php/qqml/article/view/454. Date accessed: 13 nov. 2019.

Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study: Transkribus as a case study. Journal of Documentation [online]. Emerald Publishing Limited, 2019, 5(75), 2 [cit. 2019-11-12]. ISSN 0022-0418. Dostupne z: https://www.emerald.com/insight/content/doi/10.1108/JD-07-2018-0114/full/html

Handwritten Text Recognition Results on the Bentham Collection with Improved Classical N-Gram-HMM methods. HIP (Historical Image Processing) [online]. Gammarth, Tunisia: ACM, 2015, 15-22 [cit. 2019-11-12]. ISSN 978-1-4503-3602-4. Dostupne z: https://dl.acm.org/citation.cfm?id=2809551

An Efficient End-to-End Neural Model for Handwritten Text Recognition. CoRR [online]. TCS Innovation Labs, Delhi, 2018, 1 [cit. 2019-11-12]. Dostupne z: https://arxiv.org/abs/1807.07965

Tateosian, Laura; Guenter, Rachael; Yang, Yi-Peng; and Ristaino, Jean (2017) „Tracking 19th Century Late Blight from Archival Documents using Text Analytics and Geoparsing„ Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings: Vol. 17 , Article 17. DOI: https://doi.org/10.7275/R5J964K5 Dostupne z: https://scholarworks.umass.edu/foss4g/vol17/iss1/17

Stork L., Weber A., van den Herik J., Plaat A., Verbeek F., Wolstencroft K. (2018) From Handwritten Manuscripts to Linked Data. In: Mendez E., Crestani F., Ribeiro C., David G., Lopes J. (eds) Digital Libraries for Open Knowledge. TPDL 2018. Lecture Notes in Computer Science, vol 11057. Springer, Cham

Garantující pracoviště:	Katedra pomocných věd historických a archivnictví, Filozofická fakulta
Vedoucí práce:	doc. RNDr. Štěpán Hubálovský, Ph.D.
Oponent:	Mgr. Klára Rybenská, Ph.D.

Datum zadání závěrečné práce: 21.2.2020

Prohlášení studenta

Čestně prohlašuji, že tato práce je mým vlastním autorským dílem. Práci jsem vypracoval samostatně a uvedl jsem všechny prameny, literaturu a zdroje, které jsem při vypracování práce použil nebo z nich čerpal.

V Hradci Králové dne Tomiška Ondřej

…… ……

Poděkování

Nejprve bych chtěl poděkovat své rodině za jejich podporu během mého celého studia, a to nejen při psaní této diplomové práce. Bez vás bych tam, kde jsem dnes určitě nebyl.

Dále bych chtěl poděkovat paní doktorce Borkovcové za všechny její vyučované obory a jejich vysokou úroveň, která mě nutila neustále se zlepšovat. Děkuji jí také za všechny její rady, velkou snahu a pomoc, kterou mi poskytla během mého studia, včetně nápadu na téma této diplomové práce.

Dále bych chtěl poděkovat doktorce Falátkové, za to, že se i přes to, že jsme se na začátku vlastně neznali, rozhodla semnou tuto práci řešit po odborné stránce. Velmi si cením také jejích rad ohledně stylizace a formální stránky práce, a hlavně ochotě se mi věnovat. Vím, že to se mnou neměla jednoduché.

Nakonec musím také poděkovat docentu Hubálovskému za jeho vedení mé diplomové práce, za jeho vždy rychlou odpověď, výborný přístup a jeho ochotu mi vyhovět pokaždé, když jsem po něm něco potřeboval.

Anotace

TOMIŠKA, ONDŘEJ. *Automatizovaný přepis rukopisných historických dokumentů a jejich využití pomocí moderních IT*. Hradec Králové. Filozofická fakulta, Univerzita Hradec Králové, 2020, 131 str., Diplomová práce.

Cílem práce je analyzovat možnosti využití současných informačních technologií (NLP, HTR, OCR, aj.) a jejich nástrojů (Transkribus, Quartex, Textract, NLTK, aj.) pro automatizaci procesů transkripce rukopisných i tištěných dokumentů a jejich následné využití pomocí text mining, web scraping a natural language processing metod, nástrojů a technik v kontextu archivního zpracování a metod Digital Humanities.

Tyto technologie, jejich nástroje a jednotlivé implementace pro automatizovanou transkripci a další počítačové zpracování jsou popsány a mezi sebou komparovány, kdy důraz je kladen na jejich praktické využití, jejich slabiny, výhody a reálný dopad na oblast současného českého archivnictví.

Klíčová slova:

HTR, NLP, OCR, Transkribus, Text mining

Annotation

TOMIŠKA, ONDŘEJ. *Automatized transcription of handwritten historical records and their use with modern IT*. Hradec Králové. Filozofická fakulta, Univerzita Hradec Králové, 2020, 131 p., Master Thesis.

This Master´s thesis aims to analyze current possibilites regarding the use of modern information technologies (NLP, HTR, OCR etc.) and its tools (Transkribus, Quartex, Textract, NLTK etc.) to automatically transcribe handwritten and printed records used to further process via Text Mining, Web Scraping and Natural Language Processing methods, tools and techniques in the context of archival processing and Digital Humanities methods.

These technologies, tools and implementations for automatized transcription and advanced computer processing are described and compared their practical usage, weaknesses, strengths and tangible impacts on the current Czech archival field.

Keywords:

HTR, NLP, OCR, Transkribus, Text mining

Obsah

Seznam zkratek

AI Artifical Intelligence

ALTO Technical Metadata for Layout and Text Objects

API Application Programming Interface

ASCII American Standard Code for Information Interchange

BI Business Intelligence

BMP Windows Bitmap

CAPTCHA Completely Automated Public Turing test to tell Computers and Humans Apart

CAS Computational Archival Science

CER Character Error Rate

CLI Command Line Interface

CRM Customer Relationship Management

CSV Comma-separated value

DH Digital Humanities

DLA Document Layout Analysis

DjVu Déja Vu file format

eIDAS electronic Identification, Authentication and trust Services

EPUB Electronic Publication

ERM Entity Relations Model

ESSL Elektronický systém spisové služby

ETL Extract Transform Load

FAQ Frequently Asked Questions

GDPR General Data Protection Regulation

HMM Hidden Markov Model

HOOSC Histogram of Orientation Shape Context

HTML Hypertext Markup Language

HTR Handwritten Text Recognition

ICDAR International Conference on Document Analysis and Recognition

ICFHR International Conference on Frontiers in Handwriting Recognition

ICOM International Comittee of Museum

JPEG	Joint Photographic Experts Group
JSON	Javascript Object Notation
LSTM	Long Short-Term Memory
METS	Metadata Encoding and Transmition Standard
MLIS	Master of Library and Information Science
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OCR	Optical Character Recognition
PDF	Portable Document Format
PNG	Portable Network Graphics
PVH	Pomocné vědy historické
READ	Recognition and Enrichment of Archival Documents
RNN	Recurrent Neural Network
SPSS	Statistical Package for Social Sciences
SQL	Structured Query Language
TIFF	Tagged Image File Format
TOME	Topic Model and Metadata Visualization
TXT	Text file
UI	Umělá inteligence
UTF-8	Unicode Transformation Format 8-bit
WAV	Waveform audio file format
XML	Extensible Markup Language
ZKB	Zákon o kybernetické bezpečnosti

Slovníček pojmů

Apache 2.0 – Typ softwarového licenčního ujednání poskytující daný software svobodně, kompatibilní s GPL. Umožňuje software distribuovat a upravovat.

Big Data – Obrovské množství datových souborů, se kterými právě díky svému masivnímu objemu nelze manipulovat, zpracovávat je nebo je spravovat běžnými metodami a softwarovými řešeními. Vlastní odvětví datové vědy zabývající se zpracováváním, ukládáním a správou masivního objemu dat.

Business Intelligence – Kombinace strategie a IT umožňující efektivnější postup v podniku při rozhodovacím procesu založeným na datech a jejich analýze.

Cloud – Služby umožňující poskytovat výpočetní služby skrze počítačové sítě (nejčastěji internet) nepřímo jeho uživateli bez nutnosti správy poskytovaných technických prostředků. Například poskytování výpočetního výkonu, neurální sítě, služeb umělé inteligence, webových služeb, technických prostředků aj.

Data Mining – Souhrn metod, technik a procesů umožňující získávání netriviálních, relevantních a užitečných informací a znalostí z dat.

Datová věda – Disciplína na pomezí statistiky a IT zabývající se prací s daty, extrahováním užitečných a relevantních informací ze strukturovaných a nestrukturovaných dat.

Digital Humanities – Oblast na pomezí komputace a humanitních věd, která využívá nástroje a prostředky informačních technologií pro dosažení vlastních výstupů. Zabývá se využitím komputačních metod, nástrojů a technologií pro potřeby vlastních humanitních věd.

Digital Born – Typ eDokumentu, který byl již vytvořen v plně digitálním prostředí, zpravidla počítačem.

Digital Surrogates – Typ eDokumentu, který byl vytvořen převodem analogového dokumentu do digitální podoby skrze digitalizační proces.

Dolování v textu – Specifická oblast dolování v datech zabývající se získáváním netriviálních relevantních informací a znalostí z textových dat.

eDokument – Dle zákona 499/2004 Sb. se jedná o každou písemnou, obrazovou, zvukovou nebo jinak zaznamenanou informaci, která byla vytvořena původcem, nebo byla původci doručena.

ER model – Entitně vztahový model sloužící ke konceptuálnímu zobrazení datových struktur v grafickém prostředí.

Git – Software pro záznam a stopování změn v datových souborech systému používaný pro koordinované práci mezi programátory.

Github – Nejpopulárnější internetová služba sloužící ke koordinaci vývoje aplikací skrze použití Git softwaru.

HTR – Souhrn technologií umožňující převod a rozpoznávání rukopisných textů v obrazové podobě (digitalizátu) do textové podoby.

Metadata – Jedná se o data, která slouží k popisu jiných dat s těmito metadaty souvisejícími. Například popisná, strukturální, administrativní aj.

Neurální sítě – Výpočetní modely složené z umělých neuronů, které kopírují funkce reálných biologických neuronů.

NLP – Mezidisciplinární obor na pomezí lingvistiky, informatiky a podoboru umělé inteligence, který má za cíl umožnit strojové zpracování (a porozumění) přirozeného lidského jazyka počítačům.

OCR – Technologie umožňující rozpoznávání a převod znaků v obrazovém souboru do textové či zvukové podoby.

Proprietární software – Takový typ softwaru, který má zpravidla uzavřený zdrojový kód, a jeho užívání a manipulaci s ním vytyčuje autor ve smluvních podmínkách.

Strojové učení – Disciplína spadající pod oblast umělé inteligence, která se zabývá takovými počítačovými algoritmy, které se díky vlastní zkušenosti a učení dokážou zlepšovat v rozhodovacím procesu nebo v přesnosti jejich výsledků. Takovýto algoritmus se umí sám učit a adaptovat na nová vstupní data bez toho, aniž by musel zasáhnout člověk.

Web scraping – Zpravidla automatizovaný proces sloužící k extrahování relevantních dat z webových stránek.

Workflow – Jedná se o pracovní postupy, metodiky a návody vybraných činností Např. práce s určitým nástrojem, pracovní postup při digitalizačním procesu, aj.

XML – Značkovací rozšiřitelný (upravitelný) jazyk umožňující vytváření vlastních značkovacích jazyků pro vlastní potřeby zpracování, a to často v kontextu s metadatovými standardy popisu.

Úvod

Pro humanitní vědy je, především v poslední dekádě, charakteristické určité propojení s informačními technologiemi. Právě rozvoj těchto IT – konkrétněji potom pokroky v datové vědě, a především oblasti umělé inteligence – v kontextu s humanitními obory přináší přidanou hodnotu a obohacení jejích výstupů.

Archivnictví, knihovnictví a informační věda tomuto nejsou výjimkou. Právě s rozvojem technologií strojového učení, HTR, OCR, NLP, dolování v datech, statistiky a modelů neurálních sítí souvisí pokročilé možnosti zpřístupňování archiválií a jejich následné využití pomocí komputačních metod.

Cílem této diplomové práce je provést kompletní analýzu možností využití těchto moderních informačních technologií (HTR, OCR, strojové učení, NLP, dolování v datech, aj.) a jejich jednotlivých aplikací (Transkribus, Quartex, SpaCy, Textract, aj.) v kontextu archivního prostředí. Výsledkem by měl být kompletní základní přehled o možnostech využití současných technologií a nástrojů v kontextu zpřístupňování archiválií a jejich využití pomocí komputačních metod, který v českém prostředí zatím chybí.

Práce se člení na dvě základní části, kterým předchází krátký průzkum současné podoby archivnictví jak u nás, tak v zahraničí následovaný poměrně rozsáhlým zmapováním současného stavu vědeckého bádání v oblastech využití technologií a nástrojů pro zpřístupňování archiválií, a to především v oblasti automatizovaného přepisu rukopisných a tištěných archiválií, konkrétněji potom nástroji Transkribus, OCRopus, Tesseract, Quartex, Finereader a technologiemi HTR, OCR a cloudovým možnostem masivních korporací Google, Microsoft a Amazon. Hlavním výstupem aplikování těchto technologií a jejich nástrojů je počítačem čitelný digitalizát, jehož využití se věnuje další část tohoto rešeršního oddílu. Zde je zanalyzována současná situace ve vědecké sféře pro oblasti využití a následného zpracování těchto digitalizátů pomocí technologií NLP, Scraping, dolování v textových datech a jejích jednotlivých nástrojů, frameworků a metod.

Rešeršní oddíl potom následuje hlavní část textu, ve které se autor zabývá analýzou těch nejvyužívanějších a relevantních online i offline nástrojů sloužících

k automatizovanému přepisu rukopisných a tištěných archiválií umožňující tak jejich zpřístupnění. Součástí je základní popis jednotlivých nástrojů, jejich základních procesů a funkcí, výhod a nevýhod, zakončený jejich srovnáním. Výkonnému nástroji Transkribus je věnována samostatná kapitola zabývající se jeho základním popisem následovaným úplným popisem průchodu všech jeho základních postupů a funkcí včetně úplného workflow.

Druhou hlavní částí textu je analýza a popis základních možností moderního zpracování a využití počítačem čitelných archiválií získaných nejen pomocí metod a nástrojů popsaných v předcházejícím oddílu, kdy platí, že důraz je kladen na jejich reálné využití a přínos. Jsou zmíněny výhody a nevýhody, jejich základní použití, a jednotlivé nástroje společně s tím, co nového přináší a v čem jsou problematické.

Poslední část této práce se zabývá základními výstupy, které vychází ze syntézy a analýzy předcházejících kapitol doplněná o reflexi.

Tato práce se ve výsledku snaží pomoci zodpovědět dvě otázky, se kterými se archivnictví potýká, nebo v blízké budoucnosti potýkat bude.

Jak zpřístupnit řadu prozatím nedostupných archiválií veřejnosti a jak efektivněji zpřístupnit ty již nyní dostupné?

Jak reagovat na zvětšující se objem dokumentů, se kterými archivnictví musí pracovat a jak tato data využít?

1. Současný stav vývoje a literární rešerše

Jedním ze současných problémů archivnictví je fakt, že stále existuje mnoho archiválií nezpřístupněných – nejsou na to finanční, technické prostředky a ani personál. Archiváři musí provádět selekci toho, co bude a co nebude zpřístupněno. *„Průzkum problematiky pořádání archivních souborů zjistil, že trvale udržitelný rozvoj k dosažení stavu zpracovanosti archiválií na úrovni 75 % do roku 2030 ... není reálný.“*[1]

Při procesu digitalizace[2] ve většině případů dochází k vytvoření archivní a prezentační kopie. Pokud se jedná o písemný, tištěný dokument, aplikuje se na něj také OCR, a i v tomto případě je téměř vždy potřeba provést ruční korekci. Někdy ani nelze OCR technologii využít z důvodů náročnosti textu. Zbývají nám tedy rukopisné archiválie a řada tištěných archiválií, pro které standardní OCR nelze použít. Jediný způsob, jak text zpřístupnit veřejnosti[3] je tedy ruční transkripce, která je časově velmi náročná.

Především díky pokrokům v oblasti informačních technologií v posledních 15 letech (pokroky v OCR a HTR technologiích) a poměrně rozsáhlých výzkumným projektům se v posledních několika letech objevují možné odpovědi na otázku, jak jednodušeji a efektivněji zpřístupnit[4] velké množství rukopisných a tištěných archiválií nacházejících se v archivech Evropy.

---

1 *Koncepce rozvoje archivnictví*. Ministerstvo vnitra ČR [online]. Praha: MV, 2018 [cit. 2020-09-08]. Dostupné z: https://www.mvcr.cz/clanek/koncepce-rozvoje-archivnictvi.aspx

2 Řada menších archivů (především okresních) si nemůže dovolit mít specializovaného pracovníka v digitalizační laboratoři. Archiváři v okresních archivech musí být velmi univerzálními pracovníky, řešící vše od vstupu archiválií do archivu, předarchivní péče, zpracování archiválií, inventarizaci, rešeršní činnost, aj.

3 Badatelé také často mívají problémy text vůbec přečíst – ať už kvůli jazyku, typu písma, stylu autora. Vzniklý digitalizát v obrazové podobě, nejčastěji v podobě JPEG formátu, který archiv zpřístupní skrze své různorodé webové aplikace tak má využití pouze pro část badatelů, kteří se s textem dokážou sami vypořádat.

4 Toto neplatí pouze pro oblast archivnictví, ale i pro oblast rozpoznávání rukopisných textů obecně, a to především kvůli výhodám, které toto přináší pro oblast marketingu. Viz kapitoly o Microsoft OCR Azure a Google Cloud Vision.

Je důležité si uvědomit, že pouze obrazové digitalizáty jsou hůře přístupné pro badatele nejen z hlediska jejich čitelnosti (jazyk, písmo, styl písaře), ale i z hlediska zpracovatelnosti, a to především v ohledu na moderní možnosti statistických a Big Data analýz. Text v takovéto podobě je sice pro badatele[5] okem čitelný, ale není čitelný strojově, což klade výrazná omezení na jejich zpracování badateli (pod poměrně rychle se rozvíjejí Digital Humanities), kteří chtějí provádět výzkum např. aplikováním metod dolování v datech.[6]

Je poměrně dobře známo, že díky dynamicky a rychle se rozvíjející informatice (především oblasti kryptografie, datové vědy a pokrocích v oblasti umělé inteligence) se (digitální) archivnictví potýká s řadou problémů.

Řeší se ochrana eDokumentů ze střednědobého či dlouhodobého hlediska (*Cubr – Dlouhodobá ochrana digitálních dokumentů*), systém digitálního archivu a vstup archiválií do nich společně se stále současným problémem eSSL (*Lechner, Kunt – Spisová služba)* či se řeší právní nařízení a normy (2014 eIDAS, 2016 GDPR, 2017 Velká a malá novela ZKB) anebo zastarání HW i SW. Na tato témata vzniká řada vědeckých prací společně s množstvím bakalářských a diplomových prací.

eDokumenty a digitální archiválie mají ale také mnoho velmi významných výhod. Řada z nich je poměrně dobře zmapovaná a využívaná – digitální data oproti svým analogovým protějškům nestárnou, lze je mnohem lehčeji duplikovat, rozšiřovat a prezentovat.[7] Významná vlastnost a výhoda jakýchkoliv digitálních dat (ať se jedná o archiválii, dokument, záznam), se kterou se začalo pracovat teprve až v posledních zhruba patnácti letech je potom samotná využitelnost těchto dat, a to především z pohledu datové vědy a jejích metod, technologií a postupů.

---

[5] Zde je potřeba si pod pojmem badatel představit nejen veřejnost, ale i archiváře, historika, muzejníka.

[6] MILIONI, Nikolina. *Automatic Transcription of Historical Documents* [online]. Uppsala, 2020 [cit. 2020-09-08]. Dostupné z: http://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1437985&dswid=-1804. Diplomová. Uppsala Universitet.

[7] CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. Praha: Národní knihovna České republiky, 2010. ISBN 978-80-7050-588-5.

Následující části této kapitoly se věnují současnému stavu vědeckého bádání právě těchto dvou nastíněných problémů – zpřístupňování rukopisných archiválií pomocí metod moderních OCR, HTR a jejich využití pomocí metod a postupů vycházejících z datové vědy (NLP, Text Mining) a informatiky obecně. Oběma oblastem předchází stručný nástin současného stavu archivnictví nejen v evropském prostředí, ale i u nás.

1.1. Stav archivnictví v 21. století

V mnoha humanitních oborech dochází k čím dál bližšímu kontaktu s informačními technologiemi a jejich nepopiratelnými výhodami a nevýhodami, které (do té doby poněkud statických humanitních věd) přinášejí. Archivnictví není v tomto případě výjimkou. Proto je otázka moderní podoby této vědy velmi aktuální a stále ještě neustálená.

1.1.1. České prostředí

V našem prostředí za poslední dobu vzniklo několik vědeckých článků a dalších prací, které se podobou archivnictví zabývají. Velmi zásadní je v tomto ohledu především *Koncepce rozvoje archivnictví v České republice na léta 2018 až 2028 s výhledem do roku 2035* vydané pod MV. Tato koncepce je zásadní především proto, že problémy v ní identifikované budou obsaženy v novém archivním zákoně a směru vývoje archivnictví. Některé body, které je třeba také zdůraznit jsou ty personálního charakteru. V koncepci je identifikován problém se středoškolsky vzdělanými archiváři, restaurátory, a především stále vyšší nároky na archiváře při plnění dosavadních povinností, kdy zásadní je, že „*…chybí systém následného cíleného odborného vzdělávání, kde selhávají i vysokoškolské vzdělávací programy*“[8]. Kromě běžných a již řadu let známých problémů s nedostatečným technickým vybavením a stavem archivních budov a repositářů je stojí za zmínku situace spisové služby. Jako problematickou oblast zmiňuje možnost vedení spisové služby v elektronické či listinné podobě. Možnost volby tak vnímá jako

[8] *Koncepce rozvoje archivnictví*. Ministerstvo vnitra ČR [online]. Praha: MV, 2018 [cit. 2020-09-08]. Dostupné z: https://www.mvcr.cz/clanek/koncepce-rozvoje-archivnictvi.aspx

„nepříliš šťastnou“ a hlavní problém vidí v nejednotnosti a diferenciaci předarchivní péče.[9] Můžeme tedy očekávat změny ve vedení spisové služby a jasném upřednostnění jejího vedení v elektronické podobě.

Pro tuto diplomovou práci je důležitá především sekce o digitalizaci archiválií. Celkem jasné je, že preference badatelů jsou na straně vzdáleného přístupu k archiváliím skrze webové portály. Z celé koncepce jsou důležité především tyto tři zásadní body. Za prvé – uvědomuje si neúměrné zatížení archivářů, kteří musí splňovat své základní povinnosti společně s novými a jsou neúměrně zatíženi. Za druhé – je nutné vytvořit jednotnou koncepci či rámec digitalizačního procesu. Za třetí – je potřeba vytvořit systém cíleného odborného vzdělávání (školení, vysokoškolské programy).

Kromě této koncepce rozvoje archivnictví bude velmi zásadní nový archivní zákon, který je od roku 2019 v přípravě a měl by nahradit současný zákon, původně z roku 2004 a již schválená digitální ústava.

Co se týče současné literatury tak lze zmínit hlavně knihu *Digitální archivnictví* z roku 2019, která doposud v repertoáru velmi chyběla. Důležitá je především jako určité shrnutí celé oblasti společně s elektronickou spisovou službou (pro kterou ale existuje samostatná kniha z roku 2017 – *Spisová služba 2. aktualizované vydání.)* Velmi často se stále, nejen v literatuře, odkazuje na starší dílo *Dlouhodobá ochrana digitálních dokumentů* z roku 2009[10]. Obecně lze usoudit, že současné literatury pro tuto oblast digitálního archivnictví je pomálu[11]. Chybí totiž ucelený rámec a

---

[9] *Koncepce rozvoje archivnictví*. Ministerstvo vnitra ČR [online]. Praha: MV, 2018 [cit. 2020-09-08]. Dostupné z: https://www.mvcr.cz/clanek/koncepce-rozvoje-archivnictvi.aspx

[10] Z pohledu informačních technologií je deset let poměrně dlouhá doba. Na druhou stranu u nás není zatím nic novějšího. Velmi populární v zahraničí je kniha *Digital Preservation for Libraries, Archives, and Museums*, která se zabývá stejným tématem, ale je modernější (2017) a při jejich porovnání nabízí některé odlišné pohledy na ochranu digitálních dat, a to především problematikou archivace webových stránek, sociálních sítí a potom kritické oblasti práce s tzv. Big Data.

[11] Například ze zmíněné knihy z roku 2019 – Digitální archivnictví, můžeme z informačních zdrojů a literatury čítajících 47 položek pouze 5 českých knih.

metodika právě pro procesy spojené s digitálním archivem. S příchodem nového archivního zákona možná dojde k vydání nových knih o archivnictví.

Vědeckých článků pro tuto oblast máme několik. Lze zmínit práci Marie Ryantové z Jihočeské univerzity *Training of Archivists in the 21st Century: Some Reflections*. V této práci hovoří především o problematice vzdělávání archivářů, která i po letech zůstává nevyřešená. „*Nevertheless systematic education in computer science remains unresolved and encounters various difficulties which includes unclear definition of content in this fast and spontaneously developing sub-sector, problems associated with the workload of suitable specialists, etc.*[12] "

1.1.2. Evropské prostředí

Pokud se podíváme na situaci v Evropě, zjistíme, že se zde kladou poněkud jiné otázky, i když v základu mají archivy velmi obdobné problémy i přes to, že koncepce archivnictví je v různých zemích Evropy odlišná společně se samotným uspořádáním archivů. Lze hovořit o problémech týkajících se problematiky zachování důvěryhodnosti eDokumentů[13] a archiválií a jejich zpřístupňování skrze webové portály. Rozdíl, který lze při prozkoumání různých vědeckých studiích a článcích vysledovat je přijetí a praktikování modernějších praktik Digital Humanities, kdy u nás je to záležitost posledních pěti let a teprve se začínají utvářet. Především v západních zemích jsou praktikovány již delší dobu, a tudíž před námi mají určitý náskok. Zjednodušeně lze říct, že v těchto zemích jsou humanitní obory více digitální, a tudíž více spřízněné s informační vědou.

Toto lze vypozorovat například už jen z jednotlivých sylabů univerzit, které vzdělávají budoucí archiváře. Pro porovnání lze uvést University College London, kdy zjistíme, že důraz je kladen nejen na klasické archivářské předměty, ale i na řadu předmětů právě obecně z informatiky, konkrétněji potom právě z datové vědy

[12] Ryantová, M. (2017). Training of Archivists in the 21st Century: Some Reflections. *Atlanti*, 27(2), 225-233. https://doi.org/10.33700/2670-451X.27.2.225-233(2017)

[13] V angloamerickém pojetí je slovo „document" považováno v českém archivnictví a spisové službě jako záznam, nikoliv „dokument", pokud jej takto přeložíme. Pojem pro „dokument" je v angloamerickém prostředí record.

– databáze, statistika, digitální kurátorství, aj[14]. Vidíme snahu o aplikování více technického myšlení do jednotlivých oborů. Například práce *Introducing Computational Thinking into Archival Science Education* nebo *Developing a Framework to Enable Collaboration in Computational Archival Science Education*. Naopak na Karlově univerzitě je ve studijním plánu 2020 jednoznačně kladen důraz na klasické archivnictví – PVH, archivní teorie, obecné a české dějiny, dějiny správy[15].

Podrobnější rozdíly a sledované problematiky lze vysledovat například v *The Challenge of the Digital and the Future Archive* vydaným Národním archivem Velké Británie ve stejném roce jako ona zmíněná koncepce českého archivnictví. Některé okruhy jsou totožné, řeší se problémy s udržitelností a vůbec s důvěryhodností digitálních archivů a prezentace jejich archiválií. Některých se naopak ona koncepce vůbec nedotýká.

Podrobně je potom popsána otázka hlubšího využití a návaznosti strojového učení a komputace obecně na archivní prostředí. S tím souvisí i důvěryhodnost archiválií dostupných v daném archivu. Pro řadu archivních problémů vidí odpověď společně s riziky ve využití právě těchto technologií.

„We see a future where AI and emergent technologies become part of our everyday recordkeeping practices…[16]*“* Umožnit určitou automatizaci části archivářovi práce je viděno jako cesta, jak uvolnit ruce archivářů (nepoměr archivářů k počtu nezpracovaných archiválií je problémem v řadě Evropských zemí) a tím zefektivnit mnohé kroky práce s archiváliemi.

Druhou věcí je potom samotné zpřístupňování archiválií. S rostoucím počtem digitalizátů a digital-born archiválií vznikají problémy s orientací v nich a v jejich

---

[14] Archives and Records Management MA. *UCL* [online]. London, 2020 [cit. 2020-09-10]. Dostupné z: https://www.ucl.ac.uk/prospective-students/graduate/taught-degrees/archives-records-management-ma

[15] Studijní plány. *FILOSOFICKÁ FAKULTA Univerzita Karlova* [online]. Praha, 2020 [cit. 2020-09-10]. Dostupné z: https://www.ff.cuni.cz/studium/studijni-obory-plany/studijni-plany/

[16] Goudarouli, E., Sexton, A. & Sheridan, J. The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK. Philos. Technology. 32, 173–183 (2019). https://doi.org/10.1007/s13347-018-0333-3

navigaci. Na archiválie se více a více kouká jako na data, a tak se s nimi i pracuje – „*Instead of trying to cleanse and standardise data, our techniques aim at leveraging uncertainty by quantifying and working around the 'fuzziness' found in our large-scale collections. Our aim is to enable the user to make robust, data-driven access decisions"*[17]. Jako cíl si tedy vytyčují umožnit uživatelům[18] dělat rozhodnutí (=výstupy) svých bádání v archivech na základě analýzy sesbíraných dat.

Tradičně platí, že bádání probíhá tak, že badatel dorazí do archivu, vyzvedne si archiválii nebo její fotokopii a tu v badatelně čte, a tak z ní „doluje" data, získává tak znalosti.[19]

Digitalizace a digitální archivy již řadu let umožnily zpřístupnění archiválií na dálku skrze webové aplikace. Toto přináší několik zásadních výhod – jednodušší přístup a možnosti vyhledávání v archiváliích pomocí filtrování a fulltextového vyhledávání. Badatel si tedy archiválii zobrazí, pročítá ji a tímto způsobem z ní „doluje" data, a tak získává znalosti.

Příliš se toho co se týče samotného získávání znalostí z archiválií nezměnilo. Problémem tohoto způsobu zpracování je jeho rychlost – pročítání jednotlivých archiválií zabere poměrně dost času a pokud badatel chce dělat větší analýzy nebo masivnější badatelskou činnost z mnoha archivních souborů, činí tak velmi složitě a zdlouhavě.

Právě při současných možnostech využití machine learning a deep learning technologií, sofistikovaných statistických analýz, dolování v datech, natural language processing technikách a dalších metodách přebraných z datové vědy vidí

---

[17] Goudarouli, E., Sexton, A. & Sheridan, J. The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK. Philos. Technol. 32, 173–183 (2019). https://doi.org/10.1007/s13347-018-0333-3

[18] Zde je důležité hovořit o uživatelích, což mohou být standardní badatelé z veřejnosti, historici ale i archiváři samotní pracující v daném archivu.

[19] Goudarouli, E., Sexton, A. & Sheridan, J. The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK. Philos. Technol. 32, 173–183 (2019). https://doi.org/10.1007/s13347-018-0333-3

odpověď, jak umožnit efektivnější a částečně automatizované zpracování[20] archiválií.

Velmi zajímavé jsou potom pokusy o definování nového, multidisciplinárního podoboru tzv. komputační archivní vědy (CAS). Například definice[21] z *Archival Records and Training in the Age of Big Data* z roku 2018: Multidisciplinární věda zabývající se aplikováním komputačních metod a jejích zdrojů pro masivní zpracování, analýzu, ukládání, dlouhodobou ochranu a prezentaci archivního materiálu s cílem zlepšit efektivitu, produktivitu a přesnost při rozhodovacím procesu při přijímání, uspořádání, popisu, uchovávání a zpřístupňování. Komputační archivní věda je kombinací archivního myšlení a komputačního myšlení.[22] Otázkou zůstává, zda se tento podobor uchytí. V současnosti to zatím vypadá na to, že se zatím začíná pomalu objevovat i v dalších pracích, např. *Computational Archival Science (2018)*, kde je velmi podobně definována, a co je důležitější – věnuje se velmi podobnému okruhu problémů.

Určitým úskalím, se kterým se potýká také české archivnictví je potom zpřístupnění velkého množství nezpracovaných archiválií. Samotná velikost archivních souborů je problematická.[23] S tím souvisí i proces digitalizace při kterém je použito OCR, pokud je text dostatečně jednolitý. Pokud tomu tak není, zpravidla je digitalizát dostupný pouze v obrazové podobě (prezenční kopie ve formátech

---

[20] A to nejen pro prostředí archivů, ale i muzeí, galerií a knihoven – obecně paměťových institucí. Tyto zmíněné metody a technologie umožňují mimo jiné efektivní, jednoduché a automatizované zpracování a vytváření různých typů grafiky pro prezentování získaných dat – různé typy grafů, histogramů a dalších možností prezentace dat. Toto může velmi dobře posloužit např. pro vytváření infografiky do muzejního prostředí. Zpracování je časově nenáročné, většinu práce odvede software pro statistickou analýzu – např. IBM SPSS.

[21] Jedná se o volný překlad autora.

[22] Marciano, R., Lemieux, V., Hedges, M., Esteva, M., Underwood, W., Kurtz, M. and Conrad, M. (2018), "Archival Records and Training in the Age of Big Data", Percell, J., Sarin, L.C., Jaeger, P.T. and Bertot, J.C. (Ed.) Re-envisioning the MLS: Perspectives on the Future of Library and Information Science Education (Advances in Librarianship, Vol. 44B), *Emerald Publishing Limited*, pp. 179-199. https://doi.org/10.1108/S0065-28302018000044B010

[23] Goudarouli, E., Sexton, A. & Sheridan, J. The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK. Philos. Technol. 32, 173–183 (2019). https://doi.org/10.1007/s13347-018-0333-3

JPEG). Rukopisné texty mohou být digitalizované, ale i tak mohou být do určité míry nepřístupné.

Ona nepřístupnost spočívá v tom, že badatel musí umět číst dané písmo, znát jazyk, poradit si s fyzickým poškozením a špatným stavem textu (anebo se musí obrátit na archiváře). Důležité je uvědomit si také to, že nepřístupné nejsou pouze lidskému oku, ale také automatizovaným zpracováním počítačem, což souvisí právě s předcházejícím bodem o zpracování archiválií, alespoň z části, automatizovaně.

Určitou odpovědí pro tuto masu nezpracovaných archiválií je potom využití komplexních softwarových řešení využívajících OCR a HTR technologií, a tím umožnit jejich širší zpřístupnění ve všech zmíněných významech slova. Nasazením těchto systémů a technologií přitom není tak daleko, jak by se mohlo mnohým zdát. Příkladem je projekt EU READ a jeho hlavní výstup v podobě platformy pro kompletní zpřístupňování rukopisných archiválií – Transkribus[24].

1.2. Automatizovaný přepis historických dokumentů

Problematika OCR a automatizovaného převodu dokumentů, jak ji známe nyní, sahá do 70. let minulého století[25], a i pro oblast archivnictví, do které se dostalo později, je použití OCR při digitalizačním procesu běžnou záležitostí, a to především pro tištěné[26] dokumenty.

Jednou obrovskou výhodou oblasti rozpoznávání a převodu dokumentů (rukopisných i tištěných) je její využití v komerčním sektoru[27], díky kterému je tato

[24] Platformě Transkribus je věnována samostatná kapitola 3.

[25] WILES, Rachel. Have we solved the problem of handwriting recognition?: Before Deep Learning, there were OCRs. *Towards data science* [online]. London, 2019 [cit. 2020-09-15]. Dostupné z: https://towardsdatascience.com/https-medium-com-rachelwiles-have-we-solved-the-problem-of-handwriting-recognition-712e279f373b

[26] OCR je stále relevantní i pro moderní archiválie a obecně dokumenty. Dle současného zákona a vyhlášky mají jen vybrané subjekty povinnost provádět spisovou službu čistě elektronicky, a tak je objem fyzických dokumentů stále relativně vysoký. Běžnou praxí potom je dokument si raději vytisknout, a tak se celé problematice práce s eSSL vyhnout. Určité změny bychom se mohli dočkat při vydání nového archivního zákona reagující na digitální ústavu, který je nyní v přípravě.

[27] Některými příklady využití jsou zefektivnění vnitřních procesů zpracování informací – např. vyplňování formulářů (obecně digitalizace a později jejich strojové zpracování).

problematika řešena intenzivněji, než pokud by se jednalo pouze o výzkum v oblasti archivnictví (a obecně humanitních věd).

OCR má mnoho specifických problémů a limitů při digitalizačním procesu v oblasti archivnictví, obecněji digitálních humanitních vědách. Firmy (obecně oblast podnikání) pohlíží na OCR jinak a nachází tak jiné limity a problémy, se kterými se musí vypořádat. Naštěstí pro řadu problémů pro obě oblasti (jak podnikání, tak archivní) lze nalézt průnik.

Pro obě oblasti je velmi důležitá přesnost výsledného převodu. V oblasti podnikání, kde se OCR používá, se jedná především o moderní dokumenty psané buď ručně (méně časté) nebo ve vybraném textovém procesoru (nejčastěji MS Word). Přesnost pro textové dokumenty je pro podnikání tedy už téměř zvládnutá. „*Traditional OCR could handle the easy stuff – about 80 percent of document workflows. For the more complicated stuff (like handwriting), humans had to intervene and perform manual data entry.*“[28]

Rizikem, se kterým se oblast archivnictví tolik nezabývá je potom oblast bezpečnosti. Samotný převod a přenos fyzického dokumentu a jeho obsahem mezi počítačem a člověkem anebo obráceně vytváří další bezpečnostní riziko především v oblastech, kde je informační bezpečnost důležitá (zdravotnictví, státní správa, bankovnictví, aj.). Odpovědí pro toto bezpečnostní riziko je kompletní přenos celého procesu práce s dokumentem do digitálního prostředí.

1.2.1. Současné OCR systémy a jejich možnosti

Současná praktika je taková, že se OCR při digitalizačním procesu aplikuje, pokud se jedná o tištěný text (standardní OCR systémy nedokážou převod rukopisných dokumentů). Obecně platí, že čím novější texty, tím úspěšnější OCR

Celkově problematika tzv. computer vision je velmi obsáhlá, a i největší světový giganti jako Microsoft a Google jsou v ní velmi zainteresováni.

[28] Can AI-powered OCR really read handwriting better than a human? Handwriting OCR. *SS&C* [online]. 2020 [cit. 2020-09-17]. Dostupné z: https://vidado.ai/ocr-poor-quality-docs/

je[29]. Pro přesnost je důležitá jednolitost textu, protože pokud je text nejednolitý, nastávají problémy především kvůli nekvalitnímu tisku, fyzickému poškození, vadám, škrtům a dalšími možnými vlastnostmi textu, který zásadně snižují efektivitu OCR systému a tím pádem i přesnost výsledného textu.

Po každém OCR výstupu je provedena ruční korekce[30]. Pro otázku konkrétních řešení je vypracováno mnoho výzkumů a vědeckých prací a v dnešní době je její aplikování a využití v podstatě zvládnuté[31]. Pomocí moderních OCR (Např. Tesseract) lze hovořit o využitelnosti i pro takové tištěné dokumenty, které spadají k poměrně ranným, novověkým tiskům.

Příkladem je práce *Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin,* z roku 2018, ve které autoři popisují možnosti moderního OCR s použitím trénovacích dat (tzv. data setů), které výrazně zvyšují procentuální úspěšnost jednotlivých OCR Tyto data sety totiž obsahují slovníky, které znesnadňují OCR klasifikaci jednotlivých znaků. *„Historical OCR has been advanced to a state where even very early printings from the 15th century can be recognized by individually trained models with a character recognition rate of 98% and above.“*[32]

Toto ale nereprezentuje většinu OCR systémů používaných při digitalizačních procesech jednotlivých archivů a funguje spíše jako ukázka toho, čeho všeho jsou OCR systémy schopné, pokud je proveden pečlivý výběr dat a vytvoření takového data setu[33]. K tomu, aby bylo dosaženo takovéto úrovně úspěšnosti je potřeba

---

[29] ROMEIN, Annemieke. Entangled Histories: OCR + HTR = ATR: Automatic Text Recognition. *KB LAB* [online]. 2020 [cit. 2020-09-15]. Dostupné z: https://lab.kb.nl/about-us/blog/entangled-histories-ocr-htr-atr-automatic-text-recognition

[30] Je možné mít sebepřesnější OCR, ale i tak je nutno provést korekci kvůli např. škrtnutým znakům a zmíněným vadám původního dokumentu.

[31] Toto neplatí pro samotný proces digitalizace. Doposud neexistuje žádný jednotný framework pro její kompletní proces v archivním prostředí. Výsledkem je nejednotnost napříč archivy. Každý tak má vlastní proces, což přináší určité problémy.

[32] SPRINGMANN, Uwe. Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. *JLCL* [online]. München, 2018, 1(33), 17 [cit. 2020-09-15]. Dostupné z: https://arxiv.org/abs/1809.05501v1

[33] Tento konkrétní data set má 313 173 řádků textu.

časově náročný výzkum a tým odborníků. Nejedná se tedy o něco, co by mohlo být efektivně používáno ve standardním archivu se současným pracovním vytížením.

Naopak i dnešní OCR mají stále jisté nedostatky i v oblasti tištěných dokumentů. Pro shrnutí současných možností OCR a jeho nedostatků vznikla v roce 2014[34] práce, *OCR of Historical Printings of Latin Texts: Problems, Prospects, Progress,* která přesněji prezentuje aktuálnější situaci v archivech.

Tato práce komparuje tři velmi hojně používané OCR řešení – Tesseract, OCRopus a Finereader, kdy velmi populární jsou především Tesseract a OCRopus, jelikož jsou šířeny s otevřeným zdrojovým kódem, zadarmo.

Existuje samozřejmě mnoho dalších řešení, ale v podstatě platí, že každým rokem vycházejí nové a nové aktualizace a verze již známých a populárních OCR řešení anebo vycházejí nové, které velmi často svojí výkonností předčí předcházející.

Ze zkoumaných řešení poskytuje největší přesnost a nejlepší výsledky zmíněné placené OCR Finereader.[35] Na druhém místě je potom OCRopus, který chybuje méně, ale zato stojí výrazně více výpočetního výkonu procesoru.

Table 1: Character accuracies in % for sample pages

Page	Finereader 11	Tesseract 3.03	OCRopus 0.7
15	87.79	80.88	80.70
16	82.94	77.41	76.94
17	85.25	75.98	86.07
18	85.93	79.51	85.53
19	87.94	80.09	79.09

Obr. 1 – tabulka přesnosti převodu OCR v procentech[36]

Výsledná tabulka ukazuje procenta úspěšně převedeného textu[37] na danou stranu pro jednotlivé znaky pro 5 různých stran textu. To znamená pro stranu 15

[34] Což je už poměrně starší práce, ale zatím nemá žádné obdoby.

[35] Springmann, Uwe & Najock, Dietmar & Morgenroth, Hermann & Schmid, Helmut & Gotscharek, Annette & Fink, Florian. (2014). OCR of historical printings of latin Texts: Problems, prospects, progress. ACM International Conference Proceeding Series. 10.1145/2595188.2595205.

[36] Springmann, Uwe & Najock, Dietmar & Morgenroth, Hermann & Schmid, Helmut & Gotscharek, Annette & Fink, Florian. (2014). OCR of historical printings of latin Texts: Problems, prospects, progress. ACM International Conference Proceeding Series. 10.1145/2595188.2595205.

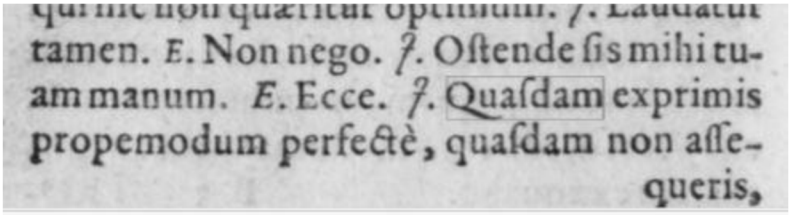
[37] Například pro stranu 15 bylo s pomocí OCR Finereader převedeno úspěšně 87,79 % textu.

byl v OCR Finereader každý jednotlivý znak programem jistě rozpoznán s přesností 87.79 %.

Přesnost OCR softwarů se udává v procentuální přesnosti na znak. Například pokud provádíme převod pro 100 znaků a máme 99 % přesnost, znamená to, že 99 znaků bylo rozpoznáno OCR jistě → OCR dokázalo úspěšně přiřadit znak k písmenu a 1 znak OCR nerozpoznalo → nedokázalo klasifikovat. (Výsledek může být správně anebo špatně; je neznámý)[38]

Ve výsledku je tedy Finereader nejefektivnější průměrnou úspěšností převedeného textu 85,97 %. Na druhém místě je OCRopus s 81,54 %. Na posledním místě je potom OCR systém Tesseract se 78,77 %.

OCR bylo aplikováno na knihu *Progymnasmata Latinatatis* z 16. století.


tamen. E. Non nego. ℣. Oſtende ſis mihi tu-
am manum. E. Ecce. ℣. Quaſdam exprimis
propemodum perfectè, quaſdam non aſſe-
queris,

Obr. 2 – ukázka textu knihy Progymnasmata Latinatatis[39]

Nedokonalost i současných OCR lze vypozorovat s prezentovaných dat. Chybovost se pohybuje mezi 12,06 % až 24,02 %, což jsou poměrně vysoká čísla.

I přes to, že se většina ze zmíněných softwarů každým rokem zlepšuje ve své výkonnosti, hlavním důvodem pro jisté obavy (a to především pro ty populární open source – Tesseract a OCRopus) je nutnost umět pracovat s příkazovou řádkou, což může být překážkou pro řadu archivářů a knihovníků.[40] Dalším problémem je potom již několikrát zmíněná nejednotnost v digitalizačním procesu jak

---

[38] OCR Accuracy Measurement. *ABBYY Technology Portal* [online]. [cit. 2020-09-15]. Dostupné z: https://abbyy.technology/en:kb:tip:ocr-accuracy

[39] Springmann, Uwe & Najock, Dietmar & Morgenroth, Hermann & Schmid, Helmut & Gotscharek, Annette & Fink, Florian. (2014). OCR of historical printings of latin Texts: Problems, prospects, progress. ACM International Conference Proceeding Series. 10.1145/2595188.2595205.

[40] MILIONI, Nikolina. *Automatic Transcription of Historical Documents* [online]. Uppsala, 2020 [cit. 2020-09-08]. Dostupné z: http://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1437985&dswid=-1804. Diplomová. Uppsala Universitet

v knihovnách (kde je situace o něco lepší) tak v archivech. Některá placená řešení nabízí přívětivější, grafické uživatelské rozhraní, ale zde je potom problémem právě finanční stránka věci, takže na jednu stranu existuje řada volně dostupných řešení, často se přibližující placeným řešením, avšak pouze s CLI, což stále je překážkou pro pracovníky působící v oblasti humanitních věd.

Vznik obdoby Základních pravidel pro zpracování archiválií nebo přímo daného frameworku pro oblast digitalizace by mohl přinést určitou jednotu umožňující efektivnější nasazení OCR systémů, obecně digitalizačního procesu a zároveň umožnit vznik a trénování zmíněných data setů, které mohou výrazně zvýšit úspěšnost převodu.

Pravděpodobně nejaktuálnějším a nám nejbližším současným projektem, zabývajícím se zpřístupněním archivních pramenů i pomocí moderních OCR technologií je spojený česko-německý projekt Porta fontium. *„Hlavním cílem projektu je zpřístupnění archivních pramenů z česko-bavorského příhraničí širokému spektru uživatelů z řad odborné i laické veřejnosti pomocí nejmodernějších informačních technologií.“*[41] Projekt působí od roku 2014 a měl by skončit v roce 2021. V současnosti je zprovozněný webový portál a na něm zpřístupněna řada archiválií, i když zatím bez OCR. Význam má tento projekt také proto, že je pořadatelem workshopů[42], které se zabývají moderním archivnictvím a současnými problémy včetně využití moderních IT. Projekt dal za vznik i několika vědeckým článkům, kdy lze zmínit například práci *Building an efficient OCR system for historical documents with little training data* z roku 2020, která se zabývá právě využitím moderního OCR za použití neurální sítě, popisuje celý proces práce, současné nástroje a pracovní postupy, výsledky dosažené chybovosti[43].

---

[41] Moderní zpřístupnění historických pramenů (2018-2021). *Porta fontium* [online]. 2020 [cit. 2020-10-13]. Dostupné z: http://www.portafontium.cz/article/moderni-zpristupneni-historickych-pramenu

[42] Lze zmínit například poslední workshop s názvem Archivní rešerše v 21. století.

[43] MARTÍNEK, J., L. LENCL a P. KRÁL. Building an efficient OCR system for historical documents with little training data: Existing tools and OCR systems. *Neural*

1.2.2. HTR systémy a jejich současný stav

Druhou, významnou technologií umožňující zpřístupnění (převod) historických dokumentů je technologie HTR. Vývoj této technologie se datuje až do 50. let minulého století, kdy ale z důvodu nedostatku výpočetního výkonu nedošlo k výraznějším posunům až doposud.

Masivní rozvoj neurálních sítí a výrazné zvýšení dostupného výpočetního výkonu během posledních 15 let umožnilo další výrazný vývoj HTR technologie pro její využití v oblasti rozpoznávání rukopisných historických dokumentů.[44]

HTR je technologie, která umožňuje pomocí modelu neurální sítě automatizovaný přepis rukopisného archivního materiálu. Ve stručnosti[45] funguje tak, že pro vybranou archivní sbírku nebo fond, je provedena částečná manuální transkripce 15 000 – 20 000 slov. Tato tréninková data jsou potom nahrána do systému, který se pomocí nich naučí automatizovaný přepis zbytku fondu/sbírky.[46]

Každým rokem[47] se pořádá konference ICDAR[48], která se obecně zabývá analýzou dokumentů a rozpoznáváním znaků. Má tedy i určitou souvislost s oblastí historických dokumentů. Hlavními položkami této každoroční konference jsou

---

Computing and Applications [online]. 2020 [cit. 2020-10-17]. Dostupné z: doi:https://doi.org/10.1007/s00521-020-04910-x

[44] Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, P., Kallio, M., Kaplan, F., Kleber, F., Labahn, R., Lang, E.M., Laube, S., Leifert, G., Louloudis, G., McNicholl, R., Meunier, J.-L., Michael, J., Mühlbauer, E., Philipp, N., Pratikakis, I., Puigcerver Pérez, J., Putz, H., Retsinas, G., Romero, V., Sablatnig, R., Sánchez, J.A., Schofield, P., Sfikas, G., Sieber, C., Stamatopoulos, N., Strauß, T., Terbul, T., Toselli, A.H., Ulreich, B., Villegas, M., Vidal, E., Walcher, J., Weidemann, M., Wurster, H. and Zagoris, K. (2019), "*Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study*", Journal of Documentation, Vol. 75 No. 5, pp. 954-976. https://doi.org/10.1108/JD-07-2018-0114

[45] Podrobný popis celého workflow pro platformu Transkribu, kde je HTR důležitou součástí, je popsán v další hlavní kapitole. Zde pouze jako základní nástin pro lepší pochopení podkapitoly.

[46] Handwritten Text Recognition Workflow: Basic Workflow. *Transkribus* [online]. 2018 [cit. 2020-09-17]. Dostupné z: https://transkribus.eu/wiki/index.php/Handwritten_Text_Recognition_Workflow

[47] Letošní konference pro rok 2020 byla zrušena kvůli Covid-19 pandemii.

[48] International Conference on Document Analysis and Recognition

analýza dokumentů a jejich rozpoznávání, využití umělé inteligence, rozpoznávání vzorů v textu, analýza rukopisů a jejich ověřování, rozpoznávání textu a jeho detekce a další související oblasti.[49]

Druhou konferencí, která se každoročně pořádá je konference ICFHR – International Conference on Frontiers of Handwriting Recognition. Tato se úžeji specializuje na rozpoznávání rukou psaných textů. Konference se účastní experti z různorodých oblastí – informatici, archiváři, knihovníci, lingvisté, datový vědci aj. Mezi hlavní oblasti této konference patří rozpoznávání rukopisů, rozpoznávání kurzivního písma, symbolů, matematických vzorců, forenzní analýzy, znaková písma aj.[50]

Výsledkem těchto dvou konferencí je řada vědeckých prací a poznatků, společně s novou testovací sadou dokumentů, které se používají pro soutěž. Každý rok je vydán jiný seznam dokumentů, pro které se soutěž bude konat (jednou kurzivní text, podruhé mandarínštinou psané texty aj.). Cílem soutěže je vytvořit takový model neurální sítě, který bude mít největší úspěšnost v rozpoznávání znaků v daných dokumentech. Výsledkem je tedy řada modelů neurálních sítí, které se snaží o co nejvyšší úspěšnost v dané soutěži.[51] Důležité je, že touto soutěží vznikne každý rok mnoho sesbíraných dat (benchmarků[52]), podle kterých se potom další rok řídí novější a novější modely neurálních sítí a tím se vývoj posouvá rychleji dopředu. Obě konference jsou tedy pro oblast jak OCR, tak HTR velmi důležité a napomáhají vědeckému snažení a bádání v těchto oblastech.

---

[49] ICDAR2019: Program Booklet. *ICDAR 2019* [online]. Sydney, 2019 [cit. 2020-09-19]. Dostupné z: http://icdar2019.org/program-booklet/

[50] ICFHR: International Conference on Frontiers in Handwriting Recognition. *WikiCFP* [online]. 2020 [cit. 2020-09-19]. Dostupné z: http://www.wikicfp.com/cfp/program?id=1366&f=International%20Conference%20on%20Frontiers%20in%20Handwriting%20Recognition

[51] A set of benchmarks for Handwritten Text Recognition on historical documents. *Pattern Recognition* [online]. 2019, (94), 133 [cit. 2020-09-19]. Dostupné z: https://www.sciencedirect.com/science/article/abs/pii/S0031320319302006

[52] Určitých standardizovaných data setů, podle kterých lze velmi přesně určit chybovost jednotlivých nástrojů, a tak je mezi sebou efektivně porovnat.

Stejně jako u OCR existují komerční i volně dostupná řešení. V tomto případě se jedná ale o mnohem složitější technologií, než kterou je OCR, a tudíž zdaleka není taková možnost výběru jako právě u OCR.

Pro paměťové instituce existují (v době psaní této práce) v podstatě dvě řešení, kdy obě nabízejí velmi podobné služby. Jedno komerční a druhé volně dostupné.

1.2.2.1. Komerční řešení Quartex

Komerčním řešením je služba Quartex od firmy Adam Matthew Digital. Toto softwarové řešení (ve formě služby) umožňuje knihovnám a archivům a dalším institucím zpřístupňovat jejich obsah právě pomocí HTR technologie. Samotná firma o sobě v *Adam Drewe, Head of Platform Services at Adam Matthew, about Quartex* tvrdí, že toto řešení umí zpřístupnit a prezentovat různé druhy archiválií, kodexy, korespondenci, tištěné knihy, noviny, videa. Podporuje všechny běžné obrazové, zvukové i video formáty.[53] Jelikož se ale jedná o komerční řešení, které je finančně náročné, není k němu dostupných toliko vědeckých prací a článků[54].

Hlavní výhody této platformy lze spatřit v tom, že podporuje komplexní zpřístupnění jakýchkoliv dat včetně zvuku a video. Druhou výhodou je možné využití modernějšího modelu neurální sítě, jelikož tato platforma začala fungovat v roce 2018. Platforma se prezentuje jako kompletní řešení prezentace archivního materiálu, se vším, co s tím souvisí – převod z analogové do digitální podoby, provedení HTR a vznik textové podoby, který lze prezentovat pomocí webové aplikace a s možností filtrovacího systému. Zde je potřeba se znovu odkázat na fakt, že prozatím k tomuto tématu neexistuje dostatek vědeckých článků a výzkumu, aby bylo možné si udělat lepší obrázek.

Určité reálné výsledky pro použití tohoto systému už ale existují. Národní archiv Spojeného království pomocí tohoto systému již zpřístupnil přes 70 000

---

[53] Head of Platform Services at Adam Matthew Digital interviewed in The Charleston Advisor. *Quartex* [online]. 2018 [cit. 2020-09-17]. Dostupné z: https://www.quartexcollections.com/news/item/head-of-platform-services-at-adam-matthew-talks-about-quartex

[54] Po zadání hesla Quartex a HTR do Google Scholar je v době psaní této práce (2020) dostupné pouze citované interview a krátká zmínka (v práci citovaném) článku Transforming scholarship in the archives through handwritten text recognition

archiválií z období kolonialismu 1606–1822. Výsledkem je nejen kompletní webové prostředí, ale také možnost úplného vyhledávání skrze všechny manuskripty pomocí fulltextového vyhledávání[55] napříč dvěma staletími textů různých písařů, rodin a typů různých písem a jazyků[56]. Vyhledávání funguje tak, že do fulltextového pole zadáte heslo (například jméno osoby, o kterou má badatel zájem). Toto vyhledané slovo se potom přímo zobrazí ve všech textech, kde se vyskytuje s tím, že si uživatel sám vybere, který text si chce prohlédnout podrobněji.

Obecně ale lze říct, že v současnosti toto řešení nabízí obdobný software jako Transkribus a umožňuje využití HTR technologie pro vyhledávání a zpřístupňování archivních sbírek a fondů.[57]

Naopak zásadním problémem s komerčními řešeními je jejich finanční náročnost – je potřeba tým odborníků, neurální síťový model a také dostatek výpočetního výkonu. Všechny tyto věci mohou velmi ztížit přístup vědců a dalších uživatelů pro oblast humanitních věd. V posledních letech navíc začínají vznikat komplexní komerční digitalizační služby, které ale mohou být pro hůře financované archivy či knihovny méně dostupné.

---

[55] Colonial America: Complete CO5 files from The National Archives, UK, 1606-1822. Adam Matthew *A SAGE Publishing Company* [online]. 2020 [cit. 2020-10-13]. Dostupné z: https://www.amdigital.co.uk/primary-sources/colonial-america

[56] Vyhledávání podporuje španělštinu a angličtinu. Demonstrační video si lze prohlédnout zde: https://www.youtube.com/watch?v=ChWMZMpivbQ&feature=emb_title

[57] Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, P., Kallio, M., Kaplan, F., Kleber, F., Labahn, R., Lang, E.M., Laube, S., Leifert, G., Louloudis, G., McNicholl, R., Meunier, J.-L., Michael, J., Mühlbauer, E., Philipp, N., Pratikakis, I., Puigcerver Pérez, J., Putz, H., Retsinas, G., Romero, V., Sablatnig, R., Sánchez, J.A., Schofield, P., Sfikas, G., Sieber, C., Stamatopoulos, N., Strauß, T., Terbul, T., Toselli, A.H., Ulreich, B., Villegas, M., Vidal, E., Walcher, J., Weidemann, M., Wurster, H. and Zagoris, K. (2019), "*Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study*", Journal of Documentation, Vol. 75 No. 5, s. 954-976. https://doi.org/10.1108/JD-07-2018-0114

1.2.2.2. Transkribus a projekt READ

Transkribus vznikl původně z projektu EU Transcriptorium, který probíhal od roku 2013 do roku 2015 a měl za cíl vytvořit inovativní, efektivní a finančně přístupné řešení pro indexování, vyhledávání a úplnou transkripci historických rukopisů v podobě obrazového digitalizátu za použití právě HTR technologie.[58]

„Within the tranScriptorium project span, we intend to apply the developed HTR technology to historical documents in cursive handwriting, for which only HTR technology can offer appropriate solutions.“[59]

V roce 2016 se projekt Transcriptorium transformoval do rozsáhlejšího projektu READ, znovu financovaný Evropskou Unií. Projektu READ předcházela masivní výzkumná činnost, kdy i z řady zmíněných vědeckých článků lze zjistit, že byli sponzorovány právě z projektu READ. Řada z nich je také dostupná zadarmo bez nutnosti využití univerzitních sítí nebo zpřístupnění článku z vlastních finančních prostředků.

Projekt má za cíl udržet, vyvíjet a zajišťovat další vzdělávání pro funkční badatelskou online infrastrukturu, kde nové technologie umožňují inovaci v archivním výzkumu.[60]

Hlavní výstupy tohoto masivního projektu jsou především tři na sebe navazující části tvořící úplné řešení celého digitalizačního procesu, od samotného převodu z fyzické do digitální podoby, k jeho automatizované transkripci, až po možnosti prezentace a vyhledávání v takto digitalizovaných rukopisných dokumentech.

---

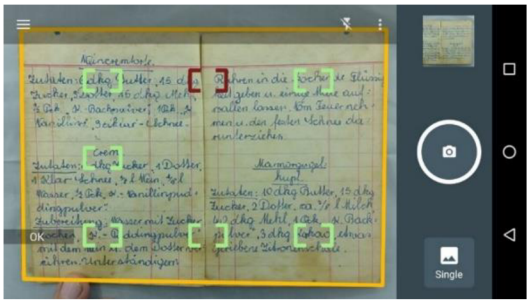
[58] TranScriptorium. *TranScriptorium* [online]. 2013 [cit. 2020-09-17]. Dostupné z: http://transcriptorium.eu/

[59] TranScriptorium. Objectives [online]. 2013 [cit. 2020-09-17]. Dostupné z: http://transcriptorium.eu/

[60] Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation* [online]. Emerald Publishing Limited, 2019, 75(5), 2 [cit. 2020-09-17]. ISSN 0022-0418. Dostupné z: https://www.emerald.com/insight/content/doi/10.1108/JD-07-2018-0114/full/html

První částí je tzv. scantent. Výsledkem je vytvoření kompletního, levného[61] řešení umožňující digitalizaci[62] fyzických archiválií.[63] První součástí tohoto řešení je volně dostupná mobilní aplikace DocScan, která umožňuje skenování historických dokumentů pomocí fotoaparátu v mobilním telefonu.

Zde potom může vyvstat legitimní otázka, zda je použití integrovaného fotoaparátu z mobilního telefonu dostatečné pro archivní potřeby.[64] Samozřejmě využití mobilního telefonu v archivním prostředí, kde se využívají profesionální skenery při digitalizaci, nemá místo. Určité využití lze nalézt pro potřeby vzdělávání studentů a školení archivářů při digitalizačním procesu jako poměrně levná a dostupná alternativa pro drahé a málo dostupné profesionální skenery.



Obr. 3 – ukázka uživatelského prostředí aplikace DocScan[65]

Aplikace umožňuje nejen focení, ale také nastavení parametrů fotoaparátu, ale také automatické rozpoznávání pohybu a focení při otáčení stránek.

Společně s aplikací tvoří scantent i tento malý, černý stan s konstrukcí, která umožňuje zachycení telefonu na jeho vrchu a s dodávaným LED osvětlením si poradí i s neoptimálními světelnými podmínkami.

---

[61] Celé řešení stojí 240 euro, přibližně 6300 Kč. Možná se toto zdá hodně, ale v porovnání s profesionálními skenery nutnými k běžné digitalizaci je to relativně malá částka. Na druhou stranu se vlastně jedná pouze o stan s LED světlem a černou podložkou pro dokument a cena 6300 Kč se může v tomto případě zdát jako poměrně vysoká.

[62] Ve smyslu převodu z analogové (fyzické) podoby do digitální podoby.

[63] ScanTent + DocScan App. *READ-COOP* [online]. 2020 [cit. 2020-09-17]. Dostupné z: https://readcoop.eu/scantent/

[64] V některých případech se kvalita fotoaparátů s použitím zmíněné aplikace dokáže vyrovnat s běžným fotoaparátem, v jiných případech ne. Velmi záleží na typu telefonu. Pro řadu telefonů sice 300 dpi není dosažitelných, ale je to spíše otázkou času, než se tomuto rozlišení průměrné telefony přiblíží. (Ty dražší už tuto hranici překonali).

[65] ScanTent + DocScan App. *READ-COOP* [online]. 2020 [cit. 2020-09-17]. Dostupné z: https://readcoop.eu/scantent/

Stan[66] má společně s aplikací DocScan a fotoaparátem integrovaným v současných mobilních zařízeních imitovat skenery a DSLR v archivním prostředí. Jedná se o poměrně levné a přenosné řešení pro převod analogového dokumentu do digitální podoby.

Obr. 4 – ukázka zmíněné stanové konstrukce pro skenování[67]

Druhou, a tou hlavní, částí je potom samotná platforma Transkribus. Transkribus je komprehensivní platforma pro digitalizaci, rozpoznávání, transkripci a vyhledávání historických dokumentů.[68]

Jádrem všeho je neurální síť, která je schopná naučit se podle poskytnutých trénovacích dat rozpoznávat a transkribovat dané rukopisné dokumenty stejného typu písma a jazyku.

Ve stručnosti[69] funguje tak, že daný uživatel si vybere archivní sbírku, kterou chce transkribovat. Obrazové soubory získané buď ze Scantent nebo samostatně nahraje do systému skrze volně dostupného klienta. Sám potom musí provést transkripci 15 000 – 20 000 slov. Takto transkribovanou část archivní sbírky potom

---

[66] Výsledné fotografie z tohoto prostředí za použití několika současných populárních chytrých telefonů si lze prohlédnout na https://readcoop.eu/scantent/

[67] The Austrian government meets READ DocScan and ScanTent! *READ-COOP* [online]. 2017 [cit. 2020-09-17]. Dostupné z: https://readcoop.eu/austrian-govt-meets-docsan-and-scantent/

[68] Transkribus. READ-COOP [online]. [cit. 2020-09-19]. Dostupné z: https://readcoop.eu/transkribus/

[69] Kompletní proces je popsán v samostatné kapitole Transkribus.

nahraje do systému, který podle těchto tréninkových dat zvládne sám zbytek celé sbírky transkribovat.

Velkou výhodou celé platformy je jeho poměrně velká uživatelská základna. Ta má více než 30 000 uživatelů[70], kdy mnoho z nich již Transkribus využilo pro vytvoření různých kompletně transkribovaných sbírek. Mezi ně patří univerzity, knihovny, archivy a další instituce zabývající se Digital Humanities po celé Evropě (zástupci v Německu, Nizozemí, Rakousku, Polsku, Velké Británii, Finsku, Švédsku aj.)

Z těch větších lze zmínit například Národní archiv Nizozemska, které pomocí platformy Transkribus digitalizují a zpřístupňují dokumenty z provenience Holandské Východoindické společnosti. Celková sbírka čítá 475 769 slov s výslednou chybovostí programu Transkribus 7.48 %.[71] Tato transkribovaná data chtějí využít při dalších digitalizačních procesech a celá je dostupná zadarmo všem uživatelům.

Nejbližší pro naše prostředí bude pravděpodobně projekt v Pasovském diecézním archivu, kde pomocí Transkribu provádí transkripci a vyhledávání v místní velké kolekci sakramentálních pramenů, skládající se z 400 000 slov psaných rukou až čtyřiceti písařů. Výsledná chybovost se pohybuje mezi 17 až 19 procenty,[72] což především z důvodu náročnosti textu a velkého množství písařů[73] je přijatelný výsledek, protože chybovost slov do 30 % ještě umožňuje člověku text pochopit a použít[74].

---

[70] Transkribus. *READ-COOP* [online]. [cit. 2020-09-19]. Dostupné z: https://readcoop.eu/transkribus/

[71] National Archives releases first version of a Dutch handwriting model. *READ-COOP* [online]. 2019 [cit. 2020-09-19]. Dostupné z: https://readcoop.eu/national-archives-releases-first-version-of-a-dutch-handwriting-model/

[72] Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation* [online]. Emerald Publishing Limited, 2019, 75(5), 954-976 [cit. 2020-09-19]. Dostupné z: https://www.emerald.com/insight/publication/issn/0022-0418

[73] Protože každý písař má trochu jiný styl psaní a další charakteristické znaky.

[74] KATUŠČÁK, Dušan. Digital Humanities a automatická transkripcia rukopisných textov. *ITlib* [online]. Ministerstvo školstva, vedy, výskumu a športu Slovenskej republiky, 2020, 2020(1) [cit. 2020-09-19]. Dostupné z:

Mezi další, velké uživatele platformy patří Národní archiv Finska, který momentálně pracuje na vytvoření webové aplikace schopné vyhledávat ve finských soudních záznamech od roku 1810 po rok 1870 právě za pomocí Transkribu.

Na oficiálních stránkách se uvádí, že se Transkribus dá využít pro dokumenty psané v arabštině, latině, angličtině, němčině, polštině, hejbrejštině, holandštině a bengálštině[75]. Tento výčet ale zcela jistě není kompletní, a zasloužil by si aktualizaci, protože už ze samotné podstaty Transkribu, který je založen na vytváření tréninkových dat, která se potom nahrají do systému a poskytnou neurální síti tréninková data, podle kterých potom dojde k transkripci, by se zcela jistě dal využít pro potřeby českého jazyka i mnoha dalších jazyků. (Běžně jej lze totiž použít pro slovenštinu a polštinu).

Otázka úspěšnosti a využitelnosti Transkribu je stále velmi populární a tak na ni vzniká řada vědeckých článků. Z našeho prostředí lze zmínit práci profesora Katuščáka[76] *Digital Humanities a automatická transkripce rukopisných textů* z roku 2020, kdy popisuje svůj postup a výsledky při použití Transkribu při digitalizaci a transkripci rukopisných listů Andreje Kmeťa. Důležitost práce navíc spočívá i v tom, že autor využil i zmíněný scantent. „*Naše skúsenosti overené experimentom potvrdzujú, že jednotlivé rukopisy možno automaticky transkribovať, … Výsledky transkripcie sú čitateľné, použiteľné a možno ich exportovať …, editovať, redigovať, korigovať. V experimente sme dosiahli chybovosť (CER) 1,76 %. …. Chybovosť slov 16,88%.*“[77] Ve výsledku se jeví Transkribus jako

---

https://itlib.cvtisr.sk/archiv/2020/1/digital-humanities-a-automaticka-transkripcia-rukopisnych-textov-digital-humanities-and-automatic-transcription-of-handwritten-texts.html?page_id=3698

[75] Transkribus. *READ-COOP* [online]. [cit. 2020-09-19]. Dostupné z: https://readcoop.eu/transkribus/

[76] prof. PhDr. Dušan Katuščák, PhD. Působí na Slezské univerzitě v Opavě při Ústavu bohemistiky a knihovnictví. Pro zmíněnou práci byl Transkribus použit pro slovenský jazyk.

[77] KATUŠČÁK, Dušan. Digital Humanities a automatická transkripcia rukopisných textov. *ITlib* [online]. Ministerstvo školstva, vedy, výskumu a športu Slovenskej republiky, 2020, 2020(1) [cit. 2020-09-19]. Dostupné z: https://itlib.cvtisr.sk/archiv/2020/1/digital-humanities-a-automaticka-transkripcia-rukopisnych-textov-digital-humanities-and-automatic-transcription-of-handwritten-texts.html?page_id=3698

použitelný a vhodný nejen pro velké a střední projekty (zmíněné Finsko, Nizozemsko, Německo), ale i projekty menšího rázu, jako například práce profesora Matuščáka.

Co se týče možností a výběru současných řešení a vědeckého snažení, tak digitalizační procesy a vytváření textové podoby obrazových digitalizátů se dělí do dvou větví – pro rukopisné dokumenty je výsledným řešením z větší části Transkribus. Pro oblast tištěných dokumentů je stále řešením využití moderních pokročilejších OCR. Například projekt IMPACT, sponzorovaný Evropskou unií poskytuje řadu nástrojů[78] právě pro digitalizaci tištěných textů. Cílem tohoto projektu je digitalizaci historických tištěných textů lepší, levnější a rychlejší.[79]

Lze zmínit i související projekt NewsEye, který je fundován z projektu Horizon 2020 (EU). Tento projekt má za cíl vývoj nových konceptů, metod a nástrojů Digital Humanities pro zpřístupňování historických novin pro široké publikum badatelů. Projekt si klade za cíl změnit přístup k digitálním historickým datům – jak jsou zkoumána, zpřístupňována, používáná a analyzována.[80] Zatím je ale spíše v počátcích (vznikl letošní rok, 2020), a tak je jeho výsledek ještě daleko.

Až donedávna byl pro uživatele klient Transkribus dostupný zadarmo, stačilo se registrovat na webových stránkách[81] a program si stáhnout. Koncem roku 2020 ale Transkribus přešel na placený model především pro udržitelnost jeho funkčnosti, protože již vypršelo financování z grantů. Model je nyní nastaven na financování podle kreditů, kdy koupí jednoho kreditu lze transkribovat jednu stranu textu. Pro univerzity a další vědecké pracovníky včetně studentů existují možnosti získání kreditů zadarmo anebo výrazně levněji. Kredity stojí méně v případě

---

[78] Až 287 nástrojů od OCR, image processing, evaluace, korekčních nástrojů, post processing, obrazové segmentace, tréninkových dat aj.

[79] About us. *Impact digitisation.eu* [online]. 2020 [cit. 2020-09-19]. Dostupné z: https://www.digitisation.eu/about/

[80] About: What is NewsEye about? *NewsEye* [online]. 2020 [cit. 2020-09-19]. Dostupné z: https://www.newseye.eu/about/

[81] www.transkribus.eu

placené subskripce a její délce trvání, kdy delší subskripce rovná se výhodnější nákup kreditů.[82]

Posledním výstupem projektu READ je samotná prezentace a vyhledávání v kompletně digitalizovaných archiváliích. Jedná se o dedikované řešení pro prezentování a vyhledávání v takto převedených dokumentech skrze webové prostředí a také vytváření digitálních edic.[83]

Nejedná se pouze o webové prostředí, ale také o vyhledávací engine schopný full-textového vyhledávání, které využívá právě textovou podobu dokumentů. Ve výsledku umožňuje vyhledávat v jakémkoliv dostupném textu přesná slova a pasáže. Celé je toto dostupné také na mobilních zařízeních a celkově se jedná o moderní, responzivní webovou aplikaci zakončující celý proces.

Projekt READ tedy přináší téměř úplné řešení celého digitalizačního procesu od samotného převodu do digitální podoby (Scantent) k samotné transkripci dokumentu a jeho vytváření jeho základního metadatového popisu (Transkribus) až k jeho prezentování a vyhledávání (Read&Search). V některých oblastech toto kompletní řešení ukazuje své nedostatky, ale na druhou stranu je využitelné pro všechny instituce, které chtějí prezentovat své historické dokumenty.

Specializované instituce budou určitě chtít nahradit alespoň některé části celého procesu, například zmíněnou stanovou konstrukci a své, takto digitalizované, dokumenty prezentovat na svých vlastních webových portálech[84].

V roce 2019 byl READ transformován do dnešní podoby READ-COOP. Tato transformace ukončila základní část výzkumu, v této nové podobě má především za úkol udržovat a dál vyvíjet platformu Transkribus.

Velmi výraznou výhodou této platformy Transkribus je především jeho dostupnost, a to především v kontrastu s komerčním řešením Quartex. Velmi dlouho byl dostupný volně a jeho platební model je ze současných možností

---

[82] Transkribus Credits: About Transkribus Credits. *Transkribus* [online]. 2020 [cit. 2020-10-20]. Dostupné z: https://readcoop.eu/transkribus/credits/

[83] Read&search. *READ-COOP* [online]. [cit. 2020-09-18]. Dostupné z: https://readcoop.eu/readsearch/

[84] I když jejich nahrání do prostředí Read&Search nic nestojí a servery provozuje READ-COOP, takže instituce musí řešit pouze nahrání a metadatový popis.

nejmírnější, a i přes to, že byl spuštěn už v roce 2016, za konkurencí zatím nijak neztrácí (v roce 2020 dostal nový HTR engine). Důležité je zdůraznit, že dokud existuje toto dostupné řešení, a to i za předpokladu, že bude méně výkonné než ta komerční, můžeme zcela jistě počítat s tím, že zůstane tím nejpopulárnějším a nejvyužívanějším[85], což jej již z charakteristiky HTR technologie velmi důležité, protože uživatelská základna je pro efektivní využití této technologie zásadní. Projekt READ se svým dostupným řešením Transkribus umožňuje prakticky využít moderní technologie, které by byli archivům jinak, z především finančních důvodů, nepřístupné.[86]

1.3. Využití textových digitalizátů pomocí moderních IT

Fakt, že moderní informační technologie prožívají obrovský a velmi dynamický rozvoj, je velmi známý, a velmi často se ho využívá jako hlavní argument pro oblast digitálního archivnictví a řešení uchovávání, ochrany a důvěry právě digitálních archiválií. Poukazuje se na nestálost a nepředvídatelnost tohoto rozvoje, jak uvádí Clayton Christensen v *Innovator´s Dilemma*[87].

Nejedná se ale pouze jen o překážku. Právě kvůli masivnímu rozvoji těchto technologií lze docílit výsledků, které by byly jen těžko myslitelné, což dokládají výsledky zmíněného Transkribu a automatizace přepisu rozsáhlých sbírek historických dokumentů.

---

[85] Pokud nevznikne nějaké další řešení, které by výrazně zrychlilo celkový proces práce s programem Transkribus.

[86] A set of benchmarks for Handwritten Text Recognition on historical documents. *Pattern Recognition* [online]. 2019, (94), 133 [cit. 2020-09-19]. Dostupné z: https://www.sciencedirect.com/science/article/abs/pii/S0031320319302006

[87] Autor se v knize zabývá tím, jak jsou nové technologie nepředvídatelné. Christensen zmiňuje, že i když je firma při výrobě produktu úspěšná a stabilní, určité disruptivní technologie jí mohou naprosto zničit, a to se ani nemusí jednat o výkonnější technologii. Jednou z ukázek je oblast fotoaparátů. Od 80. let zhruba do roku 2010 se velmi hojně rozvíjeli kompaktní fotoaparáty, které byly dostatečně výkonné, levné a dostupné. S příchodem chytrých telefonů se tento trh téměř zhroutil a velmi se propadl. Fotoaparáty v chytrých telefonech nedosahovali ani zdaleka takových výsledků jako kompakty, přesto se jejich trh propadl a dnes už se s nimi setkáme téměř výhradně v poloprofesionální a profesionální oblasti.

1.3.1. Computer Vision

Obecně lze Computer Vision definovat jako oblast umělé inteligence, kde počítače vidí a rozumějí[88] datům ze souborů v obrazovém nebo video formátu.[89] Velmi často se používá při analýze obrazu – např. pro využití v moderních autonomně řídících vozech nebo v rozpoznávání povoleného obsahu[90].

Jelikož je Computer Vision velmi populární a dotovanou oblastí umělé inteligence, tak velmi rychle a dynamicky roste a rozvíjí se. V posledních pár letech se pomalu dostává i do oblasti archivnictví a obecně humanitních věd. Stále ale platí, že velké projekty v oblasti Digital Humanities se zabývají spíše analýzou textu[91] a tato oblast se těší menšímu zájmu, i když v posledních letech roste.[92]

V současnosti se použití Computer Vision technik používá především pro řešení specifických projektů a výsledné práce fungují jako jakési ukázky technologie a jejího využití.

Konkrétním příkladem je např. práce *Analyzing and visualizing ancient Maya hierogplyhics using shape: From computer vision to Digital Humanities.* V této práci autoři právě využívají technologie Computer Vision pro analýzu Mayských hieroglyfů za pomocí nástroje HOOSC[93], a hlavně automatizaci celého procesu rozpoznávání Mayských hieroglyfů. Deskriptor dokáže jednotlivé glyfy rozpoznat

---

[88] Tj. dokážou z obrazu, ať pohyblivého nebo statického, přečíst informace a vyvodit z nich patřičný závěr.

[89] What is Computer Vision? *DeepAI* [online]. [cit. 2020-09-19]. Dostupné z: https://deepai.org/machine-learning-glossary-and-terms/computer-vision

[90] Běžným příkladem je využití na YouTube pro rozpoznávání povoleného obsahu – rozpoznává a maže příliš násilný obsah, sexuální obsah, aj.

[91] SMITS, Thomas. Illustrations to Photographs: Using computer vision to analyse news pictures in Dutch newspapers, 1860-1940. *DH* [online]. Raboud University, The Netherlands, 2017 [cit. 2020-09-22]. Dostupné z: https://www.semanticscholar.org/paper/Illustrations-to-Photographs%3A-using-computer-vision-Smits/ad6ac997dcafe9b975d5758578fd3dd19fbd2ccb

[92] KLEPPE, M., M. LINCOLN a T. SMITS. *Computer Vision in Digital Humanities. DH* [online]. 2017 [cit. 2020-09-22]. Dostupné z: https://www.semanticscholar.org/paper/Computer-Vision-in-Digital-Humanities-Kleppe-Lincoln/07b7bbe8dd59e3bd2b63babdcd43b76e019ede7d?p2df

[93] Tento nástroj se používá například pro rozpoznávání a kategorizaci čínských znaků nebo egyptských hieroglyfů.

a porovnat s dalšími dostupnými hieroglyfy.[94] Výsledkem je webová aplikace, která umožňuje rozpoznávat Mayské hieroglyfy a přiřazovat je k již známým hieroglyfům.

Dalším příkladem a pro naše prostředí bližší je práce *Illustration to Photographs: Using computer vision to analyse news pictures in Dutch newspapers, 1860-1940.*, kdy za pomocí analýzy 2 vlastností obrazu: poměru logických a reálných pixelů (poměr pixelů) a množství informací obsažených v obrazu (úroveň entropie) lze automaticky klasifikovat[95], zda se jedná o ilustraci, nebo fotografii.[96]

V současnosti neexistuje nějaký větší projekt, zabývající se oblastí Computer Vision přímo v archivním prostředí. U nás se tedy jeví její nasazení spíše také pro specifické projekty a problémy, například pro zpracování velkých sbírek novin, pečetí, razítek, ilustrací a fotografií.

Využití této technologie si lze představit například u klasifikace takových archivních a muzejních předmětů, které lze systematicky klasifikovat a rozpoznat – pečetě, razítka a například i vexilologické, faleristické a heraldické sbírky. Přitom je velmi důležité mít data, podle kterých může program předmět klasifikovat – zdigitalizované slovníky a katalogy[97].

---

[94] HUI, Rui, Carlos PALLAN, Jean-Marc ODOBEZ a Daniel GATICA-PEREZ. Analyzing and visualizing ancient Maya hieroglyphics using shape: From computer vision to Digital Humanities. *Digital Scholarship in the Humanities* [online]. 2017, 2(32), 179-194 [cit. 2020-09-22]. Dostupné z: https://www.researchgate.net/publication/322633817_Analyzing_and_visualizing_ancient_Maya_hieroglyphics_using_shape_From_computer_vision_to_Digital_Humanities

[95] Fotografie mají vysokou úroveň entropie a vysoký poměr reálných a logických pixelů. Ilustrace mají naopak nízkou úroveň entropie a nižší poměr pixelů.

[96] SMITS, Thomas. Illustrations to Photographs: Using computer vision to analyse news pictures in Dutch newspapers, 1860-1940. *DH* [online]. Raboud University, The Netherlands, 2017 [cit. 2020-09-22]. Dostupné z: https://www.semanticscholar.org/paper/Illustrations-to-Photographs%3A-using-computer-vision-Smits/ad6ac997dcafe9b975d5758578fd3dd19fbd2ccb

[97] Například pro oblast sfragistiky mít univerzální katalog s 2D obrazovou podobou pečetí a jejich co možná nejpřesnějším popisem, podle kterého může program pečetě kategorizovat a klasifikovat. (vnější znaky, typ pečetě, aj.).

Computer Vision má ale určitě budoucnost i v dalších oblastech archivnictví a spisové služby. Možné využití lze vysledovat při automatizovaném rozpoznávání a klasifikaci razítek (např. úředních). Zatím ovšem chybí dostupné nástroje, protože vesměs se jedná o specializované nástroje, které nelze použít univerzálněji.

Největšími poskytovateli služeb Computer Vision jsou Microsoft, se svým Microsoft Azure (konkrétně část Cognitive services) a Google Vision API. Obě tyto korporace poskytují svá řešení skrze cloud a své vlastní, rozsáhlé neurální sítě. Nevýhodou je jejich upravitelnost – většina těchto řešení je určena spíše pro komerční oblast – popis obsahu, rozpoznávání násilného, sexuálního obsahu, rozpoznávání obličejů, názvů značek, aj.

Mimo toto Microsoft poskytuje také OCR řešení, které dosahuje vysoké přesnosti a kvalitních výstupů. Pro české prostředí je problémem, že pro tištěný text podporuje pouze angličtinu, španělštinu, němčinu, francouzštinu, italštinu, portugalštinu a holandštinu a pro rukopisný text pouze angličtinu.[98] Pro naše prostředí[99] má tedy určité využití pouze němčina, což je určitá nevýhoda, protože Microsoft Azure´s Computer Vision API HTR dosahuje lepších výsledků pro angličtinu než Transkribus.[100]

V porovnání s oblastí zpřístupňování (Transkribus aj.) a zpracovávání (NLP, Data Mining, aj.) textového materiálu, je oblast analýzy a zpracovávání obrazového materiálu moderními prostředky méně probádaná, a zatím nemá tolik dostupných univerzálnějších nástrojů, metodik, postupů a ukázek využití.

---

[98] Optical Character Recognition (OCR). *Microsoft* [online]. 2020 [cit. 2020-09-22]. Dostupné z: https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-recognizing-text

[99] Čeština je pouze částečně podporována, více v kapitole o porovnávání technologií.

[100] HIMANIBEN P, Patel. *Archival Document Processing using Cognitive Computing* [online]. Carolina, 2020 [cit. 2020-09-22]. Dostupné z: https://thescholarship.ecu.edu/handle/10342/7489. Master thesis. East Carolina University.

1.3.2. Natural Language Processing

Naproti Computer Vision stojí Natural Language Processing. Zatímco Computer Vision se zabývá zpracováním obrazových souborů (ať dynamických nebo statických), Natural Language Processing se zabývá zpracováváním textových informací ze souborů. *„Computer vision is to images as Natural-language processing (NLP) is to words.“*[101]

Co se týče současných vědeckých článků a prací, je třeba zmínit práci *Natural language processing and machine learning as practical toolsets for archival processing* z roku 2020 shrnující a popisující základní nástroje NLP a machine learning a jejich aplikaci do archivního prostředí.

Velmi širokého užití má potom NLP v oblasti státní správy a spisové služby, a to především v kombinaci s nasazením rozsáhlejších neurálních sítí pro tuto oblast.

Můžeme hovořit o alespoň částečné automatizaci procesu celého oběhu dokumentů od jejich automatického vytváření po rozpoznávání a klasifikace agend dokumentů a jejich obsahu, což by mohlo přinést řadu výhod – rychlejší proces zpracování dokumentu, jednodušší kooperace s archivy, jednodušší dodržování norem a zákonů (GDPR, ZKB), zvýšená bezpečnost[102], snížený rizikový lidský faktor, aj.

V současnosti dochází k tomu, že stále probíhá přechod z analogových fyzických dokumentů do úplného digitálního řešení. To je také zatím největší překážkou v nasazení těchto moderních IT společně s vysokou počáteční investicí. Pro nasazení těchto technologií je potřeba nejprve vyřešit tyto problémy.

---

[101] ROBERTSON, Mikaela. 6 technologies behind AI: Type #4: Computer vision. *CodeBots* [online]. 2018 [cit. 2020-09-21]. Dostupné z: https://codebots.com/artificial-intelligence/6-technologies-behind-ai

[102] Using AI to unleash the power of unstructured government data: Applications and examples of natural language processing (NLP) across government. *Deloitte Insights* [online]. USA, 2019 [cit. 2020-09-24]. Dostupné z: https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/natural-language-processing-examples-in-government-data.html

1.3.2.1. Nástroje

Většina implementací NLP pro archivy byla experimentálního rázu anebo velmi úzce zaměřena na konkrétní problematiku. Hlavní výjimkou je ePADD, který obsahuje robustní NLP funkce skrze svůj rozpoznávací modul.[103]

Projekt ePADD vznikl v roce 2010 a je ve vývoji dodnes. Přes deset let se snaží o vytvoření řešení umožňující aplikování Machine Learning a Natural Language Processing pro dárce archivních sbírek, archiváře, vědce, badatele.[104] Je dostupný zadarmo a nejedná se o specializované řešení pouze pro NLP, ale i pro standardní práci archiváře – posuzování archiválií, zpracování, ochrana, zpřístupňování a prezentace.[105] Prozatím se jedná o to nejdéle vyvíjené řešení a slouží ke zpracování e-mailových archiválií[106].

NLP funkce této platformy slouží především v (automatizované) identifikaci entit[107]. Další NLP funkcí je vyhledávání kritických slov[108] v emailech, které mají asociaci k osobním nebo bezpečnostním informacím, obecně takových informací, které mohou vyžadovat další manuální přezkoumání. Výsledná přesnost je prozatím nízká, kdy se vyhledaná klíčová slova často nachází mimo kontext.[109]

---

[103] HUTCHINSON, Tim. Natural language processing and machine learning as practical toolsets for archival processing. *Record Management Journal* [online]. 2020, 2(30), 155-174 [cit. 2020-09-24]. ISSN 0956-5698. Dostupné z: https://www.emerald.com/insight/content/doi/10.1108/RMJ-09-2019-0055/full/html

[104] EPADD: About. *Stanford LIBRARIES* [online]. Stanford, 2020 [cit. 2020-09-24]. Dostupné z: https://library.stanford.edu/projects/epadd/about

[105] EPADD, Stanford University. *Digital Preservation Coalition* [online]. 2020 [cit. 2020-09-24]. Dostupné z: https://www.dpconline.org/events/digital-preservation-awards/epadd-stanford-university

[106] Znovu tedy spíše specializovanější řešení, i když určitý potenciál pro jeho aplikaci pro více druhů archiválií existuje, minimálně jako proof of concept.

[107] Entita je v tomto případě např. jeden email, vlastnost této entity je např. odesílatel nebo příjemce.

[108] Obecně se tomuto říká keyword searching, jedna ze základních NLP metod.

[109] HUTCHINSON, Tim. Natural language processing and machine learning as practical toolsets for archival processing. *Record Management Journal* [online]. 2020, 2(30), 155-174 [cit. 2020-09-24]. ISSN 0956-5698. Dostupné z: https://www.emerald.com/insight/content/doi/10.1108/RMJ-09-2019-0055/full/html

Dalším poměrně slibným projektem byl BitCurator NLP, jehož nástroje jsou v současné době dokončené, ale mají vysoký technický strop pro většinu archivářů.[110] Projekt byl ve vývoji během let 2016 až 2019 a v současné době se spíše udržuje a vyvíjí jeho stejnojmenná aplikace. Software používá existující knihovny NLP (SpaCy, Textract)[111] pro paměťové instituce[112]. Tento nástroj do sebe kombinuje funkce digitálních forenzních analýz a NLP. BitCurator má sloužit jako komplexní nástroj schopný úplného zhodnocení a výběru archiválií společně s digitálními forenzními analýzami. Diplomová práce *"What is on this disk?" An Exploration of Natural Language Processing in Archival Appraisal* z roku 2019 se explicitně věnuje tomuto nástroji. Ve výsledku je tento nástroj sice funkční, ale širšího uplatnění se nedočkal, jedná se spíše o jakýsi prototyp komplexnějšího nástroje archivního zpracování, který se ale příliš spoléhal právě na NLP a potom se nedokázal rovnat funkcemi a přehledností standardním zavedeným nástrojům archivního zpracování. Toto řešení je volně dostupné, lze si jej stáhnout a vyzkoušet na jeho vlastním GitHubu[113].

Dalším nástrojem, který sloužil především jako ukázka konceptu využití NLP v archivnímu zpracování, je ArchExtract. Tento program je již neudržovaný a vývoj už neprobíhá.[114]

Cílem projektu ArchExtract, který probíhal 2014 až 2015, bylo aplikovat několik NLP nástrojů a metod na velké digitální sbírky textových archiválií a

---

[110] HUTCHINSON, Tim. Natural language processing and machine learning as practical toolsets for archival processing. *Record Management Journal* [online]. 2020, 2(30), 155-174 [cit. 2020-09-24]. ISSN 0956-5698. Dostupné z: https://www.emerald.com/insight/content/doi/10.1108/RMJ-09-2019-0055/full/html

[111] Tyto nástroje jsou: Textract, textacy, spaCy, scikit-learn, GraphLab. Jedná se obecně o nástroje NLP pro všeobecné využití, nespecifické pro archivní prostředí. Více o nich v další kapitole o nástrojích NLP.

[112] BitCurator NLP. *BitCurator* [online]. 2018 [cit. 2020-09-24]. Dostupné z: https://bitcurator.net/bitcurator-nlp/

[113] Odkaz: https://github.com/bitcurator/bitcurator-nlp/wiki

[114] HUTCHINSON, Tim. Natural language processing and machine learning as practical toolsets for archival processing. *Record Management Journal* [online]. 2020, 2(30), 155-174 [cit. 2020-09-24]. ISSN 0956-5698. Dostupné z: https://www.emerald.com/insight/content/doi/10.1108/RMJ-09-2019-0055/full/html

vytvořit webovou aplikaci, která by sloučila funkce NLP nástrojů postavených na příkazové řádce do grafického prostředí a tím umožnila (a zpřístupnila) tyto nástroje archivářům a badatelům[115]. Webová aplikace umožňuje extrahovat témata, entity (osoby, místa, data, aj.) a klíčová slova z archivních sbírek a fondů.

Při zkušebním provozu bylo zjištěno, že nástroje textové analýzy ArchExtract byli úspěšné ve většině případů při identifikování témat, jmen, datumů a míst popsaných v textu archiválie a tím umožnily poskytnout archivářům bližší porozumění obsahu sbírek a fondů.[116]

Posledním nástrojem je potom TOME, tento nástroj slouží k interaktivnímu prohledávání a vizualizaci obsahu textových archiválií jednotlivých souborů. Program nejprve zanalyzuje a kategorizuje textová data, která potom vizualizuje do interaktivního modelu.[117] Tento model lze připodobnit k relačnímu datovému modelu, kterého se využívá při tvorbě databází. Tento program[118] znovu využívá Keyword spotting.

Populární Archivematica prozatím nepodporuje NLP funkce. Pro masivní nasazení NLP funkcí pro archivní zpracování by mohlo být důležité, aby byly nástroje umožňující NLP dostupné v nějaké populárním a rozšířeném systému (jakým Archivematica je).

Prozatím tedy neexistují žádná generalizovaná řešení pro archivní zpracování. Vycházejí vědecké práce a články shrnující současné možnosti, popisující

---

[115] W. ELINGS, Mary. Using NLP to Support Dynamic Arrangement, Description, and Discovery of Born Digital Collections: The ArchExtract Experiment. *SAAERS* [online]. 2016 [cit. 2020-09-24]. Dostupné z: https://saaers.wordpress.com/2016/05/24/using-nlp-to-support-dynamic-arrangement-description-and-discovery-of-born-digital-collections-the-archextract-experiment/

[116] W. ELINGS, Mary. Using NLP to Support Dynamic Arrangement, Description, and Discovery of Born Digital Collections: The ArchExtract Experiment. *SAAERS* [online]. 2016 [cit. 2020-09-24]. Dostupné z: https://saaers.wordpress.com/2016/05/24/using-nlp-to-support-dynamic-arrangement-description-and-discovery-of-born-digital-collections-the-archextract-experiment/

[117] TOME: Interactive TOpic Model and MEtadata Visualization. *Digital Integrative Liberal Arts Center* [online]. Georgia, 2018 [cit. 2020-09-24]. Dostupné z: https://dilac.iac.gatech.edu/dilac-projects/topic-model-metadata-visualization

[118] Ukázkový model si lze prohlédnout na webových stránkách: http://tome.lmc.gatech.edu/.

jednotlivé nástroje a aplikovatelnost NLP pro archivy. Lze očekávat, že v blízké budoucnosti budou NLP nástroje integrovány do současných softwarových řešení pro archivní zpracování, společně s vznikem nových NLP nástrojů, které budou ve větší míře specializované s několika pokusy o generalizovanější a kompletnější řešení.

Je možné, že pokud se například řešení Transkribus rozšíří a standardizuje v archivech jednotlivých zemí EU, mohlo by dojít k masivnějšímu vývoji právě NLP aplikací pro archivní zpracování, ale jelikož je NLP závislé na nutnosti mít archiválie v textové, strojem čitelné, podobě, musí se nejprve vyřešit tento problém.

1.3.3. Dolování v datech

První techniky manuálního dolování v datech, které jsou velmi časově náročné, se datují do 13. století, kdy Hugo de Sancto Theodorico a několik dalších mnichů poprvé využili techniku konkordance[119] pro manuální indexování Bible.[120]

S příchodem informačních technologií se procesy dolování v datech začali pomalu automatizovat, a teprve v posledních 10 letech vznikají práce zabývající se získáváním znalostí z textu pomocí počítačových nástrojů a metod spadajících pod odvětví dolování v textových datech.[121]

Dobrým příkladem je práce z roku 2019 *How to analyse non-digital historical archives of large organizations — a text-mining case study*, která poukazuje na využití dolování v datech při analýze velkého množství archiválií. (V tomto případě se jedná o materiály z let 1947–2018). Tato práce popisuje celý pracovní postup od digitalizace, po čištění a filtrování dat, standardní textové analýzy až po využití metod a technik dolování v datech a funguje tak jako určitý návod, nebo alespoň

---

[119] Konkordance je jedna ze základních technik dolování v textových datech.

[120] REITER, Brian-Patrick. How to analyse non-digital historical archives of large organisations - text mining case study. *WST Working Paper Series* [online]. Economic and Social History and History of Technology, 2019, (4) [cit. 2020-09-26]. Dostupné z: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3460121

[121] REITER, Brian-Patrick. How to analyse non-digital historical archives of large organisations - text mining case study. *WST Working Paper Series* [online]. Economic and Social History and History of Technology, 2019, (4) [cit. 2020-09-26]. Dostupné z: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3460121

pomůcku, jak tyto metody aplikovat a použít. Autoři střídají několik nástrojů, popisují jejich nedostatky a kvality, což je důležité především proto, že neexistují žádné obecnější nástroje, a proto je nutné nástroje testovat metodami pokus omyl, což je velmi náročné především časově.

Dalším příkladem je práce *Tracking 19th century late blight from archival documents using text analytics and geoparsing* z roku 2017, která využívá textovou analýzu, metody NLP a geoparsing pro vyhledání zdrojů a výskytu nákazy brambor 1845, díky které Irsko zažívalo obrovský hladomor a migraci obyvatel[122].

Práce popisuje využití nástrojů pro dolování v datech jako je Google Ngram Viewer, PDFMiner a NLTK a také celý pracovní postup. Výsledkem této práce jsou interaktivní mapy s výskyty nákazy zjištěné podle rozpoznávání slov (a jejich počtu) souvisejících právě s touto nákazou brambor. Tohoto výsledku se docílilo právě technikami dolování v datech a obecně NLP. Tato práce je dobrou ukázkou především velkého zjednodušení celého procesu práce právě pomocí těchto moderních IT[123].

Zatím největšími překážkami aplikování metod dolování v datech pro archivní prostředí je nedostatečná standardizace nástrojů a metodologie, malé množství případových studií a dalších vědeckých prací a také náročnost na zdroje (finanční a lidské).[124]

Naopak největší klady využití těchto metod a technologií je možnost rychleji zpracovávat velké množství dat a také rozšíření výstupů standardní kvalitativní

---

[122] TATEOSIAN, L., R. GUENTER, Y. YANG aj. RISTAINO. Tracking 19th century late blight from archival documents using text analytics and geoparsing. *Conference: International Conference for Free and Open Source Software for Geospatial* [online]. 2017 [cit. 2020-09-26]. Dostupné z: https://www.researchgate.net/publication/322754348_TRACKING_19TH_CENTURY_LATE_BLIGHT_FROM_ARCHIVAL_DOCUMENTS_USING_TEXT_ANALYTICS_AND_GEOPARSING

[123] V práci autoři pracovali s dokumenty z celého světa, a to především z Irska, Velké Británie a Spojených států amerických.

[124] REITER, Brian-Patrick. How to analyse non-digital historical archives of large organisations - text mining case study. *WST Working Paper Series* [online]. Economic and Social History and History of Technology, 2019, (4) [cit. 2020-09-26]. Dostupné z: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3460121

textové analýzy. Dolování v datech pomáhá ve výběru a filtrování dokumentů pro výzkum tím, že badatelům poskytuje obecnější přehled o všech sesbíraných datech najednou a zároveň rozšiřuje standardní výstupy badatelské činnosti.[125]

Obecně tedy v současné době vznikají jednotlivé práce, a to především případové studie, na specifická témata používající nástroje, které popisují a porovnávají mezi sebou a vytvářejí také návrhy na pracovní postupy. Například první zmíněná práce obsahuje návrh na pracovní framework.

1.4. Shrnutí

Celkem lze shrnout, že nasazením zmíněných moderních informačních technologií a vytvoření uceleného rámce digitalizace může přinést zásadní benefity především ve třech oblastech.

Za prvé – zpřístupnění velkého množství dosud nedostupných archiválií aplikováním moderních OCR a HTR systémů, kdy se především platforma Transkribus jeví jako nejnadějnější se svojí poměrně velikou uživatelskou základnou, relativně vysokým množstvím dokončených výzkumných prací a řadou již digitalizovaných projektů většího i menšího rázu, které mohou posloužit jako určitý návod na použití.

Za druhé – společně s růstem DH, efektivnější zpracování archiválií ve větším měřítku než doposud pomocí metod a technologií přejatých z datové vědy, a to především dolování v datech, NLP, statistických analýz a moderní textové analýzy, kdy lze předpokládat, že toto zpracování by mělo přijít především ze strany archivářů a historiků seznámených s principy Digital Humanities a technologiemi běžně využívanými právě při DH.

Za třetí – částečná automatizace standardních pracovních procesů v archivu pomocí zavedení neurální sítě s kombinovaným využitím Natural Language Processing a Computer Vision pro oblast identifikace a rozpoznávání archiválií a obecně dokumentů. Prozatím ovšem neexistuje žádné jednotné řešení, a tak se toto

[125] REITER, Brian-Patrick. How to analyse non-digital historical archives of large organisations - text mining case study. *WST Working Paper Series* [online]. Economic and Social History and History of Technology, 2019, (4) [cit. 2020-09-26]. Dostupné z: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3460121

jeví jako prozatím nedostupné. Lze počítat s tím, že v blízké budoucnosti budou vznikat pokusy o zavedení právě tohoto procesu pro automatizaci části archivářovi práce. Reálné je aplikování těchto technologií především pro archivy s větším množstvím Digital-Born archiválií, u kterých odpadá nutnost převodu archiválií do strojově čitelné podoby. Vůbec před aplikováním těchto technologií do archivnictví lze uvažovat o jejím využití v oblasti státní správy a spisové služby, které se přímo dotýkají také čistě archivního prostředí.

Pro oblast automatizovaného přepisu historických dokumentů se jeví Transkribus jako současné nejlepší řešení. Naopak pro oblast využití moderních IT při práci s textovými dokumenty zatím žádná jednotná řešení neexistují, a tak výzkum stále probíhá především na bázi specifických projektů pro specifická řešení se specifickými nástroji. Především proto lze tvrdit, že se výraznějšího rozšíření této oblasti v archivním prostředí dočkáme později. Brání tomu především nepřipravenost archiválií na strojové zpracovávání, nedostatečné technické vybavení, finanční náročnost a náročnost na lidské zdroje (IT specialisty, techničtěji vzdělané archiváře).

Celkem přinesli moderní informační technologie mnoho problémů do oblasti archivnictví, které jejich vyřešení doposud věnuje velké úsilí, avšak z jejich výhod zatím nečerpá naplno.

2. Zpřístupňování počítačem čitelných archiválií a dokumentů

Zpřístupňováním se myslí automatizované rozpoznávání a převod rukopisného archivního materiálu, ať se jedná o archiválie, nebo také řadu běžných dokumentů v rámci jejich oběhu v instituci. Tato kapitola se zabývá především takovými softwarovými řešeními, technikami a metodami, které umožňují vytváření počítačem čitelné textové podoby vybraných archiválií a obecně dokumentů, kdy hlavními řešeními jsou především OCR a HTR technologie a jejich jednotlivé aplikace, ať již s použitím neurálních sítí, či nikoliv.

V současnosti existuje velké množství[126] možných aplikací umožňující standardní OCR převody, několik aplikací s možností využití HTR technologií a s využitím neurálních sítí. Proto je v této práci proveden výběr jen těch nejpoužívanějších.

2.1. OCRopus

OCRopus je balíček nástrojů, které slouží k extrahování textu ze skenovaného obrazu.[127] Tento balíček nástrojů je napsán v programovacím jazyce Python a nemá grafické uživatelské rozhraní. Ke komunikaci se využívá standardní CLI a spadá pod Apache Licence 2.0,[128] jedná se tedy o volně dostupný software, který lze volně modifikovat a distribuovat.

OCRopus má dvě varianty – jednu psanou v Pythonu a druhou v C++, kdy podle autorů obě verze dosahují podobné chybovosti, ale verze psaná v C++ je rychlejší než ta v Pythonu, což je patrné z charakteristiky Pythonu, jako interpretovaného jazyka[129]. V současnosti je tu tedy volba mezi oběma verzemi, časem je v plánu

---

[126] Jen na webových stránkách www.digitisation.eu projektu EU Impact je k dispozici 62 volně dostupných nástrojů pro rozpoznávání a převod textu. Navíc existuje velká řada dalších nástrojů, které jsou placené.

[127] Extracting text from an image using Ocropus. *Danvk.org* [online]. 2015 [cit. 2020-10-13]. Dostupné z: https://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html

[128] OCRopus – Python-based tools for document analysis and OCR. *LinuxLinks* [online]. [cit. 2020-10-13]. Dostupné z: https://www.linuxlinks.com/ocropus/

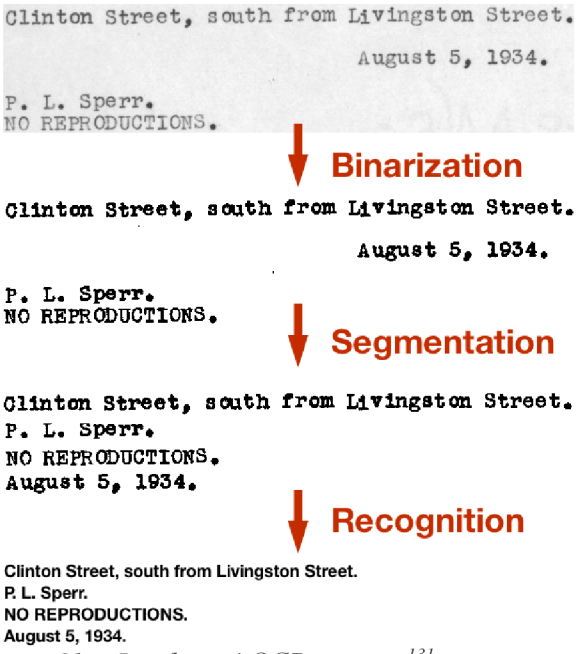
[129] Otázkou potom zůstává, zda je rozdíl v rychlosti zpracovávání obou verzí pro archivní potřeby natolik relevantní.

plně nahradit verzi v Pythonu verzí psanou v C++.[130] Neurální model sítě je také postupně nahrazován modernějším a výkonnějším typu LSTM, tento proces ještě není úplně hotový a jeho využití se plánuje společně s přechodem na verzi OCRopus aplikace psanou v C++ programovacím jazyce.

2.1.1. Základní pracovní proces OCR

Každý zdroj uvádí pracovní proces OCR trochu odlišně, ale v podstatě ho lze celý rozdělit do čtyř základních částí. Jedná se o binarizaci, segmentaci, rozpoznávání znaků a extrahování výsledného textu.

Většina OCR aplikací má velmi podobný proces fungování pouze s drobnými odchylkami. Celý proces lze obecně shrnout do následujícího obrázku pro vytvoření základní představy.



Obr. 5 – shrnutí OCR procesu[131]

[130] Ocropy: CLSTM vs OCRopy. *Github* [online]. 2020 [cit. 2020-10-13]. Dostupné z: https://github.com/ocropus/ocropy

[131] Extracting text from an image using Ocropus. *Danvk.org* [online]. 2015 [cit. 2020-10-13]. Dostupné z: https://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html

2.1.1.1. Binarizace

Prvním krokem je binarizace, při které dochází k převodu velmi informačně bohatého[132] obrazu s různými barvami a dalšími prvky do jednoduché binární podoby, kdy se v obrazu vyskytuje pouze černá a bílá[133], čímž se odstraní přebytečný informační šum. Toto umožňuje jednoduší čtení informací z obrazu, a tudíž snazší rozpoznávání znaků.

Binarizace může být někdy velmi složitá zvláště pokud se digitalizuje text, který má již vybledlý text a rozdíly mezi tmavými a světlými body jsou velmi malé. V tomto případě je někdy nutné přikročit k použití rastrového grafického editoru k úpravě dokumentu tak, aby bylo možné text řádně binarizovat.

Moderní OCR k binarizaci používají téměř výhradně adaptivního prahování (adaptive threshold), stejně jako OCRopus, který vypočítá průměrnou hodnotu odstínu šedé v celém obrazu a tu porovná s vybranými body, pokud je tento bod tmavší než průměr, bude ve výsledku černý, pokud je světlejší, bude ve výsledném binarizovaném obrazu bílý.[134] Toto je velmi důležité při práci s knihami a dalšími nerovnými digitalizovanými materiály, protože často dochází k tomu, že výsledný digitalizát má nerovnoměrné[135] nasvícení,[136] a to především v takových institucích, kde není dostupný profesionální knižní skener. Další, drobnou, úpravu, kterou

---

[132] Což je pro lidské oko samozřejmě výhodné, protože to nese další, někdy důležité informace, pro strojové zpracování ale pouze ztěžuje čtení a přidává nadbytečný informační šum.

[133] Image Binarization (1) : Introduction. *Craft of Coding* [online]. 2017 [cit. 2020-10-13]. Dostupné z: https://craftofcoding.wordpress.com/2017/02/13/image-binarization-1-introduction/

[134] EDDINS, Steve. Adaptive thresholding for binarization. *MathWorks* [online]. 2016 [cit. 2020-10-13]. Dostupné z: https://blogs.mathworks.com/steve/2016/07/25/adaptive-thresholding-for-binarization/

[135] Výstižný příklad porovnání adaptivního a jednoduššího globálního prahu lze nalézt na https://blogs.mathworks.com/steve/2016/07/25/adaptive-thresholding-for-binarization/.

[136] Extracting text from an image using Ocropus. *Danvk.org* [online]. 2015 [cit. 2020-10-13]. Dostupné z: https://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html

OCRopus dělá je srovnání obrazu a textu tak, aby byl plně horizontální pomocí odhadování statistické šikmosti.[137]

2.1.1.2. Segmentace

Dalším krokem je segmentace. Segmentace je rozdělování celého obrazu do jednotlivých částí, které lze později lépe zpracovávat.[138] Celý text se segmentuje na řádky, slova a jednotlivé znaky. V případě OCRopusu na celé řádky textu. Existují dva základní způsoby segmentace – segmentace pomocí detekce hran a segmentace hledáním oblasti, kdy druhá možnost je přesnější a více odolná proti šumu.

OCRopus používá segmentaci hledáním oblasti, kdy nejprve odhadne medián velikosti spojených částí obrazu, tedy znaků textu. Poté odstraní části textu, které jsou příliš velké anebo malé, a tudíž nemohou být jednotlivými znaky. Na celý tento výsledek potom aplikuje y-derivátové Gaussovské rozostření a podle takto rozostřeného obrazu dokáže rozpoznat jednotlivé řádky[139].

Příkladem je zobrazení na následujícím obrázku. Zvýrazněné části jsou řádky textu, světle šedé oblasti jsou prázdná místa. Oblasti s textem jsou výše než oblasti prázdné.

Obr. 6 – ukázka segmentace na řádky v OCRopus[140]

---

[137] Extracting text from an image using Ocropus. *Danvk.org* [online]. 2015 [cit. 2020-10-13]. Dostupné z: https://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html

[138] REDDY, Susmith. Segmentation in OCR: A basic explanation of different levels of Segmentation used by the OCR system. *Towards Data Science* [online]. 2019 [cit. 2020-10-13]. Dostupné z: https://towardsdatascience.com/segmentation-in-ocr-10de176cf373

[139] Extracting text from an image using Ocropus. *Danvk.org* [online]. 2015 [cit. 2020-10-13]. Dostupné z: https://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html

[140] Extracting text from an image using Ocropus. *Danvk.org* [online]. 2015 [cit. 2020-10-13]. Dostupné z: https://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html

2.1.1.3. Rozpoznání znaků

Když už jsou rozpoznány řádky textu, dojde k rozpoznávání samostatných znaků, zda se jedná o písmena, interpunkční znaménka či číslice. Software dokáže znaky klasifikovat podle jejich charakteristik popsaných v deskriptorech. Jednotlivé znaky jsou sice rozpoznány, ale téměř nikdy ne se 100% jistotou, a proto jsou jednotlivé tvary znaků porovnávány s lexikálními slovníky a na základě jejich porovnání je znak vyhodnocen na základě pravděpodobnosti s podobností znaku.[141]

Toto je standardní postup při OCR procesu bez použití modelu neurální sítě. Všechny moderní OCR umožňují ale tyto modely využít, v případě OCRopus se jedná o LSTM RNN model. Tento model umožní provést mapování dat s již naučenými vzory, což pro oblast komerčního využití a moderních tištěných dokumentů (de facto celé 20 století) není výrazným problémem a modely tak umožňují dosáhnout vysoké úspěšnosti i bez dostupnosti vlastních trénovacích data setů.

Problém nastává při jejich použití na historických dokumentech, kde je buď problém s kvalitou a nestálostí tisků anebo jednoduše nejsou dostupné dostatečné trénovací data sety pro úspěšné mapování a klasifikaci textu, a proto musí být tento trénovací data set vytvořen.

OCRopus potřebuje pro vytvoření efektivního tréninkového data setu obrazové soubory s již dokončenou transkripcí, protože neurální model se učí podle právě podle těchto data setů v porovnání s vlastní chybovostí. Svůj vlastní převod daného obrazového souboru porovná s již dostupnou transkribovanou podobou a dle ní své chyby opravuje. Toto poté několikrát opakuje, dokud se nedostane na přijatelnou úroveň přesnosti.[142] Zásadní je potom otázka velikosti tohoto transkribovaného vzorku pro naučení modelu.

---

[141] MEJZLÍK, Martin. OCR historických dokumentů [online]. Brno, 2016 [cit. 2020-10-13]. Dostupné z: https://is.muni.cz/th/hynsu/?fakulta=1433. Bakalářská. Masarykova univerzita Fakulta informatiky. Vedoucí práce RNDr. Michal Růžička, Ph.D.

[142] Training an Ocropus OCR model. *Danvk.org* [online]. 2015 [cit. 2020-10-13]. Dostupné z: https://www.danvk.org/2015/01/11/training-an-ocropus-ocr-model.html

Odpověď se v tomto případě velmi různí, protože každá archivní sbírka či fond vyžadují různě velké vzorky, a tak neexistuje univerzální hodnota, dle které by se dalo řídit.

Po výsledném převodu je tedy potřeba vždy kontrolovat správnost výsledku a zvýšit množství trénovacích dat. Velmi zásadní je také přesnost transkripce trénovacích data setů, protože pokud se v něm vyskytují chyby, neurální model se naučí tytéž chyby a potom je již velmi obtížné jej je odnaučit.

2.1.1.4. Extrahování textu

Výsledný obrazový digitalizáty s textovou podobou je extrahován celou řadou způsobů do různých formátů. V tomto se asi nejvíce liší rozdílné OCR aplikace. Konkrétně OCRopus umožňuje text extrahovat do prostého textu .txt nebo do specializovaného[143] souborového formátu hOCR postaveného na HTML. Výsledné soubory jsou v základu seřazeny podle abecedního pořadí.[144]

2.1.2. Kraken

Kraken je OCR aplikace odvozená od OCRopusu, která si klade za cíl opravit některé chyby OCRopus systému a současně udržet jeho dosavadní funkčnost.[145] Je optimalizovaný pro historické texty a také písma mimo latinský okruh a je psaný v Pythonu.[146]

Hlavním rozdílem oproti OCRopus řešení je odstranění všech ostatních nástrojů a funkcí kromě samotného OCR a neurálního modelu. Kraken přebírá stejnou licenci jako OCRopus a funkčně je na tom velmi podobně. Má ale rozšířené možnosti využití skriptů, variačních sítí neurálních architektur a také poskytuje veřejné úložiště pro jednotlivé naučené modely.

---

[143] Souborvý formát hOCR podporuje řada těch nejpopulárnějších OCR aplikací.

[144] Extracting text from an image using Ocropus. *Danvk.org* [online]. 2015 [cit. 2020-10-13]. Dostupné z: https://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html

[145] Kraken: Features. *Kraken* [online]. 2016 [cit. 2020-10-15]. Dostupné z: http://kraken.re/

[146] Kraken: Description. *Github* [online]. 2020 [cit. 2020-10-15]. Dostupné z: https://github.com/mittagessen/kraken

Z pohledu archivního hlediska je dobrá především širší podpora metadatových standardů a schémat – podporuje ALTO, abbyXML a z OCRopus také přebírá formát hOCR. Program tak není odkázán pouze na TXT a hOCR soubory na bázi HTML, ale nově i na XML metadatová schémata.

Hlavní nevýhodou je jeho rychlost, kdy nová verze OCRopusu je mnohem rychlejším řešením, protože je psaná v C++ a Kraken v Pythonu. Kraken má ale zcela jistě své uplatnění, lze zmínit práci *Building an efficient OCR system for historical documents with little training data* z projektu Porta Fontium, ve které je Kraken částečně implementován.

Experimentálně[147] dokáže také Kraken podporovat a rozpoznávat rukopisný text[148], pro který ale uživatel nutně potřebuje natrénovat vlastní neurální model, jelikož žádné předpřipravené modely neexistují, anebo nejsou veřejně přístupné.

Další, i když pro naše prostředí nepříliš zásadní výhodou, tohoto řešení je také podpora textů psaných či tištěných zprava doleva.

Otázkou potom zůstává samotný vývoj, kdy Kraken má mnohem menší tým než OCRopus, který nyní přechází plně na modernější LTSTM a je tak efektivnějším a rychlejším nástrojem, který ale zatím nedisponuje některými zmíněnými funkcemi systému Kraken.

2.1.3. Výhody a nevýhody

Mezi hlavní výhody tohoto balíčku nástrojů je jeho dostupnost – lze jej pořídit zadarmo na oficiálních webových stránkách, lze jej šířit, upravovat dle vlastních potřeb, což může pro archivní prostředí být určitým benefitem. Co se týče výstupů, tak podpora vlastního formátu založeného na HTML společně s běžným TXT je naprosto dostačující. Výhodou tohoto hOCR souborového formátu postaveném na HTML je také to, že HTML patří mezi jednodušší tagovací jazyky a jelikož má

[147] Pro toto je zatím výzkum na začátku a určitě by stálo zato, provést měření přesnosti v porovnání např. s řešením Transkribus nebo Quartex.

[148] ACHARY, S. Unleashing the Kraken for OCR. *Analytics Vidhya* [online]. 2020 [cit. 2020-10-15]. Dostupné z: https://medium.com/analytics-vidhya/unleashing-the-kraken-for-ocr-fba6bff73c8c

velmi široké použití při psaní webových stránek, tak se jedná o velmi rozšířený[149] jazyk a jeho podobnost s výstupním formátem hOCR je velmi vysoká, takže existuje určité překrytí obou.

Jelikož se nejedná o jednu aplikaci, ale spíše o balíček nástrojů, tak určitou výhodou je jeho modulárnost – lze použít jakýkoliv nástroj z dostupných, kdy základními komponentami jsou analýza rozvržení dokumentu, OCR a využití statistických neurálních modelů. Celý tento systém si tedy lze přizpůsobit vlastním potřebám. Výhodou současně využívaného neurálního modelu je jeho nezávislost na jazyku – podporuje všechny dostupné Evropské jazyky společně s latinou.[150]

Další výhodou balíčku OCRopus je jeho rychlost, a to především verze psané v C++, která je mnohonásobně rychlejší. Zde ovšem nastává problém v tom, že starší a rozšířenější verze psaná v Pythonu má být plně nahrazena výše zmíněnou, a tak dojde k postupné ztrátě podpory této verze a nutnost přesunu na jiný software nebo na C++ verzi.

Další jeho nevýhodou je nutnost zvládání využití příkazové řádky, která je uživatelsky nepřívětivá pro velkou část archivářů a obecně pracovníků v oblasti paměťových institucí.

Jako nevýhodu lze také chápat informační servis dostupný k tomuto balíčku nástrojů. Téměř všechny oficiální informace jsou v angličtině a němčině, neexistují pro něj žádné samostatné oficiální stránky – tým za OCRopusem využívá pouze github. I přes to, že se jedná o nástroje menšího rázu, autoři si dali práci s vytvořením alespoň základního informačního servisu. Na githubu jsou dostupné ke stažení starší verze, testovací nástroje, kopie licence a stručný návod na instalaci a použití. K dispozici jsou také samostatné webové stránky s popisem každé základní funkce a základního schématu celého fungování nástrojů.

---

[149] V dnešní době se výuka základů html praktikuje na téměř každé střední škole, v některých případech i na základních.

[150] SPRINGMANN, U., D. NAJOCK, H. MORGENROTH a SCHMID. OCR of historical printings of latin Texts: Problems, prospects, progress. *Document and Text Processing* [online]. 2014 [cit. 2020-10-15]. Dostupné z: http://springmann.net/papers/2014-DATeCH-Springmann.pdf

Hlavní cílem autorů je v blízké budoucnosti nahradit současný systém novou verzí CLSTM psanou v C++. Jelikož je tento systém nadále vyvíjen a udržován, lze uvažovat nad tím, zda není lepší počkat na vydání zmíněné náhrady, anebo využít jiný nástroj.

Celkem tedy tento nástroj dosahuje dobrých výsledků[151], umožňuje využití neurálních modelů pro lepší přesnost a možnost využití pro starší tisky, a to především zadarmo, má otevřený zdrojový kód a poskytuje slušný informační servis i když pouze v angličtině. Vývoj je v současné době stále silný a autoři aktivní.

Naopak hlavní nevýhodou je nutná znalost příkazové řádky k používání tohoto nástroje a také již zmíněný přechod z původní Python verze.

2.2. Tesseract

Tesseract je OCR engine s CLI, který podporuje moderní neurální sítě typu LSTM v současnosti ve verzi 4 a fungující na bázi rozpoznávání řádků. Autoři také podporují starší verzi 3.0 postavenou na Pythonu, která umožňuje rozpoznávání vzorů jednotlivých znaků.[152]

Panuje tedy obdobná situace jako v případě OCRopus, kdy zde se v současnosti udržují a vyvíjejí dvě verze – verze 4.x postavená na LSTM, která je rychlejší a má přesnější model a verze 3.x, která je sice pomalejší, ale za to rozšířenější a už mnoho let zaběhlá.

Podporuje všechny základní výstupové formáty jako je TXT, hOCR, PDF, TSV[153] a nově je přidána experimentální podpora ALTO (XML) výstupového formátu. V základu podporuje přes 100 jazyků včetně češtiny a UTF-8 kódování, kdy další jazyky lze Tesseract naučit podle dodaných trénovacích data setů. Tesseract je psán v C++ a má silnou uživatelskou základnu.

---

[151] Porovnání zmíněných nástrojů je na závěr kapitoly.

[152] Tesseract OCR: About. Github [online]. 2020 [cit. 2020-10-15]. Dostupné z: https://github.com/tesseract-ocr/tesseract

[153] Toto je vlastní formát Tesseractu.

V základu se k jeho ovládání používá CLI, avšak existují i grafická rozhraní[154] dostupná od třetí strany, takže se nejedná o oficiálně podporované GUI, ale jejich existence je důležitá pro potřeby archivů a knihoven.

Jedná se o nejpoužívanější a jeden z nejvýkonnějších OCR, který je vyvíjen společností Google, která jej používá především pro rozpoznávání textu na mobilních zařízeních, ve videích a ve svém emailovém klientu gmail k rozpoznávání textu v obrázcích jako antispamové ochraně[155] a obecně k získávání dat z obrazových souborů.

Má otevřený zdrojový kód a má podporovanou kompatibilitu s TensorFlow. Stejně jako OCRopus odpovídá licenčnímu ujednání Apache 2.0, takže jej lze svobodně šířit a upravovat.[156]

2.2.1. Výhody a nevýhody

Mezi základní výhody OCR Tesseract je jeho veliká uživatelská základna, která umožnila právě vývoj různých nástaveb na Tesseract, různých návodů a řešení pro běžně vyskytující se problémy. Jelikož se jedná o populární OCR nejen v archivech a knihovnách po celém světě, tak je vcelku dobře probádané a hojně používané, což také zapříčinilo vzniku různých vědeckých prací popisující využití tohoto OCR i s jeho variantou použití s neurálním modelem.

---

[154] Na oficiálních stránkách Tesseractu existuje výběr z více než 20 možných grafických uživatelských rozhraních. Zde je potřeba dávat pozor na typ licence, protože tato rozhraní jsou ve většině případů pod jinou licencí než samotný Tesseract a navíc je nelze často využít na jiných typech operačních systémů. Seznam lze nalézt na: https://tesseract-ocr.github.io/tessdoc/User-Projects-%E2%80%93-3rdParty.html.

[155] Tesseract OCR. *Google Open Source* [online]. 2020 [cit. 2020-10-15]. Dostupné z: https://opensource.google/projects/tesseract

[156] ZELIC, F. a A. SABLE. A comprehensive guide to OCR with Tesseract, OpenCV and Python. *Nanonets: Automate Data Capture* [online]. 2020 [cit. 2020-10-15]. Dostupné z: https://nanonets.com/blog/ocr-with-tesseract/#opensourceocrtools

Je také vyvíjen velkou a stabilní mezinárodní společností Google, což je důležité především z hlediska udržitelnosti této aplikace do budoucna, kdy Tesseract má širší použití právě v komerční oblasti sbírání dat.

Další bezespornou výhodou je i neoficiální podpora grafických uživatelských rozhraní, která výrazně zvedají dostupnost tohoto programu pro pracovníky paměťových institucí. Výhodou je také vestavěná podpora výstupního PDF formátu mimo standardní zmíněné formáty.

Oproti řešení OCRopus nemá takovou modulárnost – jedná se pouze o jeden nástroj. Tesseract také nedokáže rozpoznávat rukopisné texty oproti experimentálnímu Krakenu, který je přímo uzpůsobený historickým textům a je poměrně náchylný na kvalitní podklad[157] – nízké rozlišení skenu, chyby a vady textu či psací látky nebo jejich složitější vzory a okrasné prvky.[158] V základu také podporuje pouze ovládání pomocí příkazové řádky, což vyžaduje určité zaučení a může být překážkou.

Tesseract je zaběhlé, volně dostupné[159], velmi populární a robustní OCR s dobrou informační základnou, dostatečně dobrou přesností výsledků a díky třetí straně, dostupným grafickým rozhraním, které je pro pracovníky knihoven a archivů důležité. S dostatečně natrénovaným neurálním modelem je schopen zpracovat téměř jakýkoliv jazyk a typ písma. Je velmi hojně využíván ve vědeckých pracích a projektech současnosti a minulosti. Má stabilní zázemí a vývoj díky společnosti Google a je kompatibilní s moderním nástrojem pro práci s neurálními sítěmi založenými na TensorFlow.

---

[157] Na toto i autoři sami na oficiálních stránkách upozorňují a zmiňují, že dobrá kvalita skenu a podkladu, může zvýšit přesnost OCR.

[158] ZELIC, F. a A. SABLE. A comprehensive guide to OCR with Tesseract, OpenCV and Python. *Nanonets: Automate Data Capture* [online]. 2020 [cit. 2020-10-15]. Dostupné z: https://nanonets.com/blog/ocr-with-tesseract/#opensourceocrtools

[159] Znovu díky svým uživatelům existuje řada wrapperů, takže Tesseract lze spustit na téměř jakémkoliv systému a pracovat s ním s různými populárními nástroji.

2.3. ABBYY Finereader

Aplikace Finereader od společnosti ABBYY je robustní a komplexní software, který umožňuje práci s jak tištěnými dokumenty, tak různými typy elektronických textových dokumentů, kdy do sebe kloubí funkce zobrazování a úprav PDF společně s OCR.[160] S tímto softwarem lze tedy nahradit jakýkoliv standardní prohlížeč PDF souborů a OCR aplikaci jedním řešením.

Jedná se ale o komerční balíček s uzavřeným zdrojovým kódem. V současnosti nabízí společnost ABBYY 3 základní verze této aplikace. Standardní verzi, která je určena pro jedno zařízení a má trvalou licenci, v současnosti ve verzi 15, a vyšší, korporátní verzi, také s trvalou licencí. Základním rozdílem mezi nimi je dostupnost automatizace postupů digitalizace a limitu převodu až 5000 stránek za měsíc.[161]

ABBYY také nabízí zkušební verzi, kterou lze aktivovat na 7 dní a která umožňuje během nich převod až 100 stran dokumentů.[162] Tato verze je výhodná především při porovnávání jednotlivých OCR řešení.

Finereader obsahuje grafické uživatelské rozhraní pro řízení celého procesu OCR. Mimo to umožňuje také základní grafické úpravy obrazového dokumentu pro získání lepšího podkladu pro aplikování OCR. Většina těchto grafických úprav lze aplikovat automatizovaně anebo manuálně. Tento integrovaný editor obrázků umožňuje korekci zešikmení, vyrovnání textové linie, barevná korekce fotografie, odstranění šumu a rozmazání, převrácení barvy pozadí, oříznutí, otáčení obrazu, úprava kontrastu a jasu, úrovní[163] a další základní retušovací nástroje. Finereader je dostupný buď jako samostatná spustitelná aplikace, anebo skrze internet jako cloudová služba.

---

[160] ABBYY Finereader 14: Uživatelská příručka. *Natur.cuni.cz* [online]. 2017 [cit. 2020-10-17]. Dostupné z: https://www.natur.cuni.cz/fakulta/cit/navody/soubory/abbyy-fine-reader-uzivatelska-prirucka

[161] ABBYY FineReader PDF 15 for Windows: the smarter PDF solution. *ABBYY Finereader PDF* [online]. 2020 [cit. 2020-10-17]. Dostupné z: https://pdf.abbyy.com/

[162] ABBYY FineReader PDF 15 – free trial. *ABBYY Finereader PDF* [online]. 2020 [cit. 2020-10-17]. Dostupné z: https://pdf.abbyy.com/lp/finereader15-download-free-trial/

[163] ABBYY Finereader 14: Uživatelská příručka. *Natur.cuni.cz* [online]. 2017 [cit. 2020-10-17]. Dostupné z: https://www.natur.cuni.cz/fakulta/cit/navody/soubory/abbyy-fine-reader-uzivatelska-prirucka

Finereader podporuje výstupy v PDF, CSV, TXT, HTML, EPUB, DjVu, XLS a DOCX. Má tedy docela široký výběr standardních formátů textových souborů. Oproti Tesseractu a Krakenu ale nepodporuje ovšem ALTO. Co se týče jazykové[164] podpory, tak ta je velmi obdobná jako v případě Tesseractu, navíc rozšiřitelná. Zmíněné grafické uživatelské rozhraní stejně jako dostupná uživatelská příručka a základní informační servis poskytovaný výrobcem jsou jako jediné dostupné v češtině.

Finereader je rozšířeným softwarem s velkou uživatelskou základnou, to ale ovšem především v komerční oblasti zpracování elektronických dokumentů. Při práci na platformě Transkribus byl Finereader vybrán jako nejvhodnější OCR a do této aplikace byl integrován. Transkribus tedy disponuje tímto OCR ve svém základu.[165]

Pro Finereader také existují specializované slovníky pro historické texty, které by měli zvýšit přesnost výsledného OCR výstupu dostupné po registraci na webových stránkách[166].

Mimo tyto slovníky je v programu také integrován tzv. FrakturOCR, který obsahuje slovník pro frakturní písmo[167] z 16. století. Tento slovník podporuje němčinu a lotyšštinu. Na první pohled toto vypadá jako velmi specifické využití, ale frakturní písmo se používalo poměrně dlouhou dobu a bylo zakázané až v roce 1941[168] Josefem Goebbelsem.[169]

---

[164] Úplný výčet podporovaných jazyků lze nalézt na webových stránkách https://finereaderonline.com/en-us/Help/Recognition#fraktur

[165] MARTÍNEK, J., L. LENCL a P. KRÁL. Building an efficient OCR system for historical documents with little training data: Existing tools and OCR systems. *Neural Computing and Applications* [online]. 2020 [cit. 2020-10-17]. Dostupné z: doi:https://doi.org/10.1007/s00521-020-04910-x

[166] www.digitisation.eu

[167] Volně lze vyzkoušet online na https://finereaderonline.com/en-us/Tasks/Create

[168] Tohle ovšem neznamená, že se písmo nijak nevyvíjelo a že se využívalo výlučně. Určitě by stálo za to, v nějaké další práci otestovat přesnost tohoto dostupného slovníku s tím, jak se fraktura postupně vyvíjela.

[169] Fraktura. *Encyklopedieknihy.cz* [online]. 2020 [cit. 2020-10-17]. Dostupné z: https://www.encyklopedieknihy.cz/index.php/Fraktura

2.3.1. Výhody a nevýhody

ABBYY Finereader nabízí kompletní řešení OCR, které je z vybraných nástrojů nejpřístupnější ze všech hledisek kromě finančního a z hlediska licence – příručka a návod, stejně jako webové stránky a samotný software jsou v českém jazyce. Aplikace také má úplné grafické uživatelské rozhraní, což jako výhodu nelze podceňovat. Práce s takovýmto softwarem je mnohem jednodušší, a tak se jej může naučit téměř každý uživatel znalý základní práce s PC.

Důležitá je také poskytovaná uživatelská a technická podpora, což žádný z předchozích OCR nenabízí. Výhodou je také integrovaný základní grafický editor pro běžnou úpravu dokumentu společně s funkcí prohlížeče PDF.

Za výhodu lze považovat i existenci zmíněných slovníků z projektu EU a potom integrovaný slovník pro frakturní písmo, které by měli poskytovat přesnější výsledek OCR.

Zmínit lze i poměrně rozsáhlou podporu výstupních formátů, kdy mimo ty standardní podporuje také například EPUB, který se používá pro tvorbu e-knih, což může být relevantní především při edičním či knižním zpracování komplexního dokumentu u kterého je OCR pouze jeho součást.

Už ze své podstaty proprietárního komerčního softwaru program nemá otevřený zdrojový kód, a tak nelze uvažovat o úpravách a customizaci tohoto softwaru přesně pro podmínky daného projektu nebo digitalizačního procesu.

Druhou, poměrně jasnou nevýhodou je nutnost nákupu tohoto softwaru, zvláště potom, když každá zakoupená licence funguje pouze na jednom zařízení. Přináší tedy určité finanční zatížení pro instituci, kdy všechny ostatní předchozí OCR byly dostupné zadarmo a s otevřeným zdrojovým kódem. Oproti systému Kraken také nepodporuje rozpoznávání rukopisného textu.

Určitým problémem je také komerční řešení celého programu, kdy platí, že pokud chce uživatel použít neurální síť pro zpřesnění[170] výsledku, musí využít serverovou verzi tohoto programu a buď si pronajmout výpočetní výkon a část sítě

[170] A tím pádem vůbec získat dostatečně přesné výsledky, kdy bez použití neurální sítě takového výsledku nelze dosáhnout.

buď na určitou dobu anebo zaplatit za zpracování přesného počtu stran, což je samozřejmě poměrně vysoká finanční zátěž pro instituce, jakými jsou archivy a knihovny.[171]

2.4. Online OCR řešení

V současnosti poskytují některé mezinárodní korporace robustní řešení v oblasti získávání dat z obrazových souborů, a to se týká samozřejmě především textu. Využití těchto řešení pro zpracování a automatizovaný přepis historických dokumentů je teprve v počátcích, ale především pro možné budoucí použití je dobré se o nich zmínit. Tyto velké korporace mají obrovské zdroje jak informační, tak finanční a také personální.

2.4.1. Microsoft OCR Azure READ API

Součástí komputačních kognitivních služeb nabízených korporací Microsoft je OCR služba schopná převodu rukopisného i tištěného materiálu. Extrahuje text z obrazových souborů různých typů. Podporuje JPEG, PNG, BMP, PDF a TIFF formáty dokumentů.[172]

Pro tištěné dokumenty podporuje řadu světových jazyků včetně češtiny a němčiny, v seznamu podporovaných jazyků ale není latina. Pro rukopisné dokumenty podporuje pro naše prostředí důležitou němčinu, čeština dostupná není[173]. Za zmínku stojí určitě také podpora kombinace jazyků pro dokument, kdy lze použít vícero jazyků pro jednu stranu textu.

Microsoft nabízí 2 API, kdy prvním je standardní OCR API pro tištěné texty a modernější READ API, které umožňuje přesnější převod tištěných textů a převod

---

[171] MEJZLÍK, M. OCR historických dokumentů [online]. Brno, 2016 [cit. 2020-10-17]. Dostupné z: https://is.muni.cz/th/hynsu/?fakulta=1433. Bakalářská. *Masarykova univerzita Fakulta informatiky*. Vedoucí práce RNDr. Michal Růžička, Ph.D.

[172] Optical Character Recognition (OCR). *Microsoft* [online]. 2020 [cit. 2020-10-20]. Dostupné z: https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-recognizing-text

[173] Zde to bude pravděpodobně jen otázka času, protože počet podporovaných jazyků roste – zatím jsou dostupné jen ty největší evropské jazyky – angličtina, němčina, španělština, francouzština, holandština. Nově v září 2020 přidána japonština a předtím také čínština.

rukopisných textů. Údajně pak dosahuje toto řešení přesnějších výsledků v porovnání s platformou Transkribus, i když pouze pro texty psané v angličtině.[174]

2.4.2. Google Cloud Vision AI

Google má již nyní mnohaleté zkušenosti s OCR, vlastní a vyvíjí nejrozšířenější OCR engine Tesseract. Jejich cloudová služba Vision AI poskytuje také OCR řešení s využitím jejich rozsáhlé neurální sítě. Jako jediná ze zmíněných podporuje plně také český jazyk. Umožňuje rozpoznání jak rukopisného, tak tištěného textu. Obecně platí, že tištěný text rozpoznává mnohem lépe než ten psaný a čím novější text je tím přesnější je výsledek. Plusem je určitě také zkušební verze,[175] která je dostupná online.

Mimo extrakci textu dokáže systém také z obrazu získat informace o povaze dokumentu. Může například rozpoznat psací látku, typ písma, jestli dokument obsahuje kaligrafii, lékařský text[176] nebo obsah pro dospělé a další kategorie skrze keyword spotting.[177]. Toto je uzpůsobené především komerční sféře obsahu, ale pokud by existovalo nějaké samostatné řešení pro oblast archivnictví nebo knihovnictví, bylo by možné vytváření vlastních kategorií důležitých pro oba obory.

2.4.3. Amazon Textract AWS

Společnost Amazon nabízí svou online službu nazvanou Textract[178] založenou na strojovém učení. Tato služba umožňuje extrahovat textové informace z široké

---

[174] HIMANIBEN, Patel. *Archival Document Processing using Cognitive Computing* [online]. Carolina, USA, 2019 [cit. 2020-10-20]. Dostupné z: https://thescholarship.ecu.edu/handle/10342/7489?show=full. Diplomová. East Carolina University. Vedoucí práce Tabrizi M. H. N.

[175] Lze vyzkoušet na https://cloud.google.com/vision/docs/drag-and-drop

[176] Například zda obsahuje názvy léků, rostlin, bylin, titulů aj.

[177] Vision AI. Google Cloud [online]. 2020 [cit. 2020-10-20]. Dostupné z: https://cloud.google.com/vision

[178] Zde je potřeba zmínit také Textract, což je nástroj napsaný v Pythonu, který umožňuje extrahovat obsah z různých typů eDokumentů (PDF, PPTX, DOCX). Naproti tomu Textract od Amazonu je OCR řešení dostupné jako online služba.

řady dokumentů – formuláře, tabulky, grafy, textové dokumenty, obrazové dokumenty.[179]

Její použití spočívá především v automatizaci procesu zpracovávání dokumentů v instituci, vytváření chytrých indexů a efektivnějšího vyhledávání informací a pozdější zpracování skrze NLP a tím pádem také efektivnější BI.[180] Pro archivnictví má využití pro OCR převod tištěného materiálu, a to pouze v angličtině.

Pro české prostředí se jedná zatím o něco, co je nevyužitelné, protože nepodporuje český jazyk. Některé ze zmíněných řešení jsou ale jistou ukázkou jasného posunu k plně digitální spisové službě a obecně komunikace ve státní správě, kdy velkým krokem je už nyní schválená digitální ústava (zákon č.12/2020 Sb.), která má být základním prvkem moderního českého e-Governmentu a dnes už je její nasazení do reálné praxe jen otázkou času a k nasazení něčeho takového, jako je úplná digitální spisová služba by mohlo některé obdobné řešení jako jsou výše zmíněná být nasazeno[181].

2.4.4. Výhody a nevýhody

Tato online řešení už ze své podstaty mají hlavní výhody a nevýhody všech cloudových služeb. Všechny mají tu výhodu, že jsou velmi jednoduše dostupná a použitelná. Není třeba instalovat žádné nástroje a knihovny do počítače, stačí pouze internetový prohlížeč. Odpadá tedy nutnost mít vlastní ICT tým, nebo alespoň odpadá část zátěže. Jednoduchost používání a přívětivé grafické uživatelské rozhraní jsou také bonusem. Nároky na výpočetní výkon také klesají, protože již není potřeba trénovat vlastní modely, není potřeba provádět hlavní procesy OCR na svém zařízení. Výpočetní výkon je na straně provozovatele, což může být z některých hledisek nevýhodou. Zmínit lze také dobrý informační servis a vůbec

---

[179] Hard Limits in Amazon Textract. *Aws Amazon* [online]. 2020 [cit. 2020-10-20]. Dostupné z: https://docs.aws.amazon.com/textract/latest/dg/limits.html

[180] Amazon Textract. *Aws Amazon* [online]. 2020 [cit. 2020-10-20]. Dostupné z: https://aws.amazon.com/textract/

[181] Nemusí to ovšem být řešení od těchto korporačních gigantů, kteří s tímto přišly jako první. Tyto korporace ale fungují jako určitá předzvěst toho, jak by něco obdobného mohlo být nasazeno i u nás.

existenci zákaznické podpory, která v případě použití jakéhokoliv volně dostupného řešení vlastně neexistuje.

Naopak nevýhodou je potřeba dostatečně silného internetového připojení, software si také nelze přizpůsobit přesně podle svých požadavků. Uživatel nemá přístup ke zdrojovému kódu, hardwaru ani ostatním záležitostem. Musí tak v tomto spoléhat na poskytovatele služby. Existuje také určitá závislost na poskytovateli této služby a jejího udržování.

Určitě je třeba také zmínit problematiku ochrany osobních a dalších údajů, kdy všechny dokumenty, které chce uživatel převést do textové podoby musí nejprve nahrát poskytovateli služeb skrze webové prostředí. Toto je problém nejen z hlediska poskytování osobních údajů třetí straně, ale také bezpečnostní problém z důvodu informační bezpečnosti.

Všechny tyto služby jsou navíc finanční zátěží pro instituce, protože všechny mají platební model nastavený podle počtu převedených stran anebo využívají různých subskripčních modelů.

2.5. Srovnání nabízených řešení

Při porovnávání jednotlivých řešení je nutné použít několik základních kategorií, protože samotná přesnost při převodu z obrazové podoby do textové není dostačující, a to také proto, že rozdíly[182] v přesnosti výsledku nejsou tak vysoké.

Pro porovnání OCR řešení je k dispozici mnoho různých benchmarků, kdy je ale velmi důležitý původ dokumentů, které jsou pro porovnání použity. Většina těchto porovnání totiž neporovnává historické dokumenty, ale standardní dokumenty v oběhu většiny institucí s relativně současnými fonty a současnou jazykovou strukturou a písmem. Pro archivní oblast je tedy důležité využít historické dokumenty různých období a různých jazyků.

Kromě přesnosti je třeba také zohlednit dostupnost ve všech směrech. Prvním je finanční dostupnost – archivy a kulturně zaměřené instituce velmi často pracují

[182] Zde je potřeba upozornit na to, že zmíněná OCR řešení, kromě těch poskytovaných online, si lze přizpůsobit a dosáhnout dostatečně vysoké přesnosti díky využití vlastních slovníků a data setů pro současný projekt a díky tomu dosáhnout dostatečné přesnosti. Právě proto na té výsledné přesnosti tolik nezáleží.

s poměrně omezeným rozpočtem. Dále je potřeba zohlednit dostupnost a kvalitu informačního servisu, který je uživateli softwaru poskytován – dostupná dokumentace, doplňkové nástroje, uživatelská základna, zákaznická podpora.

Dostupnost je také potřeba řešit z pohledu uživatele – obsahuje software pouze textové rozhraní nebo i grafické? Instaluje se lokálně nebo je dostupné skrze cloud? I tyto otázky je třeba si při výběru a srovnávání OCR řešení položit.

2.5.1. Tesseract, OCRopus, Finereader

Pro porovnání těchto nástrojů a získání zpracovaných výsledků existuje řada dostupných zdrojů. Z většiny se jedná ale o takové testy, při kterých jsou použity současné standardní eDokumenty s využitím v současnosti používaných znakových sad písem a jazykovou stavbou. V těchto testech[183] jsou velmi často získávány velmi oslnivé výsledky již téměř dosahující hranice 100 % úspěšnosti. Tyto testy sice nejsou pro historické dokumenty nijak podstatné, ale pro potřeby alespoň spisové služby a digital-born archiválií jsou stále relevantní a užitečné, i když ve výsledku se jejich hodnoty úspěšných převodů tolik neliší, a to alespoň při použití latinky a standardních fontů v evropském prostředí.

Výsledná přesnost převodu je tedy ne až tak podstatným faktorem při výběru toho či onoho softwarového řešení, jak by se prvně mohlo zdát. Toto ovšem neplatí nejen pro oblast současných eDokumentů, kde je tento rozdíl ve výkonnosti nástrojů výrazně menší, ale také pro oblast převodu historických dokumentů, a to alespoň v určité míře. Samozřejmě přesnost je důležitá, ale je potřeba si stanovit určitou hranici minimální přesnosti, kterou chce daný uživatel dodržet a poté spíše zohlednit další důležité faktory při výběru.

Pro srovnání nástrojů je také důležité zmínit jeden zásadní problém, kterým je nesourodost podkladů. Už z principu mají historické dokumenty různě kvalitní digitalizáty, jsou psané či tištěné různě kvalitními tisky, různými písaři, používá se velká řada písem v různorodých oblastech. Toto je v přímém kontrastu se současnými dokumenty, které jsou stále více sourodé – používá se jedna nebo

[183] Lze zmínit například bakalářskou práci *Porovnání OCR technologií* z letošního jara 2020.

maximálně dvě abecedy (latinka a azbuka), používají se stále ty samé fonty a znakové sady, které už ze své podstaty počítačového zpracování vedou právě k sourodnosti.

Proto jsou výsledná testování sice vypovídající, ale nelze je aplikovat napříč[184] všemi historickými dokumenty. Pro převod obrazové podoby digitalizátu do textové formy je kritická individuální úprava a specifikace nástrojů, což lze vidět i v rozličných vědeckých článcích porovnávající právě tyto nástroje, kdy je aplikován individuální přístup ke každému projektu zvlášť. V zásadě se neustále opakuje využití nástrojů zmíněných v této práci, ale i v případě použití stejného nástroje pro jednu práci se defacto jedná o alespoň částečně odlišný nástroj v jiné práci díky jejich upravitelnosti, která je nutná pro zpracování právě historických dokumentů[185].

2.5.1.1. Přesnost převodu

Nicméně přesnost je stále relevantní údaj potřebný k porovnání nástrojů. Následující tabulka ukazuje výslednou přesnost pro pět stran úryvku z knihy Progymnasmata Latinitatis z roku 1589 psané v latině. Trénování pro nástroj Finereader je mnohem složitější, a tak je v tabulce uvedena původní přesnost bez použití trénovacích dat.

1589 - Pontanus - Progymnasmata Latinatitis							
Před aplikováním trénovacích dat				Po aplikování trénovacích dat			
strana	Finereader	Tesseract	OCRopus	strana	Finereader	Tesseract	OCRopus
15	87,79%	80,88%	80,70%	15	87,79%	93,90%	83,66%
16	82,94%	77,41%	76,94%	16	82,94%	85,65%	78,00%
17	85,25%	75,98%	86,07%	17	85,25%	91,56%	86,80%
18	85,93%	79,51%	85,53%	18	85,93%	92,68%	89,59%
19	87,94%	80,09%	79,09%	19	87,94%	90,15%	80,97%

Obr. 7 – porovnání úspěšnosti pro latinský text z r. 1589 a rozdíly trénovacích dat[186]

[184] Pokud provádíme testování pro latinu psané humanistickým písmem, nedosáhneme stejných výsledků při použití stejných nástrojů pro novogotické písmo psané němčinou.

[185] Jedním z mnoha příkladů je psaní velkého s, které se velmi podobá modernímu f nebo ligatura ae.

[186] Autorova tabulka vytvořená na základě výsledků práce: Springmann, Uwe & Najock, Dietmar & Morgenroth, Hermann & Schmid, Helmut & Gotscharek, Annette & Fink,

Možnost aplikování vlastních trénovacích dat je zásadní, protože v průměru dochází k 10% zpřesnění výsledků, někdy i více. Výsledné přesnosti mají mezi sebou jednotlivé nástroje rozsah maximálně 10 %, což je v některých případech poměrně markantní rozdíl. Nástroj Finereader dosahuje průměrně 85,97 %, nástroj Tesseract 78,77 % před aplikováním trénovacích dat a 90,79 % po jejich aplikování. OCRopus dosahuje před trénovacími daty 81,67 % a po aplikování 83,80 %.

V případě použití těchto OCR nástrojů pro latinský text je tedy nejvýkonnějším nástrojem Tesseract následovaný nástrojem Finereader a nakonec OCRopus. U Tesseractu sledujeme velký skok před a po aplikování trénovacích dat, u nástroje OCRopus je rozdíl mizivý. Naopak nástroj Finereader je nejvýkonnějším nástrojem před aplikováním trénovacích dat a otázkou potom zůstává, zda se potom vyplatí proces aplikování anebo vytváření trénovacích dat pro nástroj Tesseract, když rozdíl v natrénovaném[187] Tesseractu a nenatrénovaném Finereaderu je pouze 5 %.

Samozřejmě jeden latinský text nelze považovat za reprezentativní pro všechny nástroje z výše zmíněných důvodů. Následující tabulka obsahuje průměrné přesnosti několika velkých archivních sbírek z různých časových období. Výsledná data přesností jsou dostupná po aplikování trénovacích dat.

Prvním jsou záznamy ze Stormontu z Irského parlamentu mezi léty 1921–1976, reprezentují tedy modernější texty, se kterými se v archivech napříč Evropou ale stále běžně lze setkat. Druhý sloupec obsahuje přesnosti nástrojů pro skeny ze Zákonů o unii[188] (Acts of Union) z roku 1707. Poslední sloupec reprezentuje přesnosti nástrojů pro skeny ze Slovníku staré skotštiny[189], který obsahuje ukázky napříč staletími od 12. do počátků 18.[190]

Florian. (2014). OCR of historical printings of latin Texts: Problems, prospects, progress. *ACM International Conference Proceeding Series*. 10.1145/2595188.2595205.

[187] Odpovědí je pravděpodobně dostupnost těchto trénovacích dat. Pokud má uživatel lehce dostupná transkribovaná data, která může pro program použít, pak se jejich aplikování vyplatí. Pokud by naopak musel transkripci provádět sám, nemusí tomu tak být.

[188] Acts of Union 1707

[189] Dictionary of the Older Scottish Tongue

[190] BLANKE, T., M. BRYANT a M. HEDGES. Ocropodium: Open source OCR for small-scale historical archives. *Journal of Information Science* [online]. 2012, (38), 76-86

	Stormontské záznamy	Zákony o Unii	Slovník staré skotštiny
Finereader	96,00%	79,50%	87,00%
OCRopus	82,50%	20,40%	82,60%
Tesseract	93,00%	64,00%	79,30%

Obr. 8 – porovnání průměrné přesnosti nástrojů napříč staletími[191]

Finereader dosahuje průměrné přesnosti 87,5 %, OCRopus 61,83 % a Tesseract 78,77 %. Zde je patrné především to, že OCRopus se na práci se zmíněnými texty jako celku nehodí, protože pro Zákony o Unii z roku 1707 je nepoužitelný. Znovu se jedná o příklad oné individuality přístupu ke každému zvolenému projektu.

Jako obvykle je Finereader nejúspěšnějším nástrojem následovaný Tesseractem. Tato data jsou ovšem méně individuálně přesná než předcházející tabulka, ale zato poskytují širší rozhled napříč různými typy textových dokumentů, které lze digitalizovat napříč mnohými staletími.

Bez použití trénovacích dat je Finereader nejpřesnějším nástrojem napříč všemi měřeními, následuje jej Tesseract a poté OCRopus. V případě využití trénovacích dat jsou potom nástroje mnohem vyrovnanější a Tesseract se také často dostává v přesnosti měření před Finereader. Vždy ale záleží na individuálních dokumentech a výsledky se mohou poměrně razantně lišit u jiných dokumentů.

2.5.1.2. Další faktory

Kromě přesnosti je nutné zmínit i další parametry, kterými se tyto tři nástroje liší. Jak bylo zmíněno výše, mohou být právě těmi rozhodujícími faktory při výběru nástroje, pokud je výsledná přesnost pro uživatele dostačující a rozdíly v přesnosti mezi jednotlivými nástroji nejsou tak markantní.

Finereader se může zdát jako nejvhodnější řešení, jelikož má grafické uživatelské rozhraní, dobrou informační základnu, dosahuje poměrně přesných

---

[cit. 2020-11-06]. Dostupné z: https://www.researchgate.net/publication/254115552_Ocropodium_Open_source_OCR_for_small-scale_historical_archives

[191] Vlastní autorova tabulka postavená na základě dat z: BLANKE, T., M. BRYANT a M. HEDGES. Ocropodium: Open source OCR for small-scale historical archives. *Journal of Information Science* [online]. 2012, (38), 76-86 [cit. 2020-11-06]. Dostupné z: https://www.researchgate.net/publication/254115552_Ocropodium_Open_source_OCR_for_small-scale_historical_archives

výsledků napříč různými typy dokumentů a je poměrně široce podporován a také integrován do aplikace Transkribus. Jeho nevýhody jsou ale patrné v tom, že se jedná o komerční řešení s proprietární licencí, a tudíž je neohebný – nelze provádět úpravy ve zdrojovém kódu, uživatel je vázán licenčním ujednáním. Stojí také finanční obnos, zatímco ostatní dvě řešení jsou zadarmo.

Tesseract se jeví jako nejlepší volně dostupné řešení, neoficiálně podporuje uživatelská rozhraní, dosahuje uspokojivých výsledků, a to především při využití trénovacích dat, má otevřený zdrojový kód a je volně šiřitelný. Naopak v základu podporuje pouze CLI a bez trénovacích dat nedosahuje takových výsledků jako Finereader a navíc nemá zákaznickou podporu. OCRopus se jeví jako nejslabší řešení, ale poskytuje navíc některé funkce, které ostatní dva nástroje neposkytují. Mimo to má také nejmenší uživatelskou základnu, ale zato má vlastní nástavbu ve formě nástroje Kraken.

Nakonec je výběr vždy podmíněn konkrétnímu projektu digitalizace, kdy je dobré si každý nástroj nejprve předem vyzkoušet (např. Finereader má dostupnou zkušební verzi) a zvážit všechny výhody a nevýhody[192] těchto řešení.

2.5.2. Online OCR

Porovnány jsou 3 služby poskytované společnostmi Microsoft, Google a Amazon. Všechny tyto služby fungují v podstatě stejně a již ze své podstaty cloudové služby mají stejné výhody a nevýhody[193]. Základními rozdíly mezi nimi jsou přesnost výsledku, čas, dostupnost a zákaznická podpora.

Poměrně zajímavá je cenová situace – každá společnost si účtuje úplně stejně vysokou částku 1,5 dolarů (asi 35 Kč[194]) za 1000 stran textu. Rozdíl je potom v různých slevách a zvýhodněních v případě použití různých typů balíčků, subskripcí, množství převedeného textu a další. Ve výsledku záleží tedy vždy na konkrétních projektech, pro které chce uživatel těchto služeb využít.

---

[192] Které jsou popsány v jejich příslušných podkapitolách výše.

[193] Viz jejich samostatné podkapitoly výše.

[194] Kurz z 29.10.2020

Přesnost výsledku se pro české prostředí měří velmi těžce z již zmíněných důvodů a také proto, že služby se stále vyvíjejí a zlepšují. Pro úplnost jsou rozdíly mezi jednotlivými řešeními popsány v následující tabulce[195]. Důležitá je nejen průměrná chybovost, ale i maxima, minima a výkyvy.

V levém sloupci jsou jednotlivý poskytovatelé služeb – Amazon, Microsoft a Google. V druhém sloupci je minimální dosažená chybovost, následují 3 tercily[196] rozdělující řadu na 3 stejně veliké díly.

CER

Provider	Min	Q1	Q2	Q3
AWS	2.3%	9.5%	17.3%	73.2%
Azure	2.3%	15.2%	23.6%	38.8%
GCP	2.1%	11.6%	19.7%	74.1%

Obr. 9 – porovnání CER pro online OCR řešení[197]

Z tabulky lze rozpoznat, že v průměru se přesnosti mezi jednotlivými službami tolik neliší a průměrný rozsah je do 6 %. Jak řešení Textract od Amazonu, tak řešení Cloud API Od Googlu dosahují velmi podobných výsledků, kdy Textract je lehce přesnějším řešením, obě tak poráží řešení od Microsoftu v tomto ohledu. Tabulka je postavena na základě skenů a následného OCR soudních záznamů tzv. Old Baileyho kolekce z let 1674–1913.

Při bližším porovnání je ale zřejmé, že obě řešení mají mnohem vyšší kolísání (třetí kvantil), kdy řešení od Microsoftu je sice méně přesné ve výsledku, ale zato si udržuje poměrně lineární růst chybovosti a zdaleka nemá takové výkyvy jako ostatní dvě řešení.

---

[195] Tato tabulka také slouží jako dobrý příklad pro přesné porovnání, protože používá místo chybovosti uvedené v aritmetickém průměru nebo mediánu právě kvantily, které poskytují přesnější výsledek tam, kde by medián nebo aritmetický průměr byly nedostatečně vypovídající.

[196] Rozdělení chybovosti v řadě na kvantily je výhodné především proto, že lze vypozorovat určité kolísání v přesnosti, a tak jsou například oproti průměru nebo mediánu mnohem více vypovídající.

[197] The Old Bailey and OCR: Benchmarking AWS, Azure, and GCP. *ACM Symposium on Document Engineering* [online]. New York, NY, USA: Association for Computing Machinery, 2020, 4 [cit. 2020-10-29]. Dostupné z: https://doi.org/10.1145/3395027.3419595

Přesnost je ale pro naše prostředí zatím méně relevantní, protože při současné situaci lze u nás efektivněji využít pouze zmíněných služeb od Googlu a do určité míry také služeb od Microsoftu. Amazon je pro nás méně vhodný[198].

Při rozhodování výběru služby jsou v tomto případě zásadní spíše detaily než nějaké očividné a veliké rozdíly mezi nimi. Informační servis a webové stránky mají všechny tři společnosti velmi podobný. Jako jediné řešení je ale to od Microsoftu dostupné i v českém jazyce. Amazon je se svým Textractem zatím mimo české prostředí, a tak porovnání lze tedy provést jen pro služby Googlu a Microsoftu, kdy Google poskytuje některé funkce navíc[199], které Microsoft zatím nenabízí, a tak je zatím současným nejvhodnějším řešením pro česky psané texty.

Ve výsledku ale žádné řešení není dokonalé. Stále je potřeba vyřešit některé problémy. Medián CER napříč těmito třemi systémy je 17,33 %[200], což je stále poměrně vysoké číslo. Než se ale jakákoliv z těchto řešení dostanou do našeho prostředí a budou běžnější, situace může být během horizontu několika dalších let už odlišná.

Tady lze vysledovat určitou paralelu při využití řešení jako je Tesseract nebo OCRopus, které dosahují obdobných přesností, ale alespoň v případě těchto řešení lze výslednou chybovost výrazně snížit pomocí vlastních úprav v kódu programu anebo ve vytvoření vlastních data setů a využití strojového učení.

---

[198] Viz samostatné podkapitoly o těchto službách.

[199] Znovu se lze odkázat na předcházející podkapitolu o této službě.

[200] The Old Bailey and OCR: Benchmarking AWS, Azure, and GCP. *ACM Symposium on Document Engineering* [online]. New York, NY, USA: Association for Computing Machinery, 2020, , 4 [cit. 2020-10-29]. Dostupné z: https://doi.org/10.1145/3395027.3419595

3. Transkribus

Současný proces zpřístupňování archiválií a jejich digitalizátů standardně funguje skrze různé webové aplikace. V České republice existuje řada projektů zabývajících se digitalizací a zpřístupňováním archiválií, kdy hlavními problémy je jejich roztříštěnost – pro každý projekt existují jiné webové stránky a neexistuje žádný jednotný portál pro jejich prezentaci. Jejich uživatel tedy musí mít nějaké základní know-how a musí mít alespoň základní přehled o jejich existenci.

Většina těchto projektů zpřístupní své digitalizáty jako obrazová data, což je samo o sobě problematické a to především, pokud těchto digitalizátů je větší množství. Účelem těchto projektů je nakonec totožný – poskytnout uživatelům možnost získat z těchto obrazových digitalizátů informace. Tento cíl je ve většině případů naplněn – výsledné webové aplikace s přístupnými obrazovými digitalizáty a základními možnostmi jejich vyhledávání jsou zpřístupněné. Jak jsou ale tyto informace přístupné?

Průměrný uživatel nemá základní znalosti PVH, a to především paleografie a diplomatiky. Místo toho musí tedy spoléhat na badatelskou rešerši poskytovanou archiváři.

Současné trendy v Digital Humanities a informatice obecně kladou velký důraz na automatizaci procesů, a tudíž využití strojového zpracování[201], které je mnohem náročnější, pokud jsou nutná data pouze v obrazové podobě a s různě hlubokým metadatovým popisem. Transkribus je tedy velmi ambiciózním[202] projektem, který si klade za cíl archivní dědictví takto zpřístupnit.

V současnosti má Transkribus několik funkčních celků tvořících kompletní aplikaci. Skládá se ze serveru a klienta, kdy klient je dostupný volně ke stažení po registraci na webových stránkách ve formě portable Java aplikace pro Windows a

[201] Jedná se o zpracování jak pomocí statistických analýz, tak v současnosti velmi populárních Big Data analýz, ve kterých lze analyzovat velké množství dat najednou.

[202] Asi nejlepší ukázkou možností, které Transkribus poskytuje, je volně přístupná online demonstrace několika tisíc transkribovaných stran na webových stránkách http://transcriptorium.eu/demots/htr/index.php.

MacOS X. HTR model na straně serveru je postaven na populárním, volně dostupném, frameworku PyTorch pro práci s neurálními sítěmi.[203]

Prvním tímto celkem je analýza rozložení dokumentu (DLA), která umožňuje rozpoznání rozložení textu v dokumentu – odlišuje tedy text od pozadí a dokáže jej také segmentovat na řádky. Toto je umožněno vlastní U-net architekturou hlubokého učení[204]. Toto DLA P2Pala je navíc volně přístupné skrze github[205], což je veliká výhoda oproti ostatním řešením. K dispozici je také volně dostupná webová aplikace[206] kde si lze nahrát vlastní dokument a nechat si jej segmentovat na řádky.

Druhou částí aplikace je potom automatizovaná transkripce postavená na vlastní HTR technologii PyLaia, která je také volně přístupné na githubu[207].

Další částí je počítačem asistovaná transkripce, která umožňuje interaktivní hodnocení a opravu dokumentů (CATTI). Skládá se z webové aplikace s GUI a vlastního enginu.[208]

Zde potom končí oblast automatizovaného transkribování a začíná oblast zpracování počítačem čitelných dokumentů, jejichž součástí je vyhledávání klíčových slov, indexování a klasifikace dokumentu, pokročilejší dotazování, hlubší sémantický popis a automatizovaná sumarizace.

Vyhledávání klíčových slov je možné už po prvním projetí dokumentů HTR modelem. Není tedy potřeba provádět ruční transkripci před tím než lze

---

[203] How To Search Documents with the Keyword Spotting Feature. *READ COOP* [online]. 2020 [cit. 2020-11-10]. Dostupné z: https://readcoop.eu/transkribus/howto/how-to-use-keyword-spotting/

[204] Commercial Overview TranSkriptorium. *Transkriptorium* [online]. 2020 [cit. 2020-11-10]. Dostupné z: http://www.transkriptorium.com/user/images/transkriptorium/ts-presentation.pdf

[205] Odkaz na github P2Pala: https://github.com/lquirosd/P2PaLA

[206] Odkaz na demonstraci DLA: http://prhlt-carabela.prhlt.upv.es/tld/

[207] Odkaz na github PyLaia: https://github.com/jpuigcerver/PyLaia

[208] Commercial Overview TranSkriptorium. *Transkriptorium* [online]. 2020 [cit. 2020-11-10]. Dostupné z: http://www.transkriptorium.com/user/images/transkriptorium/ts-presentation.pdf

v dokumentu vyhledávat. Transkribus by totiž měl být schopný i chybná slova rozpoznat správně.[209]

3.1. Workflow

Standardní proces pracovního postupu na platformě Transkribus začíná registrací na webových stránkách a poté stáhnutí a instalací klienta. Ten je dostupný pro všechny běžné operační systémy, navíc se jedná o portable[210] aplikaci, tedy bez nutnosti instalace. K fungování je potřeba mít na pracovní stanici nainstalovanou Javu nejméně ve verzi 8.[211]

Aplikace má grafické uživatelské rozhraní a je dostupná pouze v angličtině. Z počátku se může zdát poněkud složitá, ale pro její pochopení existuje řada připravených návodů včetně video návodů[212] na YouTube. Celkově je informační servis, který je pro aplikaci poskytován velmi kvalitní a dokáže uživatele provést všemi kroky od instalace po trénování modelu až k metadatovému popisu.

Po spuštění aplikace je nutné se nejprve přihlásit. Po přihlášení se uživateli zpřístupní dostupné již natrénované HTR modely většinou anglické, německé, holandské a francouzské provenience, pokud je chce využít pro svůj projekt transkripce. K těmto modelům jsou také zpřístupněné čtyři již vypracované transkripce sloužící jako ukázkové práce.

V uživatelské části jsou také dostupné údaje o vlastních kolekcích, jak rozpracovaných, tak hotových. Lze si také prohlédnout záznamy zaznamenané aktivity přihlášeného uživatele a základní informace o svých kolekcích a o zrovna otevřeném dokumentu.

---

[209] How To Search Documents with the Keyword Spotting Feature. READ-COOP [online]. 2020 [cit. 2020-11-17]. Dostupné z: https://readcoop.eu/transkribus/howto/how-to-use-keyword-spotting/

[210] Ona přenositelnost může být dobrá například pro potřeby výuky a představení programu, vyučující si jí jednoduše může přinést na flash disku a jednoduše používat, pokud to samozřejmě bezpečnostní opatření a práva uživatele umožní.

[211] Download and Installation. *READ-COOP* [online]. 2020 [cit. 2020-11-17]. Dostupné z: https://readcoop.eu/transkribus/wiki/download-and-installation/

[212] Základní návody na ovládání aplikace: https://readcoop.eu/transkribus/knowledge-base/how-to-guides/

Je důležité zmínit, že veškeré úkony náročné na výpočetní výkon jsou prováděny na straně serveru, takže pro klientskou aplikaci Transkribus není potřeba nadstandardně výkonná pracovní stanice. Důležité je především stabilní internetové připojení s dostatečnou rychlostí uploadu, protože velké kolekce kvalitních digitalizátů mohou být objemově náročné pro standardní asynchronní spojení.

Celý proces funguje tak, že tyto operace uživatel zadává serveru, který je buď splní (pokud má uživatel dostatečný kredit) anebo nesplní. Všechny probíhající a již dokončené operace si lze prohlédnout v záložce Jobs.

Type	State	Doc...	Username	Description	Errors	Created	Started	Finished	ID
PyLaia Decoding	CANCELED	253...	o.tomiska...		0	17.11.2020 16:...			1484783
Optical Character Recognition	FINISHED	253...	o.tomiska...	Done, dura...	0	17.11.2020 16:...	17.11.2020 16:...	17.11.2020 16:...	1484769
Layout analysis (CITlabAdvanced...	FINISHED	253...	o.tomiska...	Done, dura...	0	15.11.2019 19:...	15.11.2019 19:...	15.11.2019 19:...	783700

Obr. 10 – snímek záložky Jobs se seznamem prováděných operací[213]

3.1.1. Předzpracování dat

Pro transkripci je nutné nejprve dokumenty, které chce uživatel transkribovat nahrát do systému. Tyto nahrané dokumenty jsou v této fázi neveřejné.

Po nahrání je nutné provést základní úpravy ve všech dokumentech určených pro transkripci, protože se může stát, že proces automatizované segmentace nebyl 100% úspěšný[214]. V záložce Layout Analysis lze automatizovanou segmentaci dokumentu vynechat buď úplně, anebo pro určité stránky, které mohou být problematické, například obálka knihy, velmi silně poškozené strany, aj. Pro přesnější automatizovanou segmentaci lze využít také nově přidané funkce P2PaLA, která by měla umožnit lepší segmentaci z již sesbíraných trénovacích modelů.

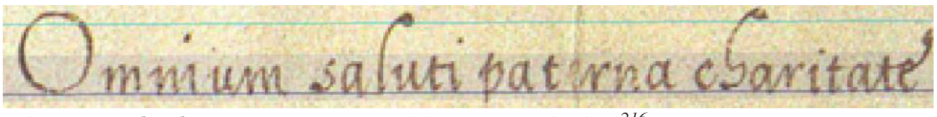
Pokud existuje již alespoň částečná transkripce k vybraným dokumentům pro transkripci, například již vydané ediční práce, lze využít funkce Text2Image

[213] Autorem vytvořený snímek seznamu operací v záložce Jobs.

[214] Program většinou dokáže alespoň základní rozpoznání a automatizovanou segmentaci úspěšně. Výsledek ale ovšem velmi závisí na fyzickém stavu digitalizovaného dokumentu a na kvalitě výsledného digitalizátu.

v záložce Tools, která umožňuje přesnější automatizovanou segmentaci na linie pomocí právě již existující transkripce.[215]

Korekci segmentace lze provést v oblasti se zobrazením dokumentu. Uživatel může ručně upravit řádky tak, aby vizuální označení odpovídalo reálnému textu – v programu jsou graficky vyznačené řádky ohraničující spodní část i vrchní část liter. Nutná je také úprava segmentování na jednotlivé textové bloky, které ohraničují obsah textu, tedy délku krát výšku.



Obr. 11 – ukázka segmentace v aplikaci Transkribus[216]

V předcházejícím snímku lze vidět tři základní prvky segmentace – spodní linii pro označení spodní hranice textu (tmavě modrá), vrchní linii označující horní hranici textu (zelená) a textové bloky vyznačující obsah reálného textu (průhledná modrá).

3.1.2. Ruční a automatizovaná transkripce

Následujícím krokem je samotná transkripce dokumentů. Zde je dobré se držet pravidel transkripce pro ediční práci. Transkribus je schopný rozepisovat také abreviace, rozpozná ligatury, diakritiku a další standardní prvky transkripce.[217]

Samotnou transkripci je možné dělat pod oknem se zobrazením dokumentu, který si lze libovolně přibližovat. Transkripční prostředí má číslované řádkování pro snazší orientaci a dokument si lze přiblížit a oddálit.

Mimo prvky standardní ruční transkripce lze dokument také obohatit o metadata při transkribování, a to v záložce Metadata, kdy lze označit tagy různě důležité prvky textu – jména, místa, názvy, data a další kontextuální informace.[218] Tyto metadata se potom dají využít pro efektivnější analýzu a vyhledávání. Při

---

[215] Text2Image. *Transkribus* [online]. 2019 [cit. 2020-11-17]. Dostupné z: https://transkribus.eu/wiki/index.php/Text2Image

[216] Autorem vytvořený snímek ukázky segmentace v programu Transkribus.

[217] Conventions for special characters. *READ-COOP* [online]. 2020 [cit. 2020-11-17]. Dostupné z: https://readcoop.eu/transkribus/howto/transkribus-transcription-conventions/

[218] Tagging. *READ-COOP* [online]. 2020 [cit. 2020-11-17]. Dostupné z: https://readcoop.eu/transkribus/howto/transkribus-transcription-conventions/

transkripci lze využít funkce virtuální klávesnice, která obsahuje řadu nestandardních znaků, které mohou být pro správnou transkripci potřeba. Například tedy znaky řecké abecedy, hebrejština, latinské ligatury (např. ae) a další.

Pro efektivní automatizovanou transkripci je oficiálně doporučeno ručně transkribovat alespoň 15 000 slov nebo 75 stran.[219]

Kritická je přesnost výsledné transkripce, protože tato data budou použita pro trénování modelu. Výsledný neurální model na straně serveru se totiž tato trénovací data naučí chybně, pokud mu jsou chybně dodána.

Před samotnou transkripcí je možné také využít integrovaného OCR řešení Abbyy Finereader 11 k vytvoření alespoň částečné transliterace[220], která může ulehčit práci. Velmi ale záleží na charakteru transkribovaného textu.

Po vytvoření tohoto souboru ručně transkribovaných dokumentů jako trénovacích dat zbývá už jen nahrát netranskribované dokumenty a využít uživatelem vytvořený trénovací model a poté počkat na zpracování serverové strany a pokud je výsledek dostatečně přesný jej exportovat. Export lze provést do formátů PDF, DOCX, METS, ALTO, TXT a XML mimo toto lze exportovat také komprimovaný balík.

Mimo právě vytvořený model lze využít některé z dostupných modelů, kdy každý má základní popis obsahující počet slov, CER, typ a druh písma a jazyk, období a také graf popisující jeho přesnost. Existující dostupné modely se pohybují mezi CER od 0,10 % po 11 %, kdy většinou se jedná spíše o jednotky či desetiny procenta. Tyto modely mají ale poměrně úzké využití pro specifické texty.

Výsledná doba pro získání automatizované transkripce na základě uživatelem poskytnutých trénovacích dat závisí na objemu textu a na vytíženosti serverové strany. V případě nedostačující CER lze trénovací model upravit a doplnit o další

---

[219] Handwritten Text Recognition Workflow: Prerequisites. *READ-COOP* [online]. 2020 [cit. 2020-11-17]. Dostupné z: https://readcoop.eu/transkribus/wiki/handwritten-text-recognition-workflow/

[220] OCR systémy nedokážou provádět transkripci oproti HTR technologii, která toto dokáže

ručně transkribované strany. Někdy je potřeba tento postup zopakovat vícekrát pro dosažení vyhovující přesnosti transkripce.

3.1.3. Možnosti po zpracování

Po získání automatizované transkripce je možné pro další zpracování vyhledávací funkce (ikona dalekohledu). Lze vyhledávat standardně fulltextově, pomocí tagů anebo využít hledání klíčových slov (záložka KWS).[221] Klíčová slova lze také kombinovat.

Výsledek tohoto hledání je potom počet výskytů, strana, kde se slovo nachází a grafická ukázka výskytu, což je ořez snímku dokumentu obsahující pouze hledanou frázi. Při vyhledávání lze také zohlednit velká či malá písmena a také vyhledávat shody pouze části[222] slov.

3.2. Výhody a nevýhody

Charakteristickým znakem Transkribu je jeho multifunkčnost. Klientská aplikace do sebe integruje funkce OCR, správce souborů a kolekcí, transkripčního prostředí, metadatového popisu a především HTR. Pro celý proces automatizované transkripce tedy není potřeba žádná další externí aplikace. Zdrojový kód pro různé části aplikace je navíc volně dostupný.

Důležitá je rozhodně také podpora ze strany poskytovatele. Poskytovaný informační servis je velmi kvalitní a různorodý – obsahuje standardní textové návody, popisy, dokumentace a video návody. Dostupná je také informační linka v podobě emailu a kontakty na všechny nejdůležitější pracovníky v případě dotazů. Mimo to jsou poskytovány dostatečné příklady využití celé platformy společně s demoverzemi.

---

[221] How To Search Documents with the Keyword Spotting Feature. *READ-COOP* [online]. 2020 [cit. 2020-11-17]. Dostupné z: https://readcoop.eu/transkribus/howto/how-to-use-keyword-spotting/

[222] Příkladem může být například vyhledávání koncovky „ovi“ v dokumentu. Funkce v tomto případě vrátí slova Petříčkovi, Holubovi, Vránovi.

Pro dosažení dostatečné kvality ruční transkripce, a tudíž lepšího výsledku automatizované transkripce poskytuje Transkribus také vzdělávací portál[223] s paleografickými cvičeními, který je dostupný v digitálním prostředí skrze webový prohlížeč. Tento vzdělávací portál je volně dostupný, a i v případě nevyužití aplikace Transkribus má svou hodnotu například pro výuku paleografie.

Platforma Transkribus je navíc poměrně dobře zastoupená univerzitami a dalšími institucemi napříč Evropou. Uživatelská základna je tedy poměrně značná. Projekt je navíc funkční od roku 2015 a tak měl již určitou dobu na to se vyvinout a vyspět. V současné době vychází stále nová rozšíření, opravy a další přidané funkce, například výše zmíněný modul P2PaLA. Projekt byl původně financován EU pod programem EU Horizon 2020 READ [224] a do podzimu 2020 byl plně dostupný zadarmo.

Od podzimu 2020 přešel na nový model, kdy byl zaveden systém kreditů. Klientská aplikace je nadále dostupná zdarma, ale pro využití HTR neurální sítě na straně serveru je nutné zaplatit právě těmito kredity. Každý nově registrovaný uživatel má zdarma 500 kreditů. V současnosti je to nastavené tak, že 1 kredit = 1 strana převedeného textu. Novější HTR+ model by měl umožnit ještě levnější převod. Celkově se Transkribus snaží o to, aby platební model byl co nejdostupnější pro co nejširší řadu uživatelů. Pro studenty a vyučující relevantních oborů je navíc možnost individuální dohody a zpřístupnění programu v celé šíři zdarma. Na webových stránkách je nabízena také možnost předplatného, takže se rozhodně nejedná o drahé řešení. V aplikaci dostupné OCR je zmíněné ABBYY Finereader, které ve výše zmíněných testech dosahovalo velmi dobrých výsledků.

Další výhody se týkají HTR technologie spíše obecněji než přímo tohoto konkrétního řešení Transkribus. HTR umožňuje převod v podstatě jakéhokoliv textu psaného libovolným písmem, stylem a jazykem, což je základem pro převod jakéhokoliv[225] historického textu, který je už z podstaty své doby velmi

---

[223] Odkaz na vzdělávací portál: https://transkribus.eu/r/learn/app/documents/

[224] The roots of Transkribus. *READ-COOP* [online]. 2020 [cit. 2020-11-19]. Dostupné z: https://readcoop.eu/transkribus/?sc=Transkribus

[225] Rozdíl je patrný především v porovnání s OCR technologiemi.

různorodého charakteru. Kompletní proces převodu textu do transkribované podoby umožňuje plnohodnotnou digitální ediční práci s výstupem navíc obohaceným o hlubší metadatový popis, a tudíž výkonnější možnosti vyhledávání v textu. S dostatečně kvalitními trénovacími daty je tedy teoreticky možné převést téměř jakkoliv[226] starý a špatně dostupný historický text, což je hlavní výhoda HTR technologie.

Z podstaty HTR technologie má celý projekt Transkribus dvě základní slabiny. První z nich je náročnost na výpočetní výkon a lidské zdroje, což je obojí poměrně drahou záležitostí. Toto je kritické především nyní, kdy je financování z EU projektu READ již podstatně omezeno. Otázkou zůstává, zda nově nasazený platební model je udržitelný a jestli o něj bude dostatečný zájem.

Druhou zásadní slabinou celé platformy je nutnost vytváření alespoň částečných ručních transkripcích, což je časově i znalostně poměrně náročný úkon, a to především, pokud jsou brány v potaz nároky na kvalitní a přesné transkripce fungující jako trénovací data pro HTR model. S tím souvisí také dosavadní dostupné modely, které jsou zatím dostupné pouze v cizojazyčné podobě. Pro české prostředí jsou víceméně využitelné jen ty latinské, německé a částečně také polské. Bez ručních transkripcí pro automatizovanou transkripci je HTR technologie nedostačující, a tudíž je závislá na uživatelské základně, která tyto transkripce dokáže produkovat v dostatečné kvalitě.

Pro celý projekt je tedy kritická dostatečně velká uživatelská základna. Pokud si Transkribus dokáže udržet dostatečně velkou uživatelskou základnu, pokryje tím jak finanční náročnost celého projektu za současného kreditového systému, a navíc bude stále získávat nové trénovací modely pro použití pro HTR model.

Mimo tyto dvě základní slabiny se navíc s přechodem na komerční platební model dostává Transkribus do konkurence s ostatními dostupnými řešeními jako je Quartex, řešení od Microsoftu, Google a Amazonu.

[226] Samozřejmě text nesmí být z větší části poškozený nebo úplně nečitelný.

To pro potřebu textů české provenience psané z velké části v češtině zatím není problém, protože další řešení češtinu[227] zatím dostatečně nepodporují, ale to může být pouze otázka času. Z výše zmíněných charakteristických slabin vyplívá také slabina na trhu, kdy pokud by se některé z výše zmíněných řešení uplatnilo více, mohlo by dojít k roztříštění trhu a tudíž uživatelské základny a HTR technologie jsou nakonec jen tak silné, jako je silná jejich uživatelská základna.

Z podstaty HTR technologie je také jasné, že vytvořené trénovací modely mají velmi často poměrně specifické využití pro konkrétní podobu textu a jejich využití mimo ni je limitované. Pro odlišný text, psaný stejným písmem, ale například jiným autorem bude potřeba buď doplnit model o další trénovací data. Původní model může posloužit ale alespoň jako určitá základna.

Kromě těchto zásadních slabin lze zmínit ještě drobnější nedostatky, které jsou jednodušeji napravitelné. V současnosti neexistuje podpora pro český jazyk jak pro klientskou aplikaci, tak webové stránky, zmíněný informační servis a ani výukové prostředí. To je mimo jiné způsobeno tím, že mezi členskými institucemi[228] zatím nefiguruje žádná česká univerzita ani paměťová instituce.

Mimo samotný software lze zmínit ještě výše zmíněné řešení ScanTent spadající pod celou platformu Transkribus. Toto stanové řešení má svoje výhody a nevýhody. Je poměrně levně dostupné, přenositelné a může být tak efektivní pro potřeby výuky a zaučení. Pro reálné použití v paměťových institucích, kde existují specializované skenery různých typů, které jsou efektivnější jak ve smyslu rychlosti, tak kvality převodu je jeho využití minimální. Rozhodně by se ale dalo zmínit jeho využití pro instituce, kde tato zařízení dostupná nejsou. Stanová konstrukce sama o sobě je také poměrně drahá.

---

[227] Němčina a latina navíc tento problém nemají, a v českém prostředí byli také velmi hojně využívané.

[228] Pro naše prostředí nejbližším členem bude Univerzita Mateja Bela v Banskej Bystrici na Slovensku. Kompletní seznam členů si lze prohlédnout na: https://readcoop.eu/members/

4. Využití počítačem čitelných archiválií a dokumentů

Na archiválie lze nahlížet nejen z klasického archivního hlediska jako něco, co je třeba uchovat, organizovat a prezentovat, ale také na ně lze nahlížet jako počítačem zpracovatelná data z pohledu datové vědy a obecně informatiky. Tento pohled na archiválie přináší nové možnosti jejich zpracování jako historického pramene. Aplikování těchto technik a metod by měla být ve výsledku obohacující, přidaná hodnota k standardní archivní nebo historické práci. Mimo tuto přidanou hodnotu umožňují efektivnější a rychlejší práci s větším množstvím dat, což je důležité především proto, že počet digitalizovaných, a v některých případech i čistě digitálních, materiálů stále roste, což je ještě více patrné, pokud vezmeme v potaz také zahraniční prameny. V současnosti se digitalizují miliony knih a dalších historických pramenů a dokumentů včetně novin.[229] Kvůli tomuto růstu je potřeba umět efektivně s těmito vyššími objemy dat nakládat, což bude také do budoucna stále více a více relevantní.

Tato kapitola si klade za cíl přiblížit[230] nástroje a technologie umožňující další využití počítačem čitelných archiválií, což umožňuje vidět prameny z jiného úhlu pohledu a přináší tak jiný typ informací a hlubší poznání. Mimo to se soustředí na klady a zápory těchto technik a postupů a také možné překážky při jejich nasazení.

---

[229] PLATILOVÁ, Mgr. Ivana. *Možnosti aplikace počítačových metod v historii* [online]. Olomouc, 2019 [cit. 2020-12-10]. Dostupné z: https://theses.cz/id/pwyry9/?zpet=%2Fvyhledavani%2F%3Fsearch%3DMo%C5%BEnosti%20aplikace%20po%C4%8D%C3%ADta%C4%8Dov%C3%BDch%20metod%20v%20historii%26start%3D1;isshlret=Mo%C5%BEnosti%3Baplikace%3Bpo%C4%8D%C3%ADta%C4%8Dov%C3%BDch%3Bmetod%3B. Diplomová. Univerzita Palackého v Olomouci. Vedoucí práce doc. Mgr. Martin Elbel, M.A., Ph.D.

[230] Dobrou praktickou ukázkou některých těchto nástrojů na reálném pramenu je diplomová práce z roku 2019 Ivany Platilové *Možnosti aplikace počítačových metod v historii*.

4.1. Data scraping

Pod pojmem scraping se rozumí extrahování dat z různých typů dokumentů.[231] Škála těchto typů dokumentů je poměrně široká, může se jednat o textové dokumenty (PDF, DOC, DOCX), tabulky (XLS, XLSX, CSV), hypertextové dokumenty (HTML, XML) a mnoho dalších typů dokumentů. Obecně je to tedy extrahování relevantních dat z různých typů elektronických zdrojů.

Data scraping lze provádět manuálně tj. uživatel sám prochází dokument za dokumentem a důležité informace sám extrahuje. Toto je ale poměrně pomalé a neefektivní. Proto se využívá různých nástrojů pro automatizaci tohoto procesu.[232] Extrahování dat z webových stránek se nazývá web scraping[233].

Data scraping je zásadní pro další práci s elektronickými dokumenty. Velmi důležitý je potom objem dat, s kterým je nutno dále pracovat. Web scraping jako technika je tím účinnější, čím více dat je pro další analýzu potřeba. Z archivního pohledu tedy závisí na velikosti archivních souborů a sbírek. Jak bylo zmíněno výše, růst eDokumentů je poměrně rychlý a digitalizátů vzniká stále více a více, a proto je potřeba umět tato data ze svých původních zdrojů efektivně extrahovat, a to ať se jedná o webové aplikace, archivní úložiště, datové sklady nebo další možnosti ukládání.

V současnosti existuje několik předních nástrojů pro data scraping. U většiny z nich je nutné počítat alespoň se základní znalostí programovacího jazyka Python[234], který se pro tyto potřeby velmi hojně používá. Naštěstí pro historiky, archiváře a další pracovníky v této oblasti je Python považován za jeden

---

[231] Data Scraping. *Techopedia* [online]. 2020 [cit. 2020-12-15]. Dostupné z: https://www.techopedia.com/definition/33132/data-scraping

[232] Web Scraping vs Data Mining: What's the Difference? *Parsehub* [online]. 2020 [cit. 2020-12-15]. Dostupné z: https://www.parsehub.com/blog/web-scraping-vs-data-mining/

[233] Což je populárnější termín často zaměňovaný s termínem Data scraping.

[234] Python je velmi relevantním jazykem pro datovou vědu, a tudíž také pro moderní archivní a historickou práci. S Pythonem se lze setkat v každé kapitole této diplomové práce. Transkribus používá PyTorch, který je postavený na Pythonu. Zmíněná OCR řešení jsou psaná v Pythonu a mnoho dalších. S Pythonem se lze setkat také jako základním nástrojem projektů Digital Humanities a Digital History napříč vědeckými články.

z nejdostupnějších jazyků – je velmi hojně využíván napříč různými odvětvími včetně Digital Humanities, existuje pro něj řada návodů, knihoven a pluginů. Navíc je považován za jeden z nejjednodušších jazyků na naučení.[235] Mimo to je Python velmi univerzálním jazykem schopným vytvářet různé typy výstupů[236].

Pro moderní digitální historii a archivnictví je scraping zásadní technika, která umožňuje efektivnější přístup a sběr informací z digitálních repositářů archivů, knihoven, vědeckých institucí a dalších zdrojů nejen českých, ale také zahraničních, což s ohledem na vcelku vysoký objem dostupných dat pro zpracování při historické analýze může poměrně efektivně šetřit čas a zvýšit množství dat, která by jinak při historické analýze zůstala opomenuta. Historik je tak schopen pro svou práci získat více dat za kratší časový úsek.

4.1.1. Textract

Je jedním z takových nástrojů. Slouží pro extrahování dat z poměrně širokého spektra[237] souborů, což je také jeho největší výhoda. Těmi základními jsou CSV, XLS, XLSX (tabulkové a databázové soubory), TXT, DOC, DOCX, PDF (textové soubory), WAV (zvukové soubory skrze speech to text), PNG, JPEG, JPG (obrazové dokumenty za pomocí Tesseract OCR) a JSON (formát pro výměnu dat).

Tento balíček je výhodný především v tom, že poskytuje jednotné prostředí a integruje do sebe funkce různých samostatných balíčků pro jednotlivé formáty.[238] Umožňuje extrahovat obsah téměř jakéhokoliv souborů, je dostupný zadarmo s otevřeným zdrojovým kódem.

---

[235] Python: Get Started. *Python.org* [online]. 2020 [cit. 2020-12-15]. Dostupné z: https://www.python.org/

[236] Jako jsou webové aplikace, projekty s použitím strojového učení, NLP, Computer Vision, Web scraping, desktopové aplikace aj.

[237] Úplný výčet je dostpuný na webových stránkách: https://textract.readthedocs.io/en/stable/

[238] Textract. Textract [online]. 2014 [cit. 2020-12-15]. Dostupné z: https://textract.readthedocs.io/en/stable/

Textract[239] lze importovat jako standardní Python balíček za pomocí příkazů:

```
import textract
text = textract.process("cesta/podsložka/..../název_souboru.přípona")
```

Nebo pomocí příkazové řádky:

```
textract cesta/podsložka/..../název_souboru.přípona
```

Pomocí tohoto balíčku lze z libovolného množství vybraných dokumentů s podporovaným souborovým formátem extrahovat textové informace, ty potom uložit dle zadaných parametrů. Při extrakci je nutné vybrat pomocí parametru extension souborový formát a lze také specifikovat výstupní jazyk. Mezi nimi většinou není latina ani další historické jazyky.

Při zavolání na knihovnu textractu pomocí výše uvedeného příkazu je vyhledán tzv. parser, což je základ pro extrahování textu z dokumentu.[240] Textract podporuje několik základních příkazů pro práci s dokumentem. První dva jsou encode a decode, které umožňují měnit kódování např. z ASCII do UTF-8 nebo Windows 1252. Příkaz decode(text) text dekóduje a encode(text) zakóduje.

Kódování je pro české prostředí důležité, protože ne každý typ kódování podporuje specifické české znaky jako jsou háčky a čárky nad písmeny. Na toto je potřeba si dát obzvláště pozor, protože pokud by uživatel změnil kódování dokumentu na takový typ kódování, který nepodporuje tato interpunkční znaménka, došlo by k neúplnému[241] převodu a tyto znaky by byly nedostupné.

Další zásadní příkaz je příkaz process(název_souboru, typ_kódování). Pomocí tohoto příkazu se text extrahuje z dokumentu. Práce s tímto nástrojem není příliš složitá, pro extrahování kompletního textu z jednoho dokumentu není potřeba napsat více než čtyři řádky. Jednoduchou ukázkou jsou následující čtyři řádky, které extrahují úplný text ze souboru dokument1 ve formátu pdf v českém jazyce.

---

[239] Možná je dobré zmínit také AWS Textract, což je naprosto odlišný nástroj od Amazonu zmíněný v kapitole 2.4.3. Jedná se o dva odlišné nástroje se stejnými jmény.

[240] Python package. *Textract* [online]. 2014 [cit. 2020-12-15]. Dostupné z: https://textract.readthedocs.io/en/stable/python_package.html

[241] Textract ale nijak nevymaže původní soubory, ty zůstávají nedotčené.

```
text = textract.process(
'archiválie/19.stol/dokument1.pdf',
'language=cz',
)
```

Obecně se nástroj vyplatí používat, pokud uživatel zpracovává větší množství dokumentů v podporovaném formátu a to především v textové podobě. Textract sice podporuj extrahování textových dat z obrazových dokumentů a zvukových nahrávek, ale jejich převod je oproti standardním textovým dokumentům méně spolehlivý a přesný.

4.1.2. Web scraping

Web scraping jak už název napovídá slouží k extrahování dat z webových stránek. Web scraping je zásadní pro sběr velkých data setů, které jsou nezbytné pro Big data analýzy, strojové učení a další statistické analýzy.[242] A to především z výše zmíněných důvodů. Nicméně web scraping má oproti klasickému data scrapingu svou charakteristickou překážku – a totiž Javascript.

Při vytváření webových stránek se velmi často využívá Javascript, který je základem moderního webového designu ať za pomocí pouze samotného Javascriptu, nebo za pomocí moderních knihoven typu Jquerry a Bootstrap. Vše ovšem závisí na šíři využití Javascriptu při generování webové stránky – pokud není využit žádný Javascript a výsledná stránka je tedy pouze hypertextový dokument, web scraping je poměrně snadnou záležitostí, protože uživatel tak má přístup ke všem datům webové stránky.

Problém nastává po použití Javascriptu při generování webové stránky, protože pro zobrazení takto generovaného obsahu vyžaduje použití Javascriptu. Web

[242] Big Data: What is Web Scraping and how to use it. *Towards data science* [online]. 2018 [cit. 2020-12-17]. Dostupné z: https://towardsdatascience.com/big-data-what-is-web-scraping-and-how-to-use-it-74e7e8b58fd6

scraping je v tomto případě poněkud složitější záležitostí, protože vyžaduje využití pokročilejších nástrojů typu Selenium.[243]

Mimo tento charakteristický problém se zobrazováním úplného obsahu hypertextového dokumentu je nutné zmínit ještě problém se sémantikou, kdy v korektně napsaném sémantické webové stránce je mnohem snazší se orientovat, ať už pro lidské oko, tak pro scrapery, které se pro web scraping používají.

S tím souvisí také další překážka, kterou web scraping často trpí a tou je použití technologií typu CAPTCHA, které mají zamezit přístupu botům, což vlastně scrapery, které web scraping provádí jsou. Znovu je zde zásadní alespoň základní znalost Pythonu.

4.1.2.1. Nástroje

Pro web scraping se používají boti, tzv. crawlers nebo scrapers. Pro web scraping existuje řada nástrojů, knihoven a doplňků. Většina je psaná v Pythonu.[244] Mezi nejpopulárnější patří Selenium a Scrapy.

Selenium je populárním nástrojem především proto, že je velmi uživatelsky přívětivé a je jednoduché s ním psát kód.[245] Proto má také velkou uživatelskou základnu, a tudíž řadu informačních videí, návodů, dalších doplňků a dalších podpůrných prvků.

Mimo to je Selenium dobrou volbou právě pro obsah generovaný Javascriptem, oproti Scrapy, se kterým je poněkud těžší takovýto obsah procházet a pracovat s ním. Naopak Scrapy je robustnějším a ohebnějším nástrojem pro větší data sety.[246]

---

[243] Big Data: What is Web Scraping and how to use it. *Towards data science* [online]. 2018 [cit. 2020-12-17]. Dostupné z: https://towardsdatascience.com/big-data-what-is-web-scraping-and-how-to-use-it-74e7e8b58fd6

[244] Big Data: What is Web Scraping and how to use it. *Towards data science* [online]. 2018 [cit. 2020-12-17]. Dostupné z: https://towardsdatascience.com/big-data-what-is-web-scraping-and-how-to-use-it-74e7e8b58fd6

[245] YIN, Michael. Web Scraping Framework Review: Scrapy VS Selenium. *AccordBox* [online]. 2019 [cit. 2020-12-17]. Dostupné z: https://www.accordbox.com/blog/web-scraping-framework-review-scrapy-vs-selenium/

[246] YIN, Michael. Web Scraping Framework Review: Scrapy VS Selenium. *AccordBox* [online]. 2019 [cit. 2020-12-17]. Dostupné z: https://www.accordbox.com/blog/web-scraping-framework-review-scrapy-vs-selenium/

Nakonec co se týče nástrojů, existuje jich opravdu velká řada a pro každý projekt může být vhodnější jiný nástroj či knihovna.

Existují také vizuální řešení, kdy je nutnost programovat minimální. Tyto nástroje ovšem mají své limity především v rychlosti zpracování, velikosti data setů a vůbec ohebnosti, kterou vlastní program umožňuje. Může se ale jednat o efektivní nástroje, které stojí za to vyzkoušet jako první, pokud chce uživatel provádět jednoduchý, méně rozsáhlý web scraping. Některé populární nástroje tohoto typu jsou DataMiner, Portia a Webscraper.io, které jsou všechny dostupné v podobě pluginů do Google Chrome prohlížeče. Lze je tedy jednoduše nainstalovat a jejich používání je poměrně jednoduchou záležitostí.

4.1.2.2. Využití

Web scraping má využití především v přípravě a sběru dat pro další práci s nimi. Lze jej využít na jednotlivých webových aplikacích oblastních archivů či Národního archivu s digitalizovanými archiváliemi právě pro jejich automatizovaný sběr pro další nejen historickou práci. Velké množství dat, které lze pomocí této techniky získat je důležité především pro statistiku a dolování v datech, kdy je právě množství těchto dat důležité pro dostatečnou přesnost výsledků těchto analýz a jejich manuální sběr by byl příliš pomalý a neefektivní.

Moderní historická analýza ale není jediným využitím této techniky. Využití má také v oblasti státní správy a předarchivní péče také za pomocí pokročilých Big Data analýz, dolování v datech a statistických šetřeních, které můžou být relevantní pro archiv jako instituci anebo archivní vědy jako takové.

4.2. Text mining

Text mining je specifická kategorie dolování v datech zabývající se prací s textovými[247] daty na pomezí lingvistiky a počítačové vědy. Zatímco scraping slouží pouze k extrakci dat do vhodného formátu, text mining slouží k samotné analýze těchto dat[248] jejíž výsledkem je získávání znalostí. V celém procesu získávání znalostí z dat je tedy scraping jakýmsi předzpracováním a přípravou a text mining je tedy souhrn procesů, metod a technik sloužící k jejich analýze, ze které lze vypozorovat požadované výsledky. Pro text mining se využívá řada nástrojů, znovu platí, že alespoň základní znalost Pythonu a SQL je velmi důležitá.

4.2.1. Základní metody a procesy

Obecný základní workflow při dolování v textu je sběr, normalizace a očištění dostupných dat. Následuje identifikace vzorů[249], jejich následná analýza a výsledné extrahování znalostních informací[250].[251] Mezi základní techniky[252] dolování v textu

---

[247] Znovu je třeba zdůraznit, že většina typů pokročilejší analytické práce s daty vyžaduje právě textovou podobu dat, proto je převod obrazového materiálu do textové podoby tak zásadní.

[248] PEREZ, M. Web Scraping vs Data Mining: What's the Difference? *Parsehub* [online]. 2020 [cit. 2021-01-19]. Dostupné z: https://www.parsehub.com/blog/web-scraping-vs-data-mining/

[249] Pod vzory si lze představit určité opakující se prvky v textu, které lze identifikovat, analyzovat a učinit z nich nějaký výstup. Jednoduchým příkladem je kombinace výskytu slov „prodává, grošů, kop" (a mnoho dalších) v textu. Tato informace, že se v textu vyskytuje tato kombinace slov je užitečná a lze z ní vyvodit závěry → pravděpodobně se jedná o listinu.

[250] Tyto informace mohou mít význam nejen pro osobu badatele, ale také další počítačové nástroje, které na jejich základě dokáží získat další, komplexnější výstupy (Například klasifikátory nebo vytváření automatizovaných obsahů dokumentů.

[251] SHILPA, Dang. Text Mining : Techniques and its Application. *International Journal of Engineering & Technology Innovations* [online]. Indie: M.M Institute of Computer Technology & Business Management Maharishi Marakandeshwar University, 2014, (4), 22-25 [cit. 2021-01-20]. ISSN 2348-0866. Dostupné z: https://www.researchgate.net/publication/273038150_Text_Mining_Techniques_and_its_Application

[252] Zmíněných technik je mnohem více, pro potřeby diplomové práce jsou zmíněny pouze ty nejzásadnější pro specifický okruh archivní a historické práce.

patří výběr a extrahování informací, klasifikace, frekvence, shlukování a sumarizace.

Základním nástrojem dolování v textu je například samotný webový prohlížeč, který používá každý. Jednoduchým příkladem je např. Google Ngram viewer[253], který umožňuje vyhledávat Ngramy, tedy počty slov či slovních spojení ve všech publikacích zveřejněných na Google Books a zasadit je do časového výseku buď do grafu, nebo tabulky. Toto je užitečné například při zjišťování vědeckého zájmu o danou problematiku.

4.2.1.1. Sběr a předzpracování dat

Prvním krokem je standardně zúžení množství sesbíraných dat pouze na ty relevantní. Po snížení objemu dat na relevantní data následuje proces předzpracování dat – základem je čištění dat, kdy se odstraní neúplné, chybné a redundantní hodnoty. Následuje transformace dat a jejich generalizace pomocí slučovacích technik a různých typů agregací, kompresních metod či redukci rozměrů. Tyto procesy se dějí buď manuální činností či automatizovaně zpravidla pomocí skriptů.

4.2.1.2. Frekvence

Nejběžnější a nejjednodušší technika dolování v textu je frekvence. Jedná se o identifikování počtů opakujících se slov v dokumentu. Výsledkem je frekvenční slovník, který obsahuje právě počty těchto opakujících se slov[254], na jehož základě je možné vyvodit určité závěry, anebo jej využít pro další dolování či analýzu.

---

[253] Nástroj si lze vyzkoušet na následujícím odkazu: https://books.google.com/ngrams/graph?content=Digital+Humanities%2CNatural+language+processing&year_start=1980&year_end=2019&corpus=26&smoothing=3&case_insensitive=true. Tento konkrétní odkaz zobrazuje výskyt slov Digital Humanities a Natural language processing v publikacích v čase a tím pádem zobrazuje zvýšení vědeckého zájmu o tyto oblasti.

[254] Důležité je provést výběr těchto opakujících se slov – spojky, částice, zájmena a další nemají pro nějakou další analýzu přílišný význam (pokud se tedy nejedná přímo o práci zabývající se např. staročeštinou).

Standardně se využívá při analýze zákaznických hodnocení, sociálních sítí či zákaznické odezvy a dalších.[255]

Zjištění počtu opakujících se slov je užitečné ale i pro využití při archivním zpracování či historické práci.

Příkladem může být zpracování rozsáhlejší korespondence. Při vysokých nebo nízkých počtech opakujících se slov lze tedy vyvodit určité závěry: lze zjistit, v jakém postavení je autor k respondentům, podle toho, jak je oslovuje. Komunikuje autor spíše s osobami níže postavenými nebo výše postavenými? Lze zjistit o jakých tématech v jakých obdobích autor dopisuje podle počtu tematických slov např. „revoluce“, „podzim“, „1917“ a další. Mimo tematické a časové určení lze z frekvencí[256] slov zjistit, na co autor klade důraz a co ho v té dané konkrétní době zajímá, čemu se věnuje a jaké z toho má pocity. Frekvence je ale ve výsledku poměrně povrchní sondou, která ale může fungovat jako důležitý prvek postavené hypotézy.

Mimo tyto zmíněné příklady určitě existuje také řada dalších využití, tuto činnost lze samozřejmě provést manuálně, ale tento postup by byl pomalejší a méně efektivní.

4.2.1.3. Kolokace a konkordance

Zjišťování samotných frekvencí opakujících se slov je samo o sobě užitečné, ale z těchto slov lze vytěžit mnohem více informací právě pomocí kolokace a konkordance.

Konkordance umožňuje zobrazení slovních spojení předcházejících a následujících vybrané slovo např. u takových slov, které se často opakují a jsou tudíž důležité. Funguje tedy buď jako doplňková část analýzy, která přidává kontext, anebo pouze jako zvýraznění daného slovního spojení, což je samo o sobě

---

[255] What is Text Mining: Methods and Techniques. *MonkeyLearn* [online]. [cit. 2021-01-21]. Dostupné z: https://monkeylearn.com/text-mining/

[256] V tomto případě je samozřejmě zásadní brát ohled na množství dopisů v poměru k výskytům slov → více slov ale také větší objem korespondence neznamená vyšší důraz na dané slovo.

poměrně užitečné. Analýzou konkordance slova může pomoci při porozumění kontextu obsahu.

Výstupem je potom grafické zobrazení hledaného slova a sousedních slovních spojení, které není příliš odlišné[257] od následující tabulky.

Předchází	Cílové slovo	Následuje
nyní	pracuji	na nové knize
dříve jsem	pracoval	v novinách
teď	pracuje	s Karlem

Obr. 12 – ukázka konkordance[258]

Konkordance je tím složitější, čím více ohebný jazyk je. Například v angličtině by toto nebyl žádný problém, protože vyhledávané slovo by bylo pořád „work" nebo „working". Čeština je ale poměrně flexivní jazyk, takže je konkordance složitější.

Kolokace filtruje vyhledané slovo společně se sousedními slovy. Kolokace jsou vlastně slovní spojení, která se v textu opakují. Znovu se jedná spíše o upřesnění kontextu a další způsob filtrování, který umožňuje zobrazovat konkrétní slovní spojení, která mají pro badatele význam. Například slovo „král" se opakovaně vyskytuje v kontextu slov „hloupý", „Francouzský", „falešný" aj[259].

Mimo to lze konkordanci, kolokaci a frekvenci analyzovat pomocí nástrojů s grafickým rozhraním – tzv. korpusových manažerů, které mohou být pro historiky a archiváře přístupnější[260]. Celkově tedy konkordance a kolokace slouží k upřesnění a zvýraznění určitého kontextu v daném souboru textů.

---

[257] Samozřejmě záleží na využité aplikaci anebo napsaném skriptu.

[258] Vlastní tvorba na základě: What is Text Mining: Methods and Techniques. *MonkeyLearn* [online]. [cit. 2021-01-21]. Dostupné z: https://monkeylearn.com/text-mining/

[259] PLATILOVÁ, Mgr. Ivana. *Možnosti aplikace počítačových metod v historii* [online]. Olomouc, 2019 [cit. 2020-12-10]. Dostupné z: https://theses.cz/id/pwyry9/?zpet=%2Fvyhledavani%2F%3Fsearch%3DMo%C5%BEnosti%20aplikace%20po%C4%8D%C3%ADta%C4%8Dov%C3%BDch%20metod%20v%20historii%26start%3D1;isshlret=Mo%C5%BEnosti%3Baplikace%3Bpo%C4%8D%C3%ADta%C4%8Dov%C3%BDch%3Bmetod%3B. Diplomová. Univerzita Palackého v Olomouci. Vedoucí práce Doc. Mgr. Martin Elbel, M.A., Ph.D.

[260] SQL a Pythonu se při dolování v textu a dalších zmíněných uživatel ale stejně nevyhne.

4.2.1.4. Klíčová slova

Cílem této metody je rozpoznávání klíčových slov v daném textovém souboru. Tato klíčová slova jsou zásadní pro další zpracování, a to především v rozpoznávání obsahu textu, jeho sumarizaci a klasifikování.

Pod klíčovými slovy si lze znovu v případě např. listiny[261] představit slova spadající do diplomatických kategorií popisu – tedy např. intitulace, invokace, devoční formule, sankce a mnoho dalších. Obecně platí, že čím více společných znaků, které se opakují u daného typu písemnosti takováto písemnost má, tím snazší je tato klíčová slova odhalit, což v případě listiny platí.

Nicméně program potřebuje určitý klíč k rozpoznávání těchto klíčových slov. Pro získání tohoto klíče existuje řada způsobů. Nejjednodušší je manuální nastavení klíčových slov, tj. že programátor manuálně nastaví sadu klíčových slov, které jsou pro danou část programu důležitá, tj. u listiny nastaví taková klíčová slova, která se v listinách opakují[262], tedy například groš, kopa, in perpetuum, z milosti Boží, kšaft aj. Tato předpřipravená klíčová slova musí buď autor kódu vymyslet sám, anebo může použít tzv. referenční korpusy, které tyto slova už obsahují.[263]

Manuální nastavení těchto slov je ale kolikrát nedostatečně přesné a také časově náročné. Mnohem vhodnější je využít strojového učení, kdy neurální síť po poskytnutí dostatečně velikého vzorku toho daného typu dokumentů, u kterých se klíčová slova sledují sám naučit. Nejprve je ale nutné takovéto síti poskytnout dostatečný vzorek s řadou dokumentů, u kterých jsou již klíčová slova vybrána.

Klíčová slova mají určitý význam i pro samotného badatele stejně jako v případě publikací[264]. Zcela zásadní jsou ale při dalším strojovém zpracování, a to

---

[261] Listina je jedna z těch jednodušších na zpracování, protože je poměrně formulační.

[262] Tedy standardní (písemné) znaky, kterými se listina charakterizuje.

[263] PLATILOVÁ, Mgr. Ivana. *Možnosti aplikace počítačových metod v historii* [online]. Olomouc, 2019 [cit. 2020-12-10]. Dostupné z: https://theses.cz/id/pwyry9/?zpet=%2Fvyhledavani%2F%3Fsearch%3DMo%C5%BEnosti%20aplikace%20po%C4%8D%C3%ADta%C4%8Dov%C3%BDch%20metod%20v%20historii%26start%3D1;isshlret=Mo%C5%BEnosti%3Baplikace%3Bpo%C4%8D%C3%ADta%C4%8Dov%C3%BDch%3Bmetod%3B. Diplomová. Univerzita Palackého v Olomouci. Vedoucí práce doc. Mgr. Martin Elbel, M.A., Ph.D.

[264] Například takové diplomové práce.

v případě vytváření automatizovaného systému schopného textové soubory kategorizovat a jejich obsah sumarizovat.

4.2.1.5. Klasifikace

Textová klasifikace je proces přidělování tagů a kategorií k textu na základě jeho obsahu.[265] Automatizovaná klasifikace textu je především pro archivní zpracování a obecně oběh eDokumentů v instituci velmi užitečným nástrojem, který efektivně zrychluje tento proces.

Standardní klasifikace obsahu se řídí manuálně nastavenou sadou lingvistických a v případě archivního zpracování také diplomatických pravidel. Cílem je ke každé sadě slov připojit tag anebo kategorii - tj. např. kněz, biskup, mnich = klérus. Po nastavení těchto pravidel algoritmus textový soubor projde a podle nich text klasifikuje.[266]

V současnosti se ale pro lepší výsledky a přesnější klasifikaci používá kombinace manuálně nastavených pravidel a strojového učení. Takovýto model se musí nejprve podle ukázkových dat, které jsou tvořeny již zpracovanými, klasifikovanými texty naučit. Kritická je přesnost těchto ukázkových tréninkových dat – pokud je chyba již v nich, model jí bude považovat za správnou klasifikaci.[267]

Metody klasifikace textu dolování v textu mají velký potenciál v archivním, předarchivním (tj. spisové služby) a knihovním zpracování.

Konkrétně při archivním zpracování je možné klasifikaci využít pro rozpoznávání typu dokumentu a jeho částí, a tudíž je alespoň částečně zpracovat automatizovaně. Příkladem může být vstup eDokumentů do instituce, kdy takovýto systém dokáže dokument automaticky klasifikovat a zařadit do určité kategorie (vysvědčení, žádost, list) a podle této klasifikace potom s nimi dále naložit například dle spisového a skartačního plánu. Takovýto systém by byl schopný

---

[265] What is Text Mining: Methods and Techniques. *MonkeyLearn* [online]. [cit. 2021-01-21]. Dostupné z: https://monkeylearn.com/text-mining/

[266] What is Text Mining: Methods and Techniques. *MonkeyLearn* [online]. [cit. 2021-01-21]. Dostupné z: https://monkeylearn.com/text-mining/

[267]What is Text Mining: Methods and Techniques. *MonkeyLearn* [online]. [cit. 2021-01-21]. Dostupné z: https://monkeylearn.com/text-mining/

rozpoznat některé další prvky textu – např. příjemce, odesílatele – a podle nich automatizovaně vyplnit alespoň část spisu. Jednalo by se především o velkou úsporu času – úředník nebo archivář již nebude muset každý spis založit a úplně vyplnit, archiválii již nebude muset kategorizovat. Toto má benefity nejen pro rychlost zpracování, ale také pro kriticky důležitou integritu eDokumentu a to ať už ve spisové službě nebo archivu.

4.2.1.6. Extrakce a sumarizace

Klasifikace dokumentu není jedinou velkou výhodou využití dolování v textu a jeho technik. Druhou zásadní technikou je sumarizace. Pomocí technik sumarizace je možné automatizovaně vytvořit krátký výtah z textu, tedy stručný regest. Sumarizace spoléhá na správnou klasifikaci textového souboru.

Stejně jako v případě klasifikace se dnes využívá především kombinace nastavených pravidel a tagů společně se strojovým učením a statistickými metodami, kdy je princip téměř totožný a platí stejná pravidla.

Nejběžnější způsoby pro sumarizaci textového souboru je využití běžných výrazů a podmíněných náhodných polí. Běžné výrazy definují sekvenci znaků tak, aby bylo možné je přiřadit k tagu. Každý takovýto vzorec je ekvivalentní k pravidlům, nastavených při textové klasifikaci.[268] Pomocí těchto pravidel program nadefinuje důležitost slov a jejich vztah k vedlejším slovům a z nich potom vytvoří sumarizaci.

Kondiční náhodná pole (CRF) jsou statistické metody, které lze použít při extrahování textu a vytváření sumarizací. Tento systém vytváří váhový systém známek, kdy každé slovní spojení dostane určitou váhu důležitosti, podle které se určí, zda se do výsledné sumarizace dostane či nikoliv. Nevýhodou CRF je jejich náročnost na výkon a složitější provedení vyžadující podrobnější znalosti této techniky.[269]

---

[268] What is Text Mining: Methods and Techniques. *MonkeyLearn* [online]. [cit. 2021-01-21]. Dostupné z: https://monkeylearn.com/text-mining/

[269] What is Text Mining: Methods and Techniques. *MonkeyLearn* [online]. [cit. 2021-01-21]. Dostupné z: https://monkeylearn.com/text-mining/

Sumarizace je z archivního hlediska důležitá při automatizovaném vytváření krátkých regestů, kdy především při digitalizačních projektech a jejich prezentaci je velmi časově náročný prvek právě psaní těchto krátkých regestů a jejich vyplňování k jednotlivým archiváliím, které jsou na webu nebo vnitřním organizačním systému dostupné v řádu stovek či tisíců. Automatizovaná sumarizace je tedy pro archiv velkou úsporou času v mnoha oblastech.

Mimo sumarizaci lze extrahovat také specifické části textu jako jsou jména, názvy budov, míst, správních jednotek aj. Patrná výhoda je tedy při prezentování a organizaci archiválií. Pro badatele je mnohem snazší do vyhledávače napsat například jméno osoby, které se mu zobrazí přímo v místě celého textu, kde se nachází[270] a to vše bez toho, aniž by samotné archiválie musel pročítat.

4.2.2. Využití

Obrovskou výhodou těchto technik je fakt, že se již hojně využívají v soukromém sektoru pro oblast marketingu, zpracování dokumentů a dalších relevantních odvětvích. Výzkum pro tuto konkrétní oblast je masivní a dobře financovaný. Je tedy z čeho těžit, archivnictví, informační věda a knihovnictví jej akorát musí použít pro řešení vlastních problémů.

Je potřeba si uvědomit, že samostatné metody a techniky jsou samy o sobě užitečné a přínosné. Jejich síla ale tkví v jejich možné kombinaci – fungují jako aktivátory pro další technologie a možnosti zpracování. Bez OCR a HTR by nebyly digitalizáty v textové podobě a bez textové podoby lze text mining využít jen velmi složitě. Bez základních procesů dolování v textu by nebylo možné provádět sumarizace, klasifikace a extrakci textu. Tyto vazby se ale nevztahují pouze k dolování v textových datech či technikám scraping, ale také k řadám dalších typů technologií, které se přesně nevztahují k textu.

Příkladem je využití dolování v textu pro akceleraci získání geoprostorových dat z množství textových archiválií archivních fondů a sbírek a jejich zasazení do mapy. Text mining je v tomto případě pouze akcelerátor celé práce. Obecně tyto

[270] A jestli se tam vůbec nachází → je pro badatele tato archiválie relevantní nebo ne?

moderní technologie přinášejí novou perspektivu do historické i archivní práce – buď práci zrychlují, něčím obohacují nebo vytvářejí úplně nové výstupy.

Pro akceleraci a podporu dalšího výzkumu, inovativních metod a technologií pro oblast zpracování a prezentace archivního a knihovního obsahu je zásadní prezentovat tato data[271] v textové podobě.

Text mining je využitelný a přínosný ve všech oblastech archivního zpracování – předarchivní péče (spisová služba), vstup archiválie do archivu a její zařazení, archivní popis, vytváření pomůcek, digitalizace, výzkumná činnost, prezentace archiválií. Celkově může přispět alespoň k částečné automatizaci a zrychlení některých procesů spojených s archivním a předarchivním zpracování a oběhu dokumentů v jakékoliv instituci, a to především v kombinaci s informačními systémy založenými na neurálních sítích a strojovém učení (NLP). Mimo toto využití lze text mining využít při jakékoliv rešeršní práci a pro testování vlastních hypotéz.

4.3. Natural Language Processing

NLP, tedy zpracování přirozeného jazyka a text mining, tedy dolování v textových datech jsou dvě velmi provázané disciplíny, které se často zaměňují. Obě se ale od sebe částečně odlišují. Zatímco text mining je subdisciplína dolování v datech, NLP je odvětví umělé inteligence, která se zabývá komunikací a zpracováním přirozeného jazyka tak, aby i přirozený jazyk byl počítačem zpracovatelný a bylo mu možné rozumět.[272]

Zásadní rozdíl mezi nimi je v tom, že text mining umožňuje extrahovat relevantní informace a data ze sbírky textových souborů. Tato data potom nějakým

---

[271] Pro jakýkoliv výzkum je zásadní mít k dispozici kvalitní data v adekvátním formátu a možnostem přístupu. Protože pokud tomu tak není a přístup k datům je mnohem složitější a časově náročnější → badatel musí sám dělat transkripce, převody do dalších formátů, doplnění metadat a další operace. Určitě by za stálo za to, vytvořit alespoň určitou selekci archivních sbírek a fondů reprezentující většinu archiválií v NAD a tu pro výzkumné potřeby plně digitalizovat a vytvořit jejich textovou podobu.

[272] TITENOK, Yaroslav. Natural Language Processing vs Text Mining. *Sloboda Studio* [online]. 2020 [cit. 2021-01-28]. Dostupné z: https://sloboda-studio.com/blog/natural-language-processing-vs-text-mining/

způsobem prezentuje, ale je již na uživateli jim porozumět. Text mining se stará tedy jen o to relevantní data z celého data setu extrahovat a prezentovat. Naopak NLP místo toho neslouží ke sběru a dolování v textových datech, ale za to umožňuje těmto již sesbíraným a relevantním informacím porozumět.[273] NLP je tedy celé o porozumění již sesbíraných relevantních informací a jejich seskupování do komplexnějších celků.

Stejně jako u dolování v textu platí, že oblast NLP se v posledních šesti letech velmi rozmohla a je velmi aktuální, což jde ruku v ruce s rozvojem umělé inteligence, strojového učení a neurálních sítí. Základní aplikace této technologie jsou v převodu textu na řeč, automatizovaných strojových offline i online překladačů[274], robotice, virtuální personální asistenti (Siri, Google, Alexa aj.) a mnoho dalších odvětví.

V oblasti humanitních věd má NLP své kořeny již v 50 letech minulého století, kdy poprvé bylo využito při vytváření počítačem čitelné kompilace děl Tomáše Akvinského – Index Thomisticus, tehdy ještě na děrných štítcích.[275] Od té doby došlo k výrazným posunům, ale zásadní průlomy přišly až s příchodem výrazného růstu výpočetního výkonu a pokrocích v oblasti umělé inteligence během posledních deseti let.

---

[273] TITENOK, Yaroslav. Natural Language Processing vs Text Mining. *Sloboda Studio* [online]. 2020 [cit. 2021-01-28]. Dostupné z: https://sloboda-studio.com/blog/natural-language-processing-vs-text-mining/

[274] Jedním z takových je např. DeepL (https://www.deepl.com/home).

[275] EGGERS, W. D., N. MALIK a M. GRACIE. Using AI to unleash the power of unstructured government data: Applications and examples of natural language processing (NLP) across government. *Deloitte Insights* [online]. US, 2020 [cit. 2021-01-28]. Dostupné z: https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/natural-language-processing-examples-in-government-data.html

4.3.1. Možnosti aplikace a využití

Základními metodami NLP je identifikace entit a vytváření tematických modelů. Entitu lze v tomto kontextu chápat stejně jako v datovém modelování (ERM), tj. entita[276] je záznam libovolného objektu, který je součástí reálného světa a tvoří základní prvek datového modelu.[277] Archivní NLP software by měl být schopen identifikovat řadu entit, které jsou pro archivní prostředí důležité – osoby, místa, organizace, budovy, předměty. Identifikací těchto předmětů dále napomáhá při dalším bádání[278], program si identifikací takovýchto entit utvoří „bližší obrázek" o obsahu archivního fondu či sbírky.

Druhou hlavní složkou NLP aplikace je vytváření tematických modelů. Výsledkem takovéhoto modelování by měl být datový tematický model ukazující vztahy mezi jednotlivými entitami (lze provést přirovnání k databázovému ER modelu) a jejich shlukování do témat. Tato témata mohou být např. noviny, psaní, rozruch, zloba, peníze aj. Všechna relevantní slova jsou takto roztříděna do tematických celků. Takovéto tematické celky lze nazvat entitní množiny. Každá takováto entita má vlastní atributy, což jsou jejich vlastnosti – např. osoba má jméno, bydliště, pohlaví a další. Vztahy těchto entit, jejich atributů a entitních množin lze potom zobrazit v grafickém uspořádání velmi obdobném ER modelování. Pro oblast archivního popisu právě za pomocí 3D modelu.

Výstupem je tedy grafický relační model archivního souboru nebo sbírky, který zobrazuje jednotlivé entity této sbírky nebo souboru (osoby, budovy, organizace, aj.) v podobě entitních množin. Každá entita má určité vlastnosti (osoba jméno, bydliště, povolání, aj.). Všechny tyto entitní množiny s entitami a jejich atributy

---

[276] Entitou může být cokoliv – osoba, místo, organizace, budova, produkt aj.

[277] NEHA, T. Difference Between Entity and Attribute in Database. *BinaryTerms* [online]. 2019 [cit. 2021-02-02]. Dostupné z: https://binaryterms.com/difference-between-entity-and-attribute-in-database.html

[278] MORGAN, Goodman. *"What is on this disk?" An Exploration of Natural Language Processing in Archival Appraisal* [online]. Chapel Hill, North Carolina, 2019 [cit. 2021-02-02]. Dostupné z: https://doi.org/10.17615/a13b-va69. Diplomová. School of Information and Library Science of the University of North Carolina. Vedoucí práce Christopher A. Lee.

mají mezi sebou určité vztahy, kdy například entita osoba bydlí v entitě budova a je součástí entity organizace.

V podstatě se jedná o odlišný pohled na způsob archivního popisu, kdy standardním výstupem bývá inventář, což je v podstatě logicky uspořádaný textový dokument popisující archivní soubor nebo sbírku. Takovýto textový dokument sice umožňuje orientaci v archivním sbírce, ale oproti grafickému ER modelu má několik nevýhod.

Největší nevýhodou je, že jej jednoduše nelze importovat do relační databáze. Z grafického entitního modelu lze pouhým importem a možnou úpravou jednoduše vytvořit relační databázi. Druhou obrovskou výhodou tohoto modelu je problematika sémantiky a počítačového zpracování. Inventář je ve výsledku dokument, který je vytvořen pouze pro potřeby člověka – je ve formě textového dokumentu a nemá dostatečně hlubokou sémantickou informací, aby jej dokázal počítač přečíst a vhodně zařadit. Grafický ER model, ze kterého lze vytvořit relační databázi tuto hlubší sémantickou informaci nese, mimo ni také je pro počítač lépe čitelný a zpracovatelný ze své podstaty rozdělení dokumentů do entit, entitních množin, atributů a jejich vztahů. Takto čitelný grafický ER model umožňuje potom mnohem jednodušší a rychlejší zpracování pomocí pokročilých statistických analýz, Big Data analýz a dalších možnostech zpracování. Grafický ER model je lehčeji zpracovatelný počítačem, než je inventář, což nese mnohé výhody. Mimo počítačové zpracování má grafický ER model výhodu také při bádání, kdy toto grafické uspořádání důležitých prvků může být pro badatele přínosné v tom, že právě tyto vztahy vidí přímo před sebou v grafické podobě.

Výhody grafického zobrazení jsou sice jasné, detailní inventární práci ale také v několika ohledech nedokážou plně nahradit, a tak by ve výsledku bylo nutné vytvořit určitý kompromis mezi standardní inventarizací a grafickým ER modelem.

Mimo tato základní archivní využití má NLP mnoho dalších aplikací v jednotlivých oblastech archivní vědy a praxe, které do sebe velmi často kombinují možnosti převodu obrazových digitalizátů do textové podoby, metody dolování v textu a právě NLP.

Pro vytváření těchto modelů existuje referenční model, který lze použít jako vzor pro vytváření vlastních modelů – CIDOC CRM[279]. Tento referenční model je volně dostupný s vlastní dokumentací. Jeho vývoj je podporován Mezinárodním koncilem muzeí (ICOM) a Mezinárodní komisí pro dokumentaci (CIDOC).

4.3.2. Nástroje

V podstatě platí to samé, co u dolování v textu – bez znalosti Pythonu nebo jiného populárního jazyka je práce s NLP poměrně obtížná. Vedoucí platformou a sadou nástrojů pro vytváření programů pro práci s NLP je NLTK. Tato sada nástrojů má také poměrně silnou uživatelskou a informační základnu. Nástroj má otevřený zdrojový kód, skrze webové prostředí zprostředkovává vlastní Wiki, FAQ, diskusní fórum a vlastní příručku[280] – *Natural Language Processing with Python*, která provádí uživatele základními funkcemi obecně programování v Pythonu a s jednotlivými částmi nástrojů NLTK (práci s korpusem, analyzování lingvistických struktur aj.).[281]

Mimo tento nástroj existují ještě dvě poměrně populární řešení. Druhým volně dostupným řešením s otevřeným zdrojovým kódem je OpenNLP od společnosti Apache. OpenNLP je poměrně podobným nástrojem jako NLTK, podporuje všechny základní operace spojené s NLP výše zmíněné. Poskytovaný informační servis je obdobný[282] s NLTK, a tak záleží spíše na preferenci uživatele.

Posledním velmi populárním řešením je SpaCy NLP, které se používá spíše v podnikovém NLP, ale i pro oblast humanitních věd má svá využití. Ze všech tří nástrojů je nejrychlejší, ale zato postrádá takto kvalitní informační servis a některé funkce NLTK. Mimo tyto nedostatky je ale snazší na naučení, protože výsledný

---

[279] http://www.cidoc-crm.org/

[280] Příručka je volně dostupná na následujícím odkazu: http://www.nltk.org/book/

[281] Natural Language Toolkit. *NLTK* [online]. 2020 [cit. 2021-01-28]. Dostupné z: https://www.nltk.org/

[282] Znovu poskytuje příručku, dostupná na: https://opennlp.apache.org/docs/1.9.3/manual/opennlp.html

kód je čistší a jednodušší[283]. V současnosti je také rychleji vyvíjeno, a to především oproti NLTK, jehož vývoj je zpravidla pomalejší.

4.3.3. Problémy

NLP trpí řadou běžných problémů typických pro aplikaci moderních informačních technologií v oblasti humanitních věd – tj. náročnost na výpočetní výkon, lidské zdroje, finance, nedostatečný výzkum, roztříštěnost, nejednotnost nástrojů a mnoho dalších popsaných v kapitolách výše.

Mimo tyto obecné problémy lze vysledovat i některé specifičtější. Prvním z nich je nedostupnost dat v textové podobě, která je pro jakoukoliv práci s NLP či dolováním v textu jednoduše nutná.

Druhým zásadním problémem je fakt, že současné nástroje jsou uzpůsobené spíše na práci se strukturovanými dokumenty a daty dostupnými na internetu, anebo v oběhu jednotlivých institucí či podniků. Tato data jsou více homogenní a využívají současnou moderní formu jednotlivých jazyků a jelikož je současné NLP založeno na použití neurální sítě, která potřebuje dostatečně obsáhlý jazykový korpus a další tréninková data pro dostatečně přesné výsledky, nastává problém s formou jazyka. Současná čeština, němčina a angličtina jsou poměrně odlišné jazyky než ty, které se používali v minulosti. Kvůli tomuto problému odlišnosti je nutné vytvářet vlastní data sety určené pro trénování těchto sítí. Mimo

Současná data jsou také více homogenní, a tak jsou nástroje přímo uzpůsobené na ně, což je přímo v kontrastu projektů v oblasti humanitních věd, které jsou většinou specifické na určitý typ archiválií se specifickým písmem a jazykovou formou.[284] K umocnění tohoto problému je třeba ještě zmínit fakt, že vlastně neexistuje žádná standardizace postupů, metodik a nástrojů.

---

[283] Industrial-Strength Natural Language Processing. *SpaCy* [online]. 2021 [cit. 2021-01-28]. Dostupné z: https://spacy.io/

[284] MCGILLIVRAY, Barbara, Thierry POIBEAU a Pablo Ruiz FABO. *Digital Humanities and Natural Language Processing: "Je t'aime... Moi non plus"* [online]. The Alliance of Digital Humanities Organizations, 2021, 2(14) [cit. 2021-01-28]. Dostupné z: http://www.digitalhumanities.org/dhq/vol/14/2/000454/000454.html

5. Reflexe a výstupy

Obecně známou skutečností je, že současný svět je stále více digitální, a od původních analogových vazeb se stále více upouští. Příkladem pro české prostředí je nastávající platná digitální ústava.

S vývojem IT roste objem obecně eDokumentů, které musí instituce zpracovat. Zde lze vysledovat dva zdroje tohoto růstu. První je nekonečně dlouhý růst množství dat, které instituce a podniky produkují. Tento růst se odráží mimo jiné ve vědeckém prostředí, státní správě, a tudíž také paměťových institucí. Druhým faktorem pro tento růst objemu jsou digitalizační projekty. V porovnání s předchozím bodem je jejich množství zanedbatelné, v prostředí archivů a knihoven ale velmi relevantní, protože z obou těchto zdrojů plynnou data právě do nich. Zvýšený objem dat je sám o sobě poměrně složitý problém, který přináší řadu otázek, které je třeba zodpovědět.

První z nich je vůbec otázka množství dat, a tudíž jejich výběru. S rostoucím objemem dat roste množství vynaložené práce na proces výběru těch relevantních. Toto přináší vyšší pracovní zatížení, která je časově náročnější. Zde je právě nasnadě alespoň částečná automatizace tohoto procesu pomocí neurálních sítí s technologiemi NLP.

Druhou otázkou je vůbec samotný výběr. Jak se v datech orientovat? Určitě je nasnadě výkonnější organizace a uspořádání dat už u samotného zdroje, tu je ale problematické[285] prosadit a vyhláškou či zákonem vynutit. Zde nastupuje právě data scraping a dolování v textových datech, kdy by archivář byl schopen se v masech dat sám orientovat a vytěžit z nich jen to podstatné, co je potřeba zachovat, anebo poskytnout badateli v rámci rešeršní práce a badatelských dotazů.

Třetím problémem je jejich efektivní organizace do výkonných logických organizační struktur. Tato překážka je poměrně efektivně řešena za pomocí databází a informačních systémů postavených na OAIS. Organizaci je ale možné zefektivnit za pomocí grafických datových modelů, zmíněných v podkapitole NLP.

---

[285] Stačí zmínit i v současném stavu problémy s vytvářením korektních SIP balíčků, vedení spisové služby v elektronické podobě aj.

Čtvrtou otázkou je jejich prezentace. Jak efektivně tato data prezentovat veřejnosti? Standardní informace dnes dostupné badateli jsou především informace o zpracování archiválie, místo uložení, základní metadatový popis, popř. další pomůcky, ne už tak často bývá dostupný také regest a obrazový digitalizát. Tento repertoár informací lze ale výrazně rozšířit a posunout prezentování archiválií mnohem dál, kdy nové možnosti prezentace jsou především vyhledávání v textu (např. jmen, institucí, spolků, aj.) díky aplikaci technologií HTR a moderních OCR. Poskytnout výše zmíněný grafický datový model pro zobrazení vztahů přináší také určité obohacení obsahu, protože právě ty vztahy institucí, osob, spolků jsou pro badatele zásadní.

5.1. Překážky

Při jakékoliv nově zaváděné technologii lze vysledovat některé společné překážky a problémy, které se objevují téměř pokaždé, když se takováto (pro obor) nová technologie začíná zkoumat a pomalu zavádět. Zmíněné technologie tomu nejsou výjimkou. Mimo tyto obecné problémy a překážky existuje ale řada takových, které jsou specifické buď pro oblast paměťových institucí a jejich výzkumných projektů, nebo pro technologie samotné.

Jelikož použití těchto technologií zatím nemá v našem prostředí žádnou jednotnou metodiku, názvosloví a patřičnou základní literaturu, většina výzkumné a projektové činnosti začíná buď od začátku, nebo se inspiruje podobnými pracemi. Takový výzkumník, anebo člověk, který tu či onu technologii chce vlastně využít nemá tedy na čem moc stavět[286] kromě právě těchto jednotlivých vědeckých pracích. Zde je potřeba říct také fakt, že většina technologií je dobře popsaná a jejich samotná informační základna je poměrně silná. Problém je, že jsou v pouze v kontextu technologií samotných, nikoliv však v kontextu s komplexní problematikou paměťových institucí, které vyžadují vlastní know-how a specifický přístup.

[286] Důkazem toho je poměr informačních zdrojů v této diplomové práci, kdy drtivá většina z nich je právě z vědeckých článků. Menší množství zdrojů se potom týká přímo té dané technologie samotné, ne však v kontextu Digital Humanities.

Druhou a tou největší překážkou je nedostupnost archiválií v digitální textové podobě a vůbec samotná nejednotnost digitalizace, kdy každý jednotlivý archiv má vlastní postupy a neexistuje žádná jednotná iniciativa. Tato roztříštěnost je jak pro badatele, tak pro samotné archivy nevýhodná – webové stránky, jako hlavní zdroj prezentace archiválií, jsou nesourodé, každý archiv má vlastní řešení, což je náročné na finance nebo na techniku. S tím, jak to funguje v současnosti si archivy buď vydržují vlastního IT pracovníka s vlastními technickými prostředky, nebo používají cloudová řešení, což je samozřejmě náročné na finance. Jednodušší možností by bylo vytvoření jednotného webového portálu (obdoba read&search, který používá Transkribus) k prezentaci, ke kterému by měly přístup všechny archivy, které by tak nemusely mít vlastní úložiště. Toto přináší výhody i v jednotnosti vstupních archiválií a udržování jednotného standardu digitalizátů. Pro badatele je samozřejmě také jednodušší navštěvovat pouze jeden portál. Finanční náklady by tak byly řešeny pouze na té vyšší, centrální[287] úrovni. Otázkou ale zůstává, zda takováto úroveň centralizace je pro české prostředí realizovatelná, či nikoliv.

Když už jsou archiválie dostupné skrze webové prostředí v digitální podobě, jedná se pouze o statický snímek, který je pro potřeby dalšího strojového zpracování nedostatečný. Především proto je možný výzkum a aplikace technologií časově náročná a pomalá. Velkou část takovéto práce a zdrojů totiž spolkne právě jen samotná transkripce.

Pro archivní vědu, a hlavně pro určitou akceleraci výzkumu a vůbec zvýšení zájmu o tuto problematiku, by bylo poměrně velkým benefitem, kdyby byl vytvořen a zpřístupněn výběr reprezentativních archivních souborů, dostupných online formou, plně popsán, s textovou podobou standardního obrazového digitalizátu. Tato ukázková sbírka vyčleněných archivních souborů či sbírek by potom sloužila jako základní vzorek pro srovnávání jednotlivých technologií, zkoušení jejich efektivity a užitečnosti, a to jak pro oblast zpřístupňování archiválií (HTR, OCR),

[287] Což má výhodu snížení byrokracie a snížení počtu problémů spojených s přerozdělováním financí.

tak pro oblast jejich dalšího strojového zpracování (HTR, text mining). Takovýto připravený reprezentativní vzorek[288] by byl efektivní při testování vlastních hypotéz při projektové a vědecké činnosti. Mimo to také pro studenty, kteří by si mohli prakticky zmíněné nástroje vyzkoušet. Jednoduše jde o to, poskytnout vědcům, badatelům, studentům a zájemcům přívětivější vědecké prostředí s připravenými, relevantními daty.

Toto už v omezené formě funguje například v Anglii, kde právě rukopisy z provenience Jeremyho Benthama z tzv. Bentham Project fungují jako takovýto vzorek. Tento, poměrně jednotvárný fond, je ale psán pouze jedním písmem, omezeným počtem písařů, což není dostatečně reprezentativním vzorkem pro šíři archiválií dostupných v České republice. Mimo to je také samozřejmě v angličtině, která v porovnání s češtinou má odlišné nuance a překážky.

Poslední překážkou je potom problematika vůbec lidí, kteří by se těmito technologiemi a jejich aplikováním v tomto specifickém prostředí paměťových institucí chtěli zabývat. Mimo standardní archivní vzdělání je totiž potřeba také znalost a orientace obecně v IT, konkrétněji potom v oblasti datové vědy, jako je statistika, databáze, Python. V současné době se u nás obory archivnictví teprve pomalu začínají uzpůsobovat těmto novým požadavkům. Většinou se těmto technologiím věnují především ze své vlastní iniciativy, nebo z iniciativy vyučujících, či vedoucích prací, nikoliv však v rámci sylabů[289].

Zde se nabízí několik variant možných řešení, kdy navrhnout lze některé následující. První variantou je provést úpravy v oborech archivnictví a PVH vyučovaných na univerzitách, a tyto základní okruhy do nich implementovat. Problémem ovšem je, že při implementaci těchto dalších znalostí do současných studijních programů, by mohl takovýto obor být příliš složitý. Tato varianta se zdá jako nereálná.

---

[288] Samozřejmě čím větší by tento vzorek byl tím lepší.

[289] Výjimkou v této oblasti je např. nově akreditovaný obor Digitální historické vědy na FF UHK.

Druhou možností je vytvoření nového oboru, založeného na základech archivnictví a informatiky. Z archivnictví by byl kladen důraz na novověk – tj. historie, správa, paleografie a diplomatika jako základní znalosti, bez kterých se archivář neobejde. Naopak středověk by byl méně dotován, tj méně znalostí středověké patrimoniální správy, historie a těch PVH, které se středověkem úzce souvisí, jako je např. heraldika a středověká paleografie. Student by pak měl na výběr, zda chce studovat standardní archivnictví v současné formě, či toto více technicky zaměřené archivnictví s důrazem na novověk.

Tato varianta by nemusela platit po celou dobu bakalářského a magisterského studia, ale mohla by být rozdělena až v specializovanějším magisterském studiu – tj. jednotné bakalářské studium a specializované magisterské s těmito dvěma odvětvími. Určitou inspirací v tomto případě můžou být studijní programy zemí jako je Estonsko nebo Spojené státy americké (MLIS). Zde je potom riziko toho, že takovýto student nebude mít dostatečné znalosti z obou oblastí.

Třetí variantou je vytvoření zcela nového oboru od archivnictví odlišného. Tento nový obor by se nezabýval pouze archivnictvím, ale prací s informačními technologiemi v oblasti paměťových institucí – tj. na pomezí informační a knihovní vědy, datové vědy, informatiky, archivnictví a muzejnictví. V tomto oboru by byl, ze všech zmíněných, kladen nejvyšší důraz na práci s digitálními daty, ať už muzejního, knihovního nebo archivního charakteru, tj. proces úplné komplexní digitalizace, uchovávání a dlouhodobá ochrana dat, jejich prezentace, možnosti zpracování a organizace. Základ by měly tvořit digitalizační techniky, webové technologie, Python, SQL jazyk, statistika a další prvky z datové i archivní vědy.

Z uvedeného lze vyvodit, že znalosti procesu digitalizace by měly být základní součástí každého studijního programu budoucího archiváře v jakékoliv jeho formě. Kritická je v tomto případě provázanost reálného využití těchto technologií a teoretické výuky. Například v současnosti se s nástrojem Transkribus musí každý naučit sám, z vlastní iniciativy. Podklady k tomu sice jsou, ale jeho výuka se určitě dá zakomponovat do standardních předmětů zabývajících se digitalizací, protože Transkribus vlastně celý základní proces digitalizace kopíruje (viz. kapitola Transkribus). Určitá představa by tedy mohla být, že student či skupina by na

předmětech, kde je transkripce standardně využívána, tedy PVH, digitalizace nebo ediční činnost, měl za úkol vypracovat jako semestrální práci hotový data set.

Pro vytvoření takovéhoto data setu je potřeba umět text manuálně transkribovat, poté dokumenty převést do digitální podoby pomocí skeneru nebo fotoaparátu[290], provést některé základní úpravy v grafickém editoru, dokument popsat a doplnit o metadata. Takto vypracovaný data set potom použít při automatizované transkripci v programu Transkribus. Výsledkem by byl kompletně zdigitalizovaný archivní fond nebo sbírka nebo její části, která by navíc měla přímý užitek pro daný archiv, odkud by archiválie pocházeli, což je také obrovská výhoda. Něco takového samozřejmě nezvládne student prvního nebo druhého ročníku, zkušenější studenti[291] by ale tyto základní znalosti měli mít.

Student by si tak ve výsledku vyzkoušel paleografii, ediční činnost, práci s fotoaparátem či skenerem, grafickým editorem a Transkribem. Vlastně by prošel kompletním procesem digitalizace s použitím moderních technologií.

5.2. Benefity

Aplikace těchto moderních technologií přináší především nové typy výstupů při analýze archiválií, které nabízejí jiný úhel pohledu na věc a slouží tak jako rozšíření standardních prací. Mimo toto obohacení přinášejí také řešení mnohých současných problémů výše zmíněných. Ve výsledku by měly být alespoň něčím, co doplňuje současné výstupy práce s archiváliemi, a zároveň tuto práci především z časového hlediska usnadňují a činí jí tak efektivnější.

Pro oblast zpřístupňování archiválií, tj. HTR, OCR, Transkribus, se jedná především o získávání textové podoby z digitálního obrazu archiválie. Tato textová podoba je benefiční především pro dvě věci. První je zpřístupnění archiválie širšímu okruhu lidí, kteří tak nemusí umět číst dané písmo a nemusí se tak orientovat v paleografii. Výsledkem by mohl být například web, který umožňuje vyhledávání

[290] Nebo jako v případě ScanTent u Transkribu v obdobné konstrukci pomocí chytrého telefonu, což by nebylo tak finančně náročné.

[291] Což se přímo hodí, protože síla Transkribu je v jeho uživatelské základně. Množství takto digitalizovaných textů by bylo nezanedbatelné se stále novými a novými studenty.

všech slov v archiváliích, např. osoby, budovy, místa, spolky, aj., což je užitečné ve všech oblastech zájmu[292] badatelů, ať profesionálních historiků, tak dalších zájemců z řad veřejnosti. To je tedy otázka prezentace archiválií. Tou druhou věcí je potom fakt, že textová podoba digitalizátu funguje jako aktivátor pro další možnosti zpracování – tj. pokročilejší možnosti vyhledávání a filtrování archiválií a jejich obsahu, možnosti využití technologií NLP, text mining, scraping, vytváření grafických datových modelů a mnoho dalších. Což lze shrnout jako čitelnost pro počítač.

Na celou věc se tedy lze dívat z pohledu přístupnosti – nejprve pro člověka, tj. archiválii si dokáže přečíst v současném jazyce a písmu, může ji, i informace v ní obsažené, vyhledávat a potom pro počítač, tj. textová podoba je zpracovatelná dále počítačem a technikami výše zmíněnými.

No a pro oblast zpracování počítačem čitelných archiválií se jedná o přínosy z hlediska zefektivnění práce s archiváliemi a vytváření nových typů výstupů. Znovu zde lze vytyčit dva zásadní přínosy těchto technologií – prvním z nich je urychlení procesů práce s archiváliemi, a to jak z pohledu badatelů, tak archivní instituce, např. automatizovaná sumarizace a kategorizace obsahu. Druhým zásadním přínosem je potom efektivnější možnosti vyhledávání a těžení relevantních dat z velkého množství dostupných dokumentů částečně po vzoru Big Data analýz, tedy např. vyhledávání v textu, grafické datové modely, nové možnosti NLP aj.

Toto začne být stále více důležité postupem času, kdy, již nyní je jasné, že přechod do plně digitálního prostředí je nevyhnutelný, a to ať z hlediska tzv. digitální ústavy pro předarchivní péči a spisovou službu, tak např. v současnosti probíhající pandemie. S tímto postupným přesunem do digitálního prostředí, který začal na začátku nového tisíciletí přichází právě mnohonásobně vyšší objem dat, se kterými je nutno pracovat. Proto je také určitá, alespoň částečná, automatizace některých procesů spojených s oběhem dokumentů v institucích (včetně archivu)

[292] Příkladem může být genealogický výzkum, kdy si badatel jednoduše zadá hledané slovo, které se mu zobrazí přímo v dokumentu, a když nezobrazí – tak to je sama o sobě užitečná informace. Již by nemusel procházet jeden dokument za druhým ručně.

důležitá. Do blízké budoucnosti je možné predikovat vědecké úsilí a pokusy v oblasti integrování neurálních sítí do jednotlivých sfér státní správy

Jak bývá pravidlem, toto všechno určitou dobu potrvá, než se to projeví i v našem prostředí. Tato výsledná doba ale nemusí být tak dlouhá, jak se může na první pohled zdát, protože EU v tomto směru funguje jako určitý akcelerátor – lze zmínit např. zavedení GDPR, eIDAS nebo normy NIS ve formě Zákony o kybernetické bezpečnosti.

Samozřejmě jak se všemi novými věcmi v tomto oboru je nutné počítat s určitou střídmostí. Žádná technologie není ale dokonalá a nedokáže plně nahradit soustavnou detailní práci profesionála, může mu ale v jeho práci výrazně pomoci, a to především v manuálních, repetitivních a časově náročných operacích.

Závěr

Platí, že eDokumenty a digitální data obecně přinášejí své vlastní specifické komplexní problémy, se kterými se archivnictví potýká už dvě desetiletí. Takováto data ale mají mnoho specifických výhod, které jednoduše jejich analogové protějšky nemají. Prozatím z nich ale archivnictví naplno nečerpá. Proti novým informačním technologiím stále panuje určitá rezervovanost z pohledu archivářů, za kterou ale nese zodpovědnost především nesprávná implementace a vůbec zatím nedostatečný posun v těchto technologiích, které často nepřinesly požadovaný výsledek. Posun ve vývoji a výkonnosti těchto technologií je ale dnes již mnohem dál, a tak si jistě zaslouží další pokus. Zatím ale stále platí, že implementace těchto technologií je prozatím u nás stále v počátcích, posun k stále více digitálnímu prostředí je ale stále poměrně intenzivní, a to především z iniciativy EU (eIDAS, NIS, GDPR, aj.).

Tato práce si kladla za cíl poskytnout relativně podrobný přehled o současném vědeckém stavu věci a možnostech využití moderních IT k automatizované transkripci rukopisných a tištěných dokumentů, jejich nástrojů a jejich následnému pokročilému zpracování a uspořádání. Snažila se poukázat na přínosy využití právě těchto obohacujících technologií a na jejich nezanedbatelné výhody, které do archivní vědy přinášejí.

V první části byla provedena poměrně rozsáhlá rešeršní práce zabývající se současným stavem archivní vědy jako takové u nás v porovnání se západními zeměmi, dále pak vědeckou sférou v oblasti OCR a HTR technologií a jejich jednotlivých nástrojů jako je Transkribus, Quartex, Tesseract, Finereader a jiné. Navazující rešeršní část se zabývala možnostmi využití takto získaných textových digitalizátů v podobě využití dolování v datech, NLP, statistických nástrojů a metodami scraping.

Druhá část se věnovala vytvoření podrobného přehledu o technologiích umožňující právě automatizovanou transkripci rukopisných a tištěných archiválií v podobě technologií HTR, OCR a jejich jednotlivých nástrojů. Samostatná kapitola je potom věnována nástroji Transkribus, který je tím zatím nejzásadnějším

nástrojem v kontextu těchto nových technologií, navíc podporovaný EU. Je popsán jeho kompletní workflow a funkce.

Třetí část se zabývala technologiemi a nástroji, které těží z dostupných textových dat jako jsou Scrapy, Selenium, Textract, Google Ngram viewer, NLTK a další. U každé technologie a nástroje je dostupný jejich základní popis funkcí, využití v kontextu archivního prostředí a jejich výhody a nevýhody, případně možnosti další aplikace.

Následující část se věnovala shrnutí a konečné syntéze všech popsaných technologií a témat v kontextu s Digital Humanities a reflexi nad dvěma na sebe velmi závislými oblastmi, jak zpřístupňování rukopisných a tištěných archiválií pomocí automatizované transkripce skrze řadu vybraných nástrojů, tak jejich následné využití pomocí moderních IT, které přináší odpovědi na řešení některých zásadních otázek současných (nebo i budoucích) problémů.

Mimo to se snažila pomoci zodpovědět dvě v úvodu zmíněné otázky, tedy *„Jak zpřístupnit řadu prozatím nedostupných archiválií veřejnosti a jak efektivněji zpřístupnit ty již nyní dostupné?“*

Odpověď v tomto případě vidí právě ve využití zmíněných moderních technologií (popisovaných v kapitole 2, 3) jako je Transkribus, což by společně s možnostmi NLP a strojového učení mohlo velmi napomoci tento problém vyřešit. Samozřejmě společně s organizovaným úsilím a korektním využití těchto technologií.

A druhou otázkou: *„Jak reagovat na zvětšující se objem dokumentů, se kterými archivnictví musí pracovat a jak tato data využít?“*

V tomto případě je odpovědí znovu využití moderních metod a nástrojů přejatých z datové vědy jako jsou dolování v datech, NLP a další nástroje popisované v kapitole 4. Zásadní je potom především zavedení alespoň částečné automatizace některých procesů spojených s archivním zpracováním a životním cyklem dokumentu. Hlavními prvky by v tomto případě měli být automatizovaná sumarizace a kategorizace obsahu společně s vlastním informačním systémem postaveným na neurálních sítích.

Využití těchto dat je potom otázkou, kterou se již několik let datová věda věnuje v podobě tzv. Big Data analýz, takže informace a nástroje je odkud čerpat. Zásadní jsou v tomto případě takové nástroje, které umožňují získávat relevantní informace z velkého objemu dat, tj. scraping a text mining metody a nástroje.

Technologie sami o sobě jsou ale nedostačující, zásadní je jejich správná, organizovaná implementace a jejich metodické využití, která tu zatím není. Individuální projekty a práce jsou povětšinou menšího měřítka a obecnější dopad tak nemají. Zásadní je vůbec položit těmto technologiím základ v našem oboru a vytvořit tak určité zázemí.

Další úvahy jsou poměrně složité, protože se jedná o poměrně novou problematiku týkající se zhruba posledních pěti let. Do budoucna by bylo určitě zajímavé zabývat se těmito technologiemi konkrétněji a každé zvlášť s určitými praktickými výstupy, například naprogramovat v Pythonu bota, který by byl schopný rozpoznávat typy dokumentů a vytvářet jejich vlastní krátké obsahy.

Mimo tento příklad by bylo velmi přínosné v budoucnu zanalyzovat konkrétní dopady těchto technologií a vůbec rozšířenost jejich využití – jakých výsledků bylo dosaženo, kde zklamaly, kde naopak přinesly nové možnosti využití nebo vyřešily nějaký dlouhotrvající problém.

Literatura a informační zdroje

A set of benchmarks for Handwritten Text Recognition on historical documents. *Pattern Recognition* [online]. 2019, (94), 133 [cit. 2020-09-19]. Dostupné z: https://www.sciencedirect.com/science/article/abs/pii/S0031320319302006

ABBYY Finereader 14: Uživatelská příručka. *Natur.cuni.cz* [online]. 2017 [cit. 2020-10-17]. Dostupné z: https://www.natur.cuni.cz/fakulta/cit/navody/soubory/abbyy-fine-reader-uzivatelska-prirucka

ABBYY FineReader PDF 15 – free trial. *ABBYY Finereader PDF* [online]. 2020 [cit. 2020-10-17]. Dostupné z: https://pdf.abbyy.com/lp/finereader15-download-free-trial/

ABBYY FineReader PDF 15 for Windows: the smarter PDF solution. *ABBYY Finereader PDF* [online]. 2020 [cit. 2020-10-17]. Dostupné z: https://pdf.abbyy.com/

About us. *Impact digitisation.eu* [online]. 2020 [cit. 2020-09-19]. Dostupné z: https://www.digitisation.eu/about/

About: What is NewsEye about? *NewsEye* [online]. 2020 [cit. 2020-09-19]. Dostupné z: https://www.newseye.eu/about/

ACHARY, S. Unleashing the Kraken for OCR. *Analytics Vidhya* [online]. 2020 [cit. 2020-10-15]. Dostupné z: https://medium.com/analytics-vidhya/unleashing-the-kraken-for-ocr-fba6bff73c8c

Amazon Textract. *Aws Amazon* [online]. 2020 [cit. 2020-10-20]. Dostupné z: https://aws.amazon.com/textract/

Archives and Records Management MA. *UCL* [online]. London, 2020 [cit. 2020-09-10]. Dostupné z: https://www.ucl.ac.uk/prospective-students/graduate/taught-degrees/archives-records-management-ma

Big Data: What is Web Scraping and how to use it. *Towards data science* [online]. 2018 [cit. 2020-12-17]. Dostupné z: https://towardsdatascience.com/big-data-what-is-web-scraping-and-how-to-use-it-74e7e8b58fd6

BitCurator NLP. *BitCurator* [online]. 2018 [cit. 2020-09-24]. Dostupné z: https://bitcurator.net/bitcurator-nlp/

BLANKE, T., M. BRYANT a M. HEDGES. Ocropodium: Open source OCR for small-scale historical archives. *Journal of Information Science* [online]. 2012, (38), 76-86 [cit. 2020-11-06]. Dostupné z: https://www.researchgate.net/publication/254115552_Ocropodium_Open_source_OCR_for_small-scale_historical_archives

Can AI-powered OCR really read handwriting better than a human? Handwriting OCR. *SS&C* [online]. 2020 [cit. 2020-09-17]. Dostupné z: https://vidado.ai/ocr-poor-quality-docs/

Colonial America: Complete CO5 files from The National Archives, UK, 1606-1822. *Adam Matthew A SAGE Publishing Company* [online]. 2020 [cit. 2020-10-13]. Dostupné z: https://www.amdigital.co.uk/primary-sources/colonial-america

Commercial Overview TranSkriptorium. *Transkriptorium* [online]. 2020 [cit. 2020-11-10]. Dostupné z: http://www.transkriptorium.com/user/images/transkriptorium/ts-presentation.pdf

Conventions for special characters. *READ-COOP* [online]. 2020 [cit. 2020-11-17]. Dostupné z: https://readcoop.eu/transkribus/howto/transkribus-transcription-conventions/

CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. Praha: Národní knihovna České republiky, 2010. ISBN 978-80-7050-588-5.

Data Scraping. *Techopedia* [online]. 2020 [cit. 2020-12-15]. Dostupné z: https://www.techopedia.com/definition/33132/data-scraping

Download and Installation. *READ-COOP* [online]. 2020 [cit. 2020-11-17]. Dostupné z: https://readcoop.eu/transkribus/wiki/download-and-installation/

EDDINS, Steve. Adaptive thresholding for binarization. *MathWorks* [online]. 2016 [cit. 2020-10-13]. Dostupné z: https://blogs.mathworks.com/steve/2016/07/25/adaptive-thresholding-for-binarization/

EGGERS, W. D., N. MALIK a M. GRACIE. Using AI to unleash the power of unstructured government data: Applications and examples of natural language processing (NLP) across government. *Deloitte Insights* [online]. US, 2020 [cit. 2021-01-28]. Dostupné z: https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/natural-language-processing-examples-in-government-data.html

EPADD, Stanford University. *Digital Preservation Coalition* [online]. 2020 [cit. 2020-09-24]. Dostupné z: https://www.dpconline.org/events/digital-preservation-awards/epadd-stanford-university

EPADD: About. *Stanford LIBRARIES* [online]. Stanford, 2020 [cit. 2020-09-24]. Dostupné z: https://library.stanford.edu/projects/epadd/about

Extracting text from an image using Ocropus. *Danvk.org* [online]. 2015 [cit. 2020-10-13]. Dostupné z: https://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html

Fraktura. *Encyklopedieknihy.cz* [online]. 2020 [cit. 2020-10-17]. Dostupné z: https://www.encyklopedieknihy.cz/index.php/Fraktura

Goudarouli, E., Sexton, A. & Sheridan, J. The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK. Philos. Technology. 32, 173–183 (2019). https://doi.org/10.1007/s13347-018-0333-3

Handwritten Text Recognition Workflow: Basic Workflow. *Transkribus* [online]. 2018 [cit. 2020-09-17]. Dostupné z: https://transkribus.eu/wiki/index.php/Handwritten_Text_Recognition_Workflow

Handwritten Text Recognition Workflow: Prerequisites. *READ-COOP* [online]. 2020 [cit. 2020-11-17]. Dostupné z: https://readcoop.eu/transkribus/wiki/handwritten-text-recognition-workflow/

Hard Limits in Amazon Textract. *Aws Amazon* [online]. 2020 [cit. 2020-10-20]. Dostupné z: https://docs.aws.amazon.com/textract/latest/dg/limits.html

Head of Platform Services at Adam Matthew Digital interviewed in The Charleston Advisor. *Quartex* [online]. 2018 [cit. 2020-09-17]. Dostupné z: https://www.quartexcollections.com/news/item/head-of-platform-services-at-adam-matthew-talks-about-quartex

HIMANIBEN, Patel. *Archival Document Processing using Cognitive Computing* [online]. Carolina, USA, 2019 [cit. 2020-10-20]. Dostupné z: https://thescholarship.ecu.edu/handle/10342/7489?show=full. Diplomová. East Carolina University. Vedoucí práce Tabrizi M. H. N.

How To Search Documents with the Keyword Spotting Feature. *READ COOP* [online]. 2020 [cit. 2020-11-10]. Dostupné z: https://readcoop.eu/transkribus/howto/how-to-use-keyword-spotting/

HUI, Rui, Carlos PALLAN, Jean-Marc ODOBEZ a Daniel GATICA-PEREZ. Analyzing and visualizing ancient Maya hieroglyphics using shape: From computer vision to Digital Humanities. *Digital Scholarship in the Humanities* [online]. 2017, 2(32), 179-194 [cit. 2020-09-22]. Dostupné z: https://www.researchgate.net/publication/322633817_Analyzing_and_visualizing_ancient_Maya_hieroglyphics_using_shape_From_computer_vision_to_Digital_Humanities

HUTCHINSON, Tim. Natural language processing and machine learning as practical toolsets for archival processing. *Record Management Journal* [online]. 2020, 2(30), 155-174 [cit. 2020-09-24]. ISSN 0956-5698. Dostupné z: https://www.emerald.com/insight/content/doi/10.1108/RMJ-09-2019-0055/full/html

ICDAR2019: Program Booklet. *ICDAR 2019* [online]. Sydney, 2019 [cit. 2020-09-19]. Dostupné z: http://icdar2019.org/program-booklet/

ICFHR: International Conference on Frontiers in Handwriting Recognition. *WikiCFP* [online]. 2020 [cit. 2020-09-19]. Dostupné z: http://www.wikicfp.com/cfp/program?id=1366&f=International%20Conference%20on%20Frontiers%20in%20Handwriting%20Recognition

Image Binarization (1) : Introduction. *Craft of Coding* [online]. 2017 [cit. 2020-10-13]. Dostupné z: https://craftofcoding.wordpress.com/2017/02/13/image-binarization-1-introduction/

Industrial-Strength Natural Language Processing. *SpaCy* [online]. 2021 [cit. 2021-01-28]. Dostupné z: https://spacy.io/

KATUŠČÁK, Dušan. Digital Humanities a automatická transkripcia rukopisných textov. *ITlib* [online]. Ministerstvo školstva, vedy, výskumu a športu Slovenskej republiky, 2020, 2020(1) [cit. 2020-09-19]. Dostupné z: https://itlib.cvtisr.sk/archiv/2020/1/digital-humanities-a-automaticka-transkripcia-rukopisnych-textov-digital-humanities-and-automatic-transcription-of-handwritten-texts.html?page_id=3698

KLEPPE, M., M. LINCOLN a T. SMITS. *Computer Vision in Digital Humanities. DH* [online]. 2017 [cit. 2020-09-22]. Dostupné z: https://www.semanticscholar.org/paper/Computer-Vision-in-Digital-Humanities-Kleppe-Lincoln/07b7bbe8dd59e3bd2b63babdcd43b76e019ede7d?p2df

Koncepce rozvoje archivnictví. Ministerstvo vnitra ČR [online]. Praha: MV, 2018 [cit. 2020-09-08]. Dostupné z: https://www.mvcr.cz/clanek/koncepce-rozvoje-archivnictvi.aspx

Kraken: Description. *Github* [online]. 2020 [cit. 2020-10-15]. Dostupné z: https://github.com/mittagessen/kraken

Kraken: Features. *Kraken* [online]. 2016 [cit. 2020-10-15]. Dostupné z: http://kraken.re/

Marciano, R., Lemieux, V., Hedges, M., Esteva, M., Underwood, W., Kurtz, M. and Conrad, M. (2018), "Archival Records and Training in the Age of Big Data", Percell, J., Sarin, L.C., Jaeger, P.T. and Bertot, J.C. (Ed.) Re-envisioning the MLS: Perspectives on the Future of Library and Information Science Education (Advances in Librarianship, Vol. 44B), *Emerald Publishing Limited*, pp. 179-199. https://doi.org/10.1108/S0065-28302018000044B010

Martínek J., P. Král a L. Lenc. Building an efficient OCR system for historical documents with little training data. *SpringerLink* [online]. Neural Computing and Applications [cit. 2020-10-13]. Dostupné z: https://link.springer.com/article/10.1007/s00521-020-04910-x

MARTÍNEK, J., L. LENCL a P. KRÁL. Building an efficient OCR system for historical documents with little training data: Existing tools and OCR systems. *Neural Computing*

and Applications [online]. 2020 [cit. 2020-10-17]. Dostupné z: doi: https://doi.org/10.1007/s00521-020-04910-x

MCGILLIVRAY, Barbara, Thierry POIBEAU a Pablo Ruiz FABO. *Digital Humanities and Natural Language Processing: "Je t'aime... Moi non plus"* [online]. The Alliance of Digital Humanities Organizations, 2021, 2(14) [cit. 2021-01-28]. Dostupné z: http://www.digitalhumanities.org/dhq/vol/14/2/000454/000454.html

MEJZLÍK, Martin. OCR historických dokumentů [online]. Brno, 2016 [cit. 2020-10-13]. Dostupné z: https://is.muni.cz/th/hynsu/?fakulta=1433. Bakalářská. *Masarykova univerzita Fakulta informatiky*. Vedoucí práce RNDr. Michal Růžička, Ph.D.

MILIONI, Nikolina. *Automatic Transcription of Historical Documents* [online]. Uppsala, 2020 [cit. 2020-09-08]. Dostupné z: http://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1437985&dswid=-1804. Diplomová. Uppsala Universitet.

MORGAN, Goodman. *"What is on this disk?" An Exploration of Natural Language Processing in Archival Appraisal* [online]. Chapel Hill, North Carolina, 2019 [cit. 2021-02-02]. Dostupné z: https://doi.org/10.17615/a13b-va69. Diplomová. School of Information and Library Science of the University of North Carolina. Vedoucí práce Christopher A. Lee.

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, P., Kallio, M., Kaplan, F., Kleber, F., Labahn, R., Lang, E.M., Laube, S., Leifert, G., Louloudis, G., McNicholl, R., Meunier, J.-L., Michael, J., Mühlbauer, E., Philipp, N., Pratikakis, I., Puigcerver Pérez, J., Putz, H., Retsinas, G., Romero, V., Sablatnig, R., Sánchez, J.A., Schofield, P., Sfikas, G., Sieber, C., Stamatopoulos, N., Strauß, T., Terbul, T., Toselli, A.H., Ulreich, B., Villegas, M., Vidal, E., Walcher, J., Weidemann, M., Wurster, H. and Zagoris, K. (2019), "*Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study*", Journal of Documentation, Vol. 75 No. 5, pp. 954-976. https://doi.org/10.1108/JD-07-2018-0114

National Archives releases first version of a Dutch handwriting model. *READ-COOP* [online]. 2019 [cit. 2020-09-19]. Dostupné z: https://readcoop.eu/national-archives-releases-first-version-of-a-dutch-handwriting-model/

Natural Language Toolkit. *NLTK* [online]. 2020 [cit. 2021-01-28]. Dostupné z: https://www.nltk.org/

NEHA, T. Difference Between Entity and Attribute in Database. *BinaryTerms* [online]. 2019 [cit. 2021-02-02]. Dostupné z: https://binaryterms.com/difference-between-entity-and-attribute-in-database.html

OCR Accuracy Measurement. *ABBYY Technology Portal* [online]. [cit. 2020-09-15]. Dostupné z: https://abbyy.technology/en:kb:tip:ocr-accuracy

OCRopus – Python-based tools for document analysis and OCR. *LinuxLinks* [online]. [cit. 2020-10-13]. Dostupné z: https://www.linuxlinks.com/ocropus/

Ocropy: CLSTM vs OCRopy. *Github* [online]. 2020 [cit. 2020-10-13]. Dostupné z: https://github.com/ocropus/ocropy

Optical Character Recognition (OCR). *Microsoft* [online]. 2020 [cit. 2020-10-20]. Dostupné z: https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-recognizing-text

PEREZ, M. Web Scraping vs Data Mining: What's the Difference? *Parsehub* [online]. 2020 [cit. 2021-01-19]. Dostupné z: https://www.parsehub.com/blog/web-scraping-vs-data-mining/

PLATILOVÁ, Mgr. Ivana. *Možnosti aplikace počítačových metod v historii* [online]. Olomouc, 2019 [cit. 2020-12-10]. Dostupné z: https://theses.cz/id/pwyry9/?zpet=%2Fvyhledavani%2F%3Fsearch%3DMo%C5%BEnosti%20aplikace%20po%C4%8D%C3%ADta%C4%8Dov%C3%BDch%20metod%20v%20historii%26start%3D1;isshlret=Mo%C5%BEnosti%3Baplikace%3Bpo%C4%8D%C3%ADta%C4%8Dov%C3%BDch%3Bmetod%3B. Diplomová. Univerzita Palackého v Olomouci. Vedoucí práce Doc. Mgr. Martin Elbel, M.A., Ph.D.

Python package. *Textract* [online]. 2014 [cit. 2020-12-15]. Dostupné z: https://textract.readthedocs.io/en/stable/python_package.html

Python: Get Started. *Python.org* [online]. 2020 [cit. 2020-12-15]. Dostupné z: https://www.python.org/

Read&search. *READ-COOP* [online]. [cit. 2020-09-18]. Dostupné z: https://readcoop.eu/readsearch/

REDDY, Susmith. Segmentation in OCR: A basic explanation of different levels of Segmentation used by the OCR system. *Towards Data Science* [online]. 2019 [cit. 2020-10-13]. Dostupné z: https://towardsdatascience.com/segmentation-in-ocr-10de176cf373

REITER, Brian-Patrick. How to analyse non-digital historical archives of large organisations - text mining case study. *WST Working Paper Series* [online]. Economic and Social History and History of Technology, 2019, (4) [cit. 2020-09-26]. Dostupné z: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3460121

ROBERTSON, Mikaela. 6 technologies behind AI: Type #4: Computer vision. *CodeBots* [online]. 2018 [cit. 2020-09-21]. Dostupné z: https://codebots.com/artificial-intelligence/6-technologies-behind-ai

ROMEIN, Annemieke. Entangled Histories: OCR + HTR = ATR: Automatic Text Recognition. *KB LAB* [online]. 2020 [cit. 2020-09-15]. Dostupné z: https://lab.kb.nl/about-us/blog/entangled-histories-ocr-htr-atr-automatic-text-recognition

Ryantová, M. (2017). Training of Archivists in the 21st Century: Some Reflections. *Atlanti*, 27(2), 225-233. https://doi.org/10.33700/2670-451X.27.2.225-233(2017)

ScanTent + DocScan App. *READ-COOP* [online]. 2020 [cit. 2020-09-17]. Dostupné z: https://readcoop.eu/scantent/

SHILPA, Dang. Text Mining : Techniques and its Application. *International Journal of Engineering & Technology Innovations* [online]. Indie: M.M Institute of Computer Technology & Business Management Maharishi Marakandeshwar University, 2014, (4), 22-25 [cit. 2021-01-20]. ISSN 2348-0866. Dostupné z: https://www.researchgate.net/publication/273038150_Text_Mining_Techniques_and_its_Application

SMITS, Thomas. Illustrations to Photographs: Using computer vision to analyse news pictures in Dutch newspapers, 1860-1940. *DH* [online]. Raboud University, The Netherlands, 2017 [cit. 2020-09-22]. Dostupné z: https://www.semanticscholar.org/paper/Illustrations-to-Photographs%3A-using-computer-vision-Smits/ad6ac997dcafe9b975d5758578fd3dd19fbd2ccb

SPRINGMANN, U., D. NAJOCK, H. MORGENROTH a SCHMID. OCR of historical printings of latin Texts: Problems, prospects, progress. *Document and Text Processing* [online]. 2014 [cit. 2020-10-15]. Dostupné z: http://springmann.net/papers/2014-DATeCH-Springmann.pdf

Springmann, Uwe & Najock, Dietmar & Morgenroth, Hermann & Schmid, Helmut & Gotscharek, Annette & Fink, Florian. (2014). OCR of historical printings of latin Texts: Problems, prospects, progress. *ACM International Conference Proceeding Series*. 10.1145/2595188.2595205.

SPRINGMANN, Uwe. Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. *JLCL* [online]. München, 2018, 1(33), 17 [cit. 2020-09-15]. Dostupné z: https://arxiv.org/abs/1809.05501v1

Studijní plány. *FILOSOFICKÁ FAKULTA Univerzita Karlova* [online]. Praha, 2020 [cit. 2020-09-10]. Dostupné z: https://www.ff.cuni.cz/studium/studijni-obory-plany/studijni-plany/

Tagging. *READ-COOP* [online]. 2020 [cit. 2020-11-17]. Dostupné z: https://readcoop.eu/transkribus/howto/transkribus-transcription-conventions/

TATEOSIAN, L., R. GUENTER, Y. YANG aj. RISTAINO. Tracking 19th century late blight from archival documents using text analytics and geoparsing. *Conference:*

International Conference for Free and Open Source Software for Geospatial [online]. 2017 [cit. 2020-09-26]. Dostupné z: https://www.researchgate.net/publication/322754348_TRACKING_19TH_CENTURY_LATE_BLIGHT_FROM_ARCHIVAL_DOCUMENTS_USING_TEXT_ANALYTICS_AND_GEOPARSING

Tesseract OCR. *Google Open Source* [online]. 2020 [cit. 2020-10-15]. Dostupné z: https://opensource.google/projects/tesseract

Tesseract OCR: About. *Github* [online]. 2020 [cit. 2020-10-15]. Dostupné z: https://github.com/tesseract-ocr/tesseract

Text2Image. *Transkribus* [online]. 2019 [cit. 2020-11-17]. Dostupné z: https://transkribus.eu/wiki/index.php/Text2Image

Textract. *Textract* [online]. 2014 [cit. 2020-12-15]. Dostupné z: https://textract.readthedocs.io/en/stable/

The Austrian government meets READ DocScan and ScanTent! *READ-COOP* [online]. 2017 [cit. 2020-09-17]. Dostupné z: https://readcoop.eu/austrian-govt-meets-docsan-and-scantent/

The Old Bailey and OCR: Benchmarking AWS, Azure, and GCP. *ACM Symposium on Document Engineering* [online]. New York, NY, USA: Association for Computing Machinery, 2020, 4 [cit. 2020-10-29]. Dostupné z: https://doi.org/10.1145/3395027.3419595

The roots of Transkribus. *READ-COOP* [online]. 2020 [cit. 2020-11-19]. Dostupné z: https://readcoop.eu/transkribus/?sc=Transkribus

TITENOK, Yaroslav. Natural Language Processing vs Text Mining. *Sloboda Studio* [online]. 2020 [cit. 2021-01-28]. Dostupné z: https://sloboda-studio.com/blog/natural-language-processing-vs-text-mining/

TOME: Interactive TOpic Model and MEtadata Visualization. *Digital Integrative Liberal Arts Center* [online]. Georgia, 2018 [cit. 2020-09-24]. Dostupné z: https://dilac.iac.gatech.edu/dilac-projects/topic-model-metadata-visualization

Training an Ocropus OCR model. *Danvk.org* [online]. 2015 [cit. 2020-10-13]. Dostupné z: https://www.danvk.org/2015/01/11/training-an-ocropus-ocr-model.html

TranScriptorium. *TranScriptorium* [online]. 2013 [cit. 2020-09-17]. Dostupné z: http://transcriptorium.eu/

Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation* [online]. Emerald Publishing Limited, 2019, 75(5), 2 [cit. 2020-09-17]. ISSN 0022-0418. Dostupné z: https://www.emerald.com/insight/content/doi/10.1108/JD-07-2018-0114/full/html

Transkribus Credits: About Transkribus Credits. *Transkribus* [online]. 2020 [cit. 2020-10-20]. Dostupné z: https://readcoop.eu/transkribus/credits/

Using AI to unleash the power of unstructured government data: Applications and examples of natural language processing (NLP) across government. *Deloitte Insights* [online]. USA, 2019 [cit. 2020-09-24]. Dostupné z: https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/natural-language-processing-examples-in-government-data.html

Vision AI. Google Cloud [online]. 2020 [cit. 2020-10-20]. Dostupné z: https://cloud.google.com/vision

W. ELINGS, Mary. Using NLP to Support Dynamic Arrangement, Description, and Discovery of Born Digital Collections: The ArchExtract Experiment. *SAAERS* [online]. 2016 [cit. 2020-09-24]. Dostupné z: https://saaers.wordpress.com/2016/05/24/using-nlp-to-support-dynamic-arrangement-description-and-discovery-of-born-digital-collections-the-archextract-experiment/

Web Scraping vs Data Mining: What's the Difference? *Parsehub* [online]. 2020 [cit. 2020-12-15]. Dostupné z: https://www.parsehub.com/blog/web-scraping-vs-data-mining/

What is Computer Vision? *DeepAI* [online]. [cit. 2020-09-19]. Dostupné z: https://deepai.org/machine-learning-glossary-and-terms/computer-vision

What is Text Mining: Methods and Techniques. *MonkeyLearn* [online]. [cit. 2021-01-21]. Dostupné z: https://monkeylearn.com/text-mining/

WILES, Rachel. Have we solved the problem of handwriting recognition?: Before Deep Learning, there were OCRs. *Towards data science* [online]. London, 2019 [cit. 2020-09-15]. Dostupné z: https://towardsdatascience.com/https-medium-com-rachelwiles-have-we-solved-the-problem-of-handwriting-recognition-712e279f373b

YIN, Michael. Web Scraping Framework Review: Scrapy VS Selenium. *AccordBox* [online]. 2019 [cit. 2020-12-17]. Dostupné z: https://www.accordbox.com/blog/web-scraping-framework-review-scrapy-vs-selenium/

ZELIC, F. a A. SABLE. A comprehensive guide to OCR with Tesseract, OpenCV and Python. *Nanonets: Automate Data Capture* [online]. 2020 [cit. 2020-10-15]. Dostupné z: https://nanonets.com/blog/ocr-with-tesseract/#opensourceocrtools

Seznam obrázků