ZÁPADOČESKÁ UNIVERZITA V PLZNI
FAKULTA PEDAGOGICKÁ
KATEDRA VÝPOČETNÍ A DIDAKTICKÉ TECHNIKY

PRINCIPY PŘEVODU TEXTU Z TIŠTĚNÉ DO DIGITÁLNÍ PODOBY A ZPŮSOB APLIKACE V OSOBNÍCH POČÍTAČÍCH

BAKALÁŘSKÁ PRÁCE

Jana Záhorová
Přírodovědná studia
Informatika se zaměřením na vzdělávání

Vedoucí práce: PhDr. Zbyněk Filipi, Ph.D.

Plzeň 2017

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně s použitím uvedené literatury a zdrojů informací.

V Plzni 30. června 2017

.....

vlastnoruční podpis

Děkuji své rodině a přátelům za dodávanou sílu a podporu.

OBSAH

SEZNAM ZKRATEK

CCD	Charge-coupled device, technologie snímače skeneru
CD-ROM	Compact Disc Read-Only Memory, nepřepisovatelné optické medium sloužící k uložení a čtení počítačových dat
CIS	Contact Image Sensor, technologie snímače skeneru
ČSÚ	Český statistický úřad
DPI	Dots per inch
e-book	electronic book – elektronická kniha, elektronická publikace obecně
EPUB	Electronic Publication – elektronická publikace, zároveň i jeden z elektronických knižních formátů
FRA	FineReader Abbyy – OCR software
ICR	Intelligent character recognition
IWR	Intelligent word recognition
KDD	Knihovna digitálních dokumentů
MOBI	Mobipocket – formát e-booku primárně vyvíjený pro firmu Kindle
NK ČR	Národní knihovna České republiky
OCR	Optical Character Recognition – optické rozpoznávání znaků
PDF	Portable Document Format – formát dokumentů, jehož zobrazení je nezávislé na platformě
QR	Quick response – kód pro obrazové uložení informace
RGB	red-green-blue – barevný model
SONS	Sjednocené organizace nevidomých a slabozrakých
SR	Slovenská republika
UNESCO	Organizace OSN pro vzdělání, vědu a kulturu
VHS	Video Home System, kazeta s magnetickou páskou k záznamu obrazu a zvuku

ÚVOD

Pryč jsou doby, kdy jste po městě, v parcích či na nádražích potkávali čtenáře mající oči zabořené do papírové knihy nebo časopisu. Ne že by úplně vymizeli, ale oči mnohým z nich teď místo do papíru hledí do mobilních telefonů, čteček či tabletů. Nebudu ve své bakalářské práci rozebírat klady či zápory papírové verze oproti digitální, ale spíš se zaměřím na proces, kterým se z papírové knihy stává právě elektronická.

V úvodní kapitole se zamýšlím nad významem digitalizace v současném světě. Přibližuji přístup jednotlivých státních organizací k této problematice v oblasti v oblasti správní, kulturní a vědecké. V následujících kapitolách se snažím rozebrat principy převodu textu z papírové verze do digitální, a představím faktory, které ovlivňují kvalitu OCR převodu. Na závěr uvádím praktickou ukázku převedení papírové knihy do digitálního formátu v komerčním softwaru FineReader Abbyy 11, který bude doplněn videonávody na přiloženém DVD. Zmíním výhody i nevýhody programu a dále popisuji export do jednotlivých typů souboru od editovatelných (docx, rtf) po needitovatelné (pdf, djvu, epub ad.) Nemohu zanedbat ani možnosti úpravy zbývajících chyb, které unikly při prvních opravách přímo v programu FineReader Abbyy.

1 VÝZNAM DIGITALIZACE V SOUČASNOSTI

Obecně termín digitalizace označuje technický proces převodu vybraných měřitelných fyzikálních veličin konkrétního objektu do binárních hodnot [1]. Nevztahuje se jen k tištěnému textu, ale i k obrazovému, audio či video materiálu jako jsou mapy, obrazy, fotografie, mikrofilmy, noty, kazety, VHS, exponáty atd. Vzniklé elektronické[1] formáty jsou mimo vlastních dat doplněny ještě o tzv. metadata, což jsou katalogizační údaje díla.

Jedná se tedy o další způsob zachovávání a přenášení informací ve světě. Vzhledem k množství elektronických zařízení, která digitální data zobrazují, by se mohlo zdát, že končí éra papíru. Není to však poprvé, co v dějinách dochází k výměně jednoho média za jiné. Nikdy však předchozí formát nebyl plně nahrazen.

Zpočátku probíhalo předávání informací ústně, což mělo své nevýhody, ať už se jednalo o dlouhou dobu doručení, docházelo ke zkreslení informace a hlavně vyvstala otázka – jak zprávy věrně uchovat. S vynálezem písma došlo k posunu v oblasti věrnosti sdělení, uchování a dokonce i jeho utajení. Za další zlom v historii lze považovat vynález knihtisku v polovině 15. století, který masově rozšířil písmo i mezi obyčejný lid. O čtyři století později vedlo sestrojení telegrafu k zrychlení předávání sdělení – zprávy se přenášely v reálném čase a na velké vzdálenosti. A dnes díky rozšířenosti sdělovacích médií neřešíme jen rychlost předávání zpráv na velkou vzdálenost, jejich obrovskou kvantitu, ale i množství lidí, které k nim má přístup. I přesto, že se vývoj stále žene dopředu, tak v současnosti vedle sebe koexistují všechny formy uchovávání informací.

Digitalizace je logickým pokračováním trendu uchovávání sdělení, dat, kulturního i historického dědictví. V dnešní době se jedná i o nezbytnou potřebu dobře fungující veřejné správy státu – zdravotnictví, školství atd. Zejména představuje přínosné a praktické řešení – digitální materiál je v rámci svého určení rychle dostupný, snadno přenositelný, zabírá výrazně méně prostoru, šetří peníze – a často bývá uváděno i obligátní tvrzení, že šetří životní prostředí.

Ekologické hledisko je však diskutabilní. Přeci jen jakákoliv forma elektronického nosiče potřebuje zdroj energie ke svému spuštění a běhu. V České republice (ČR) se na výrobě elektrické energie podle údajů Českého statistického úřadu (ČSÚ) z necelých 51 % [2]

[1] Pojem elektronické formáty, potažmo elektronická kniha, je v běžné komunikaci častěji používaný termín, než pojem digitální kniha.

podílejí parní (tepelné) elektrárny, které patří mezi činitele účastnící se znečišťování prostředí. Instituce stále nemají plnou důvěru v digitální uchovávání svých dat a tak vedle elektronické podoby zachovávají i papírovou zálohu dokumentů. Právě pro případ selhání elektronického nosiče, výpadku elektrické energie či chybného zápisu dat. Je tedy k uvážení, k jak velké úspoře papíru potažmo lesů dochází.

Jako jeden z kladů elektronických dokumentů je uváděna snadná a rychlá dostupnost. Pokud nahlédneme do údajů ČSÚ, tak zjistíme, že v roce 2016 bylo v ČR připojeno k počítačové síti téměř 76 % [3] všech domácností. Což je 111% nárůst oproti roku 2006, kdy bylo k síti připojeno pouhých 36 %. Není překvapivé zjištění, že rodiny s dětmi mají vyšší procento připojení k internetu než domácnosti bez dětí. Rozdíl mezi nimi činí 26 %, což není zanedbatelné číslo. Přeci jen je znatelné, že mladší generace se s ICT technologiemi učí pracovat již ve školních institucích.

Stojí za povšimnutí, že čísla vyjadřující počet domácností vybavených osobním počítačem (PC) jsou téměř totožná s čísly vyjadřujícími připojení k internetu. Jediný rozdíl nalezneme v domácnostech bez dětí, kdy PC vlastní 68 % domácností [4], ale připojení k internetu je ve stejné skupině o jediné procento vyšší. Z předchozích údajů lze usuzovat, že připojení k internetu se stává standardní podmínkou pro práci s osobním počítačem a dalšími elektronickými zařízeními.

Počítačová síť je v současné době živé prostředí, které zvyšujícímu se počtu svých uživatelů nabízí nepřeberné množství interakcí a služeb. Na tento trend se snaží reagovat i instituce, které ve značné míře ovlivňují život občana – ať už je to veřejná správa státu, zdravotnictví i samotné školství a vědecké instituce. Snaží se nabídnout uživatelům rychlejší a méně komplikované on-line služby.

Efektivita těchto služeb je ale závislá na několika faktorech. Patří mezi ně např. technologická vybavenost jednotlivých institucí, počítačové dovednosti nejen poskytovatele ale i koncových uživatelů, pro které jsou služby primárně určeny. V případě uchovávání dat je třeba řešit i volbu formátového standardu. Technologický vývoj postupuje rychle dopředu a současné formáty by nemusely být v budoucnosti kompatibilní.

Pro ilustraci můžeme uvést jako příklad současnou bouřlivou diskuzi na téma povinného zavádění eReceptů ve zdravotnictví. Jsou součástí schválené Národní strategie elektronického zdravotnictví na období 2016 – 2020. Tato strategie si vyčlenila několik cílů, které mají více zapojit občana při péči o své zdraví, má zvýšit efektivitu systému, zvýšit kvalitu a dostupnost zdravotních služeb a vytvořit informační infrastrukturu [5]. Samotný eRecept má přinést úlevu pro stálé pacienty, kteří potřebují pouze napsat recept. Již nebudou muset čekat u lékaře, ale specifický kód jim dorazí e-mailem nebo jako sms. Dále má poskytovat lékařům zpětnou vazbu, zda pacient léky vyzvedne a tím dodrží léčebný režim. Měl by zabraňovat zneužívání receptů [6] atd.

Lékaři zatím systému vytýkají nevyužité možnosti. V podstatě musí investovat do softwaru, který zatím nahrazuje lékařův podpis a razítko [7]. Přitom by mohl zabraňovat vzniku duplicit, sledovat interakci léků u pacienta, vést lékový záznam atd. Dále pak šéf České lékařské komory Milan Kubek upozorňuje na fakt, že absence osobního počítače či kvalitního a hlavně spolehlivého připojení k internetu se týká tisíců lékařů [8].

Národní kontrolní úřad (NKÚ) ve své květnové tiskové zprávě uvádí, že již bylo do evidence investováno přes 300 milionů korun a ve sledovaném období 2011 – 2016 bylo ze všech receptů pouhé 1,5 % vydáno elektronicky [9]. Další obavy panují z toho, že se lékař při své práci bude věnovat více administrativě než samotnému vyšetřování pacientů. Plný dopad tohoto kroku ale bude viditelný až v roce 2018, kdy bude vydávání eReceptu povinné pro všechny lékaře.

V souvislosti s digitalizací (nebo též elektronizací) se také stále řeší otázka autorských práv a ochrany osobních údajů. Vzhledem k zadání bakalářské práce je důležité znění zákona č. 121/2000 Sb., zákon o právu autorském, o právech souvisejících s právem autorským [10]. Zejména paragrafy 30 a §30a, které upravují volné užití díla i jinou osobou než je samotný autor. A to v tom smyslu, že si fyzická osoba může pořídit přepis, rozmnoženinu, napodobeninu, ale pouze jen pro svou vlastní osobní potřebu. Nesmí ji tudíž šířit, nebo z ní mít jakýkoliv hospodářský prospěch. Pokud si tedy zakoupíte papírovou knihu a chcete si vytvořit její elektronickou kopii pro svou čtečku elektronických knih, ponecháte tam veškeré údaje o autorovi a nakladateli, tak se tímto jednáním žádného přestupku nedopouštíte. Nesmíte ji však poskytnout nikomu jinému.

1.1 KULTURNÍ A HISTORICKÉ DĚDICTVÍ

Dějiny jsou plné násilných aktů či přírodních katastrof, které vedly ke zničení památek a historických dokumentů. Ať už se jednalo o války a bombardování, či následné rabování národního dědictví nebo o požáry a ničivé povodně, které po sobě zanechávaly zcela zdevastované dokumenty. Svět a spolu s ním i ČR dospěly k bodu, kdy se rozhodly, že své kulturní a historické bohatství zdigitalizují a zachovají v elektronických formátech pro budoucí generace. Ve svých programech se zavázaly zpřístupnit elektronický obsah knihoven nejen akademické obci, ale i široké laické veřejnosti v souladu s autorským právem jednotlivých děl. Přednostně jsou digitalizována díla z tzv. veřejné domény (*public domain*), kterým autorská práva již vypršela.

1.1.1 PROJEKTY VE SVĚTĚ

První projekty zabývající se digitalizací a zpřístupněním elektronických knih se objevily ve Spojených státech amerických (USA). Vznikl zde zřejmě nejstarší stále živý projekt *Gutenberg*. Na počátku 90. let 20. stol. spustila americká Kongresová knihovna ve Washingtonu projekt *American Memory Programme*, který se zaměřil na digitalizaci dokumentů týkajících se amerických dějin. Výsledná práce byla poskytnuta formou CD-ROMů školám na podporu výuky historie. Kongresová knihovna si tak vytvořila vlastní standardy a digitalizační postupy, které využívala i v dalších dílčích projektech. Cílem bylo, a stále je, vytvoření jednoho společného přístupu do všech amerických sbírek. Americkými zkušenostmi se v následujících letech inspirovaly státy v Evropě.

V Lisabonské smlouvě z roku 2000 vznikla strategie pro zvýšení produktivity a ekonomického růstu Evropy zejména v konkurenci USA a Japonska, které Evropu technologicky předstihly. Jedním z dílčích cílů bylo bezpečné využívání internetu a vytváření digitálního obsahu. Zpřístupnění náplně mělo vést k individuálnímu, profesnímu i sociálnímu rozvoji společnosti a mělo doplňovat proces vzdělání. Další dílčí akční plány eEurope řešily otázku digitalizace v rámci Evropy, jednotlivých států i samotných kulturních a historických institucí.

Jako hlavní cíl strategie Evropské komise pro digitalizaci kulturního a vědeckého dědictví vešel v platnost v roce 2005 podnět i2010: Digitální knihovny. Zaměřil se na výzkum technologií souvisejících s tvorbou a funkcí digitálních knihoven podporujících vytvoření

evropské digitální knihovny. Výsledkem je portál *Europeana*. Jejich cílem je zachování historických dokumentů vztahujících se k dějinám daného státu [11]. Další celoevropský projekt spustila Organizace OSN pro vzdělání, vědu a kulturu (UNESCO) pod názvem *Unesco Memory of the World*. V jednotlivých státech byly též spuštěny velké projekty – např. Francie se svým projektem *Gallica*. Níže následuje bližší seznámení s některými světovými projekty.

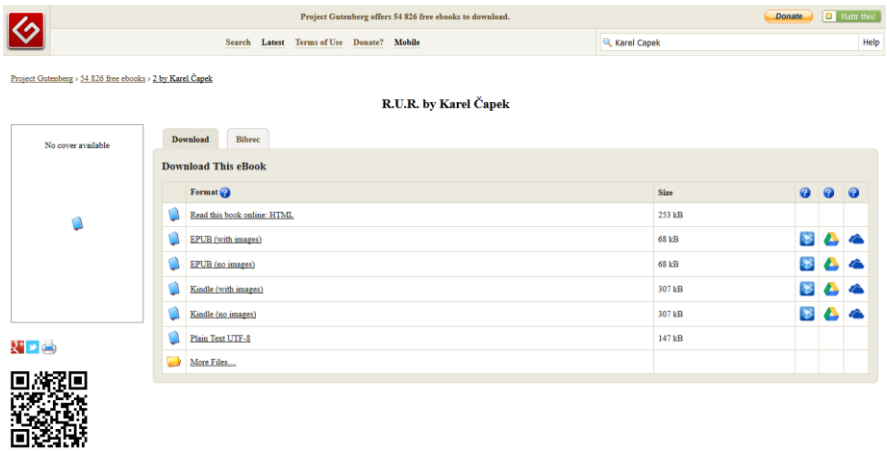
Gutenberg

Michael Hart (1947–2011) byl známý americký autor, který se proslavil založením průkopnického projektu, poskytujícího volně dostupná díla v elektronickém formátu. Tento projekt vznikl již v roce 1971, a ačkoliv i on musel reagovat na změny v amerických zákonech týkající se autorských práv, poskytoval a poskytuje volná díla, nebo díla s již ukončeným autorským právem. V roce 1998 americká vláda vydala zákon Copyright Term Extension Act (CTEA) [13], který upravoval autorské zákony ze 70. let 20. stol. a rozšířil ochranu autorského díla vydaného před rokem 1978 na 95 let od jeho publikování.

Celá databáze čítá na 54 000 volně stažitelných elektronických knih ve formátech určených pro e-book čtečky jako je epub (electronic publication) či mobi (Mobipocket) pro Kindle. Celý projekt je založený na práci dobrovolníků [14]. Digitalizace se zaměřuje na tři oblasti literatury:

- lehká literatura (dětské příběhy, pohádky, bajky),
- knihy pro náročnější čtenáře (náboženské texty, klasičtí autoři – Shakespeare, Melville atd.,
- vědecká část (slovníky, almanachy, encyklopedie).

Projekt se samozřejmě primárně zabývá anglicky psanou literaturou, ale ani ostatní jazyky zde nepřijdou zcela zkrátka. Jazyky jsou rozděleny do dvou kategorií, kdy dělící hranici tvoří limit 50 zveřejněných e-booků. Do skupiny, která zde má větší zastoupení, patří sedmnáct jazyků, mezi něž patří např. čínština, dánština, nizozemština, finština, francouzština i umělý jazyk esperanto. Do té druhé skupiny se řadí dalších padesát jazyků, jako je namátkou japonština, srbština, bretonština, hebrejština, maorština nebo i čeština. Z národních autorů je zde zastoupen Karel Čapek, ale třeba zde najdeme i českého překladatele Jaromíra Hrubého. Je zde uložen pod pseudonymem H. Jaroš jeho český překlad Dostojevského díla Zápisky z mrtvého domu.



Obrázek 1 - prostředí projektu Gutenberg (Zdroj: vlastní)

Navštívíte-li webové stránky projektu[2], naleznete vcelku přívětivé a hlavně jednoduché uživatelské prostředí v angličtině a dalších třech dostupných jazykových verzích. Tím, že se jedná o neziskový projekt, je grafická stránka projektu upozaděna a primárním požadavkem je funkčnost a obsah projektu. Vzhledem k tomu, že je u knih provedeno OCR, má Gutenberg obsah své knihovny indexován nejen pomocí metadat, ale i přímé citace textu díla, díky čemuž lze dílo dohledat pomocí internetového vyhledávače – zejména amerického Yahoo.

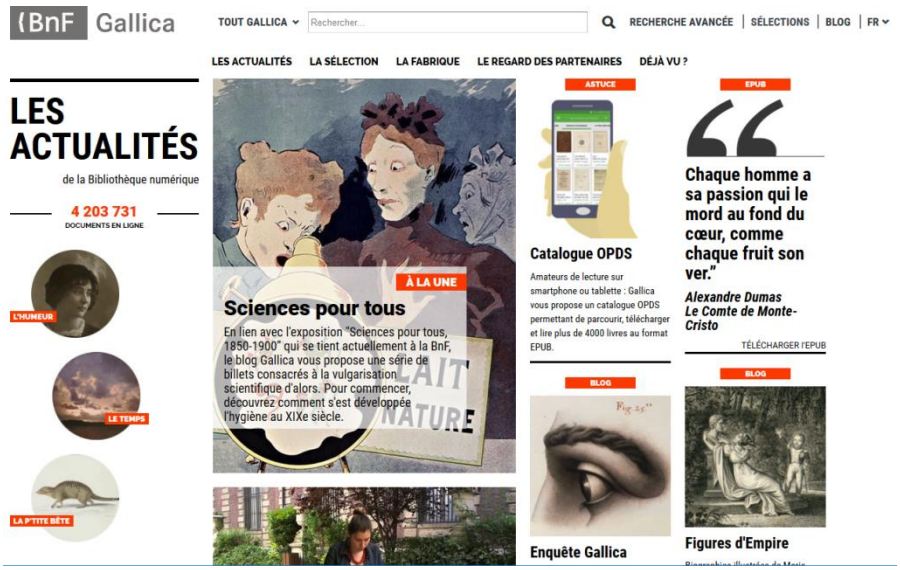
Gallica

Francouzská národní knihovna v roce 1997 převzala projekt digitální knihovny Gallica[3], který v současnosti obsahuje přes čtyři miliony digitálních děl. Nezaměřují se pouze na textové dokumenty, ale digitalizují i manuskripty, mapy, plány, audio a video nahrávky, partitury, pohlednice a obrazy [22]. Ročně do knihovny přibude okolo sta tisíce děl.

Vizuální stránka projektu je barevnější a pestřejší. Je však otázka, zda je i přehlednější. Ačkoliv v roce 2016 byla dostupná ve čtyřech jazykových verzích, jen o rok později je přístupná pouze ve francouzštině. Volná díla lze pak zobrazit na monitoru, či stáhnout v různých formátech – nejčastěji pdf. Obrazové soubory se zobrazují ve formátu jpg (komprimovaný formát rastrové grafiky) a tiff (nekomprimovaný formát rastrové grafiky). Textové dokumenty, u kterých bylo provedeno OCR (Optical Character Recognition), lze stáhnout ve formátu txt, nebo je přímo ve webové aplikaci fulltextově

[2] Dostupné na: http://www.gutenberg.org
[3] Dostupné na: http://gallica.bnf.fr

vyhledávat. Jenže na rozdíl od projektu Gutenberg není obsah knihovny Gallica indexován, ačkoliv díla obsahují metadata.



Obrázek 2 - Úvodní stránka francouzského projektu Gallica (Zdroj: vlastní)

Europeana

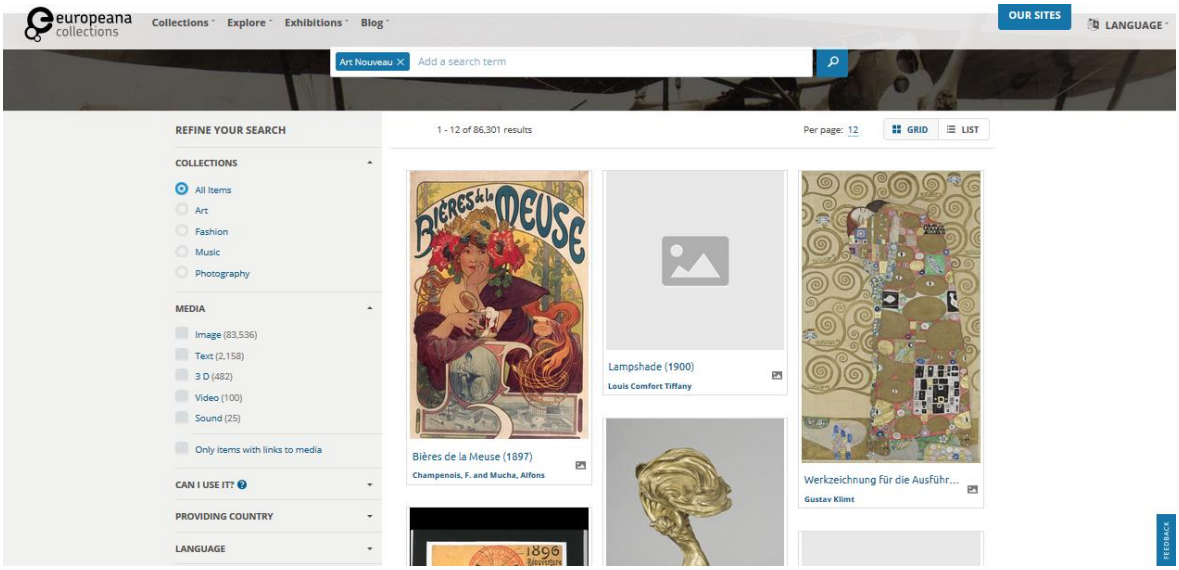
V listopadu roku 2008 byl oficiálně spuštěn projekt Europeana[4]. Jedná se o výsledek iniciativy Evropského parlamentu s názvem *i2010: Směrem k evropské digitální knihovně*. Tento pokyn zastřešil iniciativu z předchozích let *i2010: Digitální knihovny* [15].

Tato digitální knihovna, muzeum a archiv v jednom, je přístupná ve 27 jazykových mutacích, kde je zpřístupněno přes 53 milionů děl z národních a univerzitních knihoven států Evropské unie – mezi nimi se nachází i ČR s projektem Národní digitální knihovny. Jedná se tedy o nejobsáhlejší sbírku digitálních dokumentů na evropském kontinentu. Zpočátku byl do knihovny přidán již existující digitální obsah paměťových institucí jednotlivých států a Evropská komise předpokládá, že do roku 2025 bude zdigitalizováno celoevropské kulturní dědictví [17].

Každý digitální materiál je doplněn o metadata, u kterých se používá standard Dublin Core[5]. Na základě metadat je vytvořen rejstřík, který je po zadání požadavku uživatelem prohledáván. Jako výsledek se zobrazí náhledy dokumentů s možností výběru požadovaného formátu, jazyka, země aj. Plné zobrazení vybraného dokumentu proběhne až na stránkách instituce, která digitální dokument uchovává. Nedochází tak k duplicitě

[4] Dostupné na: http://www.europeana.eu
[5] Více informací na stránkách Masarykovy univerzity v Brně: http://webserver.ics.muni.cz/dublin_core/

materiálu a jsou tak vyřešena i autorská práva, která se řídí pravidly toho daného státu [18]. Cílem portálu je umožnit uživatelům nahlédnout na historické epochy, umělecké slohy či události v celoevropských souvislostech nejen pasivně, ale i aktivně formou blogů, diskuzních fór atd.



Obrázek 3 - ukázka prostředí portálu Europeana (Zdroj: vlastní)

1.1.2 PROJEKTY V ČESKÉ REPUBLICE

Digitalizace dokumentů v ČR by se dala rozčlenit do tří fází podle toho, jak byly jednotlivé digitalizační projekty chápány, kdo je organizoval, z jakých fondů byly dotovány a jaké byly jejich výsledné výstupy [19].

- experimentální (od 90. let 20. st. do roku 2002),
- ochranná selektivní digitalizace, projekty týkající se born-digital[6] materiálů,
- masové projekty (po roce 2011) .

V první fázi se digitalizace zaměřovala na dokumenty, které byly ve sbírkách nejcennější (prvotisky, staré tisky aj.). V roce 1992 se ČR připojila k projektu UNESCO – *Paměť světa (Memory of the World Programme)* a vytvořila program *Memories Mundi Series Bohemica*. Vzhledem k technologickým možnostem v této době byly výsledné digitalizace zprostředkovávány uživateli offline prezentacemi na optických médiích (CD-ROM).

V druhé fázi – ke konci 90. let – se digitalizace začala zaměřovat na dokumenty ohrožené znehodnocením. Některé dokumenty byly v takovém stádiu rozpadu, že jakékoliv

[6] označení pro dokumenty, které vznikly v digitální podobě a nemají fyzickou (např. tištěnou) podobu.

zpřístupnění širší veřejnosti by je mohlo nenávratně zničit. Vytvoření digitální kopie bylo jedinou cestou, jak takováto díla zachovat. Formálního přijetí digitalizace jako metody ochrany dokumentů se dostalo ve zprávě Americké knihovnické asociace v roce 2004 s názvem *Digitization as a preservation reformatting method*. V českém prostředí začala probíhat výběrová digitalizace nejvíce ohrožených děl. Digitalizační centra se vytvořila i v dalších knihovnách, např. v Moravské zemské knihovně, v Moravskoslezské vědecké knihovně v Ostravě atd. V této fázi se v ČR objevuje projekt *Kramerius* – jakýsi prototyp digitální knihovny, ve kterém jsou díla v tomto období převáděna pouze na mikrofilmy. Projekt WebArchiv spuštěný v roce 2001 řeší problematiku born-digital materiálů, v současnosti obsahuje téměř 300 TB dat.

Po roce 2005 přestává být cílem digitalizace ochrana papírových dokumentů, ale objevuje se snaha nabídnout veřejnosti co nejvíce děl v elektronickém formátu online. Třetí fáze se tak vyznačuje tzv. masovou digitalizací, která je třeba typická např. pro projekt Google Books. Evropská Unie masivní digitalizaci podpořila již roku 2006 ve svém *Doporučení Evropské komise členským státům* ze srpna 2006. Mezi léty 2007 až 2010 probíhala jednání mezi NK ČR a firmou Google ohledně spolupráce na projektu Google Books. V roce 2008 se začaly rýsovat náznaky největšího českého projektu Národní digitální knihovny a v roce 2011 došlo k oficiálnímu spuštění. Born-digital dokumenty jsou v této fázi ošetřeny v projektu e-deposit[7] [19].

Memories Mundi Series Bohemica

Jak už bylo zmíněno výše, v roce 1992 se ČR připojila k projektu *Paměť světa* organizace UNESCO. Už o rok později představila na konferenci v Paříži první digitalizovaný rukopis na CD-ROMu. V roce 1995 vznikla česká odnož projektu *Memories Mundi Series Bohemica*. Ve stejném roce byl vyhlášený program na národní úrovni – Národní program digitálního zpřístupnění vzácných dokumentů *Memories Mundi Series Bohemica*. V Národní knihovně ČR (NK ČR) bylo vytvořeno digitalizační pracoviště. Metadata v této době ještě neměla žádné vytvořené standardy. Zpočátku knihovna využívala popis na základě značkovacího jazyka HTML, později ve spolupráci s firmou AiP Beroun vyvinula vlastní prostředí DOBM (*Digitized Old Books and Manuscript*) zpracované na platformě SGML [20]. Později byl projekt začleněn do mladší verze *Manuscriptorium*.

[7] Dostupné na adrese: http://edeposit.nkp.cz/cs

Kramerius

Projekt *Kramerius* je vyvíjen od roku 1999. Po roce 2002 byly cíle projektu rozšířeny na popud dopadu katastrofálních povodní. Bylo poničeno velké množství vzácných knih zejména v pražských knihovnách. Má tedy za cíl reformátovat knižní zdroje kvůli jejich záchraně a zachování. Primárně je určen zejména pro papírovou tvorbu – tzn. monografie a periodika. Dále však může obsahovat i vydání, která vyšla pouze elektronicky, či mapy, notopisy atd. V současné době projekt zaštiťuje Knihovna Akademie věd České republiky.

Čtvrtá verze je vyvíjena od roku 2009 a finančně je podporována dotačními programy Akademie věd a Ministerstva kultury České republiky. Díla jsou zpřístupněna ve všech prohlížečích, ale protože jsou primárně zveřejňována ve formátu DjVu, uživatel musí mít v browseru nainstalovaný plug-in pro jeho prohlížení. Uživatelské prostředí je zpřístupněno ve dvou jazycích – čeština a angličtina. Uživatel má k dispozici nápovědu pro lepší orientaci a infoportál[8] o současném dění.

Některá díla, jejichž zveřejnění by nebylo v souladu s autorským zákonem, mohou být zobrazena pouze v prostorách Národní knihovny a Moravské zemské knihovny. Lze si objednat kopii těchto dokumentů, nebo lze též objednat kopii konkrétního článku – včetně elektronického – a to prostřednictvím služby eDDO[9].

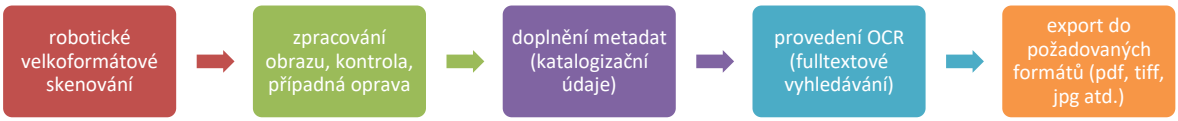
Národní digitální knihovna (NDK)

Národní digitální knihovna využívající ke zpřístupnění děl projekt *Kramerius* vzniká z fondů Národní knihovny České republiky a Moravské zemské knihovny. Obě knihovny mají právo autorského výtisku a získávají do svých fondů veškerou dostupnou bohemikální produkci (zahraniční publikace o České republice i archivní fondy). Projekt je spolufinancován ze Strukturálních fondů EU pro regionální rozvoj. Má tedy za úkol zdigitalizovat, dlouhodobě ochránit a zpřístupnit podstatnou část svých knihovních fondů.

V knihovnách fungují tzv. digitalizační linky, ve kterých probíhá proces od skenování, zpracování obrazu – jeho následné kontroly a případné úpravy (korekce stran), doplnění metadat (katalogizační údaje), a provedení OCR, kvůli fulltextovému vyhledávání, až

[8] Dostupné na: http://kramerius-info.nkp.cz/
[9] knihovní služba umožňující objednávat a přijímat elektronické kopie části dokumentů (do dvaceti stran) přes internet, kopie jsou většinou dodávány do 48 hodin

k výslednému výstupu v podobě pdf, djvu, či dalších obrazových souborů **Chyba! Nenalezen zdroj odkazů.**.



Obě digitalizační linky jsou vybaveny velmi výkonnými knižními skenery, mj. švýcarský 4DigitalBooks DL 3003[10], rakouský Treventus ScanRobot 2.0 MDS, Canon a Plustek, které mají zajistit denní produkci až 54 tisíc stránek denně z obou pracovišť. To znamená až jeden milion stránek za měsíc. Skenery jsou robotické a velkoformátové, čímž se urychluje celý proces a minimalizuje se lidská chyba. Skenery mají senzory na detekci pokrčených stran, dokáží samostatně obracet stránky a například skener 4DigitalBooks má čtecí hlavu se snímačem nahoře nad sklem (face-up), kniha se ke sklu přitiskne zespodu, čímž se stránky vyrovnávají. U stolních skenerů je to přesně naopak, čtecí hlava je pod skenovacím sklem a předloha se k němu tiskne nahoře.

Od roku 2011 v dalších knihovnách České republiky probíhají regionální projekty zaměřující se na uchování a digitalizací děl souvisejících právě s daným regionem [23]. V takovémto množství je logické, že musela vzniknout databáze již zdigitalizovaných děl, aby se zabránilo případné duplikaci. Knihovny si to mohou ověřit buď přímo na stránkách Národní digitální knihovny nebo na stránkách Registru digitalizace[11].

V současné době se ČR řídí *Strategií digitalizace kulturního obsahu na léta 2013 – 2020*[12]. V dílčích cílech se plán zaměřuje na řešení otázky autorských práv, dlouhodobého uložení a zpřístupnění elektronického materiálu a v neposlední řadě integrování obsahu do knihovny *Europeana* [12].

1.2 ZRAKOVĚ POSTIŽENÍ ČTENÁŘI

Elektronické knihy asi nejvíce ocení zrakově postižení čtenáři. Elektronická zařízení, která jim knihu předloží, nabízí funkci zvětšování písma nebo hlasité předčítání textu ženským či mužským hlasem. Přestože syntéza řeči je v dnešní době velice kvalitní, nedokáže plně

[10] Video se záběry funkčnosti skeneru jsou ke zhlédnutí zde: https://www.youtube.com/watch?v=RROgUub-cDQ&feature=youtu.be
[11] Dostupné na: http://www.registrdigitalizace.cz/rdcz/
[12] Dostupné na adrese: https://www.mkcr.cz/strategie-evropa-2020-digitalizace-kulturniho-obsahu-831.html

vyvinout tempo čtení, které zvládá čtenář zrakem. Existuje adekvátní náhrada v podobě audiobooků, jejichž nabídka je však omezena rychlostí produkce, výrobními náklady a částečně i tematickým obsahem. Vzájemně se tak oba formáty navzájem doplňují.

Digitální knihovny pro čtenáře se zdravotním postižením (dále jen knihovny) shromažďují, uchovávají a poskytují svým uživatelům rozmnoženiny elektronický knih. I tyto knihovny podléhají autorskému zákonu. Vzhledem k tomu, že níže bude srovnána knihovna z ČR s digitální knihovnou Slovenské republiky (SR), je třeba zmínit znění autorského zákona obou zemí. Je zajímavé, že oba státy mají licence děl pro potřeby zdravotně postižených občanů ošetřeny téměř totožně. Ať už v § 38 českého Autorského zákona 121/2000 Sb.[13] nebo v § 46 slovenského Autorského zákona č. 185/2015 Z.z.[14] je zde specifikován způsob veřejného rozšiřování právě pro potřeby zdravotně postižených. Není zapotřebí souhlas autora a ani nevzniká povinnost uhrazení odměny. Podmínkou však je nulový přímý či nepřímý majetkový prospěch.

Pro srovnání funkce, přístupnosti a obsahu byli vybráni zástupci jedné české a jedné slovenské digitální knihovny pro zrakově postižené. Za Českou republiku to je projekt Knihovna digitálních dokumentů (KDD) a za Slovensko projekt DigiBooks (www.digibooks.sk) založený občanským sdružením Infoblind.

1.2.1 KNIHOVNA DIGITÁLNÍCH DOKUMENTŮ

Knihovna digitálních dokumentů byla založena už v roce 1993 pod patronátem organizace Sjednocené organizace nevidomých a slabozrakých (SONS). O deset let později byla přeměněna na knihovnu plně odpovídající klasických standardům[15] a v roce 2010 proběhla její poslední modernizace[16]. Vylepšili uživatelský systém, který je daleko přehlednější s funkcí vypnutí CSS stylů, právě kvůli specializovaným počítačům pro nevidomé. Dále bylo doladěno vyhledávání publikací, listování v knihách a sledování požadovaných knih.

Do KDD se může registrovat pouze čtenář, který splňuje následující podmínky:

- musí být občanem České republiky starším patnácti let,

[13] Plné znění dostupné na: https://www.zakonyprolidi.cz/cs/2000-121#cast1
[14] Plné znění dostupné na: http://www.zakonypreludi.sk/zz/2015-185
[15] Starší verze digitální knihovny je stále přístupná na adrese http://knihovna.braillnet.cz/
[16] Zatím poslední verze knihovního systému je na http://www.kdd.cz/

- musí být držitelem karty ZTP/P nebo ZTP vydaného v ČR,
- dále musí umět pracovat s PC se speciálním programovým vybavením pro zrakově postižené a musí mít přístup k internetu.

Zdravý uživatel, ani jako zákonný zástupce zdravotně postiženého, se do knihovny registrovat nemůže. Obsah knihovny je zajišťován organizací SONS ve spolupráci s 90 českými nakladatelstvími, jako jsou např. Academia, Baronet, Portál, Grada, Euromedia Group či Metafora. Dalším zdrojem je samotná digitalizace (skenování) s převodem do prostého textu a v některých případech je jako zdroj uveden i internet.

1.3 VĚDECKÁ ČINNOST A VZDĚLÁVÁNÍ

V předchozí kapitole byly zmíněny různé vládní strategie podporující digitalizaci kulturního a historického dědictví, vytváření e-Govermentu atd. Ministerstvo školství, mládeže a tělovýchovy České republiky (MŠMT) vytvořilo *Strategii digitálního vzdělávání do roku 2020* (dále Strategie). Tento dokument je součástí Strategie vzdělávací politiky ČR do roku 2020 a vládou byla Strategie přijata v listopadu 2014 jako usnesení vlády ČR č. 927/2014[17]. Vláda tak reaguje na změny související s rozvojem digitálních technologií, zejména proto, aby studenti zlepšili své postavení na českém i mezinárodním trhu práce.

Ve Strategii jsou nastíněny tři hlavní cíle, mezi něž patří zlepšení kompetencí žáků v oblasti práce s informacemi, rozvíjení jejich informatického myšlení a zavedení nové metody a způsobů vzdělávání pomocí digitálních technologií. K naplnění cílů je třeba zajistit podmínky pro rozvoj kompetencí nejen žáků, ale i samotných pedagogů, zajistit systém podpory školám i obnovu vzdělávací infrastruktury. Z výzkumů uvedených ve Strategii vyplynulo, že ačkoliv jsou školy dostatečně technologicky vybaveny, využití ve výuce je značně omezeno – ať už je to dáno bariérami pedagoga (strach z nových technologií, nedostatek času se s nimi seznamovat atd.), nebo překážkami uváděnými školami (chybějící finance na inovace, nedostatek profesionálních správců ICT) atd. Přesto se velkou vizí do budoucna stalo propojení školního výukového prostředí s učebním prostředím žáků. Cílem by se pak mělo stát takové otevřené školní učební prostředí, které bude prostupné na všech úrovních a pro všechny [24].

[17] Dostupné na adrese: https://apps.odok.cz/attachment/-/down/VPRA9R9BDBJX

1.3.1 DIGIBOOKS

Knihovna Digibooks[18] vznikla na přelomu let 2003 a 2004, jako reakce na nepřístupný obsah KDD pro občany Slovenské republiky. Peter Grosser tehdy založil Občianske združenie INFOBLIND, které se zaměřilo na zpřístupňování informací nevidomým. Záměrem projektu digibooks se stal:

- převod knih do elektronické podoby,
- zpřístupnění těchto knih nevidomým prostřednictvím uživatelsky přívětivého internetového portálu,
- vytvoření prostředí pro vzájemné sdělování zkušeností při práci s PC i samotnou digitalizací.

Chod projektu je placen členskými příspěvky ve výši třinácti eur na jeden kalendářní rok, ale především pak příspěvkem sponzorů, kteří si odepíší 2 % z daní. Členem se může stát:

- občan SR i ČR, pokud je majitelem průkazu ZTP, ZTP/S a ZTP/P a zaplatí roční členský poplatek,
- občan SR i ČR bez průkazu ZTP do funkce korektora vázaného platnou smlouvou a zaplatí roční členský poplatek.

Obsah knihovny je naplňován jejími členy, kteří podle výše uvedené litery autorského zákona mohou vytvářet rozmnoženiny literárních děl. Další uživatelé pomáhají korekturami (opravami chyb v textu), doplňováním databáze, sháněním e-booků z dalších dostupných zdrojů. Mezi nevidomými členy převažuje slovenská národnost, ale překvapivě mezi korektory dominují Češi. Nové verze e-booků s vyšším stupněm korektury činí ročně až 15 % přírůstků, které se nezapočítávají do počtu titulů. V současné době začalo sdružení INFOBLIND spolupracovat s firmou Corvus a plánuje se spuštění nové verze knihovny, tentokrát i s použitím CSS stylů.

Obě knihovny jsou provozovány na základě obdobných autorských zákonů. Po obsahové stránce zahrnují monografie, encyklopedie i různá periodika. Jak ale pozorujeme v tabulce 1, množství zpřístupněných e-booků jednoznačně hovoří ve prospěch otevřené spolupráce v projektu Digibooks. Rozdíl v počtu členů mezi oběma knihovnami může být způsoben odlišným financováním. Čeští nevidomí mají členství v KDD placeno státem, v Digibooks může členský poplatek přeci jen vytvářet přístupovou bariéru. Dalším faktorem může být i nízká informovanost o existenci projektu Digibooks.sk na území ČR.

[18] Knihovna je dostupná na adrese: https://www.digibooks.sk

Zcela nepochopitelným se u knihovny KDD jeví věkové omezení uživatele. Jako by mladší nevidomí čtenáři nečetli nebo neuměli pracovat s počítačem a speciálním softwarem. Nastavení pravidel limituje zejména vidomé rodiče, protože je vyloučená jejich registrace – nejsou držiteli karty ZTP.

Tabulka 1 - srovnání digitálních knihoven KDD.cz a Digibooks.sk

	KDD.cz	DIGIBOOKS.sk
Počet členů / korektoři	1802 / 0	550 / cca 180
Národnost členů	česká	slovenská a česká
Věkové omezení	od 15 let	žádné
Financování	stát, sponzoři	členské příspěvky, sponzoři
Počet titulů k roku 2017	34 860 ks	106 064 ks / cca 82 000 v cz

2 PRINCIPY PŘEVODU TEXTU DO DIGITÁLNÍ PODOBY

V současné době je technologie automatického optického rozpoznávání obrazových dat běžnou součástí života člověka. Ať už se jedná o čárové kódy a QR kódy obsahující jednoznačné prvky identifikace a případně další informace, nebo o automatické vyhodnocování dotazníků na základě přesně umístěných značek v dokumentu. Technologie OCR dokáže rozpoznat rukou psaný či tištěný text a při převodu do strojem editovatelného prostředí zachovává formátování původního zdroje. Jmenované technologie jsou využívány v různých odvětvích od zdravotnictví (rozbor krevního obrazu), astronomie (popis vesmírných těles na základě různých typů snímkování), nebo v zemědělství (analýza úrodnosti v dané oblasti na základě leteckých snímků) [33][34]. Vývoj OCR systémů ale trval celé století, než se technologie dostala na současnou úroveň.

2.1 VÝVOJ V OBLASTI OCR

První pokusy proběhly již na konci 19. století, kdy se objevil první sítnicový skener, jehož přenosová soustava byla tvořena sadou fotobuněk. V roce 1912 byl vynalezen optophone. Ruční skener, který na základě tištěného znaku vytvořil tón odpovídající konkrétnímu symbolu. Už v roce 1929 v Německu si Gustav Tauschek sestavil přístroj využívající OCR. Jeho mechanický stroj pracoval se šablonou a fotosnímačem. Když se šablona překryla s obrazem, na tento stav reagoval fotosnímač vysláním signálu, že se jedná o stejný znak.

Po roce 1950 se vývoj OCR programů rapidně zrychlil. Strojové čtení již bylo na takové úrovni a zvládalo zpracovávat dostatečný objem dat, že se v běžném provozu začaly objevovat první OCR systémy. V roce 1954 Reader's Digest zavedl do praxe OCR systém zpracovávající informace z děrných štítků. Obsahem převodu byly finanční zprávy psané tiskacími stroji [27].

Následující vývoj OCR se dá rozdělit podle použité technologie do několika generací.

- první generace – speciální druhy písma pro OCR,
- druhá generace – rozpoznání ručně psaných čísel, definování standardů znakových sad,
- třetí generace – vývoj algoritmů pro rozpoznávání textu ze zdrojů nižší kvality,
- čtvrtá generace – zvyšování počtu znakových sad pro různé jazykové rodiny (např. cyrilice), vývoj multiplatformních OCR systémů.

2.1.1 PRVNÍ GENERACE OCR

Roky mezi 1960 – 1965 jsou označovány pro komerční sféru OCR systémů jako první generace. Stroje se omezovaly pouze na čtení pro ně přímo navržených znaků, které však vypadaly nepřirozeně. V následujících letech se začaly objevovat stroje, které dokázaly pracovat až s deseti druhy písma (fonty). Jejich počet byl omezen, protože stroje stále pracovaly na bázi párování šablon, které porovnávalo obraz znaku s knihovnou vzorových obrazů znaků každého písma [30].

2.1.2 DRUHÁ GENERACE OCR

Druhá generace OCR systémů spadá do období mezi lety 1965 až 1970. Nové systémy si už poradily s vyšším počtem fontů a začaly pracovat s jednoduchým rozpoznáváním ručně psaného textu – zejména detekovaly sadu číslic. V roce 1965 byl na Světové výstavě v New Yorku představen nový OCR systém IBM 1287. Ve stejném roce pak firma Toshiba vyvinula první automatický stroj na třídění dopisů, který se orientoval podle směrovacího čísla.

V roce 1966 vyšla standardizovaná znaková sada, která pro americký trh byla nazvána OCR-A, pro evropský trh OCR-B. Obě sady byly navrženy a stylizovány tak, aby bylo usnadněno optické rozpoznávání a stále svým přirozenějším tvarem čitelnější pro člověka [30]. Obě verze se pro porovnání nachází v příloze 1.

2.1.3 TŘETÍ GENERACE OCR

Po roce 1970 se vývoj systémů OCR zaměřil především na kvalitu rozpoznávání předloh nízké kvality – především ručně psaných znaků, dále na zvyšování výkonu a snižování ceny. Tyto jednoduché OCR systémy, které především ještě stále využívaly neproporcionální písma, jejichž pevně stanovená šířka zvyšovala kvalitu převodu. Koncepty psané psacími stroji, pak byly převedeny do počítače ke konečné úpravě. Zmíněný systém se využíval zejména ve zdravotnictví, žurnalistice, na poštovních úřadech atd.

2.1.4 ČTVRTÁ GENERACE A SOUČASNOST

Ke konci 70. let 20. století dokázal komerční software Kurzweil Reading Machine rozpoznat prakticky libovolný font a zároveň prováděl syntézu textu a řeči. Nahlas tak nevidomému uživateli předčítal jakýkoliv dokument. V 80. letech byl pro Ministerstvo

zahraničí USA vyvinut první skener na čtení cestovních pasů. V obchodech se začal využívat skener pro čtení cenovek zboží. V roce 1989 vstoupila na trh se svým OCR produktem ruská společnost Abbyy. Jejím cílem bylo vytvoření programového prostředí, které bude uživatelsky přívětivé a jednoduše převede tištěný text do digitální podoby.

Po roce 2000 je už technologie OCR k dispozici široké veřejnosti. Většina prodávaných skenerů i multifunkčních zařízení je dnes vybavena alespoň základními OCR programy. Mobilní zařízení disponují tak výkonnými procesory a vysoce kvalitními fotoaparáty, že byly vyvinuty aplikace OCR umožňující převod přímo v telefonu. Zpočátku takové aplikace pracovaly se strojovým textem (různé překladače), později už plnohodnotně převáděly telefonem pořízené fotografie obsahující symboly na text.

Běžný uživatel si v současnosti ke své práci může vybrat z nabídky komerčních i freewarových programů. Mezi vysoce kvalitní komerční programy patří např. Abbyy Finereader, se kterým se v této práci blíže seznámíme.

2.2 METODY OCR

Z předchozí kapitoly vyplynulo, že existuje vícero druhů metod zpracování tištěného či psaného textu, které se liší svým zaměřením a technologií postupu. Opomineme-li klasické zadávání dat do počítače klávesnicí, což je pro zpracování velkého objemu dat neefektivní, zaměříme se na metody automatického rozpoznávání informací.

2.2.1 OPTICAL MARK RECOGNITION (OMR)

OMR je metoda, která se využívá při vyhodnocování speciálních formulářů. Funguje na bázi prosvěcování papírového podkladu. Tam, kde je ve formuláři odpověď vyplněna (začerněna), zmenšuje se v daném místě průchod světla. Tmavé části jsou na předem definovaném prostoru detekovány a vyhodnoceny. Velkou výhodou OMR je značná přesnost a jednoduchost. Prakticky se tento princip používá při spotřebitelských výzkumech, testování a loteriích. Na stejném principu funguje i čtení čárových a QR kódů, se kterými se uživatel setkává v běžném životě nejčastěji [30].

Čárové kódy jsou tvořeny sledem světlých mezer a tmavých čar různé šířky. Laserové snímače vysílají červené světlo, které černá barva absorbuje, ale bílá ho odráží. Snímač zaznamenává rozdíly v odrazu a přeměňuje je v elektrické signály. Tyto příznaky jsou převáděny na číslice a písmena a jsou srovnávány se sadou schválených kombinací [35].

Obrázek 4 - Ukázka čárového kódu EAN (Zdroj [36])

QR kód může obsahovat více informací než čárový kód, protože je dvoudimenzionální. Obsahuje oddíly černých a bílých čtverců tvořících geometrické obrazce nesoucí dvě vrstvy informací. Geometrická vrstva lokalizuje bity nesoucí informace a informační vrstvu se samotnými daty. Nečitelným se stává až po poškození velké části kódu [36].

Obrázek 5 - Ukázka QR kódu url Západočeské univerzity (Zdroj: [])

2.2.2 OPTICAL CHARACTER RECOGNITION (OCR)

Optické rozpoznávání znaků je metoda, která převádí tištěný znak na digitální. Pracuje na principu porovnávání bitmapového obrazu znaku s databází, ve které je daný znak zaznamenán a doplněn o význam. Kvalita výsledného převodu se odvíjí od kvality zdrojového dokumentu. Stále však platí, že žádný OCR systém není bezchybný. Více v kapitole o fázích procesu rozpoznávání znaků.

2.2.3 INTELLIGENT CHARACTER RECOGNITION (ICR)

ICR je specifičtější technologie OCR zaměřující se na rozklíčování různých typů rukopisu. Obsahuje tzv. self-learning, algoritmus, který pomáhá programu se učit. To vede k urychlování převodu a minimalizování podílu uživatele na úpravě. Program označí neznámý znak, uživatel přiřadí a potvrdí hodnotu symbolu ze známé sady, kterou program dokáže přeložit. Při příštím rozpoznávání jej program správně vyhodnotí. Tato technologie je založena na konceptu neuronových sítí, přesto její úspěšnost není úplně 100% [37].

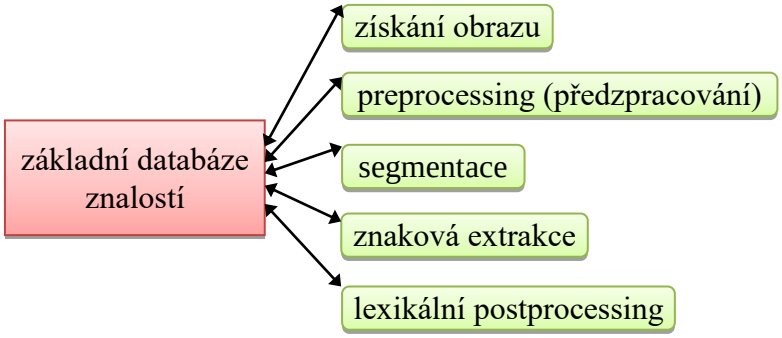
2.2.4 INTELLIGENT WORD RECOGNITIONS (IWR)

Inteligentní rozpoznávání slov tvoří další stupeň vývoje OCR. Dokáže rozpoznat ručně psané znaky včetně dalších vlastností tištěného písma jako je sklon, tučnost atd. IWR nepracuje se samostatnými symboly, ale rozpoznává celá slova dokonce i fráze.

Technologie vznikla opět za účelem minimalizování ruční práce uživatele s následnými opravami chybně rozpoznaných symbolů [37].

2.3 FÁZE PROCESU ROZPOZNÁVÁNÍ ZNAKŮ

Proces OCR převodu obsahuje komplex podprocesů, jejichž cílem je určit uspořádání dokumentu, aby ve výsledku vznikl dokument co nejvěrněji kopírující předlohu. Základ systému tvoří jakási databáze znalostí, která komunikuje se všemi podprocesy a předává jejich výstupy následující části procesu. Nejdříve se získá obraz z přímého zdroje (skener, fotoaparát) nebo nepřímého zdroje (internet). Získaný obraz je předzpracován optimalizačními nástroji, jenž eliminují kazy předlohy. Následuje proces segmentace, který vyhodnocuje oblasti jako je text, tabulka, obrázek a následně analyzuje každý symbol. Identita každého symbolu se porovná s jeho popisem ve znakové sadě, načež dochází k jeho klasifikaci a rekonstrukci slova.



Obrázek 6 - schéma návaznosti procesů v OCR softwaru

2.3.1 ZÍSKÁNÍ OBRAZU (OPTICKÉ SKENOVÁNÍ)

Podproces získávání obrazu zajišťuje vlastní vstup dat do procesu. Obraz papírového dokumentu načteme do procesu jako výstup z externích zařízení typu skener, digitální fotoaparát či videokamera. Obvykle je tento podproces součástí OCR programu jako samostatný software. Obrazový soubor lze získat i nepřímou cestou – stažením z internetu.

Výsledek procesu OCR je však velmi silně ovlivněn kvalitou získaného obrazu. Při skenování je třeba sledovat několik důležitých parametrů – hodnotu DPI (tedy počet obrazových bodů na palec) a nastavení kontrastu. V nativním prostředí skeneru nalezneme bližší specifikace pro jednotlivé typy dokumentů na základě kvality papíru (časopis, fotografie, noviny atd.), a typu obrazu (barevný, šedý, černobílý).

2.3.2 PREPROCESSING

Kvalita vstupního obrazu může být rozdílná. Různé kazy, šumy, nerovnoměrné osvětlení, různé pokřivení, zvlnění obrazu následně vedou ke snížení kvality výstupu po provedení OCR. Systém tak má zabudovaný podproces předzpracování obrazu, který má za úkol zmíněné vady eliminovat. Lze využít automatické funkce předběžného zpracování již během skenování, kdy program obraz převede do digitální podoby a upraví jej podle parametrů u dokumentů s nejvyšší úspěšností OCR. Jenže ani tato úprava nemusí stačit, takže program nabízí uživateli grafický editor, kde snímky může dál upravovat. Má možnost znovu srovnat snímek, narovnat řádky s textem, odstranit pohybový efekt, šum, zvýšit jas a kontrast či dodatečně zmenšit rozpoznávanou plochu ořezem atd.

Nejčastěji se v preprocessingu využívá filtrace obrazu. Obraz je tvořen mřížkou bodů nesoucích informace o barvě či jasu. Grafické filtry s těmito informacemi pracují. Proces pracuje se dvěma typy filtrů a to lineárními a nelineárními. První skupina pracuje se změnou světla v prostoru a její nevýhoda spočívá v určité degradaci obrazu. Při použití filtru horní propusti se zvýrazní detaily, ale zároveň se zvýší šum. Použijeme-li filtr dolní propusti – tzv. vyhlazovací filtr (Gaussův) – dojde sice k odstranění šumu, ale zároveň se obraz rozostří, čímž dochází ke snížení schopnosti rozpoznání hran při následujícím procesu segmentace.

Nelineární filtry vybírají jasovou úroveň podle okolí zvoleného bodu a danou hodnotou vybraný bod přepíší. Do těchto filtrů patří určování minima, maxima a mediánu. Poslední zmíněný je využívám asi nejčastěji. Jak vyplývá z názvu filtru, při nahrazování bodů volí střední hodnotu jasu v okolí, čímž má minimální dopad na ostatní část obrazu. Je tak vhodným nástrojem na odstraňování šumu.

2.3.3 SEGMENTACE

Segmentace tvoří jeden z hlavních pilířů OCR převodu. Jejím úkolem je rozpoznávat objekty nacházející se v popředí snímku a odlišit je od pozadí. K tomu využívá jednu ze základních metod jakou je prahování. Tzv. práh je vhodně zvolená hodnota jasu v obraze a bod, který se nachází pod nebo nad zvoleným prahem, je pak klasifikován jako popředí či pozadí obrazu. Tato metoda je samozřejmě vhodná pro optimálně nasvícený snímek. Pokud je k dispozici nerovnoměrně osvětlený obraz, je třeba použít adaptivní či

víceúrovňové prahování, kdy se obraz rozdělí na několik menších oblastí, pro které je hodnota prahu stanovována individuálně.

Na základě prahování se naskenovaný obraz rozčlení na oblasti a rozliší, zda se jedná o textové, grafické nebo tabulkové bloky. Dále určí vztahy mezi těmito seskupeními, např. rozhodne o pořadí čtení jednotlivých oddílů v dokumentu. Po rozčlenění proběhne segmentace řádků, slov a nakonec znaků. V první řadě program nachází souvislou řadu symbolů, která je od další řady symbolů oddělená souvislým pruhem světlejšího vodorovného pozadí. Takže tam, kde se hodnota jasu bodů blíží k nule, je stanoven horizontální předěl mezi jednotlivými řádky.

Pak se zaměří na detekování jednotlivých a souvislých základních znaků, rozliší diakritiku a interpunkční znaménka. Znaky se do skupin rozdělí podle své výšky a svislé pozice v řádku za pomoci referenčních linek tzv. dotažnic. V horní části se tak nachází diakritika a horní přetahy písmen, jako mají písmena b, d, t, atd. Ve spodní části detekuje dolní přetahy písmen (p, j, y, g) a interpunkci. Dělení řádku na jednotlivá slova vychází z podobného principu jako rozpoznávání řádků na stránce. Detekce se orientuje podle velikosti mezery mezi znaky a slovy, která je buď porovnávána se stanovenou hodnotou prahu, nebo na základě výpočetního algoritmu hodnot histogramu kvůli nepravidelnosti mezer mezi slovy v jiném řádku.

Při segmentaci může nastat několik faktorů, které proces naruší. Jedná se o špatnou čitelnost znaků, na jejímž základě nelze vytvořit správný popis znaku. Tečky v interpunkci i diakritice mohou označeny za šum a při segmentaci jsou ignorovány. Části textu, které se vymykají standardní velikosti řádku v textu, mohou být označeny za grafiku a naopak, část obrazového souboru může být „přeložena“ do textové podoby. Každopádně výsledkem segmentace je informace o velikosti, šířce a pozici znaku v textu.

2.3.4 ZNAKOVÁ EXTRAKCE

Výsledkem procesu segmentace je tedy jakási matice hodnot znaku a znaková extrakce jej na základě popsaných vlastností začne přiřazovat k odpovídajícímu vzoru v databázi. Sleduje se pět hlavních faktorů:

- šum (citlivost na další rušivé segmenty),
- pokřivení (rozšíření, smrštění textu),

- variace stylu (různé formy zobrazení jednoho symbolu různými fonty),
- posunutí (horní, dolní index),
- rotace (orientace symbolu).

Kvalita této extrakce je přímo úměrná počtu znaků, kterými je symbol popsán. K porovnání výše zmiňovaných znaků s vnitřní vzorovou databází OCR systém využívá několik metod. Jednou z nich je nalezení shody. Každý již známý symbol má svou binární šablonu, s níž je rozpoznávaný znak srovnáván. Míra shody se zaznamenává a šablona s největším počtem shodných pixelů je označena za rozpoznaný znak. Vzhledem k množství variant je tato metoda pomalá a navíc málo variabilní. Různé znázornění znaku odlišnými fonty zvyšuje čas rozpoznání. Další metodou je statistická klasifikace pracující s pravděpodobností. Na základě příznaků je vypočítána pravděpodobnost výskytu v dané třídě znaků a poté je symbol zařazen do třídy s nejpodobnějšími vlastnostmi.

2.3.5 LEXIKÁLNÍ POSTPROCESSING

Poslední fází automatického systému rozpoznávání OCR je lexikální postprocessing. Když člověk narazí na slovo, které je nečitelné, k jeho rozpoznání ho dovede pochopení kontextu, ve kterém je výraz použit. Stejným způsobem pracuje zmíněný podproces. Je silně ovlivněn podporou daného jazyka v OCR programu a vytvořeným slovníkem. Pracuje i na základě statistických modelů výskytu slov v daném jazyce. Slovník nejčastěji využívá k prohledávání prefixový strom. V uzlech se nacházejí slabiky a slova se stejnou předponou mají společný začátek cesty. Tento způsob značně zmenšuje nároky na paměť a urychluje prohledávání slovníku. K dalšímu urychlení dojde po rozdělení hlavního slovníku na menší slovníky obsahující slova s podobným základem a délkou. Opět prohledávání využívám strukturu prefixového stromu. Lexikální processing pomocí databáze vybere nejvhodnější slovo a nahradí jej v digitálním textu.

Výše popsaný technologický postup OCR převodu je pouze nástinem, jak obecně proces OCR pracuje. Je tvořen mnoha algoritmy, které zde rozepisovány nebyly. Podrobnější popis naleznete v publikacích [40][41][42].

3 FAKTORY PODÍLEJÍCÍ SE NA KVALITĚ OCR PŘEVODU

Než se uživatel pustí do digitalizace vybrané monografie či jiného textového dokumentu, musí si uvědomit několik skutečností. Hned zpočátku je důležité si stanovit, za jakým účelem e-book vzniká, jaká je požadovaná kvalita výsledného formátu a kolik času může strávit nad digitalizací jednoho díla. Od těchto cílů se odvíjejí prostředky, které autor digitalizace při převodu použije. Mezi ně rozhodně patří způsob získání zdrojového obrazu – zda stačí uživateli fotografie nebo potřebuje kvalitní naskenované obrazy. Dále záleží na předloze samotné – od použité vazby až po typografii. Posledním faktorem je pak výběr samotného ať už komerčního nebo volně dostupného softwaru. Nyní se podíváme na jednotlivé faktory detailněji.

3.1 PROSTŘEDKY ZÍSKÁVÁNÍ OBRAZU

Bez nadsázky se dá kvalita zdrojového obrazu označit za alfu a omegu kvalitního převodu, závisí na něm chybovost rozpoznávaného textu. Mezi nejběžnější postupy získávání obrazu patří fotografování nebo skenování tištěného dokumentu.

3.1.1 FOTOGRAFOVÁNÍ DOKUMENTU

Při pořizování kvalitního obrazu pomocí fotoaparátu řešíme čtyři podmínky:

- kvalita a nastavení přístroje (mobilní telefon, profesionální digitální zrcadlovka),
- rovnoměrné rozložení světla a úhel dopadu na fotografovanou plochu,
- manipulace s předlohou,
- strávený čas.

Pokud budeme uvažovat o standardní velikosti předlohy do formátu A4, uživatelská příručka FineReader Abbyy doporučuje obrazový snímač s 5 miliony body, minimálně pak se 2 miliony pixelů. Kvalita obrazu se odvíjí od velikosti snímače a množství megapixelů v něm obsažených. Vysoké rozlišení na malé ploše snímače pak způsobuje vyšší šum výsledné fotografie.

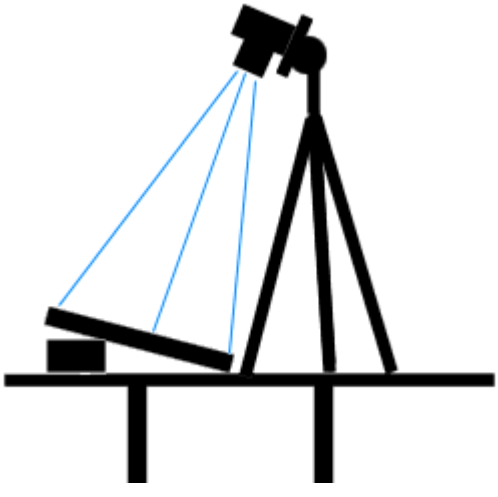
Jestliže bude zpracováván objemnější dokument, je lepší použít stativ. Zajistí stabilní pozici fotoaparátu s objektivem zaměřeným na střed dokumentu nejlépe nastavený kolmo k předloze. Přístroj by měl být nastaven nejlépe na manuální ovládání clony, optického zoomu, zaostřování a blesku. Pro získání ostřejšího obrazu je doporučováno zaostření do prostoru mezi středem a okrajem dokumentu. Zaostření na střed by

způsobovalo zkreslení okrajů stránky. Nastavení hodnoty ISO by se kvůli šumu mělo nastavit na co nejnižší hodnoty. Hodnota clony se mění v závislosti na tmavosti obrazu, když je temný, je lepší nastavit menší hodnotu. Některé fotoaparáty dokonce nabízejí režim Dokument k fotografování textu.

Světlo je velice významným faktorem ovlivňujícím kvalitu obrazu. Cíleným požadavkem je rovnoměrné nasvícení dokumentu, kterého lze nejlépe dosáhnout za dne. Pokud z nějakého důvodu nelze denní světlo využít, je nutné k osvětlení použít buď jedno silné světlo umístěné přímo nad dokumentem, nebo dvě menší světla umístěna naproti sobě. Zamezí se tím tvorbě nežádoucích stínů. Při fotografování se nedoporučuje využívat blesk, protože pak dojde k přepálení oblasti a zároveň k nerovnoměrnému osvětlení, zejména u lesklého podkladu.

Manipulace s dokumentem je dalším faktorem, který uživatel musí řešit. Největší potíže způsobuje fotografování knihy, protože stránky mají tendenci se samy otáčet. Jedním ze způsobů, jak tomu zabránit, je zatížit strany průhledným sklem – čímž se strany ještě více vyrovnají, nebo například použít klipsy. Přesto právě tato manipulace navyšuje čas strávený fotografováním dokumentu.

To je možná nejdiskutovanější část procesu, protože na jednu stranu má uživatel fotoaparát neustále k dispozici (zejména v mobilních telefonech), může další čas ušetřit na vypnutí funkce autofokus, protože stačí zaostřit pouze jednou. Nastaví samospoušť podle rychlosti, jakou zvládá otáčet stránky dokumentu. Na druhou stranu právě manipulace s dokumentem bude nejvýraznější problém [32][36].



Obrázek 7 - funkce stativu při fotografování dokumentu (Zdroj [36])

3.1.2 SKENOVÁNÍ DOKUMENTU

Druhou zmiňovanou formou získávání obrazu je skenování. Správný výběr skeneru je podstatný. Zajímáme se o tyto parametry:

- typ skeneru (stolní či ruční)
- typ snímače
- velikost skenovací plochy
- rychlost skenování
- pořizovací cena.

Když opomineme velkoformátové skenovací stroje, které mají k dispozici knihovny nebo velké firmy, tak běžný uživatel ve své praxi použije nejčastěji stolní, ruční nebo řádkový skener. Opět záleží na velikosti skenovaného dokumentu, aby k jeho skenování zvolil vhodný nástroj. Ruční a řádkové skenery se nejvíce využijí u skenování krátkého článku či jen úryvků v textu.

Stolní skenery se tedy ve velké míře využijí ke skenování vícestránkových dokumentů, protože velikost jejich skenovací plochy dosahuje standardní velikosti A4 až A3. Tudíž se klasický formát publikace může skenovat po dvoustranách, čímž dochází k úspoře času stráveného skenováním. Je třeba si uvědomit, že skenovací plocha (plocha, ze které sensor snímá předlohu) není totožná s velikostí skla. Většinou je plocha na každé straně až o 1 cm menší.

V závislosti na předloze je pro uživatele důležitým faktorem skenování integrovaný snímač. V současnosti většina stolních skenerů má zabudovaný tzv. CCD snímač (Charge-coupled device) nebo CIS snímač (Contact Image Sensor). Technologie CCD snímače funguje na bázi odraženého světla z katodové lampy pomocí soustavy čoček, zrcadel a filtrů s RGB barvami. Snímač totiž nesnímá barvu, ale intenzitu světla. U CIS technologie vyzařují světlo tři řádky LED diod v barvách RGB, které jsou součástí čtecí hlavy s řadou senzorů.

Rozdíl v technologii snímání je rozpoznatelný zejména u předloh, kde je zapotřebí zvýšená rozlišovací schopnost tmavých částí. CIS snímač má problém s prostorovým snímáním, což se projevuje při zejména při skenování knih, kdy se hřbet knihy vzdálí od skenovací plochy a obraz této oblasti je pak tmavý až černý, jak lze pozorovat v příloze 2 v porovnání se skenem stejné publikace snímačem CCD.

Při samotném skenování, ať už použijeme skener s jakýmkoliv snímačem, je důležité nastavit si parametry skenování – jako je barva, rozlišení DPI a kontrast. Nejvíce je výrobci i dalšími zkušenými uživateli doporučováno, pro co nejkvalitnější rozpoznávání textu, nastavení šedé barvy a rozlišení 300 DPI. Musíme si uvědomit, že většina publikací není tištěna na bezchybném bílém křídovém papíru, ale na materiálu, který mívá malé mikroskopické kazy či nečistoty. Při volbě černobílého skenu bychom program při procesu segmentace mátli a výsledkem by opět byla větší chybovost.

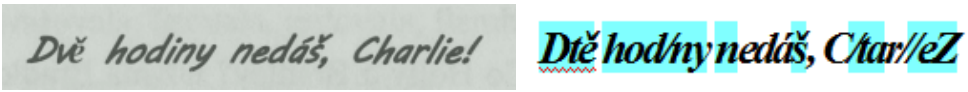
Nastavení rozlišení je neméně důležitou součástí. Pro skenování běžného textu je standardně doporučováno nastavení hodnoty DPI na 300 bodů v šedých odstínech. Hustota bodů je pro OCR proces dostatečná a rozpoznávací algoritmus má dostatečné množství jasových prahů pro segmentaci. Software sám dokáže určit, jestli použité skeny mají dostatečné rozlišení pro optimální čtení. Pokud je výskyt rozpoznávacích chyb vyšší, sám většinou stránku označí a doporučí k přeskenování ve vyšším rozlišením 600 DPI. V případě běžného textu vyšší rozlišení nejen zvětšuje velikost obrazového souboru, ale ani nedochází k výraznému zmenšení počtu OCR chyb.

3.2 PŘEDLOHA

I při zpracovávání kvalitního digitálního obrazu papírového dokumentu může uživatel narazit na zvýšenou chybovost. Je dána typem vazby a použitou typografií. Pevnost vazby silně ovlivňuje přilnutí stránky dokumentu ke skenovací ploše. I přesto, že při skenování využijeme CCD snímač, tak zejména u silné knihy jakmile sklon řádky přesáhne více než 20 stupňů, kvalita OCR značně klesá [32]. Možným řešením je silnější přitlačení předlohy k ploše, které by sice úhel o několik stupňů snížilo, ale zvýšila by se tím možnost poškození vazby. Extrémním řešením, zejména u skenerů s automatickým podavačem, bývá úplné odříznutí vazby, aby předloha plně dolehla ke sklu skeneru.

Dalším faktorem je užitá typografie v dokumentu. Zejména se jedná o použitý font a zvolené řádkování. Velké úskalí se skrývá v bezpatkovém typu písma, zejména u znaků majících velice podobný tvar – např. ve slově Illionois. Dále při zúžení mezer mezi znaky může dojít spojení dvou znaků do jednoho – např. slovo trn může aplikace OCR rozpoznat jako tm. Samostatnou kapitolu tvoří speciální fonty – ozdobné, historické či rukopisné. Většina OCR programů má ve své databázi širokou základnu druhů písma k rozpoznání,

ale přesto nemohou postihnout všechny existující fonty. V případě, že využívaný software písmo nezná a nepodporuje výuku nového fontu, nezbývá uživateli nic jiného, než danou oblast chybně rozpoznaných znaků ručně opravit.



Obrázek 8 - ukázka kvality OCR neznámého rukopisného písma (Zdroj: vlastní)

V některých případech způsobuje problémy atypické řádkování v publikaci. V kapitole 2.3.3 je rozebírán postup segmentace, tedy rozpoznávání řádků a znaků v dokumentu na základě algoritmu rozlišujícího souvislé světlé a tmavé plochy. Jakmile je světlá plocha mezi diakritikou a písmenem příliš široká, může ji program OCR označit za nový řádek a diakritiku rozpozná jako písmena. Slovo ŘÁD by tak bylo rozděleno do dvou řádků – háček jako písmeno v, čárku jako písmeno i v horním řádku, RAD jako slovo bez diakritiky o řádek níž.

3.3 SOFTWARE

Většina uživatelů stojí o to, aby program, který využívá, plnil svou funkci co nejlépe. Před pořízením jakéhokoliv softwaru je proto důležité si uvědomit, jak často jej bude uživatel využívat, jestli bude potřebovat všechny nabízené funkce, jaká je podpora českého jazyka atd. Naštěstí je k dispozici dostatečný počet OCR systémů, aby si každý našel to své.

3.3.1 KOMERČNÍ SOFTWARE

Komerční software primárně sloužil velkým firmám, kde zcela běžně probíhalo zpracovávání velkého objemu dat. Takové firmy mají vytvořené pracovní skupiny s rozdělenými činnostmi – od skenování po archivování dokumentů v databázi. Se snížením ceny hardwaru se staly komerční programy dostupné i pro běžné uživatele. Jedním z nejznámějších na českém trhu je ruský produkt FineReader Abbyy a americký OmniPage.

Abbyy FineReader

Ruská společnost byla založena v roce 1989 v Moskvě. Jejím zakladatelem byl David Young. V současnosti má po světě čtyři centrály, které se nacházejí v Kalifornii, v Mnichově, Kyjevě a Moskvě, přičemž v Moskvě probíhá hlavní vývoj. Databázi optického rozpoznávání již tvoří téměř 200 jazyků. Program je dostupný pro operační systém

Windows 7, 8, 8.1 a 10. Uživatelské prostředí je přívětivé a jednoduché, plně v českém jazyce. Má širokou nabídku výstupních formátů od docx, pdf, pptx, xmls, po podporovaný výstup v html s podporou kaskádových stylů. V současnosti na trhu existuje verze 14, jejíž cena se pohybuje okolo 5 000,- Kč. Mezi nejnovější funkce patří přehledné porovnávání dokumentů neomezené formátem, které jsou dostupné pouze ve verzi Corporate a Enterprise. Dále je zdokonalován modul uchovávání věrnosti zdrojového obrazu a podporován vývoj rozpoznávání jednořádkových matematických vzorců. Pro učitele tak může být zajímavá nabídka programu s cenou pohybující se okol 3 000,- Kč. Ta však nabízí jenom klasický modul OCR a úpravy a komentáře PDF.

Při instalaci a chodu potřebuje mít k dispozici alespoň 1 GB paměti RAM (doporučeny jsou 4 GB RAM) a vyžaduje 2,4 GB místa na pevném disku pro instalaci. [39]

Omnipage

Komerční software Omnipage je produktem firmy Nuance Communications, která je s OCR systémy spjata již velmi dlouho. Je vyvíjen pro operační systém Windows a podporuje okolo 120 jazyků. V současnosti je stále na trhu 18. verze, která je dostupná ve třech modelech – Standard, Ultimate a Server. Stejně jako FineReader Abbyy požaduje alespoň 1GHz procesor, ale nároky na paměť jsou nižší. Požadované minimum pro verzi Standard, která je cenově nejdostupnější pro běžného uživatele, je 512MB paměti RAM. Na instalaci a chod potřebuje též méně místa necelý 1 GB paměti pevného disku. Cena je stanovena na 150 dolarů, což se při současném kurzu pohybuje okolo 3 500,- Kč. Ani jedna z verzí nenabízí grafické rozhraní v českém jazyce, a speciální profesní slovníky také pro OCR v českém jazyce nejsou dostupné. Výčet výstupních formátů je ale téměř totožný s nabídkou Finereader Abbyy – docx, html či pdf. Autoři Omnipage dokonce tvrdí, že se jedná o nejrychlejší aplikaci na trhu s přesností OCR pohybující se okolo 99 % [40].

3.3.2 FREEWARE ON-LINE APLIKACE

Pro uživatele, kteří nepotřebují zpracovat mnohostránkový dokument a dokážou tolerovat o něco vyšší chybovost, existují online aplikace. Mezi ně patří *Onlineocr.net*, kdy velmi jednoduše bez jakékoliv registrace uživatel nahraje soubor ve formátech jpeg, PDF, bmp atd., z podporovaných 46 jazyků vybere češtinu, zvolí výstupní formát (docx, xml, txt) a spustí proces. Vše má přehledně na jednom místě, ale velikost souboru je omezena na 5 MB. Pokud potřebuje uživatel převést vícestránkový dokument, tak už je vyžadována

registrace. Dalšími podobnými projekty jsou newocr.com a free-ocr.com. Obě aplikace podporují český jazyk. Jako zdrojový obraz přijmou většinu obrazových souborů a poskytují totožné výstupní formáty – docx, pdf, txt. Je nasnadě, že takovéto programy kvalitě komerčních softwarů nemohou konkurovat, ale své uživatele si jistě našly.

4 POSTUP DIGITALIZACE TEXTU V PROGRAMU FINEREADER ABBYY 11

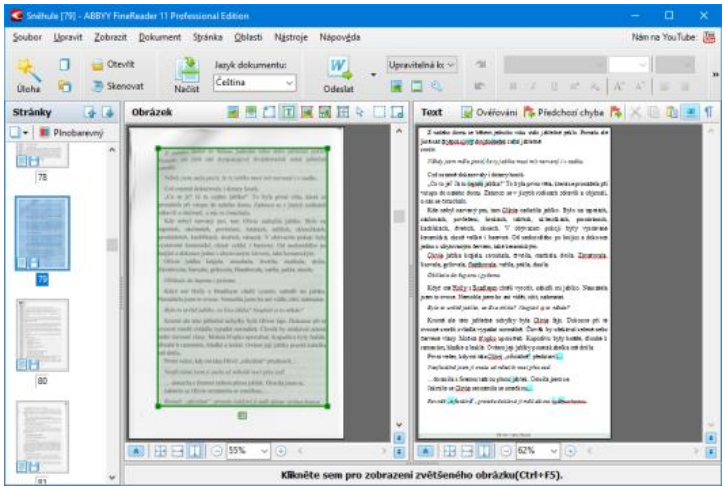
K vytvoření digitální kopie tištěného dokumentu jsem si vybrala software FineReader Abbyy (FRA) z toho důvodu, že jsou mezi uživateli často zmiňovány jeho kvality. Dále na mě působí uživatelsky velice přívětivě, podporuje velké množství jazyků, obsahuje výukový modul, je plně podporován v českém jazyce a vlastním programovou licenci.

Upozorňuji, že program je tak komplexní a mnohovrstevnatý, že nelze v této části práce plně popsat všechny dostupné nástroje, které k tvorbě elektronických publikací nabízí. V následujících podkapitolách představím jeden z možných návodů na vytvoření vlastního e-booku, který bude podpořený výukovými videi umístěnými na přiloženém DVD.

4.1 PRACOVNÍ PROSTŘEDÍ FINEREADER ABBYY 11

Pracovní prostředí programu je přehledně členěno do tří až čtyř oken zastřešených hlavním panelem nástrojů a panelem rychlého přístupu.

První okno (bráno z levé strany) zobrazuje všechny naskenované stránky v dokumentu. Má dvě volby zobrazení. Na snímku jsou zobrazeny tzv. miniatury na nichž je zvýrazněna stránka, se kterou se aktuálně pracuje. Druhou volbou je seznam, který uživatele informuje o množství chyb v převodu (procentuálně i množstevně). V horní části okna se v panelu nástrojů nachází nástroje k posunu stran v rámci dokumentu na jinou pozici.



Obrázek 9 - pracovní prostředí programu FineReader Abbyy 11

V prostředním okně s názvem Obrázek nalezneme zobrazení skenu v měřítku, v jakém si zvolíme. V horní části okna jsou panely nástrojů upravující typ oblasti a grafický editor. V dolní části se nalézají vlastnosti obrazu.

V pravém okně Text se zobrazuje výsledek procesu OCR. Nástroje v horní liště okna nabízejí kontrolu pravopisu a procházení chybně rozpoznaných znaků. V dolní části okna panel Vlastnosti textu informuje o rozložení dokumentu, použitém formátování textu atd.

4.2 SKENOVÁNÍ DOKUMENTU

Po spuštění programu se v uživatelském prostředí objeví úvodní obrazovka. Uživateli nabízí poloautomatické, nejčastěji používané úlohy. Tato nastavení vyžadují jen minimální zásah uživatele a to v podobě dodání zdrojového obrazu ke zpracování. Další úkony se již spustí automaticky podle daného algoritmu. Poslední zásah člověka do procesu je konečná fáze uložení digitálního dokumentu. Výsledný dokument dodržuje přesné rozložení zdrojového obrazu. Digitalizátor, který si raději volí způsob procesu sám, má možnost úvodní obrazovku vypnout a přejít rovnou do pracovního prostředí FRA.

Detailnější nastavení procesu skenování nalezneme v záložce Nástroje, volba Možnosti. V dialogovém okně vybereme kartu Skenovat/Otevřít viz obrázek 10. Ve spodní části okna vidíme zobrazenou volbu rozhraní, kde si uživatel vybere, zda ke skenování využije prostředí FRA nebo zvolí nativní rozhraní samotného skeneru. FRA disponuje nastavením režimu skenování (barevně, stupně šedi, černobíle), volbou rozlišení (DPI) a nastavením jasu, které lze nastavit automaticky i uživatelsky. Nativní rozhraní se většinou vybírá tehdy, pokud si digitalizátor zakládá na grafické kvalitě barevných zobrazení, má k dispozici více funkcí.

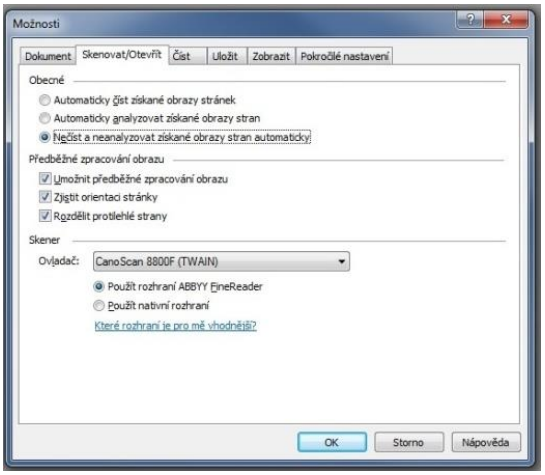
V oddílu Předběžné zpracování obrazu si uživatel nastaví, co se má s obrazem při skenování dít:

- umožnění předběžného zpracování obrazu (automatická vylepšení obrazu)
- zjistit orientaci stránky (automatické otočení dokumentu, nerozpozná u grafiky)
- rozdělit protilehlé strany (orientuje se podle nižšího jasu u hřbetu knihy)

Tyto volby lze nastavit i v dialogovém okně skenování v prostředí FRA. Jejich účinnost je vysoká, přesto jsou situace, se kterými si neporadí. Pokud je v dokumentu celostránková ilustrace nedokáže rozpoznat, zda není obraz otočený o 180°. U dokumentů, které mají text velmi blízko hřbetu, má problém detekovat hranici k rozdělení stran. Nebo na druhou stranu některé obálky knih rozdělí do několika částí. Všechny zmíněné potíže lze snadno

ručně upravit v editoru obrázků. Jen obálka musí být znovu naskenována s vypnutou volbou rozdělení stran.

V oddílu Obecné lze deaktivovat některé procesy, které by mohly vést ke zpomalení činnosti programu. Vypnutí vybraných funkcí zkracuje čas při skenování předlohy. Mám na mysli automatické čtení a analyzování dokumentu během skenování, protože je lze spustit samostatně po doskenování.

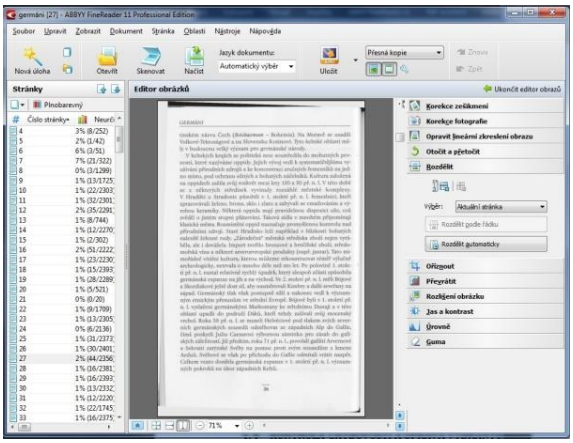


Obrázek 10 - dialogové okno pro nastavení skenování (Zdroj: vlastní)

Program umí načíst i obrazy z externího zdroje v různých grafických formátech. V záložce Soubor > Otevřít soubor / obraz PDF si v dialogovém okně zvolí složku a označí soubory, které chce načíst. Pro tyto obrazy platí stejné nastavení v oddílech Předběžného zpracování obrazu i Obecné, jaké bylo nastaveno při skenování.

4.3 ÚPRAVA OBRAZŮ

Při skenování může být zapnuta funkce předběžného zpracování obrazu. Myslí se tím procesy vedoucí ke zkvalitnění obrazového zdroje pro proces OCR, jako je automatické narovnávání stran, narovnávání řádků, odstraňování šumu atd. Ne všechny nedostatky jsou odstraněny při prvním projetí úpravy, a proto FRA nabízí vlastní grafický editor.



Obrázek 11 - editor obrazů pro úpravu předlohy (Zdroj: vlastní)

Nástroje editoru umožňují dodatečné uživatelské vylepšení zdrojového obrazu. Přesto jednotlivá nastavení nástrojů může digitalizátor ovlivňovat pouze částečně. Např. v nástroji Korekce fotografie při použití efektu Redukce ISO šumu nenajdeme možnost výběru, jakým mechanismem a v jaké síle bude efekt využit. Dále mají všechny nástroje volbu použití efektu na aktuální stránku nebo celý dokument. Aplikace na celý dokument může v některých případech napáchat více škody než užitku. Např. při automatickém rozdělování dvoustran dokumentu, kdy naskenované strany nemají výrazně jasově odlišený střed obrazu. Po aplikaci je jedna strana výrazně užší než druhá, protože se nástroj orientoval podle nejtmavší části obrazu a to jsou konce textových polí. Proto je lépe editor používat jako doplňující funkci.

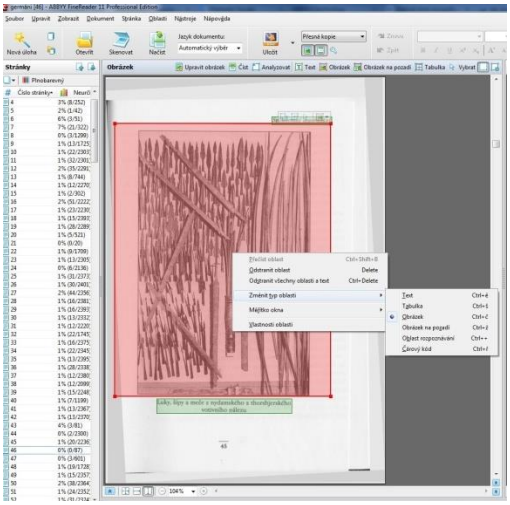
4.4 KONTROLA CHYBOVOSTI A ÚPRAVA VYBRANÝCH OBLASTÍ

Jsme ve fázi, kdy v programu máme naskenované obrazy, zkontrolovali jsme jejich kvalitu, případně jsme odstranili drobné nedostatky v grafickém editoru. Všechny obrazy jsou rozdělené, žádný se nezobrazuje obráceně, nikde nechybí část textu a řádky jsou pěkně srovnané. Můžeme spustit analýzu a čtení dokumentu.

Ačkoliv analýza barevně rozčlení dokument na oblasti obsahující text (zelená barva), obraz (červená), tabulku (modrá) atd., dál s nimi nepracuje. Je tedy možné přistoupit k procesu čtení. Máte na výběr spuštění procesu pro celý dokument z hlavního panelu nástrojů, nebo z panelu okna obrázek pro jednotlivé obrazy. Po průchodu operace zjistíte, že samo čtení si stejně dokument nejdřív zanalyzuje a zároveň hned provádí rozpoznávání znaků. Je tedy logické spouštět ihned proces čtení celého dokumentu z hlavního panelu.

Text je rozpoznán, ale je na uživateli, aby v pracovním prostředí FRA zkontroloval, že byly všechny strany správné analyzovány a přečteny. Blízkost oken obrazu a skenu pro porovnání je uživatelsky velice přívětivé. Stává se, že část grafiky je rozpoznána jako text a obráceně. Vybereme chybně označenou oblast a smažeme ji. V oknu Obrázek se zvolí vhodný typ oblasti a označí se jím nerozpoznaná oblast. Stále v panelu Obrázek se zvolí příkaz číst, protože se tento pokyn omezí pouze na upravovanou stránku v dokumentu.

Je důležité zkontrolovat, zda textové oblasti mají přiřazený správnou textovou funkci. Zda je zobrazený text označený za základní, jako záhlaví a zápatí či text v rámečku (textboxu). Zmiňuji to zde proto, že se u některých typů dokumentů stává, že je název kapitoly (ať už číselný nebo slovní) rozpoznán jako záhlaví. Při exportu se tato rozložení zachovává a může pak ve výsledku působit dojmem, že se kapitola „ztratila“ a dojde ke ztížení orientace v digitálním dokumentu.

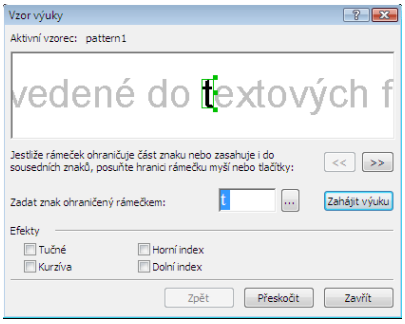


Obrázek 12 - oblasti OCR (Zdroj: vlastní)

4.5 TVORBA NOVÝCH UŽIVATELSKÝCH VZORŮ

V kapitole Faktory podílející se na kvalitě OCR převodu byl neznámý font (zejména zdobný, rukopisný či obsahující matematické symboly) uveden jako jeden z činitelů selhání rozpoznání znaků. Některé OCR programy to řeší vestavěným výukovým modulem. Tak jako FineReader Abbyy. Jedná se však o pokročilejší funkci a je na uživateli, jestli pro něj bude rychlejší nerozpoznaný text upravit v textovém poli ručně, nebo se dané písmo bude objevovat při digitalizaci častěji a vyplatí se písmu vytvořit vlastní vzor.

V případě druhé varianty otevřeme záložku Nástroje, nástroj Možnosti. Na kartě Číst musíme nastavit volbu Používat vestavěné i uživatelské vzory, zaškrtnout pole Číst s výukou a pak kliknout na tlačítko Editor vzoru. Založíme nový vzor a přiléhavě jej pojmenujeme např. názvem fontu. Volbu potvrdíme a zavřeme dialogová okna. Spustíme čtení oblasti, a pokud proces narazí na neznámý znak, otevře se nám editační okno Výuky vzoru. V horní části okna vidíme část textu se znakem v zeleném rámečku. Okraje rámce lze rozšiřovat tak, aby bylo označeno celé písmeno. Ve střední části okna pak zvolíte symbol, které rozpoznávané písmeno představuje, a stiskneme tlačítko Zahájit výuku. Někdy jsou symboly tak blízko u sebe, že je lepší je do uživatelského slovníku uložit jako jeden složený znak. Výuka vzoru zahrnuje i zachování kurzívy, tučnosti a indexů. Po skončení výuky se vrátíme do záložky Nástroje > Možnosti > Číst > Editor vzorů a tam daný vzor nastavíme jako aktivní. Při digitalizaci jiné publikace můžeme vzor deaktivovat a v případě potřeby i upravit.



Obrázek 13 - ukázka tvorby uživatelského vzoru v prostředí FRA 11 (Zdroj: vlastní)

4.6 VÝBĚR FORMÁTU PRO EXPORT TEXTU

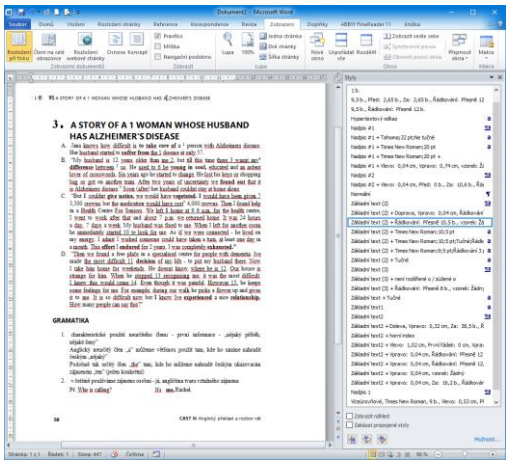
FineReader Abbyy nabízí uživateli širokou škálu formátů, do kterých může být výsledný digitální dokument odeslán. Může si zvolit formát, který je nejpřijatelnější pro zařízení, ve kterém si bude elektronickou knihu zobrazovat. Nabídka formátů obsahuje:

- editovatelné formáty (výsledný dokument lze dál opravovat – docx, odt, rtf, excel
- soubory využívající jazyk html a kaskádové styly (html, epub)
- obrazové soubory (Djvu, PDF, jednosouborový tiff atd.)

V záložce Nástroje > Možnosti > karta Uložit zvolíme další specifika, která mají výše zmíněné formáty v nabídce. Soubory typu Djvu a PDF zobrazují totožnou kopii originálu v přijatelné souborové velikosti s možností přidání OCR vrstvy. Ostatní formáty mají

v možnostech volbu zachování přesného formátování, které má omezená práva úpravy, nebo naopak úplné zrušení formátování a zachování čistého textu.

U editovatelných souborů lze nastavit variantu Upravitelný soubor, který spojuje zachování formátu původního souboru s možností zásahu uživatele. Pro další práci s exportovaným souborem se jedná o nejvhodnější volbu. Když si prohlédneme obrázek 14, zjistíme, že exportovaný jednostránkový soubor je tvořen množstvím aktivních stylů, mezi nimiž se nachází i typograficky (důležitá pro děj, zdůraznění) důležitá kurzíva a tučnost písma. V konečné fázi úpravy dokumentu by takové množství stylů mohlo činit obtíže při konverzi do dalších specializovanějších formátů.



Obrázek 14 - přehled OCR stylů v jednostránkovém dokumentu (Zdroj: vlastní)

4.7 NÁSLEDNÁ KOREKTURA EDITOVATELNÉHO FORMÁTU

V poslední části procesu vzniku digitální knihy pracujeme s neupraveným výstupem z OCR programu. Jak bylo zmíněno výše, takový dokument obsahuje velké množství stylů a i po opravě rozpoznávacích chyb přímo v programu se nejedná o bezchybný dokument. To si uvědomoval i jeden z korektorů slovenské knihovny Digibooks a vytvořil sadu maker[19], která měla za úkol co nejvíce usnadnit konečnou práci korektorů. Byla a jsou vyvíjena primárně pro textový dokument Microsoft Office Word (dále jen Word).

Na svých stránkách má přesné popisy a návody, jak s makry zacházet, aby bylo dosaženo co nejoptimálnějšího výsledku. Aby se nám makra vůbec ve Wordu zobrazila, je třeba je nainstalovat pomocí souborového manažera do adresáře Po spuštění. Součástí instalovaného balíčku je šablona knizka_menu, knizka_cz_slovnik a knizka_zz_slovnik.

[19] Dostupná na adrese: http://santiago.wz.cz/

Po instalaci spustíme uložený wordový dokument, ve kterém se nám díky šabloně knizka_menu objevil nový ribbon. Santiagova makra lze ještě spouštět přes kartu Zobrazení > Makra a následně vybírat ve správném pořadí ze seznamu. Aby makra správně pracovala, je důležité při exportu OCR textu do Wordu nechat zaškrtnutá pole Zachovat konce stránek a konce řádků, protože se podle těchto značek orientuje makro na spojení dokumentu. Ačkoliv je Santiagem doporučený určitý postup, je na uživateli, v jakém pořadí některá makra spustí. Jeden z možných postupů je:

Tabulka 2 – tabulka základních maker

Název makra	Účel makra
Seskupeny_objekt	vyhledá spojený obrázek s textem, pravým tlačítkem myši zvolit oddělit, nutné spustit před makrem na spojování řádků, text uvnitř seskupení by byl ignorován
Odstraneni_textboxu	extrahování textu z testových polí
Spojovani_stranek_radku	spojí rozdělená slova a spojí dokument v jeden celek
Uprava_textu	automaticky opravuje řadu typografických chyb - odstraňuje nadbytečné mezery - sjednocuje uvozovky - doplňuje mezeru za interpunkčními znaménky
Nahrazeni_odrazek	v souvislém textu může být předložka V, I na začátku odstavce rozpoznána jako odrážka, a při konečném odstraňování stylů mizela z textu, takto se z použitých odrážek stává prostý text
Nahrazeni_tabulatoru	po odstranění odrážek zůstává za zmiňovanými předložkami tabulátor, toto makro jej nahrazuje mezerou
Nahrazeni_CZ(SK)_slovnik	soubor .txt, který slouží jako zásobník velmi často se opakujících OCR chyb každý uživatel si ho může plnit vlastními výrazy, jen musí dodržovat následující syntax: [hledané slovo];[nahrazené slovo];[způsob hledání] Př. _ted;_teď;a (a = vyhledává pouze celé shodné slovo) trn;tm;n (n = vyhledává v části slova (zatrní = zatmí))
Nahrazeni_ZZ_slovnik	definuje pokročilé uživatelské úpravy za pomoci regulárních výrazů jednou z funkcí je např. vyhledávání neukončené věty uprostřed odstavce na základě detekce velkého písmene následujícího za slovem neukončeném tečkou
Odstraneni_stylu	poslední spouštěné makro, které odstraňuje přebytečné styly a zachovává pouze ty, které jsou označeny v nabídce (kurzíva, kapitálky, indexy, velká písmena, tučnost atd.)

Výsledkem je čistý dokument připravený k formátování a korektuře čtením. Ještě před těmito kroky lze spustit ještě tzv. kontrolní makra, která vyhledávají OCR relikty a spouští se klávesovými zkratkami:

Tabulka 3 - seznam kontrolních maker

Klávesová zkratka	Účel makra
CTRL + num1	vyhledává začátky řádků když: je na začátku malé písmeno a následující znaky ,..– když je na začátku velké písmeno, ale na konci předchozího řádku je čárka
CTRL + num2	vyhledává uvozovky, které jsou z obou stran obklopeny mezerou nebo naopak obklopeny textem
CTRL + num3	vyhledává pomlčku, která není oddělena mezerami od okolního textu, ignoruje -li
CTRL + num4	vyhledává chybně rozpoznané znaky uvozovek v kombinaci: ., ,. “ .. atd.
CTRL + num6	hledání dlouhé pomlčky, která je z jedné strany spojena s malým či velkým písmenem
CTRL + num7	hledá tečku nebo čárku obklopenou písmenem z pravé strany nebo čárku obklopenou mezerou z levé strany
CTRL + num8	vyhledává chybějící uvozovky do páru
CTRL + num9	vyhledává ď a ť s použitým apostrofem místo háčku
CTRL + ALT + num6	vyhledává nestandardní, atypické znaky
CTRL + ALT + num7	vyhledává francouzské uvozovky » «, které jsou z obou stran obklopeny mezerou nebo naopak obklopeny textem
CTRL + ALT + num8	vyhledává chybějící francouzské uvozovky do páru
CTRL + ALT + num9	hledá chybějící kulaté závorky do páru
CTRL + num,	vyhledává jednoduché uvozovky do páru

Po projetí dokumentu základními i kontrolními makry [38], pokračujeme použitím wordového nástroje kontroly gramatiky a pravopisu. K naformátování dokumentů použije uživatel styly, které jsou užitečné nejen k přehlednému členění dokumentu, ale automatické konvertory na jejich základě převádějí dokument do formátu, které jsou podporovány elektronickými čtečkami.

ZÁVĚR

Jak z textu mé bakalářské práce vyplynulo, digitalizace tištěných dokumentů je již nedílnou součástí života každého z nás – ať už v roli koncového uživatele nebo autora. Pro širokou veřejnost je zde dost dostupných komerčních i freewarových aplikací či programů, ať už pro vlastní potřebu nebo pracovní využití. O volbě některého z nich bude rozhodovat cílový požadavek uživatele, v jaké kvalitě mu daný převod stačí.

Bylo překvapivé zjištění, jak hluboce jsme v běžném životě digitalizací silně ovlivněni. Vyplývá to mimo jiné i z nařízení jednotlivých ministerstev a dalších státních orgánů a institucí včetně Evropské Unie. V rámci zachování kulturního a historického dědictví probíhá množství celosvětových i evropských, případně českých projektů, které spolu navzájem spolupracují. Digitalizace ve velké míře zasáhla i práci ve školství, vědě a kultuře, a to jak v běžné činnosti, tak především v oblasti zdravotně znevýhodněných jedinců – zejména zrakově i sluchově postižených.

Pokusila jsem se seznámit s historickým vývojem programů zabývajících se převodem tištěného textu do digitální podoby, který probíhá už od počátku 20. století. Rozepsala jsem jednotlivé fáze procesu OCR s tím, aby byla pochopitelná i pro laiky. Zmínila jsem literaturu, ze které lze daný proces nastudovat odborněji.

Následně jsem přistoupila k vysvětlení, které faktory ovlivňují kvalitu výstupu rekognice. Nastínila jsem možná řešení, jak předejít zvýšenému množství chyb. Informuji o programech, které se OCR procesem zabývají, ať se jedná o nástroje komerční nebo volně dostupné.

V poslední kapitole pak uvádím jeden z možných konkrétních postupů, jak si může běžný uživatel vytvořit vlastní kopii papírové knihy, aniž by se dopustil porušení zákona.

RESUMÉ

Česky

Tato bakalářská práce pojednává o problematice digitalizace v současném světě. Práce je rozdělená do čtyř částí.

První tři kapitoly se zabývají teorií, čtvrtá část je praktická. V teoretických kapitolách je pojednáváno o využití digitalizace v běžném životě člověka, existujících projektech týkající se digitalizace. Následuje popis vývoje programů OCR a je doplněn o rozbor fází procesu OCR. Praktická část pak přestavuje možný způsob převodu tištěné publikace do digitální podoby. Praktická část je doplněna videonávody na přiloženém DVD.

English

This bachelor thesis deals with the problems of digitization in the contemporary world. The work is divided into four parts.

The first three chapters deal with theory, the fourth part is practical. The theoretical chapters deal with the use of digitization in everyday life of people, existing digitization projects. The following is a description of the development of OCR programs and is accompanied by an analysis of the OCR process phases. The practical part represents the possible way of transferring the printed publication into digital form. This section is complemented by a video tutorial on the enclosed DVD.

SEZNAM LITERATURY

[1] VRBENSKÁ, Františka. Digitalizace dokumentů. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR, 2003- [cit. 2017-06-04]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001728&local_base=KTD.

[2] ČESKÝ STATISTICKÝ ÚŘAD. Tab. Bilance elektrické energie. *Český statistický úřad* [online]. 23.05.2017 [cit.2017-05-23]. Dostupné z: https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&z=T&f=TABULKA&katalog=30835&pvo=ENE04&c=v3~8__RP2015

[3] ČESKÝ STATISTICKÝ ÚŘAD. Tab. Vybavenost domácností informačními a komunikačními technologiemi. *Český statistický úřad* [online]. 23.5.2017 [cit.2017-05-23]. Dostupné z: xhttps://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&z=T&f=TABULKA&katalog=31031&pvo=ICT03&str=v149

[4] ČESKÝ STATISTICKÝ ÚŘAD. Tab. Vybavenost domácností informačními a komunikačními technologiemi. *Český statistický úřad* [online]. 23.5.2017 [cit.2017-05-23]. Dostupné z: https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&z=T&f=TABULKA&katalog=31031&pvo=ICT03&str=v146

[5] *Národní strategie elektronického zdravotnictví* [online]. Praha: Ministerstvo zdravotnictví ČR, 2016 [cit. 2017-06-30]. Dostupné z: http://www.nsez.cz/

[6] Elektronický recept: výhody a přínosy pro lékaře i pacienta. *Www.prolekare.cz* [online]. Praha, 2016 [cit. 2017-06-30]. Dostupné z: http://www.prolekare.cz/novinky/prolekare/elektronicky-recept-vyhody-a-prinosy-pro-lekare-i-pacienta-6780?confirm_rules=1

[7] JANÍKOVÁ, Simona. Elektronické recepty jsou od příštího roku povinné. Za vytištění kódu na papír už stát zaplatil 300 milionů. *Hospodářské noviny* [online]. Praha, 2017 [cit. 2017-06-30]. ISSN 1213-7693. Dostupné z: https://byznys.ihned.cz/c1-65627060-elektronicke-recepty-jsou-od-pristiho-roku-povinne-vyjdou-na-300-milionu-nic-ale-neumi-rikaji-experti

[8] Systém „eRecept" stál přes 300 milionů korun. Jeho využití bylo mizivé. *Www.nku.cz* [online]. Praha, 2017 [cit. 2017-06-30]. Dostupné z: https://www.nku.cz/cz/pro-media/tiskove-zpravy/system-_erecept_-stal-pres-300-milionu-korun--jeho-vyuziti-bylo-mizive-id8693/

[9] IDNES.CZ. E-recepty budou od příštího roku povinné. Skončíme, hrozí mnozí lékaři. *IDNES.cz* [online]. Praha: MAFRA, 2017 [cit. 2017-06-30]. Dostupné z: http://zpravy.idnes.cz/lekari-e-recepty-elektronicke-recepty-ministerstvo-zdravotnictvi-1fy-/domaci.aspx?c=A170215_200842_domaci_fer

[10] Autorský zákon: Zákon č. 121/2000 Sb., zákon o právu autorském, o právech souvisejících s právem autorským a … *Bussiness.center.cz* [online]. Praha: HAVIT, 2017 [cit. 2017-06-30]. Dostupné z: http://business.center.cz/business/pravo/zakony/autorsky/

[11] VORLÍČKOVÁ, Blanka. *Online zpřístupnění kulturního dědictví v kontextu vybraných témat informační politiky*. Praha, 2012. Disertační. Univerzita Karlova. Vedoucí práce Doc. PhDr. Richard Papík, Ph.D.

[12] ČR. *Strategie Evropa 2020 - Digitalizace kulturního obsahu: Strategie digitalizace kulturního obsahu na léta 2013 - 2020*. In: . Praha: MK ČR. Dostupné také z: https://www.mkcr.cz/strategie-evropa-2020-digitalizace-kulturniho-obsahu-831.html

[13] WIKIPEDIA CONTRIBUTORS. Copyright Term Extension Act. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2017-06-30]. Dostupné z: https://en.wikipedia.org/wiki/Copyright_Term_Extension_Act

[14] *Free ebooks - Project Gutenberg* [online]. Salt Lake City: Gutenberg [cit. 2017-06-30]. Dostupné z: http://www.gutenberg.org/wiki/Gutenberg:Contact_Information

[15] Zpráva o „i2010: směrem k evropské digitální knihovně". *Evropský parlament* [online]. 2007 [cit. 2017-06-30]. Dostupné z: www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A6-2007-0296+0+DOC+XML+V0//CS

[16] ČESKÝ STATISTICKÝ ÚŘAD. Tab. Vybavenost domácností informačními a komunikačními technologiemi. *Český statistický úřad* [online]. 23.05.2017 [cit.2017-05-23]. Dostupné z:

https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&z=T&f=TABULKA&katalog=31031&pvo=ICT03#w=

[17] COMMISSION RECOMMENDATION of 27 October 2011 on the digitisation and online accessibility of cultural material and digital preservation. *Eur-lex.europa.eu* [online]. 2011 [cit. 2017-06-30]. Dostupné z: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:283:0039:0045:EN:PDF

[18] VORLÍČKOVÁ, Blanka. 2009a. Europeana: online přístup k evropskému kulturnímu a historickému dědictví. *Ikaros* [online]. Roč. 13, č. 3 [cit. 10.12.2012]. URN-NBN:cz-ik5313. ISSN 1212-5075. Dostupný z World Wide Web: http://www.ikaros.cz/node/5313

[19] MELICHAR, Marek; HUTAŘ, Jan. České paměťové instituce a digitální data – historický exkurz, současný stav a předpokládaný vývoj III.. *Duha* [online]. 2014, roč. 28, č. 2 [cit. 2017-06-17]. Dostupný z WWW: <http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj-1>. ISSN 1804-4255.

[20] MELICHAR, Marek; HUTAŘ, Jan. České paměťové instituce a digitální data – historický exkurz, současný stav a předpokládaný vývoj I.. *Duha* [online]. 2013, roč. 27, č. 4 [cit. 2017-06-17]. Dostupný z WWW: <http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj>. ISSN 1804-4255.

[21] Projekty digitalizace. *Artslexikon* [online]. 2012 [cit. 2017-06-30]. Dostupné z: http://www.artslexikon.cz//index.php?title=Projekty_digitalizace

[22] *Gallica* [online]. [cit. 2017-06-30]. Dostupné z: http://gallica.bnf.fr/accueil/?mode=desktop

[23] VORLÍČKOVÁ, Blanka. Europeana: online přístup k evropskému kulturnímu a historickému dědictví. *Ikaros* [online]. 2009, ročník 13, číslo 3 [cit. 2017-06-30]. urn:nbn:cz:ik-13047. ISSN 1212-5075. Dostupné z: http://ikaros.cz/node/13047

[24] Strategie digitálního vzdělávání do roku 2020. *MŠMT ČR* [online]. Praha, 2014 [cit. 2017-06-30]. Dostupné z: http://www.msmt.cz/uploads/DigiStrategie.pdf

[25] Digitalizace. *Národní digitální knihovna* [online]. Praha: Národní knihovna ČR, 2017 [cit. 2017-06-30]. Dostupné z: http://www.ndk.cz/digitalizace-1

[26] Koncepce rozvoje knihoven ČR na léta 2011 – 2015 včetně internetizace knihoven. *Ministerstvo kultury ČR* [online]. Praha: MK ČR, 2017 [cit. 2017-06-30]. Dostupné z: https://www.mkcr.cz/assets/literatura-a-knihovny/Koncepce_rozvoje_knihoven_2011-2015.pdf

[27] EIKVIL, Line. *Optical Character Recognition* [online]. 1993 [cit. 2016-06-30]. Dostupné z: http://www.nr.no/~eikvil/OCR.pdf

[28] Technology. *Abbyy* [online]. Moskva: Abbyy, 2017 [cit. 2017-06-30]. Dostupné z: http://www.abbyy.co.il/?categoryId=63424

[29] SOBOTKA, Zdeněk a Martin SOBOTKA. *Základy číslicového zpracování obrazu*. Praha: Dům techniky ČSVTS, 1990. ISBN 80-02-00736-0.

[30] SOBOTKA, Zdeněk a Martin SOBOTKA. *Počítačová analýza a rozpoznávání obrazu*. Praha: Dům techniky ČSVTS, 1990. ISBN 80-02-00739-5.

[31] ABBYY FineReader: *User's Guide for ABBYY FineReader 11: uživatelská příručka k aplikaci FineReader 11*. 1. vyd. ABBYY software, 2011, 110 s. Dostupné z: http://www.abbyy.com/fr11guide_cz.pdf

[32] Čárový kód - základní prostředek automatické identifikace zboží. *Kodys* [online]. Praha: Kodys, 2017 [cit. 2017-06-30]. Dostupné z: http://www.kodys.cz/technologie/carovy-kod

[33] *Wikipedie: Otevřená encyklopedie: QR kód* [online]. c2017 [citováno 25. 06. 2017]. Dostupný z WWW: https://cs.wikipedia.org/w/index.php?title=QR_k%C3%B3d&oldid=14851385

[34] *Wikipedia: Intelligent character recognition* [online]. 2016 [citováno 25. 06. 2017]. Dostupný z: https://en.wikipedia.org/w/index.php?title=Intelligent_character_recognition&oldid=740663848

[35] DOLEJŠÍ, Tomáš. *Jak fotografovat dokumenty a listiny*. *Fotorádce* [online]. 2007 [cit. 2017-06-27]. Dostupné z: https://www.fotoradce.cz/jak-fotografovat-dokumenty-a-listiny

[36] PAVLÍK, Vladimír. Scannery - CIS nebo CCD? *Svět hardware* [online]. 1999 [cit. 2017-06-27]. Dostupné z: https://www.svethardware.cz/scannery-cis-nebo-ccd/826

[37] Sada maker na úpravu digitalizovaných textů pro MS Word. *Santiago.wz.cz* [online]. 2015 [cit. 2017-06-29]. Dostupné z: http://santiago.wz.cz/index.html

[38] ABBYY FineReader: *User's Guide for ABBYY FineReader 14: uživatelská příručka k aplikaci FineReader 14*. 1. vyd. ABBYY software, 2014, 268 s. Dostupné z: http://www.abbyy.cz/files/download/Guide_Czech.pdf

[39] OmniPage Standard. *Nuance. com* [online]. 2017 [cit. 2017-06-29]. Dostupné z: https://www.nuance.com/print-capture-and-pdf-solutions/optical-character-recognition/omnipage/omnipage-standard.html

[40] VESELÝ, Peter. *OCR analýza*. Brno, 2005. Diplomová práce. Masarykova univerzita. Vedoucí práce Mgr. Lukáš Svoboda.

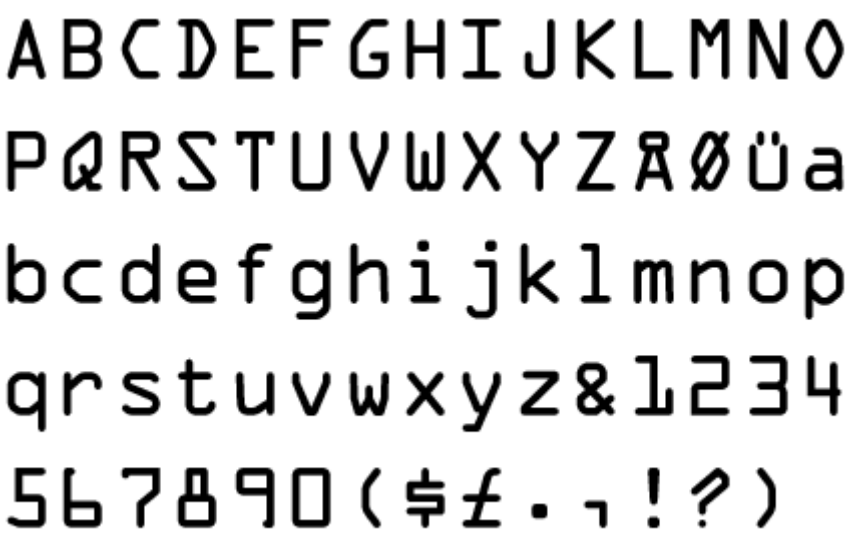
[41] MARINIČ, Michal. *Rozpoznávání textu z obrazových dat*. Brno, 2014. Diplomová práce. Vysoké učení technické. Vedoucí práce Ing. Radim Burget, Ph. D.

[42] HAVRÁNEK, Petr. *Optimalizace procesu rozpoznávání textu pomocí Vision Builder*. Plzeň, 2012. Diplomová práce. Západočeská univerzita. Vedoucí práce Ing. Radek Holota, Ph. D.

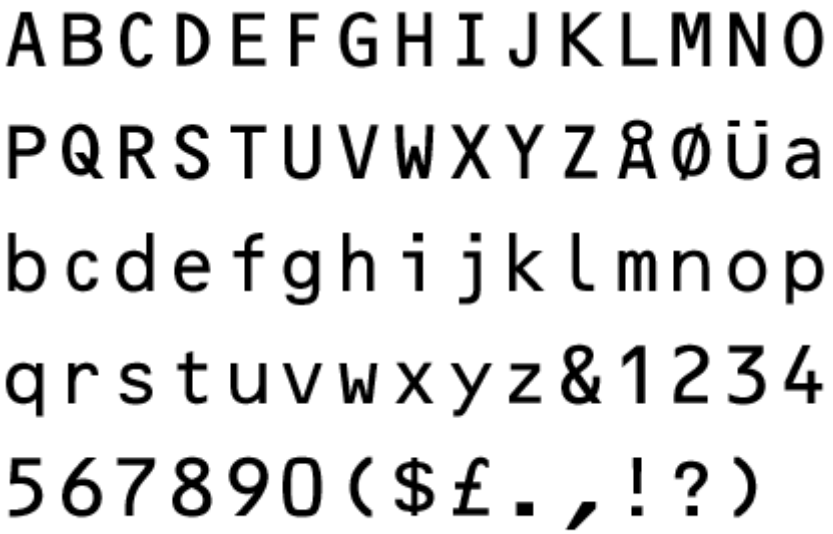
SEZNAM OBRÁZKŮ, TABULEK, GRAFŮ A DIAGRAMŮ

PŘÍLOHY

Příloha 1 – ukázky standardizovaných znakových sad pro USA a Evropu



znaková sada OCR-A



evropská znaková sada OCR-B

Příloha 2

Stručné shrnutí postupu digitalizace

- naskenování dokumentu ve stupních šedi s hodnotou DPI 300
- kontrola zpracování obrazu
    - dodatečné rozdělení protilehlých stran
    - kontrola, zda jsou naskenovány všechny strany dokumentu
    - oříznutí (zmenšení rozpoznávací plochy)
    - případné srovnání řádků s textem, využití dalších grafických nástrojů
- čtení dokumentu
    - kontrola rozpoznaných oblastí (snižování chybovosti správnou volbou typu)
    - kontrola textové funkce (záhlaví, zápatí atd.)
- export do editovatelného formátu a uložení projektu FRA
    - základní a kontrolní Santiagova makra
    - naformátování dokumentu pomocí stylů
    - případná konverze do jiného formátu

Příloha 3

Obsah DVD:

Bakalářská práce ve formátu .docx

Bakalářská práce ve formátu .pdf

Složka Výuková videa

1. Pracovní prostředí FineReader Abbyy
2. Automatizované úlohy
3. Nastavení parametrů skenování
4. Načtení obrazového zdroje
5. Dodatečná úprava naskenovaného obrazu
6. Úprava typu oblasti dokumentu
7. Úprava textové funkce
8. Tvorba výukového vzoru
9. Export souboru
10. Instalace Santiagových maker
11. Použití Santiagových maker
12. Formátování dokumentu pomocí stylů