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: Mgr. Zbyněk Filipi

**Plzeň 2015**

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

V Plzni 30. června 2015

.....

vlastnoruční podpis

Děkuji Mgr. Zbyňku Filipimu za podporu a trpělivost při vedení mé práce.

# OBSAH

SEZNAM ZKRATEK ..... 2
ÚVOD ..... 3
1 VÝZNAM DIGITALIZACE V SOUČASNOSTI ..... 4
1.1 KULTURNÍ A HISTORICKÉ DĚDICTVÍ ..... 5
1.1.1 Projekty ve světě ..... 5
1.1.2 Projekty v České republice ..... 7
1.2 VĚDECKÁ ČINNOST A VZDĚLÁVÁNÍ ..... 9
1.3 ZRAKOVĚ POSTIŽENÍ ČTENÁŘI ..... 11
1.3.1 KDD (Knihovna digitálních dokumentů) ..... 11
1.3.2 Digibooks ..... 12
2 PRINCIPY PŘEVODU TEXTU DO DIGITÁLNÍ PODOBY ..... 14
2.1 HISTORIE OCR ..... 14
2.1.1 První generace OCR ..... 14
2.1.2 Druhá generace OCR ..... 14
2.1.3 Třetí generace OCR ..... 15
2.1.4 Současnost ..... 15
2.2 POJMY ..... 15
2.3 FÁZE PROCESU ROZPOZNÁVÁNÍ ZNAKŮ ..... 16
2.3.1 Získání obrazu (optické skenování) ..... 16
2.3.2 Preprocessing ..... 16
2.3.3 Lokalizace a segmentace ..... 17
2.3.4 Znaková extrakce ..... 17
2.3.5 Lexikální postprocessing ..... 17
3 FAKTORY PODÍLEJÍCÍ SE NA KVALITĚ OCR PŘEVODU ..... 18
3.1 PROSTŘEDKY ZÍSKÁVÁNÍ OBRAZU ..... 18
3.2 PŘEDLOHA ..... 18
3.3 SOFTWARE ..... 18
3.3.1 Komerční software ..... 18
3.3.2 Freeware ..... 18
4 POSTUP DIGITALIZACE TEXTU V PROGRAMU FINEREADER ABBYY 11 ..... 20
4.1 NASTAVENÍ SKENOVÁNÍ ..... 20
4.2 ÚPRAVA OBRAZŮ ..... 20
4.3 KONTROLA CHYBOVOSTI A ÚPRAVA VYBRANÝCH OBLASTÍ ..... 21
4.4 TVORBA UŽIVATELSKÝCH VZORŮ ..... 21
4.5 VÝBĚR FORMÁTU PRO EXPORT TEXTU ..... 22
4.6 NÁSLEDNÁ KOREKTURA EDITOVATELNÉHO FORMÁTU ..... 22
ZÁVĚR ..... 23
RESUMÉ ..... 24
SEZNAM LITERATURY ..... 25
SEZNAM OBRÁZKŮ, TABULEK, GRAFŮ A DIAGRAMŮ ..... 26
PŘÍLOHY ..... I

# SEZNAM ZKRATEK

OCR Optical Character Recognition – optické rozpoznávání znaků

EPUB Electronic Publication – elektronická publikace

PDF Portable Document Format – přenosný formát dokumentů

QR Quick response – kód pro obrazové uložení informace

# 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 mnohým z nich teď oči hledí místo do papíru tak 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 následujících kapitolách se pokusím přiblížit význam digitalizace v současnosti, rozebrat principy převodu textu z papírové verze do digitální, a představím faktory, které ovlivňují kvalitu OCR převodu. Na závěr se zaměřím na 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, možnosti převodu do jednotlivých typů souboru od editovatelných (docx, rtf) po needitovatelné (pdf, djvu, epub ad.), schopnost se „učit“ – vytvořit si vlastní uživatelské vzory pro rozpoznání textu.

# 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 (mapy, obrazy, fotografie, mikrofilmy, noty, kazety, VHS, exponáty atd.).

Není to poprvé, co by v historii došlo k záměně jednoho média za jiné, nikdy však to předešlé nebylo nahrazeno plně. Od ústního podání, ručního přepisování v klášterech až po vynález knihtisku, který rozšířil písmo i mezi obyčejný lid. S příchodem telegrafu, díky kterému se zprávy přenášely v reálném čase a na velké vzdálenosti, se objevily i první předpovědi o konci éry papíru. Ale jak víme, v současnosti vedle sebe koexistují všechny formy uchovávání informací, v menší či větší míře využívané, vedle sebe.

Digitalizace může působit dojmem módního výstřelku současné doby, ale pravdou zůstává, že je logickým pokračováním trendu uchovávání informací, dat, kulturního i historického dědictví. Jedná se i o nezbytnou potřebu dobře fungující veřejné správy státu, protože představuje přínosné a praktické řešení – digitální materiál je v rámci svého určení rychle dostupný, snadno přenositelný, zabírá méně prostoru, šetří peníze – a z ekologického hlediska šetří i životní prostředí – lesy.

Právě snadný přístup k údajům, vždyť je v České republice připojeno k internetu, podle ČSÚ 2013[1], skoro 70% domácností, je důležité řešit i jejich bezpečnost a zachování důvěrnosti osobních údajů. Nemluvě o autorských právech jednotlivých děl, kdy zákon nestíhá reagovat na současnou situaci.

Pro téma mé bakalářské práce, je ale důležité znění zákona č. 121/2000 Sb., zákon o právu autorském, o právech souvisejících s právem autorským[2], který upravuje 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. Tzn. pokud si zakoupíte papírovou knihu a chcete si vytvořit její elektronickou kopii pro svou čtečku

[1] Uvedeno na stránkách Českého statistického úřadu – informační technologie, dostupné z: https://www.czso.cz/csu/czso/informacni_technologie_pm
[2] Znění zákona je dostupné z: http://business.center.cz/business/pravo/zakony/autorsky/

elektronických knih, ponecháte tam veškeré údaje o autorovi a nakladateli. – žádného přestupku se tímto jednáním 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 Česká republika dospěl k bodu, kdy se rozhodl, že své kulturní a historické bohatství zdigitalizují a zachovají v elektronických formátech pro budoucí generace.

### 1.1.1 PROJEKTY VE SVĚTĚ

První projekty, které propagovaly elektronické knihy, či digitalizaci papírových děl se objevily v USA, kde vznikl asi nejstarší projekt, který trvá dodnes, a tím je *Gutenberg*. Americká Kongresová knihovna spustila projekt *American Memory Programme*, který je obdobou evropského projektu *Europeana*. Jejich cílem je zachování historických dokumentů vztahujících se k dějinám daného státu. V Evropě byly též spuštěny velké projekty – např. Francie se svým projektem *Gallica*, celoevropský projekt *Europeana*, nebo evropský projekt *Unesco Memory of the World*. Těm nejvýznamnějším projektům se budu věnovat níže.

#### Gutenberg

Michael Hart (1947–2011) byl známý americký autor, který se proslavil vynálezem elektronické knihy (e-booku) a založením prvního 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.

Celá databáze čítá na 100 000 publikací určených k prohlížení a skoro 50 000 volně stažitelných elektronických knih ve formátu epub či mobi pro Kindle. První knihy digitalizoval sám ještě v době, kdy i samotný internet byl v plenkách, ale postupně se počet dobrovolníků rozšiřující řady digitalizátorů rozrostl na několik tisíc.

Projekt se zaměřuje na tři oblasti literatury, které digitalizuje. Lehkou literaturu v podobě dětských příběhů, pohádek, či bajek. Knihy pro náročnější čtenáře jako jsou náboženské

texty či klasičtí autoři jako Shakespeare či Melville, a pak vědeckou část tvořenou slovníky, almanachy a encyklopediemi.

Projekt se samozřejmě primárně zabývá anglicky psanou literaturou, ale ani ostatní jazyky zde nepřijdou zcela zkrátka. Ostatní jazyky jsou rozděleny do dvou kategorií – kdy dělící hranici tvoří limit 50 zveřejněných ebooků. Do skupiny, která zde má větší zastoupení, patří sedmnáct jazyků, mezi něž patří např. čínština, dánština, nizozemština, finština, francouzština, němčina, řečtina, italština, latina 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 H. Jaroše, protože je zde uložen jeho český překlad Dostojevského díla Zápisky z mrtvého domu.

Pro uživatele je k dispozici vcelku přívětivé a hlavně jednoduché uživatelské prostředí v angličtině a dalších třech dostupných jazykových verzích – portugalština, němčina a francouzština. Uživatel má k dispozici možnost vyhledávání v databázi, nebo procházení katalogu tříděného na různé kategorie. Tím, že se jedná o neziskový projekt, je grafická stránka projektu upozaděna a primární požadavkem je funkčnost a obsah projektu[2][3].

### Gallica

Francouzská národní knihovna v roce 1997 převzala projekt digitální knihovny Gallica, který v současnosti obsahuje přes tři miliony digitálních děl – od manuskriptů, map, plánů, audio a video nahrávek, knih, partitur, pohlednic, obrazů a periodik. Z toho jen knih bylo v roce 2015 přes pět set sedmdesát tisíc kusů [4]. Ročně do knihovny přibude okolo sto tisíc děl.

Stránky projektu jsou ve čtyřech jazykových verzích – od francouzštiny pro angličtinu. Hned na úvodní straně je vyhledávací pole pro zadání názvu knihy či jména autora. Na stejné straně nalezneme ukryté menu, které uživatele přesměruje na požadovaný katalog – rukopisy, mapy, obrazy atd. Volná díla lze pak zobrazit na monitoru, či stáhnout v různých formátech – nejčastěji pdf, pak obrazové soubory jpg, tiff. Textové dokumenty lze díky procesu OCR stáhnout i ve formátu txt, nebo je přímo fulltextově vyhledávat.

Knihovna obsahuje díla nejenom týkající se francouzských dějin, ale i zahraniční díla v angličtině, němčině, italštině a latině. Obsah knihovny není indexován, ačkoliv obsahuje metadata, tudíž nelze dílo dohledat pomocí internetového vyhledávače.

### Europeana

V roce 2008 na popud Evropské komise vznikl projekt souboru evropských digitálních knihoven Europeana. Ředitelem projektu je Jill Sousins. Tato digitální knihovna, muzeum a archiv v jednom je přístupná ve 27 jazykových mutacích, kde je zpřístupněno přes 15 milionů děl z národních a univerzitních knihoven států Evropské unie [5].

## 1.1.2 PROJEKTY V ČESKÉ REPUBLICE

Ani Česká republika nezůstává pozadu a na popud dopisu Evropské komise[3] se u nás vytvořil projekt Národní digitální knihovny, a samozřejmě existují projekty k zachování evropského a českého kulturního dědictví, týkající se starých knih, vzácných rukopisů, historických map a digitalizaci v oblasti živého umění a folkloru. NDK není prvním projektem, jen je první takto masivním v českém prostředí.

### ČTE

V roce 2000 vznikl na popud manželů Praksových projekt Č.T.E. ( české texty elektronicky), který si dal za cíl zpřístupnit texty sloužící zejména pro vzdělávací účely. Pomocí softwaru pro katalogizaci knih v digitálním archivu použili software E-Library. Časem se počet spolupracovníků rozrostl na celkem 20 lidí.

Stránky podle poslední aktualizace ukončily svou činnost k roku 2002. Na stránkách ale dodnes najdeme funkční odkazy na další české a zahraniční archivy poskytující elektronické texty.

### NDK

Národní digitální knihovna využívající ke zpřístupnění děl prostředí Kramerius vzniká z fondů Národní knihovny České republiky a Moravské zemské knihovny, jelikož mají obě knihovny právo autorského výtisku, mají dostupnou veškerou 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.

[3] Viz dopis: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2005:0465:FIN:CS:HTML

V knihovnách fungují tzv. digitalizační linky, ve kterých probíhá proces od skenování, zpracování obrazu – jeho následné kontrole a případné úpravě (korekce stran), doplnění metadat (katalogizační údaje), a provedení OCR, kvůli fulltextovému vyhledávání, až k výslednému výstupu v podobě pdf, djvu, či dalších obrazových souborů [6].

Obě digitalizační linky jsou vybaveny velmi výkonnými knižními skenery mj. švýcarský 4DigitalBooks DL 3003[4], 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 na minimum. 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), a 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.

Do roku 2019 by obě knihovny měly zdigitalizovat okolo padesáti milionů stránek, a pokud vezmeme v úvahu, že průměrné množství stránek v publikaci se pohybuje okolo 200 stran, jedná se tedy přibližně o 250 tisíc publikací.

Od roku 2011 do 2015 v dalších knihovnách České republiky probíhaly regionální projekty, zaměřující se na uchování a digitalizací děl souvisejících právě s daným regionem [7]. V takovémto množství je logické, že musela vzniknout databáze již zdigitalizovaných děl, proti případné duplikaci. Knihovny si to mohou ověřit buď přímo na stránkách Národní digitální knihovny nebo na stránkách Registr digitalizace[5].

Národní knihovna zaštiťuje dva podprojekty Manuskriptorium[6], které se zabývá vzácnými tisky, rukopisy, listinami a mapami, takže vytváří virtuální badatelské prostředí a zároveň vytváří zálohu kvůli případnému poničení.

Projekt *Kramerius* je zde delší dobu, než se rozjel digitalizační boom fondů knihoven. Byl iniciován katastrofálními povodněmi roku 2002, kdy bylo poničeno velké množství knih. Tudíž si dal za cíl reformátovat knižní zdroje kvůli jejich záchraně a zachování. Primárně je určen zejména pro papírovou tvorbu – tzn. monografie a periodika. Dále však může

[4] Video se záběry funkčnosti skener je ke zhlédnutí zde: https://www.youtube.com/watch?v=RROgUub-cDQ&feature=youtu.be
[5] http://www.registrdigitalizace.cz/rdcz/
[6] http://www.manuscriptorium.com/cs

obsahovat i vydání, která vyšla pouze elektronicky, či mapy, notopisy atd. V současné době projekt zaštiťuje Knihovna Akademie věd České republiky.

Současná verze 4 je vyvíjena od roku 2009, a finančně je podporována dotačními programy Akademie věd a Ministerstva kultury. Díla jsou zpřístupněna ve všech prohlížečích, ale protože jsou primárně zveřejňována ve formátu DjVu, uživatel musí mít v browseru nainstalovaný plugin pro jeho prohlížení. Uživatelské prostředí je zpřístupněno ve dvou jazycích – čeština a angličtina. Uživatel má k dispozici i nápovědu pro lepší orientaci, dokonce i infoportál[7] o současném dění. Celkově mi přijde tento projekt, vedení databáze daleko přehlednější než zahraniční obdoby.

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.

## 1.2 VĚDECKÁ ČINNOST A VZDĚLÁVÁNÍ

Digitalizace textů vnesla i do vzdělávací a vědecké oblasti závan čerstvého vzduchu. Vědci v pohodlí svého domova mohou bádat ve zpřístupněných věrných faksimiliích[8] historických dokumentů. Co jim však elektronický zdroj neposkytne – jsou jakékoliv informace o fyzické stránce informačního zdroje – např. kvalita papíru, vazby atd.

Ve vzdělávací oblasti by mohla mít digitalizace textu daleko větší uplatnění. Od zpřístupnění učebních textů, které se prakticky od papírové učebnice nezničí, včetně toho, že díky digitálnímu obsahu může být doplněna nejen obrazy ale i výukovým videem, interaktivním obsahem testů, které díky vyhodnocení poskytnou okamžitou zpětnou vazbu uživateli. Takovéto učebnice je pak samozřejmě daleko jednodušeji aktualizovat, aby odpovídaly současnému pojetí, bez nutnosti opravy nákladně tisknout.

Jistě – díky zavádění interaktivních dotykových tabulí do škol – se na trhu objevil prostor pro interaktivní učebnice, jejichž významným producentem je plzeňské nakladatelství Fraus. Takovéto učebnice jsou omezeny množství licencí a samozřejmě cenou.

[7] http://kramerius-info.nkp.cz/
[8] Věrná faksimile přesně zobrazuje daný zdroj – včetně poškození, vybledlosti, ručně psaných poznámek atd.

Další možností, kterou školy mohou využít, je e-learningový software – např. moodle, který je freewarový. Na druhou stranu, obsah si uživatel musí doplňovat sám a samozřejmě nesmí odporovat autorskému zákonu. Velkou výhodou je, že je přístupný i pro žáky či studenty, kteří z různých důvodů absentují ve škole, a potřebnou látku se mohou doučit doma.

V poslední době vycházejí vědecké studie, které se snaží osvětlit příčinu klesající úrovně schopnosti porozumět čtenému textu. V roce 2014 proběhla konference Mezinárodní společnosti pro empirická studia literatury a médií v italském Turíně, kde zazněla myšlenka, že e-reading vede k povrchnímu vnímání textu. Ziming Liu již v roce 2005 napsal studii, že čtenáři při čtení z obrazovky mají tendenci číst selektivně – tzn. jen ho očima procházejí a hledají klíčová slova. V důsledku to znamená, že jim uniká kontext a nemusí nad čteným textem přemýšlet [8].

Norská vědecká pracovnice Anne Mangenová potvrdila, že čtení online a z papíru má odlišný proces. Navíc rychlejší čtení se vypořádává s velkým množstvím dostupných informací, které jsou v současnosti online k dispozici.

Elektronické čtivo je jednodimenzionální [9]. To znamená, že čtenáře ochuzuje o vjemy, které prožívá během čtení papírové knihy – váha, vůně papíru, otáčení stránek. Tyto prožitky pak mají vliv i na paměť, protože si mozek daný text lépe uspořádá, oddělení textu na stránky mu pomáhá se v textu lépe orientovat – vnímá konec.

Navíc samozřejmě čtení souvisí i s fyzickou stránkou čtenáře – především zrakem. Elektronická zařízení jako mobil, tablet, iPad, monitor k očím příliš šetrná nejsou. Tím že vyzařují světlo, dochází k tomu, že čtenář méně mrká, čímž si zvlhčuje oči. Dále může způsobovat bolest hlavy a oční vady.

Alternativou jsou elektronické čtečky pracující s E-linkem. Čtečka není podsvícená, pracuje s denním světlem, tudíž šetří zrak čtenáře. I když se nové typy snaží být multifunkční – přehrávání hudby, připojení k internetu, nasvícení displeje z rámu čtečky pro večerní čtení, stále platí, že se jedná o úzce zaměřený výrobek.

Navíc ve spojení se vzděláváním, kdy je třeba zobrazit velkoformátovou učebnici, je třeba zvolit vhodnou velikost displeje ať už tabletu nebo čtečky.

## 1.3 ZRAKOVĚ POSTIŽENÍ ČTENÁŘI

Elektronické knihy asi nejvíce přivítají zrakově postižení čtenáři. Elektronická zařízení, která jim knihu předloží, jim nabízí funkci zvětšování písma případně hlasité předčítání textu ženským či mužským hlasem. Jistě je tu adekvátní náhrada v podobě audiobooků, jejichž pořizování ale není levnou záležitostí ať už při jejich výrobě či jejich koupi. Každopádně syntéza řeči, ač už v dnešní době je velice kvalitní, nedokáže plně vyvinout tempo čtení, které zvládá čtenář zrakem.

Existují samozřejmě digitální knihovny, které podléhají autorskému zákonu, ale na občany se zdravotním postižením – zejména zrakovým – je udělena výjimka, jak ve slovenské verzi autorského zákona č. 618/2003 Z.z., tak české verze – 121/2000 Sb. Slovenská specifikuje způsob veřejného rozšiřování, právě jen pro potřeby zdravotně postižených, aniž by byl třeba souhlas autora, nebo aby mu byla uhrazena odměna, s podmínkou nulového přímého či nepřímého majetkového prospěchu. Znění českého zákona je téměř totožné.

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

### 1.3.1 KDD (KNIHOVNA DIGITÁLNÍCH DOKUMENTŮ)

Tato knihovna byla založena už v roce 1993 pod patronátem organizace SONS (Sjednocené organizace nevidomých a slabozrakých). O deset let později byla přeměněna na knihovnu plně odpovídající klasických standardům[9] a v roce 2010 proběhla její poslední modernizace[10]. 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. Navíc byl systém, kvůli výpadkům v předchozích letech, přemístěn na výkonnější server, který je pod celodenním dohledem.

[9] Starší verze digitální knihovny je stále přístupná na adrese http://knihovna.braillnet.cz/
[10] Zatím poslední verze knihovního systému je na http://www.kdd.cz/

Aby byl čtenář zaregistrován do této knihovny, musí být občanem České republiky starším patnácti let, 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. V otázkách a odpovědích se dozvíte, že pokud toto nesplňujete, nemáte žádnou možnost, aby vám bylo dovoleno se zaregistrovat – ani jako případný dobrovolník s digitalizováním starších publikací.

Knihovna má zaregistrováno přes jeden a půl tisíce uživatelů, na které připadne téměř dvacet tisíc monografií (knih) a sedmdesát titulů periodik. S KDD spolupracuje celkem devadesát nakladatelství, mezi něž patří například Academia, Baronet, Portál, Grada, Euromedia Group či Metafora.

### 1.3.2 DIGIBOOKS

Tato knihovna vznikla na přelomu let 2003 a 2004, vlastně díky striktním pravidlům přijetí uživatelů do české KDD, která zakazují přijetí uživatele jiné národnosti než české. Peter Grosser tedy založil Občianske združenie INFOBLIND, která se zaměřila právě na digitalizaci textů.

Chod projektu je vlastně placen členskými příspěvky ve výši třinácti eur na jeden kalendářní rok, ale především pak příspěvkem sponzorů, kteří si odepíší 2% z daní. Knihovna pak nabízí možnost placení členského příspěvku formou ztráty 800 bodů, které uživatel získává například za korektury zdigitalizovaných textů, přidání nového souboru, založení nového autora či publikace.

Pravidla přijetí členů nejsou zdaleka tak přísná, jako má české obdoba. Členem se může stát občan Slovenska i České republiky, pokud je majitelem průkazu ZTP, ZTP/S a ZTP/P. Navíc se do knihovny může zaregistrovat i vidící člověk, který chce pomáhat s korekturami a digitalizací. Pro ty je připravený jiný typ smlouvy. Samozřejmě pro všechny členy platí přísné pravidlo zákazu jakéhokoliv vynášení zdigitalizovaných děl na jiné – pirátské – servery. Toto porušení se trestá doživotním vyloučením z knihovny.

Porovnáme-li množství zpřístupněných ebooků, tak čísla jednoznačně hovoří ve prospěch takovéto otevřené spolupráce, protože obsahově jsou obě knihovny totožné. Obsahují monografie, encyklopedie i periodika. V roce 2004 měla knihovna založeno 2000 autorů a 4 700 titulů. K lednu 2015 se jedná o skoro 23 tisíc autorů a 82 tisíc ebooků. Překvapivě

z toho obrovského množství je valná většina v češtině – 66 tisíc titulů. Šestnáct tisíc je slovenských.

Rozdíl mezi oběma knihovnami je zřejmě způsoben odlišným financováním, protože jinak pro obě platí totožné autorské zákony. Zcela nepochopitelným se u české verze jeví věkový limit uživatele, který předpokládá, že mladší nevidomí jedinci nečtou nebo neumějí pracovat s počítačem a speciálním softwarem. Nastavení pravidel vlastně limituje i rodiče, kteří mohou být vidomí, protože tím se opět vylučuje jejich registrace – nejsou držiteli karty ZTP.

# 2 PRINCIPY PŘEVODU TEXTU DO DIGITÁLNÍ PODOBY

Pokud si čtenář potřeboval knihu převést do digitální podoby, nezbylo lidem nic jiného, než co mnichům ve středověkých klášterech – byť k tomu používali zcela jiné nástroje, a tím byl mechanický přepis. Jednalo by se o činnost časově náročnou, s vysokou pravděpodobností výskytu lidské chyby a ne příliš produktivní. Proto přišli na řadu nové patenty, které měli v budoucnu tuto činnost zjednodušit.

## 2.1 HISTORIE OCR

Už v roce 1929 v Německu si Gustav Tauschek podal patent na 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.

Od roku 1950 se vývoj OCR programů rapidně zrychlil. V roce 1954 Reader's Digest přišel s OCR systémem pracujícím na bázi děrných štítků, které dokázal následně počítač zpracovat. Obsahem převodu byly finanční zprávy psané tiskacími stroji [10].

### 2.1.1 PRVNÍ GENERACE OCR

Roky mezi 1960 – 1965 jsou označovány jako pro komerční sféru OCR systémů jako první generace. Stroje se omezovali 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 fonty. Tento počet byl omezen, protože stroje stále pracovaly na bázi párování šablon, které porovnávalo obraz znaku s knihovnou prototypových obrazů znaků každého písma.

### 2.1.2 DRUHÁ GENERACE OCR

V roce 1965 byl na Světové výstavě v New Yorku představen nový OCR systém – IBM 1287. Tento stroj již dokázal číst běžně dostupné tištěné texty a nově i texty psané rukou – zejména detekci čísel. 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. Tyto sady byly navrženy a stylizovány tak, aby bylo usnadněno optické rozpoznávání a stále svým přirozenějším tvarem čitelnější pro člověka. Obě verze nalezneme v příloze 1.

### 2.1.3 TŘETÍ GENERACE OCR

Po roce 1970 se vývoj zaměřil především na kvalitu rozpoznávání i předloh nízké kvality – především ručně psaných znaků, dále zvyšování výkonu a snižování ceny. Tyto jednoduché OCR systémy, které především ještě stále využívaly neproporcionální písma, jejichž pevně stanovená šířka zvyšovala kvalitu převodu. Tyto koncepty psané psacími stroji, pak byly převedeny do počítače k finální úpravě. Tento systém se využíval ve zdravotnictví, žurnalistice, poštovním úřadě atd.

### 2.1.4 SOUČASNOST

Se snižováním nákladů na pořízení OCR softwaru, se programy staly dostupnějšími pro běžné uživatele osobních počítačů. Asi nejznámější a vysoce kvalitní komerční program je Abbyy Finereader, který podporuje až 200 jazykových verzí. V mobilních telefonech již existují aplikace, které provedou OCR analýzu na vyfotografovaném obrazu přímo v telefonu.

## 2.2 POJMY

Ve výše zpracované kapitole o historii vyplynulo, že existuje vícero druhů metod zpracování tištěného či psaného textu, které se liší svým zaměřením a technologií postupu. Takže se v této kapitole zaměřím na přiblížení některých pojmů souvisejících s daným tématem.

**OCR – Optical Character Recognition**

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.

**ICR – Intelligent Character Recognition**

Jedná se o pokročilejší technologii, která se dokáže učit. Uživatel označí neznámý znak, potvrdí jeho hodnotu vybraným písmenem, které program dokáže rozpoznat, a program jej při příštím rozpoznávání správně vyhodnotí. Jedná se o tvorbu vlastních uživatelských vzorů. Tato technologie je založena na konceptu neuronových sítí. Přesto její úspěšnost není 100%.

**IWR – Intelligent Word Recognitions**

Inteligentní rozpoznávání slov tvoří další stupeň vývoje OCR. Dokáže rozpoznat ručně psané znaky i další vlastnosti tištěného písma – sklon, tučnost atd. Shluk znaků (slovo, frázi) dokáže porovnat s databází a případné chybné znaky nahradí správnými.

**OMR – Optical Mark Recognition**

Jedná se o metodu, kterou se využívá při vyhodnocování speciálních formulářů. Prosvítí se papírový podklad, tam kde je ve formuláři odpověď vyplněna (zaškrtnuta, začerněna), tak v daném místě se zmenšuje průchod světla. Tmavé části jsou detekovány a vyhodnoceny. Prakticky se tento princip používá při spotřebitelských výzkumech, testování, loteriích, nebo na stejném principu funguje i čtení čárových a QR kódů.

## 2.3 FÁZE PROCESU ROZPOZNÁVÁNÍ ZNAKŮ

Obecně se systém OCR skládá z několika částí – kroků, postupů. Nejdříve se získá obraz, který je následně dolaďován optimalizačními nástroji. Následuje lokalizace a segmentace, které vyhodnotí, jak se které oblasti obrazu budou číst, a každý symbol se pak extrahuje procesem segmentace. Identita každého symbolu se porovná s jeho popisem ve znakové sadě. Nakonec se provede rekonstrukce slova, většinou v programech zobrazena v novém okně.

### 2.3.1 ZÍSKÁNÍ OBRAZU (OPTICKÉ SKENOVÁNÍ)

Je několik způsobů, jak se k digitálnímu obrazu předlohy dostat. Jedním z nich je fotografování dokumentu klasickým fotoaparátem nebo mobilním telefonem. Výsledky jsou však pak velmi ovlivněny kvalitou osvětlení plochy. Další možností je získání z jiných zdrojů dostupných např. z internetu. Nejpoužívanějším způsobem je ale skenování ať už robotickým, skenerem s podavačem nebo domácím stolním skenerem. Na kvalitě tohoto skenování závisí úspěšnost následujících kroků OCR. Více tento důležitý faktor proberu v kapitole Faktory ovlivňující kvalitu OCR převodu.

### 2.3.2 PREPROCESSING

Nasnímaný obraz lze ve většině OCR programů ještě upravit, aby byl výsledek co nejlepší. Tento proces má dvě části – tu první ovlivňuje uživatel, který pomocí editoru obrázku

ovlivní naskenovaný dokument – rozdělení dvoustran (někdy to program špatně rozpozná a nechá strany spojené), upravení kolinearity, kontrastu, orientace stránky atd.

Ta druhá je plně automatizovaná a jedná se právě o převod skenu do binárního obrazu.

### 2.3.3 LOKALIZACE A SEGMENTACE

V následujícím kroku se naskenovaný obraz rozčlení na oblasti (vychází z rozdílu barevnosti obrazu) – rozčlení se na textové, grafické, tabulkové bloky. Následuje segmentace – po řádcích a pak po jednotlivých znacích. Výsledkem je informace o pozici, výšce a šířce znaku.

### 2.3.4 ZNAKOVÁ EXTRAKCE

Slouží k získání základních charakteristických rysů jednotlivých symbolů. Má charakter křivky opisující vybraný – popisuje jeho typické znaky. 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) a rotace (orientace symbolu). Kvalita této extrakce je přímo úměrná počtu znaků, kterými je symbol popsán. Tyto znaky (příznaky) porovnává se svou databází a shodu pak dosadí do textového dokumentu.

### 2.3.5 LEXIKÁLNÍ POSTPROCESSING

Je ovlivněn podporou daného jazyka v OCR programu. Pracuje se statistickými modely výskytu slov v daném jazyce. Pomáhá tedy v případě chybně rozpoznaných znaků ve slově pomocí databáze vybrat ten nejvhodnější symbol.

# 3 FAKTORY PODÍLEJÍCÍ SE NA KVALITĚ OCR PŘEVODU

## 3.1 PROSTŘEDKY ZÍSKÁVÁNÍ OBRAZU

fotografie

skenování

- CCD x CIS snímač
- hodnota dpi
- kontrast

## 3.2 PŘEDLOHA

Použitý font – zdobné písmo, tabulky, čísla, vzorce, slévání fontů rn=m atd.

Řádkování – příliš široké vytvoří dva řádky textu – diakritiku a písmena rozdělí – háček jedna řádka, r druhá řádka

Spojování dlouhých dialogů do jednoho odstavce

Vazba a možnost přitisknutí ke sklu skeneru

## 3.3 SOFTWARE

### 3.3.1 KOMERČNÍ SOFTWARE

Abbyy FineReader

Microsoft Office Digital Imaging

Novodynamics verus

Omnipage

Readiris

SimpleOCR

Tesseract

### 3.3.2 FREEWARE

GOCR

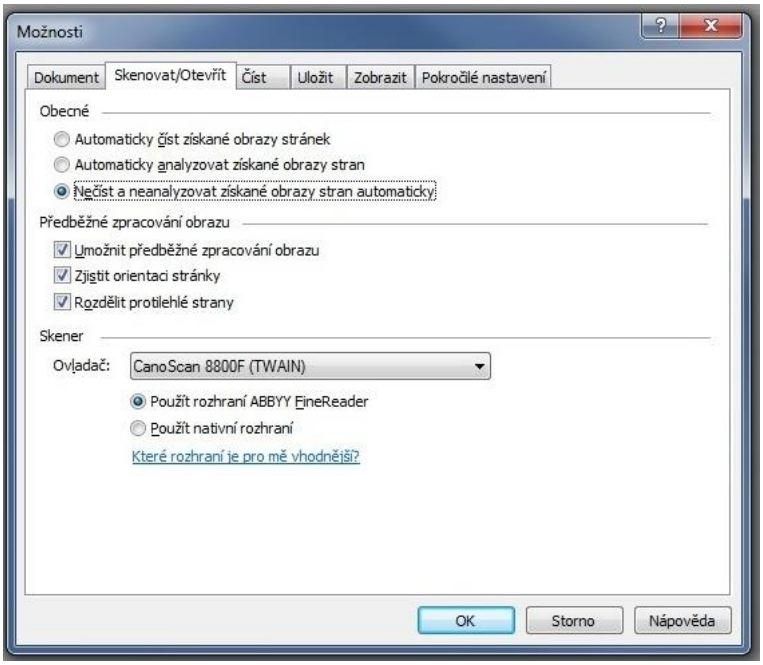
SimpleOCR

OCRAD

Online aplikace SharpEye

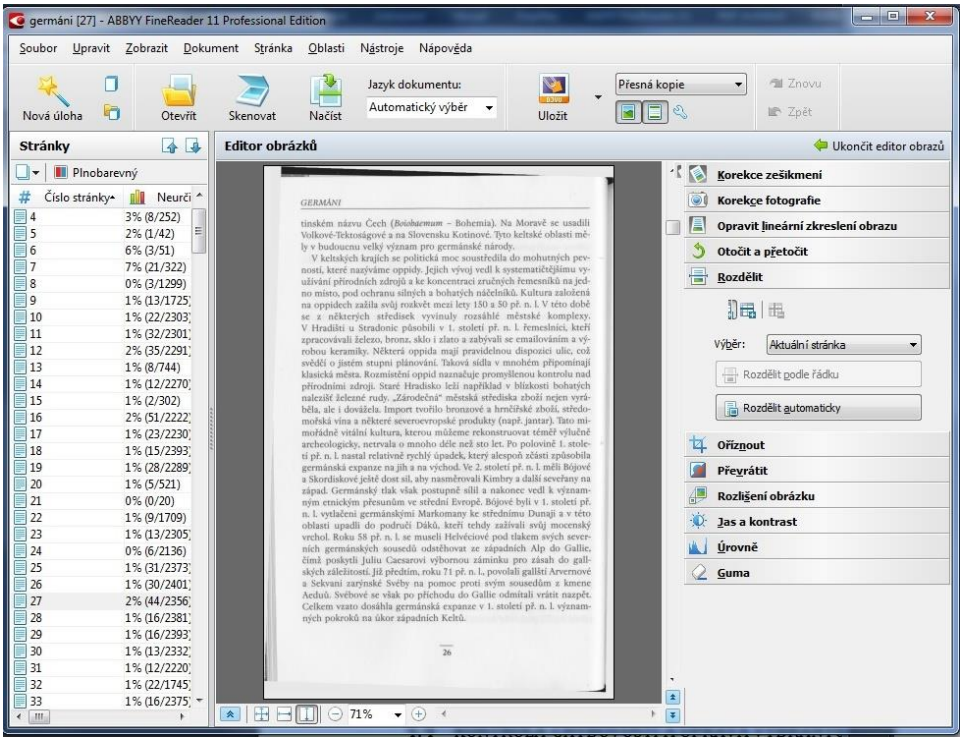
# 4 POSTUP DIGITALIZACE TEXTU V PROGRAMU FINEREADER ABBYY 11

## 4.1 NASTAVENÍ SKENOVÁNÍ



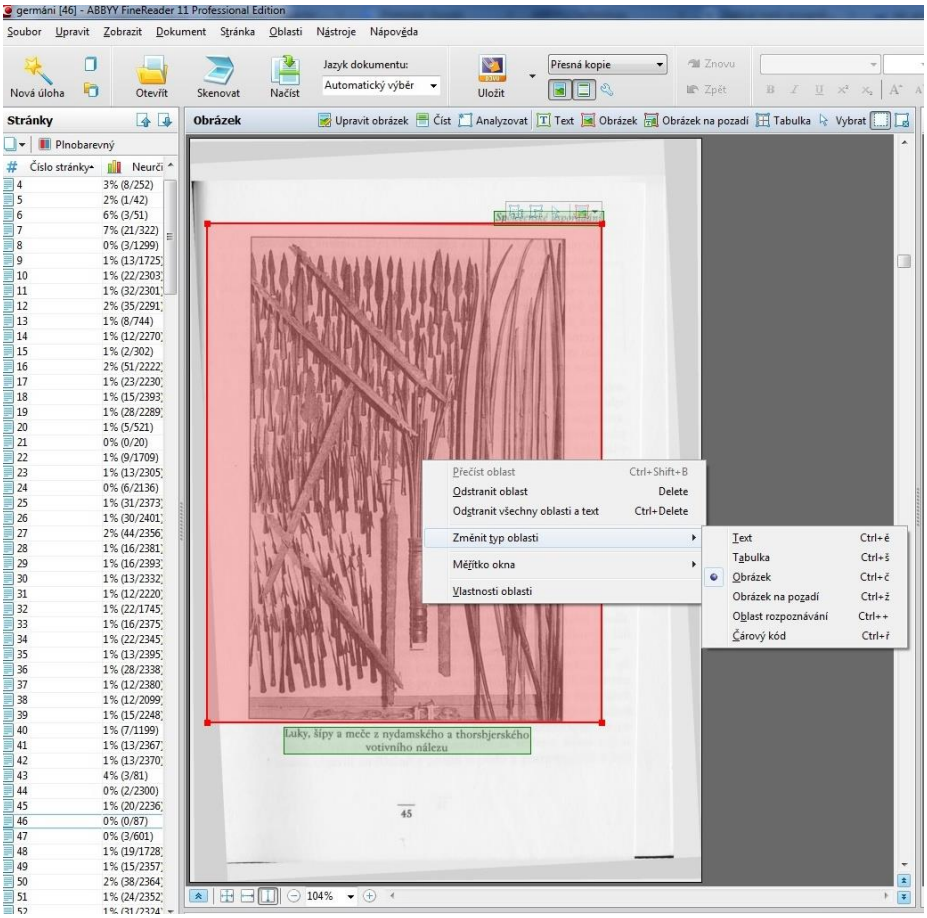

Obrázek 1 - dialogové okno pro nastavení skenování

## 4.2 ÚPRAVA OBRAZŮ



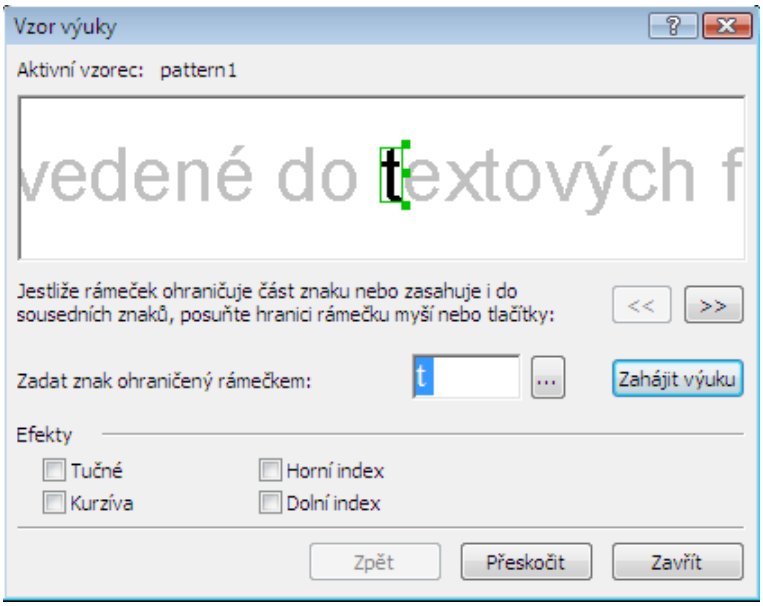

Obrázek 2 - editor obrazů pro úpravu předlohy

## 4.3 KONTROLA CHYBOVOSTI A ÚPRAVA VYBRANÝCH OBLASTÍ




Obrázek 3 - oblasti OCR

## 4.4 TVORBA UŽIVATELSKÝCH VZORŮ




Obrázek 4 - ukázka tvorby uživatelského vzoru v prostředí FRA 11

## 4.5 VÝBĚR FORMÁTU PRO EXPORT TEXTU

FineReader nabízí širokou škálu formátů souborů pro uložení digitálního textu. Mezi editovatelné patří docx, rtf, odt, a pak klasické čtečkové epub – případně odeslat do čtečky Kindle, obrazové DjVu, od

## 4.6 NÁSLEDNÁ KOREKTURA EDITOVATELNÉHO FORMÁTU

# ZÁVĚR

Jak z textu mé bakalářské práce vyplynulo, že 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.

## RESUMÉ

The aim of this work was an introduction to the process and principles of digitizing printed text, its current use in society, and a practical demonstration of such a transfer in the program FineReader 11th text was properly completed video excerpts on the 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. 2015-06-27]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001728&local_base=KTD

[2] http://www.gutenberg.org/wiki/Main_Page

[3] http://artslexikon.cz/index.php/Projekty_digitalizace

[4] http://gallica.bnf.fr/?&lang=EN

[5] http://ikaros.cz/europeana-online-pristup-k-evropskemu-kulturnimu-a-historickemu-dedictvi

[6] http://www.ndk.cz/digitalizace-1

[7] http://www.mkcr.cz/assets/literatura-a-knihovny/Koncepce_rozvoje_knihoven_2011-2015.pdf

[8] http://literarky.cz/civilizace/89-civilizace/18087-vdci-e-reading-vede-k-povrchnimu-vnimani-textu-zaveme-pomale-teni

[9] http://psychologie.doktorka.cz/z-monitoru-si-pamatujeme-mene-nez-pri-cteni-z-tisteneho-media/

[10] EIKVIL, Line. *Optical Character Recognition* [online]. 1993 [cit. 2015-06-30]. Dostupné z: http://www.nr.no/~eikvil/OCR.pdf

[11] http://www.abbyy.co.il/?categoryId=63424

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

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

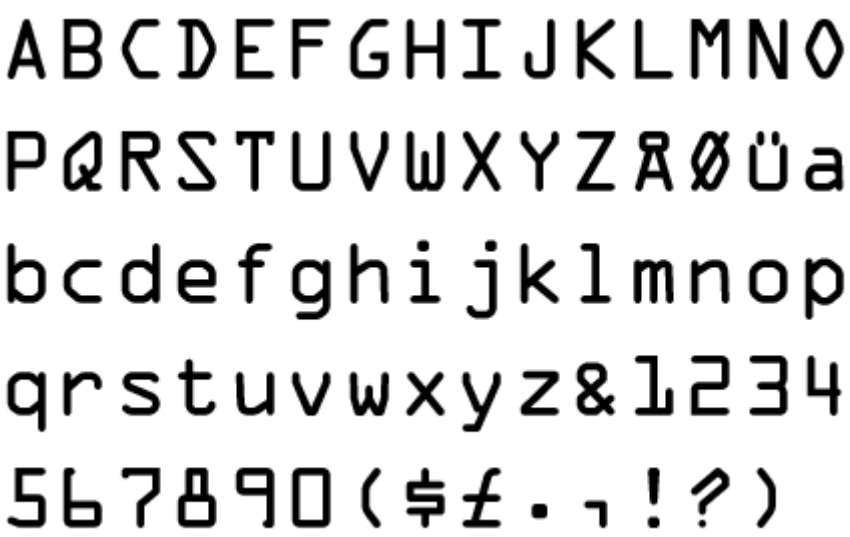
[14] ABBYY FineReader: *User's Guide for ABBYY FineReader 11: uživatelská příručka k aplikaci FineReader 11*. 1. vyd. ABBYY software, 2011, 110 s. Dostupné z: http://www.abbyy.com/fr11guide_cz.pdf

# SEZNAM OBRÁZKŮ, TABULEK, GRAFŮ A DIAGRAMŮ

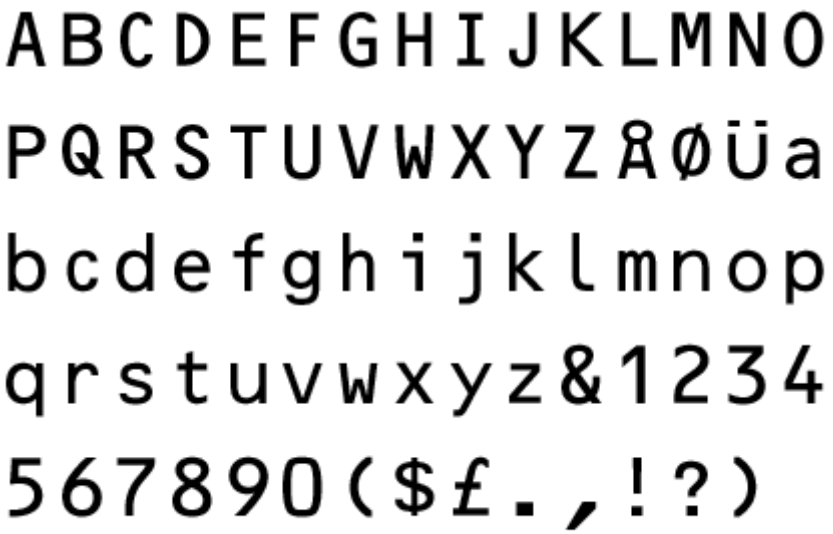
Obrázek 1 - dialogové okno pro nastavení skenování ..... 20
Obrázek 2 - editor obrazů pro úpravu předlohy ..... 20
Obrázek 3 - oblasti OCR ..... 21
Obrázek 4 - ukázka tvorby uživatelského vzoru v prostředí FRA 11 ..... 21

# PŘÍLOHY

Příloha 1 – ukázky standardizovaných znakových sad pro USA a Evropu




znaková sada OCR-A




evropská znaková sada OCR-B