Softwarová aplikace pro elektronickou archivaci textových dokumentů

A software application for electronic archiving of text documents

Bc. Michal Provazník

Diplomová práce
2010

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky
akademický rok: 2009/2010

ZADÁNÍ DIPLOMOVÉ PRÁCE
(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Michal PROVAZNÍK**
Studijní program:	**N 3902 Inženýrská informatika**
Studijní obor:	**Informační technologie**
Téma práce:	**Softwarová aplikace pro elektronickou archivaci textových dokumentů**

Zásady pro vypracování:

1. **Vypracujte literární rešerši na dané téma.**
2. **Vytvořte aplikaci na převod tištěných materiálů do elektronické podoby (například metodou OCR) dle normovaný rozměrů.**
3. **Vytvořte univerzálního rozhraní mezi USB Scan a PC, dále vytvořte aplikace v jednom z prostředí (DotNet, Visual C++, Delphi, SQL, Oracle).**
4. **Vytvořte nástroje pro napojení databáze na další aplikace.**
5. **Navhněte optimální hardwarovou konfiguraci.**

Rozsah práce:

Rozsah příloh:

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

1. **ULLMAN, L. PHP a MySQL. Computer Press, Brno, 2004, ISBN: 80-251-0063-4.**
2. **CHERIET, Mohamed. Character Recognition Systems: A Guide for Students and Practitioners. Wiley-Interscience, 2007. 360 s. ISBN 978-0471415701.**
3. **LIBERTY, Jesse. Programming .NET 3.5. OReilly Media; 1 edition, 2008. 480 s. ISBN 978-0596527563**
4. **HANAK, Jan. C 3.0 - Programování na platformě .NET 3.5. 2009. ISBN 978-80-7413-046-5.**
5. **TOM, Archer. Myslíme v jazyku C : knihovna programátora, Grada publishing, Praha, 2002. 308 s. ISBN 80-247-0301-7**
6. **MOLINARO, Anthony. SQL : Kuchařka programátora, Computer Press, 2009. 576 s. ISBN 978-80-251-2617-2**
7. **KALINA, Tomáš, PhDr., KUNT, Miroslav, Ing. Elektronická archivace -- výzva pro odborníky více oborů. Národní archiv [online]. 2005 [cit. 2010-02-09]**

Vedoucí diplomové práce: **Ing. Roman Šenkeřík, Ph.D.**
Ústav informatiky a umělé inteligence

Datum zadání diplomové práce: **19. února 2010**

Termín odevzdání diplomové práce: **8. června 2010**

Ve Zlíně dne 19. února 2010

prof. Ing. Vladimír Vašek, CSc.
děkan

prof. Ing. Vladimír Vašek, CSc.
ředitel ústavu

ABSTRAKT

Cílem této diplomové práce je návrh a implementace softwarové aplikace, která převádí obsah osobních dokladů do textové podoby pomocí skeneru s využitím technologie OCR. Takto získaná data se pomocí této aplikace ukládají do navržené databáze. V práci je dále rozebírána otázka archivace a její důsledky pro lidstvo.

Klíčová slova: Aplikace, Archivace, USB, OCR, Skener, SQL, C#, Občanský průkaz

ABSTRACT

This thesis proposes a software application that converts the contents of personal documents into text format by using a scanner with OCR technology. Collected data is stored to database. This document analyzed the issue of archiving and its implications for humanity.

Keywords: Application, archiving, USB, OCR, scanner, SQL, C#, Identity card

Na tomto místě bych rád poděkoval svému vedoucímu Ing. Romanu Šenkeříkovi, Ph.D. za vedení této diplomové práce, cenné rady, připomínky, trpělivost a také čas věnovaný mé práci. Dále bych rád poděkoval Ing. Miroslavu Macovi za nápady, náměty a inspiraci.

V neposlední řadě chci poděkovat své rodině za podporu při studiu a trpělivost.

Motto:

Kdo ovládá minulost, ovládá budoucnost: kdo ovládá přítomnost, ovládá minulost.

George Orwell

Prohlašuji, že

- beru na vědomí, že odevzdáním diplomové/bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že diplomová/bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk diplomové/bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně a jeden výtisk bude uložen u vedoucího práce;
- byl/a jsem seznámen/a s tím, že na moji diplomovou/bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – diplomovou/bakalářskou práci nebo poskytnout licenci k jejímu využití jen s předchozím písemným souhlasem Univerzity Tomáše Bati ve Zlíně, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše);
- beru na vědomí, že pokud bylo k vypracování diplomové/bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky diplomové/bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem diplomové/bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji,

- že jsem na diplomové práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze diplomové práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně …….

podpis diplomanta

OBSAH

ÚVOD

Soudobá koncepce světa nás všechny stále přesvědčuje o tom, že se nic neděje izolovaně od celku. Většina událostí a jevů má dnes svůj původ v nezměrném množství rozmanitých skutečností, které spolu souvisejí, navzájem se ovlivňují, doplňují se a společně vytvářejí komplexní obraz světa, kde žijeme. Díky dané skutečnosti dospějeme k faktu, že žijeme v malém světě, kde je vše spojeno se vším ostatním.

Z bohaté historie naší civilizace je třeba uchovávat nejen náš genotyp, ale i vědění, zkušenosti a myšlenky. Již po staletí si lidé tyto informace vyměňovali z generace na generaci. Každá generace vnesla do světla poznání svou hodnotu, naopak některé starší hodnoty a myšlenky byly navždy vrženy do propasti zapomnění. Ale které hodnoty jsou špatné s které dobré? Historie nás učí, jak chápat historii. Pochopení vyžaduje naučit se myslet podobně, jako lidé v dané době.

Před 5000 lety bylo vynalezeno písmo. Vynález písma byl obrovský objev. Miliony let historie pravěku skončily a započala skutečná historie člověka – starověk. Konečně již bylo možné myšlenky největších soudobých myslitelů a vynálezců uchovávat. Před tímto vynálezem nemusely být některé velké myšlenky lidí pochopeny, a tak se na ně zapomnělo, neboť nebyl nikdo, kdo by tuto dále předával. Dalším ohromným objevem byl knihtisk, vynalezený Johannem Gutenbergem kolem roku 1447. Masové rozšíření knih do celého světa bylo epochální – vzdělaní lidé se již mohli velmi snadno se svými myšlenkami dělit s ostatními.

Archivace informací ve formě knih a svazků s lidstvem vydrželo až do 80 let minulého století. Výpočetní technika společně s vynálezem Internetu, nacházela a stále nachází nové možnosti využití. Je tedy přirozené, že tyto technologie našly své uplatnění i v archivaci dokumentů.

Otázkou, zda dokumenty, jako jsou knihy, ztratí své opodstatnění a budou zcela nahrazeny digitálními nebo digitalizovanými protějšky, se této práci zabývat nebudu, nýbrž chci poukázat na některé atributy digitálních dokumentů a z toho odvozené metody archivace za účelem jejich dlouhodobého zpřístupnění, tj. problému, kterému musí čelit každá instituce, která se rozhodne zařazovat digitální dokumenty do svého fondu. Dále se budu zabývat digitalizací.

Myšlení nás, lidí, je ale pořád stejné. Mění se jen naše možnosti. Díky tomu půjde vývoj světa stále dopředu, neustále budeme zkoušet nové věci, a z těch starých si vezmeme to nejlepší. Jak se asi naši následovníci budou dívat na dnešní naše snažení?

I. TEORETICKÁ ČÁST

1 DIGITÁLNÍ DOKUMENT

Digitálním dokumentem, též elektronickým dokumentem (dále jen ED) se rozumí veškeré soubory digitálních záznamů, které zprostředkovávají textové (alfanumerické), obrazové a zvukové informace a které vznikají, uchovávají se a využívají se pouze prostřednictvím výpočetní techniky. ED (označované někdy v souladu s terminologií prosazovanou v mezinárodní standardizaci v oblasti bibliografického popisu také jako "elektronické zdroje" [1]) začínají být od druhé poloviny 90. let považovány (jak z teoretického, tak z praktického hlediska) za legitimní součást publikační produkce v národním i globálním měřítku. Tato situace pochopitelně znamená zásadní změnu ve fungování informačních institucí, jejíž důsledky se zatím neprojevily v plném rozsahu. V širším kontextu by se měla zařadit do druhé etapy procesu implementace informačních a komunikačních technologií do informačních služeb. V první fázi usnadňovaly správu a využívání knihovních fondů a umožňovaly vzdálený či lokální přístup k informačním systémům. Pro druhou fázi charakteristické, že se uplatňuje jednak při konverzi analogových dokumentů do jejich elektronické reprezentace, která je nabízena jako jejich náhrada na základě individuální objednávky nebo zpřístupněna plošně prostřednictvím počítačové sítě, jednak při vytváření digitálních dokumentů ve hmotné či nehmotné podobě jako výstupů elektronického publikování, které nemají analogovou předlohu (v této souvislosti se objevuje koncept digitální nebo hybridní knihovny).

Díky rozšíření prostředí World Wide Web (dále jen WWW) do podvědomí široké veřejnost, roste přirozeně i zájem knihoven o digitální dokumenty. WWW nabízí možnost poměrně snadno uskutečňovat soukromé publikační aktivity, ale rovněž ho lze efektivně využívat pro oficiální komunikaci (ve vědecké, akademické a podnikové sféře a ve veřejné správě), která měla dosud z různých důvodů v dominantní míře psanou nebo tištěnou formu (tj. formu fyzicky existujících dokumentů složených z hmotného nosiče informací a "množinou dat nebo informací, které jsou na něm (nebo v něm) fixované a formálně a obsahově uspořádané" [2]). Tomuto tématu je oprávněně věnována značná pozornost a z diskusí, které probíhají na různých úrovních, vyplývá, že knihovny (zejména ty, které mají depozitní funkci) se zaměřují právě na druhou skupinu digitálních online distribuovaných dokumentů.

1.1 Vlastnosti digitálního dokumentu

Obecné vlastnosti ED jsou odvozeny z podstaty digitálního záznamu. Analogový záznam je založen na signálu se spojitě proměnlivým průběhem (např. elektromagnetický záznam zvuku na magnetofonovém pásku), přičemž k jeho fixaci je třeba aplikovat specifický typ nosiče podle druhu informací a příslušné dekódovací mechanické, optické, elektrické nebo elektronické zařízení (v tomto smyslu jde tedy o strojem čitelný záznam, resp. dokument). Toto vymezení z fyzikálního hlediska se však v praxi pokládá za příliš rigidní, a proto se k analogovému záznamu řadí také text, který je zapsaný nebo vytištěný na papíře či jiném materiálu a který je tak vnímatelný zrakem.

1.2 Digitální záznam

Reprezentace digitálního záznamu je v terminologii počítačových věd zastoupena posloupností znaků binární soustavy ("0" a "1"), tj. kódu, se kterým umí pracovat pouze počítač, a teprve pomocí výstupních zařízení (monitor a tiskárna) lze převést digitální záznam do analogové, člověku srozumitelné podoby. K hlavním rysům tohoto kódu patří univerzálnost (slouží k vyjádření textových, obrazových, zvukových a audiovizuálních informací, resp. dat) a maximální redukce znakové sady. Digitální záznam je principiálně nezávislý na konkrétním nosiči, který tak přestává plnit svou původní roli prostředku k časoprostorovému transferu informací, a informace takto zachycené jako by z pohledu uživatele ztratily "hmotnou" podobu.[3]

Typické odlišnosti digitálního a analogového dokumentu ukazuje následující tabulka:

Digitální dokument	Analogový dokument
dekódování počítačem nebo jím řízenými periferiemi	dekódování strojem nebo lidskými smysly
proměnlivost	stálost
hypertextová / hypermediální struktura	lineární struktura
multimedialita	unimedialita
stavebnicový charakter	celistvost a sekvenčnost
neztrátová reprodukce	ztrátová reprodukce
snadná formální transformace	obtížná formální transformace
distribuovanost (možnost on-line přístupu)	lokalizovanost
snadná kontrola integrity záznamu	obtížná kontrola integrity dat
interaktivnost	jednostranné působení

Tab. 1. Odlišnosti digitálního a analogového dokumentu

1.3 Skladba digitálního dokumentu

Samotný ED - a nejen on - se skládá ze tří složek, tvořících úplnou informaci, na které musíme brát ohled:

1. obsah, tj. vlastní informace zaznamenané v ED
2. struktura, tj. organizace obsahu
3. kontext, tj. vazby na jiné ED

1.4 Dělení digitálních dokumentů

Digitální dokumenty lze rozdělit podle tří kvalitativních hledisek do těchto kategorií:

1.4.1 Dělení podle původu

- dokumenty primárně digitální (např. multimediální aplikace, některé elektronické časopisy)

- dokumenty existující paralelně v tradiční a digitální formě (většina odborných časopisů)
- dokumenty převedené z tradiční do digitální formy (fotografie vystavené na Internetu, pokud nebyly pořízeny digitálním fotoaparátem)

1.4.2 Dělení podle stupně proměnlivosti [4]

- statické dokumenty (fixní forma a obsah – např. dokumenty ve formátu PDF)
- dynamické dokumenty (fixní forma a proměnlivý obsah – dynamické www stránky)

1.4.3 Podle vztahu k digitální (hybridní) knihovně [5]

- externí on-line dokumenty (webové stránky)
- externí off-line dokumenty (audionahrávky na CD-DA)
- dokumenty, které vznikly při realizaci projektu digitalizace (digitální kopie středověkých rukopisů)

2 PARADIGMATA ARCHIVACE DIGITÁLNÍCH DOKUMENTŮ

Díky rozdílům mezi analogovým a digitálním záznamem nelze na digitální dokumenty aplikovat stejný koncept ochrany jako na dokumenty analogové. P. Conway tento koncept označuje jako "odpovědnou správu" (responsible custody) [6]. Tento koncept se primárně soustřeďuje na uchování jejich nosiče (jakožto faktoru ovlivňujícího informační hodnotu zaznamenaných informací) v takovém fyzickém stavu, který umožňuje zpřístupnění jejich intelektuálního obsahu (důraz se klade na preventivní opatření, která mají omezit působení degradačních činitelů - vhodné klimatické podmínky úložného prostoru a pravidla využívání dokumentů).

Digitální archiv sice zpřístupňuje spravované dokumenty, ovšem předmětem ochrany se namísto ochrany celého dokumentu, jakožto i nosiče informace, stává pouze integrita samotného digitálního záznamu, kterou nelze zúženě interpretovat jako pouhý přenos záznamu z jednoho média na druhé (strategie archivace digitálních dokumentů) - nosič ustupuje do pozadí, což je patrné především, ale nikoliv výhradně, u dokumentů šířených po síti. Přidá-li se i závislost digitálních dokumentů na technickém prostředí, pak se tento problém stává mnohem komplexnějším.

Ochrana či archivace digitálního záznamu (v anglické terminologii "digital preservation"), je definována jako soubor vzájemně provázaných opatření a metod technické a organizační povahy týkajících se uložení, administrace a zpřístupnění digitálního záznamu (digitálních objektů), jejichž smyslem je zabezpečit, že bude možné jeho dekódování v dlouhodobé perspektivě (tj. po dobu, která není předem ohraničena) s vědomím, že vlastnosti technických prostředků, které budou k tomuto účelu aplikovány v budoucnosti, nelze v současnosti dostatečně popsat [7].

ED a jejich interpretace v informačních technologiích, mají velkou výsadu, a sice jejich snadná, rychlá a flexibilní možnost modifikace a možnost zpřístupnění. Je třeba mít na paměti, že tyto dokumenty jsou i velmi křehké. Zatímco knihy přečkaly několik stovky let a stále jsme schopni je přečíst (navzdory technologickým změnám v jejich výrobě jde o týž informační systém), digitální dokumenty bez ohledu na jejich fyzickou životnost mohou snadno ztratit svou funkčnost tím, že přestanou být k dispozici dekódovací zařízení. K. Rusell konstatuje, že digitální dokumenty jsou pouze "signály", které je třeba obnovovat, jinak navždy zmizí [8].

V informačním věku stoupá význam a četnost elektronického publikování. Díky tomuto faktu může vzniknout iluze, která může vést k podcenění problému ochrany digitálních dokumentů. Informace má stále menší časovou platnost, navíc slouží pouze k "okamžité spotřebě", je třeba se smířit s tím, že jejich životnost je a priori krátká, a v případě nutnosti ji lze libovolně prodloužit pořízením nekonečného množství kopií bez ztráty kvality. Instituce, zabývající se touto problematikou, by k tomuto problému měly pohlédnout a zbytečně na tento typ informací neplýtvat svými prostředky.

3 STRATEGIE ARCHIVACE DIGITÁLNÍCH DOKUMENTŮ

Proces dlouhodobé archivace a zpřístupnění digitálních dokumentů v rutinním režimu tvoří z technického a také z administrativního hlediska jednu ze čtyř strategií, která je implementována s cílem umožnit překlenout morální stárnutí technologií, aniž by došlo k nežádoucímu narušení integrity digitálního dokumentu (vedoucí v praxi k redukci jeho informační hodnoty), který lze chápat jako soustavu elementárních digitálních objektů. Při formulování strategie je třeba brát v úvahu disponibilní finanční zdroje, technické a personální zázemí a v neposlední řadě druh dokumentů, které mají být v digitálním archivu uchovány.

3.1 Metadata

Velmi důležitá role v předmětu archivace ED patří metadatům. Metadata jsou obecně definována jako strukturovaná data o jiných datech. Metadata jsou sice data od primárních dat odvozená, ale nehrají podružnou roli, neboť za prvé činí primární data srozumitelnými, tj. zajišťují jejich dekódování, a za druhé stanoví rámec, v němž je možné je využívat. Z tohoto hlediska je hlavním smyslem metadat poskytovat přidanou informační hodnotu k primárním datům. Dlouhou dobu byla metadata na prvním místě spojována s dokumenty zpřístupněnými v prostředí WWW, a to v souvislosti s několika iniciativami, u jejichž zrodu stály knihovny a další informační instituce. Cílem těchto projektů je usnadnit jednak jmennou a věcnou klasifikaci těchto zdrojů se specifickými vlastnostmi a jejich účinné vyhledávání.

Struktura „dat o datech“ by měla být čitelná multiplatformě, dále by měla být čitelná i pro člověka – nejlépe je tedy mít uloženy jako prostý text podle všeobecně přijatelného standardu (např. Unicode).

Díky věcnější interpretaci informace jsou metadata popisující informační hodnotu dokumentu označována jako popisná metadata a metadata popisující softwarovou platformu, hardwarovým vybavením a v neposlední řadě s přístupovým softwarem se označují jak technická metadata.

Popisná metadata je třeba doplnit technickými metadaty, která především zabezpečují integritu složeného dokumentu, specifikují mapu vyskytujících se digitálních objektů, přesně definují vztah tohoto dokumentu k nosiči, na kterém (kterých) je aktuálně fixován,

aby mohl být dokument v extrémním případě interpretován. Tento typ metadat může dále zahrnovat údaje o původním softwarovém a hardwarovém prostředí pro jejich eventuální budoucí emulaci a - v případě digitalizovaných dokumentů - technické parametry snímacího zařízení, pomocí něhož byly digitální kopie pořízeny, za účelem optimalizace jejich pozdějšího dekódování.

3.2 Strategie archivace

K základním strategiím, patří migrace, emulace digitálního prostředí a technologické muzeum a konverze digitálních dokumentů do analogové formy. Přestože první z nich je jednoznačně považována za nejperspektivnější, ani její stoupenci ji neoznačují za optimální.

3.2.1 Migrace

Nejčastěji v reálných podmínkách uplatňovanou archivační strategii (nejen v informačních institucích, ale i v podnikové sféře) představuje migrace, která sice vyžaduje značné investice, ale z dlouhodobého hlediska je finančně efektivní. Při migraci dochází k periodickému transferu digitálních dokumentů ze starší generace digitálního prostředí, které je morálně zastaralé, do generace mladší. Primárně je tak věnována pozornost obsahové složce těchto dokumentů. Termínem "digitální prostředí" se rozumí hardwarová a softwarová platforma a aplikační software.

Migrace je proces v podstatě nevyhnutelný, což je dáno kontinuálními změnami v oblasti digitálních informačních technologií. Těmito změnami se v oblasti informačních technologií rozumí vývoj jak formátů souborů, jež definují datovou část digitálních dokumentů, tak i znakových sad textových dokumentů. Z hlediska ochrany digitálního záznamu ideální, avšak v praxi nedosažitelný stupeň univerzálnosti, kdy je zaručena dlouhodobá čitelnost digitálního dokumentu bez ohledu na dané digitální prostředí, v němž je využíván, nemá smysl uvažovat.

Migrace zpravidla zahrnuje dílčí operaci - kopírování digitálního záznamu, aniž by bylo nutné jej modifikovat, na nový nosič (tzv. refreshment), které se dnes provádí v podstatě v rutinním režimu ze dvou důvodů: buď fyzická životnost konkrétního nosiče (např. CD-ROM) se chýlí ke konci, a proto hrozí nebezpečí, že záznam bude ztracen, nebo se oprávněně předpokládá, že aktuální typ nosiče se výhledově stane morálně zastaralým

(např. náhrada magnetooptického disku CD-ROM). Tato konverze se může týkat i datového formátu podle toho, k jakému účelu se digitální kopie mají využívat. Dalším opatřením, které je podporováno producenty softwarových aplikací a hardwarových zařízení, je zpětná kompatibilita (modernější systémy jsou schopny dekódovat starší digitální dokumenty - např. MS Word 2007 > MS Word 2003).

3.2.2 Emulace

Strategie emulace digitálního prostředí, která od počátku vzbudila značnou pozornost odborné veřejnosti, avšak její přínos nebyl dosud praktickými zkušenostmi potvrzen ve větším měřítku, je založena na opačném principu než migrace.

Zásady emulace publikoval S. B. Robertson v modelu Digital Rosetta Stone. J. Rothenberg [9]. Domnívá se, že univerzálním a potenciálně nejméně složitým způsobem ochrany digitálního záznamu je jeho archivace v originálním formátu spolu s originálním aplikačním softwarem a zajištění jeho funkčnosti a chování prostřednictvím imitace vlastností digitálního prostředí, v němž vznikl. Na rozdíl od migrace, kterou podrobil nesmlouvavé kritice (považuje ji za finančně, organizačně a časově náročnou a především riskantní strategii, neboť vykazuje příliš vysokou chybovost), emulace je postavena na jednotném softwarovém principu, který lze aplikovat kdykoliv a nezávisle na formátu daného dokumentu. Díky tomu lze překonat trvalou nejistotu z dalšího vývoje technologií, které jsou nezbytné pro dekódování digitálních dokumentů.

Aby bylo dosaženo zamýšleného efektu, je třeba identifikovat a zapsat množinu metadat, která jsou rozdělena do tří skupin a která jsou buď zapouzdřena v dokumentu, nebo uložena v externí databázi. V první skupině budou obsažena technická metadata reprezentující prvky původního digitálního prostředí (aplikační software a hardwarová a softwarová platforma), zatímco v druhé skupině bude specifikován samotný emulátor v takové podobě, aby mohl být interpretován jakýmkoliv překladačem, který bude v budoucnu vyvinut a instalován jako nadstavba nového digitálního prostředí, tj. aby mohl být spuštěn v aktuálním operačním systému a pomocí aktuálního hardwarového vybavení. Třetí skupinu tvoří technická dokumentace o využití emulátoru, popisná metadata vztahující se k danému digitálnímu dokumentu (včetně okolností jeho vzniku a jeho úpravách) ve formě prostého textového souboru. Dokument, emulátor a aplikační software

musí být fixován na nosiči, který je podporován stávajícím prostředím, z čehož vyplývá, že emulace počítá přinejmenším s kopírováním záznamu [9].

3.2.3 Technologické muzeum

Pod pojmem technologické muzeum se rozumí deponování digitálního záznamu v podstatě jako artefakt v originálním formátu a prostředí a na originálním nosiči. [5] Tato strategie, která předpokládá, že originální digitální prostředí nelze přesně reprodukovat, a proto je nanejvýš užitečné, aby nebylo konvertováno (migrace) nebo emulováno, se týká nejen digitálních dokumentů (zejména se ukazuje jako nezbytná v případě, že nedošlo včas k jejich migraci nebo kopírování), ale rovněž analogových dokumentů, k jejichž dekódování je nezbytné technické zařízení (např. diapozitivy, vinylové gramofonové desky atd.). Nevýhody tohoto řešení jsou patrné: počet kombinací hardwaru a softwaru pro jednotlivé dokumenty nebo jejich druhy z hlediska správy by nutně časem přesáhly únosnou mez. K zachování jejich funkčnosti by navíc bylo třeba komponentů, které nelze již na trhu získat. Naprostá svázanost s originálním prostředím by bránila zpřístupnění těchto dokumentů jinak než v lokálním režimu. [9 s. 12-13]

3.2.4 Konverze do analogové formy

Strategie konverze digitálních dokumentů do analogové formy se opírá o skutečnost, že ochranné metody aplikované u analogových dokumentů jsou dostatečně ověřeny. Nesporně vyšší stabilita tradičních materiálů v čase (pokud jsou uloženy v odpovídajících mikroklimatických podmínkách) a jejich nezávislost na proměnlivosti technických prostředků pro účely dlouhodobého uchovávání je vykoupena tím, že analogové kopie jsou ochuzeny o všechny přednosti svých předloh, kvůli nimž byly digitální dokumenty vytvořeny a které se projevují především při jejich zpracování a zpřístupnění (strojové vyhledávání, hypertext, multimedialita, databázová struktura, dynamičnost, interaktivnost aj.). Z tohoto důvodu nejde podle J. Rothenberga [8] o způsob, se kterým by se mělo v širším měřítku zabývat jako seriózním alternativním řešením problému ochrany digitálního záznamu. V případě statických textových dokumentů (uložených např. v rozšířených formátech PDF) může být tato strategie aplikována s cílem zajistit kopii, která bude k dispozici i tehdy, když elektronický originál nebude dále přístupný.

V úvahu přicházejí dva typy nosičů: papír a mikrofilm. V prvním případě (papír) má smysl v této souvislosti se zabývat pouze tiskem na tzv. permanentní papír, který ve srovnání s dosud běžně používaným, průmyslově vyrobeným papírem, který je méně odolný vůči vlhkosti a plynným exhalátům, obsahuje menší podíl kyselotvorných substancí a zásaditou složku o určitém množství jako alkalickou rezervu. Hodnota pH se pohybuje v rozmezí 7.5-10, díky čemuž by se životnost papíru splňujícího tato kritéria měla prodloužit až na stovky let. Druhou možností je převod digitálního záznamu na mikrofilm, který umožňuje zachytit černobílé digitální obrazy (bitonální mód) ve vysokém rozlišení (600 dpi). Stejně jako v předchozím případě platí, že při této konverzi dojde k zákonitému ochuzení dokumentu. Tento systém, při němž se uplatňuje opačný postup než hybridní technologie (produkce archivních mikrofilmů a digitalizace mikrofilmů druhé generace určené ke zpřístupnění), je znám pod zkratkou COM (computer-output microfilm). [10]

4 DIGITALIZACE ANALOGOVÝCH DOKUMENTŮ

Z předchozích kapitol jsou výhody ED oproti analogovým záznamům jasně patrny. Proces převodu proces převodu analogového záznamu na záznam digitální se nazývá digitalizace. Digitalizace, stejně jako většina technických aplikací, podléhá matematickým a fyzikálním zákonům, které tento proces limitují.

4.1 Lidské vnímání barev

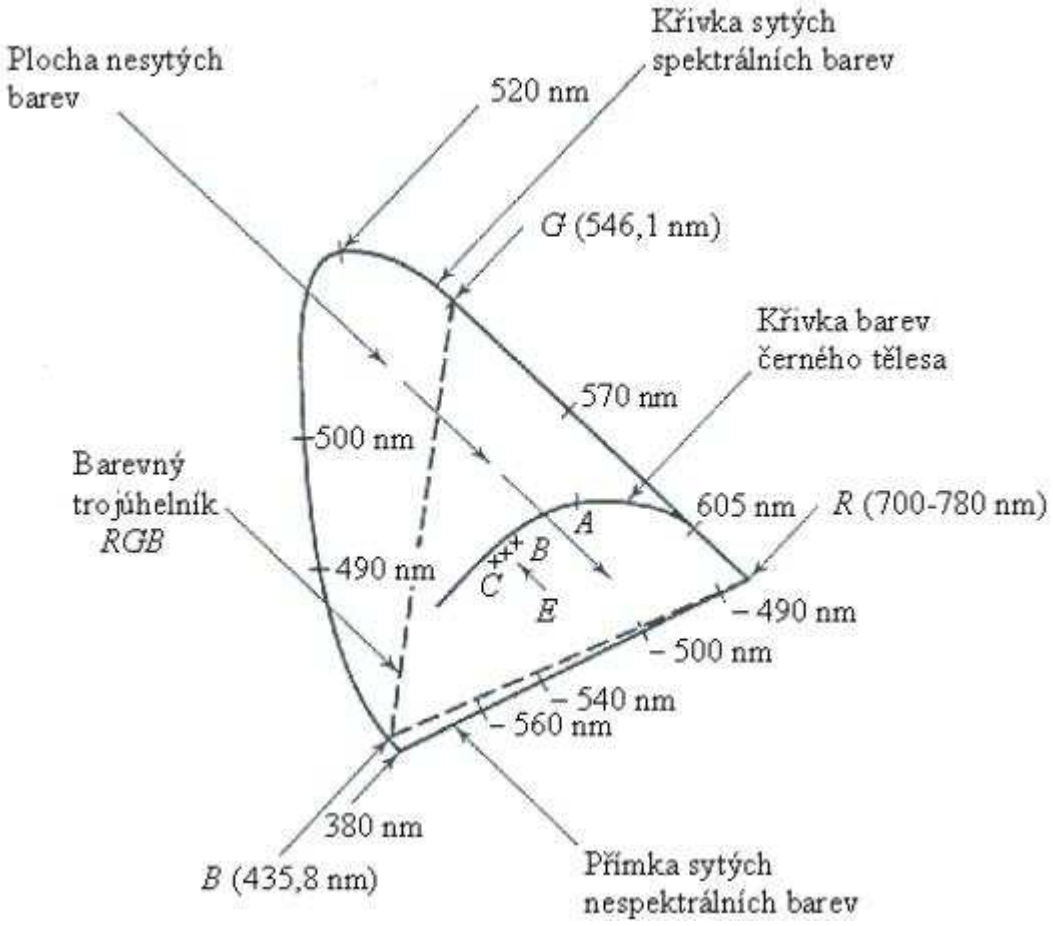
Lidské i zvířecí vidění je vysoce komplexní proces. I přes veškeré pokroky vědeckých znalostí v této oblasti, ale není dosud dopodrobna známo, jak lidský zrak přesně pracuje, a to zdaleka nejen třeba co se týká vyšší sémantické analýzy zrakových signálů mozkem (tj. rozpoznávání tvarů, objektů apod.), ale i co se týká procesů, které stojí na úplném počátku vidění, neboli vidění v jeho nejprimitivnější formě – formace zrakových signálů okem. Speciálně sítnice v oku, se svými světlocitlivými receptory a komplikovanými nervovými spojeními, vývojově patřící k mozku, je nesmírně složitý orgán. Jak funguje, je stále ještě do značné míry předmětem teorií a dohadů [11].

Barevné vidění je především záležitostí čípků, i když existují důkazy, že i tyčinky se na něm mohou za jistých okolností částečně podílet. To, že vidíme barevně, je způsobeno tím, že existují tři druhy iodopsinu – fotoaktivního pigmentu, jež čípky obsahují. Tyto pigmenty jsou spektrálně selektivní a každý druh je citlivý na jiný rozsah vlnových délek. Maximum citlivosti „modrých“ čípků se pohybuje kolem vlnové délky 440 nm, zatímco u „zelených“ čípků je to asi 540 nm a u „červených“ asi 570 nm. Červené a zelené čípky jsou si navzájem hodně podobné - většina savců je na rozdíl od člověka dokonce vůbec nemá takto rozlišené a místo nich má pouze jeden typ „žlutých“ čípků (takže vidí pouze dvojbarevně - podobně jako někteří barvoslepí lidé). Vlastnosti modrých čípků jsou podstatně výrazně odlišné. V sítnici je jich mnohem méně, odhadem jen asi 4%. Zelených čípků je asi 32% a zbylých 64% je čípků červených. Pro vysvětlení, proč tomu tak je, existují různé teorie, jedna např. tvrdí, že to snižuje vliv chromatické aberace čočky, jiná zase, že to kompenzuje vyšší podíl kratších vlnových délek v denním světle [11]. Lidské oko tedy vnímá barvu třemi základními barvami, a sice červenou, modrou a zelenou. Z těchto barev lze složit všechny možné, pro člověka viditelné, barvy.

Barevné modely interpretují chápání barev pro lidské oko. Modely pro digitální dokumenty ovšem vychází z fyzikálního popisu barev – diagramu barevnosti.

4.1.1 Diagram barevnosti

Všechny skutečné (reálné) barvy leží uvnitř tzv. plochy nesytých barev, která je v barevné rovině ohraničena křivkou sytých barev diagramu barevnosti (obr.1). Bodům, které leží mimo uvedenou plochu nesytých barev a jejího ohraničení, neodpovídá žádná skutečná barva.



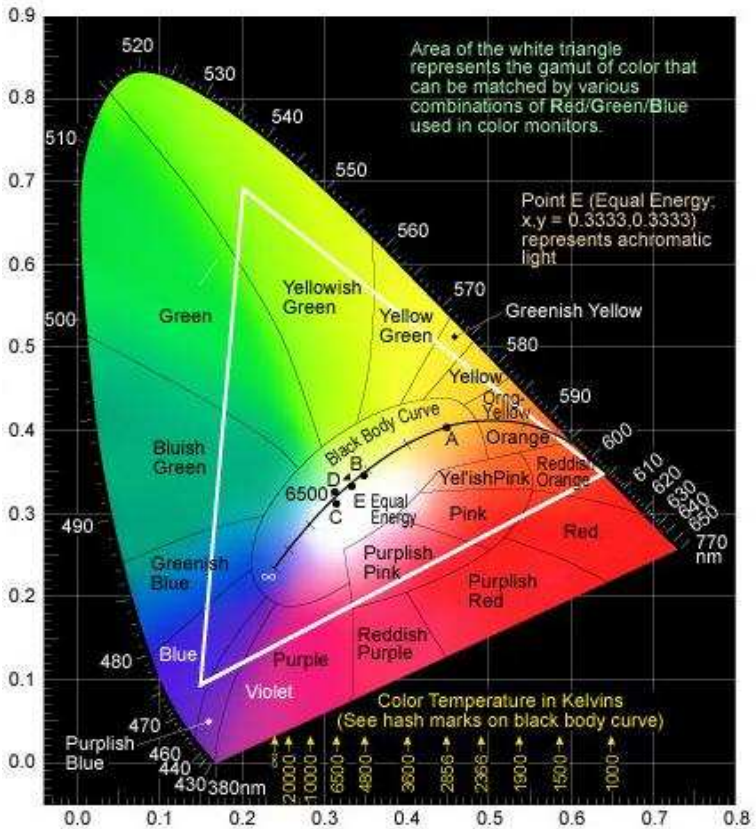
Obr. 1. Diagram barevnosti [13]

Křivka sytých barev má tvar zakřiveného trojúhelníku, leží na ní základní spektrální barvy *R*, *G*, *B* a uvnitř je barevný trojúhelník *RGB*. Její přímkovou základnu tvoří přímka sytých nespektrálních barev (přímka čistých purpurů), obsahující purpurové barvy charakterizované vlnovými délkami jejich doplňkových barev (s formálně připojeným znaménkem mínus), a její horní část reprezentuje křivka sytých spektrálních barev –

křivka spektrálních světel. Vzhledem k tomu, že všechny spektrální barvy o spektroskopických vlnových délkách λ = 700-780 nm prakticky způsobují stejný barevný vjem, bývá základní bod barevnosti *R* vztažen k tomuto celému vlnovému intervalu. Barvy světla, které vydává černé těleso (černý zářič) při daných absolutních teplotách, leží na křivce barev černého tělesa, v jeho blízkosti jsou body, které reprezentují smluvní bílá světla (na obr. 1 jsou znázorněny body příslušející smluvním bílým světlům *A*, *B*, *C*, *E*). [13]

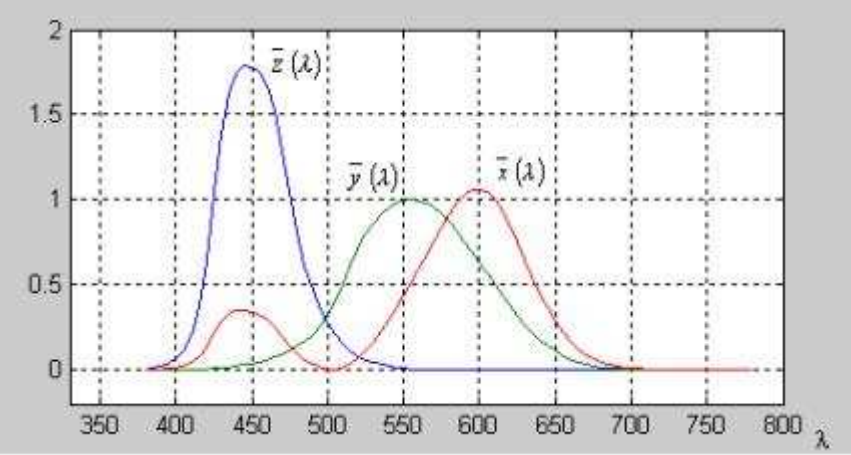
Každou barvu lze vyjádřit pomocí tří základních barev, jestliže jsou známé světelné toky základních barev $\Phi_1(R), \Phi_2(G), \Phi_3(B)$. Světelný tok se však nepoužívá jako barevná souřadnice. Podle mezinárodní normy používáme na kvantitativní určení barvy barevné souřadnice *X*, *Y*, *Z*, které jsou zvolené tak, aby souřadnice všech barev byly kladné a aby bod odpovídající bílému světlu ležel uprostřed barevné roviny. Protože na určení barvy stačí znát poměr souřadnic $X:Y:Z$, je výhodné používat raději relativní souřadnice, tzv. barevné (trichronometrické) koeficienty.

Na sestrojení barevného trojúhelníku (Obr. 2) je třeba znát spektrální složení světla (rozložení světelného toku podle vlnových délek), dále redukované barevné souřadnice



Obr. 2. Barevný diagram RGB [14]

pro každou barvu, které se dají vypočítat ze standardních křivek tj. z křivek na skládání spektrálních barev (Obr. 3).



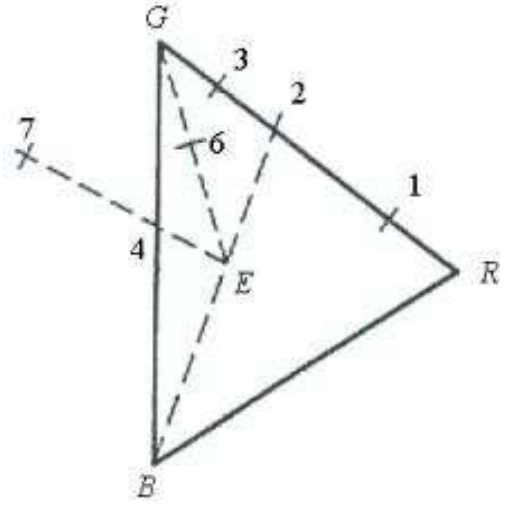
Obr. 3. Standardní barevné křivky [14]

4.1.2 Barevný trojúhelník RGB – aditivní mísení barev

Barva se obyčejně určuje třemi veličinami: barevným tónem, sytostí a světelným tokem. Barevný tón určuje spektrální barvu, která se musí složit s bílou, aby se dostala požadovanou barvu. Vlnová délka, která této spektrální barvě přísluší, se nazývá převládající (dominantní) vlnová délka. Sytost barvy je definována poměrem světelného toku příslušejícím světlu daného barevného tónu k celkovému světelnému toku (tj. součtu světelného toku syté a bílé barvy). Světelný tok vyjadřuje schopnost světla vyvolat zrakový vjem, ale na světelném toku nezávisí barva světla. Závisí však na poměru světelných toků světla o různých frekvencích. Na přesné určení barvy proto stačí znát jen dominantní vlnovou délku a sytost [12].

Barvu světla lze znázornit bodem v barevné (kolorimetrické) rovině. Při znázorňování se vychází ze tří barev, které se volí za základní (základní body barevnosti). Podle mezinárodní normy jsou těmito barvami tři spektrální barvy: *R* = červená (red, purpurově červená) základní barva – λ = 700 nm, *G* = zelená (green, žlutozelená) základní barva – λ = 546,1 nm, *B* = modrá (blue, fialově modrá) základní barva – λ = 435,8 nm. V oblasti lidského vidění odpovídají barevným citlivostem čípků sítnice lidského oka při trojbarevném vidění.

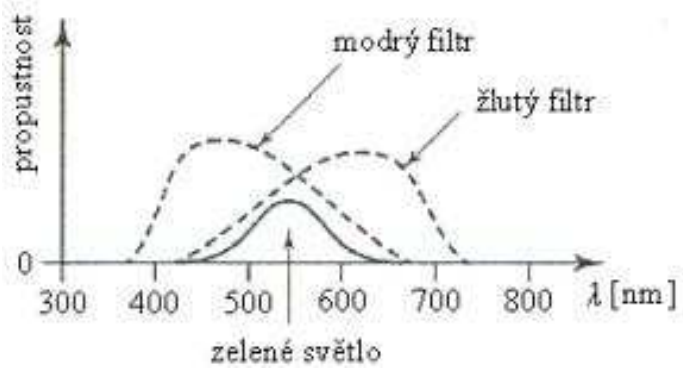
Třem základním barvám přísluší na barevné rovině body *R*, *G* a *B* (obr. 4). Na znázornění barev, které vyjdou aditivním skládáním dvou základních barev, se použije pravidlo, podle kterého bod znázorňující složenou barvu, leží na spojnici bodů těch barev, jejichž složením barva vznikla. Všechny barvy, které můžeme získat složením dvou základních barev, leží na stranách trojúhelníka *RGB*, a to blíže k té základní barvě, které větší část obsahují. Například míšením červené barvy R a zelené barvy G vznikne podle jejich zvoleného poměru barva oranžová (bod 1 na obr 21), žlutá (bod 2) a žlutozelená (bod 3).



Obr. 4. Barevný trojúhelník RGB [13]

4.1.3 Subtraktivní mísení barev

Při subtraktivním mísení barev se ze spektra daného mnohobarevného (polychromatického) světla odebírají (odečítají) některé jeho spektrální složky (nezaměňovat se zápornými hodnotami barev při aditivním míšení), proto má výsledná barva chudší spektrální složení a jeví se tudíž obecně jiná, než jaká je původní barva světla.



Obr. 5. Subtraktivní skládání barev. [13]

Subtraktivní mísení lze realizovat například za pomocí barevných optických filtrů zároveň zařazených za sebou před jediný zdroj mnohobarevného světla. Provede-li se subtraktivní mísení například tak, že se před reflektor vyzařující bílé světlo zařadí modrý a žlutý filtr, tak vznikne zelené světlo (obr. 5), kdežto při aditivním mísení modrého a žlutého světla vznikne barva bílá [13].

Obdobně jako v případě aditivního mísení lze při subtraktivním míšení získat ze tří barev různé barevné tóny. Při něm jsou však základem tzv. normální barvy (barviva), tj. barvy, které jsou doplňkové k základním barvám *R*, *G*, *B*, konkrétně jde o barvy azurovou, purpurovou a žlutou. Subtraktivním mísením (složením) těchto tří normálních barev v různých poměrech (hustotách) vzniknou rozličné tóny různých barev. Při stejných hustotách příslušných subtraktivních optických filtrů vznikne šedá barva, a jsou-li filtry dostatečně syté, neprojde jimi vůbec žádné světlo a výsledkem je černá barva. Omezíme-li se na filtry stejné hustoty (sytosti), pak při subtraktivním míšení barev (barviv) obecně platí: azurová + žlutá = zelená, purpurová + žlutá = červená, azurová + purpurová = modrá, azurová + purpurová + žlutá = černá (šedá).

4.1.4 Aplikace aditivního a subtraktivního mísení barev

U barvotisku se subtraktivně mísí dvě barvy tak, že se tisknou přes sebe, světlo pak prochází vrstvami barviva, odráží se od bílého podkladu a opět prochází vrstvami barviva. Tímto způsobem vzniká dvojnásobná optická filtrace světla a tím se zvětšuje sytost výsledné barvy barvotisku. Barvotisk někdy též mísí barvy aditivně, a to tehdy, když se barevné tečky (případně plošky), tvořící barvotisk, umístí těsně vedle sebe. V takovém případě každá tečka (ploška) odráží světlo své barvy, a tato světla se mísí aditivně ve výslednou barvu. Někdy se u barvotisků využívá i kombinace aditivního a subtraktivního mísení barev a to tak, že barevné elementy barvotisku jsou natištěny na bílém podkladu částečně vedle sebe a částečně na sobě.

Subtraktivní mísení barev se projevuje i při osvětlování těles. Osvětlí-li se těleso světlem určité barvy, pak tato barva působí na jeho povrch jako optický filtr na odraz světla, jehož barva se subtraktivně mísí s barvou dopadajícího světla. Výsledná barva záleží proto jak na barvě tělesa, tak i na zbarvení dopadajícího světla [13].

Aditivní model se využívá na zobrazovacích zařízeních (např. monitor). V IT se tento model velmi často využívá (zjednodušen na model RGB) jako nejsnazší interpretace barvy z palety (např. při programování).

4.1.5 Další barevné modely a kolorimetrie

Fyzikální podstata světla ani lidské vnímání takovou symetrii, jako barvy u aditivního nebo subtraktivního míšení barev, nevykazují. Z těchto představ vznikly jednoduché barevné prostory RGB, CMY, HSV a další. Měly však mnoho nevýhod. Nevystihovaly citlivost lidského oka a byly závislé na použitých základních barvách (a barvivech). A tak vznikla věda o vnímání barev – kolorimetrie.

4.1.6 Gama křivka

Gama je číslo, pod kterým se skrývá mocnina, podle níž je světelná intenzita závislá na vstupní veličině. Podle fyzikálních zákonů o vlivu elektromagnetického pole na elektrony, putující z elektronového děla na stínítko obrazovky, je jas funkcí přibližně 2,5. mocniny přivedeného napětí. Napětí je pak (většinou) lineárně závislé na hodnotě, která je zapsaná ve videopaměti (není však přímo úměrné, neboť napětí při hodnotě nula nemusí být nulové). Nekorigované PC má tedy hodnotu gama 2,5. Některé platformy se však snaží korigovat toto zkreslení přímo v hardwaru nebo na nízkých úrovních operačního systému.

Nelineární je však i tiskový proces. Většina tiskových procesů pracuje s drobnými černými či barevnými tiskovými body v plné barvě. Tyto body může na jedné straně zmenšovat potlačení tiskového bodu (podleptání tiskové desky, potlačení osamělého náboje u laserového tisku či tepelná setrvačnost u tepelného tisku), na druhé straně je zvětšuje nárůst tiskového bodu (nahromadění barvy v místě tiskového bodu).

Skenery (a snímače digitálních fotoaparátů a kamer) jsou přibližně lineární – jejich hodnota gama odpovídá přibližně 1. Podobně jsou na tom i filmové osvitové jednotky.

4.2 Skenery jako digitalizační zařízení

K převodu analogového dokumentu (nejčastěji na papíře) do elektronické podoby je nejjednodušší metodou přepis (či obkreslení). Tato metoda je časově nejvíce náročná a v praxi se příliš nepoužívá (až na výjimky rukopisů nebo velmi cenných dokumentů, které

by se mohly poškodit i světelnými zdroji ze snímačů)[15]. K této proceduře se používají skenery (další možné zařízení jsou například fotoaparát, čtečky apod.).

Skener je elektronické zařízení, které převádí grafickou informaci do elektronické, počítači srozumitelné podoby. Skener ze vstupního podkladu vytvoří bitmapu za účelem jejich dalšího zpracování, uložení či tisku [16].

4.2.1 Typy skenerů

- ruční - tímto scannerem je nutno ručně přejíždět po snímané předloze. Nevýhodou je malá kvalita nasnímaného obrazu způsobená jak nízkým rozlišením snímače, tak nutností přesného ovládání ze strany uživatele. Používá se tam, kde je třeba rychle snímat malé plochy, případně při nemožnosti umístění předlohy do stolního scanneru. Dnes téměř vymizel vzhledem k masivnímu rozšíření stolního typu.
- stolní - předloha se pokládá na sklo, pod nímž projíždí strojově ovládané snímací rameno, princip je tedy podobný jako u kopírovacího stroje. Dnes jsou už velmi levné. Nevýhodou je zejména možnost snímání jen relativně tenkých předloh. Velkoformátové scannery jsou schopné snímat předlohu po sloupcích. Dražší modely často snímají pomocí přídavných nástavců také diapozitivy a negativy.
- Bubnové (Drum) - předloha je nalepena na rotujícím válci a je snímána paprskem. Jejich nevýhodou je vysoká cena, a proto jsou využívány zejména pro snímání velmi velkých předloh, případně tam, kde je potřeba velice vysoká kvalita výsledku (např. z předlohy – diapozitivu je potřeba vytisknout plakát rozměru A2). Tato technologie je zároveň nejstarší.
- Filmové - slouží pro snímání jednotlivých políček filmu. Vzhledem ke svému specifickému účelu jsou vesměs používány pouze profesionálně.

4.2.2 Skenování

Skenováním se rozumí procedura, která převádí analogovou předlohu do digitální bitmapy pomocí skeneru. Dokument je nutné ve skeneru nejdříve nasnímat. Základním požadavkem je dobré a rovnoměrné osvícení předlohy po celé její ploše. To zajišťovala u plošných skenerů donedávna tzv. *"Chladná katodová lampa"*, neboli zářivka. Výhodou tohoto řešení je vysoká intenzita produkovaného světla, nevýhodou pak nerovnoměrné osvětlení (nejvíce

světla je vyzařováno uprostřed). Aby byl tento nedostatek v co možná největší míře odstraněn, je zářivka obvykle doplněna systémem zrcadel, které vrací odražené světlo na místo, kde je ho potřeba. Novější řešení u tzv. *CIS* technologie využívá řadu luminiscenčních *LED diod*. Všechny použité diody jsou přirozeně stejné a to zaručuje maximální možnou stejnoměrnost osvětlení po celé šíři snímaného dokumentu. Osvětlovací a snímací mechanizmus se postupně posouvá po předloze a snímá jeden řádek za druhým. [17]

CCD

Kombinace zářivka - optická soustava - snímací prvek *CCD* je klasická technologie, nazývaná CCD. Skenery vybavené tímto způsobem snímání jsou trochu dražší, choulostivější na poškození, ale mají lepší barevnou citlivost.

CIS

V poslední době velmi rozšířená technologie. Jedná se o dvě řady diod, jednu vysoce svítivých LED diod a řadu diod snímacích. Kladem jsou nižší výrobní náklady a tudíž nižší cena "CIS" skenerů, menší rozměry a větší odolnost. Nevýhodou je naopak nižší svítivost a citlivost (to se projeví například při snímání jemných barevných odstínů nebo třeba u silnější rozevřené knihy ve hřbetu).

Převod obrazové informace na elektronickou

Snímač pracuje tak, že intenzita světla, které dopadá na jeho jednotlivé buňky je přeměněna na elektrický náboj o různé síle. Každý bod elektronické podoby obrazu je složen ze tří informací - intenzity tří základních barev - R (červená), G (zelená) a B (modrá). Každý bod snímané předlohy je tedy měřen třemi buňkami snímače - každá buňka pomocí speciálních filtrů vyhodnocuje jednu z uvedených barevných složek bodu. V plošných skenerech jsou použity tzv. řádkové CCD nebo CIS snímače, použitý snímač tedy určuje maximální možné optické rozlišení skeneru.

4.2.3 Parametry skenerů

Prvním, velmi důležitým parametrem, je tzv. optické rozlišení. Udává se v pixelech na palec (DPI). Výrobci často uvádějí až šestinásobnou hodnotu tohoto rozlišení, nazývané též jako „softwarové rozlišení“, které patří do kategorie mediální masáže. Nejsou-li buňky

CCD snímačů ve čtvercové síti (všechny tři barvy v každém bodu) a mají-li jinou geometrii, výrobce někdy uvádí pod různými názvy další nevěrohodné „rozlišení“ podle počtu snímacích bodů. Věrohodnější údaj o rozlišení poskytne celkový počet snímacích elementů vydělený třemi (počet barev).

Bitová hloubka je další parametr, který souvisí s gama křivkou (kvalita skenování v podání nejtmavších odstínů). Lze-li nahrát gama křivku přímo do skeneru, pak je to velmi výhodné. Existují dva typy těchto skenerů – s digitální převodní tabulkou – zde může gama korekci zajistit převodní tabulka mezi A/D převodníkem a datovou částí; a s analogovou gama korekcí (někdy i s analogovým řízením jasu a kontrastu), kde může gama korekci zajistit analogový prvek vřazený před A/D převodník.

4.2.4 Barevná kalibrace skeneru

Barevná kalibrace je postup, kdy se vytváří a používá tzv. *barevný profil*, který popisuje barevné zkreslení daného zařízení. K úpravě barev lze použít barevné kalibrační profily, které převádějí barevný prostor zařízení do nezávislého barevného prostoru nebo barevného prostoru jiného zařízení. Při kalibraci obrázku je možné profily zřetězit (je to výhodné z hlediska minimalizace výpočetních chyb – s profily se pracuje v plovoucí řádové čárce, zatímco obrazová data jsou omezená na malý počet bitů).

4.2.5 Rozhraní skenerů

Typ připojení skeneru k počítači souvisí s rychlostí skenování.

- Datový tok skenerů připojené přes paralelní port (LPT) je maximálně 12000 kbit/s.
- SCSI (Ultra640 SCSI) skenery mohou data odesílat rychlostí až 640 MB/s.
- USB skenery (nejčastější připojení) mají propustnost 1.5 Mb/s (USB 1.1), 480 Mb/s (USB 2) a až 4.8Gb/s (USB 3)[18].

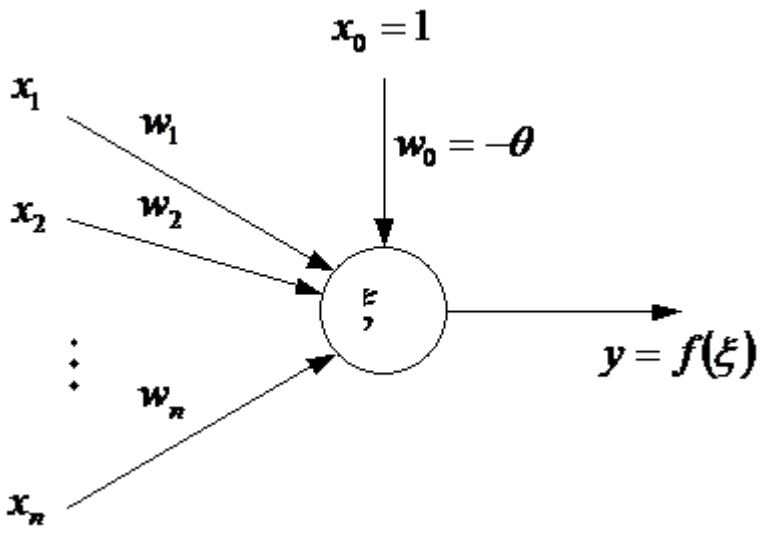
4.3 Vytěžování textu metodou OCR

OCR neboli optické rozpoznávání znaků (z anglického Optical Character Recognition) je metoda, která pomocí vstupní bitmapy (ze skeneru či externího souboru) umožňuje digitalizaci tištěných textů, s nimiž pak lze pracovat jako s normálním počítačovým textem. Počítačový program převádí obraz buď automaticky, nebo se musí naučit rozpoznávat

znaky. Převedený text je téměř vždy v závislosti na kvalitě předlohy třeba podrobit důkladné korektuře, protože OCR program nerozezná všechna písmena správně. Tato metoda využívá Hopfieldovy neuronové sítě.[19]

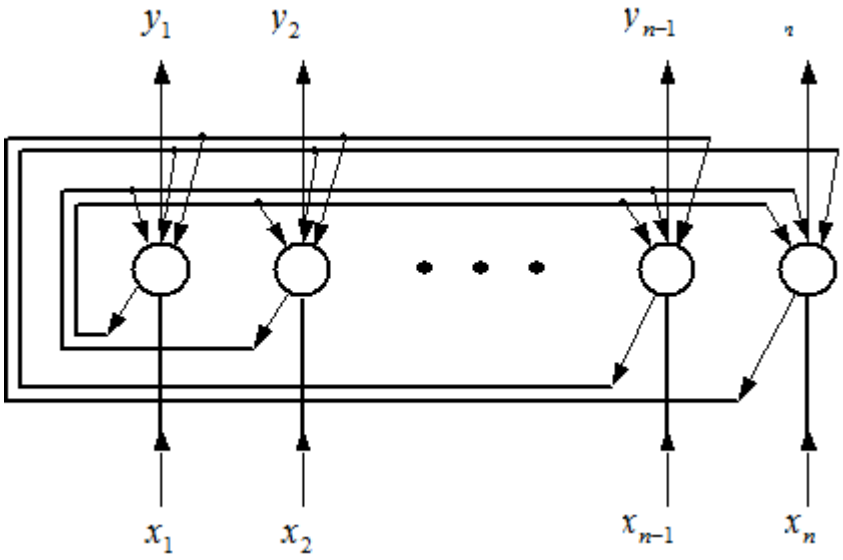
4.3.1 Popis Hopfieldovy neuronové sítě

Hopfieldova síť identifikuje znak z matice pixelů, kterou lze zjednodušeně interpretovat logickými hodnotami (černá/bílá). Tato síť ke své funkci potřebuje asociativní paměť, kterou lze chápat jako množinu váhových hodnot (vzorů). Hopfieldova síť se skládá z tolika neuronů, kolik má vstupů. Všechny signály jsou bipolární (nabývají hodnot -1 nebo 1). Práh q je pro všechny neurony nulový - přenosová funkce se nebude posouvat. Dle topologie sítě se vede výstup každého neuronu zpět na vstupy ostatních neuronů přes tzv. váhy. Pro výpočet nových výstupních hodnot při vybavovacím procesu lze použít dvě metody: synchronní aktualizování výstupů a asynchronní aktualizace výstupů. Z hlediska softwarové realizace se jeví být vhodnější první metoda, při které jsou všechny aktuální hodnoty napřed uschovány a pomocí nich se počítají nové hodnoty. Druhá metoda je též implementovatelná, např. pomocí vláken (threads), avšak v tomto případě hodnoty výstupu neuronů závisí i na pořadí výpočtu a celý proces je tedy nedeterministický.



Obr. 6. Základní model umělého neuronu

Na *Obr. 6* je znázorněn základní model umělého neuronu s vyznačenými vstupy ($x_0 - x_n$), výstupem y. *Obr. 7* ukazuje schéma zapojení umělých neuronů do hopfieldovy sítě.



Obr. 7. Schéma hopfieldovy sítě

4.3.2 Metodika Hopfieldovy sítě

Učení této sítě probíhá zpravidla jednorázově - na začátku procesu. Vybavování (interpretace výsledné množiny) je iterační proces a proto je smysluplným výstupem až poslední stav neuronů. Na vstupy sítě se přivede neznámý obrazec a počítají se výstupy. Po každém průběhu se výstupy poopraví a přivedou opět na vstupy sítě. Tento postup se opakuje tak dlouho, až se hodnoty na výstupech během dvou cyklů nezmění ani na jednom neuronu.

4.3.3 Důsledky řešení

Hopfieldova síť patří svou strukturou mezi asociativní paměti, její odpovědí na předložený vzor je přímo nalezený vzor. V tomto případě se tedy nejedná o OCR ve smyslu rozpoznání textu stylu komerčních programů. Klasický klasifikátor vznikne z Hopfieldovy sítě až po rozšíření o komparátor, který odpověď sítě porovná s naučenými vzory.

Úspěšnost vybavování je v případě Hopfieldovy sítě dána Hammingovou vzdáleností vzorů. Pokud je tato vzdálenost malá, může dojít k chybě při vybavování, nebo vzniku fantomů. Pokud je naopak Hammingova vzdálenost velká např. vlivem posunutí vzoru, dojde též ke špatnému rozpoznání. Teoreticky je tedy nutné, aby počet vzorů byl menší než 15% bodů obrazce. Tuto hodnotu je však ještě třeba snížit, protože rozpoznávané znaky (číslice, resp. písmena latinské abecedy) se vyznačují poměrně malou Hammingovou

vzdáleností. Praktické zkušenosti ukazují, že v případě rozpoznávání znaků je počet vzorů, které lze síť naučit, dán pouze cca 7% z celkového počtu neuronů sítě.

5 POUŽITÉ TECHNOLOGIE

Technologie v kontextu této část práce jsou technologie programovacího jazyka, metody programování a postupů, které se aplikují v praktické části práce.

5.1 Programovací jazyk C#

Programovací jazyk je prostředek pro zápis algoritmů, jež mohou být provedeny na počítači. Zápis algoritmu ve zvoleném programovacím jazyce se nazývá program. Programovací jazyk je komunikačním nástrojem mezi programátorem, který v programovacím jazyce formuluje postup řešení daného problému, a počítačem, který program interpretuje technickými prostředky. Programovací jazyk je vlastně soubor pravidel pro zápis algoritmu, odborně řečeno se jedná o formální jazyk.[20]

Jazyk C# je vysokoúrovňový objektově orientovaný programovací jazyk vyvinutý firmou Microsoft zároveň s platformou .NET Framework. Microsoft vytvořil C# na bázi jazyků C++ a Java (a je tedy nepřímým potomkem jazyka C, ze kterého čerpá syntaxi), což je výhodné pro programátory programující v jazycích C++ nebo C. Příkazy a funkce, které se již osvědčily letitými zkušenostmi v psaní kódu, byly převzaty a úspěšně se používají i v tomto jazyku.

5.1.1 Specifikace a rozdíly oproti jazyku C a C++

Pan Archer [21] specifikuje rozdíly a inovace tohoto poměrně nového jazyka následovně:

- postupné vylepšování – míní zlepšení, jež redukuje počet chyb, které se často objevovali v jazycích C nebo C++, například inicializací poměnných, podmíněné příkazy vyžadují logické hodnoty a další – o to vše se stará překladač, takže program se bez ošetření těchto aspektů nespustí.
- Typový systém tohoto jazyka využívá automatickou správu paměti, proto se vývojáři nemusí zdržovat s ruční správou, která vedla k velké chybovosti.
- Typový systém je jednotný – Common Type System je unifikovaný typový systém, používaný všemi jazyky pod .NET Framework, tedy i jazykem C# (dále například VB.NET). Všechny typy, včetně primitivních datových typů jako je Integer, jsou potomky třídy System.Object.

- z dalších z mnoha vylepšení je výhoda v používání vlastností, metod a událostí, které vedou k efektivnější práci s jazykem, dále podporu atributů, které umožňují definici a použití deklarativních informací o komponentách.

5.1.2 Aplikace jazyka C#

Jazyk C# lze využít k tvorbě databázových programů, webových aplikací a stránek, webových služeb, formulářových aplikací ve Windows, softwaru pro mobilní zařízení (PDA a mobilní telefony) atd.

Díky možnosti použití ODBC (Open Database Connectivity), což je standardizované softwarové API pro přístup k databázovým systémům, je možné snadno a efektivně psát aplikace v libovolném jazyce, v libovolném operačním systému a databázovém systému.

5.1.3 Soubory XML

XML (Extensible Markup Language) je obecný značkovací jazyk, který byl vyvinut a standardizován konsorciem W3C. Je zjednodušenou podobou staršího jazyka SGML. Umožňuje snadné vytváření konkrétních značkovacích jazyků (tzv. aplikací) pro různé účely a různé typy dat. Používá se pro serializaci dat. Zpracování XML je podporováno řadou nástrojů a programovacích jazyků vč. jazyk C#.

Jazyk je určen především pro výměnu dat mezi aplikacemi a pro publikování dokumentů, u kterých popisuje strukturu z hlediska věcného obsahu jednotlivých částí, nezabývá se vzhledem. Prezentace dokumentu (vzhled) může být definována pomocí kaskádových stylů. Další možností zpracování je transformace do jiného typu dokumentu, nebo do jiné aplikace XML.

XML hned od samého počátku myslel na potřeby i jiných jazyků než je angličtina. Jako znaková sada se implicitně používá ISO 10646 (také Unicode). V XML proto můžeme vytvářet dokumenty, které obsahují texty v mnoha jazycích najednou – můžeme přepínat mezi různými jazyky v jednom dokumentu. Současně je přípustné i jiné libovolné kódování (např. windows-1250, iso-8859-2), musí však být v každém dokumentu přesně určeno. Odpadají tak problémy s konverzí z jednoho kódování do druhého.

5.2 Databáze

Databáze je určitá uspořádaná množina informací uložená na paměťovém médiu. V širším smyslu jsou součástí databáze i softwarové prostředky, které umožňují manipulaci s uloženými daty a přístup k nim - systém řízení báze dat (SŘBD). Běžně se označením databáze – v závislosti na kontextu – myslí jak uložená data, tak i software (SŘBD).

Předchůdcem databází byly papírové kartotéky. Umožňovaly uspořádávání dat podle různých kritérií a zatřiďování nových položek. Veškeré operace s nimi prováděl přímo člověk. Správa takových kartoték byla v mnohém podobná správě dnešních databází. Poté se data zpracovávala strojem. S rozmachem výpočetní techniky byla vyvinuta snaha o implementaci báze dat do počítačového prostředí.

5.2.1 Databázový jazyk SQL

SQL (Structured Query Language) je standardizovaný dotazovací jazyk používaný pro práci s daty v relačních databázích. Relační databáze je databáze založená na relačním modelu. Je založena na tabulkách, jejichž řádky obvykle lze chápat jako záznamy a eventuelně některé sloupce v nich (tzv. cizí klíče) uchovávají informace o relacích mezi jednotlivými záznamy v matematickém slova smyslu.

V databázi je možné definovat i několik tabulek, které mohou mezi sebou být provázány vztahy - ty slouží ke svázání dat, která spolu souvisejí a jsou umístěny v různých databázových tabulkách. V zásadě rozlišujeme čtyři typy vztahů.

1. mezi daty v tabulkách není žádná spojitost, proto nedefinujeme žádný vztah.
2. 1:1 používá se, pokud záznamu odpovídá právě jeden záznam v jiné databázové tabulce a naopak. Takovýto vztah je používán pouze ojediněle, protože většinou není pádný důvod, proč takovéto záznamy neumístit do jedné databázové tabulky. Jedno z mála využití je zpřehlednění rozsáhlých tabulek.
3. 1:N přiřazuje jednomu záznamu více záznamů z jiné tabulky. Jedná se o nejpoužívanější typ relace, jelikož odpovídá mnoha situacím v reálném životě.
4. M:N je méně častým. Umožňuje několika záznamům z jedné tabulky přiřadit několik záznamů z tabulky druhé. V databázové praxi bývá tento vztah z praktických důvodů nejčastěji realizován kombinací dvou vztahů 1:N a 1:M, které

ukazují do pomocné tabulky složené z kombinace obou použitých klíčů (třetí resp. tzv. vazební tabulka).

5.2.2 Databázové prostředí MSSQL

Microsoft SQL Server je relační databázový a analytický systém pro e-obchody, byznys a řešení datových skladů vyvinutý společností Microsoft.

II. PRAKTICKÁ ČÁST

6 SPECIFIKACE ZADÁNÍ

Základem archivace dokumentů pro tuto práci je vytěžování informací z osobních dokladů pomocí skeneru metodou OCR. Myšlenka archivace je založena na uchovávání získaných informací např. v databázi, což je tento případ. Vývojové prostředí bylo zadáno Visual Studio a programovací jazyk C#. Databáze byla zadána jako relační; zadaný jazyk pro databázi je SQL. Doporučená struktura zdrojového kódu je modulová z důvodu možnosti rozšiřování o další typy dokladů. Jediným podporovaným osobním dokladem je občanský průkaz. Řešení má zahrnovat i další aplikaci, která využívá databázi.

Aplikace (stejně jako návrh databáze) byla vyvíjena v prostředí MS Visual Studio 2008 ve výukové variantě v rámci programu MSDN AA na FAI UTB pro nekomerční využití.

6.1 Občanský průkaz

Pro skenování občanského průkazu a získáváním informací v něm, se uvažují průkazy vydávané od 1. ledna 2005. Platnost tohoto průkazu je po dobu v něm uvedenou. Tento průkaz obsahuje strojově čitelnou oblast, kterou program zpracovává.

Strojově čitelná oblast je v občanském průkazu (*Obr. 8*) umístěna na čelní straně pod číslem občanského průkazu (*Obr. 9*).

Obr. 8. Vzor občanského průkaz - čelní pohled (vlevo), zadní pohled (vpravo)[22][1]

Obr. 9. Detail čelní strany občanského průkazu se zvýraněnými strojově čitelnými daty

[1] Podle českého Autorského zákona (Zákon č. 121/2000 Sb., § 3, odstavec a) jde o dílo vyňaté z ochrany dle práva autorského. Z toho vyplývá, že je tento obrázek uvolněn jako volné dílo.

6.1.1 Datová část občanského průkazu

Zvýrazněná oblast na Obr. 8 ukazuje oblast, kterou lze strojově číst. Tato oblast usnadňuje automatizované zpracování pomocí OCR. Tyto údaje mají podobu dvou textových řádků u spodního kraje dokladu (tak, aby bylo možné protažení čtečkou). Údaje jsou vytištěny fontem OCR-B ve velikosti Size1 (cca 14 bodů; přesně viz ISO 1073-2:1976). Obsah těchto dvou řádek je definován v ISO 7501 respektive ICAO 9303. Délka jednoho řádku může být u každého typu dokumentu různá. U občanského průkazu náleží jednomu řádku 36 znaků, čili celkem je třeba vytěžit 72 znaků (u cestovního pasu to je 44*2 znaků, u víza 36*2 znaků atd.).

První řádka obsahuje informace o dokladu, vydavatelském státu dokladu, jméně a příjmení vlastníka. Tyto informace jsou koncipovány následovně:

- 2 znaky – typ a podtyp dokladu (občanský průkaz – typ = „I“, podtyp = „D“, cestovní pas – typ „P“, podtyp „S“ – služební, „D“ – diplomatický, „<“ – pas bez speciálního určení)
- 3 znaky - vydavatelský stát (dle normy ISO 3166 – v ČR „CZE“)[23].
- Příjmení držitele ukončené sekvencí znaků „<<“.
- Jméno držitele ukončené sekvencí „<<“ nebo koncem řádku. Zbytek vyplněn znakem "<". Samostatný výskyt znaku „<“ u jména, anebo u příjmení, má význam mezery (používá se třeba u lidí, kteří mají více křestních jmen).

Druhá řádka obsahuje tyto informace:

- 9 znaků – číslo dokladu + 1 znak - kontrolní součet (viz. kapitola Kontrolní součet)
- 3 znaky – občanství (dle normy ISO 3166 – v ČR „CZE“)[23].
- 6 znaků – datum narození (formát YYMMDD) + 1 znak – kontrolní součet
- 1 znak – pohlaví – „M“ – muž, „F“ – žena
- 6 znaků platnost do (formát YYMMDD) + 1 znak – kontrolní součet
- Obecně je zde umístěno volitelné pole proměnně délky. Podle přítomnosti/nepřítomnosti kontrolního znaku volitelného pole a souhrnného kontrolního znaku končí na konci řádku nebo jeden či dva znaky před koncem

řádku. U občanského průkazu se zde nachází druhá část rodného čísla doplněná „<<<“

- 1 znak - kontrolní znak volitelného pole - na některých typech dokladů chybí zcela, na jiných není vypočítán a obsahuje znak '<'
- Volitelný souhrnný kontrolní znak (poslední znak) přítomný pouze na některých dokladech. Pokud U občanského průkazu je kontrolní součet počítán z posloupnosti vytvořené z hodnot data narození, platnosti a čísla dokladu.

U některých typů dokladů může délka čísla dokladu přerůst 9 znaků - v tom případě je v poli "číslo dokladu" prvních 9 znaků, místo kontrolního znaku je '<', zbývající část čísla je ve volitelném poli. V takovém případě volitelné pole neobsahuje obsah obvyklý pro daný typ dokladu a zemi vydání.

Veškeré údaje se kódují výhradně velkými písmeny anglické abecedy, arabskými ciframi a znakem '<'. Pro vkládání dalších znaků jsou definované transkripce (např. 'Ä' → 'AE'). Jiné znaky nejsou dovoleny a nahrazují se znakem '<'.

6.1.2 Kontrolní součet

Každý znak se nahradí číselnou hodnotou. Cifry jsou přímo touto hodnotou, znak '<' má hodnotu 0, znaky A-Z mají hodnoty 10 - 35.

Z takto získané řady se za pomoci cyklicky opakovaných vah '7','3','1' vypočte vážený součet. Kontrolní číslicí je zbytek po dělení deseti tohoto součtu.

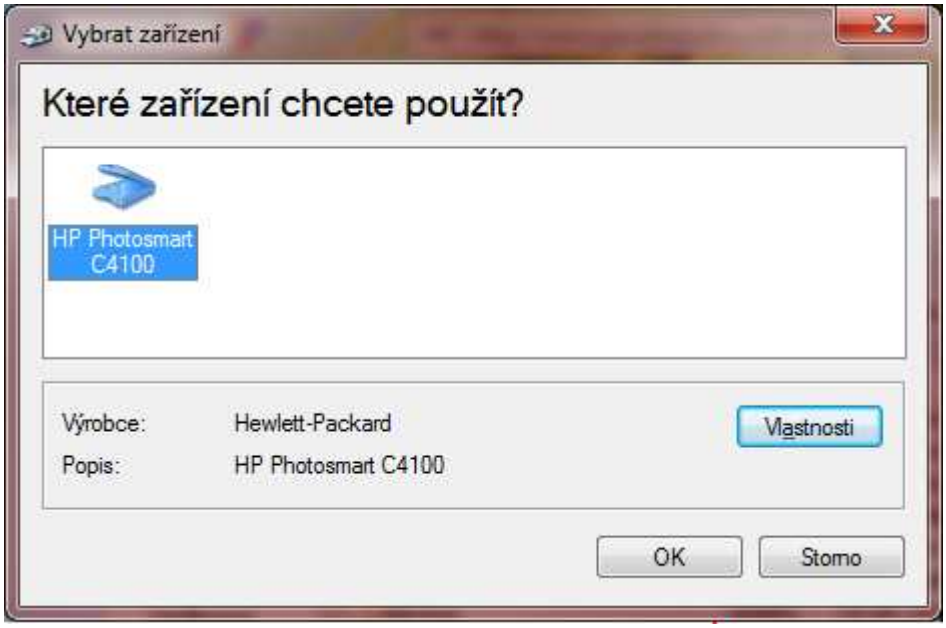
Příkladný výpočet kontrolní součtu pro posloupnost „170420“, započne výpočtem sumy: 1*7 + 7*3 + 0*1 + 4*7 + 2*3 + 0*1 = 62, ta se podělí 10 a zbytek po dělení je kontrolní součet - 62 mod 10 = 2.

7 IMPLEMENTACE ZADÁNÍ

Postup vývoje aplikace využívající převod dokumentu do elektronické podoby zahrnuje modul, který komunikuje s vstupním zařízením (skener); získaná data (bitmapa) vyhodnotí jako text, který aplikace analyzuje. Výstupní informací je takto prosévaná informace ve formě dokumentu do podoby určení osobních informací z dokumentu.

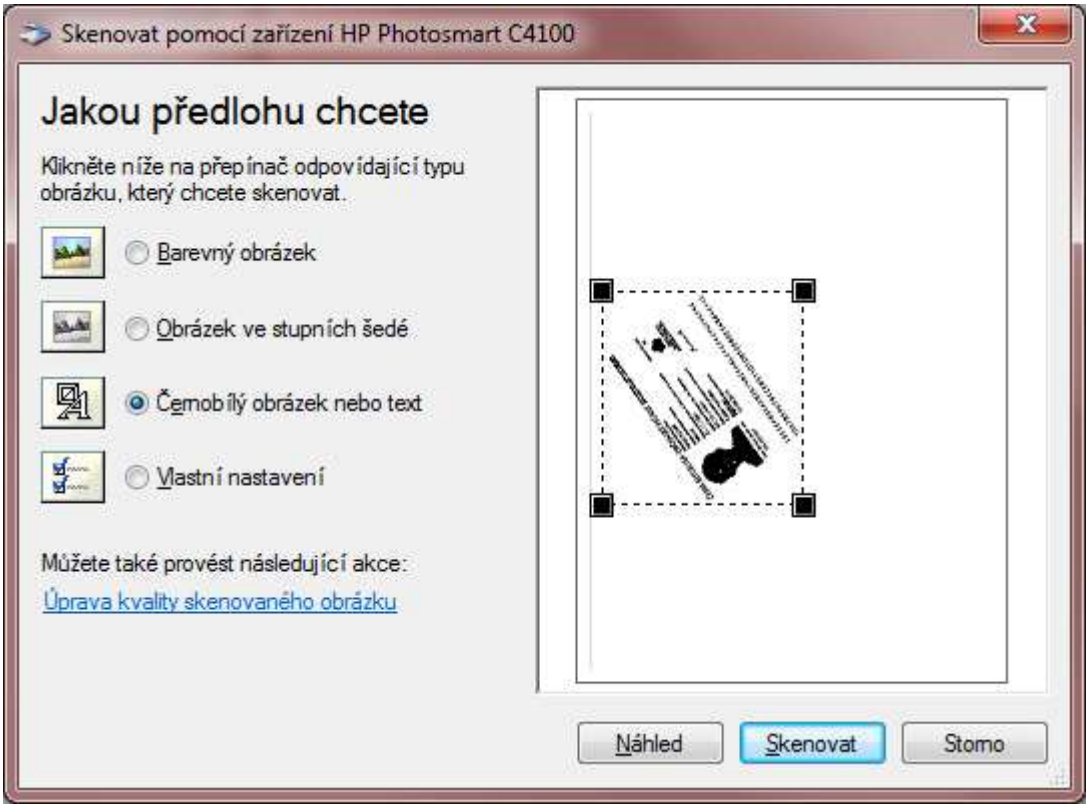
7.1 Rozhraní mezi skenerem a PC

Pro komunikaci mezi skenerem a PC byla vybrána knihovnu WIA od společnosti Microsoft. Výhodou této knihovny je podpora všech skenerů pro Windows bez nutnosti instalace podpůrného softwaru pro skener; stačí mít jen nainstalované ovladače pro tento skener.



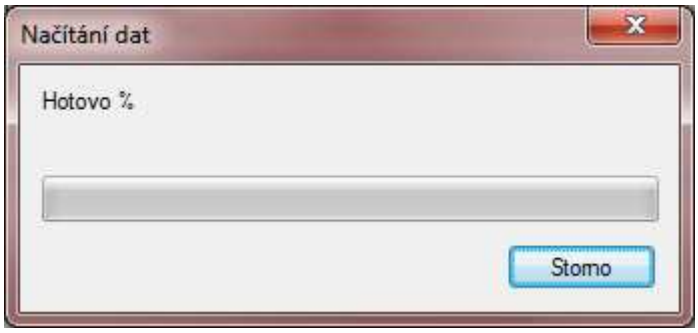
Obr. 10. Výběr skeneru

Po vyvolání metody skenování uživatel vybere vstupní zařízení.



Obr. 11. Nastavení skenování

V tomto dialogu (viz. *Obr. 11*) lze nastavit parametry výstupní bitmapy, rozsah skenování a kvalita skenování.



Obr. 12. Průběh skenování

Po dokončení procesu skenování se výstup uloží jako bitmapa do paměti, která se bude vyhodnocovat metodou OCR.

7.2 Vytěžování textu

Ze vstupní bitmapy získanou metodou OCR se vyseparuje textová informace. Tuto proceduru nabízí společnost Microsoft jako knihovnu pod názvem MODI (Microsoft

Office Document Imaging). Knihovna nabízí nejen podporu více jazyků jako vzory pro OCR, ale i identifikaci textu z otočené bitmapy.

Pro výstupní aplikaci byla zvolena identifikace pro anglický jazyk, což pro strojově čitelnou oblast u dokumentů dostačuje.

Obr. 13. Ukazatel průběhu identifikace textu

Tuto funkci byla naprogramována tak, aby běžela v samostatném vlákně (viz. zdrojové kódy v příloze P II), jinak by mohlo být hlavní vlákno obsazeno touto časově náročnou procedurou a uživateli by se mohlo zdát, že aplikace zamrzla.

7.2.1 Příklad výstupního textu

Po identifikaci a úpravě textu (odstraněním bílých znaků) z bitmapy získané skenerem, může výstupní řetězec vypadat například následovně:

CZECHUPUSUCAIDENTIFICATIONCARDKTPUILIQJJFTCH#QIJE-
CART!DIDENTITTP&IJM!NI/SURNAME/N0MSPECIMENJUENO/GIVINNAMES/PR
TNOMVZOR29.02.1979RODNTC!SLO/PELSONALNO.INOPELSONNEL795229/0292
STATNIORLA.NSTVI/NATIONALRRY/NATIONAUTTCESKÁREPUBLIKAFLATNO
STDO/OP,DOC.NO/NO.DCC,DAT!OFEXPIRY/DATED'EXPIP.ATION99000610010.01
.2015IDCZESPECIMEN<<VZOR<<<<<<<<<<<<<<<9900061003CZE7902295F1501
1020292<<<5CESKAREPUBLIKAOBCANSKVRPRUKAZ•.•\\%\•I..ISCISLODATUMN
AROZENI/DATEOPIRTH/DATEDENAISSANCEPOHULVI/SEX/SWFP0DHDK2ITILE
HOLDER'S$IGNATURESIGNATUREDUTITUL'DR!4Uxxx

Obr. 14. Získaný text ze vzoru občanského průkazu po aplikaci OCR

Zvýrazněná posloupnost znaků na *Obr. 14* je hledaný text. Zde jsou umístěna strojově čitelná data z občanského průkazu.

7.3 Analýza získaného textu

Protože knihovna MODI nemusí vracet text vždy ve správném pořadí, tak byl algoritmus analýzy navrhnut a naprogramován jako iterační test podřetězců. Vychází se z předpokladu, že jsou vždy znaky jednoho řádku umístěny za sebou, a že jeden řádek má 36 znaků. Přirozeně je dále známo, jaká data jsou očekávána.

Kvůli podpoře více typů dokumentů bylo třeba sestavit metadata (šablony) skládající se z tabulky prefixů (počáteční podřetězce každého dokumentů – u občanského průkazu „ID“) a tabulky počtů znaků (občanský průkaz – 72 znaků). Dále byla sestavená tabulka států (dle normy ISO 3166 – u nás „CZE“)[23], se kterou se porovnává řetězec sestavený z 3. – 5. znaku řetězce. Na zbytku řádku (6. – 36. znak) se nacházejí jen znaky „A“ – „Z“ nebo znak „<“. Dále se postupuji individuálně dle typu dokumentu. Občanský průkaz obsahuje na 2. řádku v prvních deseti znacích čísla, následuje kód státu (opět se porovnává s tabulkou), dalších 7 znaků jsou čísla následované znakem „M“nebo „F“ určujícím pohlaví držitele. Dalších 11 znaků jsou čísla, dále řetězec „<<<“. Posledním znakem je číslo. Po splnění těchto podmínek je řetězec identifikován. Posledním volitelným krokem je ověření správnosti identifikovaných dat dle kontrolních součtů (je možné tento krok přeskočit – metoda OCR nemusí 100% identifikovat všechny znaky správně, a tak by nebyla data z dokumentu identifikována).

7.4 Návrh databáze

Protože se uchovávají data získaná z dokladů, tak dostačuje pouze jedna tabulka. Sloupce tabulky byly zvoleny takto:

- ID – primární klíč záznamu
- s_name(varchar(100)) – jméno vlastníka
- s_surname(varchar(100)) – příjmení vlastníka
- s_personal(varchar(11)) – rodné číslo uložené ve tvaru „123456/7890“
- s_doc(varchar(50)) – číslo průkazu. Protože s číslem průkazu se nepracuje jako s číslem a navíc není vyloučené, že se zde může objevit jiný znak (jiný doklad) než číslo, tak je tato hodnota deklarována jako řetězec.
- s_state(varchar(3)) – zkratka státu (dle normy ISO 3166)

- s_gender(varchar(1)) – pohlaví vlastníka „M“ – muž, „F“ – žena
- type(int) – typ průkazu (pro občanský průkaz j zde hodnota 1)

7.5 Připojení k databázi

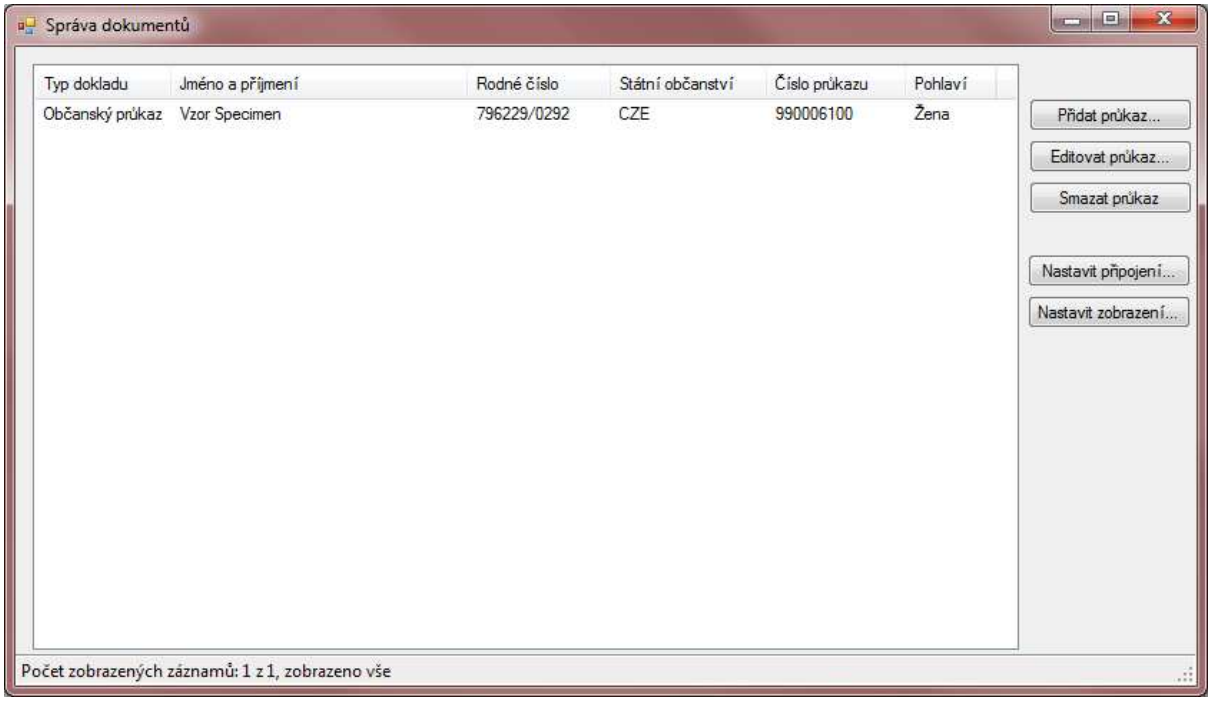
Způsob připojení k databázi využívá prostředí MS SQL. Přístup je navrhnut tak, aby bylo možné se k databázi připojit jak z lokálního serveru, tak ze serveru vzdáleného (je nutné zadat IP adresu serveru, port, uživatelské jméno a heslo). Před připojením je nutná konfigurace autentizace (na straně vzdáleného MS SQL serveru).

8 SOFTWAROVÉ APLIKACE

Vyvinuté aplikace mohou využívat stejnou databázi (dle nastavení dané aplikace). První aplikace pro kompletní správu dokumentů má název „Documents“. Tato aplikace umožňuje spravovat (přidávání, úprava, mazání) data uložené v databázi. Dále má v sobě implementovaný modul pro připojení k databázi. Druhá aplikace jménem „Documents_view“ umožňuje pouze prohlížení záznamu uložené v databázi, včetně nastavení způsobu připojení.

8.1 Aplikace Documents

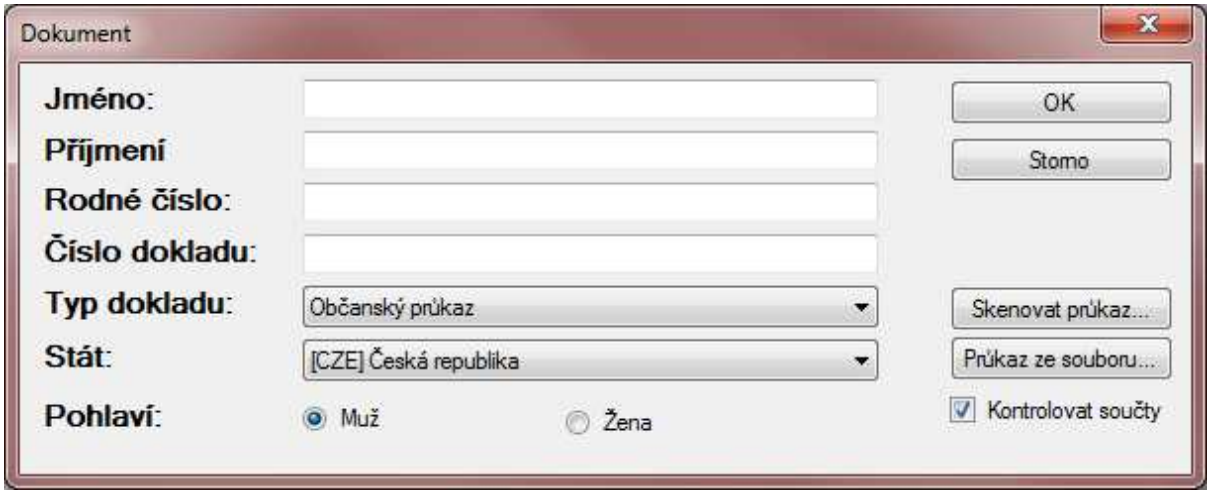
Aplikace obsahuje metadata zahrnující šablonu občanského průkazu a cestovního pasu.



Obr. 15. Hlavní okno aplikace. V levé horní části je seznam uložených záznamů v databázi, v pravé horní části jsou tlačítka pro správu dat a nastavení programu, v dolní části je stavový řádek informující uživatele o zobrazených záznemech.

8.1.1 Přidání nového záznamu

V hlavním okně aplikace (viz. *Obr. 15*) uživatel stiskne tlačítko „Přidat průkaz…“, který vyvolá dialog pro přidání nového záznamu do databáze.



Obr. 16. Dialog pro přidávání záznamů do databáze. Obsahuje vstupní pole pro zadávání informací (vlevo) a ovládací prvky pro vyvolání vytěžovacího procesu informací z dokumentu (vpravo dole).

Do vstupních polí dialogu pro přidávání záznamů (viz. *Obr. 16*) uživatel může zadat informace o dokladu ručně, nebo budou vyplněny informacemi získanými přímo ze skenovaného dokumentu (tlačítkem „Skenovat průkaz…“), nebo již naskenovaného dokumentu uloženého jako obrázek (tlačítkem „Průkaz ze souboru…“) ve formátu (BMP, PNG nebo JPG). Zaškrtávacím polem „Kontrolovat součty“ bude program kromě separace informací z dokladu kontrolovat i kontrolní součty v dokumentu (viz. Kapitola 6.1.2). Po neúspěšné identifikaci (text buď nebyl rozpoznán, nebo nebyla nalezena strojově čitelná oblast nebo kontrolní součty jsou chybné) se tato data nezobrazí.

Validace vstupních dat

Před stiskem tlačítka „OK“ v dialogu pro přidávání záznamů (viz. *Obr. 16*) se provádí validace zadaných dat.

- Kontrola jména a příjmení spočívá v ověření, že se tato data skládá pouze s písmen české abecedy (tedy znaků „A“-„Z“ a diakritiky), navíc délka jména a příjmení musí obsahovat alespoň 2 znaky.
- Číslo dokladu musí obsahovat pouze čísla „0“ – „9“ nebo písmena „A“-„Z“.
- Rodné číslo je ve tvaru „123456/7890“ (tedy 6 číslic + „/“ + 3 nebo 4 číslice). Navíc po odstranění znaku „/“ se získá číslo, které musí být beze zbytku dělitelné číslem 11. Třetí číslice v rodném čísle, obsahuje informaci o pohlaví držitele průkazu. Je-li na tomto místě číslo „0“ nebo „1“, pak patří průkaz muži. Číslem „5“

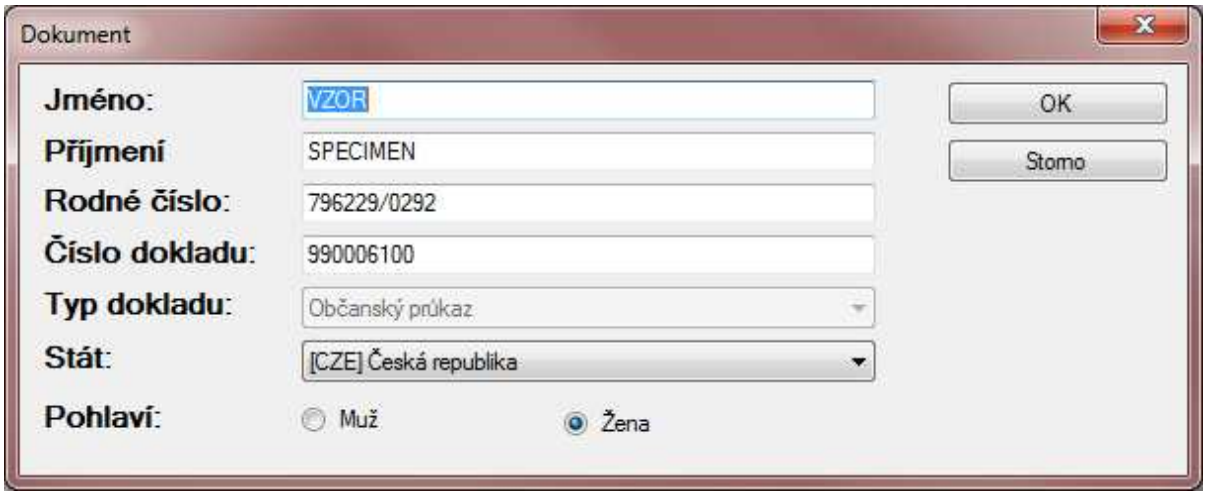
nebo „6“ na této pozici vyznačuje majitelku průkazu. V případě vyplnění jiného státu než „CZE“ se výše uvedený postup kontroly neprovádí.

Vložení informace do databáze

Po úspěšném ověření vstupních dat se tato uloží do databáze jako nová položka. Dialog se uzavře a v hlavním okně aplikace (viz. *Obr. 15*) se objeví v seznamu.

8.1.2 Editace exitujícího záznamu

V okně aplikace (viz. *Obr. 15*) po stisku tlačítka „Editovat průkaz…“ se vyvolá dialog pro editaci existujícího záznamu v databázi (v seznamu hlavního okna musí být položka označena).



Obr. 17. Editace existujícího záznamu v databázi. Obsahuje vstupní pole pro zadávání informací (vlevo).

Funkcionalita tohoto dialogu je oproti dialogu pro přidávání záznamu (viz. *Obr. 16*), mírně redukována, a sice o funkci vytěžování informací přímo z dokumentu. Je zde také implementovaná validace vstupních dat. Po stisku tlačítka „OK“ se v databázi přepíše stará informace za data nová.

8.1.3 Smazání záznamu

V hlavním okně aplikace (viz. *Obr. 15*) po stisku tlačítka „Smazat průkaz“ (položka musí být označena v seznamu) program vyvolá dotaz pro uživatele. Po potvrzení se tato položka smaže z databáze a seznam se aktualizuje.

8.1.4 Nastavení připojení databáze

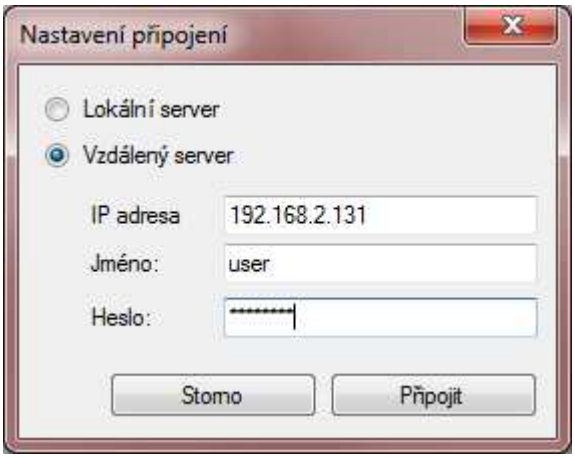
Byla již zmíněna možnost připojení aplikace k databázi buď na lokálním serveru, nebo na vzdáleném serveru. Konfigurace možnosti připojení je uložena v externím XML souboru, konkrétně v souboru „access.xml“. Struktura tohoto souboru je následující:

```
<?xml version="1.0"?>
<ROOT>Data Source=.\SQLEXPRESS; Initial Catalog=Documents; Integrated Security=True</ROOT>
```

Data konkrétně uvedená výše databázi připojí na lokální server. Struktura tohoto XML souboru pro síťovou konfiguraci připojení by vypadala například takto:

```
<?xml version="1.0"?>
<ROOT>Data Source=192.168.2.131,1433; Network Library=DBMSSOCN;Initial Catalog=Documents; User ID=user; Password=password</ROOT>
```

Soubor XML generuje dialog, který se zobrazí po stisku tlačítka „Nastavit připojení…“ v hlavním okně aplikace.

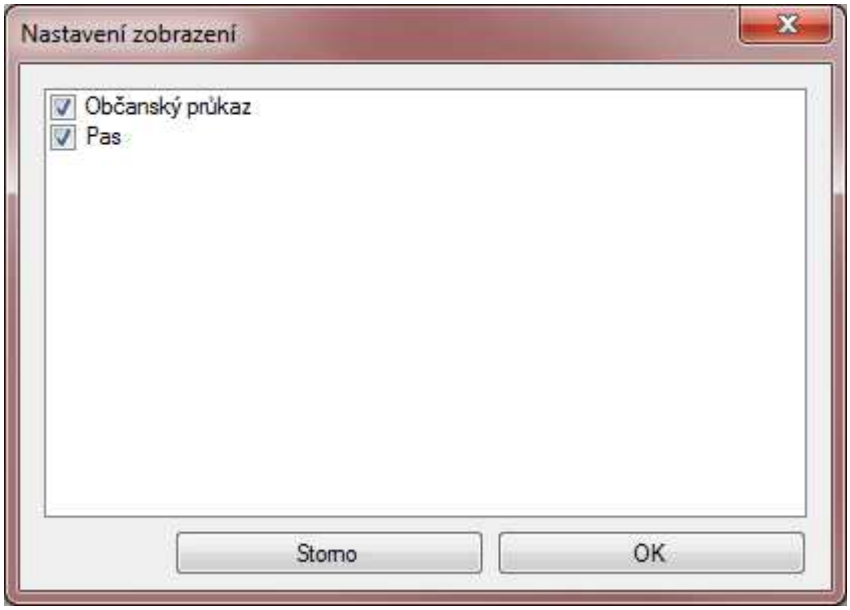


Obr. 18. Nastavení připojení aplikace k databázi.

Tlačítkem „Připojit“ v dialogu (viz. *Obr. 18*) se otestuje připojení, a po úspěšném testu se tato informace uloží do XML souboru.

8.1.5 Nastavení zobrazení

Uživatel si může vybrat typ dokumentů, které se budou v seznamu hlavního okna zobrazovat. Nastavení se vyvolá tlačítkem Nastavit zobrazení…“ v hlavním okně aplikace.

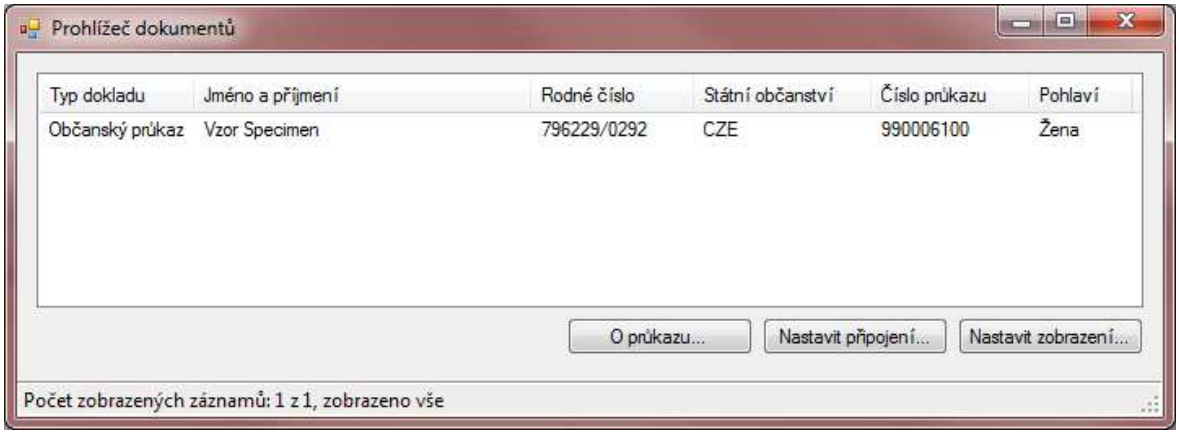


Obr. 19. Nastavení zobrazení typů dokumentů v seznamu.

Stiskem tlačítka „OK“ v dialogu nastavení zobrazení (viz. *Obr. 19*) se dialog zavře a v hlavním okně se aktualizuje seznam dle nově nastavených parametrů.

8.2 Aplikace Document_Viewer

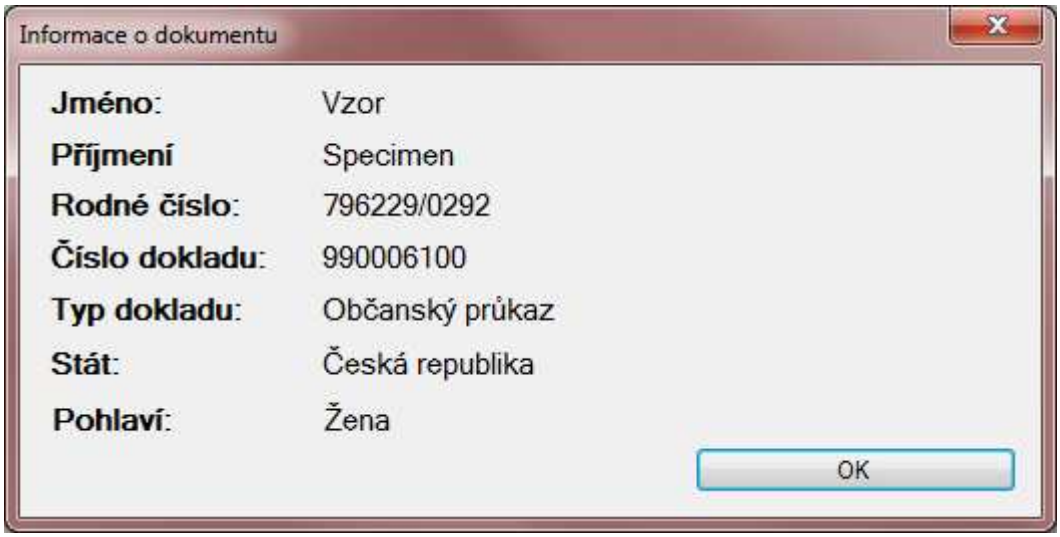
Softwarová aplikace „Document_Viewer“ demonstruje způsob připojení na databázi. Aplikace má pouze elementární funkci čtení dat, nastavení zobrazení (analogie jako 8.1.5 - Nastavení zobrazení u první aplikace), nastavení připojení databáze (využívající stejného principu jako první aplikace).



Obr. 20. Hlavní okno aplikace. V horní části je seznam uložených záznamů v databázi, v dolní části jsou tlačítka pro nastavení programu, v dolní části je stavový řádek informující uživatele o zobrazených záznemech.

8.2.1 Informace o průkazu

V hlavním okně aplikace (viz. *Obr. 20*) si uživatel může zobrazit informace o vybraném dokumentu v seznamu kliknutím na tlačítko „O průkazu…“.



Obr. 21. Dialog zobrazující informace o vybraném dokumentu

V tomto dialogu (viz. *Obr. 21*) jsou uvedeny jen detaily vybraného dokumentu uloženého v databázi.

9 NÁVRH OPTIMÁLNÍ HARDWAROVÉ KONFIGURACE

Při Návrhu optimální konfigurace se vychází z Liebigova zákona zákona ekologie (známý jako Liebigův zákon minima), který říká, že organismus je tak silný, jako jeho nejslabší článek[24]. Návrhu tedy vychází z nejnáročnějšího zařízení

9.1 Používaný hardware

Jako skener byla pro vývoj aplikace použita multifunkční tiskárna HP Photosmart C4100. Systémové a hardwarové požadavky k tomu zařízení jsou uvedeny v příloze P I.

9.2 Návrh konfigurace

Předpokládá se, že hardwarové nároky používaného zařízení jsou větší jako nároky samotného operačního systému, který si již při instalaci hardware testuje. Vývoj informačním technologií dospěl do stadia paralelizace algoritmů a vývoje softwaru využívajících více vláken. I nejslabší procesor, který je kompatibilní s operačním systémem, kde pracujeme, umí (buď pomocí instrukcí, nebo zásahem operačního systému) pracovat s více vlákny.

Časově nejnáročnější procedura softwarové aplikace (viz. příloha P II) je vytěžování informací pomocí metody OCR. Paměťově nejnáročnější je výstupní bitmapa ze skeneru. Je-li třeba mít skenování rychlé, tak se tím dá dosáhnout buď rychlým rozhraním (pakliže se skenuje více kvalitně), nebo bude dostačovat nízká kvalita výstupní bitmapy. Čím větší kvalita, tím větší je paměťová náročnost pro výstupní bitmapu (pohybuje se řádově okolo 30 MiB pro rozlišení 300 DPI – uloženo v BMP).

Navrhovaná hardwarová konfigurace:

- procesor Intel® Core™ Duo T2700
- paměť RAM – alespoň 256 MiB
- HDD – 50 MiB volného místa
- skener kompatibilní s rozhraním alespoň USB 2.0
- rozhraní USB 2.0

ZÁVĚR

Studium starých poznatků a pravd, patří dle mého názoru ke vznešeným lidským činnostem, které vede k nalézání smyslu poznání. Problémem archivace se člověk zabývá celou svou existenci. Uchování, znovuobjevování, studium a vymýšlení nového je alfa a omega lidského vědění.

V této práci jsem se zabýval archivací současnosti, konkrétně byl úkol zadán jako archivace textových informací vytěžených z osobních dokladů za použití moderních technologií.

Při řešení problému jsem využil již hotových nástrojů (rozhraní mezi skenerem a počítačem a metody OCR). Důvodem použití je velká časová náročnost na splnění zadání (v případě vývoje vlastních knihoven). V inženýrské praxi se tento postup běžně aplikuje. Vytížení textu metodou OCR je procedura vykazující určitou chybu, což se dá akceptovat ruční editací. Získáním a separací informací z dokladu se nový záznam hned nevloží do databáze, ale uživatel má možnost si data poupravit.

Napojení databáze na další aplikace jsem navrhnul pomocí implementace pomocí XML souboru. Tento soubor je možné editovat buď ručně, nebo se generuje přímo programy.

Návrh optimální konfigurace v současném stavu vývoje informačních technologií nezáleží až tak moc na procesoru; takt procesoru se již nezvyšuje, zvyšuje se počet jader. Velikost operační paměti se dnes pohybuje řádově v jednotkách GiB. Jedinou stanovenou podmínkou bylo užití rozhraní USB 2.0 mezi skenerem a počítačem (skener toto rozhraní musí podporovat). S přicházejícím rozhraním USB 3 se propustnost zvýší a zrychlí se přenos dat (za předpokladu kompatibility skeneru s tímto rozhraním).

ZÁVĚR V ANGLIČTINĚ

Study of old knowledge and truth, are in my opinion important of human activities. This study leads to finding the meaning of knowledge. People are still dealing with archiving. Storage, re-inventing, inventing a new study is the alpha and omega of human knowledge.

In this work, I dealt with today's archiving. The task was of archiving the textual information extracted from personal documents with using modern technologies.

To resolve this issue I used a ready-made tools (interface between the scanner and computer methods and OCR). The reason is the large time required to fulfill the assignment (in the case of development of their own libraries). In engineering practice, this procedure is normally applied. Getting text using OCR is a procedure that has a little error, which can edit manual edits. Getting and separation of information from the document doesn't need to insert new record to the database, but the user can modify data.

Connection a database to other applications using the XML file. This file is edited either manually or using the application.

Design of the optimal configuration for current state of development of information technology does not matter so much on the CPU. CPU clock is not increasing, increasing the number of cores. Memory is now in GiB. The only condition was set using USB 2.0 interface between the scanner and the computer (scanner must support this interface). With the coming of a USB interface 3 is to increase throughput and accelerates data transfer (assuming compatibility with the scanner interface).

SEZNAM POUŽITÉ LITERATURY

[1] ISBD(ER) : *mezinárodní standardní bibliografický popis pro elektronické zdroje : revidované doporučení ISBD(CF) : mezinárodní standardní bibliografický popis pro počítačové soubory*. Z ang. orig. přeložila, č. příklady opatřila a k tisku připr. L. Celbová. 2. revid. vyd. Praha : Národní knihovna ČR, 1998. viii, 114 s

[2] KATUŠČÁK, Dušan; MATTHAEIDESOVÁ, Marta; NOVÁKOVÁ, Marta. *Informačná výchova : terminologický a výkladový slovník : odbor knižničná a informačná* veda. Bratislava : SPN, 1998, s. 75.

[3] FEATHER, John. Preservation and the management of library collections. 2nd ed. London : Library Assocation Publishing, 1996, s. 52-53.

[4] BARKER, P. *Electronic documents and their role in future library systems. In Libraries for the new millenium : implications for managers. Ed. by D. Raitt*. London : Library Association Publishing, 1997, s. 89-113.

[5] KNOLL, Adolf. *Problematika elektronických publikací*. Národní knihovna. 1999, roč. 10, č. 4, s. 173-177.

[6] CONWAY, Paul. *The relevance of preservation in a digital world [online]. In Preservation of library & archival materials : a manual. Ed. by Sheledyn Ogden. 3rd rev. ed. Andover (Ma.) : Northeast Documnet Concervation Center*, 1999 [cit. 2010-05-20]. Dostupné na WWW: < http://www.clir.org/pubs/reports/conway2/ >.

[7] VOJŤÁŠEK, Filip. *Dlouhodobá archivace digitálních dokumentů* [online]. [cit. 2010-05-25]. Dostupné na WWW: < http://www.ikaros.cz/node/675 >.

[8] RUSSELL, Kelly. *Digital preservation : ensuring access to digital materials into the future.* [online] University of Leeds, June 1999 [cit. 2010-05-20]. Dostupné na WWW: <http://www.leeds.ac.uk/cedars/Chapter.htm>.

[9] ROTHENBERG, Jeff. *Avoiding technological quicksand : finding a viable technical foundation for digital preservation : report to the Council on Library and Information Resources* [online]. Washington, D. C. : CLIR, January 1999 [cit. 2010-05-20]. vi, 35 s. Dostupné na WWW: <http://www.clir.org/pubs/reports/rothenberg/pub77.pdf>. ISBN 1-887334-63-7.

[10] GAVITT, Sharon *COMPUTER OUTPUT MICROFILM (COM).* In [online]. [2002-03-14 [cit. 2010-05-25]. Dostupné z WWW: <www.archives.nysed.gov/a/records/mr_pub52.pdf>.

[11] PARRAMÓN, Jose M. *Problematika Teorie barev*. 1995, 1. vydání, 112 s

[12] ŠTRBA Anton: *Všeobecná fyzika 3:OPTIKA*; 1. vydání; Alfa Bratislava a SNTL Praha, 1979, s. 65-70.

[13] POSPÍŠIL Jaroslav: Mísení *barev a jejich grafické znázornění*; *Rozhledy matematicko-fyzikální*; Praha: SPN, 1998, roč. 75, s. 160-166

[14] *Teorie - PALADIX foto* [online]. 2005 [cit. 2010-06-04]. Dostupné z WWW: <http://www.paladix.cz/sekce/teorie/>.

[15] *Přepisy textů, smluv, dokumentů, opisy, rukopisy* [online]. 2010 [cit. 2010-06-04]. Dostupné z WWW: < http://www.prepisy.eu/prepisy-textu>.

[16] *Scanner - Wikipedie, otevřená encyklopedie* [online]. 2010 [cit. 2010-06-04]. Dostupné z WWW: < http://cs.wikipedia.org/wiki/Skener>.

[17] *MUSTEK.CZ - O SKENOVÁNÍ* [online]. 2010 [cit. 2010-06-04]. Dostupné z WWW: < http://www.mustek.cz/?sekce=skenovani&stranka=skenovani1>.

[18] *10 otázek o USB 3.0 - CHIP online* [online]. 2010 [cit. 2010-06-04]. Dostupné z WWW: < http://www.chip.cz/clanky/trendy/2010/03/10-otazek-o-usb-3.0>.

[19] *OCR - Hopfieldova síť* [online]. 2010 [cit. 2010-06-04]. Dostupné z WWW: < http://www.cgg.cvut.cz/members/cadikm/school/nan/Teorie.html>.

[20] *Programovací jazyk. In* Wikipedia *: the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 18.5.2004, last modified on 24.5.2010 [cit. 2010-06-05]. Dostupné z WWW: <http://cs.wikipedia.org/wiki/Programovac%C3%AD_jazyk>

[21] TOM, Archer. *Myslíme v jazyku C# : knihovna programátora*, Grada publishing, Praha, 2002.308 s. ISBN 80-247-0301-7

[22] *Osobní doklady - Ministerstvo vnitra České republiky* [online]. 2010 [cit. 2010-06-04]. Dostupné z WWW: < http://www.mvcr.cz/clanek/osobni-doklady.aspx>.

[23] *ISO - Maintenance Agency for ISO 3166 country codes - English country names and code elements* [online]. 2010 [cit. 2010-06-04]. Dostupné z WWW: < http://www.iso.org/iso/english_country_names_and_code_elements>.

[24] *Liebig's law of the minimum: Information from Answers.com* [online]. [cit. 2010-06-05]. Dostupné z WWW: < http://www.answers.com/topic/liebig-s-law-of-the-minimum>

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

.NET	dot NET
BMP	Windows Bitmap
C#	C Sharp
CCD	Charge Coupled Device
CD-DA	Compact Disc Digital Audio
CD-ROM	Compact Disc Read-Only Memory
CIS	Contact Image Sensor
CMY	cyan magenta yellow
COM	computer-output microfilm
CZE	Czech Republic
ED	elektronický dokument
HDD	hard disk drive
HSV	Hue, Saturation, Value
ISO	International Organization for Standardization
JPG	File Interchange Format
LPT	Line Printer Terminal
MODI	Microsoft Office Document Imaging
MSDN AA	Microsoft Developer Network Academic Alliance
MSSQL	Microsoft SQL Server
OCR	Optical Character Recognition
ODBC	Open Database Connectivity
PDF	Portable Document Format
PNG	Portable Network Graphics
RAM	Random access memory

RGB	red gren blue
SCSI	Small Computer System Interface
SQL	Structured Query Language
SŘBD	Systém řízení báze dat
USB	Unversal seriál bus
WIA	Windows Image Acquisition
WWW	World Wide Web
XML	Extensible Markup Language
.NET	Dot NET

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM PŘÍLOH

- P I Požadavky na systém pro HP 4100
- P II CD obsahující soubory se zdrojovými kódy aplikace, databáze a elektronickou podobou diplomové práce.

PŘÍLOHA P I: POŽADAVKY NA SYSTÉM PRO HP 4100

Minimální požadavky na systém:

- Procesor Intel® Pentium® II, Celeron® nebo kompatibilní (doporučen procesor Pentium III nebo vyšší)
- Microsoft® Windows® 98 Second Edition, Me, 2000 (Service Pack 3 nebo vyšší), XP Home, XP Professional, XP Media Center Edition, XP Starter Edition a Tablet PC Edition pro základní softwarová řešení
- Microsoft® Windows® 2000 (Service Pack 3 nebo vyšší), XP Home, XP Professional, XP Media Center Edition, XP Starter Edition a Tablet PC Edition pro plná softwarová řešení
- 128 MB paměti RAM pro systémy Windows 98 SE, ME, 2000 a XP (pro všechny operační systémy je doporučeno 256 MB nebo více)
- 750 MB volného místa na pevném disku pro instalaci softwaru (475 MB pro základní instalaci softwaru)
- Microsoft® Internet Explorer® 5.01 s aktualizací Service Pack 2 nebo vyšší

(Nepodporuje systémy Windows 95, 98, 3.1, NT® 4.0, DOS, 2003 Server nebo Windows Vista Edition.)