Univerzita Karlova v Praze

Filozofická fakulta

Ústav informačních studií a knihovnictví

Studijní program: informační studia a knihovnictví

Studijní obor: informační studia a knihovnictví

BBS Radek Bitrman, DiS.

Techniky vyhledávání v databázích s ohledem na různé typy informačních objektů

Diplomová práce

Praha 2007

Vedoucí diplomové práce: Mgr. Jan Pokorný

Oponent diplomové práce:

Datum obhajoby:

Hodnocení:

Prohlášení:

Prohlašuji, že jsem diplomovou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje.

V Praze, 12. prosince 2007

.....

Podpis diplomanta

Identifikační záznam

BITRMAN, Radek. *Techniky vyhledávání v databázích s ohledem na různé typy informačních objektů [Techniques of information retrieval in databases in reference to different types of information objects]*. Praha, 2007. 97 s. Diplomová práce. Univerzita Karlova v Praze, Filozofická fakulta, Ústav informačních studií a knihovnictví 2007. Vedoucí diplomové práce Mgr. Jan Pokorný.

Abstrakt

Předmětem diplomové práce je na příkladech databází obsahující vybrané typy informační objektů analyzovat možnosti vyhledávání. V první části práce jsou blíže charakterizovány informační objekty z hlediska typologie MIME.

V druhé kapitole jsou přiblíženy možnosti a techniky indexování informačních objektů, zmíněny jsou také techniky vyhledávání vybraných typů informačních objektů. Uvedeny jsou také zajímavé možnosti vyhledávání v podobě různých experimentů a prototypů vyhledávacích systémů. Představeny jsou také dle autora zajímavé příklady databází zabývající se vybranými typy informačních objektů.

Kapitola čtyři se poté již zabývá analýzou vyhledávacích možností u vybraných databází, porovnává jednotlivé specifika a také podobnosti. Jednotlivé vlastnosti a s tím související skutečnosti jsou zmíněny v rámci shrnutí a také SWOT analýzy jednotlivých informačních objektů.

Klíčová slova

vyhledávání informací, informační objekty, digitální objekty, databáze, textové informační objekty, grafické informační objekty, zvukové informační objekty, image retrieval.

Obsah

Předmluva **3**

1. Úvod **5**

2. Informační objekty **6**

2.1. Typologie informačních objektů dle specifikace MIME 7

2.2. Identifikace informačních objektů 14

2.3. Metadata 17

2.3.1. Obecné formáty metadat 19

2.4. Textové informační objekty 24

2.5. Grafické informační objekty 27

2.6. Zvukové informační objekty 30

3. Informační objekty v databázích **32**

3.1. Vyhledávací stroje 32

3.1.1. Vyhledávací stroje a internet 33

3.1.2. Hodnocení efektivity vyhledávání 36

3.2. Vyhledávání informačních objektů 38

3.2.1. Indexování 38

3.2.2. Vyhledávání 43

3.2.3. Prezentace výsledků vyhledávání 51

3.3 Databáze textových informačních objektů 51

3.3.1. arXiv.org 52

3.3.2. CERN Document Server 53

3.3.3. E-LIS 55

3.4. Databáze grafických informačních objektů 57

3.4.1. Corbis 57

3.4.2. Databáze JSC NASA 59

3.5. Databáze zvukových informačních objektů 61

3.5.1. The Sound Effects Library 61

3.5.2. The Freesound Project 63

3.5.3. The Philharmonia Sample Library 64

4. Analýza vyhledávacích možností u vybraných příkladů databází 66

4.1. Postup analýzy 66

4.2. Vyhledávání dle jména autora 67

4.3. Vyhledávání dle názvu 68

4.4. Vyhledávání dle textové anotace 69

4.5. Vyhledávání dle kategorie 70

4.6. Specifické možnosti vyhledávání 71

4.7. SWOT analýza 74

4.7.1. Textové informační objekty 74

4.7.2. Grafické informační objekty 76

4.7.3. Zvukové informační objekty 78

4.8. Shrnutí a návrh možných zlepšení 79

5. Závěr 85

Slovník použitých pojmů 87

Seznam použité literatury 89

Předmluva

Předkládaná práce se zabývá problematikou vyhledávání v databázích se zaměřením na různá specifika, vlastnosti a problémy různých typů informačních objektů v digitální podobě. Samozřejmě není možné zahrnout úplný popis existujících databází informačních objektů, a proto jsou v této práci popisování pouze vybrané příklady databází.

Toto téma jsem si vybral, neboť problematika vyhledávání informačních objektů mě zajímá samotného z pohledu uživatele. S některým aspekty vyhledávání se také setkávám v profesním prostředí.

V rámci příkladů databází textových informačních objektů byly zařazeny archivy elektronického tisku nebo volně dostupné databáze textových objektů. Důvodem bylo zejména profesionální zpracování, propracované možnosti vyhledávání a také trvalý a bezplatný přístup k textovým objektům, a to zejména v podobě úplných textů. Mezi příklady grafických informačních objektů byly zařazeny jak databáze fotografií, tak také odborná databáze s propracovanými možnosti vyhledávání technického rázu. Specificky zaměřené příklady databází se zajímavými možnostmi vyhledávání byly vybrány v případě zvukových informačních objektů.

V první části je nastíněno pojetí informačních objektů použité při zpracování této práce. Kapitola 2.1 charakterizuje typologii informačních objektů z hlediska specifikace MIME. Některé typy informačních (digitálních) objektů se vyskytují také ve fondech digitálních knihoven, které jsou zmiňovány v podobě příkladů databázi (otevřených archivů) zaměřující se na textové informační objekty. Nejen proto je určitý prostor věnován také digitálním knihovnám. S vyhledáváním informačních objektů také souvisí identifikace těchto objektů a identifikátory (kapitola 2.2.). Pro doplnění popisu informačních objektů je také stručně představena problematika metadat (kapitola 2.3.) Další kapitoly jsou věnovány detailnějšímu popisu hlavních skupin informačních objektů. Textové informační objekty v rámci kapitoly 2.4., grafické informační objekty jsou popsány v kapitole 2.5., zvukové informační objekty v kapitole 2.6.

Druhá část této práce se již zaměřuje na informační objekty v databázích a techniky indexování a vyhledávání. V rámci kapitoly 3.1. je stručně přiblížena problematika

vyhledávacích strojů. Kapitola 3.2. představuje jak problematiku indexování a vyhledávání jednotlivých informačních objektů, tak také charakterizuje některé prototypy u vyhledávání zejména grafických a zvukových informačních objektů. Dle autora zajímavé příklady databází jednotlivých informační objektů jsou představeny v rámci kapitol 3.3 až 3.5.

Možnosti vyhledávání uvedených příkladů databází jsou podrobněji charakterizovány ve třetí části této práce. Na představených příkladech databází se tato část zabývá možnostmi vyhledávání společnými pro uvedené typy informačních objektů (kapitoly 4.2. až 4.5.) a navzájem porovnává. Kapitola 4.6. uvádí specifické možnosti vyhledávání. Kapitola 4.7 je věnována shrnutí a autorem jsou zmíněny některé možnosti zlepšení.

Způsob citace použité literatury vychází z norem ISO 690 a ISO 690-2. Citace v textu jsou prováděny v podobě číselných odkazů na seznam literatury v závěru práce.

Užívaná spojení „textové informační objekty“ je považováno za ekvivalent spojení „textové objekty“. Obdobně je nahlíženo na „grafické informační objekty“, „zvukové informační objekty“, které se v této práci mohou vyskytnou bez přídavného jména informační. V některých případech je použito jednodušší označení pro tyto objekty – text (textové informační objekty), obrázek, fotografie (grafické informační objekty), zvuk, audio, hudba, hlasový záznam (zvukové informační objekty) z důvodu vyplývajícího z kontextu zabývající se daným objektem. V rámci této práce jsou uvažovány pouze zvukové informační objekty v digitální podobě.

Rád bych na tomto místě také poděkoval svému vedoucímu Mgr. Janu Pokornému za dobře míněné připomínky, vstřícný přístup a také trpělivost při vedení této práce.

1. Úvod

Vyhledávání informací patří mezi každodenní lidské činnosti. S potřebou informací se můžeme setkat v různých oblastech – od informací o dopravním spojení při cestě do školy či zaměstnání, informace o počasí v destinaci pro strávení dovolené, informace potřebné pro management při rozhodování o dalším vývoji firmy až po informace posílané z družic zkoumající neznámé planety a planetky.

Běžným prostředkem pro uspokojení informačních potřeb patří také internet. Obrovská informační exploze vyvolaná rozšířením internetu nebyla bohužel doprovázena odpovídajícím vývojem technologií pro vyhledávání v tomto prostředí. Při vyhledávání se stále spoléhalo jen na vyhledání zadaného klíčového slova v textu dokumentu (textového objektu) nebo například procházení tématicky uspořádaného seznamu odkazů (předmětový katalog). Zejména pro informační specialisty, kteří poměřují kvalitu procesu vyhledávání informací dle přesnosti a úplnosti, bylo toto vyhledávání nepřijatelné.

Často není relevantní informace nejen v prostředí internetu nalezena například z toho důvodu, že je uložena v jiném typu informačního zdroje (obrázek, zvukový záznam, databáze), než je typ zdrojů prohledávaných vyhledávacími stroji. V jiných případech je potřebná informace uložena v lokálních repozitářích, které nejsou z www prostředí vůbec dostupné.

Množství zvukových a grafických informačních objektů v digitální podobě vzrostlo exponenciální rychlostí během posledních let. Gigabajty nových obrázků, hudebních nahrávek, zvukových záznamů, video nahrávek jsou vytvářeny a ukládány každý den. Tyto informační objekty vytvářejí neuspořádané databáze, které nejsou z globálního hlediska vůbec strukturovány a připraveny pro opravdu efektivní vyhledávání. Potřeba využití celého potenciálu takovýchto databází, je potřeba nástrojů pro automatické indexování a inteligentní vyhledávací mechanismy. Význam těchto informačních objektů je velký, neboť vývoj těchto nástrojů je uskutečňován a podporován jak v komerční sféře, tak také v akademické.

2. Informační objekty

Tato kapitola je zaměřena na problematiku informačních objektů, charakteristiku jednotlivých typů informačních objektů, vysvětluje způsob nazírání na typologii informačních objektů dle specifikace MIME a možné typy informačních objektů v oblasti digitálních knihoven.

Často se setkáváme s pojmy informace, informační objekt anebo informační zdroje (zdroje informací), a to v různých souvislostech. Stručně si přiblížíme jednotlivé pojmy. Samotné slovo informace má původ v latinském slovu ”informare”, což znamená utvářet nebo sdělovat představu. Zkráceně lze říci, že na informace můžeme pohlížet jako na data, která jsou použitelná a zároveň srozumitelná a jsou v daný okamžik potřeba.

Nyní se zaměříme na informační zdroje a pro srovnání si zde uvedeme různé definice. Česká terminologické databáze knihovnictví a informační vědy (TDKIV) [1] definuje pojem informační zdroj jako: ”Informační objekt, který obsahuje dostupné informace odpovídající informačním potřebám uživatele., ale můžeme zde již nalézt nastínění možného způsobu dělení, které přibližuje následující odstavec. Samozřejmě se nejedná o jedinou definici tohoto pojmu.

Sklenák například ve své monografii [2] vysvětluje informační zdroj následovně: ”Informačním zdrojem budeme rozumět systém, který je reálným nebo potenciálním nositelem, zprostředkovatelem nebo šiřitelem informací”. Tomuto vymezení pak vyhovují knihovny, databázová centra, informační střediska, televize, rozhlas apod.

Na informační zdroje lze nahlížet podle různých hledisek. Například podle toho, na jakém médii jsou prezentovány nebo poskytovány. Tím se myslí zejména rozdělení na zdroje informací elektronické (v elektronické podobě) a tzv. klasické (neboli v tištěné podobě). Toto rozdělení se objevilo v důsledku technologického pokroku a zejména neuvěřitelného rozmachu v oboru výpočetní techniky. Informační zdroje v elektronické podobě usnadňují následné zpracování a používání daných informací.

Elektronický informační zdroj (nebo také elektronický zdroj) může být definován [24] jako informační zdroj, který je uchováván v elektronické podobě a je dostupný

v prostředí počítačových sítí nebo prostřednictvím jiných technologií distribuce digitálních dat (např. na discích CD-ROM).

2.1. Typologie informačních objektů dle specifikace MIME

Nyní se zaměříme na informační objekty a typologii z hlediska specifikace MIME, které se využívají ve stále větší míře nejen pro potřeby elektronické pošty, ale také ve světě World Wide Web (WWW) zejména pro tzv. „MIME-typ“. Nejdříve si přiblížíme důvody existence tohoto formátu a pouze stručně nastíníme některá specifika. Z důvodů zaměření této práce budou technické detaily popsány pouze stručně a informativně, a to pouze pro přiblížení dané problematiky.

Standard MIME neboli Multipurpose Internet Mail Extensions je původně rozšiřující internetový standard formátování emailových zpráv zahrnující [4]:

- přibalování nejrůznějších příloh v netextovém formátu,
- definování sady jednotlivých typů v rámci Content-Type,
- používání samotného textu zpráv také v jiných jazycích a znakových sadách, než jen v podobě ASCII,
- přidávání vícedílných zpráv
- používání informací v hlavičce také v jiné znakové sadě než jen v ASCII.

Hlavním důvodem pro existenci tohoto standardu MIME je dodnes používaný mechanismus pro přenos zpráv označovaný jako SMTP (Simple Mail Transfer Protocol), který garantuje pouze přenos zpráv tvořený sedmibitovými znaky. Tyto sedmibitové znaky vlastně představují znaky uvedené v kódovací tabulce ASCII. V případě potřeby poslat elektronickou poštou cokoliv nemající podobu sedmibitových ASCII znaků, musíme takovouto informaci do dané podoby nejprve přenést. Toto přenesení je možné v podobě zakódování dat právě do ASCII znaků. Standard MIME vymezuje možnosti přípustných typů, které je možné použít pro zakódování přílohy na straně odesílatele. Ale zároveň je nutné, aby o použité volbě kódování byl informován také příjemce pro úspěšné rozkódování přílohy. Mechanismus MIME tak umožňuje posílání netextových informačních objektů v podobě obrázků, zvukových (hudebních) souborů, filmů anebo datových souborů v podobě aplikací.

Pro úspěšné rozpoznání typu přiložených dat slouží MIME Content-Type (můžeme se také setkat s používaným pojmem MIME type), která je dílčí konvencí standardu MIME a určuje typ přenášených dat dle konkrétní hodnoty. Podstata MIME Content-Type v rámci elektronické pošty je v identifikaci typu dat, určenou jen pro potřeby jejich přenosu. O jednotlivých typech tohoto standardu pojednává další část.

Hlavička standardu MIME označovaná *Content-Type* představuje znakový řetězec určující základní povahu dat v samotném těle MIME entity (emailová zpráva, případně její tělo) identifikátory. Tyto identifikátory se rozdělují na dvě složky. První identifikátor určuje základní typ dat, druhá část označována jako podtyp slouží pro specifikaci základního typu [5]. Jednotlivé typy MIME byly již od začátku navrhovány s ohledem na další rozšiřitelnost, neboť právě rozšiřitelnost a možnost neustálého doplňování o nové typy respektive podtypy lze považovat za význačnou vlastnost a také důvod pro globální využitelnost.

Zajištění korektního doplňování novými typy MIME obsahující podrobné specifikace nového typu je spravováno mezinárodní organizací, která je zejména autoritou pro přidělování internetových identifikátorů, IP adres a dalších číselných identifikátorů pro služby v síti internet (Internet Assigned Numbers Authority, IANA, http://www.iana.og) [6]. Oficiální informace týkající se MIME jsou poskytovány ve formátu RFC (Request For Comments) vydávané otevřenou organizací IETF (The Internet Engineering Task Force, http://www.ietf.org) [7] starající se o rozvoj internetu a definice standardů pro jeho provoz.

V současné době vymezuje standard MIME tyto hlavní typy (které jsou dále upřesňovány konkrétními podtypy):

1) Jednoduché (diskrétní) typy:

a) **text** - je určen pro posílání informací v textové formě jako například strukturovaný text - HTML, SGML, XML, CSV, RTF, nebo samotný „čistý text“ (plain text). V podobě čistého textu se jedná o lineární sekvenci znaků přerušené případně oddělovačem řádků nebo stránek. Může také obsahovat specifikaci použité znakové sady. Definované znakové sady jsou například ASCII, UTF-8 nebo ISO-8859-X (kde X představuje poslední číslice zvolené znakové sady). Textové informace vyžadující pro zobrazení speciální program jsou zařazovány

k typu „application“. Interpretace jednoduchého textového formátu vypadá následovně:

```
Content-type: text/plain; charset=iso-8859-1
```

- **image** – definovaný typ pro statické obrázky (k jejichž zobrazení je třeba specifické zařízení v podobě obrazovky, tiskárny atd.) zahrnuje například podtypy ve formátu GIF, JPG, JPEG, TIFF. Příklad záznamu:

```
Content-type: image/tiff
```

- **audio** – obecně zahrnuje hlasová či zvuková data ve formátu, který je schopný interoperability, neboť neexistuje „ideální“ audio formát. Základním podtypem je formát „basic“. Můžeme se také setkat s formáty jako například MPEG-1 Layer 3 známý pod zkratkou MP3. Dále se můžeme setkat s formáty AAC, MP4 (viz příklad níže).

```
Content-type: audio/mp4
```

- **video** – představuje formát v podobě pohyblivých obrázků, videosekvencí atd., které mohou být doplněny také zvukem. Termín video je v tomto případě chápán jako všeobecný pojem, kdy se neodkazuje na jednotlivé technologie či formáty. Základním podtypem je formát „mpeg“.

```
Content-type: video/mpeg
```

- **application** – tento formát se používá pro nespojitá (discrete) data, aplikace, které nelze zařadit do jiných kategorií a zejména ty data, která se musí před samotným zobrazením uživateli nejprve zpracovat určitou aplikací. Do této kategorie lze tedy zařadit například přenos dat, plánovací systém v rámci emailového klienta nebo tabulkový procesor (spreadsheet). Definovány jsou dva základní podtypy: 1. „*Octet-stream*“ (obsahuje binární data) a 2. „*Postscript*“ a další specializované podtypy, kterých je velké množství.

```
Content-Type: application/zip
```

- **model** - tento typ nazývaný také jako „virtuální realita“ je určen pro posílání souborů pro troj- a vícerozměrné systémy, ve kterých lze zavést pravoúhlou soustavu souřadnic. Model se skládá z jednoho nebo více objektů, které se pak skládají z prvků, jež mají mezi sebou definovány vztahy.

```
Content-Type: model/vrml
```

- **example** – typ formátu „příklad“ je určen pouze pro použití v dokumentech poskytující příklady zahrnující specifikaci některých typů formátů, kdy použití skutečného typu formátu by bylo naprosto irelevantní. Definován byl jak typ, tak také podtyp formátu. Tento typ lze zařadit jak mezi jednoduché formáty, tak i mezi složité.

```
Content-Type: application/example
```

2) <u>Složené (kompozitní) typy</u>

- **multipart** – v případě kombinací různých typů formátů nebo více různých částí do jednoho celku se používá tento formát, který přesně definuje hranice jednotlivých části a také označení typu v rámci dané části. Tímto je zajištěno, že jednotlivé části se rozkódují dle správného typu.

```
Content-Type: multipart/mixed; boundary= gc0p4Jq0M2Yt08j34c0p
                                         [označení hranice]
                    --označení hranice

                Content-type: text/plain

                    --označení hranice

                Content-Type: image/jpeg
```

- **message** – zpráva jako samostatný typ je určen pro přiložení další emailové zprávy (například ve formátu HTTP nebo MSG v případě MS Outlook) do nové emailové zprávy.

```
Content-Type: message/http
```

Globální rozšíření standardu MIME specifikující jednotlivé typy se samozřejmě projevilo nejen v oblasti elektronické pošty, ale také v dalších oblastech internetu. Jakmile je zavedena vhodná konvence, která je také i dodržována, není důvod nevyužívat také pro samotný přenos WWW stránek webovými servery a jejich klienty (v podobě webového prohlížeče). Webový prohlížeč na straně klienta tak dostává od serveru informaci o typu posílané informace – zda se jedná o text, který se má interpretovat jako HTML kód požadované webové stránky či se jedná o čistý text nebo obrázek atd. Proto webový server při zasílání www stránky v HTML tvaru nějakému prohlížeči (klientovi) připojuje ke každé stránce také údaj, že se jedná o data s MIME-typem text/html.

Některé typy informačních objektů dle typologie MIME lze najít také v oblasti digitálních knihoven. V rámci digitálních knihoven se můžeme setkat s označením digitální objekt. V této práci zaměřující se na problematiku vyhledávání vybraných typů objektů je pojem digitální objekt považován za synonymum pojmu informační objekt. Nyní si ale stručně představíme problematiku digitálních knihoven.

Samotný pojem digitální knihovna patří v poslední době mezi frekventované, ale je provázen také řadou nejasností. Jeden z důvodů těchto nejasností lze spatřovat v neustálém technologickém vývoji výpočetní techniky. Jak zmiňuje mimo jiné Bartošek [8], problematikou digitálních knihoven se zabývá mnoho různorodých odborných komunit, z nichž každá si vytváří vlastní náplň tohoto pojmu v souladu se svým zaměřením: z pohledu databázového specialisty představuje digitální knihovna informační systém využívající architekturu federativních databází, pro odborníky zabývající se hypertextem a šířením informací je to jen jedna z nadstavbových aplikací webu, knihovník vidí v digitální knihovně další krok v automatizaci na cestě od knihovny analogové (papírové) přes automatizovanou či hybridní (fyzické sbírky s automatizovaným katalogem) až po digitální (většina či veškeré informace a služby knihovny jsou elektronické). V neposlední řadě přispívá ke zmatení pojmů i skutečnost, že pojmem digitální knihovna jsou někdy označovány systémy, které – přinejmenším z pohledu informačního specialisty – představují úplně jiné entity (např. soubory algoritmů a procedur, systémy na správu dokumentů apod.).

Existuje samozřejmě mnoho definic digitálních knihoven. V počítačovém prostředí se můžeme setkat také s definicí digitální knihovny jako spravované sbírky informací spolu s odpovídajícími službami, přičemž informace jsou uloženy v digitální podobě a jsou dostupné prostřednictvím sítě [9].

Společné znaky digitální knihovny lze však vyvodit z mnoha existujících definic [8]:

- digitální knihovna se primárně zabývá organizací elektronické sbírky za účelem lepšího přístupu.

- digitální knihovna obvykle není jedna uzavřená entita (pro zdůraznění tohoto aspektu mnozí autoři používají zásadně a výhradně množné číslo – digitální knihovny).

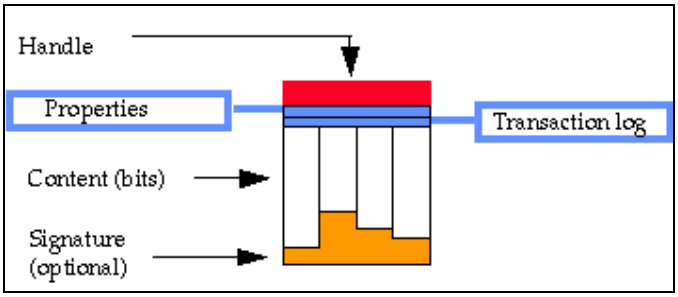
- informační zdroje v rámci digitální knihovny jsou heterogenní (způsobem organizace objektů a použitými platformami), dynamické (přidáváním a odebíráním objektů do/ze struktury digitální knihovny) a také multimediální v případě povahy dat.

- realizace digitální knihovny vyžaduje technologie pro transparentní propojení různých (autonomně spravovaných) informačních komponent.

- cílem digitálních knihoven je zajistit uživateli jednotný přístup k relevantním digitálním informacím, a to bez ohledu na jejich formu, formát, způsob a místo uložení.

Zřejmě nejznámější obecnou strukturu digitální knihovny představil Kahn a Welinsky [10], dle kterých je základním prvkem digitálních knihoven digitální objekt, který v tomto případě nepředstavuje pouze množinu bitů. Obsah téměř většiny základních digitálních objektů má určitou strukturu a informace (například autorská práva) musí být přidružena k digitálnímu objektu. Digitální objekt se tak skládá ze dvou částí – obsahu (content) a přidružených dat (informací) neboli metadat, které jsou v tomto případě tvořeny globálním jednoznačným identifikátorem digitálního objektu (označován jako „handle“), a dalšími nespecifikovanými metaúdaji. Metadaty se podrobněji zabývá kapitola 2.3.

Možná struktura digitálního objektu v oblasti digitálních knihoven dle [58] je znázorněna na obrázku 1. K samotnému obsahu digitálního objektu (content) je potřeba přiřadit také informace jako autorská práva. Digitální objekt se tak podle této navržené struktury skládá ze dvou částí – obsahu a přidružených dat (metadat).



Obrázek 1: Struktura digitálního objektu převzatý z [58]

V rámci možné struktury digitálního objektu znázorněné na obrázku 1 dle [58] jsou zahrnuty vlastnosti (properties), transakční log (transaction log) nebo signatura (signature). Obsahem digitálního objektu tak může být sekvence bytů reprezentující specifický digitální materiál, množina více datových objektů, množina identifikátorů objektu anebo jiné datové typy. Digitální objekty mohou být dle typu rozděleny do kategorií, kdy pro každou kategorii mohou být stanovena pravidla pro převod do jednotlivých typů digitálních objektů.

Typy digitálních (informačních) objektů používané v digitálních knihovnách se do určité míry podobají typům informačních objektů popsané v rámci specifikace MIME popsaným v předchozí kapitole. Stejně jako u typologie MIME se můžeme setkat s formáty typu text, obraz, zvuk nebo video. Digitální knihovny používají také jako specifický typ digitálního objektu software, který je v rámci MIME typu „application“ dále rozdělen dle jednotlivých podtypů dle charakter. Trojrozměrné digitální objekty lze přirovnat k MIME formátu typu model, který je charakterizován jako troj- a vícerozměrný systémů, ve kterém lze zavést pravoúhlou soustavu souřadnic.

Nejčastěji se můžeme v oblasti digitálních knihoven setkat s těmito typy digitálních objektů (viz také [34]):

a) **Textové digitální objekty**
b) **Obrazové digitální objekty**
c) **Zvukové digitální objekty**
d) **Video digitální objekty**
e) **Software digitální objekty**
f) **Trojrozměrné digitální objekty**

Další částí architektury digitální knihovny je tzv. handle-system představující distribuovaný mechanismus, který při obdržení identifikátoru digitálního objektu (například od uživatele zadávajícího dotaz ve vyhledávacím rozhraní v rámci digitální knihovny) vrátí seznam repozitářů (úložišť), které požadovaný objekt obsahují. Jistou podobnost tohoto handle-systému lze spatřovat například v indexu (datová struktura umožňující rychlé vyhledávání ve větších objemech dat) v rámci klasické databáze. Při vyhledávání určité hodnoty se použije index, který ukazuje na přesné umístění požadovaného objektu (nebo záznamu).

Pro zajímavost si můžeme uvést také typologii objektů často zmiňovaná koncem 90. let 20. stol. Vyvinuta byla kanadskou firmou Vancouver Webpages zabývající se službami v oblasti internetu. Tato firma v rámci vyhledávacího systému vyvinula vlastní robot „VWbot“ a speciální interaktivní formulář k tvorbě metadat nazvaný „Meta tag builder“. Typologie se skládala z 21 typů se stručnou charakteristikou. Jak uvádí Bratková [18] tato typologie označovaná také jako „VW96“ byla neúplná, účelově orientovaná a nevyvážená, neboť zahrnovala na jedné straně obecné typy a na druhé straně velmi specifické typy z oblasti obchodu a informačních technologií.

Je zřejmé, že typy informačních objektů specifikované standardem MIME definují přesné pravidla pro používání a zpracovávání. Standard MIME se díky jasné specifikaci (zajištěné zastřešením této problematiky mezinárodní organizací), interoperability a také možnosti snadné rozšiřitelnosti o nové typy informačních objektů, stal rozšířeným standardem nejen pro elektronickou poštu, ale také jako komunikační pro širokou oblast internetu. V rámci této diplomové práce se proto zaměříme na tyto vybrané typy informačních objektů specifikované právě typologií MIME:

- Textové (text)
- Grafické (image)
- Zvukové (sound)

2.2. Identifikace informačních objektů

Před samotným přiblížením vybraných typů informačních objektů si nejprve obecně představíme problematiku popisu informačních objektů. Právě problematika popisu in-

formačních objektů (zejména síťově dostupných informačních objektů v elektronické podobě) pomocí metadat úzce souvisí s vyhledáváním.

Se samotným popisem informačních objektů také souvisí potřeba dané objekty identifikovat a tuto identifikaci dynamicky a jednoznačně propojit s informačním objektem. Jedním z konceptů identifikátoru je identifikátor URI (Uniform Resource Identificator, viz [49]), který byl poprvé vydán v roce 1994 a jehož autorem byl Tim Berners-Lee.

Jak uvádí například Hilse a Kothe [51] termín URI lze klasifikovat jako a) název (name), b) vyhledávač (locator), c) oboje – název i vyhledávač. Oba tyto koncepty (name, locator) se poté vyvíjely samostatně jako URN (Uniform Resource Name, viz [50]) a URL (Uniform Resource Locator, viz [48]). Tradiční způsobem identifikace byl v prostředí například URL V polovině 90. let 20. století byla představena myšlenka použití jmenných prostor pro zaznamenání jména a umístění objektu. Tyto identifikátory objektů jsou registrovány centrálně, a proto jsou dané objekty vyhledány, přestože jejich umístění bylo změněno.

Účelem a funkcí konceptu URN je poskytovat unikátní (jednoznačný na globální úrovni) a trvalý identifikátor, který umožní rozpoznání a zpřístupnění jak charakteristik daného objektu, tak objektu samotného. Jak uvádí Bartošek [8], požadavky pro koncepci URN jsou následující:

- Globální rozsah (jednoznačné jméno nezávisle na umístění)
- Perzistence (trvale přidělené jméno, platící i po zániku popisované entity)
- Škálovatelnost (jméno lze použít pro jakýkoliv typ entity)
- Legacy support (systém musí podporovat existující identifikační systém)
- Rozšiřitelnost (možnost budoucího rozšíření systému)

Syntaxe konceptu URN je v podobě:

```
URN:nid:nss
```

kde nid (Namespace Identifier) představuje identifikátor určitého identifikačního systému (např. DOI) a nss (Namespace-specific String) je konkrétní identifikátor v rámci daného systému.

Mezi skutečnosti omezující globální rozšíření tohoto konceptu jsou například [8]:

- Směrovací systém URN není rozšířen v rámci internetu
- Budování a udržování směrovacích služeb představuje značné finanční náklady
- Nejasnosti s přidělováním URN entitám (problém verzí, různých formátů)
- Podpora URN webovými prohlížeči není taková jako v případ URL.
- Problém jednoduchého a násobného směrování – směrovací mechanismy vrací jediné URL. V případě URN mohou obecně vyžadovat identifikaci více odpovídajících služeb.

Jedním z identifikátorů vycházejícím z URN specifikace je například PURL neboli perzistentní URL. Tento systém realizovaný organizací OCLC (Online Computer Library center, http://www.oclc.org) se snaží odstranit problém s identifikací pomocí URL neboli závislost objektu a jeho umístění. PURL je tak URL poskytující nepřímou adresaci objektu. Informační objekt (zdroj) na internetu dostane přidělený identifikátor PURL, který obsahuje skutečnou URL adresu objektu [52]. Obecný tvar adresy PURL tak vypadá následovně: `http://purl.oclc.org/net/objekt`.

Technologie nazývaná „The Handles System“ [53] vyvinutá korporací CNRI (Corporation for National Research Initiatives, http://www.cnri.reston.va.us) představuje infrastrukturu obecné architektury digitální knihovny pro jednoznačnou identifikaci digitálních objektů. Syntaxe identifikátoru handle je v podobě:

```
hdl: Handle Naming Authority/Handle Local Name
```

kde, „handle naming authority“ představuje tzv. pojmenovávací autoritu, která je přidělována hierarchicky (nejvyšší úrovní je přidělována centrální autoritou, zbylé části oddělené tečkou již lokálně). „Handle local name“ již představuje libovolný název v rámci dané pojmenovávací autority.

V této části je třeba také zmínit systém pro identifikaci digitálních objektů DOI (Digital Object Identifier, http://www.doi.org), který je spravován a kontrolován Mezinárodní nadací DOI (International DOI Foundation, IDF). Jedná se o systém, jehož základem se stal směrovací mechanismus handle zmíněny výše. DOI představuje jak nový kon-

cept metadat pro objekty v podobě také aplikací nebo služeb, tak také poskytuje strukturu pro vzájemnou součinnost [57]. Příklad Syntaxe DOI vypadá následovně:

doi: *10.1234/NP5678*

kde 10 představuje pojmenovávací autoritu (číslo 10 je přiděleno nadaci IDF), dále po tečce následuje numerický identifikační kód registrující organizace. Po oddělujícím znaku „/“ následuje existující identifikátor (například lze použít i identifikátor ISBN) [58].

2.3. Metadata

Pojem metadata velmi souvisí se vznikem a rozvojem internetu [11], a to především služby World Wide Web (WWW). Výraz metadata se začal objevovat v polovině 80. let v souvislosti s budováním kolekcí či archivů digitalizovaných textů. Poté na ně navázalo v průběhu 90. let budování komplexnějších digitálních knihoven.

Samotné slovo metadata pochází z řeckého meta = mezi, za a latinského data = to co je dáno. Termín data si lze rozšířeněji představit jako vyjádření skutečnosti ve formátu, který lze následně zpracovat. Metadata jsou nejčastěji charakterizována jako data o datech. Můžeme se také setkat s definicí metadat jako informace o datech. Metadata můžeme také označit jako strukturovaná a zakódovaná data popisující vlastnosti entit (information-bearing entities), která pomáhají dané entity identifikovat. Účelem metadat je zejména klasifikace informací.

Tim Berners – Lee uvádí [12], že metadata jsou data, neboť v obecné rovině představují informace o informacích. Metadata mohou být uložena jako každá jiná data v nějakém zdroji.

Formát metadat je tvořen množinou nezávislých výroků, které reprezentují údaje o zdroji. V počítačovém systému nabývají výroky formu jména či typu výroku a souboru dalších parametrů. Výroky mohou existovat samostatně a nezávisle. Ke každé vlastnosti náleží konkrétní hodnota, obecně pak hovoříme o modelu dvojice (páru) vlastnost / hodnota.

Za zmínku také stojí další definice charakterizující metadata jako stroji srozumitelné informace o webovských zdrojích a jiných objektech [12]. Klíčové v této charakteristice metadat je, že jde o stroji srozumitelné informace. Předpokládá se jejich zpracování v/pro nejrůznější aplikace pomocí inteligentních programů označovaných výrazem „agenti“. Znamená to, že informační jazyk uložených metadat musí být naprosto formálně logicky správný a jednoznačný.

Bartošek [8] zmiňuje možnost nahlížení na metadata také v kontextu digitálních knihoven jako na počítačově zpracovatelné strukturované informační objekty, které popisují vlastnosti jiných informačních objektů.

Z hlediska obecného použití se metadata rozdělují na popisná (obecný popis pro účely vyhledávání zdrojů, identifikaci a selekci), strukturální (charakteristika formátu a struktury zdroje pro dosažení správného uložení a zobrazení zdroje) a administrativní (sloužící pro účely správy zdroje, archivaci atd.) [13].

Existuje velmi mnoho odlišných metadatových schémat. Důvodem je široký záběr služeb, které mohou metadata v prostředí sítově dostupných informačních objektů podporovat. Metadata tak plní celou řadu funkcí [76]:

- **Dokumentační** (popis důležitých charakteristik zdroje)
- **Identifikační** (jednoznačná identifikace nezávisle na čase a prostoru)
- **Lokační** (užití dat ke zjištění umístění zdroje)
- **Vyhledávácí** (zjištění existence zdroje)
- **Hodnotící** (klasifikace elementů, jež přispívá k rozhodnutí, zda se jedná o relevantní zdroj)
- **Selekční** (vyhledání informačního objektu na základě jmenných či věcných charakteristik)
- **Administrativní** (například v podobě informací o autorských právech, e-business, elektronický podpis atd.)
- **Archivační** (zajišťuje identitu a kontext zdroje za účelem správného uložení a zpracování)
- **Technická** (hlavičky HTTP, konfigurace snímacího zařízení, podstatné znaky grafických souborů, specifikace hardwarové platformy pro emulaci)
- **Sjednocující** (jednotný formát metadat umožňuje řízení a správu různorodých kolekcí).

K zajištění důvěryhodnosti a spolehlivosti tvorby metadat v případě služby World Wide Web budou v celkové architektuře hrát významnou roli jmenné prostory - dokumenty s příslušnými identifikátory URI, které budou obsahovat slovníky se jmény či typy vlastností, jež jsou definovány podle stanovené metodiky v rámci příslušných specifikací metadat [11]. Předpokládá se jejich hypertextové propojení s tvořenými záznamy metadat jednotlivých webovských zdrojů. V tomto případě půjde o specifické výroky typu vztahu mezi dvěma zdroji, které budou realizovány přes hypertextové odkazy. Celý takový výrok bude tvořen typem výroku a dvěma identifikátory URI. Příklady formátů metadat jsou uvedeny v další kapitole 2.4.1.

2.3.1. Obecné formáty metadat

Zřejmě nejznámějším formátem metadat v prostředí internetu je formát tzv. Dublinského jádra (Dublin Core, dále jen DC, http://purl.org/dc). DC představuje soubor metadatových prvků, jehož záměrem je usnadnit vyhledávání elektronických zdrojů [16]. Původně byl vytvořen jako popis zdrojů na WWW sestavený přímo autorem, postupně ale zaujal instituce zabývající se formálním zpracováním zdrojů, jako jsou muzea, knihovny, vládní agentury a komerční organizace. Sklenák [2] uvádí tyto charakteristiky, které řadí DC mezi prominentní formáty:

- **Jednoduchost** – všeobecně srozumitelná sémantika umožňující laikům i specialistům používat tento formát.

- **Sémantická interoperabilita** – prosazování globálního souboru deskriptorů, které umožní unifikovat ostatní standardy pro určování obsahu.

- **Mezinárodní konsensus** – působení na mezinárodní úrovni je rozhodující pro efektivní vývoj infrastruktury pro zpřístupňování elektronických zdrojů v prostředí internetu.

- **Rozšiřitelnost** – DC umožňuje ekonomickou alternativu k podrobnějším a složitějším modelům popisu.

- **Modularita metadat v prostředí WWW** – implementování architektury RDF pro metadata konsorciem W3C umožnilo vzájemnou koexistenci vzájemně se doplňujících, samostatně udržovaných souborů metadat.

- **Modifikovatelnost** – každý prvek DC má přesnou definici, ale zároveň uspokojuje potřeby různých komunit možností modifikace pomocí tzv. kvalifikátorů.

Metadata byla také implementována do jazyka HTML v podobě META tagu pro uložení vnořených metadat v rámci hlavičky <HEAD> HTML dokumentu. V každém META tagu byla specifikována dvojice „jméno/hodnota“, která byla kódována pomocí atributu „NAME“ a atributu „CONTENT“. V hlavičce mohly být obsaženy vícenásobné údaje.

```
<META NAME=''title'' CONTENT=''Techniky vyhledávání''>
```

Pro hodnotu atributu NAME byla navržena zvláštní tečkovací notace, pomocí níž se udávalo jméno souboru metadat a (po tečce) jméno údaje. Později se tečky uplatnily i pro vyjádření typologie jmen údajů. Metadata Dublinského jádra obdržela jméno „DC“. Konkrétní příklad dvou údajů:

```
<META NAME=''DC.Title'' CONTENT=''Techniky vyhledávání''>
<META NAME=''DC.Autor'' CONTENT=Radek Bitrman''>
```

Jednotlivé používané pojmy v rámci metadat DC vysvětluje [14]. Aktuální struktura formátu Dublinského jádra se skládá z níže uvedených položek [17]:

1. **Název (Title)** – definováno jako jméno dané zdroji, typicky představuje název, pod kterým bude daný zdroj znám.

2. **Tvůrce (Creator)** – entity primárně odpovědné za vytvoření zdroji, kdy výraz entita může představovat osobu, organizaci nebo službu.

3. **Předmět (Subject)** – definován jako předmět obsahu zdroje. Může být reprezentován klíčovým slovem, frází nebo klasifikačním tříděním. Při popisu územních nebo dočasných témat zdroje se používá spíše Pokrytí (Coverage).

4. **Popis (Description)** – výčet obsahu zdroje a zahrnuje například abstrakt, obsah, grafické znázornění nebo i textový výčet zdrojů.

5. **Vydavatel (Publisher)** – entita zodpovědná za zpřístupnění zdroje. Vydavatelem tak může být fyzická osoba, organizace nebo služba (systémy apod.).

6. **Přispěvatel (Contributor)** – entita odpovědná za provedení příspěvku k obsahu zdroje. Jako příklad lze uvést osobu, organizaci nebo službu.

7. **Datum (Date)** – datum spojené s nějakou událostí v životním cyklu zdroje. Doporučuje se užívat normu ISO 8601.

8. **Typ (Type)** – definováno jako povaha nebo žánr obsahu zdroje. Doporučeno je používat řízených slov, například prostřednictvím DCMI Type Vocabulary [DCMITYPE]. Pro popis formátu, fyzického nosiče nebo vyměření zdroje se používá Formát (Format).

9. **Formát (Format)** – fyzická nebo digitální manifestace zdroje. Formulace definice je, kromě jiného, ovlivněna současnými náhledy a závěry IFLA (International Federation of Library Associations and Institutions, http://www.ifla.org) k problematice informačních zdrojů a dokumentů, které byly často v diskusích brány v potaz; komentář k definici doplňuje některé další údaje, jako jsou např. rozměry potřebné pro popis některých typů informačních zdrojů (obrazové, trojrozměrné apod.).

10. **Identifikátor (Identifier)** – jednoznačný odkaz na zdroj v rámci daného kontextu.

11. **Zdroj (Source)** – odkaz na zdroj, ze kterého byl popisovaný zdroj odvozen.

12. **Jazyk (Language)** – definování jazyku zdroje s použitím schválených norem, například v podobě RFC 3066.

13. **Vztah (Relation)** – definice odkazu na příbuzné zdroje.

14. **Pokrytí (Coverage)** – rozsah nebo záběr zdroje. Příkladem může být místo nebo lokace v geografickém vyjádření. Dočasné období lze vyjádřit jako datum, časové období. Doporučeno je používání řízených slovníků v podobě například Tezauru Geografických Názvů (TGN).

15. **Práva (Rights)** – informace o právech udržovaná ve zdroji nebo mimo něj.

DC má ambice stát se všeobecně rozšířeným a používaným standardem, který může díky své jednoduchosti posloužit i jako základna pro sémantickou interoperabilitu mezi jinými složitějšími formáty.

V zásadě se rozlišují dva způsoby propojení metadatového záznamu a informačního zdroje (objektu) [15]:

1. metadatový záznam je uložen samostatně a odděleně od zdroje (například katalogizační záznam v knihovně).

2. metadata jsou vnořena (embedded) přímo do samotného zdroje (používá se například značkovacích jazyků SGML, HTML apod.).

Pro doplnění si uvedeme další zajímavé formáty metadat. Formát tzv. nezávislých hlaviček TEI (Text Encoding Initiative Independent Headers) představuje komplexní formát pro kódování úplných textů všech typů se zaměřením na detailní textový rozbor. Pro fondy archiválií vznikl formát EAD (Encoding Archival Description), který měl zajišťovat zpřístupnění registrů v rámci Kongresové knihovny ve Washingtonu. Příklad systému metadat z oblasti informací státní správy USA je formát GILS (Government Information Locator Service).

Do tohoto přehledu formátů metadat lze také zařadit ještě známý formát pro popis technických zpráv, který byl zveřejněn v roce 1992 jako internetová norma RFC 1807. Tento formát byl určen technické komunitě v USA pro tvorbu a přenos metadat přes internet. Připravená specifikace je jednoduchá a účelná. Záznam zahrnoval základní úda-

je o dokumentu, důležité věcné údaje, nezbytné údaje administrativního charakteru pro přenos dat a také URI dokumentů (dostupných z lokálních systémů).

Formáty metadat pro označení informačních objektů se uplatňují také v případě grafických a zvukových objektů anebo se používají také u kancelářských programů. Kancelářské aplikace jako například Microsoft Word umožňují přidávat k samotnému dokumentu (souboru) dodatečné informace - jméno autora, název, klíčová slova, status, komentář, datum vytvoření, datum revize atd.

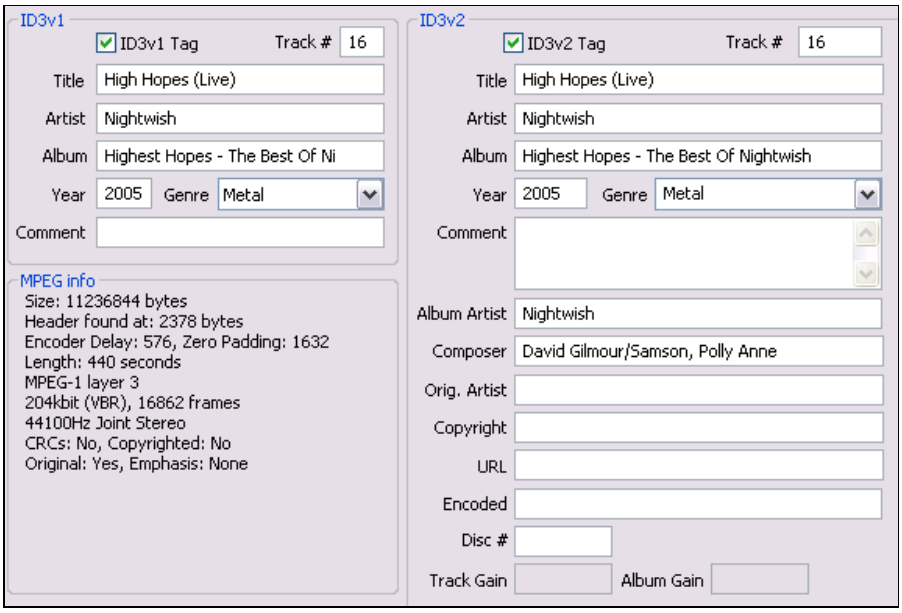
Popis obrazových informačních objektů může být obsažen již v samotném objektu, například prostřednictvím formátu Exif (Exchangeable image file format), který byl vyvinut asociací JEITA (Japan Electronics and Information Technology Industries Association, http://www.jeita.or.jp/english). Formát Exif představuje informační soubor obsahující informace o fotografii a který je připojitelný k samotnému grafickému objektu. V případě formátu Exif se používají metadata jako například:

- Výrobce a model fotoaparátu
- Datum a čas zhotovení snímku
- Clona
- Citlivost ISO
- Použití blesku
- Expoziční čas
- Ohnisková vzdálenost
- Rozlišení snímku
- Náhled snímku (do hlavičky Exif formátu se přidává také náhled o velikosti několika málo kilobajtů)
- Informace o místu pořízení (lze získat z GPS přijímače připojeného k fotoaparátu)

K přečtení údajů Exif se používá software, který tento formát podporuje jako například Exif Reader (http://www.takenet.or.jp) nebo Stuffware Photo Studio (http://www.stuffware.co.uk/photostudio/). Protože jsou metadata vložena přímo v souboru, může při editaci v programu, který s jejich přítomností nepočítá, dojít k jejich ztrátě. V současné době většina novějších programů metadata zachovává, ale při použití starších programů je dobré dávat na tuto možnost pozor.

Firma Adobe Systems Incorporated (http://adobe.com) vyvinula také standard Extensible Metadata Platform (XMP) pro lepší zpracování a ukládání metadat pro fotografie a obrázky. Avšak podpora tohoto formátu není příliš rozšířená.

Zvukové informační objekty obsahují také popis již v samotném objektu. Příkladem metadat popisující audio formát je populární ID3, který je spojován s formátem MP3. Metadata v případě ID3 jsou tak přidány k samotné zvukové informaci a doplňují jej o další informace. Ukázka metadat ID3 pro hudební soubory MP3 v rámci programu Nullsoft Winamp (http://www.winamp.com) je uvedeno na obrázku 2.



Obrázek 2: Ukázka metadat ID3 v zobrazená rámci programu Nullsoft Winamp.

V případě hudební sklady tak ID3 umožňuje přidat informace o názvu, autora, názvu alba, číslo skladby a další informace, které jsou poté uloženy v samotném hudebním souboru.

2.4. Textové informační objekty

Vyhledávání textových informačních objektů je prostřednictvím několika odlišných možností. Jednou z cest pro nalezení požadované textové informace je použití vyhledávání zadáním slova nebo případně kombinací více slov. Výsledkem je pak často výpis několika desítek, stovek nebo dokonce tisíců vyhledaných informačních objektů

(zdrojů), které dle vyhledávače odpovídají zadanému termínu. Další možností k uspokojení informačních potřeb uživatele je použití předmětových kategorií.

V případě vyhledávání textových informačních objektů můžeme narazit také například na problémy způsobené charakterem:

- Formátu
- Kódování
- Jazyka

Formát

U textových informačních objektů se můžeme setkat s čistě textovými formáty v podobě plain text (čistého textu), HTML, XML. Z hlediska indexování je důležité správné strukturování (například struktura tagů – značek XML ve zdrojovém kódu, kdy je přehledně vyznačeno záhlaví, hlavní část, autor, název právě použitím značek XML). Dalšími formáty pro uchovávání, distribuci, využívání a zpracování také textových informačních objektů jsou například i textové objekty ve formátu používající specifické aplikace jako například Microsoft Office (DOC, DOCX, PPT, PPTX, XLS, XLSX, MDB, MDBX atd.), otevřené formáty Open Office (ODT, OTS, ODG atd.). Samozřejmě lze v některých případech kombinovat textové i grafické objekty. Příkladem oblíbeného dokumentového formátu je mimo jiné formát PDF, který se používá pro prezentování informačních objektů jak textových, tak i grafických, a to zejména v prostředí internetu.

Kódování

Můžeme se setkat s textovými informačními objekty snad ve všech světových jazycích. Uživatel je tak při samotném vyhledávání omezován znalostmi cizích jazyků, ze kterých může vybírat při vyhledávání požadovaného informačního objektu. Použité kódování má také značný vliv na možnosti vyhledávání. Význam kódování pro vyhledávání lze spatřovat například v tom, že podle pořadí v kódové tabulce se postupuje pro třídění, tj. ovlivňuje složení indexů. Kódování znaků je také rozhodující pro výsledek jejich porovnávání při vyhledávání (string matching). Mezi nejfrekventovanější způsoby kódování patří například ASCII, CP 1250 (Windows-1250), ISO 8859-x (například ISO 8859-2 pro kódování latinské abecedy včetně češtiny), UTF (Unicode Transformation

Format) – nejčastěji se používá například UTF-8. Znázornění některých způsobů kódování zobrazuje tabulka 1.

Možnost vyhledávání je často odkázána na rozhodnutí správců (nebo organizace) zajišťujících technickou část vyhledávání dané databáze, funkce vyhledávání nebo také samotný text. Jako příklad si můžeme uvést problematiku češtiny a problémy s nejednotným kódováním. Nabízí se tak například možnost vyhledávání s českou diakritikou nebo bez. Často se ale můžeme v řadě případů setkat pouze s řešením v podobě vyhledávání bez diakritiky. Další možností je konverze daného dotazu (slov nebo slovního spojení) do různých typů kódování a tyto tvary jsou poté porovnávány s textovými objekty v databázi. Podobný princip lze například využití při paralelním vyhledávání, kdy se daný dotaz pošle do několik databází a výsledek se poté prezentuje jednotným způsobem.

Znaková sada	Počet znaků	Způsob kódování	Kódovací jednotka (počet bitů)
ISO 646 (ASCII)	128	ISO 7	7
	256	ISO 8	8
ISO 8859	256	v prvních 128 znacích kódování stejné jako ISO 646	8
		následujících 128 znaků je kódováno různě	
Windows	256	v prvních 128 znacích kódování stejné jako ISO 8	8
		následujících 128 znaků je kódováno různě	
ISO 10646/Unicode UCS-4	4.249.967.296	UTF-32	32
		UTF-16	16
ISO 10646/Unicode UCS-2	65.536	UTF-32	32
		UTF-16	16
		UTF-8	8
		UTF-7	7

Tabulka 1: Znázornění nejfrekventovanějších způsobů kódování znaků dle Kučerové [54]

Jazyk

Samotná čeština je, zejména oproti angličtině, jazykem velmi ohebným. Při plnotextovém vyhledávání může uživatel narazit na problémy spočívající se samotným jazykem – například skloňování, časování, morfologické tvary atd. Proto se používá například stemming, derivace a lemmatizace.

Při vyhledávání a indexování textových informačních objektů se můžeme setkat nejen se zmíněnými příklady problémů, ale také s řadou dalších překážek, které brání nalezení daných informačních objektů. Jednou z takových překážek může být například tzv. neviditelný web (deep web, invisible web). Používá se také pojem hluboký nebo skrytý web. Jedná o obrovskou oblast internetu obsahující informační objekty, které jsou nepřístupné klasickému vyhledávání pomocí internetových vyhledávačů. Neviditelný web obsahuje například ověřené informace zpracované v databázích přístupné na komerčním základu. Přístup k těmto objektům je pro autorizované uživatele, chráněn heslem. Problematika neviditelného webu a vyhledávání zmiňuje například Wright [viz 55] Jednotlivé informační objekty se mohou generovat dynamicky až na základě interakce uživatele se systémem. Velikost neviditelného webu byla například na základě studie Michaela Bergmana [56] odhadována na přibližně 400 až 550x větší než velikost viditelného webu. Přesnou velikost viditelného i neviditelného webu nebude patrně možné určit nikdy.

2.5. Grafické informační objekty

Informační objekty v podobě grafického vyjádření (obrázek, obrazová data v podobě například fotografie) se díky neustálému technologickému pokroku stále více objevují v digitální podobě pro každodenní získávání, zpracovávání, využívání a další distribuci (a to nejen v prostředí internetu). Grafické znázornění je v určitých případech nejvýstižnějším vyjádřením. Nyní si stručně přiblížíme některé aspekty informačních objektů v podobě statického nebo dynamického obrázku (například geografické informačních systémy - GIS).

V zásadě se rozlišují dva základní způsoby ukládání obrazových informací – vektorová grafika a bitmapová (rastrová) grafika. Vektorová grafika je způsob ukládání obrazových informací v elektronickém prostředí pomocí vektorů (geometrických objektů

v podobě bodů, přímek, křivek, polygonů) spojujících tzv. kotevní body. Tyto křivky mohou mít barevnou výplň formou jednolité plochy nebo barevného přechodu.

Základem vektorové grafiky jsou tzv. Bézierovy křivky pojmenované podle francouzského matematika Pierra Bézierea. Ten vyvinul matematickou metodu, pomocí níž byl schopen popsat libovolný úsek křivky pouze za pomoci čtyř bodů. Křivka je popsána pomocí dvou krajních bodů (tzv. kotevní body), které definují danou úsečku, a dvou tzv. kontrolních bodů, které určují vlastní tvar křivky [19]. Spojnice mezi kontrolním a kotevním bodem je tečnou k výsledné křivce. Tímto způsobem lze popsat i tu nejsložitější křivku, jakou jsme schopni nakreslit. Křivka nám vytvoří cestu, která může být otevřená nebo zavřená, s výplní či bez výplně. Hlavní výhodou vektorové grafiky je možnost libovolného zvětšování nebo zmenšování již vytvořeného obrázku bez sebemenší ztráty kvality, neboť vektorový obrázek se přepočítává a přizpůsobuje. Dále je možné pracovat s každým objektem v obrázku odděleně a v neposlední řadě výsledná velikost vektorového obrázku je obvykle mnohem menší než u rastrového obrázku.

Obecně lze říci, že vektorová grafika je vhodná pro tvorbu log, diagramů, sazbu, animace a jednoduché ilustrace. Editorů pro tvorbu vektorové grafiky existuje poměrně dost, ale mezi nejznámější patří Adobe Illustrator (http://www.adobe.com), Corel Draw (http://www.corel.com), anebo v případě animací v prostředí Macromedia Flash (produkt Macromedia též od firmy Adobe). Soubory jsou vytvářeny většinou s konkrétní příponou jako např. CDR (pro Corel DRAW) nebo AI (pro Adobe Illustrator). Některé programy umějí číst i soubory s cizími příponami, ale v zásadě je lepší mít data uložena v nějakém univerzálním formátu, jakým je například EPS (Encapsulated PostScript), který byl speciálně vytvořen pro přenos obrazových dat určených pro tisk. Velmi populární je pro vektorový přenos dat také komplexněji pojatý formát PDF.

Bitmapová (rastrová) grafika je složena z rastru (neboli pomyslné sítě bodů, tzv. bitmapy), kde každý bod má definovanou svou barvu a jas. Obraz je tak definován pomocí zpravidla obdélníkového rastru pixelů nebo bodů, z nichž každý nese svou vlastní informaci o vzhledu. U obrázků v obrazovém prostoru RGB má každý pixel alespoň tři bajty – pro každou z barev (R – red, G – green, B – blue) je definována její intenzita. Stejně tak se používá barevný prostor RGB i u vektorové grafiky. Čím hlubší barevný prostor (čím více možných barevných tónů), tím datově objemnější informace o každém bodu. Nejmenší barevnou hloubku má černobílá grafika, kde pro vyjádření stavu bílá a černá stačí každému pixelu pouze jeden bit. Každá bitmapa musí mít definovánu

svou výšku (počet pixelů vertikálně), šířku (počet pixelů horizontálně) a barevnou hloubku (počet bitů na pixel). Tento způsob popisu obrázků používá například televizní přijímač, digitální fotoaparát nebo počítačový monitor. Hlavní výhodou bitmapové grafiky je její široká podpora v podobě formátů jako BMP, GIF, TIF nebo JPEG. Tyto formáty je možné zobrazit téměř na každém prohlížeči nebo operačním systému (OS). Mezi nevýhody rastrové grafiky patří například zhoršení obrazové kvality obrázku při změně velikosti.

Nejpoužívanějším programem pro tvorbu a úpravy rastrové grafiky pro internet i pro tisk je v současné době Adobe Photoshop. Jeho nativní formát PSD (Photoshop Document) podporuje ukládání rastrové grafiky ve vrstvách spolu s vektorovými objekty a editovatelným textem.

Pro uložení obrázků či fotografií v odpovídající kvalitě se nejčastěji používá rastrový formát TIF (Tagged Image File Format, příp. TIFF), který je však, stejně jako většina bezztrátových rastrových formátů, pro svou datovou náročnost nevhodný pro použití na webu či v digitálních fotoaparátech. V prostředí internetu je zřejmě nejrozšířenějším rastrovým formátem GIF (Graphic Interchange Format) a JPEG (Joint Photographic Experts Group).

Mezi specifické obrazové informační objekty můžeme také zařadit tzv. interaktivní mapy neboli grafické informační systémy (GIS), které představují informační systém pro získávání, ukládání, analýzu a následnou vizualizaci dat, která mají prostorový vztah k povrchu Země [46]. Hlavní výhodou GIS v podobě například interaktivních map je spojení běžných map s informacemi o objektech. K libovolnému místu na mapě lze připojit další informace (např. fotografii, text, video apod.). Součástí interaktivních map je možnost vyhledávání konkrétních objektů.

Zajímavou oblastí v případě obrazových informačních oblastí je také samotná ochrana práv – copyright. Bitmapové obrázky vystaveny v prostředí internetu jsou neustále využívány dalšími uživateli pro jejich vlastní potřebu nebo také pro další využití v oblasti komerční i privátní, a to bez doplnění informace ohledně vlastnických práv. Kvůli tomuto zneužívání se používají různé způsoby ochrany grafických informačních objektů. Jednou z cest je například snížení velikosti nebo kvality daného obrázku, aby se předešlo případnému zneužití. Často se také můžeme setkat s ochrannou v podobě zabránění uložení obrázků z webového prohlížeče pomocí pravého tlačítka myši. Ale

zřejmě nejpopulárnější formou ochrany je tzv. vodoznak (watermark). Jde o speciální ochranný znak, který aplikace zakomponuje do bitmapového souboru tak, že je patrný běžným okem a navíc poměrně statečně odolává změnám obrázku např. v grafických editorech. Účelem tohoto vodoznaku je, aby uživatel na první pohled poznal, že daný obrazový informační objekt je chráněn autorských právem.

Další možností je přidání digitálního vodoznaku, což představuje techniku přidání skrytých informací o autorských právech nebo dalších informací přímo do obrazového nebo (také i zvukového) informačního objektu. Tyto přidané informace představují skupinu bitů popisující informaci týkající se samotného obrázku nebo případně autora (jméno autora, místo, rok atd.). Účelem digitálního vodoznaku je možnost vysledování obrázků chráněných autorskými právy. Tyto dodatečné informace lze v jistém směru považovat za metadata.

2.6. Zvukové informační objekty

Informační objekty v podobě zvukového nebo hlasového vyjádření patří také mezi stále častější objekty používané pro uchovávání a zpracovávání informací v digitální podobě. Audio díla mohou v některých případech mnohem efektivněji umožnit vyjádření, uchování a další zpracování informace než například textová. Zvukové informační objekty můžeme rozdělit na dva typy – statické a tzv. „stream“. Mezi nejčastější formáty zvukových informačních objektů můžeme zařadit MP3, Ogg (oga) nebo AAC. Dalším příkladem je také formát MIDI (Musical Instrument Digital Interface), který umožňuje digitálním hudebním nástrojům a počítačům komunikovat v reálném čase. Tento formát byl rozšířen zejména v 80. a 90. letech minulého století.

Nejznámějším formátem je samozřejmě MP3 (neboli MPEG-1 Audio Layer 3), což představuje formát ztrátové komprese zvukových souborů, založený na kompresním algoritmu MPEG (Motion Picture Experts Group). Při zachování vysoké kvality umožňuje zmenšit velikost hudebních souborů v CD kvalitě přibližně na desetinu, u mluveného slova však dává výrazně horší výsledky. Formát MP3 se stal oblíbeným při uchovávání a přehrávání hudby na počítačích a internetu. MP3 formát využívá nedokonalost lidského sluchu, kdy ze vstupního signálu se odeberou informace, jež člověk neslyší, nebo si je neuvědomuje. Využívá se principů časového a frekvenčního maskování. Komprese zvuku podle standardu MPEG-1 obsahuje 3 vrstvy, jež se liší kvalitou a obtížnos-

tí implementace. Advanced Audio Coding (zkráceně AAC) je ztrátový zvukový kodek. Byl vyvinut jako logický následovník formátu MP3. Datový formát Oga (v případě audio formátu) byl vyvinut v rámci otevřeného standardu pro vytvoření volně dostupného softwaru pro kódování a dekódování multimediálního formátu.

Za zmínku také stojí tzv. „streaming“, který je specifickým typem zvukového informačního objektu. Streaming („proudění“) lze chápat jako způsob (označované podle samotného způsob přenášení dat) pro nepřetržité přijímání a okamžité zobrazování multimediálního obsahu zejména prostřednictvím sítě internet. Mezi příklady typu „stream“ lze zařadit internetové rádio.

3. Informační objekty v databázích

Tato část diplomové práce se nejprve věnuje stručnému přiblížení pojmu vyhledávací stroj (kapitola 3.1). Kapitola 3.2. poté charakterizuje indexovací techniky a také techniky vyhledávání se zaměřením na specifika dle typu vyhledávaného objektu V dalších kapitolách si uvedeme příklady databází volně dostupných na celosvětové síti internetu obsahující uvedené typy informačních objektů zmíněných v kapitolách 2.5 až 2.7. Blíže si představíme možnosti, nabízené funkce a také různé aspekty dané různými typy informačních objektů.

Samozřejmě není možné uvést všechny představitele databází, neboť jich existuje nepřeberné množství. Proto se zde zaměříme pouze na příklady, které autor této práce považuje za zajímavé. Mezi představitele různých databází nejsou zahrnuty databáze komerčního charakteru (například komerční databázová centra), která nabízejí přístup k informačním objektům na bázi placeného přístupu.

3.1. Vyhledávací stroje

Vyhledávací stroj (VS) nebo také označovaný anglickým termínem search engine (SE) představuje systém (aplikaci), který je určen pro vyhledávání uložených dat v indexu objektů, čímž minimalizuje čas potřebný a množství dat potřebné pro nalezení odpovídající. Vyhledávací stroj je zpravidla tvořen čtyřmi základními částmi [1]:

- Robot (sběr informací)
- Indexační program (zpracovává informace získané roboty)
- Vyhledávací program (vyhledávací algoritmus)
- Grafické rozhraní (umožňuje zadávání dotazů uživateli, předávání dotazu vyhledávacímu stroji a také zobrazování výsledku hledání uživateli).

Pojem vyhledávací stroj neboli *search engine* je znám především v souvislosti s internetem a vyhledáváním v prostředí WWW. Zmíníme tedy některé aspekty vyhledávacích stroj v prostředí internetu.

3.1.1. Vyhledávací stroje a internet

Vyhledávací stroj neboli vyhledávač lze také v prostředí internetu definovat jako aplikaci, schopnou vyhledávat v rámci velkého počtu informačních objektů pomocí předběžné indexace jejich obsahu. Vyhledávač může být v širším smyslu chápán i pro webová sídla, která užívání této aplikace pomocí dialogového rozhraní umožňují [20].

Vyhledávací stroje jsou velmi populární a také užitečným prostředkem pro lokalizaci informací v prostředí internetu. Vyhledávacímu stroji je zadán dotaz neboli textový řetězec, který formou určité syntaxe vyjadřuje informační požadavek uživatele. Dotaz je předkládán vyhledávacímu stroji prostřednictvím dialogového rozhraní. Základním stavebním prvkem dotazu je výraz (term), který se může skládat z jednoslovných i víceslovných jazykových konstrukcí. Vyhledané dokumenty odpovídající zadanému dotazu se označují jako hit.

Zapotřebí je také další komponenty v podobě robota (crawler, spider), který pravidelně shromažďuje informační objekty, ze kterých je poté vytvářen vlastní index. Prioritním cílem robota je reprezentování informačních dokumentů v aktuální a nejúplnější podobě. Předmětem je převážně plný text, v menší míře poté obrázky, zvuky a další obtížně indexovatelné mediální formy. Robot tak vlastně plní danou databázi informačními objekty. Vytvářený index představuje způsob organizace veškerých informací (nashromážděny právě činností robota) tak, aby bylo umožněno rychlé a efektivní vyhledávání. Jak uvádí Sklenák [2] index si můžeme zjednodušeně představit jako seznam klíčových slov, u kterých je připojen seznam informačních objektů, kde se dané klíčové slovo vyskytuje.

Schémat vyhledávacích strojů existuje nepřeberné množství, každý vyhledávací stroj funguje na poněkud odlišných principech. Rozdíly existují ve způsobu úložišť, přístupu k datům, možnostem budování indexů apod. Z tohoto důvodu si zde blíže uvedeme v současné době bezesporu nejpopulárnější internetový vyhledávač Google (http://www.google.com), který od svého vzniku v roce 1996 výrazně předstihl konkurencí. Někteří roboti shromažďují nepřetržitě, jiní mohou být spouštěni v pravidelných intervalech. Počet dotazů, které mohou být vyhodnoceni prostřednictvím indexu, je limitován pouze výpočetními kapacitami serveru.

Google používá ke shromaždování dat dva typy robotů – tzv. deep boty a fresh boty. Hlavními nástroji Googlu pro získávání dat jsou deep boti, neboť procházejí opakovaně určitý webový prostor, ukládají datový obsah navštívených webovských sídel (případně zaznamenávají informace o změnách v tomto obsahu). Toto vyhodnocování informací získaných od deep botů je velmi náročně po stránce hardwarové, proto se aktualizace hlavních indexů uskutečňuje zhruba jednou do měsíce. Hlavním cílem fresh botů není shromažďování nových dat, ale pouze průběžné zachycování změn v obsahu určité skupiny stránek. Tyto stránky jsou vybírány pravděpodobně podle významnosti a frekvence změn. Změny hlavního indexu na základě nových informací od fresh botů probíhají neustále [20].

Stále ale zůstává značná část prostoru internetu skryta v oblasti souhrnně nazývané jako neviditelný web. Do této oblasti můžeme zařadit:

- Informační objekty (informační zdroje) poskytované za úplatu, kdy je nutné se nejdříve přihlásit a vstoupit do jinak nepřístupné části webovského sídla.

- Dynamicky vytvářené informační objekty, které neexistují v podobě stálých souborů. Příkladem mohou být výstupy nejrůznějších databázových služeb – jízdní řády, výpis obchodních informací (např. z obchodního rejstříku), plánování trasy atd.

- Informační objekty obsažené příliš hluboko ve struktuře webového sídla. Roboti procházejí každý web pouze do určité hloubky.

- Informační objekty, jejichž vyhledání robotem vyhledávačů správce (případně autor) webového sídla zakázal (např. konfiguračním souborem robots.txt).

Vyhledávací stroje shromažďují a indexují informační objekty v nejrozšířenější textové podobě, ať už ve formátu značkovacího jazyka HTML (a jeho dalších rozšířeních), nebo také stále častěji ve formátu PDF, DOC nebo PS. Bohužel ale opravdu sémantické vyhledávání informačních objektů na základě obsahu grafických nebo zvukových informačních objektů nejsou ve standardních vyhledávacích mechanismech v prostředí internetu zcela rozšířeny.

Netextové informační objekty jsou u standardních vyhledávačů hledány na základě popisu obrázku nebo názvu obrázků. Při vyhledávání tak můžeme obdržet nepředstavitelné množství nerelevantních výsledků. Princip vyhledávání podle sémantického obsahu objektu (CBIR) je naprosto logickou a potřebnou cestou do budoucna.

Zajímavou skupinou vyhledávačů jsou také tzv. metavyhledávače, kdy je pro vyhledávání informací v internetu využíváno více vyhledávacích systémů s cílem dosažení vyšší úplnosti. Tyto systémy automaticky a simultánně provedou za uživatele dotazy do vyhledávacích systémů a zobrazí celkové výsledky vyhledávání. Při využívání metavyhledávacích systémů se stačí naučit pracovat pouze s jedním rozhraním a uživatel se již nemusí řešit rozdíly (způsob zadávání dotazů, uspořádání formuláře, syntaktická pravidla pro formulaci dotazu atd.) mezi jednotlivými rozhraními různých vyhledávacích systémů. Uživatel metavyhledávacího systému také nemusí sledovat neustálý vývoj v oblasti vyhledávacích systémů.

Za zmínku stojí také vyhledávání obrázků vyhledávače Google, který nabízí v rámci rozšířeného vyhledávání obrázků také možnost zvolit typ obsahu – a) jakýkoliv obsah, b) obsah zpráv nebo c) obličeje. Dále nabízí možnost vyhledávat podle velikosti, typu souboru, barvy obrázků nebo dle domény [22]. S podobnými možnostmi se můžeme setkat ve stále větší míře u dalších známějších vyhledávačů. Například vyhledávač Altavista (http://www.altavista.com) nabízí vyhledávání obrázků nebo dokonce hudby s určitou možností detailního vyhledávání. Altavista nabízí omezit vyhledávání obrázků dle barvy (barevné, černobílé, jednobarevné), zdroje nebo dle velikosti (rozlišení). Specializovaný vyhledávač obrázků Picsearch (http://www.picsearch.com) nemá tak velkou databázi obrázků jako zmiňovaný Google nebo Altavista [23], ale dokáže například rozlišovat mezi nehybnými obrázky a animacemi (obrázky typu GIF například). Technologicky zajímavou je služba Visoo (http://www.visoo.de), která umožňuje rozpoznávat nápisy v obrázcích prostřednictvím OCR technologie (Optical Character Reading).

Možnost vyhledávání zvukových informačních objektů v rámci standardních internetových vyhledávačů není příliš rozšířená. Altavista (http://www.altavista.com) nabízí možnost vyhledávání audio objektů na základě názvu souboru nebo přidruženého popisu. Uživatel může specifikovat například typ souboru (MP3, WAV, Windows media) nebo délku stopy (délka stopy více / méně než 1:00 minuta anebo vše). Dalším příkladem vyhledávače je AlltheWeb (http://www.alltheweb.com).

Můžeme se také setkat s řadou dalších vyhledávačů, ale jejich životnost je velmi krátká a relevantnost získaných výsledků je velmi tristní [23]. V případě audio objektů se tedy spíše můžeme setkat se službou nabízející přímo kolekce nahrávek. Tyto služby fungují na komerčním základu a soustřeďují hudební soubory ve své databázi. Uživatel poté může vyhledávat a také přímo nakupovat vybrané hudební nahrávky.

Zajímavým příkladem prototypu interaktivního vyhledávacího stroje pro grafické informační objekty v prostředí internetu je ImageScape [32], který vyvinul Michael S. Lew s kolegy. ImageScape (http://skynet.liacs.nl/) umožňuje uživateli interaktivně specifikovat požadovaný informační objekt pomocí náčrtku (kresby), přímek nebo připravených ikon.

3.1.2. Hodnocení efektivity vyhledávání

Jak uvádí Pokorný [21], efektivita vyhledávacích systémů je jednak dána rychlostí zpracování požadavku a uživatelským komfortem (forma interakce se systémem), jednak ve způsobu kladení dotazů a poskytování odpovědí, ale hlavně schopností systému poskytnout informaci o relevantních dokumentech (informačních objektech).

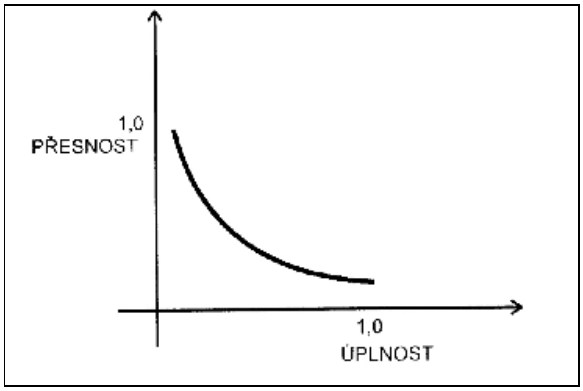
Koeficient úplnosti R (recall, výtěžnosti, rozsahu) udává pravděpodobnost, že relevantní informační objekt bude vybrán. Jinými slovy úplnost udává poměr nalezených relevantních informačních objektů a všech relevantních informačních objektů v systému [43].

$$\mathbf{R} = \frac{\text{počet vyhledaných a relevantních informačních objektů}}{\text{celkový počet všech relevantních objektů v systému}}$$

Koeficient přesnosti P (precision) představuje pravděpodobnost, že vybraný informační objekt bude relevantní neboli udává poměr nalezených informačních objektů a všech nalezených informačních objektů v systému (databázi). Při nízké přesnosti je také vyhledáno mnoho irelevantních informačních objektů.

$$P = \frac{\text{počet vyhledaných a relevantních informačních objektů}}{\text{celkový počet vyhledaných informačních objektů}}$$

Mezi přesností a úplností platí vztah nepřímé úměrnosti, nebolí čím vyšší přesnost, tím nižší úplnost a naopak. Proto není možné dosáhnout ideálního stavu – co nejvyšší hodnoty obou koeficientů [42]. Pokud tedy všechny vyhledané informační objekty se přesně týkají daného tématu, pravděpodobně jsme přišli o nějaké relevantní informační objekty. A pokud obdržíme co nejúplnější zachycení daného tématu, budou zahrnuty irelevantní informační objekty. Na obrázku 3 je znázorněn reálný funkční vztah obou koeficientů. Výpočet přesnosti vyhledaných objektů v rámci internetu je ale poněkud problematický, neboť počet relevantních dokumentů je neznámý.



Obrázek 3: Reálný vztah koeficientů přesnosti a úplnosti [42].

Pojem relevance (závažnost, důležitost) je nejčastěji definován jako vztah vyhledaného informačního objektu a informačního požadavku, případně informační potřeby [43]. Kučerová například odlišuje tři přesnější druhy relevancí:

- **formální** – představuje relevanci na úrovni syntaxe, která je určována vyhledávacím programem. Vyjadřuje tedy shodu textu dotazu s textovým obsahem informačního objektu.

- **věcná** – relevance na úrovni sémantiky vyjadřující úroveň shody významové blízkosti mezi obsahem informačního objektu a položeným dotazem.

- **pertinence** (osobní) – subjektivně určená relevance, která vyjadřuje vztah mezi obsahem informačního objektu a subjektivní informační potřebou uživatele.

3.2. Vyhledávání informačních objektů

3.2.1. Indexování

Indexování objektů patří k důležitým faktorům ovlivňující pozdější vyhledávání. V zásadě lze říci, že cílem indexování neboli vytváření indexů je optimalizace výkonu a rychlosti při vyhledávání relevantních objektů odpovídajících položenému dotazu. Bez vytvořeného indexu by vyhledávací stroj prohledával veškeré objekty v dané databázi, což by bylo časově náročné a také spotřebovalo značnou část výkonu procesoru.

Pokorný [21] charakterizuje index (invertovaný soubor) jako datovou strukturu, ve které uchováváme termíny charakterizující dané informační objekty. V prostředí databází se nejčastěji používají tyto hlavní typy indexů:

- primární (primary) - primární index obsahuje primární klíč, který se může v každé tabulce vyskytovat pouze jednou)
- unikátní (unique) - je tvořen ze sloupců obsahujících unikátní klíče)
- vedlejší index (secondary index) – Definicí jednoho či více indexů tohoto typu bývá v tabulce zajištěna optimalizace vyhledávání podle jednoho či více sloupců, pro které jsou dané indexy definovány.
- full-test – tento typ indexu je určen pro optimalizaci plnotextového vyhledávání v daném sloupci u dané tabulky.

Typů indexů existuje samozřejmě celá řada. Vysvětlení principů fungování z technického hlediska a podrobná charakteristika by byla mohla být námětem pro samostatnou práci. Pro účely této práce si tedy uvedeme pouze informativně některé typy indexů.

Většinou se u databází můžeme setkat s indexováním v podobě variant algoritmu označovaného jako B-Tree, které se staly de facto standardem pro přístup k záznamům v databázových souborech [77]. Dále se můžeme setkat s bitmapovými indexy, které se uplatňují při potřebě velice rychlého přístupu k velkému množství objektů (záznamů) s určitým klíčem. Nevýhodou je naopak jejich údržba při změnách v datové základně. Jako alternativu B-Tree indexů nabízí většina databázových systémů také hashované indexy, které používají tzv. hashovací tabulku. Hashování je technika organizace souboru poskytující velice rychlý přístup k řádkám souboru.

Textové informační objekty

V případě textových informačních objektů se můžeme setkat s automatickou indexací. Automatická indexace vychází z toho, jaké technologie zpracování textů jsou efektivně k dispozici, a aplikovat tyto technologie v souladu s určitými odpozorovanými vlastnostmi přirozených jazyků, jak uvádí Sklenák [2]. V podstatě jsou veškeré způsoby automatického indexování textů v přirozených jazycích založeny na výběru slov, kterými se bude indexovat, přímo z textu. Vybírají se tak především podstatná a přídavná jména, méně již slovesa. Pro indexování je vhodné používat pouze základní tvary slov neboli kmeny (anglicky označováno jako stemm). Jednou z technik je tzv. lemmatizace, která umožňuje určit morfologickou příbuznost porovnávaných slov.

Při automatickém indexování se jednak vybírají termíny k zařazení do indexu, jednak se často těmto termínům přiřazuje váha vyjadřující míru významu termínu pro obsahovou identifikaci informačního zdroje.

Jednoduchý prvek, který se při automatickém indexování textů obvykle používá pro vyloučení slov absolutně nevhodných pro indexování je tzv. slovník stop-slov (stop word list), označovaný také jako slovník nevýznamových slov. Jedná se o slovní druhy, které naprosto nemohou vyjadřovat věcný obsah (zájmena, předložky, spojky). Pro specializované databáze se do slovníku stop-slov zařazují i slova, která v běžném životě nejsou příliš frekventovaná, ale v dané oblasti tak všeobecná, že nejsou prakticky využitelná jako vyhledávací kritérium.

Nedostatky současného indexování Deerwester [59] rozděluje na dvě skupiny – tzv. „*synonymy*“ a „*polysemy*“. Pojem synonomy poté označuje problém, kdy různá slova

popisují stejný objekt. Uživatelé s různými znalostmi, potřebami a v různém kontextu mohou popisovat stejnou informaci za použití různých termů. Naproti tomu pojem polysemy představuje skutečnost, že většina slov má více než jeden přesný význam. U různých uživatelů nebo v různém kontextu se tak může daný term nabývat odlišných významů.

Deerwester [59] poprvé představil přístup k automatickému indexování nazvaný Latent Semantic Indexing (LSI). Tato technika se snaží překonat nedostatky spojené s vyhledáváním na základě termů, a to použitím statistických metod pro odvozování podobnosti mezi dokumenty a slovy. K samotnému zaznamenání která klíčová slova daný textový objekt obsahuje, LSI ještě zkoumá sbírku objektů jako celek pro zjištění dalších objektů, která obsahují některá shodná slova. LSI tak považuje objekty mající mnoho společných slov za sémanticky (významově) blízké a objekty s málo společnými slovy za sémantický vzdálená. Technika LSI používá matici, jejíž řádky odpovídají termům a sloupce dokumentům (objektům). Typicky se používají slova objevující se v dokumentu upravená na kořenový tvar. Mnoho různých experimentů bylo prováděny s využitím technik LSI, například Bradford [60].

Pro účely této práce lze u textových objektů částečně vycházet z [75] a techniky automatické indexace rozdělit takto:

- **slovní indexace** (automatická extrakce) – jedná se o automatickou indexaci založenou na extrakci indexačních termínů přímo z plného textu. Jednodušší a snáze programovatelná metoda, která se používá jako základ pro větší část funkčních systémů automatické indexace. V rámci této práce lze do této skupiny zařadit také indexování „plných textů“ s rozlišováním struktury dokumentu (plain text) nebo bez rozlišování struktury (HTML), indexování popisných polí záznamu apod.

- **pojmová indexace** (automatické přiřazování) - automatická indexace založená na určování indexačních termínů z řízeného slovníku nebo znalostní báze na základě jejich srovnání s výrazy z plného textu. Náročnější metoda, která není zcela běžná ve funkčních systémech.

Grafické informační objekty

Vedle klasických textových informačních objektů je potřeba indexovat také rovněž obrázky, zvuky, animace, signály atd. Potřeba indexovat velké kolekce multimediálních objektů je více než zřejmá. Uveďme si například rentgenové snímky, sekvence DNA, spektrální snímky, družicové záběry a nespočet dalších dat.

Pokud bychom chtěli vytvořit databázi obrázků, která umožňuje vyhledávání obrázku s ohledem na různé uživatele a jejich různorodé interpretace dotazu, museli bychom přidat ke každému obrázku anotaci popisující naprosto všechny prvky (zobrazené na obrázku) a jejich vzájemné vztahy. Je zřejmé, že tento způsob popisu obrázků je naprosto nereálný pro praktické využití.

V případě indexování grafických informačních objektů Pentland [25] zmiňuje dva typy tzv. sémantického indexování, a to "t*exture-like description of stuff* a "*object-like description of things".* Tyto způsoby indexování spočívají v předem vypočítaných grafických prvcích, které hrají obdobnou roli jako písmena a slova při vyhledávání textového řetězce. Uživatel může tedy použít tyto grafické prvky k popisu důležitých vizuálních jevů, které jsou brány jako dotaz a porovnávány s podobnými jevy v dané databázi. Uživatel tak může například vybrat obrázek kamenného mostu a počítač bude vyhledávat obdobné obrázky obsahující toto vyobrazení. Další možností je vyznačení například obličeje a zadat počítači vyhledání obrázků, kde se daný obličej vyskytuje.

Technika v podobě Latent Semantic Analysis (LSA) k překonání mezery mezi textovým a „obrázkovým" popisem byla použita v experimentu [viz 62], kde autoři experimentu používají k indexování vektory reprezentující předběžně vypočítanou část textur v daném obrázku. Obeid a kolektiv v tomto experimentu [62] také zmiňují využití vizuálního obsahu obrázku jako index (například barva, struktura a tvar hlavních rysů obrázku) v oblasti systémů vizuálního vyhledávání informací (Visual Information Retrieval, VIR). Nejen problematice indexování se věnuje také Lynch [61].

Indexování obrázků v rámci databáze grafických objektů na základě tzv. „2D-string" reprezentace (reprezentace na základě dvourozměrného řetězce) zmiňuje Chang [63]. Tento způsob indexování používá uložení objektů a jejich „prostorových vztahů". Další způsob indexování v podobě tzv. „graph-based" ukládání objektů zmiňuje Nabil [64].

Nejen v oblasti geografických informačních systémů (GIS) nebo data warehousingu, kde je potřeba vyhledat velké objemy dat, se můžeme setkat s tzv. bitmapovým indexováním. Bitmapové index si lze představit jako sady předem vypočítaných odpovědí na jednoduché otázky. Bitmapové indexy pracují se dvěma hodnotami (*true* nebo *false*) a označují tak, zda je odpověď na danou otázku pro ten který záznam v databázi pravdivá či nikoliv [69].

Techniky indexování grafických objektů v rámci této práce lze rozdělit na dva základní typy:

- **„klasické“ indexování** – indexování na základě popisných polí reprezentující název objektu, jméno autora, stručný popis, formát objektu (např. GIF, JPEG), klíčová slova apod.

- **sémantické indexování** – techniky indexování, které jsou stále středem výzkumu a vývoje pro dosažení opravdu efektivních metod. Jedná se například o již zmíněné indexování na základě reprezentace dvourozměrného řetězce, vektorů reprezentující předběžně vypočítanou část textur v daném obrázku, indexování vizuálního obsahu (například barva, struktura atd).

Zvukové informační objekty

Podobný způsob lze také využití pro indexování zvukových záznamů, kdy jsou jednotlivé noty, melodie, tempo indexovány a uživateli je nabídnuta možnost vyhledávat obdobné zvukové objekty shodující se s „dotazem“.

Za zmínku určitě stojí prototyp systému *Rough'n'Ready* [65], ve kterém jsou implementovány technologie pro indexování a následné prohlížení audio dat. Tento prototyp byl vyvinut pro automatické indexování a následné vyhledávání zpráv rozhlasového vysílání. Autoři tohoto projektu zmiňují technologie pro rozpoznávání jednotlivých řečníků (speaker), klasifikaci jednotlivých témat nebo rozeznávání jednotlivých jmen.

Příklad systému pro online indexování audio objektů, který vytváří index obsahu řeči (projevu) vložený do digitalizovaného zvukového objektu zmiňuje Ajmera [67]. Tento systém u kontinuálně přicházejících audio dat (objektů) nejprve zjistí hranice akusticky

homogenních částí, poté jsou jednotlivé části klasifikovány jako řeč, hudba nebo kombinace (tzv. mixture). Poté jsou převedeny do textové podoby pomocí systému automatického rozpoznávání řeči (Automatic Speech Recognition, ASR) a výsledná slova jsou metadaty.

Technika LSA se také používá v případě indexování zvukových informačních objektů. Systém indexování audio objektů v podobě mluvené řeči v rámci rozhlasového vysílání zmiňuje například Kurimo [66]. Zajímavým příkladem úspěšných technik pro budování indexů pro zvukové objekty zejména rozhlasového vysílání je systém THISL [68], který již nabízí dostatečnou přesnost a rychlost pro budování indexování.

Techniky indexování u zvukových objektů můžeme v rámci této práce rozdělit obdobně jako v případě indexování grafických informačních objektů:

- **„klasické“ indexování** – indexování na základě textových polí reprezentující název objektu, jméno autora, stručný popis, formát objektu, klíčová slova apod.

- **sémantické indexování** – techniky indexování v podobě LSA, rozpoznávání mluvené řeči, melodie, dále na zabarvení tónu, rytmu nebo podrobné strukturální analýzu zvukového objektu.

3.2.2. Vyhledávání

Tato část se zabývá přiblížením některých možností vyhledávání s ohledem na různé typy informačních objektů. Zmíněny budou zajímavé systémy vyhledávání, které jsou stále více zaměřeny na multimediální obsah a vytěžování významového obsahu objektu.

Textové informační objekty

Mezi nejběžnější modely používané pro vyhledávání textových informačních objektů patří booleovský model a vektorový model. Oba modely pracují na základě zadání termů v textové podobě pro vyhledávání. Podobné možnosti v případě vyhledávání ob-

razových nebo zvukových objektů jsou do určité míry také možné. Nyní si stručně představíme zmíněné již zmíněné nejběžnější modely vyhledávání.

Booleovský model

Vyhledávání pomocí booleovského modelu a jeho dalších rozšířeních patří mezi nejstarší a dosud nejrozšířenější. Tento model je také hlavním nástrojem vyhledávacích strojů v prostředí internetu, které používají základní operace booleovské logiky konjunkci (AND), disjunkci (OR) a negaci (NOT). Jako implicitní základní operací se používá konjunkce, pokud u výrazu není uveden žádný operátor.

Vyhledávací stroje používají také alternativní podoby pro některé operátory, například Google používá znak `-` `(mínus)` pro NOT nebo znak `|(kolmice)` pro OR. Booleovská logika obsahuje také operátor XOR, který je podobný operátoru OR (disjunkci), ale liší se v tom, že nevybírá výrazy současně.

Používá se také seskupování logických operací pomocí závorek, které stanovují prioritu při vyhodnocování dotazu. Pořadí vyhodnocování stejných operátorů se zpravidla provádí zleva doprava.

Kromě standardních logických operátorů se k dotazovacím jazykům založeným na booleovském modelu přidávají tzv. proximitní operátory (ADJ, NEAR, FAR, PARAGRAPH, SENTENCE, PHRASE atd.), pomocí nichž lze stanovit vzdálenostní omezení na výskyt daných termínů v textu. Používají se také numerické operátory (např. =, <, >, <= atd.)

Tento rozšířený model má přesto ve své základní podobě několik omezení, které zmiňuje například Pokorný [21]. Efektivita a jednoduchost modelu je důsledkem striktní interpretace booleovské logiky, kdy každá operace má pouze dva výsledky – „true“ a „false“. Stejně tak neexistuje možnost řízení velikosti výstupu, který obdrží uživatel po zadání dotazu. Již také existují modely, které zamezují striktní interpretaci booleovských operátorů a pokouší se ohodnotit vyhledané dokumenty podle jejich relevance.

Dalším omezení lze spatřovat v tom, že všechny obdržené výsledky dotazu jsou chápany jako stejně důležité, není možné je smysluplně uspořádat podle relevance. Toto je částečně řešeno, když algoritmy internetových vyhledávačů již nyní přisuzují

jednotlivým výrazům v rámci indexovaného dokumentu různé váhy (na základě četnosti, umístění v rámci dokumentu, obsažení v názvu dokumentu atd.).

Zmírnit striktní interpretaci operátorů (Pokorný používá pojem „změkčení „) lze například změnou modelu například na vektorový.

Vektorový model

Jak uvádí Pokorný ve své monografii [21], podstata vektorového modelu spočívá v reprezentaci záznamu představujících dokumenty i uživatelského dotazu pomocí vektorů. Tímto je umožněno pomocí skalárního součinu vektorů dokumentu a vektoru dotazu vyjádřit nejen shodu, ale také jakési kvantitativní ohodnocení míry této shody. Ve vektorovém modelu nelze rozlišit konjunktivní a disjunktivní dotaz. Operátor NOT je tak v tomto modelu velmi špatně realizován. Do jisté míry lze tuto operaci nahradit použitím intervalu pro váhy termů v dotazu <-1;1> namísto obvyklého <0; 1>.

Nevýhody tohoto modelu spočívají zejména v nejasné interpretaci významu vah termů vyskytujících se v požadavku. Není zřejmé, co udělá malá změna váhy s množinou dokumentů na výstupu. Koeficient podobnosti nemá jasný význam, což přináší jistou nesrovnalost do procesu vyhodnocování. Vyjadřovací síla vah je omezená. Například nelze formulovat dotaz ve tvaru (A OR B) AND (C OR D) pomocí vážení.

Za zmínku také stoji techniky pojmového vyhledávání v prostředí internetu. Průkopníkem pojmového vyhledávání na WWW lze dle Sklenáka [2] označit službu Excite, jejíž vyhledávání je založeno na metodě Intelligent Concept Extraction (ICE). Tyto technologie používají vyhledávané termíny jako pojmy lze charakterizovat sloganem „Hledat to, co je myšleno, nikoliv to, co je řečeno“. Dle Sklenáka je problémem běžného přístupu vyhledávání podle klíčových slov (znakové řetězce), nekonečná bohatost přirozeného jazyka. Právě díky této metodě jsou nalézány také dokumenty neobsahující klíčová slova, ale přesto jsou pro uživatele relevantní k danému tématu. Oblast pojmového vyhledávání reprezentují také systémy Oingo (http://www.oingo.com) a Simpli.com (http://www.simpli.com) uvedené do provozu koncem roku 1999.

Automatizované vyhledávací systémy používají vyhledávání na úrovni formální relevance. U těchto systémů je tedy potřebný další výzkum a také rozvoj technologií, kte-

ré by umožnily pokročilé a efektivní formy sémantické analýzy nebo umělé inteligence. To by například umožnilo vyhledávacím mechanismu pochopit souvislost mezi slovy „déšť“ a „pršet“. Jako příklad systému, který využívá sémantickou analýzu, lze uvést systém UCREL Semantic Analysis Systém (USAS). USAS představuje systém pro automatickou sémantickou analýzu textu (anglického jazyka. Podrobnosti a další informace naleznete v [73]. Zajímavým příkladem je také vícejazyčný systém agregace informací (M-CAST, Multilingual Content Aggregation System, viz http://www.m-cast.infovide.pl/cesky/index.html). Systém M-CAST je založen na výsledcích projektu TRUST - Multilingual Semantic and Cognitive Search Engine for Text Retrieval Using Semantic Technologies (Vícejazyčný. Vyhledávací stroj TRUST umí v současnosti vyhledávat ve čtyřech jazycích (francouzštině, italštině, polštině a portugalštině). Přidána byla také angličtina a čeština.

Jak uvádí Balíková [74], M-CAST umožňuje dotazování v přirozeném jazyce a odpovědi jsou ve formě přímé odpovědi nebo také úryvku odpovědi. Uživatel může zadat dotaz češtině například a výsledek může obdržet jak v češtině nebo v jiném jazyce implementovaným v tomto systému.

Projekt M-CAST je financován Evropskou unií v rámci víceletého komunitárního programu na podporu rozvoje a užívání evropského digitálního obsahu v globálních sítích a současně překonávání jazykové různorodosti v informační.

Grafické informační objekty

Nejčastější způsob vyhledávání grafických informačních objektů spočívá ve vytvoření textové anotace (která popisuje obsah grafického objektu) a poté vložení těchto textových informací do databáze. Samotné obrázky tak nejsou ve skutečnosti součástí databáze, ale je na ně pouze odkazováno textovým řetězcem anebo ukazatelem [25]. Tento způsob bývá označován jako „*text-based image retrieval“ [31].*

Takovýto způsob je značně neefektivní, neboť některé grafické objekty mohou obsahovat desítky nebo stovky fyzických objektů. Samotné zadávání metadat k jednotlivým grafickým informačním objektům jsou prováděny v rámci každodenní rutiny manuálně, což také znamená značně zvýšené náklady tohoto procesu. Různí lidé také mohou doplnit různé popisy (anotace) k jednomu a tomu samému obrázku. Výsledkem tak

může být minimální vypovídající hodnota metadat vztahující se k danému obrázku. V případě například ohromných databází s fotografiemi získaných z vesmírných stanic, sond a také observatoří představuje manuální zpracování nepředstavitelné finanční zatížení.

V případě vyhledávání textových informačních objektů se používá porovnávání textových řetězců dotazu uživatele a sémantického obsahu databáze. Stejný princip by se tedy měl aplikovat i pro databáze obsahující grafické objekty. Vstupem pro vyhledávání obrázků by byl také obrázek, který by bylo možné efektivně porovnat a zjistit, zda zahrnuje stejný (nebo podobný) sémantický obsah [25].

Usilovně se vyvíjejí technologie schopné hledat skutečně podle obsahu obrázků neboli technologie schopné rozpoznat elementy v obrázku a přiřadit odpovídající klíčová slova pro vyhledávání [23]**.** Tyto technologie by měly být schopné rozpoznat jednotlivé objekty (například strom, auto) na daném obrázku a poté tomuto obrázku přiřadit klíčová slova.

V této souvislosti se můžeme setkat s vyhledávání grafických informačních objektů označovaný pojmem „*content-based image retrieval*“ (*CBIR*) představující sadu technik pro získávání relevantních obrázků z databáze na základě automaticky „vytěžených“ vlastností grafického objektu [31]. Samotná funkcionalita CBIR se značně liší od vyhledávání na základě textové anotace. Charakteristiky popisující obsah obrázku jako například histogram (graf popisující kolik pixelů je na fotografii obsaženo ve škále od nejtemnější do největšího jasu, bývá ve výbavě lepších editorů a velmi dobrých fotoaparátů), barevné rozvržení, struktura, tvar nebo předmět obrázku jsou vypočítány jak pro obrázky uložené v databázi, tak také pro obrázek sloužící jako dotaz. Na základě těchto vlastností jsou vyhledány nejshodnější obrázky z databáze. Opravdu velmi obtížné je například v rámci CBIR extrahovat charakteristiky obrázku jako typ objektu zobrazený na obrázku. CBIR proto zůstává na počátku svých možností.

Jak dále uvádí Wang ve své disertační práci [31], využití technik CBIR je opravdu široké, od lékařství, vojenství, kriminalistika (otisky prstů, rozpoznávání obličeje), výzkum vesmíru, autorských práv (obchodní značka, softwarové pirátství), návrhů ve stavebnictví, klasifikaci obrázků na internetu (filtrování „nevhodných“ obrázků na internetu), zábavě (fotografie, video, filmy) i komerční sféře (módní katalogy, novinářství) atd.

Vyhledávání grafických informačních objektů se tak zaměřuje například na techniky vyhledávání podle [35]:

- textových objektů (metadat, popisných polí) přidružené ke grafickým objektům [viz 26, 27],
- vyhledávání na základě podobnosti hlavních charakteristik obrázků [viz 28, 29, 30],
- vlastností různých segmentů obrazu [viz 33],
- konceptu (náčrtu, sémantických ikon) obrázku [viz 32].

Jednotlivé možnosti vyhledávání grafických informačních objektů v databázi dle Lewa v rámci [32] znázorňuje obrázek 4.

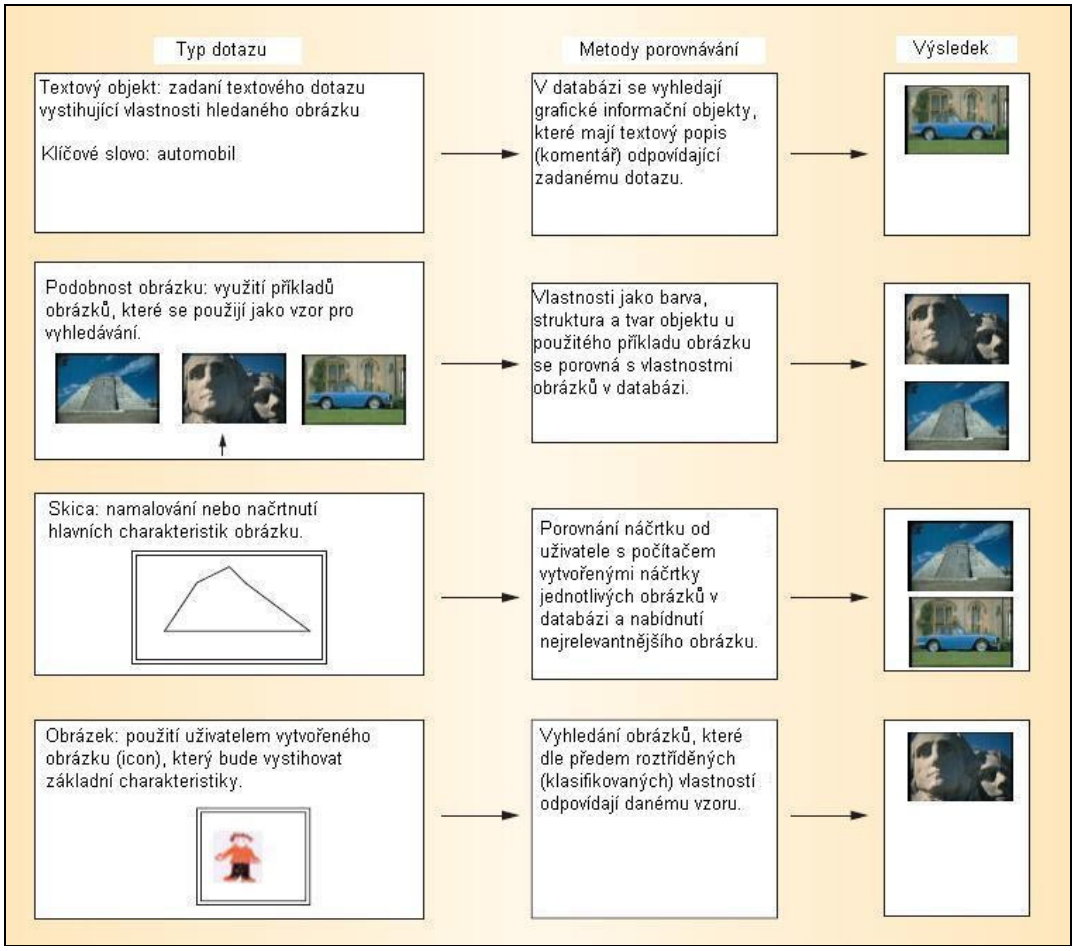
Zajímavým projektem v oblasti „*image retrieval*“ je například systém *Photobook* [viz 25], který představuje soubor interaktivních nástrojů pro procházení a vyhledávání grafických informačních objektů a sekvencí grafických objektů. Systém se soustřeďuje na vyhledávání podle sémantického obsahu obrázků, kdy jednotlivé obrázky redukuje na malý soubor koeficientů významné pro percepci. *Photoboook* používá tři typy podrobných charakteristik, na základě kterých jsou grafické objekty vyhledávány:

- vzhled (podoba)
- 2-D tvary
- textové anotace

Další zajímavým projektem v oblasti CBIR je například Tiltomoto Beta (http://www.tiltomoto.com) představující „vývojovou“ webovou stránku pro testování algoritmů pro vyhledávání obrázků na základě analyzování podobnosti a vztahů mezi obrázky. Tiltomoto Beta používá pro testování vyhledávacích algoritmů portál fotografií Flickr.com (http://www.flickr.com), ze kterých vytváří dvě oddělené databáze. První databáze obsahuje okolo 140.000 různorodých obrázků ze zmíněného portálu. Druhá databáze obsahuje obrázky „vizuálně ohromující“ neboli obsahující zajímavou vizuální stránku [44]. Tiltomoto nabízí dva odlišné způsoby vyhledávání, a to podle:

- Téma (vyhledávání obecných obrázků, kdy uživatele zajímá obecný obsah obrázku. Například budovy, příroda, zvířata atd.)

- Barevnosti / struktury (uživatel vyhledává na základě podobnosti schématu barevnosti obrázku, struktury atd.)



Obrázek 4: Znázornění technik vyhledávání grafických informačních objektů dle M.S. Lewa.

<u>Zvukové informační objekty</u>

Vyhledávání zvukových informačních objektů představuje velmi problematickou oblast. Zvukové záznamy jsou obvykle indexovány podle názvu, skladatele, předmětové klasifikace nebo hudebního žánru. Stále více ale roste potřeba po identifikaci vyhledávání zvukových informačních objektů v databázích na základě ukázky zvukového zá-

znamu nebo části písně atd. Stejně jako v případě grafických informačních objektů se touto problematikou zabývá řada výzkumných týmu v komerční i akademické sféře.

Sytém vyhledávání zvukových objektů v databázi na základě záznamu zvukové stopy pořízené mikrofonem nebo i jiných způsobem popisuje například McNab a kolektiv [36]. McNab sám předpokládá, že takovýto systém bude v blízké budoucnosti důležitou součástí úložiště digitální hudby a umožní tak analyzovat hudební záznamy a uživateli získat nahrávku na základě zvukové ukázky (vzorku), který může být často i nekvalitní nebo nepřesný. Systém zpracovává vstupní zvukový signál nezbytný pro automatické přeložení melodie pomocí technologie „optického rozpoznávání hudby“ neboli Optical Music Recognition (OMR). Analogový akustický signál je nasnímán do digitální podoby, noty jsou vysegmentovány z akustického záznamu, identifikuje se frekvence not a každá nota je označena intonací.

Experimentovalo se s různými kritérií vyhledávání a také použitými algoritmy. Je zřejmé, že vyhledávání založené na tomto základu vyžaduje pokročile technologické vybavení. Experiment mimo jiné prokázal, že takovéto vyhledávání zvukových (hudebních) objektů je v současné době nerealizovatelné.

Ale sám McNab předpovídá, že jednoho dne bude možné vyhledávat zvukové záznamy na základě zapískání určité melodie a systém poté tento vstupní „dotaz“ porovná se zvukovými objekty v databázi a vrátí odpovídající výsledky. Takovéto rozhraní pro databáze hudebních objektů budou dle McNaba tvořit základní součást digitálních knihoven budoucnosti.

Vyhledávání v audio databázích na základě zahrání melodie uživatelem, kdy zvukové objekty v dané databázi jsou vyhledány na základě shody s touto melodií zmiňuje také Ghias [37]. V této souvislosti se můžeme setkat s pojmem „Query by humming“ (QbH), což představuje způsob vyhledávání zvukových objektů, kdy se jako vstupní dotaz používá právě uživatelem zahraná melodie a porovnává se s objekty uloženým v databází. QbH tak odbíhá od klasického způsobu vyhledávání na základě názvu, jména skladatele (interpreta) nebo žánru. Vyhledáváním hudby na základě zabarvení tónu nebo rytmu tempa se zabývají další práce [viz 38, 39]. Tzanetakis [40] navrhuje alternativní prostředí, kde by bylo možné procházení a vyhledávání v dané databázi přímo na základě postupu uživatele.

Existuje mnoho standardních technik pro vyhledávání zvukových informačních objektů, které mohou být integrovány do vyhledávacího mechanismu databáze. Aucouturier [41] například zmiňuje některé techniky vyhledávání zvukových záznamů (objektů) zaměřující se na:

- textová metadat (popisná pole, doplňující informace)přidružená ke zvukovému informačnímu objektu
- podobnosti melodie zvukové ukázky uživatele a záznamu v databázi
- rytmus zvukového objektu
- zabarvení tónu zvukového objektu
- strukturální analýzu zvukového objektu

3.2.3. Prezentace výsledků vyhledávání

Součástí procesu vyhledávání je také prezentace výsledků vyhledávání odpovídající zadanému dotazu. Databáze nabízejí mnohdy opravdu uživatelsky příjemné rozhraní pro prezentování výsledků vyhledáván. Kromě prezentování vyhledaných relevantních objektů například v podobě uspořádaného seznamu jsou často uvedeny odkazy na související témata na základě klíčových slov nebo předmětových kategorií apod. Možnost vyhledávání je u některých databází až pouze po povinné autentikace uživatele. V některých případech se můžeme setkat také s platbami za zobrazení úplného záznamu objektu (v různých formátech) nebo vytištění. Podrobněji se touto problematikou u komerčních databázích zabývá [70].

3.3 Databáze textových informačních objektů

Neustále přibývá volně dostupných textových informačních objektů uložených na webech tzv. povrchového internetu nebo v databázích (repozitářích) uložených v rámci tzv. neviditelného webu. Zejména oborově zaměřené textové objekty v podobě například konferenčních materiálů, výzkumných a technických zpráv, legislativních a výukových materiálů jsou předmětem zájmu vznikajících systémů digitálních knihoven, portá-

lů a informačních bran. Jedná se o obrovský zdroj informací, jak úzce specializovaných na různé obory, tak také obecně zaměřené. Právě v tomto prostředí lze najít široké možnosti vyhledávání textových informačních objektů, využívajících pokročilých funkcí a také technologického pokroku.

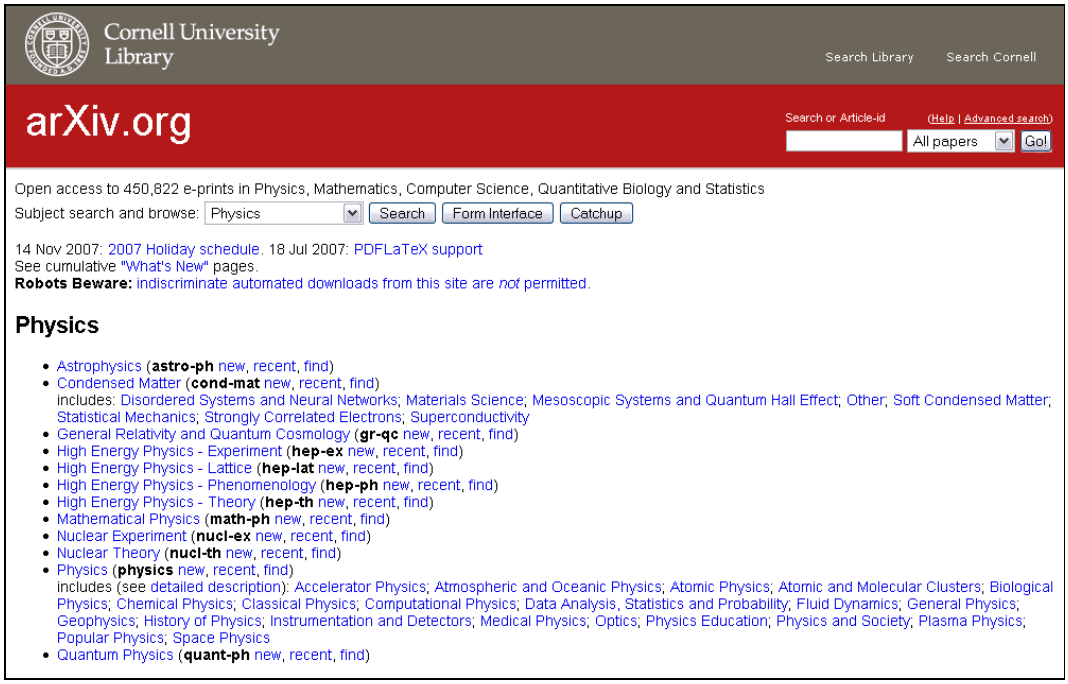
Databáze textových informačních objektů v rámci otevřeného přístupu k informacím v prostředí internetu jsou například tzv. archivy elektronických tisků. Je třeba zdůraznit pojem otevřený přístup, který zahrnuje trvalý, bezplatný online přístup k informačním objektům, zejména úplným textům. Tyto informační objekty (především textové) jsou bezplatně dostupné online, i když nemusejí být bezplatně vytvářeny. Vlastník autorských práv tak dává jednoznačné svolení ke čtení, stahování, kopírování, tištění a vyhledávání těchto textových informačních objektů, jak uvádí Bratková [45]. V této části tedy nejsou brány v úvahu komerční databázová centra skýtající neuvěřitelný zdroj informačních zdrojů, které jsou přístupné především na základě poplatků. Nyní si tedy představíme příklady databází textových informačních objektů, které autor této práce považuje za zajímavé z hlediska vyhledávání.

3.3.1. arXiv.org

Významným představitelem zmiňovaného online přístupu k informacím ve vědě je mezinárodní archiv (repozitář) arXiv.org (http://arxiv.org). ArXiv.org obsahuje textové informační objekty (plné texty) elektronických preprintů vědecko-výzkumných a technických zpráv, disertací apod. Preprint představuje prvních návrh textu článku před formálním publikováním neboli text článku, který neprošel klasickou vydavatelskou cestou. Tento repozitář byl založen jako automatizovaný sytém pro ukládání a distribuci informačních objektů v roce 1991 v rámci Národní laboratoře v Los Alamos. Nyní je vlastněn a provozován Cornell University (http://www.cornell.edu). ArXiv.org zahrnuje pět vědních oblastí – fyziku, matematiku, nelineární vědy, počítačovou vědu a kvantitativní biologii. Úvodní obrazovka tohoto repozitáře prezentuje obrázek 5.

Pokročilé vyhledávání nabízí specifikace vědní oblasti, roku, jména autora, názvu informačního objektu, slov uvedených v abstraktu, komentáři nebo v plném textu. Dále je možné vyhledávat podle předmětové klasifikace, identifikátoru nebo čísla zprávy.

V experimentálním provozu je také plnotextové vyhledávání s možností specifikovaní určitých vědních kategorií (popřípadě ve všech najednou). ArXiv.org nabízí využití booleovských operátorů AND, OR a AND NOT. Dále je využíváno zástupných znaků, převádění na základní tvary slov, dělení slov.



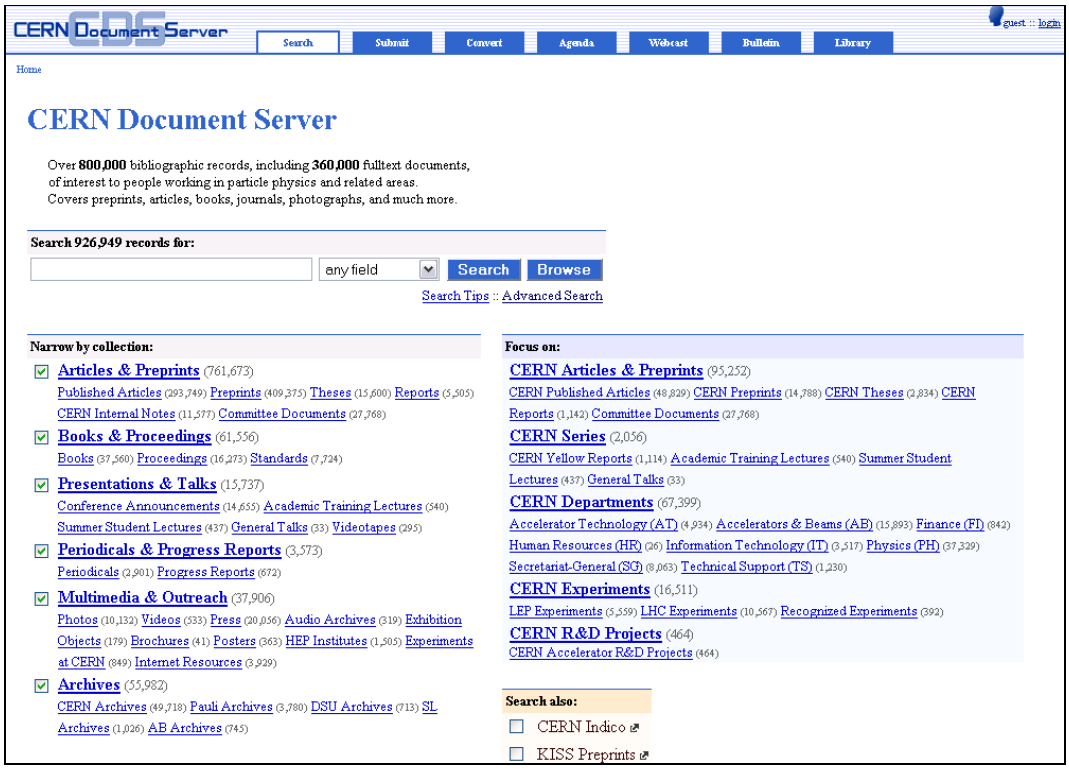
Obrázek 5: Úvodní strana archivu elektronických preprintů arXiv.org

3.3.2. CERN Document Server

Zajímavým systémem je také CERN Document server (http://cdsweb.cern.ch), který nabízí přes 360.000 plnotextových informačních objektů z oblasti fyziky a příbuzných oborů. Jedná se o textové informační objekty v podobě preprintů, článků z časopisů, knih, disertačních a jiných vysokoškolských kvalifikačních prací atd. Provozován je Evropskou organizací pro nukleární výzkum (European Organization for Nuclear Research, CERN), která byla založena v roce 1954 a sídlí v blízkosti Ženevy. Úvodní strana CERN Document Server s různými kategoriemi je znázorněna na obrázku 6.

Vyhledávání informačních objektů v rámci tohoto systému je nabízené v jednoduchém rozhraní zadáním hledaného slova nebo termínu do dotazu. Rozhraní pokročilé-

ho vyhledávání nabízí zadání jména autora, názvu objektu, čísla zprávy nebo roku. U všech uvedených možností lze specifikovat, zda vyhledávat všechny zadané výrazy najednou, nebo vyhledávat jednotlivě. Databáze CERN Document Server obsahuje také mimo jiné kategorii Multimedia, kde jsou obsaženy grafické informační objekty (obrázky).



Obrázek 6: Domovská strana systému CERN Document Server

Zajímavostí je možnost zvolit nejen vyhledávání podle přesného výrazu, částečné fráze, ale také podle regulárního výrazu (regular expression), zkráceně regexp, regex nebo jen RE. Regulární výraz představuje řetězec popisující celou množinu řetězců, konkrétně regulární jazyk. Regulární výraz se skládá z posloupnosti znaku nebo symbolů v textu, které se mají shodovat a speciálních znaků, sloužících pro popis alternativ, množin, počtů výskytů a přepínačů, které nejsou součástí hledaného textu [71]. Zřejmě nejrozšířenější normu regulárních výrazů poskytuje jazyk *Perl*. Vyhledávání na základě regulárního jazyka tak poskytuje možnost vyhledat například informační objekty podle jména autora (například John Smith), které může být vyjádřeno jak v podobě Smith, J. nebo Smith, John. Pro toto vyhledání se poté použije tato syntaxe:

```
author:/^Smith, (J|John)$/
```

Vyhledávat lze také prostřednictvím vymezení rozmezí, které lze použít jak pro vyjádření časového intervalu (vyhledat informační objekty vydané mezi roky 1980 a 1990), ale také pro jména autorů odpovídající stanovenému rozmezí. Pro toto vyhledávání se využívá znak „->“. Příkladem může být vyhledání informačních objektů podle jména autora, které se vyskytuje mezi těmito jmény: Smith, J a Smith, Oswald. Syntaxe tohoto dotazu vypadá následovně:

```
author:"Smith, J"->"Smith, Oswald"
```

3.3.3. E-LIS

K největším mezinárodním archivům v oblasti informační a knihovní vědy patří například archiv Elektronické tisky v knihovní a informační vědě E-LIS (Eprints in Library and Information Science, http://eprints.rclis.org/). Informační objekty v tomto archivu zahrnují články z novin a magazínů, knihy, konferenční sborníky, technické a výzkumné zprávy, instruktážní materiály, vysokoškolské kvalifikační práce atd. Pro zajištění kvalitního plnotextového vyhledávání jsou preferovány textové informační objekty zejména ve formátu PDF, HTML nebo TXT.

Autoři mohou ukládat textové informační objekty, ale jsou povinni doplnit požadované metadata dle webového formuláře. Každý typ informačního objektu má jinou strukturu metadat. Základními formálními prvky metadat v rámci archivu E-LIS patří název informačního objektu, jména autorů, editoři, publikační stav, veřejná doména, nakladatel nebo vydavatel, země původu informačního objektu, použitý jazyk, komentář aj. Věcná metadata zahrnují abstrakt, předmětové termíny dle klasifikace JITA (viz další odstavec) a klíčová slova. Metadata systému E-LIS jsou automaticky využívána také v dalších informačních systémech a služeb, což umožňuje také vyhledání informačních objektů v této databázi také i „zvenčí“. Obrázek 7 prezentuje úvodní webovou stránku mezinárodního archivu E-LIS.

E-LIS nabízí uživatelům vyhledávání textových informačních objektů prostřednictvím prohlížení nebo přímého vyhledávání. Klasifikační systém JITA pro obor knihovní

a informační vědy představuje základní systém pro věcné pořádání informačních objektů a také vyhledávání. Klasifikace JITA zahrnuje 12 hlavních předmětových kategorií označených velkými písmeny anglické abecedy A – L. Kategorie A a B zahrnují obecné a teoretické otázky oboru. Kategorie C – G jsou zařazovány do střední úrovně, kdy zahrnují otázky uživatelů informací, otázky publikování, otázky managementu informací atd. Specifickou úroveň prezentuje oblast H – L, které zahrnuje informační zdroje, informační technologie apod. Základní kategorie jsou na druhé úrovni děleny do dalších 135 podkategorií.

Obrázek 7: Online přístup k archivu E-LIS

Uživatel může využít vyhledávání prostřednictvím funkce prohlížení (browse), která nabízí šest rejstříků:

- Konference (Conference)
- Knihy a časopisy (Book or Journal)
- Autoři a editoři (Author or Editor)
- Předmětové termíny (Subject)
- Země (Country)
- Rok vytvoření (Year)

Využitím pokročilých možností (Advanced Search) může uživatel vyhledávat podle slov v názvu, jména autora, slov vyskytujících se v plném textu informačního objektu nebo pouze abstraktu, klíčových slov, předmětových kategorií, typu dokumentu, jména konference, jmen editorů, jazyka, citací, publikačního statusu, názvu zdrojového informačního objektu nebo roku zveřejnění.

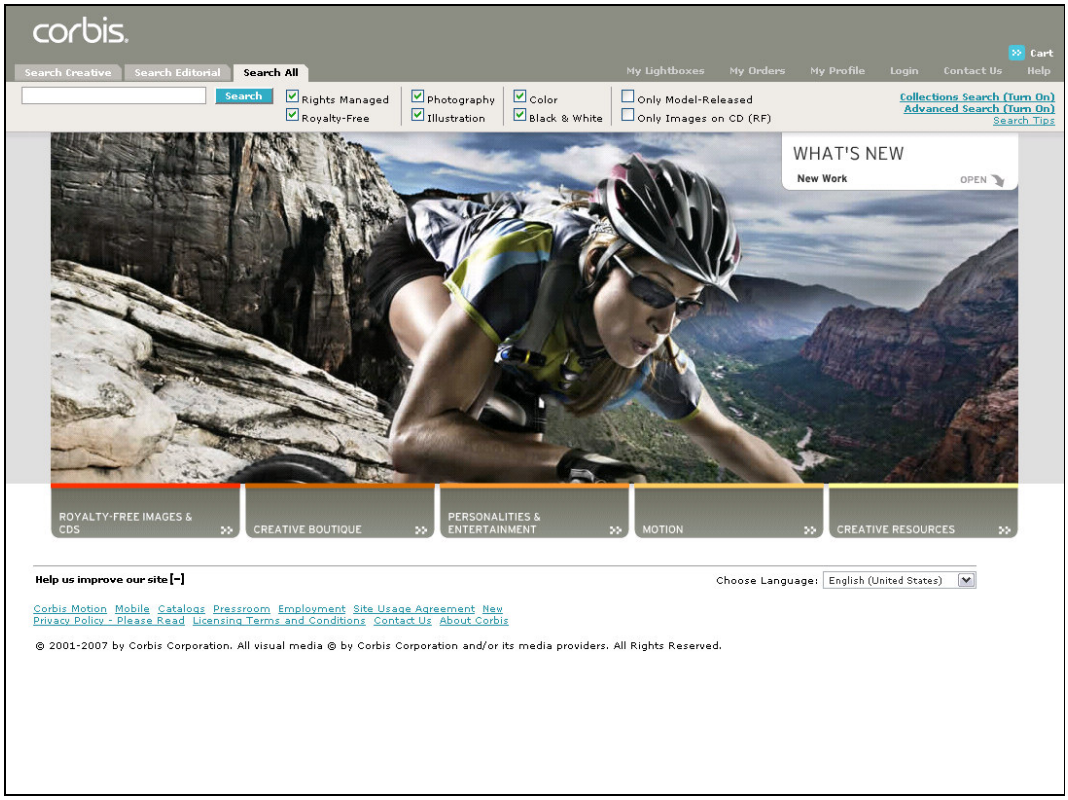
3.4. Databáze grafických informačních objektů

3.4.1. Corbis

Jako příklad databáze grafických informačních objektů si můžeme uvést databázi největší světové agentury obhospodařující autorská práva k dílům významným fotografům - Corbis (http://www.corbis.com). Online přístup k databázi Corbis je znázorněn na obrázku 8.

Vyhledávání grafických informačních objektů v rámci agentury Corbis je opravdu velmi rozmanité. Od základních možností jako název obrázku, jméno autora (fotografa), datum vytvoření, místo vytvoření až podle zařazení obrázků do kategorie obrázků nebo na základě autorských práv. Vyhledávací rozhraní databáze Corbis je zobrazeno v rámci obrázku 9. Jednotlivé obrázky v rámci této databáze jsou opravdu důkladně popsány doplněny odpovídajícími metadaty, což poskytuje hodně možností pro vyhledávání a také nalezení opravdu relevantních informačních objektů. Například vyhledávání obrázků měst pořízené z leteckého pohledu, fotografické snímky pořízené z vesmíru nebo dokonce podle počtu lidí na snímku je opravdu velmi užitečné. Jednotlivé grafické informační objekty jsou zařazovány také do kategorií jako umění, obchod, historie, zpravodajství nebo sport. Uživatel tak má možnost si požadovaný informační objekt vyhledat i touto cestou.

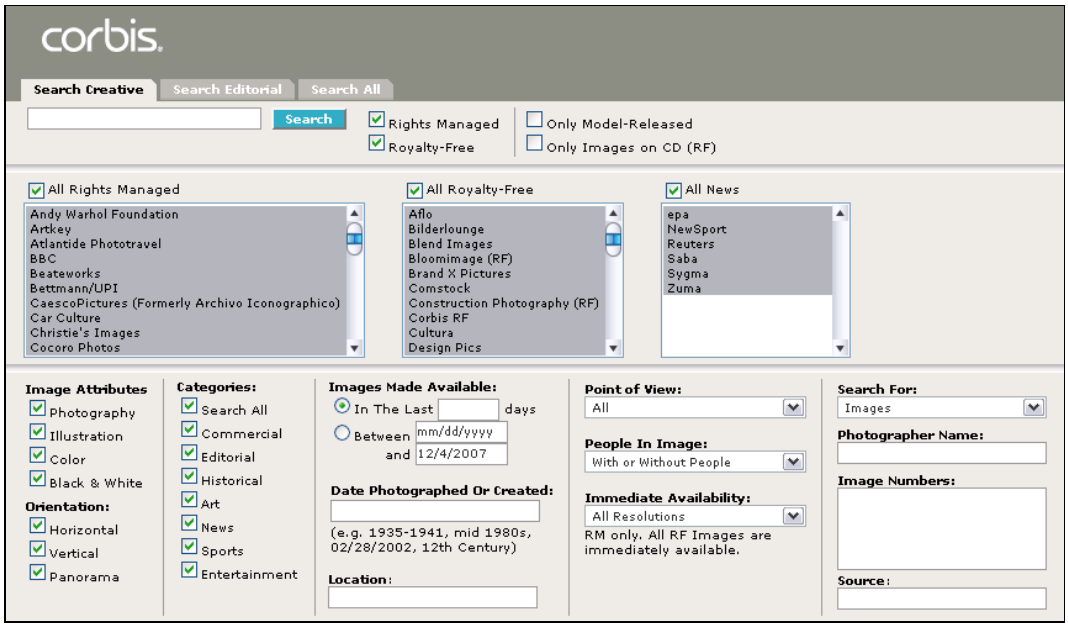
Zařazování informačních objektů do odpovídajících kategorií a také doplnění detailních popisků k těmto objektům nabízí docela efektivní vyhledávání s relevantními výsledky. Pokud zkusíme vyhledat grafické informační objekty podle klíčového slova `star` (hvězda) obdržíme grafické informační objekty znázorňující předměty ve tvaru hvězdy, obrázky obsahující slova „star“ v názvu, komentáři či dalších přidružených textových objektech nebo také známo osobnosti.



Obrázek 8: Domovská stránka pro online přístup k databázi Corbis

Pokud ale specifikujeme objekty pořízené z detailního záběru a zároveň, že musí obsahovat lidské bytosti, obdržíme jiný výsledek vyhledávání. Převážně budou zastoupeny snímky celebrit nebo objekty zobrazující předměty ve tvaru hvězd. Stačí ale změnit detaily vyhledávání na grafické objekty pořízené z vesmíru a výsledkem vyhledávání jsou fotografie nebo ilustrace vesmírných planet. Výsledky dotazu lze dále zúžit specifikováním kategorie – obchod, vydavatelství, umění aj.

Podobných databází jako Corbis.com je celá řada. Dalšími příklady můžou být Clipart.com (http://www.clipart.com), Getty Images (http://www.gettyone.com) nebo Imagestate (http://www.imagestte.com).



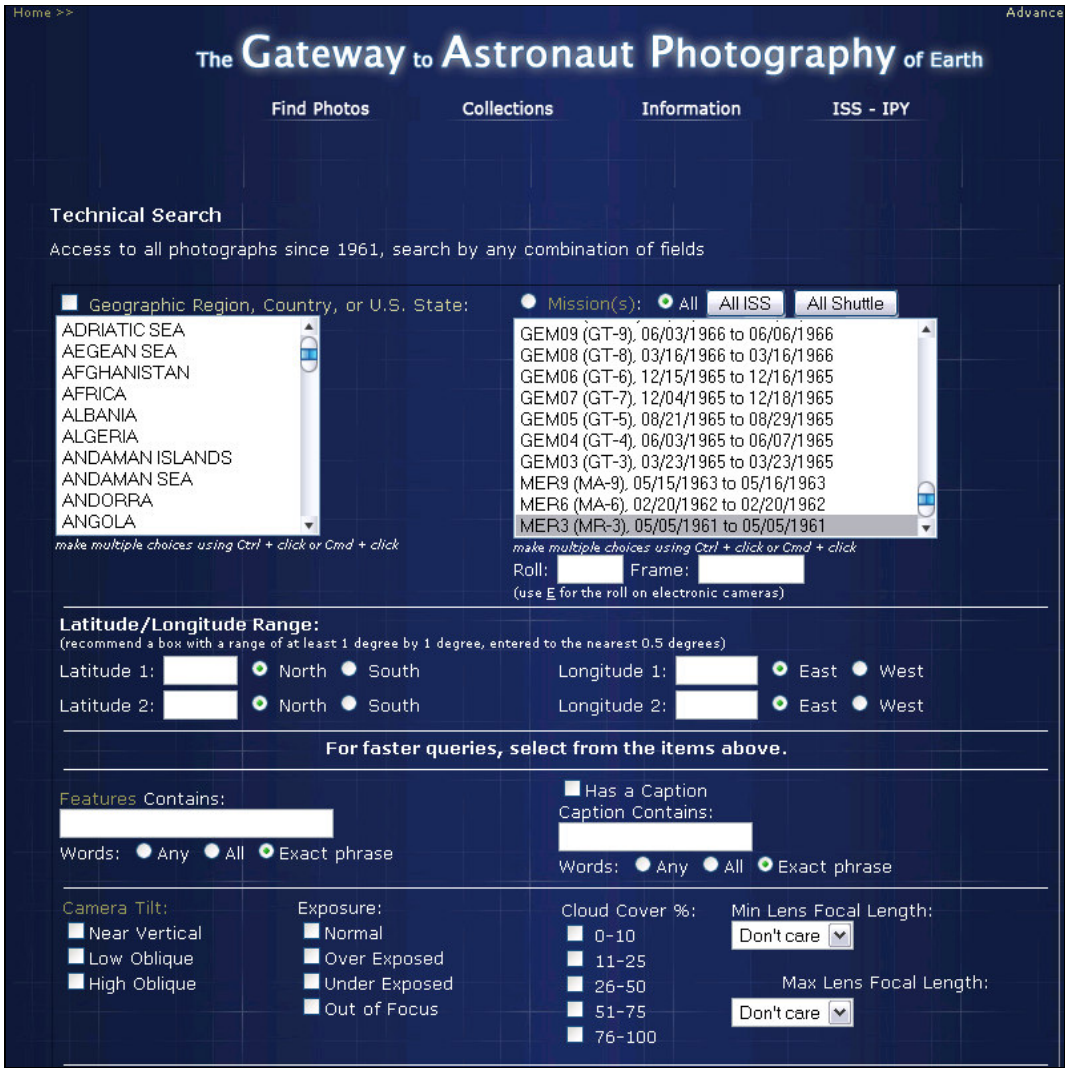
Obrázek 9: Vyhledávací rozhraní databáze Corbis.com

3.4.2. Databáze JSC NASA

Mnoho databází volně dostupných prostřednictvím internetu jsou úzce zaměřeny. Databáze s úzkým zaměřením většinou nabízejí detailní popis grafických informačních objektů, což umožňuje uživatelům opravdu detailní vyhledávání. Například se může jednat o databáze fotografií pořízených z vesmíru od vesmírných výprav, družicemi nebo od různých sond zkoumající tajemný vesmír.

Jednou z takových databází je například databáze budované vědeckou laboratoří Johnsonova vesmírného střediska v rámci NASA (National Aeronautics and Space Administration, http://eol.jsc.nasa.gov) obsahující fotografie Země.

Opravdu zdařile zpracovaná databáze grafických informačních objektů – fotografií Země nabízí uživatelům velmi mnoho způsobů vyhledávání. Uživatelé mohou využít interaktivní mapu, kde se zvolí oblast na naší planetě, která je zajímá. Další možností je vyhledávat pouze fotografické snímky pořízené na určité vesmírné výpravě.



Obrázek 10: Ukázka vyhledávacího rozhraní databáze fotografií Země v rámci JSC NASA

Databáze obsahuje přístup k fotografiím pořízené z výprav od roku 1961. Dále je možné využít tzv. technické vyhledávání (technical search), kdy uživatel specifikuje jednotlivé parametry jako například geografickou oblast, zeměpisnou délku a šířku, název fotografie nebo popis obsahující dané slovo, datum pořízení snímku, barevnost snímku nebo dokonce procentuální úroveň mračen na daném snímku. Ukázka vyhledávacího rozhraní pro technické vyhledávání je prezentována na obrázku 10.

Podobných databází existuje opravdu velké množství. Profesionálně budované databáze nabízí opravdu vysokou úroveň zařazování informačních objektů do odpovídajících kategorií a také doplňování podrobných metadat.

3.5. Databáze zvukových informačních objektů

Zvukové informační objekty jsou v prostředí internetu zastoupeny v ohromném množstvím, a to především v podobě hudebních souborů ve formátu MP3. Běžný uživatel se například na internetu může setkat s nepřeberným množstvím volně dostupných databází nabízející široký výběr hudby nabízených jak zdarma, tak i za poplatek.

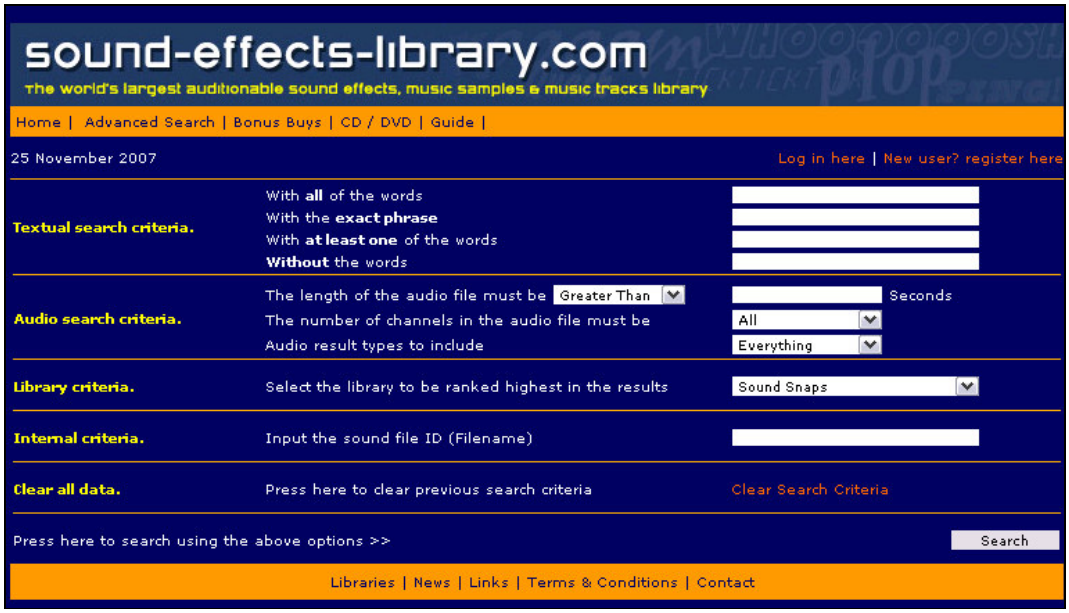
V současné době existují velmi rozšířené internetové obchodní domy, které nabízí na komerční bázi k vyhledávání databáze vyhledávání zaměřené pouze podle metadat k reálným zvukovým informačním objektům. Uživatel si tak může vyhledat potřebný produkt ve fyzické podobě CD, DVD nebo audio kazety. Samotné zvukové informační objekty zde mohou být zastoupeny pouze ukázkou ke zvolenému produktu. Nejznámějším představitelem databáze, který nabízí mimo jiné vyhledávání převážně indexovaných zvukových informačních objektů ve své databázi je Amazon.com (http://www.amazon.com). Oddělení zabývající se hudebními objekty nabízí pokročilé vyhledávání na základě metadat připojených k jednotlivým zvukovým informačním objektům. Vyhledávat je tak možné například podle klíčových slov, jména autora, názvu, fyzického nosiče (CD, audio DVD atd.). Je zřejmé, že vyhledávání v rámci podobných internetových obchodních domů nabízí pouze klasické vyhledávání pro nalezení požadovaných zvukových informačních objektů. Tyto komerční domy se snaží nabízet své produkty a oslovit co nejvíce zákazníků. Podobně fungují například MP3.com (http://www.mp3.com) nebo Allmusic.cz (http://www.allmusic.cz).

Nyní si představíme některé zajímavé příklady na internetu volně dostupných databází zvukových informačních objektů, které jsou uložené přímo v dané databázi a uživatel si může vyhledat a přímo poslechnout zvukové ukázky.

3.5.1. The Sound Effects Library

Vyhledávání v databázi obsahující více než 250.000 zvukových informačních objektů v podobě zvukových efektů, hudebních ukázek jak volně pro další použití, tak také dostupné za určitý poplatek nabízí The Sound Effects Library (Kolekce zvukových efektů, http://www.sound-effects-library.com). Zvukové informační objekty jsou využívány

ze zhruba 90 kolekcí zvukových objektů od různých společnosti, filmových studií, skladatelů apod.



Obrázek 11: Ukázka rozhraní pokročilého vyhledávání sbírky zvukových efektů Sound-effects-library.com

Jak je patrné z obrázku 11, vyhledávací rozhraní umožňuje vyhledávat podle různých kritérií:

- textových (Textual search criteria) – klasické vyhledávání na základě zadaných slov, kombinace více slov, přesné fráze nebo ještě možnosti vyloučení některých slov. Textové vyhledávání je použito vždy. Pouze se nabízí kombinace s dalšími kritérii.

- zvukových (Audio search critera) - tato možnost nabízí vyhledávání podle délky zvukového objektu (v sekundách), podle kvality (stereo, mono) nebo podle zvukových informačních objektů zahrnující pouze zvukové efekty, hudební nahrávku nebo oboje.

- sbírky (Library critera) – uživatel si může vybrat ze které sbírky zvukových informačních objektů se bude vyhledávat.

- interních (Internal criteria) – vyhledávání podle označení zvukového objektu.

Vyhledané zvukové objekty si může uživatel ihned přehrát a zjistit, zda opravdu vyhovují jeho požadavku.

3.5.2. The Freesound Project

Zajímavým a poněkud odlišným projektem zabývajícím se databází zvukových objektů je například The Freesound Project (http://freesound.iua.upf.edu/index.php, dále jen Freesound). Jedná se o spolupracující databázi licencovaných zvuků v rámci Creative Commons (http://creativecommons.org), přičemž Freesound se zaměřuje pouze na zvukové úryvky, ukázky, nahrávky apod. Veškeré tyto zvukové objekty jsou vydávány v rámci licence Creative Commons Sampling Plus 1.0 License (http://creativecommons.org/licenses/sampling+/1.0/). Samotní uživatelé (fyzické osoby nebo různé instituce) tak mohou ukládat, stahovat a využívat zvukové informační objekty.

Zvukové objekty jsou v různých formátech (například MP3, WAV atd.) a doplněny jsou metadaty popisující kvalitu zvukového objektu, délku ukázky, velikost digitálního objektů (v MB) a také doplněny klíčovými slovy (označované jako tag). Tyto klíčová slova přidávají jednotlivý uživatele, a proto ne vždy mohou být relevantní a jsou zastoupeny u jednotlivých zvukových objektů v různé míře. Freesound poskytuje nový a zajímavý způsob zpřístupňování těchto ukázek zvukových objektů a nabízí jednotlivým uživatelům:

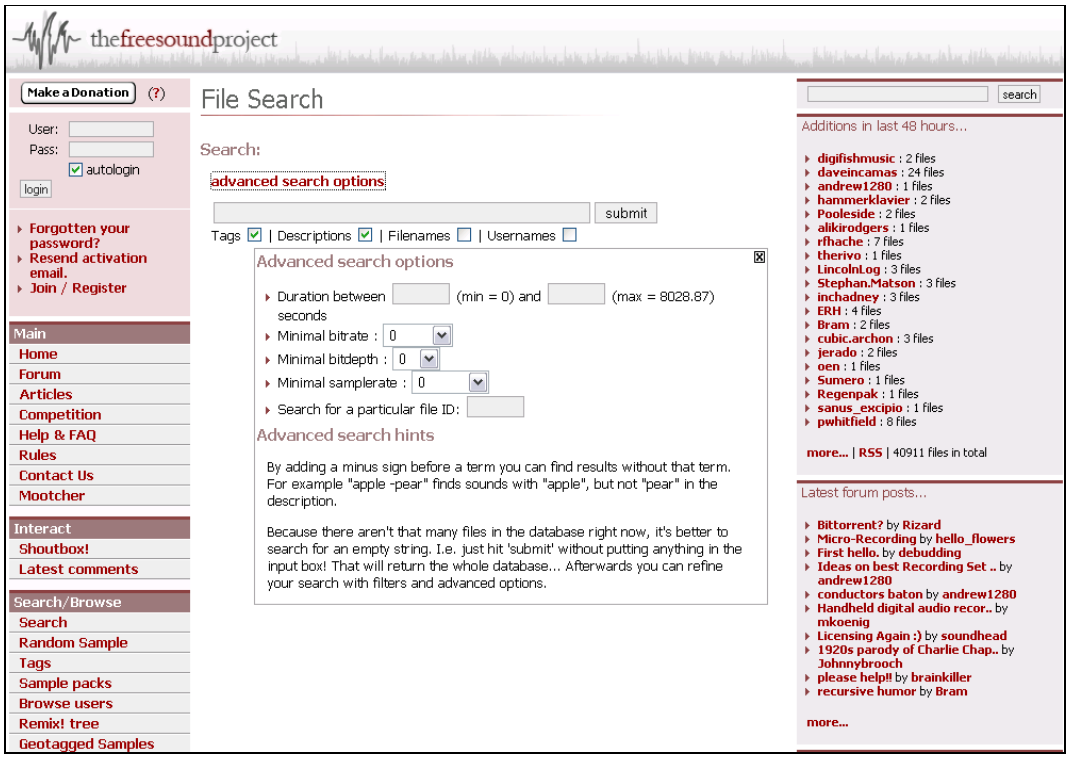
- Vyhledávat zvukové objekty na základě dotazu v textové podobě nebo specifikováním kvality těchto zvukových objektů.

- Listování zvukovými objekty použitím klíčových slov, podobnosti ukázek, podle uživatelů atd.

- Ukládání a stahování zvukových informačních objektů jednotlivými uživateli.

Cílem projektu Freesound je také vybudování volně dostupné databáze zvukových informačních objektů, jenž pak může být dále použita pro vědecký výzkum [47]. Mnoho výzkumných institucí narazilo na problémy s databází korektně licencovaných zvuko-

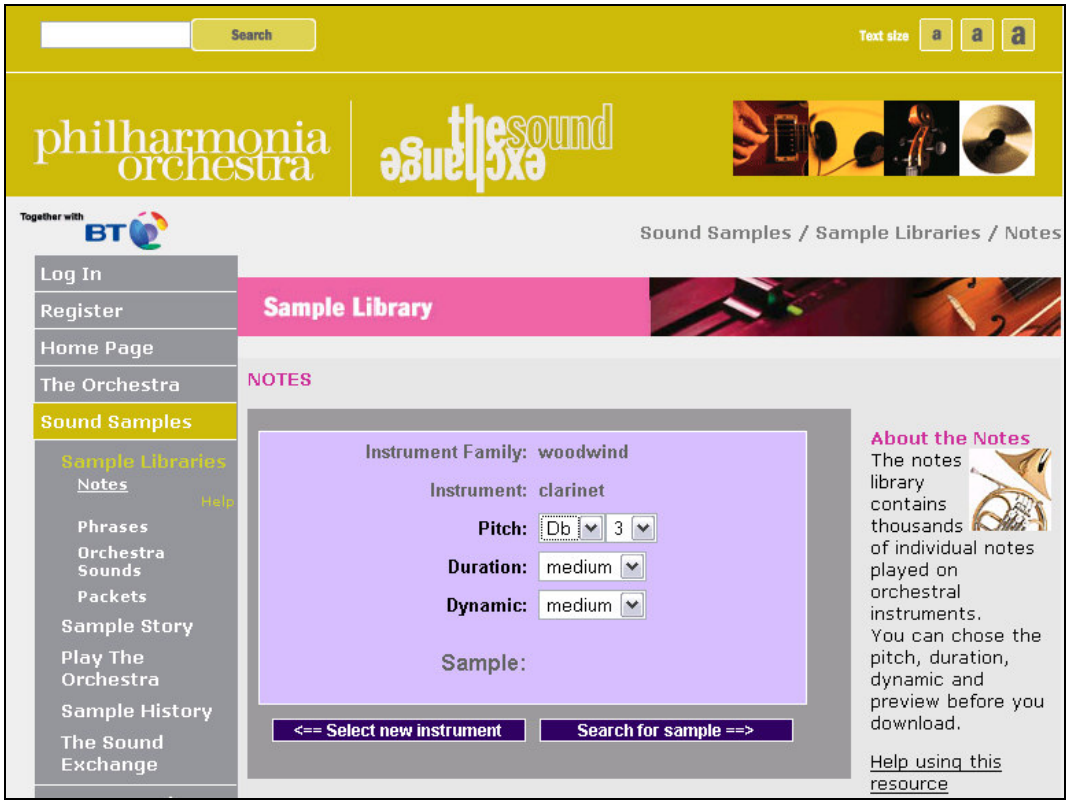
vých informačních objektů pro testování svých algoritmů. Vyhledávácí rozhraní databáze The Freesound Project s pokročilými možnostmi je zobrazené na obrázku 12.



Obrázek 12: Ukázka rozhraní pokročilého vyhledávání databáze projektu Freesound

3.5.3. The Philharmonia Sample Library

Databáze zaznamenaných zvukových objektů v podobě hudebních tónů některých hudebních nástrojů je dostupná v rámci sbírky ukázek zvukových objektů Filharmonické orchestru Velké Británie. Tato sbírka ukázek zvukových objektů je volně dostupná (http://www.philharmonia.co.uk/thesoundexchange/sound_samples/) a obsahuje například tisíce jednotlivých hudebních tónů zahrané na různé hudební nástroje. Uživatel si tak může vyhledat a poslechnout potřebný zvukový objekt na základě specifikování hudebního nástroje (smyčcové, dřevěné dechové nebo dechové nástroje), tónu (pitch), tempa a doby trvání. Vyhledávací rozhraní této databáze ukazuje obrázek 13. Vyhledaný zvukový objekt si poté uživatel může stáhnout a dále sdílet s ostatními, použít pro další hudební kompozici. Ale je striktně zakázáno tyto ukázky zvukových objektů dále prodávat.



Obrázek 13: Ukázka vyhledávacího rozhraní
podle zaznamenaných tónů hudebních nástrojů

4. Analýza vyhledávacích možností u vybraných příkladů databází

V předchozí kapitole byly představeny dle autora zajímavé příklady databází různých typů informačních objektů (textové, zvukové a grafické) s vyhledávacím rozhraním dostupným online. Uvedené příklady databází nemusí představovat nejznámější databáze informačních objektů. Samozřejmě se také nejedná o úplný výčet databází, což také nebylo ani autorovým úmyslem. Přesto lze z uvedených příkladů databází zaměřující se na různé informační objekty vypozorovat určité rozdíly v možnostech vyhledávání informačních objektů. Tato část práce se podrobněji zaměřuje na možnosti vyhledávání u vybraných příkladů databází obsahující zvolené informační objekty a také nastínění charakteristických rozdílů.

4.1. Postup analýzy

Při analýze vyhledávání se zaměříme na možnosti, které nabízí zmíněné příklady databází jednotlivých typů informačních objektů. V případě vyhledávání různých typů informačních objektů lze nalézt společné prvky jako například jméno autora (fotografa, interpreta, skladatele), název objektu (určený autorem), textové anotace (míněny zejména doplňující údaje v podobě abstraktu, doplňujících údajů u obrazových nebo zvukových informačních objektů atd.), zařazení do kategorie (stanovené kategorie mohou být v různých databázích odlišné). Dále se můžeme setkat se specifickým způsobem vyhledávání vlastní danému typu informačního objektu (plnotextové vyhledávání, vyhledávání dle tónu, vyhledávání grafických informačních objektů dle způsobu zachycení předmětu atd.).

Určitá základní metadata popisující daný informační objekt lze při vyhledávání využívat jak u textových, tak i zvukových nebo grafických informačních objektů. Možnosti vyhledávání dané charakterem informačních objektů budou představeny v rámci specifik dané typem informačního objektu. Při analýze budou tedy u jednotlivých příkladů databází porovnány možnosti vyhledávání informačních objektů, a to dle metadat (bod 1 – 4) a také vyhledávání dané charakterem informačního objektu (bod 5):

1. Jméno autora (fotografa, interpreta skladby)
2. Název objektu
3. Textové anotace
4. Zvolená kategorie v rámci databáze
5. Specifické možnosti vyhledávání

Je předem zřejmé, že uživatelé nejčastěji používají vyhledávání informačních objektů zadáním textového řetězce v podobě slov nebo kombinace více slov. Tento způsob se používá u zmíněných typů informačního objektu. Popisné údaje mohou představovat termíny nebo slovní spojení, které jsou obsažená jak v názvu objektu nebo jménu autora.

4.2. Vyhledávání dle jména autora

Naprostou samozřejmostí u textových informačních objektů je možnost vyhledávání na základě jména autora. Všechny uvedené příklady databází (repozitářů) textových informačních objektů - arXiv, CERN Document Server i E-LIS – nabízí prostřednictvím vyhledávacího rozhraní také možnost podrobnějšího specifikování použitím porovnání jednotlivých slov v různém pořadí nebo v přesně daném pořadí (přesné fráze), booleovských operátorů (případně proximitních či numerických).

Na velmi obecný dotaz v podobě jména autora `John` u arXiv.org se nám vrátí ohromné množství nalezených výsledku, kdy je zobrazeno pouze prvních 1.000 nalezených. Obdobně je tomu i u dalších příkladů repozitářů (databází). CERN Document Server nám ale ještě při volbě přesné fráze dotazu u jména autora `John` nabídne možnost zvolit z indexu autorů, pokud tvar zadaného jména autora není nalezen v indexu autorů. Tímto zužováním dotazu se dostáváme k daleko přesnější množině možných výsledků. Obměnou tvaru dotazu na `JOHN` dostáváme opět stejné výsledky, kdy se vyhledají objekty se jménem autora bez ohledu na velká či malá písmena. CERN Document Server nabízí také užší vymezení dotazu použitím regulárních výrazů.

Stejně obecné vyhledávání podle jména autora lze použít i pro grafické informační objekty. Corbis přímo nabízí vyhledávání na základě jména autora (neboli fotografa). Výsledkem zadání dotazu v podobě obecného jména fotografa `John` obdržíme veške-

ré grafické informační objekty od fotografa, v jehož jméně se vyskytuje zadané slovo. Corbis u vyhledávání dle jména fotografa také nabízí možnost dalšího specifikování na základě booleovských operátorů, použití zástupných znaků nebo vyhledávání přesné fráze. Databáze JSC NASA přímo nenabízí vyhledávání dle jména autora ze samotné podstaty pořízených fotografií.

Zvukové informační objekty v rámci databáze Sound Effects Library přímo nenabízí vyhledávání dle jména autora, ale nabízí možnost vyhledání slov či slovního spojení, přesné fráze nebo výčtem slov anebo vyloučením nežádoucích slovních výrazů. Dotaz v podobě slova `John` tak přesto vyhledá zvukové objekty, jejichž autor má uvedené jméno John. Vyhledané výsledky nám ale může vrátí také zvukové objekty, kde se slovo John vyskytuje i v samotném názvu objektu. The Freesound Project nabízí jednoduché vyhledávání na základě jména (Username), pod kterým jednotlivý uživatelé na této databázi vystupují a ukládají své zaznamenané zvukové objekty. Databáze ukázek zvukových objektů Filharmonického orchestru (The Philharmonia Sample Library) vyhledávání dle jména autora ze samotné podstaty nenabízí, neboť veškeré zvukové objekty jsou zaznamená právě Filharmonickým orchestrem.

4.3. Vyhledávání dle názvu

Vyhledávání dle názvu objektu funguje u textových informačních objektů na stejném principu jako u jména autora. Například zadáním dotazu `animal` obdržíme veškeré textové informační objekty, v jejichž názvu se vyskytuje slovo odpovídající dotazu. Pro další upřesnění dotazu lze opět použít booleovské operátory, přesnou frázi apod. CERN Document Server při zadání obecného dotazu pro název objektu nabídne možnost výběru dle indexu názvu objektů, které se nejvíc podobají zadanému dotazu.

Možnost vyhledávání grafických objektů na základě názvu v rámci databáze Corbis není přímo nabízena. Fotografie jsou do této databáze zařazovány s určitým názvem, který lze také považovat za popis daného obrázku. Název objektů v rámci této databáze představuje stručnou charakteristiku obsahu daného objektu. Proto záleží na daném odpovědném pracovníkovi a jeho uvážení, jaký stručný popis zvolí pro daný obrázek. Vyhledávání grafických informačních objektů obsahující v názvu například pojem `animal` si systém „přeloží“ jako vyhledávání grafických objektů na základě klíčového slo-

va apod. Databáze JSC NASA přímo nenabízí možnost vyhledávání dle názvu objektu. Uživatel může využít vyhledávání dle záhlaví (caption), které ovšem nemusí být uvedeno u všech obrázků.

Sound Effects Library přímo nenabízí vyhledávání zvukových objektů na základě názvu. Uživatel tak opět vyhledává daný objekt bez ohledu na to, zda se jedná o název, jméno autora apod. Opět tak záleží na pojmenování zvukových objektu v dané databázi. Databáze The Freesound nabízí sice vyhledávání dle názvu objektu (zaškrtnutím pouze možnost „Filename“), ale vyhledávání přesto probíhá i v rámci popisků jednotlivých objektů. Zde je patrné úskalí, kdy názvy zvukových objektů vybírají samotní uživatelé, kteří daný objekt ukládají v této databázi. Nejsou tak zde žádná pravidla pro pojmenovávání daných zvukových objektů a vyhledávání dle názvu by bylo v některých případech problematické. Specifická a úzce zaměřená databáze ukázek zvukových objektů Filharmonického orchestru vůbec nenabízí vyhledávání dle názvu objektu na základě uživatele uvedeného dotazu, ale pouze nabízí uživateli zvolit název tónu vybrat z připravené nabídky. Toto opět vychází ze samotné podstaty této databáze.

4.4. Vyhledávání dle textové anotace

Vyhledáváním dle textových anotací u textových informačních objektů je myšlena možnost vyhledávat podle znakových řetězců obsažených v abstraktu, komentářích, doplňujících údajích apod. Například v případě repozitáře arXiv.org je možné vyhledávat na základě slov nebo slovních spojeních obsažených v abstraktu nebo komentářích. Zadáním slova `ocean` a zvolením vyhledávání v abstraktu (abstract) nebo komentáři (comment) jsou tak zobrazeny textové informační objekty, které obsahují zadaný dotaz v komentářích nebo abstraktu. CERN Document Server nenabízí přímo specifikování vyhledávání pouze v abstraktu nebo komentáři, ale použitím vyhledávání bez specifikování (any field) jsou do vyhledávání zařazeny také komentáře nebo abstrakt informačního objektu. Vyhledávání přímov abstraktu umožňuje také E-LIS, který ještě nabízí možnost specifikování na vyhledávání všech zadaných slov v jakémkoliv pořadí anebo jednotlivých zadaných slov.

V případě obrazových informačních objektů patří vyhledávání na základě textových anotací k velmi obvyklým, neboť název obrazových informačních objektů nemusí být zcela vypovídající o samotném obsahu. Právě proto textové anotace nabízí další mož-

nost pro specifikování samotného obsahu obrazového informačního objektu. Databáze Corbis neumožňuje vyhledávat pouze přímo na základě textových anotací. Proto uživatel při zadání dotazu v podobě slova `ocean` obdrží jak obrazové informační objekty, v jejichž názvu se vyskytuje dané slovo, tak i obrazové informační objekty, u kterých je dané slovo zahrnuté v textové anotaci, komentáři nebo dalších polích popisující obsah daného obrázku. Databáze JSC NASA nabízí vyhledávání na základě přesné fráze nebo slov obsažených v charakteristikách (význačných rysů) daných obrazových objektů. Doplňující informace jsou vytvářeny odpovědnými pracovníky, na jejichž vlastním rozhodnutí do jisté míry záleží, jaké význačné rysy přiřadí jednotlivým objektům.

The Freesound projekt nabízí také vyhledávání dle doplňujících údajů v podobě textové anotace. Jednotlivé zvukové informační objekty jsou zde doplňovány charakteristikami jednotlivými uživateli, kteří mohou právě v rámci projektu The Freesound uložit určitý zvukový informační objekt. Je zde proto zřejmé, že jednotlivé doplňující údaje se budou značně lišit, neboť nevychází z předem daných pravidel nebo principů, ale záleží pouze na uvážení daných uživatelů. Sound Effects Library opět přímo nenabízí vyhledávání dle textových doplňujících údajů. Proto při zadání dotazu v podobě slova `ocean` se v této databázi vyhledají veškeré zvukové informační objekty, v jejichž názvu se zadané slovo vyskytuje. Databáze zvukových objektů v podobě jednotlivých tónů hudebních nástrojů nahrané Filharmonickým orchestrem opět nenabízí možnost vyhledávání dle doplňujících textových anotací. Toto je dáno samotným zaměřením této databáze.

4.5. Vyhledávání dle kategorie

Jednotlivé databáze zařazují informační objekty do různých kategorií či oblastí, které nabízí uživateli jak procházení daných kategorií v případě, kdy uživatel nemá přesně stanovený cíl vyhledávání anebo je v řadě případů možné právě díky daným kategoriím postupně zužovat vyhledávání. Kategorie odpovídají zejména zaměření databáze, proto se dané kategorie v uvedených příkladech databází liší.

ArXiv.org umožňuje vyhledávání pouze ve vybraných kategoriích (oblastech) jako například matematika, počítačová věda, statistika atd. V případě oblasti fyziky může uživatel zvolit přímo repozitář (archiv), ve kterém bude daný dotaz proveden (například astro-fyzika nebo kvantová fyzika). ArXiv.org dále nabízí možnost blíže specifikovat

předmětové třídy pro oblast matematiky nebo fyziky (například klasická fyzika, geofyzika, historie fyziky, chemická fyzika apod.). E-LIS používá jak pro věcné pořádání, tak i pro vyhledávání klasifikaci JITA, která obsahuje dvanáct hlavních předmětových kategorií. Uživatel tak může při vyhledávání zvolit pouze ty předmětové kategorie, ve kterých chce vyhledávat. Implicitně vyhledávání probíhá ve všech předmětových kategoriích. CERN Document server nabízí možnost specifikovat vyhledávání pouze na vybrané kolekce (collections), ve kterých bude daný dotaz proveden. Využitím možnosti „Narrow by Collection“ může uživatel přímo zužovat kategorie. Výsledky vyhledávání jsou poté přehledně rozděleny ještě dle typů (například články a preprinty, knihy, prezentace nebo výzkumné zprávy).

V rámci databáze grafických informačních objektů Corbis je možné specifikovat vyhledávání na obecné kategorie jako například obchod, historie, umění, zpravodajství, zábava. Uživatel tímto omezením obdrží požadované oblasti. V případě potřeby lze zobrazit výsledky také dalších oblastí anebo všechny výsledky bez ohledu na kategorii. Odborné vyhledávání v případě JSC NASA nenabízí přímo specifikování podle stanovených kategorií. Určité podobnosti kategorií lze spatřovat v možnosti vyhledávání například podle speciálních kolekcí – města (Cities), fotografie Země z vesmíru (Earth from Space) anebo mořské útesy (Reefs).

Sound Effect Library neumožňuje specifikovat vyhledávání zvukových informačních objektů podle kategorií. Nabízí vyhledávání pouze zvolením knihovny (Library), ve které se daný dotaz provede. Databáze The Freesound Project ani databáze Filharmonického orchestru obsahující tóny hudebních nástrojů nenabízí vyhledávání na základě kategorie.

4.6. Specifické možnosti vyhledávání

Nyní se zaměříme na možnosti vyhledávání, které jsou dány typem informačního objektů v rámci uvedených příkladů databází. Do specifických možností vyhledávání je zařazeno také vyhledávání na základě charakteristik obsahu objektu anebo vlastností, které jednotlivé databáze nabízí na základě „vstupního zpracování“ informačního objektu anebo doplněním potřebných textových informací vyjadřující tyto vlastnosti. Toto vyhledávání je dle autorova názoru vychází ze specifických vlastností jednotlivých informačních objektů.

Jako specifické lze v případě textových informačních objektů zmínit vyhledávání v úplných textech informačních objektů. V rámci tohoto vyhledávání je daný dotaz v podobě slova nebo slovního spojení vyhledáván nejen v názvech, jménech autorů, abstraktech, komentářích nebo klíčových slovech, ale vyhledávání probíhá v úplném (nezkráceném) textu informačního objektu. ArXiv.org nabízí experimentální vyhledávání v úplných textech informačních objektů, lze použít také booleovských operátorů, fráze (použitím uvozovek). Nelze specifikovat dotaz pouze na určitou část archivu nebo použít specifické vyhledávání dle názvu objektu, jména autora apod. ArXiv.org ale upozorňuje, že v rámci vyhledávání v úplných textech se jedná o experimentální službu, která může nabízet méně aktuální výsledky hledání. Vyhledané objekty jsou seřazeny na základě shody daného dotazu a objektů. Uživatel ale může znovu setřídit dané výsledky dle data. V rámci vyhledávání v úplných textech je možné v databázi CERN Document Server použít také regulární výrazy (zmíněno v rámci kapitoly 3.3.2.), která uživateli nabízí více možnosti při zadávání dotazu.

V případě vyhledávání obrázků (grafických informačních objektů) není možné použít tzv. „full-text“ vyhledávání. Databáze Corbis se snaží poskytnout uživateli více možností pro vyhledávání na základě obsahu obrázku. Toto je umožněno také díky detailnímu zpracování daných obrázků. Uživatel může při vyhledávání zvolit, zda se jedná o obrázek v podobě fotografie nebo ilustrace, zda se jedná o barevný či černobílý obrázek. Při vyhledávání je také možné zvolit orientaci obrázků – horizontální, vertikální nebo panorama. Corbis také nabízí upřesnění vyhledávání zvolením úhlu pohledu (Point of View) z předem daných možností. Uživatel může specifikovat obrázky zachycující letecký pohled (aerial), blízký záběr (close-up), pohled shora (above), pohled zespodu (below) nebo z vesmíru (From Space).

Corbis dále nabízí uživateli možnost určit, zda mají vyhledané obrázky obsahovat lidi, případně počet lidí zachycených na obrázku. Veškeré tyto možnosti vyhledávání ale závisí na úrovni „vstupního zpracování“ daných grafických informačních objektů v rámci dané databáze. Zodpovědní pracovníci musí jednotlivé obrázky zařadit do správných kategorií, opatřit odpovídajícím názvem a textovou anotací a v neposlední řadě také přiřadit klíčová slova vystihující obsah obrázku a jednotlivé objekty. V případě fotografie zachycující závodní automobil na závodní dráze spolu s diváky tak bude pravděpodobně daný obrázek opatřen mimo jiné klíčovými slovy jako například: lidé, diváci, auta, závod, automobilový závod, závodník, sport, sportovní auto, rychlost, soutěž, soutěžící, pohled shora atd. Uživatel může také vyhledávat podle kvality obráz-

ku – rozlišení. Corbis nabízí volby rozlišení: vše (All Resolutions), nízké rozlišení (Low), střední rozlišení (Medium), vysoké rozlišení (High) a velmi vysoké rozlišení (Ultra). Propracované možnosti vyhledávání jsou tedy do jisté míry také závislé na úrovni vstupního zpracování jednotlivých informačních objektů.

Také databáze JSC NASA nabízí opravdu řadu specifických možností vyhledávání. Uživatel může zvolit vyhledávání pouze u vybraných geografických oblastí nebo zemí (např. Evropa, Jaderské moře), zvolením pouze některých vesmírných výprav. Zajímavou možností je vyhledávání grafických informačních objektů na základě rozmezí zeměpisné šířky a délky zachycené na daných obrázcích. Při vyhledávání je možné specifikovat také typ použitého fotoaparátu a také detailnější klasifikace filmu (nabízené možnosti jsou například Color negative, Color infrared, Digital media apod.). Obrázky v rámci této databáze je možné také blíže specifikovat dle úhlu zachycení daných objektů na obrázku, dále dle expozice nebo také na základě procentuální úrovně pokrytí mraky na daném snímku.

Databáze Sound Effects Library nabízí v rámci vyhledávání dle audio kritérií (Audio search criteria) uživateli specifikovat dotaz zadáním délky trvání zvukových objektů (v sekundách), dále rozlišením kvality záznamu (stereo, mono). Uživatel může dále zvolit, zda chce vyhledávat pouze objekty obsahující hudební stopu nebo zvukové efekty.

The Freesound Project nabízí jak možnost stanovení určité délky zvukového záznamu, tak také rozmezí (v sekundách). Opravdu zajímavá je ale možnost vyhledávání digitálních zvukových objektů na základě minimální hodnoty přenosové rychlosti (bitrate) neboli kolik jednotek informace se přenese za jednotku času. Základní jednotkou přenosové rychlosti je bit za sekundu (bit/s). Jednotka tak udává, kolik bitů informace je přeneseno za jednu sekundu [72]. Dále je možné z nabízených možnosti zvolit minimální hodnotu bitové hloubky (bit depth) neboli počet úrovní použitelných pro digitalizaci zvukového vzorku. Vyhledávat lze také zvolením minimální hodnoty vzorkovací frekvence (sample rate).

Databáze zvukových ukázek not zahraných na hudební nástroje Filharmonického orchestru nabízí specifické vyhledávání pouze zvolením požadovaného tónu, doby trvání a dynamiky. Jiné možnosti nejsou v této databázi možné.

4.7. SWOT analýza

Prostřednictvím SWOT analýzy si můžeme uvést některé vlastnosti textových, grafických a zvukových informačních, založené také na základě poznatků uvedených v této práci.

4.7.1. Textové informační objekty

Silné stránky:

- Textové informační objekty patří k nejběžněji používaným objektům pro ukládání, zpracovávání, distribuci, vyhledávání a následné využívání informací.

- Možnost indexování a vyhledávání textových informačních objektů v rámci plných textů. Automatické indexování textů v přirozených jazycích je založeno na extrakci indexačních termínů přímo z plného textu objektu (dokumentu).

- Snadné převádění textových informačních objektů do jiných formátů a také strukturování.

- Výborné možnosti bezztrátové komprese dat u textových objektů při použití různých komprimačních metod (statistických, slovníkových atd.).

- Zobrazování textových objektů lze snadno přizpůsobit pro různé zobrazovací nebo výstupní zařízení (monitor, tiskárna atd.).

Slabé stránky:

- Uživatel je při vyhledávání textových objektů omezen pouze na své znalosti rodného (popřípadě úrovně znalosti cizích jazyků), který může použít pro vyhledávání.

- Vyhledávací mechanismy neumožňují při vyhledávání dávat do souvislostí různá slova nebo pojmy. Jedním z důvodů je používání přirozeného jazyka

v rámci textových informačních objektů a používání selekčních jazyků na straně systémů automatické indexace.

- Nedostačující pojmové vyhledávání textových informačních objektů, které je limitováno především technologickým vývojem a neefektivními algoritmy založená na určování indexačních termínů z řízeného slovníku nebo znalostní báze na základě jejich srovnání s výrazy z plného textu.

- Specifika textových objektů daná kódováním, jazykem nebo formátem redukuje množinu obdržených výsledků dotazu.

Příležitosti:

- Vyhledávání textových objektů na základě selekčního jazyka transformující lidské myšlenky do počítačem srozumitelných dat (například SQL v případě relačních databází apod.)

- Při vyhledávání textových informačních objektů lze používat booleovské, numerické, proximitní operátory nebo také regulárních výrazů.

- Technologický vývoj v oblasti automatizovaných vyhledávacích mechanismů pro textové informační objekty může přinést pokročilé a efektivní formy sémantické analýzy nebo umělé inteligence rozeznávající jednotlivé souvislosti mezi slovy. Příkladem může být třeba technika Latent Semantic Indexing (LSI), která se snaží překonat nedostatky spojené s vyhledáváním na základě termů, a to použitím statistických metod pro odvozování podobnosti mezi dokumenty a slovy.

Hrozby:

- Používání zvukových nebo grafických objektů pro rychlejší vyjádření informace (myšlenek) při komunikaci s počítačem.

- Možnost přehlcení nerelevantními textovými objekty na úkor relevantních.

4.7.2. Grafické informační objekty

Silné stránky:

- Vyjadřovací schopnost grafických informačních objektů v určitých případech mnohem převyšuje vyjadřovací schopnost u textových informačních objektů. Například v případě geografických informačních systémů, družicových snímků apod.

- Podobně jako u textových informačních objektů lze použít booleovské operátory, ale pouze u textových anotací a dalších doplňujících údajů. Nelze tedy vyhledávat přímo v samotném obsahu obrázku, ale pouze v textové anotací popisující obsah obrázku.

- Možnost využití speciálního ochranného znaku - vodoznaku (watermark), který je zakomponován do bitmapového souboru a na první pohled je patrné, že daný obrazový informační objekt je chráněn autorských právem.

- Možnost přidání digitálního vodoznaku neboli techniky přidání skrytých informací o autorských právech nebo dalších informací přímo do grafického objektu za účelem vysledování obrázků chráněných autorskými právy.

- Sémantický obsah obrázku je na rozdíl od textových objektů srozumitelný bez ohledu na jazykovou bariéru.

Slabé stránky:

- Neexistují pokročilé algoritmy pro indexování a vyhledávání grafických objektů na základě sémantického obsahu, struktury, objektů na daném obrázku apod. S tím také souvisí neexistující efektivní indexovací techniky schopné indexovat tento sémantický obsah obrázků.

- Vyhledávání grafických informačních objektů je „odkázáno“ především na využívání textových doplňujících informací popisující obsah obrázku. Pouze u těchto textových informací lze použít například booleovské operátory. Na rozdíl tex-

tových objektů nelze booleovské operátory použít pro vyhledávání v samotném obsahu objektu.

- Doplňující informace v textové podobě jsou závislé často na subjektivním pohledu odpovědného pracovníka. Zpracovávání a zařazování grafických informačních objektů s přidáváním doplňujících údajů je značně časové náročné.

- Komprimace bitmapových grafických informačních objektů je možná pouze ztrátovou metodou, čímž se snižuje kvalita daného obrázku (i když dané snížení kvality nemusí být patrné pouhým okem).

Příležitosti:

- Neustálý technologický rozvoj umožní vyhledávat na základě sémantického obsahu obrázku.

- Zdokonalování indexovacích algoritmů umožní efektivně využívat předem vypočítaných grafických prvcích, které budou hrát obdobnou roli jako písmena a slova při vyhledávání textového řetězce.

- Vyhledávání grafických objektů na základě náčrtku, příkladu objektu znázorněném na obrázku, struktury barev apod.

Hrozby:

- Vývoj technologií rozšíří prolomení ochranných prvků a tím pádem zneužívání autorských práv, což může k omezení volného přístupu k vyhledávání databázím grafických informačních objektů. Samotné používání, ukládání apod. nebude samozřejmě zdarma.

- Finančně náročný výzkum a vývoj a s tím související určitou nedostupnost indexovacích a vyhledávacích algoritmů v případě vysokých pořizovacích nákladech (patenty, licence apod.) značně zpomalí rozšíření těchto technologií.

4.7.3. Zvukové informační objekty

Silné stránky:

- Zvukové objekty nabízejí přesné zachycování melodie, rytmu, zvuku, zabarvení tónu, mluvené řeči.

- Možnost převádění zvukových objektů do textové podoby například za pomocí systému automatického rozpoznávání řeči (Automatic Speech Recognition, ASR).

Slabé stránky:

- Zvukové informační objekty je možné komprimovat pouze ztrátovou metodou, při které se snižuje kvalita.

- Doplňující informace v textové podobě jsou závislé často na subjektivním pohledu odpovědného pracovníka podobně jako u grafických objektů.

- Možnost použití například booleovských operátorů pouze u doplňujících textových informací popisující obsah zvukového objektu. Tyto operátory nelze použít při vyhledávání tempa, rytmu, zabarvení tónu apod.

- Možnost indexování pouze na základě textových polí reprezentující název objektu, jméno autora, stručný popis, formát objektu, klíčová slova apod.

- Neexistující algoritmy pro indexování sémantického obsahu zvukových informačních objektů.

Příležitosti:

- Technologický vývoj umožní efektivnější vyhledávání zvukových objektů na základě zahraní ukázky prvních tónů, podobnosti melodie zvukové ukázky uživatele a záznamu v databázi. Případně také vyhledávat podle rytmu zvukového objektu, zabarvení tónu nebo dokonce podrobnější strukturální analýzou.

- Rozpoznávání mluvené řeči, melodie, dále na zabarvení tónu, rytmu nebo podrobné strukturální analýzu zvukového objektu využitím technik LSA

- Umožnění indexování zvukových záznamů způsobem, kdy jsou jednotlivé noty, melodie, tempo indexovány a uživateli je poté nabídnuta možnost vyhledávat obdobné zvukové objekty shodující se s dotazem v podobě audio objektu.

Hrozby:

- Finančně náročný výzkum a vývoj a s tím související určitou nedostupnost indexovacích a vyhledávacích algoritmů v případě vysokých pořizovacích nákladech (patenty, licence apod.) značně zpomalí rozšíření těchto technologií.

4.8. Shrnutí a návrh možných zlepšení

V předchozích částech byly analyzovány možnosti vyhledávání společné pro jednotlivé typy informačních objektů v rámci uvedených příkladů databází, aby bylo možné vzájemné porovnání. Specifické možnosti vyhledávání dané typem objektů byly poté zmíněny jako samostatný bod analýzy.

Databáze textových informačních objektů nabízí možnosti vyhledávání nejen na základě jména autora, názvu díla nebo klíčových slov, ale také v případě propracovaného „vstupního“ zpracování jednotlivých textových informačních objektů, kdy textové informační objekty jsou zařazovány do jednotlivých kategorií, předmětových klasifikací, doplňovány abstraktem nebo komentářem a dalšími doplňujícími informacemi (slouží k přesnějšímu a snadnějšími vyhledání požadovaného textového objektu uživateli).

V rámci jednotlivých databází jsou výsledky vyhledávání prezentovány ve strukturované podobě. Například ArXiv.org i E-LIS jednotlivé textové informační objekty doplňují informacemi v podobě názvu objektu, jménu autora (autorů), abstraktu, komentáři, klíčovými slovy, předmětovou klasifikací (kategorie) v rámci dané databáze, identifikačním kódem (zejména specificky používány v rámci dané databáze), datem (například datum uložení objektu do databáze). E-LIS dále nabízí podrobné možnosti vyhle-

dávání na základě editora textového objektu, jazyku, specifikaci typu (např. kniha, technická zpráva, článek v časopisu, diplomová práce apod.).

Uvedené databáze textových informačních objektů umožňují specifikovat jméno autora v přesném tvaru, v případě nenalezení zadaného jména autora je nabídnut index autorů uvedených v databázi, čímž je uživateli umožněno zúžení dotaz. CERN nabízí použití regulárních výrazů, které umožní specifikovat daný dotaz. Samozřejmostí je použití booleovských operátorů, stemmingu neboli automatického převádění tvaru anglických slov na jejich kmeny odstraňováním koncovek, specifikování přesné fráze atd.

Pokud bychom porovnali databáze různých typů informačních objektů a možnosti vyhledávání, můžeme spatřit jistou výhodu u vyhledávání textových informačních objektů. Jedná se o možnost prohledávání plných textů dokumentů neboli v samotném obsahu objektu a nejen v názvu, abstraktu nebo klíčových slovech.

V případě příkladů databází grafických informačních objektů je vyhledávání v zásadě založeno na principu porovnávání zadaného dotazu s názvem nebo doplňujícím popisem obrázků. Databáze zvukových objektů nabízejí u uvedených příkladů daleko omezenější vyhledávání. Využívá se v zásadě zadávání klíčových slov, poté jsou nabízeny specifické možnosti vyhledávání, které se mohou lišit v jednotlivých databázích.

Podporované techniky vyhledávání jednotlivých databází prezentuje tabulka 2, kde znak „x“ značí, že databáze použití dané techniky vyhledávání umožňuje, oproti tomu znak „-" značí, že daná databáze danou možnost nenabízí. Důvodem malé podpory různých technik vyhledávání v případě zvukových objektů je dána specifickým zaměření těchto databází. Specifické možnosti vyhledávání nebyly do tohoto přehledu zahrnuty, neboť popsány v kapitole 4.6 a také se dané možnosti značně liší i mezi databázemi obsahující stejný typ informačních objektů.

V případě grafických informačních objektů je zřejmé, že určitý význam hraje „vstupní“ zpracování daného grafického objektu, zejména v podobě doplnění textových anotací a dalších informací. Grafické informační objekty musí být opravdu kvalitně zpracovány a opatřeny doplňujícími informacemi, které vystihují samotný obsah obrázku. Odpovědní pracovníci musí daný obrázek zařadit do správných kategorií, opatřit odpovídajícími klíčovými slovy.

Podle mého názoru databáze Corbis používá opravdu kvalitní zpracování a jednotlivé obrázky jsou opatřovány odpovídajícím popisem, obsah obrázku je co nejvíce vystihován klíčovými slovy. Uživatel tak má opravdu velké možnosti vyhledávání a specifikování svého dotazu.

Název databáze	Podporované techniky (možnosti) vyhledávání					
	Název objektu	Jméno autora	Textová anotace (klíčová slova)	Kategorie	Logické operátory	Přesná fráze
arXiv.org	x	x	x	x	x	x
E-LIS	x	x	x	x	x	x
CERN Document server	x	x	x	x	x	x
Corbis	x	x	x	x	x	x
NASA JSC	-	-	x	x	x	x
The Freesound Project	-	x	x	-	-	-
The Sound Effetcts Library	-	-	-	-	-	x
The Philharmonia Sample Library	-	-	-	-	-	-

Tabulka 2: Přehled podporovaných technik vyhledávání u vybraných databází

Na druhou stranu lze spatřit určité potíže v případě, kdy uživatel potřebuje vyhledat obrázky (fotografie) podobné struktury. Například obrázky se stejnou barevnou strukturou. V tomto případě jsou už nabízené techniky vyhledávání méně dostačující. Uživatel tak může použít klíčové slovo, ale poté obdrží mnoho nerelevantních obrázků, které neodpovídají. Problémem je samozřejmě také to, že zatím neexistuje technologie, která by umožnila opravdu efektivní vyhledávání obrázků na základě dotazu taktéž v podobě obrázku.

Vyhledávání textových informačních objektů používá porovnávání textových řetězců dotazu uživatele a sémantického obsahu databáze. Stejný princip by se tedy měl aplikovat i pro databáze obsahující grafické objekty. Vstupem pro vyhledávání obrázků by byl také obrázek, který by bylo možné efektivně porovnat (rozpoznáním jednotlivých objektů) a zjistit, zda zahrnuje stejný (nebo podobný) sémantický obsah.

Databáze zvukových informačních objektů jsou v řadě případů úzce specializovány a k tomu účelu je také přizpůsobena možnost vyhledávání. Obvyklým způsobem je vyhledávání na základě textových anotací, názvu zvukového objektu nebo jména autora. Stejně jako u databází grafických informačních objektů zde záleží na vstupním zpracování a doplnění odpovídajících textových informací.

Databáze The Freesound Project nabízí také možnost vyhledávání na základě vlastností vycházející z podstaty zvukových objektů v digitální podobě. Uživatel tak může zvolit minimální hodnotu vzorkovací frekvence, hodnotu přenosové rychlosti (bitrate) apod. Možnosti vyhledávání jsou značně rozdílné mezi jednotlivými databázemi zvukových objektů. Například databáze Sound Effects Library nabízí vyhledávání vycházející z vlastností zvukových objektů (Audio Search) pouze na základě kvality (stereo, mono) nebo délky záznamu. Stále ale tyto informace musí být doplněny k jednotlivým objektům v podobě textové informace. Obdobně jako u grafických informačních objektů neexistuje zatím efektivní systém a technologie, která by umožnila opravdu efektivní vyhledávání na základě ukázky zvukové záznamu apod.

Myslím, že je potřeba také globálního rozšíření základních možností vyhledávání pro grafické informační objektů, stejně jako je samozřejmé vyhledávání podle jména autora nebo názvu objektu (nejen u textových informačních objektů). V prostředí internetu se můžeme setkat s možností vyhledávání multimediálních objektů (viz Google nebo AlltheWeb), ale právě technologické možnosti vyhledávání multimediálních objektů ne na základě textových anotací, ale na základě sémantického obsahu jsou hlavním omezujícím faktorem efektivního vyhledávání.

Zajímavým příkladem může být již zmiňované Tiltomoto Beta představující „vývojovou“ webovou stránku pro testování algoritmů pro vyhledávání obrázků na základě analyzování podobnosti a vztahů mezi obrázky. Řekl bych, že takovéto možnosti vyhledávání (v databázích i v samotném prostředí internetu) budou přispívat většímu povědomí široké veřejnosti uživatelů.

Podle mého názoru je problémem také značná rozdílnost možností vyhledávání u jednotlivých typů informačních objektů a také databází. Zejména databází grafických nebo zvukových informačních objektů. Na základě uvedené analýzy a srovnání, bych spatřoval největší nedostatky v případě databází zvukových informačních objektů.

Přestože se v některých případech jedná o specificky zaměřené databáze, navrhoval bych propracovanější vstupní zpracování.

Související zlepšení bych viděl také ve stanovení kategorií (předmětové klasifikaci), která by se v obecné formě mohla použít v rámci zvukových databází. Jednotlivé databáze by poté mohli obecné kategorie rozšířit dle svého uvážení a zaměření. To by mohlo usnadnit uživateli orientaci a jakousi „standardní“ techniku vyhledávání aplikovatelnou v různých databázích. Stanovení obecných kategorií by mohlo jisté východisko z možných standardů respektive doporoučení v podobě RFC, které zavedení těchto kategorií podpořily.

Jedno z možných zlepšení by mohlo být také v prezentování vyhledaných informačních objektů. Výsledky vyhledávání u textových informačních objektů jsou prezentovány v určité formě, která v přehledné podobě uvádí jméno autora, název objektu, popřípadě abstrakt, klíčová slova, komentář atd. Podobně zavedený (pokud možno jednotný) způsob prezentování názvu objektu, jména autora, kategorie, klíčových slov, textové anotace apod. by byl u zvukových nebo grafických informačních objektů možným zlepšením. Odpovídající působ prezentování výsledků vyhledávání bych vyzdvihl u databáze fotografií Corbis.

Z uvedených výsledků srovnání možných technik vyhledávání bych také navrhoval klást větší důraz na implementování logických operátorů v rámci vyhledávání. Booleovské, proximitní nebo numerické operátory by se jistě uplatnili v oblasti zvukových a také grafických informačních objektů, jak je patrné opět na příkladu databáze Corbis, která například booleovské operátory podporuje.

V současné době jsou patrné velké nedostatky v případě vyhledávání zvukových a grafických objektů. Podle mého názorů je ale podstatné zlepšení technik vyhledávání těchto informačních objektů závislé na technologickém pokroku a vývoji nových technologií, která by umožnily vyhledávání na základě sémantického obsahu.

V případě textových informačních objektů bych viděl značná zlepšení v možnostech, které nabízí odvozování a sémantická analýza. Systémy, které by byly schopné „chápat text v souvislostech“ jsou tedy předmětem dalšího výzkumu a vývoje. Překážkou zde zůstává přirozený jazyk, s jehož zpracováním v rámci sémantické analýzy mají takovéto systémy stále problémy.

V případě vylepšení indexování textových informačních objektů bych považoval za důležitý další vývoj v oblasti tzv. pojmové indexace. Důležitou oblastí pro další vývoj je také indexování grafických objektů na základě reprezentace dvourozměrného řetězce, vektorů reprezentující předběžně vypočítanou část textur v daném obrázku, indexování vizuálního obsahu (například barva, struktura atd).

Vyhledávání grafických informačních objektů CBIR (content-based image retrieval*)*, kdy se charakteristiky popisující obsah obrázku jako například histogram, barevné rozvržení, struktura nebo obrázku je z mého pohledu jedním z hlavních zlepšení v oblasti „image retrieval“. Technika CBIR se ale stále potýká se problémy v případě extrahovat charakteristik obrázku jako například objekty zobrazené na obrázku. To je jedním z důvodů, proč je CBIR stále na počátku svých možností a praktické nasazení v globální úrovni internetu je stále dlouhodobou záležitostí. Značně zlepšení přinese možnost vyhledávání grafických objektů použitím konceptu obrázku jako samotného dotazu.

U zvukových informačních objektů bych označil za opravdu podstatné zlepšení již zmiňované „Query by humming“ (QbH) neboli způsob vyhledávání zvukových objektů, kdy se jako vstupní dotaz používá uživatelem zahraná melodie a porovnává se s objekty uloženým v databází. Určitým zlepšením v budoucnu bude určitě vyhledáváním hudby na základě zabarvení tónu nebo rytmu tempa. Tyto techniky ale naráží na problémy spojené s efektivním indexováním potřebným pro tento způsob vyhledávání a samozřejmě na nedostatky vyhledávacích algoritmů, které by umožnily vyhledávat například na základě strukturální analýzy zvukových objektů.

5. Závěr

Množství informačních objektů v digitální podobě se neustále zvyšuje, a to nejen v případě textových, ale i grafických a zvukových informačních objektů. V současné době se odhaduje, že zhruba 73 % obsahu World Wide Web představují grafické informační objekty neboli obrázky. A stále existuje pouze několik málo metod vyhledávání těchto informačních objektu v prostředí internetu [26, 32]. Uživatelé se proto setkávají s ohromným rozšířením grafických a zvukových objektů, ale na druhé straně se potýkají s nedostatečným vývojem na straně indexování a vyhledávání těchto objektů. Na základě uvedených příkladů databází a možností vyhledávání uvedených příkladů databází, lze tento rozdíl v technikách a možnostech vyhledávání textových, zvukových a grafických informačních objektů také pozorovat.

Tato práce se zaměřovala na jednotlivé typy informačních objektů a snažila se přiblížit problematiku vyhledávání různých typů informačních objektů na příkladech volně dostupných databázích. Cílem této práce nebylo najít možná řešení problémů, ale především poukázat na současné techniky a nedostatky vyhledávání u vybraných typů informačních objektů. Databází daných typů informačních objektů existuje samozřejmě celá řada a mohou nabízet i jiné specifické způsoby vyhledávání. Účelem ale nebylo připravit přehled existujících databází informačních objektů a jejich podrobné analyzování. Problematika indexování a vyhledávání různých typů informačních objektů je samozřejmě velmi rozsáhlá a nebylo možné v rámci této práce zahrnout veškeré aspekty.

Nejčastější způsob vyhledávání grafických informačních objektů spočívá ve vytvoření textové anotace popisující obsah tohoto grafického objektu. Techniky vyhledávání zvukových typů informačních objektů jsou do určité míry podobné, neboť se při vyhledávání používá název objektu, jméno autora anebo textové anotace. Při tomto procesu je důležitá zejména úroveň odborného vstupního zpracování daného objektu. V současné době se pro specializované databáze používá manuální zpracovávání, i když je to časově i finančně náročnější než automatické zpracovávání, které bohužel v některých případech nedosahuje takových kvalit díky nedostatečným algoritmům a technologickému pokroku.

Zajímavé možnosti indexování a vyhledávání zvukových nebo grafických objektů v podobě prototypů systémů založené na technikách Latent Semantic Analysis se ob-

jevují již v současné době. Systémy používající převod zvukových objektů (digitálních) do textové podoby pomocí systému automatického rozpoznávání řeči (Automatic Speech Recognition, ASR) nebo používající techniky LSA pro indexování zvukových informačních objektů jsou stále předmět výzkumu. Prototypů vyhledávání obrázků za použití metod CBIR nebo také již zmíněné analýzy LSA je také velké množství. Podle výsledků různých prototypů systémů je zřejmé, že je potřeba dalšího výzkumu a vývoje v těchto oblastech. V případě vyhledávání textových informačních objektů se používá porovnávání textových řetězců v dotazu (zadaného uživatelem) se sémantickým obsahem databáze. Snahou je tedy dosažení stejného principu také u grafických i zvukových objektů. Vstupem pro vyhledávání obrázků by byl také obrázek, pro vyhledávání audio objektů by byl audio objekt. Podstatným problém tohoto způsobu vyhledávání je zjištění sémantického obsahu daných objektů.

Jak již bylo zmíněno, požadavky na možnosti vyhledávání daleko převyšují možnosti a dostupné technologie, která by vyhledávání na základě sémantického obsahu objektu umožnily. Důležitost pokročilých technologií a opravdu efektivních algoritmů, umožňující vyhledávání obrázků na základě ukázky nebo náčrtku, zahráním pár tónů melodie pro vyhledání zvukových objektů shodující se s touto ukázkou apod., nikdo nezpochybňuje. Řada odborníků proto neustále pracuje na vývoji a zdokonalování těchto algoritmů a myslím, že je jen otázka času, kdy se stane takovéto vyhledávání realitou. Proto můžeme doufat, že postupem času se uživatelé dočkají efektivního vyhledávání obrázků na základě sémantického obsahu a hudby na základě zahrání pár tónů melodie.

Myslím, že právě vyhledávání na základě zadání dotazu v podobě právě toho typu objektu, který chceme vyhledat, by mohl být zajímavou možností do budoucna. Stejný princip jako u textových objektů, kdy vyhledáváme na základě dotazu v textové podobě, by v budoucnu mohl být běžně používán i v případě grafických a zvukových objektů. Uživatel při vyhledávání obrázků by tak zadal dotaz v podobě obrázku, který mu systém nabídne například jako vzor anebo bude uživatel moci načrtnout základní tvar objektu, které by měl požadovaný obrázek obsahovat. Obdobný způsob by mohl být také v případě zvukových objektu, kdy uživatel zadá vstupní dotaz v podobě tónu nebo ukázky určité melodie (například pomocí rozhraní systému umožňujícího kombinaci tónů nebo komponování dané ukázky).

Slovník použitých pojmů

Informační objekt – informace nebo skupina informací tvořících jednotný celek. V rámci této práce jsou informační objekty nazírány pouze ve formátu elektronickém [1].

Informační zdroj –informační objekt, který obsahuje dostupné informace odpovídající informačním potřebám uživatele. Pojem informační zdroj je v této práci považován za odpovídající pouze ve formátu elektronickém (včetně zdrojů dostupných online).

Digitální objekt - objekt zaznamenaný nebo přenášený v digitálním formátu, tj. jako řetězce bitů, který byl v této formě vytvořen nebo do ní konvertován. Umožňuje kompresi dat a přidání dodatečné informace určené buď k opravě chyb nebo k popisu obsažených údajů (metadata). Digitální objekt je považován za elektronickou podmnožinu pojmu informační objekt. V rámci této práce může být digitální objekt považován za ekvivalent pojmu informační objekt.

Databáze – normativní výklad charakterizuje databázi jako soubor datových objektů v elektronické formě uložených společně podle jednoho schématu a zpřístupňovaných počítačem [3]. V rámci této práce je pojem databáze definován jako soubor dat, který je organizován pro jejich snadné vyhledávání a manipulaci s nimi. Z hlediska vyhledávání je databáze obvykle interně rozdělena na vlastní data (jsou předmětem vyhledávání) a datové indexy, které se používají pro rychlé vyhledávání na základě daných kriterií [2].

Digitální knihovna – komplexní systém zahrnující elektronické informační zdroje a služby umožňující získávání, zpracování, vyhledávání a následné zpracování informací.

Metadata – data sdružené s objekty, která zbavují jejich potenciální uživatele nutnosti předběžné znalosti existence či charakteristik těchto objektů. V této práci je tento pojem používán především v souvislosti s elektronickým zdroji (datové soubory, textové informace, grafika, hudba) vyjádřená v digitální formě. Obecný koncept metadat zahrnuje údaje pro věcné indexování a identifikační údaje[2].

MIME *(***M**ultipurpose **I**nternet **M**ail **E**xtensions) – internetový standard vytvořený roku 1992 organizací Internet Engineering Task Force (IETF) popisující rozšiřující formátování zpráv v jiném formátu než ASCII. Pomocí tohoto standard je možné posílat a přijímat grafické, zvukové a video objekty prostřednictvím nejen emailových zprav, ale také v prostředí WWW a dalších služeb.

ASCII (A*merican* ***S****tandard* ***C****ode for* ***I****nformation* ***I****nterchange*) - americký standardní kód obsahující 256 znaků pro výměnu informací představuje v podstatě znakovou sadu (kódovou tabulku), která definuje znaky anglické abecedy a další znaky používané v informatice za účelem ukládání textu v počítači nebo přenosu textu telekomunikačními sítěmi.

CBIR *(****C****ontent-****B****ased* ***I****mage* ***R****etrieval)* - sada technik pro získávání relevantních obrázků z databáze na základě automaticky „vytěžených“ vlastností grafického objektu [31].

LSI *(****L****atent* ***S****emantic* ***I****ndexing)* - technika vyhledáváním používající statistických metod pro odvozování podobnosti mezi dokumenty a slovy. LSI tak považuje objekty mající mnoho společných slov za sémanticky (významově) blízké a objekty s málo společnými slovy za sémantický vzdálená. Technika LSI používá matici, jejíž řádky odpovídají termům a sloupce dokumentům (objektům). Vychází z analýzy LSA (Latent Semantic Analysis), což představuje techniku pro zpracování přirozeného jazyka za použití vektorů reprezentujících jednotlivé dokumenty.

QbH *(****Q****uery* ***b****y* ***H****umming)* - způsob vyhledávání zvukových objektů, kdy se jako vstupní dotaz používá právě uživatelem zahraná melodie a porovnává se s objekty uloženým v databází.

OMR *(****O****ptical* ***M****usic* ***R****ecognition)* – technika optického rozpoznávání hudby, kdy analogový akustický signál je nasnímán do digitální podoby, noty jsou vysegmentovány z akustického záznamu, identifikuje se frekvence not a každá nota je označena intonací.

Seznam použité literatury

1. *KTD - Česká terminologická databáze knihovnictví a informační vědy* (TDKIV) [databáze online]. c2002 [2007-11-30]. Praha : Národní knihovna ČR. Dostupný z WWW:<http://sigma.nkp.cz:4505/F/?func=file&file_name=find-b&local_base=ktd>.

2. SKLENÁK, V. a kol. *Data, informace, znalosti a Internet*. Praha: C. H. Beck, 2001, 507 s. ISBN 80-7179-409-0.

3. ČSN ISO 690-2 (01 0197). *Informace a dokumentace – Bibliografické citace – Část 2: Elektronické dokumenty nebo jejich části.* Praha : Český normalizační institut, 2000. 22 s.

4. FREED, N., BORENSTEIN, N. *RFC 2045 : Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies*. 1996. Dostupný z WWW: <http://www.ietf.org/rfc/rfc2045.txt>.

5. FREED, N., BORENSTEIN, N. *RFC 2046 : Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types*. 1996. Dostupný z WWW: <http://www.ietf.org/rfc/rfc2046.txt>.

6. *IANA Home Page* [online]. The Internet Corporation fo Assigned Names and Numbers, 2006, Page Updated 27-Jun-2007 [cit. 2007-11-12]. Dostupný z WWW: <http://www.iana.org>.

7. *IETF Home Page* [online]. The Internet Engineering Task Force [cit. 2007-11-12]. Dostupný z WWW: <http://www.ietf.org>.

8. BARTOŠEK, M. Digitální knihovny – teorie a praxe. *Národní knihovna* [online]. 2004, roč. 15, č. 4, s. 233-254 [cit. 2007-12-10]. Dostupný z WWW: <http://knihovna.nkp.cz/NKKR0404/0404233.html>.

9. ARMS, W. Y. *Digital Libraries.*Cambridge : MIT Press, 2000. ISBN 0-262-01880-8.

10. KAHN, R., WILENSKY, R. *A Framework for Distributed Digital Object Services. Technical Report hdl:cnri.dlib/tn95-01.* CNRI, 1995. Dostupný též z WWW: <http://www.cnri.reston.va.us/home/cstr/arch/k-w.html>.

11. BRATKOVÁ, Eva. Metadata jako nový nástroj pro komunikaci webovských informačních zdrojů. Národní knihovna : knihovnická revue. 1999, č. 4, s. 178-195. Dostupné též z WWW: <http://full.nkp.cz/nkkr/Nkkr9904/9904178.html>.

12. BERNERS-LEE, Tim. *Metadata Architecture* [online]. W3C, 1997, last edited 2000-09-21 [cit. 2007-11-30]. Dostupný z WWW: <http://www.w3.org/DesignIssues/Metadata.html>.

13. BRETHERTON, F. P., SINGLEY, P. T. *Metadata: A User's View, Proceedings of the International Conference on Very Large Data Bases (VLDB).* 1994. 1091-1094 s.

14. DCMI. *DCMI Metadata Terms* [online]. Dublin : DCMI, c1995-2007 [cit. 2007-11-25]. Dostupný z: <http://dublincore.org/documents/dcmi-terms>.

15. BARTOŠEK, M. Vyhledávání v Internetu a DUBLIN CORE. *Zpravodaj ÚVT MU*. 1999, roč. IX, č. 4, s. 1-4. ISSN 1212-0901. Dostupný též z WWW: <http://www.ics.muni.cz/zpravodaj/articles/153.html>.

16. DCMI. *Dublin Core Metadata Initiative (DCMI) : Home Page* [online]. Dublin : DCMI, c1995 - 2007 [cit. 2007-11-30]. Dostupný z WWW: <http://purl.org/dc/>.

17. DCMI. *Dublin core metadata Element Set, Verison 1.1* [online]. Dublin : DCMI, c1995-2007 [cit. 2007-11-30]. Dostupný z WWW: <http://dublincore.org/documents/dces/>.

18. BRATKOVÁ, Eva. K otázkám pojmu, třídění a typologie internetových a webovských informačních zdrojů. *Národní knihovna : knihovnická revue*. 1998, roč. 9, č. 5, s. 262-276. ISSN 0862-7487. Dostupný též z WWW: <http://eprints.rclis.org/archive/00003558/01/ResourceTypes05.pdf>.

19. Free On-line Dictionary of Computing (FOLDOC). *Bezier curve* [online]. [cit. 2007-11-30]. Dostupný z WWW: <http://foldoc.org/?Bezier+curve>.

20. SEDLÁČEK, Vojtěch. *Fenomén Google*: diplomová práce. Praha: FF UK, 2007. 110 s.

21. POKORNÝ, J., SNÁŠEL, V., KOPECKÝ, M. *Dokumentografické informační systémy*. 2. vyd. Praha : Karolinum, 2005. 184 s. ISBN 80-246-1148-1.

22. Google. *Google Image Search* [online]. c2007. Dostupný z WWW: <http://images.google.com>.

23. HLAVENKA, Jiří. *Mistrovství ve vyhledávání na internetu*. 1. vyd. Praha : Computer Press, 2002. 195 s. ISBN 80-7226-759-0.

24. WOODLEY, Mary S. Glossary [online]. Dublin (OH, USA) : *Dublin Core Metadata Initiative*, 2000 [cit. 2007-11-30]. Dostupný z WWW: <http://www.dublincore.org/documents/2001/04/12/usageguide/glossary.shtml>.

25. PENTLAND, A., PICARD, R. a kol. Photobook: Content-based manipulation of image databases. In *SPIE Storage and Retrieval for Image and Video Databases*, february 1994.

26. LAWRANCE, S. , GILES, C. Searching the World Wide Web. *Science.* 1998, roč. 280, č. 5360, s. 98 – 100.

27. COX, J. I., MILLER, L. M. at al. The Bayesian Image Retrieval System, PicHuter: Theory, Implementation and Psychophysical Experiments. *IEEE Transactions on image Processing*. 2000, roč. 9, č. 1, s. 20 - 37.

28. FLICKNER, M. at al. Query by image and video content: The QBIC system. *IEEE Computer*. 1995, roč. 28, s. 22-32.

29. WANG, Z. J., WIEDERHOLD, G. FIRSCHEIN, O. aj. Content-based image indexing and searching using Daubechies' wavelets. *International Journal on Digital Libaries*. 1997, roč. 1, s. 311-328.

30. WANG, Z. J. SIMPLIcity: A region-based image retrieval system for picture libraries and biomedical image databases. In *proceeding of the ACM Multimedia Conference*, Los Angeles, 2000.

31. WANG, Z. J. *Semantics-sensitive integrated matching for picture libraries and biomedical image databases*. 2000. Disertační práce. Stanford University. Dostupný z WWW: <http://infolab.stanford.edu/~wangz/project/thesis/wang.pdf>.

32. LEW, M.S. Next-generation web searches for visual content. *IEEE Computer*. 2000, roč.. 33, č. 11, s. 46-53. ISSN 0018-9162.

33. CARSON, C., BELONGE, S. GREENSPAN, H. aj. Blobworld: Image segmentation using Expectation-Maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2002, roč. 24, č. 8, s. 1026 – 1038. ISSN 0162-8828.

34. BRATKOVÁ, E. *Digitální knihovny : Typy digitálních objektů (1) : [WWW prezentace]* [online]. Praha : ÚISK FF UK, 2002- [cit. 2007-12-08]. Dostupné z WWW: <http://www1.cuni.cz/~brt/dvk/dk/dk1.htm>.

35. BARNARD, K., FORSYTH, D. A. Learning the Semantics of Words and Pictures. In *International Conference of Computer Vision*. 2001, s. 408 - 415. Dostupný též z WWW: <http://luthuli.cs.uiuc.edu/~daf/papers/WAP-fin.pdf>.

36. MCNAB, R.J., SMITH, A. L., WITTEN, I.H. *Music Information Retrieval Using Audio Input.* Dostupný též z WWW: <http://www.cs.waikato.ac.nz/~ihw/papers/97LAS-RJM-IHW-Music-Info.pdf>.

37. GHIAS, A. a kol. Query by Humming. Musical Information Retrieval in An Audio Database. In *Proceeding of the ACM Multimedia*, s.231–236, 1995. Dostupný též z WWW: <http://www.cs.cornell.edu/Info/Faculty/bsmith/query-by-humming.html>.

38. CHEN, J., CHEN, A. Query by Rhythm: An Approach for Song Retrieval in Music Databases. In *Proceedings of the Eighth International Workshop on Re-*

search Issues in Data Engineering, Continuous-Media Databases and Applications, 1998, s. 139-146.

39. AUCOUTURIER, J. J., PACHET, F. Improving Timbre Similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences.* 2004. Dostupný též z WWW: <http://www.csl.sony.fr/downloads/papers/uploads/aucouturier-04b.pdf>.

40. TZANETAKIS, George, ERMOLINSKYI, Andreye, COOK, Perry. Beyond the Query-by-Example Paradigm: New Query Interfaces for Music Information Retrieval. In *Proceeding of the International Computer Music Conference (ICMC)*, Gothenburg, Sweden 2002. Dostupný též z WWW: <http://citeseer.ist.psu.edu/cache/papers/cs/26772/http:zSzzSzwww-2.cs.cmu.eduzSz~gtzanzSzworkzSzpubszSzicmc02gtzan.pdf/beyond-the-query-by.pdf>.

41. AUCOUTURIER, J. J., PACHET, F., HANAPPE, P. *From sound sampling to song sampling.* Sony Computer Science Laboratory Paris. Dostupný z WWW: <http://www.csl.sony.fr/downloads/papers/2004/aucouturier-04d.pdf>.

42. PAPÍK, Richard. Vyhledávání informací I. Umění či věda? *Národní knihovna : knihovnická revue*. 2001, roč. 12. č. 1, s. 18-25. Dostupný též z: <http://full.nkp.cz/nkkr/NKKR0101/0101018.html>.

43. KUČEROVÁ, Helena. *Metody a techniky vyhledávání informací* [online]. Praha: Vyšší odborná škola informačních služeb, datum poslední aktualizace: 7. 1. 2006 [cit. 2007-11-20]. Vyhledávání informací. Dostupný z WWW: <http://web.sks.cz/users/ku/MTI/vyhled.htm>.

44. Tiltomo. *About : Tiltomo : Content Based Visual Image Search* [online]. [cit. 2007-11-30]. Dostupný z WWW: <http://www.tiltomo.com>.

45. BRATKOVÁ, Eva. Otevřený přístup, digitální knihovny a citační služby. In *INFORUM 2006 : 12. ročník konference o profesionálních informačních zdrojích, Vysoká škola ekonomická, Praha, 23.-25. května 2006* [online]. Praha : Alberti-

na icome Praha, 2006 [cit. 2007-11-19]. [21 s.]. Dostupný též z: <http://www.inforum.cz/inforum2006/pdf/Bratkova_Eva.pdf>.

46. Wikipedia, The Free Encyclopedia. *Geographic information system* [online]. c2007 [cit. 2007-11-29]. Dostupný z WWW: <http://en.wikipedia.org/w/index.php?title=Geographic_information_system&oldid=176836709>.

47. Freesound :: what is fressound? [online]. [cit. 2007-12-01]. Dostupný z WWW: <http://freesound.iua.upf.edu/whatIsFreesound.php>.

48. BERNERS-LEE, T. *RFC 1738 : Uniform Resource Locator (URL)* [online]. 1994 [cit. 2007-11-29]. Dostupný z WWW: <http://www.ietf.org/rfc/rfc1738.txt>.

49. BERNERS-LEE, T. *RFC 3986 : Uniform Resource Identificator (URI): Generic Syntax* [online]. 2005 [cit. 2007-11-29]. Dostupný z WWW: <http://www.ietf.org/rfc/rfc3986.txt>.

50. MOATS, Ryan. *RFC 2141 : URN Syntax* [online]. 1997. Dostupný z WWW: <http://www.ietf.org/rfc/rfc2141.txt>.

51. HILSE, Hans-Werner; KOTHE, Jochen. *Implementing persistent identifiers : overview of concepts, guidelines and recommendations*. London : Consortium of European Research Libraries; Amsterdam European Commission on Preservation and Access, 2006. 57 s. Dostupný také z WWW: <http://www.knaw.nl/ecpa/publ/pdf/2732.pdf>.

52. OCLC. <dehn@oclc.org>. *Persistent URL Home Page* [online]. Dostupný z WWW: <http://purl.oclc.org>.

53. Corporation for National Research Initiatives (CNRI). *HANDLE.NET – The Handle System* [online]. 2007. Dostupný z WWW: <http://www.handle.net>.

54. KUČEROVÁ, Helena. *Kódování dat, znakové sady* [online]. Praha: Vyšší odborná škola informačních služeb, datum poslední aktualizace: 5. 6. 2007 [cit. 2007-12-02]. Dostupný z WWW: <http://web.sks.cz/users/ku/MTI/vyhled.htm>.

55. WRIGHT, A. *In search of the deep Web*. Dostupný z WWW: <http://dir.salon.com/story/tech/feature/2004/03/09/deep_web/index.html>.

56. BERGMAN, Michael K. *The Deep Web: Surfacing Hidden Value Deep Content* [online]. BrightPlanet, c2001. [cit. 2007-12-07]. Dostupný z WWW: <http://www.brightplanet.com/pdf/deepwebwhitepaper.pdf>.

57. *The Digital Object Identifier System* [online]. International DOI Foundation, 2004. Dostupný z WWW: <http://www.doi.org/>.

58. ARMS, W. Key Concepts in the Architecture of the Digital Library. *D-Lib Magazine* [online]. July 1995 [cit. 2007-12-04]. Dostupný z WWW: <http://www.dlib.org/dlib/July95/07arms.html>. ISSN 1082-9873.

59. DEERWESTER, S. a kol. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* . 1990, roč. 41, č. 6, s. 391-407. Dostupný též z WWW: <http://lsi.research.telcordia.com/lsi/papers/JASIS90.pdf>.

60. BRADFORD, B .R. Relationship Discovery in Large Text Collections Using Latent Semantic Indexing. In *Proceedings of the Fourth Workshop on Link Analysis, Counterterrorism and Security*, Bethesda 2006. Dostupný též z WWW: <http://www.siam.org/meetings/sdm06/workproceed/Link%20Analysis/15.pdf>.

61. LYNCH, C. A. The technologies of electronic imaging. *Journal of the American Society for Information Science*. 1991, roč. 42, č. 8, s. 578 - 585.

62. OBEID, M., JEDYNAK, B.,DAUODI, M. Image Indexing & Retrieval Using Intermediate Features. In *Proceedings of the ninth ACM international conference on Multimedia,* Ottawa 2001. s. 531 – 533. ISBN 1-58113-394-4.

63. CHANG, S. K. a kol. A Unified Approach to Iconic Indexing, Retrieval, and Maintenance of Spatial Relationships in Image Databases. *Journal of Visual Communication and Image Representation. 1996, roč. 7, č. 4, s. 307 – 324.*

64. NABIL, M. a kol. Indexing and retrieval of video based on spatial relation sequences. In *Proceedings of the seventh ACM international conference on Multimedia (Part 2)*, Orlando 1999. s. 119 – 122. ISBN 1-58113-239-5.

65. MAKHOUL, John a kol. Speech and language technologies for audio indexing and retrieval. In *proceedings of the IEEE.* 2000, roč. 88, č. 8. Dostupný z WW: <http://www.bbn.com/resources/pdf/Audio-Indexing-Retrieval.pdf>.

66. KURIMO, M. Indexing spoken audio by LSA and SOMS. *Kohonen Maps.*1999, s. 363 – 374.

67. AJMERA, J., McCOWAN, I., BOURLARD, H. *An Online Audio Indexing System*. IDIAP, 2002. Dostupný též z WWW: <ftp://ftp.idiap.ch/pub/reports/2003/rr03-39b.pdf>.

68. ABBERLEY, D. , KIRBY, D., RENALS, S. aj. The THISL broadcast news retrieval system. In *ESCA workshop on Accessing Information in Spoken Audio.* 1999, s. 1419.

69. KUTÁČ, Daniel. Vyspělé databázové technologie III. : Bitmapové indexy. *Databázový svět* [online]. 2004 [cit. 2007-12-10]. Dostupný z WWW: <http://www.dbsvet.cz/view.php?cisloclanku=2004042701>. ISSN 1213-59.

70. MACHONSKÁ, J. *Cenová politika vybraných světových databázových center: historie a současnost cenových relací za služby*: diplomová práce. Praha: FF UK, 2001. 97 s., příl.

71. *Wikipedie: Otevřená encyklopedie: Regulární výraz* [online]. c2007 [cit. 2007-12-10]. Dostupný z WWW: <http://cs.wikipedia.org/w/index.php?title=Regul%C3%A1rn%C3%AD_v%C3%BDraz&oldid=2039802>.

72. *Wikipedie: Otevřená encyklopedie: Přenosová rychlost* [online]. c2007 [cit. 2007-12-10]. Dostupný z WWW: <http://cs.wikipedia.org/w/index.php?title=P%C5%99enosov%C3%A1_rychlost&oldid=1818431>.

73. UCREL, Lancaster University. *UCREL Semantic Analysis System (USAS)* [online]. [2007]. Dostupný z WWW: <http://www.comp.lancs.ac.uk/ucrel/usas>.

74. BALÍKOVÁ, Marie, STROSSA, Petr. Systém M-CAST : vícejazyčný systém agregace informací . In *INFORUM 2006 : 13. ročník konference o profesionálních informačních zdrojích*, *Vysoká škola ekonomická, 23.-25. května 2006* [online]. Praha : Albertina icome Praha, 2007 [cit. 2007-12-11]. Dostupný též z WWW: <http://www.inforum.cz/pdf/2007/balikova-marie1.pdf>.

75. SCHWARZ, Josef. *Současný stav a trendy automatické indexace dokumentů* [online]. Ikaros, 2003 [cit. 2007-12-11]. Dostupný z WWW: <http://www.ikaros.cz/node/1300>. ISSN 1212-507.

76. KUČEROVÁ, Helena. *Sekundární informace, metadata* [online]. Praha: Vyšší odborná škola informačních služeb, datum poslední aktualizace: 18. 11. 2007 [cit. 2007-12-12]. Dostupný z WWW: <http://web.sks.cz/users/ku/ZIZ/metadata.htm>.

77. ZELENÝ, Jindřich. *Metody pro přístup k datům v databázových systémech* [online]. [2002] [cit. 2007-12-12]. Dostupný z WWW: <http://nb.vse.cz/~zelenyj/it380/eseje/jzeleny/indexy/indexy.htm#_Toc468875867 >.

Evidence výpůjček

Prohlášení:

Dávám svolení k půjčování této diplomové práce. Uživatel potvrzuje svým podpisem, že bude tuto práci řádně citovat v seznamu použité literatury.

V Praze, 12.12. 2007.

Radek Bitrman ( )

Jméno	Katedra / Pracoviště	Datum	Podpis