ÚSTAV INFORMAČNÍCH STUDIÍ A KNIHOVNICTVÍ
FF UK V PRAZE

Petr Boldiš

Pořádání informací a znalostí na internetu: analýza a trendy

Verze 1.0

Praha

Listopad 2008

1 PŘEDMLUVA

Žijeme ve světě, který se zásadním způsobem mění na základě stále se zvyšující důležitosti informací v každém aspektu našeho života. Tato změna se projevuje také v oblasti práce s odbornými a vědeckými informacemi, kde se elektronická média stávají nejdůležitějším médiem pro komunikaci.

V této souvislosti se naskýtají také otázky, jak změnilo toto elektronické prostředí způsoby, metody a perspektivy pořádání informací. Jestliže člověk bude ke svému životu potřebovat pracovat s informacemi, tak bude problematika pořádání informací stále aktuální. Růst objemu nově publikovaných informací začíná být značným problémem a tak se automaticky objevují otázky jako: Jak uspořádat stávající informace pro pozdější využití? Je možné zachytit a zatřídit všechny publikované informace? Lze i v tomto prostředí aplikovat stávající metody pro pořádání informací nebo nastává čas radikálních změn?

Hlavní oblastí, kde se tyto otázky objevují, je počítačová síť internet. Patrně žádný jiný informační zdroj neprodělal tak dynamický vývoj v nárůstu objemu publikovaných informací a zároveň v rychlosti nárůstu počtu uživatelů. Z tohoto důvodu se tato práce soustřeďuje na internet jako na experimentální prostředí, které může velmi dobře naznačit problémy i možnosti řešení pro pořádání informací.

Název práce byl zvolen s ohledem na zažité označení této problematiky (knowledge organization), které je do češtiny překládáno jako „pořádání informací“, a odráží její vývoj, který je viditelný v činnosti mezinárodní organizace Interntional Society for Knowledge Organization (ISKO).

Problematika pořádání informací na internetu je velmi široké téma, a tak bylo nezbytné tematicky tuto práci omezit. Kapitola první – *„Specifika internetu pro vyhledávání informací“* popisuje prostředí internetu z hlediska objemu, životnosti a vyhledávání informací. Druhá kapitola *„Přístupy k pořádání informací na internetu*“ je úvodem k metodám, používaným pro pořádání informací na internetu a kapitoly tři až osm. Ty analyzují jednotlivé metody, které jsou pro ilustraci doplněny konkrétními projekty. Existují stovky různých projektů pro pořádání informací v prostředí internetu a jejich výčet nebyl cílem této práce. Seznamy různých projektů, uvedené v jednotlivých kapitolách, mají proto pouze ilustrativní charakter a nelze je považovat za vyčerpávající přehled všech existujících. Hlavním kritériem pro výběr těchto projektů byla jejich volná dostupnost, tj. volný přístup ke konkrétním projektům na internetu (např. tezaury nebo systémy automatizované klasifikace a kategorizace). Především z tohoto důvodu jsou vynechány komerční systémy pro pořádání informací (např. produkty pro automatizovanou klasifikaci ve firmách[1]), o nichž jsou dostupné informace spíše propagačního charakteru.

Některé uváděné projekty (např. Geoviser) mají spíše dokumentační hodnotu a jejich konkrétní aplikace už nemusí být přístupné. V takovém případě je (je-li to možné) odkazováno na články, které o nich pojednávají a vysvětlují jejich koncepci a přínos.

Vzhledem k rozsahu této problematiky jsou vynechány některé nejnovější přístupy k pořádání informací, jež jsou zatím ve fázi experimentování. Jde například o problematiku digitálních knihoven nebo o koncept tzv. sémantického webu, který je značně obsáhlý a vyžádal by si samostatnou práci. Tato témata jsou proto zmíněna pouze v souvislosti s jinými, příbuznými projekty.

Pořádání informací se často prolíná s problematikou tzv. identifikátorů, které můžeme také označit za formu metadat. Pro tematickou ucelenost práce jsou zmíněny pouze takové identifikátory, které kombinují funkci nalezení (identifikace) s obsahovým popisem. Zmíněny jsou tak např. identifikátory URI nebo PURL a naopak vynechány ty, které mají za úkol prosté zajištění přístupu k dokumentům (např. DOI).

Informace v prostředí internetu zastarávají velmi rychle, a je proto možné, že některé projekty již nebudou funkční, nebo zmizí úplně. Přes veškerou snahu zajistit aktualitu a funkčnost všech odkazů a použitých zdrojů, tak některé stránky již nemusí být dohledatelné.

Terminologie k problematice, která je zmíněna v této práci, není zatím v češtině ustálená. V některých případech je tak nutné používat doslovné překlady s uvedením originálního výrazu (např. v části o identifikátorech URI je používán výraz

1 Např. produkty firem *Verity* a *Autonomy*

„podprostory“ – subspaces). Pro metody vyhledávání na internetu se používá v angličtině i češtině několik synonymních výrazů. V této práci jsou používány termíny „vyhledávače“ (search engines) pro metodu hledání informací klíčovými slovy, a „webové katalogy“ (subject directories) pro hierarchicky tvořené seznamy se záznamy dokumentů.

V práci používám termíny „internet“ a „world wide web“ („www“ nebo „web“), jejichž vztah je třeba vyjasnit. Internet je označením počítačové sítě, v níž je world wide web jednou z nabízených služeb. Jde jednoznačně o nejpopulárnější službu, která stále více překrývá zdroje, poskytované ostatními službami (např. webová rozhraní pro přístup ke službě FTP, nebo k e-mailovému účtu), a tak dochází k záměně této služby za celou síť internet. Pokud jsou tedy popisovány metody a projekty pro pořádání informací v rámci internetu, **jedná se ve většině případů o prostředí world wide webu**, což ale neznamená, že zde nejsou zmíněny projekty pořádání informací z jiných služeb internetu.

Vzhledem k multidisciplinárnímu charakteru některých částí této práce mohou být některé termíny užity ve významu, který se v jiných oborech liší. To se týká především části o automatizované klasifikaci a kategorizaci, kde jsou tyto termíny odlišeny na rozdíl od pojetí matematiky a informatiky, ve kterém se tyto termíny do značné míry překrývají.

2 SPECIFIKA INTERNETU PRO POŘÁDÁNÍ INFORMACÍ

Toneme v informacích, ale žízníme po vědění.
John Naisbilt

Internet se v průběhu několika let od svého zpřístupnění veřejnosti stal fenoménem, který lze srovnat snad pouze s objevem knihtisku. Stal se dalším zpodobněním postmoderní doby, kdy se individualismus stal měřítkem všech věcí. Žádné pevné struktury a navigace zde nejsou. Zdálo se, že internet bude konečným řešením informační (publikační) exploze, která roste rok od roku geometrickou řadou.

2.1 PUBLIKAČNÍ EXPLOZE

Důvody současného zahlcení informacemi jsou dané prakticky ideálním stavem pro jejich šíření. Tyto podmínky můžeme shrnout do tří okruhů:

- komunikovatelné znalosti – tj. informace
- poptávka po těchto informacích
- technologie pro šíření informací

První dva body jsou úzce spojeny se samotnou lidskou existencí a touhou poznávat. Předpoklad třetí – technologie pro (masové) šíření informací byla naplněna pravděpodobně kolem roku 1453 Johannem Guttenbergem a jeho vynálezem knihtisku. Na ten pak navázala další technická zlepšení, která zajistila dostupnost informací po technické stránce, a zároveň se stále více zvětšoval i zájem lidí o vzdělání a další informace. Kombinace těchto dvou faktorů - technického řešení a masové poptávky vedly k stále rostoucí publikaci informací v tištěné formě.

Počítačová síť internet se veřejnosti ukazuje v polovině 80. let, kdy informační zahlcení už začíná přinášet první vážné problémy. Hlavním problémem není nedostatek informací, či jejich nedostupnost – je to, zcela absurdně, neschopnost danou informaci najít. Důležitější než samotná tvorba informací se tak stává organizace a uspořádání informací.

O informačním zahlcení se mnoho mluvilo a psalo, nicméně nikdo nebyl schopen tyto informace konkretizovat. V roce 1999 byl ale na University of Berkeley zahájen projekt *„How much information[2]“*, který si kladl za cíl alespoň přibližně vyčíslit objem informací, jež jsou uchovávány různými typy médií. Podle této studie byl v roce 2002 celkový objem nově uchovaných informací v tištěné podobě, na magnetických a optických médiích a na filmu přibližně 5 exabytů[3]. Studie dále odhaduje, že mezi léty 1999 až 2000 každoročně rostl objem nově uchovávaných informací o 30%. Na základě těchto údajů je možné předpokládat, že každým dalším rokem bude tato produkce exponenciálně narůstat, a to jak v tištěné tak i v elektronické formě.

2.2 FRAGMENTACE SOUČASNÉ VĚDY

V roce 1994 uveřejnil Peter Jaenecke článek [JAENECKE, 1994], ve kterém se zamýšlí nad současnou publikační explozí a nad možností předefinování záběru pro věcné pořádání vědeckých informací. I když v době publikování tohoto článku nebyl internet jeho předmětem, lze závěry článku přenést i na toto prostředí. Podle autora přispívá k publikační explozi zejména:

2 http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

3 přibližně 10^{15} bytů

1. Fragmentovanost současné vědy

Současná věda se stává čím dál více fragmentovanou, což ztěžuje orientaci ve specializované problematice. Tím roste i podíl „překladových dokumentů“, které nepřináší nic nového, pouze zprostředkovávají poznatky pro uživatele vně jednoho oboru.

2. Tvorba znalostí v procesu nového poznávání

Existující znalosti jsou pouze materiálem pro budování znalostí nových. Proto podle autora vznikají tzv. „pomocné informace“ (auxillary information), které přináší pouze dílčí poznatky a jsou publikovány jako určitý vedlejší produkt hlavního výzkumu. Po dokončení hlavní linie výzkumu už nejsou potřeba.

3. „Pseudoznalosti“ (pseudoknowledge)

Publikování se často mění pouze ve spojování vzájemně nesourodých poznatků, které nepřináší skutečné poznání, ale pouze „informační smog.“

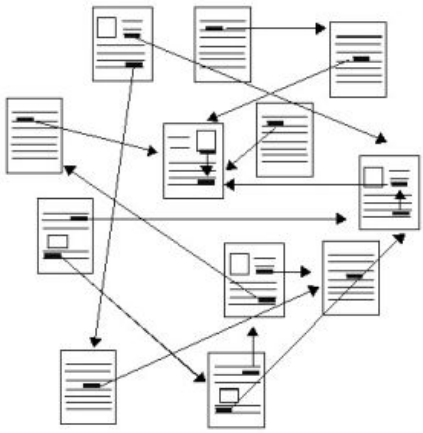
4. Věda přestává být praktickou a vytváří si svůj vlastní svět

Vědecké výsledky často nelze aplikovat do reálného života a celá věda se tak stává výzkumem pro výzkum samotný. Výsledkem jsou i tzv. „vědecké monology“, což je označení pro dokument, který autor uveřejňuje pouze s jediným cílem, aby totiž byl publikován.

Tyto změny se bytostně dotýkají veškerých komunikačních médií – internet nevyjímaje. Nové informace stále přibývají, ale porozumění těmto informacím se ztrácí. Publikování tak paradoxně nepřináší více znalostí, ale více zmatku a nejistoty. Naše společnost tápe a hledá způsob, kterým je možné řešit tento problém – zavedení řádu do komunikačního procesu. Na počátku 90. let se objevila celosvětová počítačová síť internet, ke které se ihned upnuly naděje na konečné řešení pořádání znalostí lidstva.

2.3 SPECIFIKA INTERNETU

Internet je jméno pro počítačovou síť, která nabízí různé služby, ačkoli se mnohdy tento termín zužuje na službu jedinou a sice World Wide Web (www). Důvodem obrovské popularity této služby je jedna z prvních funkčních navigací hypertextem. Tato forma navigace se velmi podobá způsobu lidského myšlení, která se označuje jako volné asociace. Hypertext nabízí velmi malou strukturovanost, avšak poskytuje rozhodující výhodu ve volbě vlastní navigační cesty, jež se ukázala být pro další rozvoj jako rozhodující.



Obrázek 1: Schéma navigace v hypertextu

Hlavním podnětem pro převod publikování na internet (prostřednictvím služby www) je absence jakékoli oficiální kontroly. Patrně poprvé v dějinách se objevilo médium, které postrádá centrální cenzurní mechanismus. Autorem se tak může stát každý, kdo má přístup k internetu. To se projevilo i v masivním nárůstu informací v tomto prostoru. Podle průzkumu společnosti *Netcraft* bylo k 1.6.2004 aktivních[4] celkem 51 635 284 domén (doménových jmen). Stejný průzkum zjistil k 1.11.2006 už 101 435 253 a k 1.8.2008 celkem 176 748 506 existujících domén. Průzkum počítá pouze počet domén druhého řádu (subdomén) s koncovkou *.com (např. http://www.domena.com) a některých dalších TLD – Top Level Domény (uk, mx, bz, sv, gt a mil). Jedná se tedy pouze o část celkového množství domén, ačkoli subdoména doména *.com tvoří většinu všech existujících domén. Tento průzkum nesrovnával počet celkových stránek, ani počet subdomén a reprezentuje tak pouze zjednodušený přístup k vyčíslení celkového počtu stránek na www.

Přesto srovnáme-li pouze data z tohoto průzkumu do číselné řady, zjistíme k jak obrovskému ročnímu nárůstu dochází ve sledovaném období. Během devíti let průzkumu se počet domén zvýšil více než tisícinásobně. Trend obrovského nárůstu počtu dokumentů na internetu potvrzují i další provedené průzkumy.

2.3.1 Souhrn nejdůležitějších zjištění z průzkumů o internetu

Výzkum, který provedl Matthew Gray z MIT,[5] uvádí v červnu roku 1993 pouze 130 domén; pokud tato data přijmeme jako výchozí, narostl jen počet zaregistrovaných domén za posledních sedm let o osm miliónů procent [Grey, 1996].

Vědci v projektu, který vedl Steve Lawrence v roce 1999 [Lawrence-Giles, 1999] dospěli k závěru, že existuje přibližně 800 miliónů stránek, které jsou přístupné indexování (jsou viditelné pro vyhledávače) o datovém objemu přibližně patnáct terabytů. Pouze 6 terabytů (cca 40%) po odstranění formátovacích značek jazyka HTML obsahovalo použitelnou textovou informaci. Stejný výzkum zjistil, že na jednom serveru se nacházelo průměrně 289 stránek. Jejich vyhledatelnost závisí na tom, zda na ně odkazují i jiné stránky.

Významným mezníkem ve studiu internetu a odhadu jeho velikosti byla studie Michaela Bergmana ze společností Bright Planet *nazvaná „The Deep Web: Surfacing the Hidden Value"* [Bergman, 2001]. Studie začala jako první rozdělovat World Wide Web na stránky, které jsou bez větších problémů dohledatelné (tzv. *„viditelný web"*) a stránky, které jsou z různých příčin obtížně dohledatelné (tzv. neviditelný web). Neviditelný web obsahuje velmi cenné informace je dostupný ve formě různých interních stránek, databází, a jiných specializovaných publikací. Michael Bergman velikost tohoto prostoru odhaduje na cca 7 500 Terabytů, což je ve srovnání s odhadem viditelného webu (19 Terabytů) přibližně 400-500x více. Tato studie byla obecně přijata odbornou veřejností, a dnes je možné se k ní odkazovat jako k jednomu ze stěžejních zdrojů.

Čtyřletý projekt společnosti OCLC „*Web Characterization Project"*[Lavoie, 2003] vyčísluje počet samostatných webových prezentací na 3,08 miliónu, což je přibližně 35% všech webových stránek. Podle těchto zdrojů tak byl celkový počet viditelných (tj. obecně přístupných) stránek v roce 2005 cca 1,4 miliardy. Průměrný počet stránek na jednu prezentaci byl 441, což je oproti výzkumu [Lawrence-Giles, 1999] nárůst o 52%.

---

4 tj. servery odpověděly na dotaz na základě HTTP požadavku na identifikaci

5 Massachusetts Institute of Technology

Výzkum z ledna 2005 [GUILLI, 2005] uvádí počet stránek, indexovatelný vyhledavači na minimálně 11,5 miliardy stránek, což by dokládalo výrazné technické zlepšení vyhledatelnosti stránek, a zároveň pokračující exponenciální nárůst počtu webových stránek.

Poslední údaje o počtu stránek se posunuly do vyšších řádů. Za všechny uveďmě alespoň The Official Gogole Blog,[6] který k červnu 2008 uvádí odhad přibližně jednoho bilionu unikátních URL adres.

2.3.2 Životnost informací

Prostředí internetu a digitálních médií paradoxně značně zkrátilo dobu životnosti dokumentů. Klasická média jako papír, film, nebo fotografie mohly být uchovávány po desetiletí, při dobré konzervaci i po staletí ve velmi dobrém stavu. V případě informací, uchovávaných na digitálních médiích se situace drasticky změnila.

Hlavním důvodem je nesmírně rychlý vývoj technologií, který ve svém důsledku komplikuje archivaci dat. Dokladem o rychlosti změn může být i prohlášení předního výrobce počítačů Dell Computers o tom, že od února roku 2004 přestává ve svých počítačích instalovat disketové mechaniky pro 3,5“ diskety [CNN, 2003]. Disketa 3,5“ se stala standardním archivačním a přenosným médiem na začátku 90. let 20. století. Období, kdy byla tato technologie dostačující – „produktivní věk“ trval čtrnáct let. Její starší předchůdce – disketa 5.25“ byla představena společností Shugart Associates v roce 1976 a její produktivní věk končí až v polovině 90. let. Pokud srovnáme jenom tyto dvě technologie archivace pro uchovávání dat, vidíme, že doba použitelnosti každé technologie se zkracuje. Pokud produktivní věk technologie 5,25“ diskety trval přibližně dvacet let, produktivní věk technologie 3,5“ diskety už dosahuje pouze 70% doby životnosti svého předchůdce.

V případě internetu je doba životnosti média, na kterém jsou informace uchovávány velmi důležitá, nikoli však rozhodující. Mnohem závažnější jsou v tomto ohledu změny datových formátů. Základní norma pro publikování dokumentů na internetu – SGML je sice normou poměrně adaptabilní a dosud veškeré konkrétní interpretace tohoto jazyka – HTML, XHTML a XML postačovaly měnícím se požadavkům pro publikování. Základní problém není v technických podmínkách pro publikování informací, ale v použití jiných, nepůvodních datových formátů, které se začínají na internetu objevovat. Především prostředí www začíná být zahlcováno datovými formáty jiného typu, než jsou nativní. Můžeme spatřit i následující důvody, které k tomu vedou:

1. Interaktivita a multimédia

Multimediální formáty Flash, Java, JavaScript a další nabízí atraktivní zábavu, která je důležitá především pro komerčně zaměřené stránky (reklama, marketing, obchod). Některé další informace (např. interaktivní mapy) není možné uveřejnit v ne-interaktivním formátu.

2. Ochrana autorských práv

Dvěma nepsanými standardy se v této oblasti staly jazyk PostScript (*.ps) a Portable Document Format. Dokumenty v obou formátech informace zpřístupňují, ale zároveň poskytují (nedokonalou) obranu proti plagiátorství.

[6] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

3. Neschopnost uživatelů publikovat v nativních datových formátech

Přes jednoduchost základních (tzv. nativních) formátů pro publikování v prostředí www je mnoho uživatelů, kteří nejsou schopni (možná ani ochotni) své dokumenty v těchto formátech zveřejňovat. Obzvláště problematický je převod množství dokumentů ve formátech textového editoru (*.doc, *.wpd, *.rtf apod.), který není pro technickou, časovou a finanční náročnost účelný.

4. Interaktivní výstupy

Problematickou stránkou se jeví popis a organizace interaktivních informací, které jsou zobrazovány na základě konkrétního požadavku uživatele (tzv. on-demand). Každý požadavek je jedinečný a pravděpodobnost jeho opakování klesá s rostoucím množstvím zpřístupněných dat (v datovém souboru – např. databázi). Výše uvedené databáze tak mohou v nativních formátech zveřejňovat pouze odpověď na dotaz uživatele a jen jako dynamicky generovanou stránku, která bude v této podobě zobrazena nejvýše několik minut. Z klasického pohledu na organizaci informací je tento stav špatný, neboť informace není dohledatelná. Na druhé straně ale není technicky možné ani logicky účelné zachycovat informace, které mají význam pro krátký časový úsek (např. kurzovní lístek, předpověď počasí).

Prospěšná stránka věci se objevuje u některých jiných – ne-nativních datových formátů jako např. PDF a PS. Tyto formáty se staly standardem pro multiplatfomní publikování dokumentů a tak jsou používány i pro publikování kvalitních informací z ověřených zdrojů (vládní agentury, univerzity, nevládní organizace apod.). Z tohoto hlediska se formát stal určitou známkou kvality pro informace publikované v tomto prostředí. Tento fakt může pomoci v lepším zaměření dotazu při vyhledávání v prostředí internetu, nicméně vyhledatelnost těchto formátů zůstává nadále komplikovaná[7].

V případě prostředí internetu je problémem i zachytitelnost informace sama o sobě; rychlost „poločasu rozpadu“ informace je dosud nevídaná. Průměrná doba životnosti dokumentu na internetu je podle studie [LAWRENCE-GILES, 1999] přibližně dva roky. Aby tento údaj mohl být vypovídající, je nutné zvážit i povahu publikované informace – některé mizí v řádu minut, jiné jsou dohledatelné i po letech.

2.4 INTERNET JAKO PROSTŘEDÍ PRO POŘÁDÁNÍ INFORMACÍ

2.4.1 Internet jako mýtus

Internet jako nové médium provázela řada velkých očekávání, která nebyla vždy naplněna. Téměř absolutní svoboda publikování na internetu přinesla také řadu závažných problémů, které si mnoho uživatelů ani neuvědomuje. Mezi nimi především:

- kvalita a důvěryhodnost informací a dohledatelnost původu je často velmi obtížná
- organizace informací, resp. jejich nalezitelnost
- životnost informací
- problematika originality, resp. duplikovatelnost informace

Internet se stal novodobým „idolem“, ke kterému se upínají nereálná očekávání. Svým způsobem je internet revolučním prostředkem pro individuální řešení informačních potřeb, nicméně není jediným a univerzálním prostředkem pro získání informace. Nejčastější mýty o internetu můžeme shrnout takto:

7 Tyto typy souborů umí prohledávat například vyhledávače *Adobe PDF Online* nebo *Google*.

„Všechno je na internetu“

Nejrozšířenější mýtus předpokládá, že internet jakoby „pohltil“ veškeré dostupné informace z dalších médií (především tištěných) a stal se tak prakticky univerzálním informačním zdrojem. Tento přístup je postaven na faktu, že internet skutečně může poskytovat alternativní přístup k informacím, původně zveřejněným v jiné formě. Přesto ani tato „duplikace“ zdrojů nezaručuje, že veškeré informace takto budou zveřejňovány (vždy budou limitující další podmínky – ekonomická náročnost, účelnost, autorská práva apod.)

„Stereotypy v přístupu k vyhledávání informací“

Nejpoužívanější přístup k vyhledávání informací má následující scénář: uživatel otevře stránku vyhledávače, který má v čase hledání nejlepší reputaci, do okénka dotazu zadá klíčové slovo nebo frázi a čeká na výsledek. Na základě výsledku má dojem, že získal úplný přehled o hledané tématice.

Tento přístup je z velké části výsledkem opomíjení informační přípravy uživatelů internetu. Zvykli jsme si na jednoduchá řešení, která již nevyžadují další námahu, a tak kvalita získané informace odpovídá času vynaloženém na její získání. Změna tohoto přístupu pravděpodobně nebude nikdy možná, neboť stále dokonalejší systémy pro hodnocení stránek vyhledávači řeší tento problém místo uživatelů.

„Informace na internetu jsou pravdivé a ověřené“

V éře tištěných médií byl skoro vždy dohledatelný zdroj informací. Toto bylo možné pouze díky formě publikování – při vydávání fungoval systém určitého „filtrování informací“ – publikovat nemohl nikdo jiný než člověk s přístupem do této komunity. S nástupem internetu může publikovat každý a takto zveřejňované informace se liší kvalitou i zpracováním. Uživatelé bohužel nepředpokládají potřebu nalezené informace ověřovat, a tak se výběr kvalitních a správných informací stává doménou oborových komunit.

2.4.2 Změny účelu komunikace

Prostředí počítačové sítě internet se od svého vzniku změnilo takřka k nepoznání. Původně vojenská, později akademická počítačová síť byla na počátku zaměřena hodně odlišným směrem a tomu odpovídaly i (dnes již nedostačující) technické standardy a protokoly. Zásadní změny můžeme vidět v následujících směrech:

1. změna povahy komunikace
2. otevření sítě široké veřejnosti

změna tzv. „demografie internetu“ (viz strana 10)

Změna povahy komunikace

Hlavním záměrem projektu tvorby této počítačové sítě byla decentralizovaná komunikace. Tomu odpovídala i jedna z prvních nabízených služeb – elektronická pošta (e-mail). Dnes se povaha užití výrazně změnila – e-mail zůstává nadále jednou z nejpopulárnějších služeb, nicméně primárně je internet používán hlavně jako informační zdroj a zábavní médium.

Otevření sítě široké veřejnosti

Internet byl původně velmi přehlednou počítačovou sítí, neboť vzhledem ke strategickému potenciálu této sítě byla otevřena pouze pro potřeby armády, později i vědeckých organizací. Otevření této sítě pro veřejnost znamenalo již výše zmíněnou změnu povahy komunikace a také nárůst publikační aktivity.

Nejvíce internet ovlivnila komerční sféra, která použila internet jako ideální nástroj pro obchod a marketing. Zde je také vysvětlení obrovského nárůstu počtu stránek – podle studie Steva Lawrence a C.Lee Gilse [LAWRENCE-GILES, 1999] **tvoří komerční stránky plných 83% obsahu world wide webu**. Z tohoto údaje je patrné, že klíčovým pro organizování obsahu

na webových stránkách bude kvalitativní vyhodnocení jejich obsahu, na základě kterého je možné omezit výběr informačních zdrojů (určení povahy hledané informace a podle ní směřovat další postup).

Změna tzv. „demografie internetu“

Tím, že byl internet otevřen široké veřejnosti, byl na místě předpoklad o širokém spektru uživatelů a jejich informačních potřeb. O internetu se mluvilo jako o mezinárodní síti, která by mohla naplnit ideál knihovny celosvětových poznatků na jednom místě. Demografie internetu však zcela nenaplnila původní předpoklady a struktura uživatelů odpovídá spíše profilu obyvatel ekonomicky a technologicky rozvinutých zemí než jejich poměrnému geografickému zastoupení.

2.4.3 Demografie internetu

Demografií internetu označujeme soubor informací o lidech, kteří toto médium využívají. Níže uvedený přehled je sestaven z různých pramenů.

Jazyk

Průzkum z roku 1998 [GVU, 1998] uvádí angličtinu jako hlavní jazyk používaným v rámci www (cca 92.2%). S velkým odstupem následovaly další jazyky – němčina (1.5%), francouzštinou (0.8%) a dánštinou (0.8%) .

Web Characterization Project“ [LAVOIE, 2003] uvádí v roce 1999 jako vůdčí jazyk angličtinu se 72%, následovaný němčinou (7%), francouštinou, japonštinou a španělštinou (po 3%). Stejný průzkum v roce 2002 potvrdil vůdčí roli angličtiny (72%), následovanou němčinou (7%). V dalším pořadí došlo k výraznému nárůstu japonštiny na 6% (tj. za čtyři roky zdvojnásobení objemu stránek) a stabilní podíl francouštiny a španělštiny (po 3%).

Rasové složení

Většina uživatelů je bílé rasy (87.2%), další etnické skupiny jsou zastoupeny mnohem méně – Afroameričané (3.8%) a další pod kategorií „ostatní“ (1.7%) [GVU, 1998].

Internet Population 2004	934 million
*Projection for 2005	1.07 billion
*Projection for 2006	1.21 billion
*Projection for 2007	1.35 billion

***Pozn.** Odhady do budoucnosti

***Tabulka 1:** Počet uživatelů internetu Zdroj: ClickZ Stats staff, 2005*

Geografické rozložení uživatelů

Celkový počet uživatelů internetu se nedá přesně určit, existují pouze odhady, založené na různých uživatelských průzkumech. Pro přibližnou představu odhadu počtu uživatelů internetu na světě je zajímavé uvést odhad renomované auditorské společnosti z oblasti výpočetní techniky Computer Industry Almanac (viz tabulka č.1). Tyto odhady jsou ilustrační a další zdroje s nimi nemusí korespondovat. V roce 2006 společnost Computer Industry Almanac potvrdila, že v roce 2005 počet uživatelů internetu vzrostl přes jednu miliardu [COMPUTER INDUSTRY ALMANACH, 2006]. Stejná zpráva uvádí, že dosažení dvou miliard uživatelů lze očekávat kolem roku 2011.

Za podrobnější rozbor stojí geografické rozložení uživatelů internetu. Jejich rozložení podle kontinentů ukazuje relativně nová statistika *International Telecommunication Union (ITU)* z roku 2007 (viz tabulka č. 2). V těchto údajích se ukazuje významný posun od roku 2000. Podle tehdejších průzkumů [COMPUTER INDUSTRY ALMANAC, 2001] tvořili obyvatelé USA v roce 2000 téměř 33% uživatelů internetu. Podle uvedené tabulky ITU (č. 2) je nyní tento poměr „pouze" 17,39%. Přesto je to nejvyšší procentuální i reálný (159 000 000) počet uživatelů. Na druhém místě je podle ITU Čína – 11,7%, což potvrzuje trend, zjištěný výzkumem Nielsen/Netrankings [NIELSEN/NETRANKINGS, 24.3.2004] (odhad 56,6 milionů uživatelů). Poslední dostupné údaje [COMPUTER INDUSTRY ALMANACH, 2006] (viz. tabulka) s drobnými odchylkami potvrzují průzkum ITU. Počet uživatelů v USA mírně vzrostl, zatímco v Číně mírně poklesnul. Tyto výkyvy však nijak výrazně neovlivňují podíl na celkovém počtu uživatelů.

Z hlediska pořádání informací je tento poznatek nesmírně důležitý, neboť objasňuje rozvrstvení uživatelů internetu a dává do kontextu i nabídky různých informačních služeb. Jejich nabídka pořád do značné míry odráží dominanci amerických uživatelů, kteří tvořili kolem roku 2000 přes čtyřicet procent uživatelů celkem. Proto existuje mnoho vyhledávacích a pořádacích systémů (portály, vyhledávače, akademické projekty), které odráží tuto kulturu. Tento fakt se odráží i ve formě nabídek, navigace, případně jiných – převzatých pořádacích soustav (např. adaptace Deweyho desetinného třídění nebo Library of Congress Subject Headings). Vzhledem k velikosti těchto projektů je nutné vnímat je jako projekty s „celosvětovým" zaměřením, ačkoli se takto původně neprofilovaly.

Příjem

Dalším významným faktorem, který určuje zájmy uživatelů je jejich příjem. Tyto údaje jsou většinou velmi citlivé a proto existují pouze průzkumy menšího rozsahu (menší počet respondentů), což se může projevit i na vypovídací hodnotě těchto dat.

Průzkum [GVU, 1998] z roku 1998 ukázal, že dominuje skupina s vysokými příjmy. Průzkum zahrnoval celkem 5022 respondentů, z nichž 17.3% na tuto otázku neodpovědělo. Ze zbývajícího vzorku vyplývá, že průměrný příjem je 57 300 amerických dolarů, což je více, než bylo zjištěno při výzkumech v letech 1996 a 1997. Respondenti starší padesáti let vykazují vyšší příjmy ve srovnání s dalšími věkovými skupinami.

Entire Sample	All	57.3	Gender	Female	55.1

	11- 20 yrs	53.2			
	21- 25 yrs	41.5	Location	USA	58.7
	26 - 50 yrs	59.1		Europe	47.6
	50+ yrs	62.8		Other	51.0

Tabulka 2: *Průměrný příjem domácností podle kategorií (v tis. dolarů)* *Zdroj: GVU, 1998*

Podle tohoto průzkumu vypadá průměrný uživatel internetu jako člen vyšší střední vrstvy, který je bohatý a vzdělaný. Další průzkum však ukazuje, že tento trend byl patrně jen dočasný a počet uživatelů s nižšími příjmy bude narůstat na úkor této příjmové skupiny.

Nielsen/Netrankings v roce 2001 uveřejnili svou analýzu o tom, že uživatelé internetu s nízkými příjmy představují (alespoň ve Spojených státech amerických) nejrychleji rostoucí skupinu. Podle této analýzy narostla skupina uživatelů internetu v USA s příjmem nižším než 25 000 dolarů ročně od roku 2000 o celých 46%.

Prudký nárůst uživatelů v nižších přijmových vrstvách dokumentuje také průzkum projektu *Pew Internet & American Life Project* [MADDEN, 2006], která uvádí, že k dubnu 2006 je k internetu připojeno cca 73% americké populace (cca 147 milionů), ve srovnání s 66% (cca 133 millonů), zjištěnými v průzkumu k lednu 2005.

Příjmová skupina příjem domácnosti v USD za rok	Uživatelů k únoru 2000	Uživatelů k únoru 2001	Procentní nárůst
0 – 24,999	4333000	6336000	46,00%
25,000 – 49,999	18777000	26363000	40,00%
50,000 – 74,999	21372000	30426000	42,00%
75,000 – 99,999	12301000	16241000	32,00%
100,000 – 149,999	7800000	10448000	34,00%
150,000 – 999,999	3751000	4804000	28,00%

Tabulka 3: *Příjmové skupiny uživatelů v USA a jejich přístup k internetu* *Zdroj: NIELSEN/NETRANKINGS, 13 MARCH 2001*

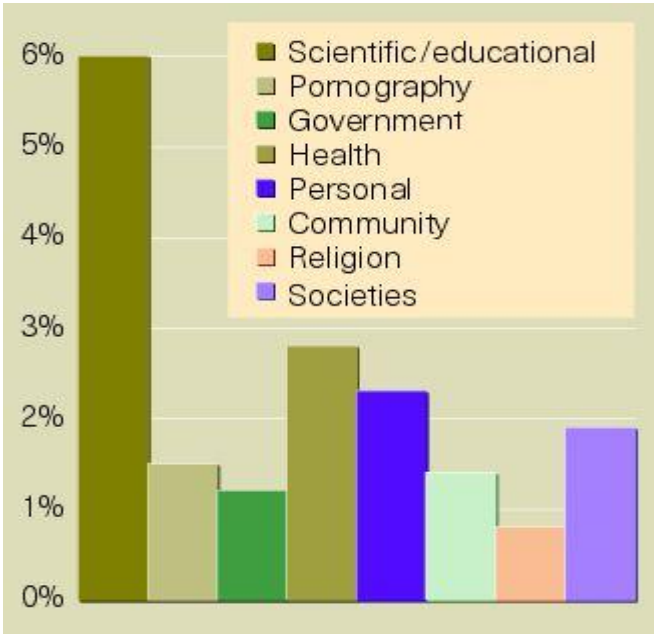
Zda lze vztáhnout výsledky této analýzy i na další neamerické uživatele je sporné. Dostupnost internetu můžeme přičítat sociálním programům a specifickému ekonomickému prostředí v USA, nicméně tento nárůst lze vidět i jako všeobecné rozšíření této technologie díky klesajícím cenám potřebného vybavení a služeb pro přístup k internetu.

Jestliže lze vztáhnout tyto výsledky obecně na celou populaci uživatelů internetu, pak zjistíme, že z původně úzce uživatelsky profilované technologie se stalo masové médium. Z toho vyplývá, že veškeré informační potřeby se příliš nebudou lišit od těch, které uživatelé měli již před nástupem internetu, pouze s drobnými specifiky tohoto prostředí (anonymita, celosvětový dosah média apod.).

2.5 CO ORGANIZOVAT: PROBLEMATIKA KVALITY A ZÁBĚRU INFORMACÍ

Internet je prostředí, které již nelze dostat pod úplnou kontrolu. Je to systém, který se vyvíjí samovolně a proto nelze předpokládat ani vstupní zpracování dokumentů. V této souvislosti je velmi aktuální otázka pořádání informací jako formy cenzury nebo filtrování informací. Cenzura v tomto významu nemá znamenat nic apriori negativního, jde spíše o oddělení dokumentů s mizivou informační hodnotou ze skupiny určené pro pořádání a věcné zpracování.

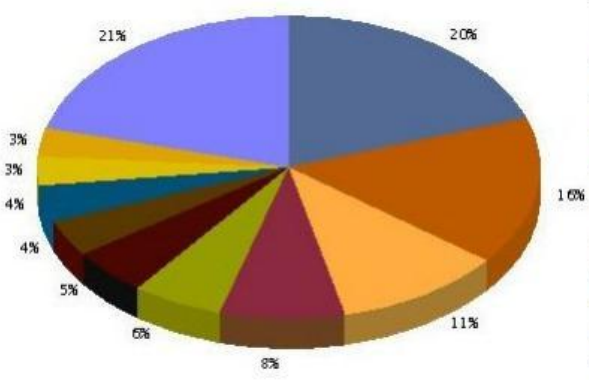
Podle několika průzkumů má určitou vypovídací schopnost typologie typů a formátů dokumentů. Nejvýznamnější studií je v tomto případě výzkum Steva Lawrence a Lee Gilse, [LAWRENCE-GILES, 1999] (viz část 1.4.2.). Z výzkumu vyplývá, že celých 83% informací na webových stránkách je publikováno s komerčním záměrem – jedná se o stránky s prezentací firem, stránky propagační nebo přímo o součást prodejního či distribučního systému (on-line obchod apod.). Toto zjištění jen potvrzuje, že internet se stal spíše marketingovým nástrojem než celosvětovou knihovnou poznatků, jak to



Obrázek 2: *Typologie informací podle S. Lawrence Giles, 1999* *Zdroj: Lawrence-*

jeho tvůrci původně zamýšleli.

Tematické rozdělení zbylých 17% webových stránek je uvedeno v grafu na obrázku č. 3.



Genre	Impressions
Email	3,862,593,000
Portals & Search Engines	3,128,558,000
General\National News	2,135,096,000
Finance	1,569,899,000
General Community	1,101,674,000
Sports & Recreation	935,223,000
Local\Regional	818,757,000
Games	716,796,000
Entertainment	649,664,000
Weather	546,694,000
Other	4,107,310,000
Total	19,572,264,000

Obrázek 3: *Využívání různých typů služeb, přístupných přes world wide web v týdnu 10 -16.1. 2005*

Zdroj: Nielsen/Netrankings (http://direct.www.nielsen-netratings.com/news.jsp?section=dat_an)

Legenda: 6% věda, 1% sex, 1% státní správa, 3% zdraví, 2,5% osobní stránky, 3,5% komunity, náboženské skupiny, spolky.

Jak ukazuje tento graf, internet se stává odrazem lidských zájmů a potřeb. Z tohoto hlediska je všeobecná klasifikace internetu věcí velmi problematickou, neboť se bude opět jednat o univerzální filozofické vidění světa, které řešily již klasifikace a třídění před stovkami let.

Možnou nápovědu poskytuje graf společnosti Nielsen/Netrankings (obr. č.1.4), který pravidelně monitoruje využívání různých typů služeb, přístupných přes www. **Přibližně polovinu všech zobrazených stránek představuje využívání služeb e-mailu, portálů a vyhledávačů a zpravodajství**. Znamená to, že důvody pro využívání internetu můžeme rozdělit do skupin:

komunikace

informační služby

obchod a služby

zábava

Na základě tohoto výzkumu se nabízí otázka, co vlastně pořádat? Možnou odpovědí je pořádat pouze informace z akademického prostředí, které by měly mít největší informační hodnotu. To ale řeší informační potřeby pouze pro úzkou část uživatelů a zhodnotit kvalitativní rozdíl mezi informací o příspěvku na vědecké konferenci a údaji o odjezdu vlaku není objektivně možné.

Problémem internetu je to, že často lze nalézt informace, ale nikoli jejich smysl, se kterým byly publikovány. Je otázkou, kolik procent dokumentů, publikovaných v rámci internetu (především webových stránek), přináší skutečně originální obsah. Určitě však můžeme tvrdit, že převážná většina dokumentů je několikanásobně duplikovaná.

Zdá se, že objektivně nelze omezit pořádání informací na internetu pouze na některou úzce vymezenou oblast. Hodnota informace je svázána s individuálním kulturním prostředím a osobním horizontem poznání. Ewald Kiel [KIEL, 1994] k této problematice uvádí: *„Náš zájem o poznatky/informace závisí na naší omezené kapacitě a na našich potřebách a zájmech v (informačně) problémové situaci.“* Autor zároveň upozorňuje i na neoddělitelnost znalostí od naší osobnosti a kulturního pozadí. Z toho vyplývá, že třídění informací má vycházet vstříc individuálním informačním potřebám, které nelze zcela předvídat. Ke své argumentaci uvádí Ewald Kiel i závěry britského sociologa Thompsona, který zdůrazňuje, že hodnota, kterou jednotlivým předmětům připisujeme, je společenským procesem, a nelze proto objektivně význam informace zhodnotit.

Pořádání informací přichází vždy pozdě – jeho cílem není rozhodovat, která informace je důležitá a která ne, ale zvolit systém pro jejich organizaci. Zdá se tedy, že otázka kvality informace je z důvodu odlišných informačních potřeb velmi těžko měřitelná mezi tématy a obory, ale velmi dobře srovnatelná v rámci jednoho tématu. Tím se celá problematika přesouvá k jiné otázce – jakým způsobem omezit prostor pro pořádání? Prvním přístupem je široký tematický záběr, ale velmi malá podrobnost témat. Druhým přístupem je pořádání úzce specializovaného tématu bez jeho zařazení do širšího kontextu. Dostáváme se tedy opět ke klasické otázce věcného pořádání – do jaké míry podrobností mají být dokumenty pořádány?

Odpověď na tuto otázku přináší patrně Ingetraut Dahlberg [DAHLBERG, 1974], která nabízí řešení tzv. **dvojího pořádání** – jedna soustava pořádání zachycuje vztahy témat mezi sebou, druhá soustava podrobně organizuje dokumenty v rámci jednoho, často úzce specializovaného tématu. Zdá se, že tento model je nezbytný i pro pořádání informací na internetu – již v současné době existují obecné rozcestníky (*Yahoo, Open Directory*) společně s úzce vymezenými soustavami zaměřenými na odborné informace (digitální knihovny, tematicky zaměřené servery apod.).

3 PŘÍSTUPY K POŘÁDÁNÍ INFORMACÍ NA INTERNETU

Uživatelé hledají řád a organizaci v prostředí chaosu

a dynamicky se měnících informací

David Eichmenn

V okamžiku zpřístupnění počítačové sítě internet veřejnosti si patrně málokdo uvědomil potřebu organizace znalostí v tomto prostředí. Zprvu se zdálo, že toto prostředí bude schopno samoregulace a že objem dokumentů nebude veliký. Proto ani navigační infrastruktuře nebyla věnována tak veliká pozornost jako technickým a bezpečnostním protokolům.

Myšlenka na globální informační infrastrukturu na internetu se objevila se zrodem konceptu komplexních identifikátorů URI (URL) (viz kapitola 4.9.1, s.27). Tento koncept navrhoval velmi zajímavé a pragmatické řešení identifikace a obsahové charakteristiky dostupných dokumentů (především webových stránek) na internetu. Návrh tohoto konceptu nebyl zatím přijat a v současnosti funguje pouze jeho část – identifikátor Uniform Resource Locator (URL), který ovšem plní spíše roli identifikátoru. Jeho použití pro obsahovou charakteristiku závisí na rozboru jeho adresové cesty a nemůže tak v žádném případě dostačovat pro účel pořádání informací.

Kvůli neexistující globální navigační infrastruktuře se proto objevují náhradní řešení, která se snaží poskytnout přehled existujících dokumentů všem uživatelům internetu. Používané přístupy představují pouze částečné řešení v tomto složitém prostředí, které má vlastní, často málo známá pravidla a zákonitosti vývoje. Internet je platforma, fungující na stejných principech komunikace jako reálný fyzický svět. Komunikační proces zde zahrnuje tři složky: autora, zprostředkovatele a uživatele.

Metody, užívané k pořádání informací na internetu můžeme rozdělit několika způsoby:

a) podle komunikačních etap

b) podle stáří koncepce

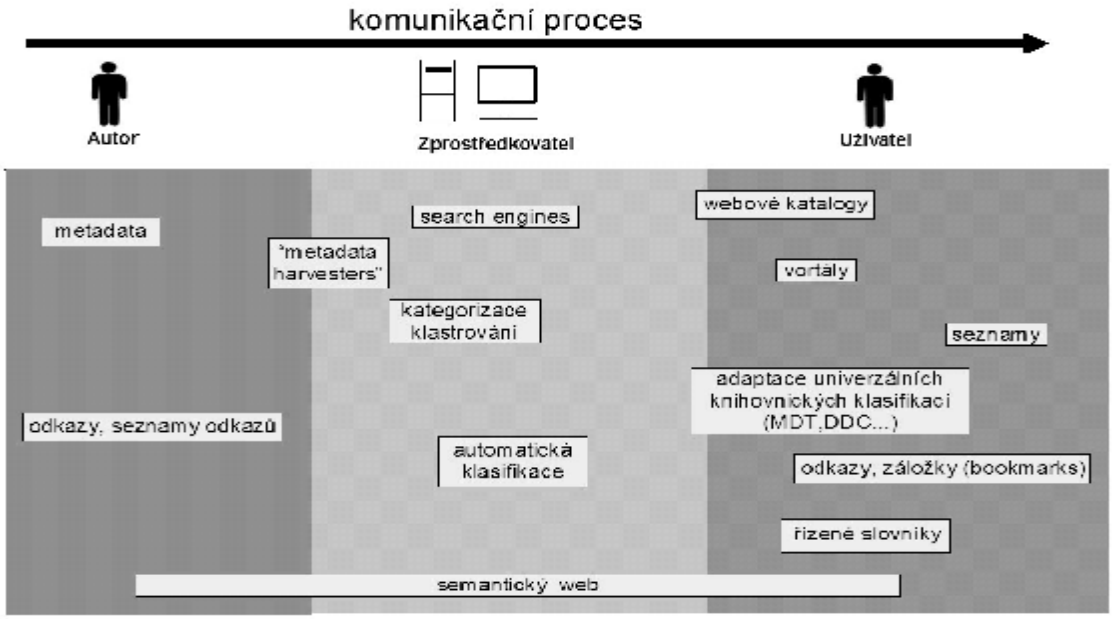
podle vzorů lidského chování při vyhledávání informací

3.1 ROZDĚLENÍ PODLE KOMUNIKAČNÍCH ETAP

Obrázek č. 2.1 ukazuje metody používané k pořádání informací na internetu, rozdělené podle etapy, ve které jsou informace pořádány. V tomto rozdělení metod pořádání informací podle etap komunikace často nelze pevně stanovit hranice, kterými by bylo možné určit, kam která metoda přesně patří. Některé z nich probíhají několika etapami komunikačního procesu, sémantický web – metoda, která je zatím ve fázi experimentování, zamýšlí dokonce zapojení všech účastníků komunikačního procesu.

Tvůrce dokumentu – autor může uživatele navigovat pomocí správného popisu svých dokumentů doplňkovými informacemi – metadaty, případně dalšími (bibliografickými) seznamy a odkazy, které uživatele mohou uvádět do kontextu.

Druhým krokem komunikačního procesu je navigace a vyhledání informací řešeny zpravidla za pomoci automatického zprostředkovatelského systému. Systémy fungují pouze s minimálními nároky na práci člověka či přímo bez jeho zapojení do procesu vyhledání, zpracování a zařazení. Tato skupina zprostředkovatelských systémů je téměř výhradně řešena pomocí metod informatiky, počítačové lingvistiky a umělé inteligence.
Nejrozšířenější metody jsou založené na třetím, konečném prvku celého komunikačního procesu – na uživateli.



***Obrázek 4: Metody pořádání informací na internetu rozdělené podle komunikačních** etap*

Uživatelsky orientované systémy pro pořádání informací jsou nejrozšířenější neboť vždy zohledňují konkrétního uživatele, případně skupinu uživatelů a jejich informační potřeby. Neexistuje pouze jediná, univerzální skupina uživatelů, a proto vzniklo množství pořádacích systémů, které jsou zamýšleny pro různé skupiny a různé potřeby.

3.2 ROZDĚLENÍ PODLE STÁŘÍ KONCEPCE

Druhou možností je rozdělení metod pořádání informací na adaptace starých metod a na metody nové. Staré metody pořádání existovaly již před vznikem internetu a po jeho zrodu jsou nově adaptovány do tohoto prostředí. Mezi aplikace starých metod pořádání informací patří především úpravy univerzálních knihovnických klasifikací (především Deweyho desetinné třídění – DDC a Mezinárodního desetinného třídění – MDT), specializovaná oborová třídění nebo tezaury.

Nové metody jsou především z oblasti metod informatiky a dalších aplikovaných oborů a k organizaci informací přistupují především z hlediska počítačové lingvistiky a statistického vyhodnocení. Některé nové metody jsou koncepčně příbuzné se starými, i když se nejedná se o jejich přesnou adaptaci (např. webové katalogy).

3.3 ROZDĚLENÍ PODLE VZORŮ LIDSKÉHO CHOVÁNÍ PŘI VYHLEDÁVÁNÍ INFORMACÍ

Všechny používané metody můžeme rozdělit také podle dvou vzorců lidského chování při vyhledávání informací [Dood, 1996] na prohlížení předmětů (subject-oriented) a vyhledávání podle klíčových slov (search-oriented).

Technika **prohlížení předmětů** se dále dělí na dvě podskupiny – na prosté prohlížení předmětů, jak to umožňuje struktura hypertextu (tzv. browsing), a na pokročilé pořádání informací v pevné struktuře ve formě různých webových katalogů (tzv. subject-oriented directories).

Podle definice na stránkách University of Berkeley znamená prosté prohlížení stránek (browsing) *„následovat odkazy na stránce, potulovat se po ní a zkoumat její obsah“*
[Barker, 2004]. Tuto techniku můžeme označit jako povrchní, nahodilé čtení a pohyb mezi jednotlivými stránkami. Alespoň na počátku byla tato technika pravděpodobně nejrozšířenějším způsobem navigace na webových stránkách. Jde o technologii nejjednodušší, úzce propojenou se stejně prostou metodou pořádání informací na internetu – se seznamy odkazů.

Pokročilé pořádání informací do struktur se objevuje jak v podobě webových katalogů stránek (viz. kapitola 6, s. 73), tak i v podobě adaptací starších univerzálních hierarchických třídění (viz. kapitola 5.3, s. 52). Tyto systémy uvádějí veškeré informace do kontextu a uživatel tak objevuje nové vztahy a struktury témat. To „může vést k novému vědomí neočekávaných nebo nových cest k obsahu v tomto prostoru“ [Dood, 1996] a prohlížení předmětů ve struktuře tak získává ještě větší význam pro orientaci a pochopení informace.

Vyhledávání podle klíčových slov je cílené a úzce zaměřené. Používá se především při pořádání odborných dokumentů, kde je pro práci s odbornými informacemi specializovaná terminologie nezbytná. Hlavním předpokladem úspěchu této metody je kladení dotazů. Většina chyb je způsobena nedostatečnou schopností uživatelů pracovat se syntaktickými možnostmi dotazovacího systému, nesprávnou rešeršní strategií nebo nedostatečnou znalostí odborné terminologie (většinou jsou pokládány nevhodné dotazy či příliš úzké termíny) – viz. kapitola 9.6, s. 129.

Každý z uvedených přístupů je pro vyhledávání efektivní:„Bylo zjištěno, že pro známá témata je optimální jejich přímé vyhledávání. Pro předmětově orientované dotazy, kde konkrétní témata nejsou známa, je ale často vyhledávání méně efektivní než prohlížení předmětů“ [Dood, 1996, s. 281]. Úspěšnost těchto metod závisí na jejich použití které obecně vychází z uvedeného citátu – známe-li odbornou terminologii, rychlejší cestou bude přímo vyhledat odpověď na konkrétní problém. Pokud se v problematice neorientujeme, znamená prohlížení témat ve struktuře uvedení do kontextu a doplnění znalostí hledajícího člověka.

Ani v případě pořádání informací na internetu nelze zatím očekávat integraci stávajících metod do jediné, univerzální, která by vyhovovala všem uživatelům. Nejnovější koncept sémantického webu se této myšlence blíží, nicméně se zatím jedná pouze o návrh, jehož budoucí realizace si vyžádá ještě dlouhý čas a u kterého není zaručené praktické uplatnění. Pravděpodobně je tak možné očekávat stejný scénář, jaký se objevil v případě tištěných dokumentů – různé metody pořádání pro různé skupiny uživatelů.

4 METADATA

Implementace metadat zvyšuje hodnotu popisovaných dat možností sdílet je v čase a prostoru[8]

4.1 ÚVOD DO PROBLEMATIKY

Základní definice metadat, která ovšem není příliš přínosná, zní *„data o datech“* nebo *„informace o informaci“*. Jedná se o termín, který je používán v mnoha oborových komunitách, kde se jeho význam může lišit. Patrně nejvýstižnější definicí pro přístup z hlediska organizace informací napsal Gail Hodge [HODGE, 2001]: *„Metadata jsou strukturované informace, které popisují, vysvětlují, identifikují umístění nebo jinak usnadňují vyhledání, použití nebo organizaci informačních zdrojů. Metadata jsou často nazývána daty o datech nebo informacemi o informacích.“*

8 http://www.fgdc.gov/publications/documents/metadata/metabroc.html

V komunitě knihovníků, která se problematikou metadat zabývá nejvíce, se termín *„metadata"* používá pro popis formálního schématu, určeného k popisu zdrojů bez ohledu na typ objektu. V užším slova smyslu se tento termín používá speciálně pro popis digitálních (nebo digitalizovaných) objektů. Metadata jsou primárně vytvořena jako strojem čitelné informace pro vyhledávače, nicméně jsou upravena tak, aby je byl schopen číst a interpretovat i člověk.

Mimo knihovnickou komunitu má tento termín daleko širší význam – metadata nás obklopují i v běžném životě, kde je ovšem pod tímto označením neznáme. Jako metadata můžeme například označit čárový kód, rodné číslo nebo identifikační číslo organizace – všechny tyto údaje popisují objekt a podávají o něm další informace. V odbornějším významu můžeme jako metadata označit informace o struktuře textu v jazycích na bázi normy SGML – HTML a XML.

Potřeba zorientovat se v různých typech metadat vedla k zahájení projektu **MetaMap**. James Turner a Veronique Moal se v tomto projektu snaží podat přehled o všech významných standardech metadat v grafické formě. Pomocí speciálního prohlížeče je dostupná tzv. *„subway map"* [TURNER-MOAL, 2003], která ukazuje jednotlivé standardy a jejich vzájemné vztahy. Mapa je organizována podle tematických linií, které přibližně dělí standardy podle jejich účelu, formátu popisovaného objektu a typu zájmové skupiny do následujících větví, které se navzájem prolínají:

Tematické linie pro rozdělení typů metadat

Účel standardu metadat

Tvorba

Podpora v procesu vytváření dokumentu

(**Příklad** SGML – Standard Generalized Markup Language)

Organizace

Zachovává organizační schema uchovávaných objektů. Může podávat

informace o hierarchických vazbách a dalších informačních odkazech.

(**Příklad** MESH – Medical Subject Headings)

Dlouhodobé uchování

Uložení a dlouhodobé uchování materiálu a zabezpečení přístupu k němu včetně otázek počítačové bezpečnosti.

(**Příklad** PANDORA Preserving and Accessing Networked Documentary Resources of Australia)

Rozšiřování

Hlavním účelem této skupiny standardů je zajistit dostupnost dat. Standardy této skupiny se zabývají především navigací, organizací strojem čitelných informací a dalšími – především technicky orientovanými otázkami.

(**Příklad** XPATH – XML Path Language)

Zájmové skupiny (standardy, které jsou vyvíjeny na základě potřeb oborových komunit)

Knihovny

Organizace informací, podpora vyhledávání včetně tvorby pomůcek a informace pro ověřování dokumentů.

(**Příklad** DLI – Digital Libraries Initiative)

Organizace

Zájmová sdružení, která vyvíjejí standardy pro vlastní potřeby. Příkladem mohou být specifické kódované informace – International Standard Recording Code (ISRC) Mezinárodní federace hudebního průmyslu pro identifikaci nahrávek.

(**Příklad** CURL – Consortium of University Research Libraries)

Archivy

Zájmem archivů je dlouhodobé uchování objektu, včetně dalších informací ve formátu, který bude čitelný/převoditelný i v budoucnu.

(**Příklad** RAD – Rules for Archival Description)

Muzea

Popis a uchování neobvyklých objektů, které jsou v jejich sbírkách. Ty mohou zahrnovat 3-D objekty, multimedia apod. Cílem není uchovat přímo fyzický objekt, ale kompletní informace o něm.

(**Příklad** CHIO – Cultural Heritage Information Online)

Formát popisovaného objektu

Text (**Příklad** OLIF - Open Lexicon Interchange Format)

Obrázky (**Příklad** JPEG - Joint Photographic Expert Group)

Video (**Příklad** MPEG - Moving Pictures Expert Group)

Zvuk (**Příklad** SpeechML - Speech Markup Language)

Tento přehled zdůrazňuje vzájemnou provázanost jednotlivých tematických větví. Tyto aspekty jsou také základním východiskem pro vnímání metadat z pohledu organizace informací.

4.2 TYPY METADAT

Každý ze současných standardů má svou specifickou roli v procesu zpracování (popisu, uchování a vyhledání) objektu. Tato metadata můžeme dělit podle typu – tj. účelu, který plní.

Deskriptivní metadata

Nejrozšířenější skupina metadat má za cíl popsat cílový objekt. Popis má zpravidla část formální (obecný popis objektu) a část obsahovou (popis obsahu). Formální část popisu zachycuje informace o fyzických rozměrech, velikosti, tvůrci originálního objektu, názvu apod. Obsahová část se snaží popsat významové vazby k dalším objektům (*relations*) a doplnit záznam o obsahový popis – anotaci nebo různá předmětová hesla, která mají v budoucnu sloužit pro vyhledání objektu.

Administrativní metadata

Do této skupiny patří údaje technického a administrativního charakteru – informace o zpracovateli záznamu, datu jeho vytvoření apod. Důležitou částí může být i identifikace organizace, kde záznam vznikl, nebo informace o fyzickém uložení popisovaného objektu (označení vlastníka a sbírky).

Určitou podskupinou administrativních metadat jsou metadata o právech, souvisejících s uchovávanými objekty (*right management*). Ta uchovávají informace o tvůrcích, majitelích autorských práv či o vlastníku originálního objektu. Některé standardy jsou vytvářeny hlavně z důvodu identifikace majetkových práv k objektu.

Strukturní metadata

Jednou z významných součástí informace o objektu je jeho strukturování. Strukturní skupina metadat se soustřeďuje na uchování vnitřních vazeb jak v objektu samotném (skladba jednotlivých částí objektu – např. text s obrázky), tak v kontextu většího celku (hierarchie, vazby a odkazy na další dokumenty). Strukturní metadata jsou jednou z nejdůležitějších částí archivních projektů, kde může ztráta informace této povahy znamenat i ztrátu informace samotné.

Většina projektů je vytvořena tak, aby bylo možné kombinovat vlastnosti všech tří výše uvedených typů. Zpravidla se jedná jen o kombinaci dvou typů metadat – deskriptivních a administrativních. Existují i komplexně pojaté projekty, které se snaží vytvořit popisný rámec, který by kombinoval vlastnosti všech tří typů metadat. Častějším přístupem bývá kombinace existujících projektů, u kterých to v rámci interoperability lze.

Takto například lze kombinovat standard pro popis archivních objektů *Electronical Archival Description (EAD)* s projektem *Metadata Encoding and Transmission Standard (METS)*, který řeší i problematiku uložení a manipulace s objekty a poskytuje komplexní infrastrukturu pro digitální archiv. Kombinace standardů se dá ve větším měřítku předpokládat i do budoucna, a proto vzrůstá důležitost vzájemné převoditelnosti jednotlivých datových elementů.

4.3 UMÍSTĚNÍ METADAT

Metadata popisují cílový objekt a jsou na něm nezávislá. Jejich umístění může být ale pevně svázáno s objektem samotným. Pro umístění metadat se používají následující přístupy:

- „zapuštění“ (*wraping*) metadat do binárního kódu digitálního objektu
- uchování metadat odděleně od objektu a vzájemné propojení odkazy

4.3.1 „Zapuštění” metadat do binárního kódu

Veškerá popisná data jsou uchovávána v kódu digitálního objektu – jako například hlavička v obrázku, která popisuje jeho formát a verzi. Postup je možné aplikovat také na textové dokumenty a některé další datové formáty. Tento přístup se z hlediska vyhledávání zdá být nejvhodnějším. Popisná data i objekt samotný jsou neoddělitelné a tím by mělo být zajištěno dohledání objektu na základě „metainformací”, které jsou s tímto objektem svázány.

Další výhodou je aktualizace objektu – proces aktualizace proběhne u zdrojového objektu, a zároveň se může změnit jeho metadatový soubor (je-li to žádoucí). Uchování metadat společně s popisovaným objektem je také technicky náročné a provádí se většinou pomocí specializovaného programu. Z tohoto důvodu se s tímto přístupem nesetkáme tak často, jako s odděleným uchováváním metadat.

4.3.2 Uchování metadat odděleně od objektu

Metadata mohou být uložena odděleně od popisovaného objektu a provázána pouze odkazem. Tento postup je technicky nenáročný, ale zároveň přináší řadu problémů. Tím zásadním je spojení souboru metadat s objektem tak, aby byl tento odkaz pevný a stabilní. V síťovém prostředí je stabilita velmi relativní, a tak snadno může dojít k rozdělení jednoho celku na dva nezávislé objekty – popisná metadata a popisovaný objekt. Dalším problémem je aktualizace obou objektů – v praxi bude vždy údržba těchto oddělených objektů problematická.

Pro vyhledávání je tento způsob uchovávání dat matoucí – popisná metadata mohou být vyhledána, nicméně přístup k objektu samotnému závisí na stabilitě odkazu (jsou možné výpadky serverů nebo jiné hardwarové a softwarové problémy) a na jeho aktuálnosti. Ta se v tomto případě ukazuje jako nejčastější příčina problémů.

4.4 SCHÉMATA METADAT

Každý standard metadat je schématem – tj. soustavou datových elementů, které jsou navzájem propojeny určitými pravidly. Podle definice Gaila Hodge [HODGE, 2001] je schéma metadat: *„soubor metadatových prvků, vytvořený za určitým účelem, například pro popis speciálního typu informačního zdroje.”*

Každý standard se skládá z následujících tří částí, které dohromady tvoří nedílný celek:

Sémantika

Sémantika definuje elementy, které se ve schématu používají, včetně významu každého z nich. Jako ilustraci můžeme použít element „creator“ standardu Dublin Core. Jeho význam je definován jako *„entita, primárně odpovědná za*

vytvoření obsahu zdroje"[9]. Aby se zamezilo možným nedorozuměním, je tato definice pro vložený obsah závazná pro každého, kdo tento standard využívá.

Syntax

Soubor pravidel pro kombinování datových elementů mezi sebou. Pravidla stanovují podmínky pro kombinaci datových prvků a jejich hodnot – například element popisující jazykovou sadu musí být vyplněn. Syntax stanovuje především vztah základních datových elementů a jejich rozšiřujících informací – kvalifikátorů. U každého elementu je definováno, s kterými kvalifikátory může být kombinován, případně další doplňkové informace (opakování jednoho kvalifikátoru několikrát s různými hodnotami, sled jednotlivých kvalifikátorů apod.).

Pravidla pro obsah

Určují zásady pro vlastnosti elementů a jejich hodnot. Pokud použijeme stejný element jako v předchozím případě – element popisující znakovou sadu, tak pravidlo pro obsah může omezovat hodnotu na kód jazyka podle normy ISO639-2. Jiným příkladem může být stanovení podmínky pro psaní jména a příjmení s velkým počátečním písmenem.

Každé schéma se vyvíjí v těchto třech částech. Zpravidla se nejvíce mění sémantická část, kde mohou přibývat další prvky pro přesnější popis, ale změnit se může i syntax nebo pravidla pro obsah – i když to už není tak obvyklé.

Většina současných metadatových schémat je založena na bázi jazyka XML, přesněji řečeno na základě přizpůsobení jeho definičního souboru – *Document Type Definition (DTD)*. DTD popisuje, jakým způsobem bude vytvořen jakýkoli XML dokument, bez ohledu na jeho povahu. Tato definice jako celek pak určuje strukturu a vzhled dokumentu a určuje význam sémantických, syntaktických i obsahových prvků.

4.5 PŘIZPŮSOBENÍ STANDARDŮ METADAT

Každý standard pro metadata naplňuje potřeby určité zájmové skupiny. Přes existenci několika desítek těchto standardů je stále potřebné upravit sadu metadat pro specifické potřeby. Může se jednat o použití několika standardů ve vzájemné kombinaci (popis specifického dokumentu nebo kombinace metadat s různým účelem – vyhledávání, uložení, strukturace dokumentu), nebo o úpravu jednoho standardu uživateli „na míru".

Úpravy a přizpůsobení je možné udělat dvěma způsoby – tzv. *extenzí schématu* nebo vytvořením „profilu" schématu.

9 **Zdroj:** http://dublincore.org/documents/dces/

extenze schématu

Extenze znamená úpravu standardní sady metadat přidáním dalších informací, které mohou detailně popsat obsah dokumentu. Tyto přidané prvky mají nejčastěji podobu tzv. **kvalifikátorů**, které zpřesňují informaci v základním elementu.

Příklad extenze v sadě metadat Dublin Core:

Základní element	Základní kvalifikátor	Extenze kvalifikátoru
Creator (Tvůrce)	Name (Jméno)	Surname (Příjmení)

Tento postup je nejjednodušší, neboť pro přidání dalších základních datových elementů je zapotřebí dosáhnout konsenzu všech institucí, které daný standard vytvářejí. Takto existuje například ve skupině vyvíjející Dublin Core spor mezi zastánci minimální datové sady – tzv. *minimalisty* a tzv. *strukturalisty*, kteří prosazují větší flexibilitu ve vytváření formálních významů prvků a jejich kvalifikátorů [WEIBEL,1997].

vytvoření „profilu" schématu

Profily jsou určitými „podskupinami" schémat, které zpřesňují použití jednotlivých prvků schématu pro potřeby konkrétního uživatele. Většinou se jedná o zúžení sady popisných elementů tak, aby všichni, kdo v komunitě tento standard používají, dodržovali stejnou datovou sadu. Tento přístup je pragmatický – datová sada je zúžena pouze na ty prvky a elementy, které se skutečně užívají, ne na všechny dostupné.

Příklad:

V datové sadě Dublin Core je celkem 16 elementů. Zájmová skupina si může vytvořit profil, ve kterém bude používat pouze pět z nich – název, předmět, popis, formát a datum.

Gail Hodge [HODGE, 2001] dodává, že tyto dva postupy úpravy metadatových schémat jsou často kombinovány tak, aby bylo možné využít výhod obou přístupů.

4.6 INTEROPERABILITA METADAT

4.6.1 RDF Framework

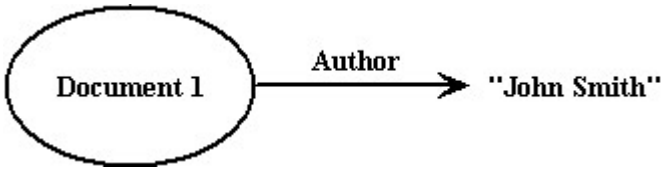
S rostoucím počtem standardů je stále aktuálnější otázka převoditelnosti dat. Se vzrůstajícími investicemi do projektů digitalizace a vytváření archivů je jednou z nejdůležitějších vlastností standardů metadat. Většina standardů je postavena na základě jazyka XML (DTD), což ale nevylučuje komplikace při převodu původních dat do jiného formátu. Z tohoto důvodu vznikly projekty, které měly umožnit vzájemnou převoditelnost datových elementů, známé jako **rámce** (frameworks).

Nejvýznamnějším projektem je **Resource Description Framework (RDF)**, vyvinutý konsorciem World Wide Web (W3C). Tento rámec je výsledkem širšího projektu **PICS: Platform for Internet Content Selection**, který byl zahájen za účelem vytvoření *„specifikace, která by umožnila lidem distribuovat metadata o obsahu digitálního materiálu ve formě „štítků"* (labels) [W3C, 1996A]. Tyto štítky měly vyjadřovat informace o obsahu dokumentu v jednoduché, strojem čitelné podobě.

Resource Description Framework (RDF) je „infrastrukturou, která umožňuje kódování, výměnu a opětovné použití metadat" [Miller, 1998]. Podle podrobnější definice konsorcia W3C 3 [W3C, 1998] je RDF „deklarativní jazyk, který

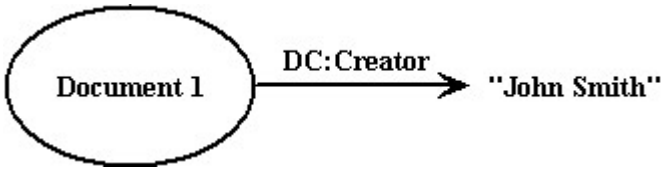
poskytuje standardní způsob pro použití XML k reprezentaci metadat ve formě výroků o hodnotách a vztazích prvků na webu. Tyto prvky, označované jako zdroje mohou mít jakoukoli formu, pokud mají webovou adresu [URI]. To znamená, že lze asociovat metadata s webovou stránkou, grafikou, zvukovým souborem, filmovým klipem atd.“

Základním konceptem RDF je **popis objektu** (resource) **prostřednictvím skupiny vlastností** (property-types), jejichž soubor označujeme jako RDF popis (description). RDF model a specifikace syntaxe reprezentují vztahy mezi zdroji, vlastnostmi a jejich hodnotami.



Obrázek 5: *Jednoduchý RDF model* *Zdroj: MILLER, 1998*

Objekt představuje „Document 1“, který má **vlastnost** „Author“, jejíž **hodnotou** je „John Smith“.



Obrázek 6: *Příklad syntaxe RDF* *Zdroj: MILLER, 1998*

RDF užívá syntax XML, aby mohla definovat sémantiku jiných metadatových modelů. Mezi jednotlivými metadatovými standardy dochází k určitým významovým odlišnostem, a tak například element „AUTHOR“ může být v jednom standardu definován jako fyzická osoba, zatímco v jiném zahrnuje definice i korporaci. Je tedy potřeba přesně definovat význam každého elementu. Proto používá RDF mechanismus tzv. **jmenných prostorů** (namespaces) jazyka XML. Tyto jmenné prostory umožňují jednoznačně určit sémantiku a slovník používaných prvků.

Pro popis objektu (Document 1) je použito jmenného prostoru **DC** (zkratka pro sadu Dublin Core), který **odkazuje na sémantiku a definici všech možných prvků této znakové sady**. Prvek „Creator“ – tvůrce má ve smyslu sady Dublin Core hodnotu „John Smith“ (viz obrázek 3.2).

Ve zdrojovém kódu jsou tyto vztahy zachyceny následovně:

<RDF xmlns=http://www.w3.org/1999/02/22-rdf-syntax-ns#

// specifikace RDF – začátek dokumentu

xmlns:dc="http://purl.org/dc/elements/1.1/">

// specifikace definice použitých „jmenných prostorů“

// metadat“(namespaces).

//V tomto případě se jedná o metadata Dublin Core

<Description about="http://www.w3.org/Press/99Folio.pdf">

// informace o popisovaném objektu

<dc:title>The W3C Folio 1999</dc:title>

<dc:creator>W3C Communications Team</dc:creator>

<dc:date>1999-03-10</dc:date>

<dc:subject>Web development, World Wide Web

Consortium, Interoperability of the Web</dc:subject>

// elementy a hodnoty metadat Dublin Core

</Description>

</RDF>

//konec dokumentu

Zdroj: W3C, 1998

Rámec RDF nabízí velmi silný nástroj, který umožňuje přizpůsobovat popis objektu podle specifických potřeb. Tento koncept se ukázal jako velmi praktický, a proto je od svého vzniku v roce 1997 dále rozvíjen. Jeho další vývoj směřuje k použití v poslední technologii popisu zdrojů v elektronické podobě – k sémantickému webu. Hlavní těžiště zájmu je proto soustředěno na vývoj RDF schémat, která by v rámci sémantického webu byla vyhledatelná specializovanými vyhledávacími roboty.

4.6.2 Mapování metadat do jiného standardu

Množství standardů metadat vzniklo z důvodu odlišných potřeb pro jejich použití. I když RDF poskytuje obecný rámec pro jejich převod a použití, neřeší ještě otázku fyzického datového převodu.

Pro převod dat slouží různé převodníky (metadata crosswalks), které umožňují převod informací z jednoho standardu do jiného. Tento způsob se označuje jako tzv. **mapování**. V praxi probíhá zobrazením obou datových sad v tabulce, kdy jsou proti sobě postaveny datové prvky stejného významu (synonyma).

Categories for the Description of Works of Art	Cataloging Cultural Objects	USMARC	Dublin Core
Object/Work-Type (core)	Work Type	655 Genre-Form	Type
Object/Work-Components		300a Physical Description-Extent	Format.Extent
CLASSIFICATION (core)	Class	050	Subject (classificationschema)
TITLES OR NAMES (core)	Title	24Xa Title and Title-Related	Title

Tabulka 4: Příklad převodníku pro metedata humanitního zaměření Zdroj: Getty, 2000

Jako při každém převodu se i zde vyskytují problémy. Největší z nich je převod méně strukturovaných metadatových elementů do elementů s větší strukturovaností – například převod elementu Dublin Core "TITLE" do katalogizačního formátu MARC s množstvím podpolí.

4.7 PROJEKTY STANDARDIZACE METADAT

4.8 STANDARDIZACE STRUKTURNÍCH PRVKŮ NA BÁZI SGML

4.8.1 SGML a HTML

V roce 1969 Charles Goldfarb, vedoucí výzkumného projektu IBM na integrovaný právní informační systém, společně se svými spolupracovníky Mosherem a Loriem vyvinuli **GML – Generalized Markup Language** (mimochodem název se shoduje s iniciálami zúčastněných pracovníků) [SVOBODA, 1997].

První pracovní verze tohoto jazyka byla zveřejněna v roce 1980 a na základě dalšího úsilí a mezinárodní spolupráce byl tento jazyk v roce 1986 přijat jako mezinárodní standard International Standards Organization pod označením **ISO 8879**. Tento jazyk umožňuje **strukturovat výslednou podobu textu** pomocí párových značek – tagů. Jejich význam je určen definičním souborem, na který se odkazují.

Aplikace SGML se z důvodu složitého strukturování definičního souboru masově nerozšířily. Teprve když byla v květnu roku 1996 představena konkrétní interpretace této normy, zaměřená na vytváření webových stránek – *Hyper Text Markup Language* (HTML), začaly být oba jazyky středem pozornosti. Současně s jazykem se objevily prohlížeče, které uměly pracovat s definičním souborem jazyka a dokumenty v HTML správně zobrazit.

Jazyk HTML ve své verzi 1.0 poprvé umožňuje popsat soubory pomocí značek (tagů) META. Tyto značky umožňují značně omezený popis – většinou pouze popis obsahu stránky (description), jména autora (author) a klíčových slov (keywords).

Deklarace obsahu vypadá následovně:

<META HTTP-EQUIV="name" CONTENT="content">

<META NAME="name" CONTENT="content">

Část „NAME“ definuje popisovaný prvek (klíčová slova, popis, jméno...) a část „CONTENT“ konkrétní obsah prvku – tedy jméno, klíčová slova apod. [W3C, 1996B]. Poslední zveřejněnou verzí jazyka je HTML 4.01, která nabízí větší podporu multimédií, skriptovacích jazyků a užití stylů – tj. doplňkových definičních souborů. [W3C, 1999]

Další vývoj směřuje k hybridnímu jazyku, který stojí na pomezí mezi HTML a XML – **eXtensible Hyper Text Markup Language (XHTML)**. Tento jazyk zahrnuje pouze minimální požadovanou definiční sadu a umožňuje úpravy definičního souboru na míru. Zároveň nabízí širokou podporu pro multimédia a práci s různými objekty (Java, JavaScript apod.). Zmíněný jazyk byl vytvořen také s ohledem na programy – klienty, kteří nepodporují plnou definiční sadu XHTML, jako jsou prohlížeče v mobilních telefonech, PDA nebo pagerech. Pro uvedené mobilní komunikátory se předpokládá užití tohoto jazyka, dokud nebudou hotové sofistikovanější aplikace pro nový jazyk – XML [W3C, 2000] .

4.9 METADATA SPOJENÁ S PROTOKOLEM HTTP A JAZYKEM HTML

Se vznikem prostředí www a technickou realizací protokolu HTTP se v oblasti internetu objevily hybridní identifikátory, které popisují přístup k dokumentu a zároveň poskytují i informace o obsahu dokumentu. Je možné je označit za určitý přechod mezi „čistými“ identifikátory a metadatovým popisem. V současné době existují dva velké projekty, které se snaží pro pořádání informací využít tyto hybridní identifikátory – **Uniform Resource Identifier (URI)** a **Persistent Uniform Resource Identifier (PURL)**.

4.9.1 Koncept Uniform Resource Identifier (URI)

Při tvorbě navigačního systému pro internet byly identifikátory přímo zahrnuty do standardní adresné soustavy. Konsorcium W3C a jeho pracovní skupiny proto vytvořili model adresování, který měl plnit dvě úlohy:

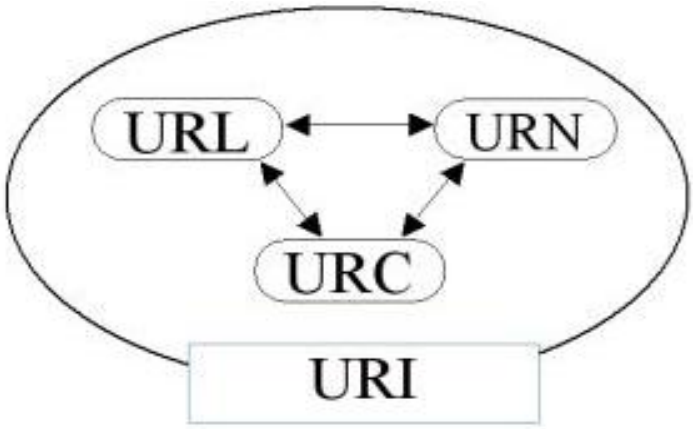
- bezpečně identifikovat cestu k dokumentu
- zajistit popis dokumentu a jeho specifika oproti jiným, podobným dokumentům.

Celý systém funguje pomocí součinností tří prvků – URL, URN a URI. První část – **Uniform Resource Location** (URL) obsahuje informace o přístupu k popisovanému objektu. Tento prvek *„nespecifikuje logický obsah, ale pouze instrukce, jak objektu dosáhnout“* [LYNCH, 1998]. URL obsahuje ve své deklaraci typ služby, kterou je objekt přístupný (služba world wide web jako http://; služba FTP jako ftp:// apod.) a samotnou adresu objektu. Tato forma adresování je používána již od roku 1990 [BERNERS-LEE, 1993].

Dynamika prostředí internetu je natolik veliká, že pouhá informace o přístupu k objektu přestala stačit. Ukázalo se, že objekty často mění své umístění, a tak se informace o přístupu stávají rychle zastaralými. Informaci o přístupu bylo potřeba doplnit jedinečným identifikátorem objektu a jeho popisem. Proto na URL navazují další části: URI a URN.

URN má za cíl přesné pojmenování/označení („persistent labeling“) [ISOC, 1998]. Koncept URN měl fungovat následovně: každý digitální objekt by byl označen „registrační autoritou“, která by spravovala centrální registr jmen. Konkrétní označení (URN) by nemělo jiný význam než unikátní označení objektu (jednoznačné jméno).

Druhým úkolem identifikátorů měla být jednoznačná identifikace objektu v síťovém prostředí (myšlenka podobná současnému systému rozpoznávání adres v prostředí www – systému DNS), který by prostřednictvím databáze identifikoval objekt, přiřadil k němu informace o jeho umístění a tyto informace odeslal uživateli. Toto řešení označuje tvůrce tohoto standardu – *Internet Engineering Task Force* (IETF) jako **„rozlišovače“** (resolvers), případně „rozlišovací databáze“ (resolving databases) [LYNCH, 1998].



Obrázek 7: *Původně zamýšlené schéma vztahu URI – URN – URC – URL*

Úplný model adresování měl tedy fungovat následovně: Každý objekt měl mít vlastní identifikaci – **Uniform Resource Identifier (URI)**, která se skládá z **Uniform Resource Name (URN)** trvalého a jednoznačného jména objektu společně s **URL** – definicí přístupové služby a přesnou adresou objektu. Tento model se ale do praktického využití nikdy nedostal.

Původní myšlenkou (přibližně do poloviny 90. let) bylo rozdělení typů identifikátorů do dvou (či více) úrovní. Identifikátor měl specifikovat buď umístění zdroje (URL) anebo jedinečné jméno zdroje (URN), nezávisle na jeho umístění. Jedinečný identifikátor – URI se tak skládal buď pouze ze jména (URN) anebo z adresy (URL) a neposkytoval tak souhrnné informace o objektu.

Identifikátor URI byl zamýšlen i pro jiné užití, než pro přímý odkaz ke zdroji. Zamýšlenou variantou byl i odkaz na soubor metadat, které by zdroj popisovaly. Pro tento soubor se začalo užívat označení **Uniform Resource Citation (URC)**. Původní koncept zařazoval URC do skupiny s dalšími identifikátory. V praxi se však nikdy neujal a zůstává otázkou, zda tento identifikátor v původní podobě bude někdy implementován.

Původní koncept se v praxi výrazně změnil. Identifikátor URI a původní – nadřazený vztah k dalším identifikátorům již není vnímán jako nejdůležitější. **Dnešní pohled již akceptuje vnímání jednotlivých identifikačních schémat samostatně,** bez toho, aby byly spojeny v jednu ucelenou skupinu. Všechny tyto typy identifikátorů jsou nyní označovány jako URI schémata, která mohou být specifikována podle konkrétního typu – např. URL.

Dokument pracovní skupiny [W3C, 2001] předkládá koncept „podprostorů“ (subspaces) identifikátorů URI, které jsou označovány jako *„prostor pro jméno/pojmenování části“* (namespace). Například identifikátor URN ve formě „urn:isbn:n-nn-nnnnnn-n“[10] je URN podprostor (podčást). Nejedná se ani o „URN schéma“ ani o „URI schéma“. Pro URN podprostory se předpokládá v budoucnu široké využití, neboť by se mohly stát hlavními identifikátory pro internet. Konkrétní užití se již plánuje pro kódy International Standard Book Number (ISBN), International Standard Serial Number (ISSN) nebo pro Digital Object Identifier (DOI).

Uvedený dokument [W3C, 2001] se zabývá i změnou ve vnímání identifikátoru URL. Ten již není vnímán jen jako součást schématu komplexního identifikátoru URI, ale stal se spíše neformálním konceptem identifikátoru, který označuje zdroj pomocí způsobu přístupu k němu (např. http jako označení přístupu pro službu World Wide Web).

Další vývoj identifikátorů je zatím předmětem jednání skupiny *IETF - Uniform Resource Identifiers (URI) Working Group*. Poslední jednání k datu revize tohoto textu (únor 2005) proběhlo v srpnu 2004 v San Diegu v rámci 60. setkání Internet

10 „ISBN“ je označení identifikátoru pro pojmenování identifikátoru.

Engineering Task Force. Na tomto setkání se projednávaly tři body: revize normy RFC 2396, vytváření metadatových schémat a především nový koncept mezinárodního identifikátoru zdrojů – Internationalized Resource Identifiers (IRI).

V lednu 2005 vznikla nová norma RFC 3986 *„Uniform resource Identifier (URI): Generic syntax"*[11], která nahrazuje předchozí normu RFC 2396 (Uniform resource locators) s cílem poskytnout jednoduchou obecnou syntaxi. **Identifikátor URI se tak stává jediným unikátním identifikátorem pro adresování v prostředí internetu**. Tato změna byla ovlivněna především zavedením nové verze komunikačního protokolu IPv6[12]. Ten přinesl do stávající infrastruktury velké změny, které souvisí především s dostatečným rozsahem adresního prostoru (odstraněna hrozba vyčerpání stávajícího počtu IP adres) a tím ovlivňuje také adresování.

Výše zmíněná norma – **Internationalized Resource Identifiers (IRI)** také navazuje na zavedení nové verze protokolu IPv6. Tento nový standard rozšiřuje možnosti použití nových znakových sad v identifikátorech URI. Dosud bylo možné pro adresy zdrojů používat pouze znaky ASCII soustavy, což tato nová norma odstraňuje a pomocí základní znakové sady (ASCII) umožňuje zápis adres zdrojů v různých jiných národních znakových sadách (např. podpora diakritiky, azbuka). Norma IRI by měla nahradit URI například pro účely identifikace protokolu nebo formátu zdrojového objektu. Další budoucnost identifikátoru URI je tak úzce spojena s vývojem kolem protokolu IPv6 a souvisejícími normami.

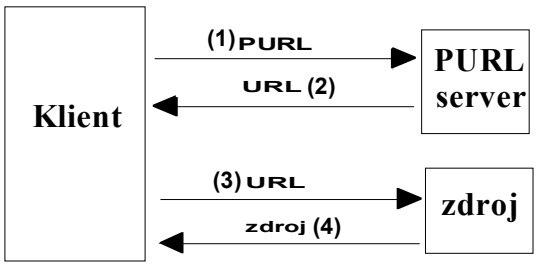
4.9.2 Persistent URL (PURL)

Koncept sdružených identifikátorů – URI, URL, URC a URN nebyl dosud do praxe implementován. Vzhledem k problémům, které přináší dynamika prostředí webu (především změny umístění stránek) začalo být problematické užívání konceptu URL – cesty k dosažení objektu.

Problémy měl v původním konceptu vyřešit další identifikátor – URN, který (jak už je uvedeno výše) měl plnit i funkci jednoznačného identifikátoru v centrální databázi [LYNCH, 1998].

Globální řešení tohoto identifikátoru nebylo v dohledné době reálné a tak společnost OCLC vyvinula v roce 1995 vlastní formu tohoto identifikátoru – **Persistent Uniform Resource Locator (PURL)**. PURL není v podstatě nic jiného než URL. Jeho struktura je viditelná na obrázku č. 3.4.Rozdíl oproti URL je v odkazování na zprostředkující server (resolution service).Zprostředkující server PURL najde k příslušné adrese URL a tu pošle klientovi[WEIBEL, 1995].

Funkci PURL blíže ukazuje obrázek č. 3.5. Klient pošle PURL serveru PURL adresu (1). PURL server vrací klientovi adresu zdroje – URL (2). Na základě této adresy může klient navázat komunikaci se zdrojem (3,4).



Obrázek 8: Funkce PURL Zdroj: http://purl.oclc.org/docs/purl_faq.html

11 http://www.ietf.org

12 Internet Protocol ve

Tento koncept vyžaduje spolupráci tvůrce zdroje, který se musí zaregistrovat v systému PURL, zvolit si jméno svého zdroje (například *„Dublin Core“*) a vložit současnou adresu URL, případně URL upravit na novou adresu. Systém je výhodný pro stránky, které často mění svou adresu – uživatel pak potřebuje znát pouze PURL stránek, aniž by musel hledat adresu novou.

Služba PURL vznikla v Online Computer Library Center (OCLC), kde je také umístěn hlavní zprostředkující server (http://purl.oclc.org). V rámci projektu ale vznikl také podpůrný software, který je možné kdekoli implementovat a vytvořit si vlastní PURL server.

Projekt PURL je pouze přechodným řešením, které má za cíl překlenout dobu, kdy není možné používat koncept URN. Je omezen také technicky a nepředpokládá se, že by mohl fungovat i po plánované úpravě technických standardů pro internet.

4.9.3 xISBN

Jedním z dalších průkopnických projektů, které se snaží prakticky oživit koncept URI je projekt OCLC nazvaný xISBN. Tento projekt se snaží na základě využití kódu ISBN nalézt příslušný bibliografický popis v katalogu OCLC – *WorldCat*. Zároveň prakticky testuje algoritmus OCLC pro model nové generace bibliografického popisu, založeného na deklaraci vztahů objektu k jiným – model **FRBR** (Functional Requirements for Bibliographic Records). Hlavní přínos projektu spočívá v možnosti dotazů na základě identifikačního kódu ISBN – pokud by se tento přístup ujal, k zavedení URI, respektive URN je zapotřebí již „pouze“ technicky tento protokol aplikovat do prohlížečů.

Služba funguje v experimentálním provozu podle následujícího schématu:

http://labs.oclc.org/xisbn/[ISBN]

Na základě vložení kódu ISBN v současné době vrací server OCLC seznam, kde je uveden hledaný kód a další ISBN, které s hledaným dílem souvisí (vícedílná kniha, jazykové mutace apod.). Tento projekt v současnosti nemá příliš veliké praktické uplatnění, ale jakmile by byl propojen do katalogu OCLC, stane se velmi důležitým rešeršním nástrojem. Tento vývoj lze v budoucnu očekávat a je jen otázkou, zda tato služba bude i nadále bezplatně přístupná [OCLC, 2004].

4.9.4 XML

Nejnovějším zástupcem z rodiny jazyků, založených na SGML, je **eXtensible Markup Language (XML)**. Na rozdíl od svých předchůdců není závislý na žádné pevné definici. Formátování dokumentu i definice formátovacích značek (Document Type Definition) je zcela v rukou programátorů.

Jazyk XML je velmi flexibilní a má široké uplatnění. Na bázi XML se tak začala vytvářet i schémata pro metadata. Většina projektů již převádí svoje sady do XML za účelem větší interoperability i vyjadřovacích možností sady samotné (sémantika a syntax). Pro tento jazyk zároveň existuje množství nástrojů, které ulehčují zpracování a validaci dat, a tak zlepšují programování v tomto jazyce.

Hlavní výhodou XML pro pořádání informací jsou tzv. jmenné prostory (namespaces – viz s.28). Jmenné prostory umožňují odlišit jména, používaná v XML dokumentech bez ohledu na jejich původ [TOMAIUOLO, 1996]. Tato definice trochu abstraktně přibližuje podstatu jmenných prostorů – XML struktura umožňuje používat i prvky jiných datových sad, které nejsou v XML definovány.

Příklad podprostoru v XML

```
<xs:schema target xmlns="http://purl.org/dc/elements/1.1/"
xmlns:x="http://www.w3.org/XML/1998/namespace">
<xs:annotation>
<xs:documentation xml:lang="en">
 DC XML Schema, 2004-07-15
</xs:documentation>

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdfsyntaxns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description>
<dc:title>DC XML Schema with DR refinements</dc:title>
<dc:date>2004-07-12</dc:date>
<dc:description>
 Schéma Dublin Core v XML namespace
</dc:description>
```

<dc:relation rdf:resource="http://purl.org/dc/elements/1.1/"/>

</rdf:Description>

</rdf:RDF>

</xs:annotation>

</xs:group>

<xs:complexType name="elementType"> <xs:simpleContent>

<xs:extension base="xs:string"> <xs:attribute ref="x:lang"

use="optional"/> </xs:extension>

</xs:simpleContent> </xs:complexType>

</xs:schema>

V tomto příkladu se vyskytují dvě datové sady – XML (elementy začínající „xs“) a schémata RDF Dublin Core (elementy začínající „rdf“ a „dc“). Do XML dokumentu je vnořen metadatový popis, vytvořený podle standardu Dublin Core pomocí RDF rámce, který data upraví do syntaxe přijatelné pro deklaraci XML. V prvních dvou řádcích jsou informace o definicích elementů – pro interpretaci uvedených elementů se program obrátí na uvedené adresy, kde je obsažena jejich definice a další informace o jejich významu.

Tímto způsobem se jazyk XML může používat jako rámec, který může obsahovat jakékoli části, které jsou kompatibilní s jeho strukturou párových značek. To znamená, že XML může být velmi flexibilně využito pro reprezentaci dokumentů, převodní rámce, ale i pro interpretaci metadat samotných.

U tohoto jazyka se předpokládá další využití především v souvislosti se standardem pro vyhledávání Z.39.50 a jeho nástupci. Další významné uplatnění XML je plánováno v oblasti tématických sítí – v rámci projektu sémantického webu.

4.10 PROJEKTY METADAT ZALOŽENÉ NA JAZYKU XML

V oblasti zaměřené pouze na metadata se HTML a především XML používá v různých projektech, zaměřených na vytváření a směnu metadat. Většina těchto projektů je postavena na definování vlastní datové sady, i když některé z nich (např. PRISM) mohou používat části jiných projektů odlišného typu. Není záměrem jmenovat zde všechny, a proto zde budou uvedeny jako ilustrace jen některé z nich.

4.10.1 Platform for Internet Content Selection (PICS)

PICS vznikl v roce 1996 za účelem filtrování materiálů, ke kterým mají přístup děti. Celý systém je založen na hodnocení materiálu obsahovým štítkem (label), který informuje o obsahu a jeho vhodnosti pro uživatele. Celý projekt byl a nadále je podporován W3C konsorciem. Tento projekt je jediným uvedeným, který od začátku nepoužíval XML, ale v době svého vzniku aktuální jazyk HTML. Teprve později byl pomocí rámce RDF převeden do jazyka XML.

Systém umožňuje producentovi stránek označit štítkem v podobě metadat obsah a distribuovat ho dále s touto informací. Na základě tohoto štítku je hodnotící služba na serveru schopna zařadit stránku do věkové kategorie a vytvořit tak profily zakázaných stránek pro různě staré děti a dospívající [W3C, 2004]. Samotné štítky jsou zakódovány v hlavičce HTML dokumentu jako technická data nebo prostřednictvím následující deklarace META tagu. Část „PICS-Label“ je označení stránky, „labellist“ pak uvádí jméno, případně adresu hodnotící služby – např. hodnocení americké filmové federace MPAA.

Příklad PICS:

<META http-equiv="PICS-Label" content='labellist'>

Dnes je tento systém užíván především pro software, určený k filtrování přístupu na internet, nicméně jeho popisné možnosti by umožňovaly i jiné uplatnění [W3C, 1996A].

4.10.2 XML Metadata Interchange Format (XMI)

Jedním z projektů, založených na XML je Metadata Interchange Format (XMI). Ten byl v roce 1997 zahájen za podpory významných firem jako jsou např. Oracle, IBM, Fujitsu nebo Unisys, sdruženými ve výzkumné skupině *„Object Management Group" (OMG)*. Cílem tohoto projektu je *„umožnit jednoduchou směnu metadat mezi softwarem pro modelování a mezi repozitáři metadat a s ním spojeným softwarem."* XMI je postaven na třech oborových standardech:

XML – eXtensible Markup Language

UML – Unified Modeling Language, – standard OMG pro modelování objektů

MOF – Meta Object Facility – standard OMG pro vytváření repositářů pro metadata.

Sdružení těchto tří prvků do nového standardu –XMI by mělo umožnit vývojářům sdílet informace o objektových modelech přes internet [XMI, 2002].

4.10.3 Projekt LIMBER

LIMBER (*Language Independent Metadata Browsing of European Resources*) se snaží vytvořit několikajazyčný tezaurus pro oblast sociálních věd v evropských zemích. Tento projekt vznikl jako součást programu Evropské unie IST – (Information Society Technology), který se zabývá řešením problémů spojených s lingvistikou a jejími hraničními obory. Hlavní myšlenkou celého projektu je dostupnost informací stejného charakteru bez ohledu na používaný jazyk. K tomu, aby byla data takto interoperabilní slouží metadatový popis, který umožňuje převádět tyto informace do dalších jazyků. K dispozici je také rozhraní v několika jazycích a tezaurus, které mohou převést slovo či koncept (např. doprava – definice názvů dopravních prostředků a další výrazy) do jiného jazyka. Informace z každé stránky by pak byla dostupná všem uživatelům, bez ohledu na jazyk, který ovládají [MILLER, 2002]. Původní sada metadat byla nekompatibilní s ostatními, nyní projekt používá RDF schéma.

4.10.4 Publishing Requirements for Industry Standard Metadata (PRISM)

Definice Publishing Requirements for Industry Standard Metadata (PRISM) popisuje standard pro výměnu a užití obsahového popisu v oblasti tradičního i elektronického publikování především v oblasti internetu. Na vývoj projektu dohlíží pracovní skupina, která je součástí organizace *IDEAlliance (International Digital Enterprise Alliance)*, která se zaměřuje na zavádění nových technologií do publikačního průmyslu.

Celý projekt je založen na kombinaci obecných standardů – XML, RDF, Dublin Core a ISO norem pro geografická místa, jazyka a také formáty pro datum a čas. Kromě toho vytváří jmenné prostory v XML a kontrolované slovníky pro specializované užití.

PRISM definuje metadatovou sadu v XML pro popis, sdružování a další užití v oblasti knih, katalogů, časopisů a dalších typů publikací. Tento standard je zároveň i rámcem pro výměnu a uchování těchto metadat.

Další vývoj odpovídá potřebám této uživatelské skupiny a zaměřuje se především na:

obecný popis zdroje jako celku

specifikace zdroje ve vztahu k jiným zdrojům

definování intelektuálních práv ke zdroji [PRISM, 2003].

4.10.5 NewsML

NewsML je standard založený na XML, který zobrazuje a řídí zprávy po celou dobu jejich životnosti, tj. produkce, směny a užití. Vývoj tohoto standardu začal v roce 1999 ve sdružení zpravodajských agentur *The International Press Telecommunications Council (IPTC)*. První verze standardu 1.0 byla představena v roce 2000; nyní je aktuální již verze 1.2. Vývoj navazuje na předchozí standardy této organizace – IIM (Information Interchange Model) a NITF (News Industry Text Format).

Tento standard podporuje multimediální dokumenty spolu s různým kódováním znaků a jazyka. Je značně flexibilní a umožňuje stejný obsah znázornit v různých formátech – např. titulní článek, hlavní článek strany, krátká tisková zpráva, a zároveň umí spravovat verze jedné zprávy – nová zpráva automaticky nahradí starou verzi. Standard se zabývá i autentifikací metadat a elektronickým podpisem jejich tvůrce.

Na nejnižší úrovní popisu (ContentItem) jsou data o zprávě samotné, která popisují fyzický charakter reprezentace zprávy. Vyšší úroveň popisu umožňuje přidat metadata: stejně jako u ostatních standardů mají administrativní a deskriptivní část. Samostatně existuje část popisující autorská práva k dokumentu a část „NewsLines“. Tato část je zamýšlena jako reprezentace některých metadat ve formě, kterou je člověk schopen přečíst a interpretovat. Většina metadat v tomto standardu je čitelná strojově i člověkem, nicméně některá mohou být čitelná pouze strojem [IPTC, 2004]. Části „NewsLine“ odpovídají povaze celého standardu – jsou zde elementy typu HeadLine, SubHeadlines, DateLine nebo CopyrightLine. Tato část metadat a její prvky mohou být použity pouze jednou v celém záznamu.

4.11 PROJEKTY ZÁJMOVÝCH SKUPIN

Počet zde uvedených projektů není v žádném případě reprezentativní ani vyčerpávající, nicméně ty, které jsou zde uvedené, mají zatím největší počet uživatelů a zdají se být klíčovými pro celý vývoj standardů metadat. Většina z uvedených projektů nyní používá také XML pro reprezentaci datových elementů, nicméně původně se tyto projekty vyvíjely odděleně - nabízely sofistikovanější nadstavbu tagu META, který se objevuje v jazyku HTML od verze 1.0.

Každý projekt je individuálně zaměřen na potřeby zájmové skupiny, pro kterou původně vznikl a proto může být i jeho užití limitováno. Stejně se liší i typ metadat – od projektu deskriptivních metadat (Dublin Core) až po komplexní infrastrukturu digitálního archivu (METS).

4.11.1 Dublin Core Metadata Iniciative (DCMI)

Domovská stránka: http://www.dublincore.org

Vznik: 1995

Určení: deskriptivní popis k usnadnění vyhledávání elektronických zdrojů

Projekt Dublin Core vznikl v roce 1995 jako jednoduchá sada metadatových elementů k popisu elektronických zdrojů – především webových stránek. Tento účel se časem rozšířil i na popis jiných než elektronických dokumentů a fyzických objektů. Projekt DCMI je spravován společností OCLC jako jeden z hlavních výzkumných úkolů. V dlouhodobém horizontu mají tvůrci standardu vizi rozvoje v těchto směrech:

Vývoj standardu metadat pro usnadnění hledání mezi doménami.

Definovat rámec pro interoperabilitu metadatových sad.

Usnadnění vývoje pro komunity – neboli tvorba specifických metadatových sad, které budou slučitelné s body 1 a 2 [DCMI, 1995].

Základní popisnou sadu verze 1.1 tvoří následujících 15 elementů:

Title: A name given to the resource.

Creator: An entity primarily responsible for making the content of the resource.

Subject: A topic of the content of the resource.

Description: An account of the content of the resource.

Publisher: An entity responsible for making the resource available

Contributor: An entity responsible for making contributions to the content of the resource.

Date: A date of an event in the lifecycle of the resource.

Type: The nature or genre of the content of the resource.

Format: The physical or digital manifestation of the resource.

Identifier: An unambiguous reference to the resource within a given context.

Source: A Reference to a resource from which the present resource is derived.

Language: A language of the intellectual content of the resource.

Relation: A reference to a related resource.

Coverage: The extent or scope of the content of the resource.

Rights: Information about rights held in and over the resource

Zdroj: DCMI, 2003

Komunita vyvíjející Dublin Core má velmi blízko k dalším organizacím, které vytváří standardy pro internet jako *Internet Engineering Task Force (IETF)* nebo *World Wide Web Consorcium (W3C)*. Metadatová sada je tak podporována i těmito organizacemi a byla v různých verzích přeložena do téměř čtyřiceti jazyků.

Hlavními přednostmi používání standardu Dublin Core jsou:

- jednoduchost vytváření a správy záznamu
- srozumitelná sémantika
- přizpůsobitelnost existujícím i nově se objevujícím standardům
- mezinárodní záběr a aplikovatelnost
- rozšiřitelnost
- interoperabilita mezi jednotlivými sbírkami a indexačními systémy [DCMI, 1998].

Tento standard byl mezinárodní komunitou velmi dobře přijat a o jeho úspěchu svědčí i množství projektů, které na jeho základě vznikly. Více než sedmdesát projektů používá metadata Dublin Core pro katalogizaci, identifikaci a jako navigační pomůcku v oblastech vědy a výzkumu, veřejně přístupných a vládních informacích a jiných specializovaných projektech.

Některé z těchto programů jsou **podporovány Evropskou unií** (např. *BIBLINK* pro vytváření a výměnu metadatových záznamů mezi vydavateli a národními knihovnami) **nebo jinými vládami** (např. australský projekt *Australian Government Locator Service (ALGS)* pro katalogizaci vládních dokumentů). Jednotlivé projekty často překračují původní rámec a vytvářejí nový – sofistikovanější produkt.

To se týká projektu *CORC (Cooperative Online Resources Cataloging)*, který OCLC zahájila v roce 1998 za účelem integrace katalogizačního formátu MARC s novými metadatovými standardy jako *Dublin Core*, *Text Encoding Iniciative (TEI)* nebo *Encoded Archival Description (EAD)*. Významným projektem, který využívá sadu Dublin Core, je také *Nordic Metadata Project*[13], který v rámci širšího strukturálního projektu NORDINFO rozšířil původní datovou sadu a vytvořil nové softwarové pomůcky pro práci s metadaty – mezi jinými i konverzní program mezi formáty Dublin Core a MARC [HAKALA, 1998].

Na konferenci v Šanghaji byl nastíněn další směr vývoje tohoto standardu. Plány do budoucna počítají s rozšířením povinných elementů (např. „Accessibility“ – dostupnost). V oblasti schémat se objevila dvě důležitá doporučení. První z nich doporučuje podrobněji rozdělit stávající schémata podle typu na schémata s definovanou syntaxí a na schémata s definovaným slovníkem. Druhé doporučení navrhuje zlepšit možnosti propojení s jinými metadatovými schématy, založenými na schématu URI tak, aby je bylo možné propojit se specializovanými terminologickými slovníky uživatelů. Toto propojení má být podrobněji projednáváno s *Národní lékařskou knihovnou USA (NLM)* [BAKER ET.AL., 2004], což naznačuje snahu o větší integraci standardu Dublin Core do stávajících modelů pro pořádání informací a také do zamýšleného konceptu sémantického webu.

V prosinci roku 2006 proběhla další revize základní popisné sada, která především vyjasňuje terminologii popisu a blíže definuje její význam. Na ní navazuje na začátku roku 2007 další významný krok – nová revize abstraktního modelu (DCMI Abstract Model). K této revizi navrhl její tvůrce Andy Powell také návrh slovníku tříd a jejich rozsahu pro metadatové elementy DCMI. Cílem tohoto návrhu je implicitní rozlišení popsaných zdrojů a jim přiřazených hodnot takovým způsobem, aby bylo možné zpracování těchto údajů strojově; tj. bez přítomnosti člověka. Tento vývoj naznačuje, že

standard DCMI bude mít nadále velkou budoucnost především v souvislosti s dalším vývojem schopností specializovaných vyhledávačů.

4.11.2 Text Encoding Initiative (TEI)

Domovská stránka: http://www.tei-c.org

13 http://www.lib.helsinki.fi/meta/index.html

Vznik: 1987

Určení: deskriptivní popis, správa, uložení a archivace materiálů z oblasti humanitních věd; především zaměřené na lingvistické texty

TEI je nejstarším projektem v oblasti popisných metadat vůbec. Projekt byl zahájen v roce 1987 jako projekt Evropské komise, National Endowment of the Humanities a Mellonovy nadace. První verze metadatové sady byla zveřejněna v roce 1990, v současnosti (od roku 2002) je k dispozici již čtvrtá verze označená jako TEI P4 (XML). Iniciativa TEI vznikla za účelem vytvoření metadatového formátu, který by umožňoval popis vytvořených a uložených textů, byl převoditelný a nezávislý na platformě. Základní požadavky byly v roce 1987 stanoveny následovně:

reprezentace textových pomůcek, nezbytných pro výzkum

jednoduchost formátu i užití, přehlednost a konkrétnost

možnost přidání uživatelských doplňků

kompatibilita se současnými i nově se objevujícími standardy.

Výsledkem je metadatová sada, která má pouze minimum povinných prvků a umožňuje uživateli stanovit si míru podrobnosti popisu podle vlastního uvážení.

Metadatová sada TEI se stala nepsaným standardem pro obory jako je historie, literární vědy, dějiny umění nebo lingvistika. Pro tyto obory má mnoho předností – například určení literárního žánru díla nebo podporu různých znakových sad včetně antických jazyků. Standard zároveň podporuje strukturní členění dokumentu na části (kapitoly, scény apod.), umožňuje zachytit speciální typografické elementy (změna fontu, speciální znaky) nebo zachycení jiných vlastností textu (gramatická struktura, umístění ilustrací apod.) [BAKER ET.AL., 2004].

Obrázek 9: Logo „TEI Pizza Chef“ Zdroj: http://www.tei-c.org/xpizza.html

Samotná sada metadat je založena na definičním souboru DTD a řídí se XML syntaxí.

Každý soubor má dvě části:

1. hlavičku (TEI header)

Hlavička obsahuje kompletní bibliografický popis elektronického souboru, uchovává data o vztahu mezi elektronickým textem a jeho původním zdrojem, popisuje další – nebibliografické aspekty textu jako použité jazyky a podává přehled o historii změn dokumentu.

2. tělo (TEI body)

Vlastní tělo dokumentu popisuje dokument po stránce formální i obsahové.

Sada TEI elementů a jejich atributů se pohybuje v řádu několika stovek a celou sadu potřebuje pouze minimum uživatelů. Z důvodu jednoduchosti použití byl proto vytvořen profil „**TEI lite**“ – minimální znaková sada pro popis dokumentu. Podle dokumentace projektu [BURNARD, 2002] profil TEI lite pro popis používá asi 90% všech zúčastněných organizací. Tento profil je přednastaven tak, aby vyhovoval základním potřebám popisu zdroje. Tvůrce dokumentu má k dispozici základní popisnou sadu, se kterou je možné popsat široké spektrum textů, s nimiž se tvůrci setkají. Tuto sadu lze použít také v kombinaci s některými typy softwaru určeného pro tvorbu XML dokumentů. Z důvodů jednoduchosti a zároveň velké flexibility je profil TEI Lite používán i u rozsáhlých projektů jako například *Oxford Text Archive*[14].

Projekt TEI nabízí také jednu unikátní možnost – **„TEI Pizza koncept“**, kterou autoři označují jako upečení „pizzy na míru“. Myšlenkou je přizpůsobení TEI co nejvíce individuálním potřebám uživatelů. Celý koncept popisuje Susan Schreibman následovně: *„směrnice TEI definují několik set SGML elementů a asociovaných atributů, které mohou být zkombinovány do mnoha odlišných DTD souborů, vhodných pro různé účely – jednoduché nebo složité. S pomocí „Pizza Chef“ si můžete vytvořit DTD soubor, který obsahuje elementy, které chcete, vhodné pro použití s jakýmkoli systémem, kompatibilním se SGML nebo XML.“* [SCHREIBMAN, 2003].

14 http://ota.ahds.ac.uk/

V rámci tohoto konceptu si uživatelé mohou přizpůsobovat části, které jsou označovány jako náplň (toppings). Jedná se především o prvky jako:

odkazy

tabulky

analýzy (lingvistické analýzy)

přepisy a další.

V prosinci roku 2000 bylo založeno nové konsorcium, které bude nadále TEI vyvíjet – k zakládajícím členům se přičlenily další organizace (např. *University of Oxford*, *Research Technologies Service* nebo *Brown University*). Vzhledem k rozšíření počtu významných organizací, podporujících tento standard, se dají očekávat nové verze datové sady.

Budoucí vývoj bude směřovat k větší interoperabilitě datové sady – především prostřednictvím RDF. Dalším cílem je rozvoj softwaru pro jednodušší popis dokumentů, který je velmi důležitý pro mnoho přistupujících institucí [BURNARD, 1995]. TEI má jako standard velikou budoucnost a vzhledem k dosavadním výsledkům projektu lze očekávat jeho větší rozšíření.

4.11.3 Global Information Locator System (GILS)

Domovská stránka: http://www.gils.net/

Vznik: 1995

Určení: Systém popisných metadat a dalších navigačních pomůcek pro podporu vyhledávání.

GILS je podle jeho tvůrců doslova „decentralizovaná sbírka lokátorů a přiřazených informačních služeb, používaných veřejností buď přímo nebo přes prostředníky za účelem vyhledání informace.“ [MOEN, 1997].

Global Information Locator System (GILS) je jedním z konkrétních výsledků zákona *Paperwork Reduction Act* z roku 1995 a stal se standardem pro výměnu informací o vládních dokumentech v elektronické formě (*Federal Information Processing Standard – FIPS Pub192*) [HODGE, 2001]. Důvodem pro jeho vytvoření byla snaha popsat vládní dokumenty, publikované především na internetu tak, aby je bylo možné rychle a přesně najít a zároveň omezit duplicitu, která zákonitě vzniká při publikaci materiálů na různých úrovních (stát-město apod.). Primární určení pro vládní dokumenty odráží i původní akronym - **G**overnment **I**nformation **L**ocator **S**ystem.

GILS není originální sadou metadat, která by byla definována vlastním slovníkem nebo syntaxí – je spíše **konkrétním profilem normy Z39.50** (ISO 23950), který je primárně určen pro vyhledávání a výměnu informací. Z důvodu zpřístupnění informací se tento standard orientuje spíše na informace o dostupnosti a distribuci informací; deskriptivní popis je až na druhém místě.

Základní datová sada GILS pro komunikaci s vyhledávacím protokolem Z39.50 se skládá z několika desítek elementů a kvalifikátorů – např. jméno autora, jméno korporace, pole klasifikace (DDC, LCC, MDT, Blissovo třídění), identifikátory (ISBN/ISSN), cena nebo dostupnost. Tyto prvky je možné kombinovat pomocí syntaxe a skládat tak dotazy na cílový systém – zpravidla server s databází [CHRISTIAN, 2003]. Jedna z výhod GILS je, že odkazy registrují sémantické datové prvky místo specifické datové struktury. To umožňuje GILS interoperabilitu, která není svázána s žádným formátem pro strukturovaná metadata [VELLUCCI, 1998].

GILS definuje vyhledávací službu pouze na úrovni komunikace mezi počítači a tato definice je převedena do architektury klient-server. Pro podporu GILS je nezbytný server, který podporuje protokoly a standardy, na jejichž základě funguje.

GILS podporuje databázové standardy jako je Structured Query Language (SQL), adresářové struktury Lightweight Directory Access Protocol (LDAP) standard a další, především založené na XML.

GILS je jedním z prvních projektů, který se zabývá metadaty v kontextu vyhledávání informací na základě protokolu Z39.50. Tento protokol se již stal neoficiálním standardem pro knihovnické systémy a je využíván i pro jiné účely. Projekt GILS se svým pojetím blíží myšlence sémantického webu, který předpokládá využívání metadat i podporu sofistikované vyhledávací soustavy.

Standard GILS byl zatím implementován v USA, Kanadě, Japonsku a Austrálii. Zároveň se stal součástí projektu sedmi nejvyspělejších zemí světa – G7 *„Global Information Society"* pro sdílení dat především z oblasti ochrany a využívání přírodních zdrojů a ekologie [GILS, 1997].

4.11.4 Encoding Archival Description (EAD)

Domovská stránka: http://lcweb.loc.gov/ead/

Vznik: 1993

Určení: Metadata pro podporu vyhledávání.

Projekt EAD vznikl na *University of California* v Berkeley, kde se tým odborníků snažil vytvořit volně šiřitelný standard pro kódování strojově čitelných metadat určených pro podporu vyhledávání. Po dvou letech tak vznikl nový standard na bázi SGML, který je souborem DTD (Document Type Definition) [VELLUCCI, 1998].

EAD je primárně určen pro popis objektů v různých katalozích a registrech, což ho odlišuje od jiných metadat pro podporu vyhledávání – ta jsou určena hlavně k vyhledávání v knihovních sbírkách, muzejních kolekcích v digitální formě nebo v archivních projektech. Přestože se EAD jako standard na tvorbu podpůrných informací pro vyhledávání liší od jiných projektů, stále se jedná o metadata.

Archivní popis se v několika směrech značně liší od popisu bibliografického [PITTI,1999]:

archivní popis reprezentuje fondy, komplex materiálů, které jsou často dostupné na různých médiích

zahrnuje hierarchickou analýzu, tj. začíná popisem celku, dále popisuje části, ze kterých se celek skládá, a poté dalšími, podřízenými částmi. Archivní popis často končí na úrovni nejmenších částí celku (rukopisů, stran, knih), ale není to pravidlem

archivní metadata jsou primárně určena pro popis na úrovni celku – fondu a proto popis zahrnuje detailní analýzu hierarchických vztahů v kolekci

ve srovnání s deskriptivním popisem jsou archivní záznamy mnohem delší.

Významným prvkem EAD je zdůraznění hierarchických vazeb v souboru a **„dědění" popisu v hierarchii** – tj. obecné informace, kterými je popsána např. sbírka knih se automaticky objeví i v popisu jedné knihy, případně její části.

Standard EAD se skládá ze tří částí – definičního souboru DTD, knihovny elementů (tag library) a z aplikačních pravidel. Definiční soubor DTD má tři části:

<eahedher> – hlavička, která dokumentuje archivní popis

<frontmatter> – část, která slouží pro publikační informace (titulní stránka, další informace)

<archdesc> – samotný archivní popis, jádro EAD

EAD má podobně jako TEI pouze minimum požadovaných povinných prvků, většina elementů je volitelná. Každý si tak sám může určit míru podrobnosti popisného schématu. Popis zdůrazňuje intelektuální strukturu a obsah materiálu spíše než jeho fyzickou charakteristiku [PITTI, 1999]. Přestože EAD striktně neurčuje intelektuální obsah metadat pro vyhledávání, definuje formu obsahu a předpokládá se využití s jinými standardy především v oblasti popisu a archivace – např. *General International Standard Archival Description (ISAD(G))*.

Nejnovější verze – EAD 2002 přinesla změny ve zrušení některých starých elementů a přidání některých nových prvků, především za účelem větší kompatibility s *General International Standard Archival Description (ISAD(G))*, s XML a jeho příbuznými technologiemi [LIBRARY OF CONGRESS, 2003B].

V současnosti spravují EAD dvě organizace – *Library of Congress (LOC)* a *Society of American Archivist (SAA)*. Tyto organizace směřují další vývoj k tomu, aby se národní oborový standard mohl stát mezinárodním. K tomu by mělo přispět i přizpůsobení a přeložení EAD DTD a dokumentace do národních jazyků. Dalším směrem vývoje je kontrola autorit pro popis objektů. V současné době probíhá práce na projektu, který se snaží vytvořit DTD soubor, kompatibilní s EAD, který by splňoval mezinárodní standard *International Standard Archival Authority Record for Corporate Bodies, Persons, and Families (ISAAR(CPF))*. Tento projekt by měl napomoci vybudování budoucí bibliografické a historické databáze o právnických osobách [PITTI, 1999].

4.11.5 Metadata Encoding and Transmission Standard (METS)

Domovská stránka: http://www.loc.gov/standards/mets

Vznik: 2001

Určení: Komplexní řešení pro popis, vyhledávací pomůcky a správu digitálního archivu.

Metadata Encoding and Transmission Standard (METS) je komplexním standardem pro kódování deskriptivních, administrativních a strukturních informací o digitálním objektu, který používá schéma jazyka XML. Tento standard je vyvíjen oddělením *Network Development and MARC Standards Office* Kongresové knihovny Spojených Států. Celý projekt byl zahájen jako iniciativa Digital Library Federation, jejímž hlavním cílem je vytvoření standardu pro digitální knihovny, který by umožňoval kódovat deskriptivní, administrativní a strukturní informace dohromady s primárním objektem popisu.

Tento standard má několik částí [LIBRARY OF CONGRESS, 2003A]:

slovník a syntax pro identifikaci digitálních částí, které dohromady tvoří jeden celek, pro specifikaci fyzického umístění a pro vyjádření vztahů mezi těmito částmi

syntax pro převody a výměnu dat o digitálních entitách (dokumentech)

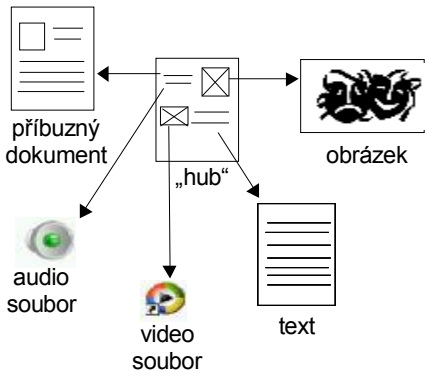
funkční syntax jako základní možnost pro orientaci v kódu pro lidské uživatele (je možné identifikovat jednotlivé části dokumentu a popisné prvky)

archivní syntax pro kódování archivních digitálních celků.

METS je používán pro popis a kódování tzv. **„hub dokumentů“** tj. dokumentů, které se mohou skládat z několika částí nebo odkazují na části příbuzné.

Popisná sada METS určuje vztahy mezi nadřazeným souborem a jeho částmi a definuje hierarchické a další vazby mezi těmito částmi. Dokument se může skládat z různých částí a typů souborů – textu, audiovizuálních souborů, obrázků nebo multimédií. Strukturní informace je možné vyjádřit mezi všemi typy objektů, nehledě na jejich formát. METS objekt může „zabalit“ obsah metadatového souboru do digitálního objektu v podobě binárních dat. Stejně tak je možné takto vložit jakákoli další metadata, která jsou použita jako doplněk k základní sadě METS.

Metadatový popis METS má sedm základních částí:



Obrázek 10: Příklad schématu „hub dokumentu“

1. METS hlavička (metsHDR)

Obsahuje základní informace o popisovaném dokumentu.

2. Deskriptivní sekce metadat (dmdSec)

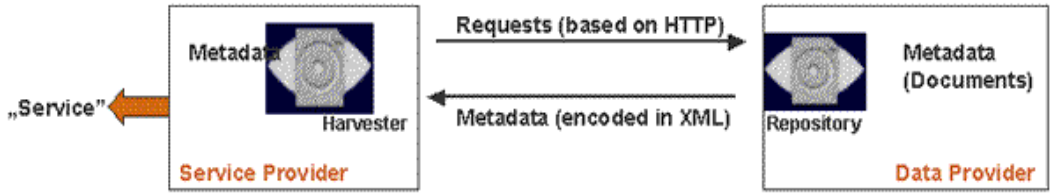
Data v jakékoli formě včetně standardů MARC, Dublin Core nebo jiných pro popis obsahu dokumentu. Všechna tato data musí používat syntaxi specifikovanou XML standardem.

3. Administrativní metadata (admSec)

Informace technické a administrativní povahy (datum vytvoření, formát souboru, in-formace o zdrojovém-nedigitálním objektu...) a informace o právech k objektu (držitel autorských práv digitálního a originálního objektu, informace k licencování apod.).

4. Souborová část (fileSec)

Zde jsou uvedeny všechny části dokumentu, které spolu tvoří jeden celek. Jednotlivé soubory jsou většinou ukládány do adresářů a skupin na základě datového formátu (obrázky, text) a základní dokument na tyto části odkazuje.



Obrázek 11: Komunikace mezi harvestorem a digitálním archivem Zdroj: http://www.oaforum.org/tutorial/english/page3.htm

5. Strukturní mapa (structMap)

Tato část popisuje (hierarchickou) strukturu všech částí digitálního dokumentu a specifikuje způsob, jak se mají soubory zobrazit v jeho struktuře.

6. Strukturní odkazy (structLink)

Odkazy na části dokumentu, specifikované ve strukturní mapě.

7. Část akcí - Behavior Section (behaviorSec)

Odkaz na externí rozhraní, které definuje chování multimediálního dokumentu – např. spuštění videa nebo zvuku.

METS je velmi flexibilním modelem, neboť dokáže informace popsat vlastním slovníkem nebo použít popis v jiném standardu, který splňuje požadavek strukturace podle specifikace jazyka XML. Kromě standardních metadatových standardů jako Dublin Core, které lze použít bez dalších úprav, lze do METS pomocí dalších rozšíření včlenit i jiná metadata. **Katalogizační formát MARC** je možné do METS přidat prostřednictvím nadstavby **MODS (Metadata Object Description Schema)** nebo ve formě schématu MARC21 – MARCXML. METS je plně kompatibilní se standardem RDF a tak ani včlenění jiných metadat není problémem.

Tento standard začíná být široce podporován pro komplexnost řešení. V některých projektech (např. *Center for digital Initiatives – Brown University*) je preferován oproti jiným standardům (Dublin Core) z důvodu větších možností popisu v deskriptivní části.

S podporou institucí jako je Kongresová knihovna, Harvard University nebo University of California se tento projekt nadále vyvíjí. Jeho vývoj směřuje i k podrobnější specifikaci pro popis autorských práv a je tak možné, že tento standard převezme úlohu nejpoužívanější popisné sady pro digitální sbírky a archivy.

4.12 HARVESTERY METADAT

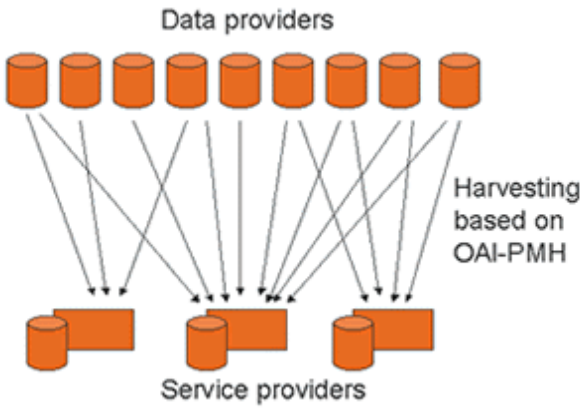
Používání metadat v digitálních archivech a na webových stránkách vedlo ke zrodu specializovaných vyhledávačů tzv. **harvesterů metadat**[15]. Jejich zrod je spojen s iniciativou *„The Open Archives Initiative"*, která v roce 2001 vytvořila standard **Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)**. Tento standard existuje již ve verzi 2.0 a jeho poslední revize byla provedena v říjnu roku 2004. Protokol stanovuje rámec pro výměnu metadat na základě tzv.

[15] Nejbližším českým výrazem je patrně „sklízení metadat".

harvestingu metadat, který je zcela nezávislý na aplikaci. Tento rámec definuje dvě části, na jejichž spolupráci je celý systém postaven:

poskytovatelé dat (administrují systém, který podporuje OAI-PMH ve smyslu vystavování metadat)

poskytovatelé služeb (používají metadata, získaná protokolem OAI-PMH jako základ pro poskytování dalších služeb) [OPEN ARCHIVES INITIATIVE, 2004].



Obrázek 12: Model harvestingu
Zdroj:http://www.oaforum.org/tutorial/english/page2.htm

Základní architektura vypadá následovně: uživatel vyhledá záznam dokumentu, který chce získat. Poskytovatel služby na základě těchto údajů vyhledá metadata tohoto dokumentu a najde archiv, ve kterém je dokument uložen. Následně si vyžádá kopii dokumentu, kterou zašle uživateli.

Specializované vyhledávače, založené na tomto protokolu, nalezená metadata použijí ke zjištění:

identifikátoru

formátu dokumentu

formátu metadat, případně dalších dostupných informací.

Koncept harvestingu je velmi zajímavý a výrazně zvyšuje použitelnost metadat, především zjednodušeného formátu Dublin Core, který je stanoven jako závazný.

Harveting metadat přináší také řadu problémů při zpracování, které Roy Tennant [TENNANT, 2004] shrnuje takto:

sady metadat

Metadata jsou shromažďována ve skupinách tak, jak je poskytovatelé dat zpřístupnili v archivu. Pro tuto organizační část neexistují žádná pravidla.

formáty metadat

V archivech se používají různé sady metadat, které nemusí být navzájem kompatibilní. Závazný formát – zjednodušená sada Dublin Core je velmi jednoduchá, a jako převodní formát nedostatečná.

artefakty metadat

Pro harvestor jsou obtížně odlišitelné různé fragmenty nebo nahodilé shluky znaků od kódu metadatového formátu. Tak může dojít k záměně např. s HTML kódem nebo s různými příklady zápisu metadat, které neobsahují informace o existujících dokumentech.

granularita metadat

Nevyřešeným problémem je i otázka granularity metadat, tj. co má být základní jednotkou pro popis. Je jí kniha, nebo jedna stránky z knihy? Stejným problémem je převod strukturovanějších údajů do formátu, který je méně strukturovaný a tak se některé údaje musí sloučit do společných polí.

variace v užívání standardů

Řada standardů metadat umožňuje úpravy a interpretace kódu, a tak se může lišit definice údaje, jeho typ, nebo způsob zápisu. Příkladem může být zápis data (2004-01-23 versus 23/1/2004).

Pro podporu projektu OAI-PMH vznikla řada programů, které jsou volně dostupné například v rámci *Virginia Tech Perl Harvester*[16], nebo *The University of Illinois at Urbana-Champaign Open Archives Initiative Metadata Harvesting Project*[17], kde je dostupný i vyhledávač metadat v on-line archivech.

V současné době využívá technologii harvestingu přes 500 archivů z celého světa. Vyhledávače – harvestory, jsou dostupné především na stránkách univerzit a různých výzkumných organizací, které se projektu účastní. Nejvýznamnějším metadata harvesterem je **OAIster.org**. K únoru 2005 indexuje 5 195 319 záznamů z 444 institucí.

Koncept metadata harvesterů je perspektivní i pro možné vyhledávače na internetu, ale vzhledem k rozšíření metadat a především jejich pokročilejších schémat (viz studie[LAWRENCE-GILES, 1999]), by jejich aplikace byla předčasná.

4.13 DALŠÍ VÝVOJ METADAT

Kromě zastánců metadat se objevili i jejich odpůrci, kteří poukazují na některé hrubé nedostatky v konceptu metadat. Cory Doctorow [DOCTOROW, 2001] uvádí následující důvody, proč koncept metadat nemůže nikdy fungovat:

záměrné dezinformace o obsahu dokumentu v metadatech

neschopnost uživatelů popisovat dokumenty metadaty – lenost?

tvůrci standardů metadat sami nevědí, co chtějí

schémata metadat nejsou neutrální – ovlivnění významu prvku v hierarchii

„metrics influence results“ – forma měření ovlivňuje výsledky

existuje více způsobů, jak popsat dokument metadaty.

16 http://www.dlib.vt.edu/projects/OAI/software/harvester/harvester.html

17 http://oai.grainger.uiuc.edu/

Z těchto bodů se dá objektivně souhlasit se čtyřmi, které je vhodné blíže rozebrat.

Záměrné dezinformace o obsahu dokumentu v metadatech

Tato praxe se objevila v souvislosti s metadaty v hlavičce HTML dokumentů. Vyhledávače tato data používaly pro hodnocení stránek a jejich zařazení v seznamu výsledků ve větší míře než dnes, a proto začala být metadata zneužívána. Uživatelé zjistili, že hodnotící algoritmus se zaměřuje na příbuznost konceptů slov (slova podobného významu, sdílející stejné téma) a hlavně na četnost klíčových slov. Začaly se objevovat dokumenty s naprosto irelevantním obsahem, nicméně vyhodnocené vyhledávačem jako nejlepší jen proto, že klíčové slovo bylo v metadatech obsaženo padesátkrát. Tato praxe, označovaná jako tzv. *„index spamming"* (viz. kapitola 9.2.3, s. 120), značně zdiskreditovala hodnotu údajů META v hlavičce HTML dokumentu a vyhledávače již nepřikládají těmto údajům tak veliký význam.

Neschopnost uživatelů popisovat své dokumenty metadaty

Tato výtka je více než oprávněná. Praxe zjednodušit si život i v oblasti publikování se rozšířila také na internetu – od vynechávání základních popisných údajů (název stránky apod.) přes nefunkční odkazy až po publikování ve formátech, které pro toto prostředí nejsou určeny.

Lidé si vždy život zjednodušovali a v této oblasti se to negativně projevuje. Celý koncept metadat je postaven na modelu popisování dokumentu samotným autorem, který by měl nejlépe vědět, o čem dokument je. Popis je vzhledem ke své pracnosti a ke složitosti zdrojového kódu současných metadatových schémat (přestože existují i uživatelsky orientované grafické editory) stále časově náročnější záležitostí; většina autorů tento čas není ochotna do své práce investovat.

V současné době někteří (především firemní) uživatelé metadata znovu objevují v rámci boomu, tzv. optimalizace stránek pro vyhledávače *(Search Engine Optimalization –SEO)*, ale nedá se předpokládat masová změna chování u většiny uživatelů. Tato nectnost může být v konečném důsledku výhodou – kvalitní obsah se projevuje i dobrým popisem a tak se metadata mohou stát i určitým indikátorem kvality dokumentu.

Schémata metadat nejsou neutrální – ovlivnění významu prvku v hierarchii

Každá hierarchie má význam pro orientaci v prostoru, ale také v hodnotovém žebříčku. Schéma – stejně jako jiná klasifikace nezbytně přináší i otázku hodnot a jejich pořadí. Tento problém je ovšem nerozlučně spjat s jakoukoli hierarchickou soustavou pořádání a není možné se ho vyvarovat.

U metadat je tato funkce oslabena – hierarchické struktury jsou vyjadřovány spíše v oblasti strukturních informací – tj. z jakých částí se popisovaný objekt skládá, ale již neurčují hodnotu jednotlivých částí.

Existuje více způsobů jak popsat dokument metadaty

Stejně jako předchozí bod je toto problémem každého pořádacího systému a katalogizace obecně. Každý člověk může zvolit popisné termíny, případně sadu metadat podle svého úsudku. Tento problém se řeší řízenými slovníky popisných termínů, které, ačkoli nemusí být natolik specifické, standardizují popisnou sadu. Pro standardy metadat jsou to často různé profily, které kromě syntaxe určují i možný slovník popisných termínů.

Přes tyto nedostatky mají metadata před sebou ještě velkou budoucnost. **Největším současným problémem metadat je jejich standardizace**. Metadata jsou podle studie [LAWRENCE-GILES, 1999] poměrně rozšířena – **ve zkoumaném vzorku mělo metadata jakékoli formy 34,2% serverů**. Výzkum ukázal, že nejčastěji jsou užívána metadata jazyka HTML (tj. nejjednodušší forma zápisu) a **užívání různých metadatových schémat** (viz. uvedené projekty) **je velmi nízké**[18]. Přesto bylo potvrzeno, že různá schémata metadat jsou užívána zájmovými komunitami, což přes malou rozšířenost dokazuje jejich přijetí v praxi. Na úkor těchto metadat by měl klesat počet stránek, které jsou popsány základními meta tagy

[18] Podle této studie používá např. metadatové schéma Dublin Core pouhých 0,3% zkoumaných stránek.

jazyka HTML. Studie také ukázala, že existuje veliká rozmanitost i v užití těchto jednoduchých metadat – celkem bylo na stránkách nalezeno 123 různých tagů pro metadata.

Autoři studie poukazují ve svých závěrech na tyto výsledky s tím, že v této oblasti chybí větší standardizace. Vzhledem k roku publikování studie – 1999 lze říci, že mnohé v této oblasti se zlepšilo, nicméně problém přetrvává a některé otázky (např. granularita metadat) se řeší velmi obtížně.

Funkčnost organizace informací formou metadat závisí na fungování komplexního modelu, který je znázorněn na obrázku č. 3.10, ve kterém jsou zapojeny všechny zúčastněné strany – autor dokumentu, zprostředkovatel (např. vyhledávač) i uživatel.

Velkým otazníkem je podpora sofistikovanějších metadatových standardů současnými vyhledávači. Pokud nebudou metadata široce akceptována napříč celým spektrem uživatelů internetu, nedá se předpokládat ani jejich větší podpora vyhledávači. Proto je potřeba přesvědčit tvůrce o efektivitě tohoto popisu a postupně tak získávat větší podporu tomuto řešení. Zatím metadata a jejich standardy zůstávají středem pozornosti jednotlivých zájmových skupin, knihovníků a informačních specialistů, ale hlavního proudu „internetové populace“ se příliš netýkají.

Metadata jsou spíše podporována vyhledávači v uzavřeném prostoru – ve firemních archivech, portálech či repositářích. Zde mají tyto informace nesmírnou cenu, která se projevuje v rychlosti nalezení informace. Firmy začínají být i schopné vyčíslit tuto hodnotu finančně. Podle studie IDC a organizací jako je AIIM, Ford Motor Company, a Reuters [FELDMAN, 2003] bylo zjištěno následující:

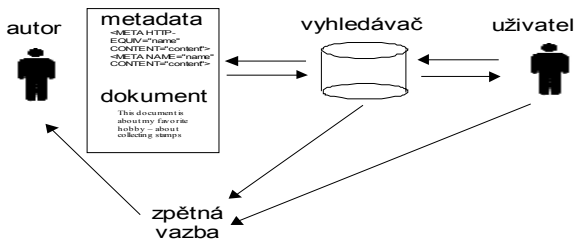
duševně pracující zaměstnanci stráví 15% až 35% svého pracovního času hledáním informací

kolem 40% uživatelů korporativních intranetových portálů uvedlo, že nejsou schopni najít zde informace, které potřebují pro svoji práci.

Tyto závěry ukazují cestu, kterou by se celý koncept metadat měl ubírat – užití pouze v omezené skupině uživatelů (podnik, oborová komunita), kde budou všichni akceptovat stejný způsob zpracování informací, standard, kterými budou datové objekty popisovány a zároveň vyhledávací systém, který tento standard bude podporovat. Jak ukazují např. výsledky projektu GILS, tato snaha se vyplácí a tento model s nezbytnými technologickými úpravami (nové technické standardy, metadata...) může ještě dlouho úspěšně fungovat.

Další vývoj metadat a jejich schémat se přesouvá do projektu tzv. sémantického webu, který metadata užívá jako základní konstrukční prvky pro vyjádření významu dokumentu a jeho vztahu ke skupině jiných, ale příbuzných dokumentů. Součástí této snahy je i budování tzv. *tématických sítí* (topic maps – viz kapitola 7, s.110), které umožní lepší přiřazení významu a kontextu dokumentu.

V současnosti již existují některé projekty, které tímto způsobem metadata využívají, nicméně všechny jsou ve stádiu experimentálního provozu a jejich zavedení do běžné praxe bude ještě několik let trvat.



Obrázek 13: Model efektivního využívání metadat

Tento vývoj zároveň znamená větší provázanost metadat s dalšími identifikátory – především s konceptem URI. První projekt (CORES), který se zabývá rozlišením typu metadat na základě části URI již probíhá pod záštitou technologického programu Evropské unie. Projekt CORES tak přispívá i k větší provázanosti URI a metadat, pro které se předpokládá široké praktické uplatnění. Kromě výše zmíněných schémat ISBN/ISSN by toto řešení bylo aplikovatelné i pro obchodní identifikátory a využití v průmyslové a jiné klasifikaci [BAKER-DEKKERS, 2003]. Na konkrétní výsledky si však ještě musíme počkat.

5 KLASIFIKAČNÍ SCHÉMATA

Klasifikace je smysluplné seskupování zkušeností.

Klasifikace je způsob vidění.

Barbara H. Kwasnik [1999]

5.1 PŘÍSTUPY KE TVORBĚ KLASIFIKACÍ

Hierarchické klasifikační systémy jsou historicky nejstarší formou pořádání informací. Pomocí systému nadřízených a podřízených částí je možné vytvářet jednotnou strukturu poznatků, která sama o sobě přináší cenné kontextové informace – především umožňuje jednotlivé předměty tzv. vizualizovat, tj. zobrazit je v souvislosti nejbližších příbuzných konceptů.

Podle Hanne Albrechtsen [ALBRECHTSEN, 1998] existují ke tvorbě klasifikačních systémů dva přístupy:

Racionálně-empirický

Sociálně-konstruktivistický

Racionálně-empirický přístup předpokládá, že znalosti jsou omezeny na smyslové vnímání faktů. V klasifikaci emprické to znamená budování struktury poznatků/vědění na základě osobního přesvědčení, vzniklého během individuálního procesu poznání světa.

Racionalismus naopak redukuje poznání na vše zahrnující strukturu konceptů, která by měla být univerzální a vyčerpávající [ALBRECHTSEN, 1998]. Empirismus i racionalismus se z hlediska tvorby klasifikací shodují v tom, že předpokládají přirozenou pravdu a objektivitu poznání.

Sociálně-konstruktivistický přístup (neboli také historicismus) vnímá poznání jako „produkt historických, kulturních a společenských faktorů, kde je rozdělení tříd a základní koncepty výsledkem dělení vědecké/kulturní/společenské práce do oblastí poznatků [in knowledge domains]“ [ALBRECHTSEN, 1998]. Podle tohoto přístupu jsou koncepty a struktura v klasifikaci neoddělitelné, a proto klasifikace musí odrážet vývoj a dynamiku, aby obojí – koncepty i struktura, vystihovalo vědění v určité oblasti. Znamená to zároveň, že klasifikační systém je vždy alespoň do jisté míry individuálním produktem, a dosažení objektivity klasifikace není pravděpodobné.

Tento přístup, jak komentuje Hanne Albrechtsen, vede ke změně role tvůrců klasifikací, z nichž se stávají „inženýři teorie poznání“ (epistemology engineers). Jejich roli popisuje následovně: *„Toto předpokládá, že designéři schémat v první řadě nemají hledat cesty jak aplikovat na znalosti jedinou strukturu, která by zohledňovala různé aspekty. Tito designéři by spíše měli působit jako „inženýři teorie poznání“* [epistemology engineers]*, kteří se budou snažit o vyjádření a reprezentaci dynamiky poznání způsobem, který umožní člověku hledajícímu informaci, postupovat od tématu jeho*

počátečního dotazu k jiným aspektům nalezené informace nebo k podobným materiálům ve stejné tematické oblasti." [ALBRECHTSEN, 1998].

Klasifikace by měla sloužit jako schéma, ve kterém je možné řadit koncepty poznání podle jejich zamýšleného účelu, funkce a rozsahu. Klasifikační schémata se mohou lišit podle svého určení. Zatímco univerzální klasifikační schéma se snaží zachytit velmi zobecněné univerzální poznání, dílčí klasifikační schéma (tezaurus, oborové třídění) si klade za cíl pouze vytvoření navigační pomůcky pro orientaci v části takto zachyceného poznání.

Struktura klasifikace má být vždy pragmatická a flexibilní neboť, jak uvádí Hanne Albrechtsen [1998]: „ ***... klasifikace vždy slouží pragmatickým účelům stejně jako věda slouží pro lidskou činnost",*** tj. měla by být schopna pomoci při řešení rozmanitých informačních potřeb s ohledem na účel klasifikace.

Většina dosud užívaných univerzálních klasifikačních systémů vychází z názorů, potřeb a pojetí 19. století, od kterého se dodnes nedokázala odpoutat. To znamená, že byly vytvářeny ještě racionálně-empiristickým přístupem, který předpokládal konsenzus v oblasti poznání a jeho strukturování. Od počátku 20. století dochází k zásadní změně pohledu na roli a účel klasifikace, kde se začíná uplatňovat sociálně – empirický model, jež chápe klasifikaci pouze jako navigační pomůcku s omezenou platností a významem. Tento přístup je více vidět u nových, pragmatických systémů jako jsou oborová třídění a stromové katalogy na internetu.

5.2 TAXONOMIE: HIERARCHIE A JEJÍ VÝZNAM PRO KLASIFIKAČNÍ SOUSTAVY

Výraz „taxonomie" pochází z řeckého slova *taxis*, které znamená „uspořádat". Tento termín se používá pro označení klasifikačního systému, tvořeného hierarchickou posloupností jednotlivých větví a termínů v nich uspořádaných. Taxonomické systémy jsou nejstarší formou klasifikačního schématu a ideově vychází z Aristotelovy filozofie, která předpokládá, že svět jako celek je možné rozdělit do jednotlivých „přirozených" tříd. Aristotelův pohled vycházel z domněnky, že existuje pouze jeden správný pohled na dělení a funkci částí tohoto světa. Tento náhled na zachycení světa je však dnes již překonaný – exkluzivita vnímání světa byla v postmoderním světě zcela zničena.

Předpoklad exkluzivního pohledu na svět přetrvává v hierarchických klasifikacích dodnes – výchozí úrovní pro třídění je u většiny knihovnických systémů celosvětové poznání, které se člení do dalších úrovní na podrobnější skupiny a třídy.

Pořádací systémy, založené na taxonomii (častěji označované jako „hierarchické" nebo „univerzální" klasifikace), mají řadu pravidel, která v kontextu také určují význam předmětů a jejich zařazení v některé z úrovní klasifikace. Barbora Kwasnik uvádí jako základní následující charakteristiky [KWASNIK, 1999]:

Inkluzivita – hlavní třída (univerzum poznání) zahrnuje i všechny své podtřídy.

Oddělení druhů – pravá hierarchie má pouze jeden typ vztahů mezi nadřazenými a podřazenými třídami, který se nazývá „oddělení druhů"; pod nadřazenou třídou jsou termíny řazeny podle svých specifik do podrobnějších kategorií.

Dědičnost – vlastnosti nadřazené třídy jsou i vlastnostmi jejích podřazených tříd.

Převoditelnost – protože vlastnosti jsou děděny, všechny podtřídy jsou nejen částmi jejich bezprostředně nadřazené třídy, ale i všech dalších nadřazených.

Systematická a předvídatelná pravidla pro asociace a odlišení – pravidla pro seskupování termínů do tříd jsou dána předem a všechny prvky musí sdílet určité vlastnosti. Předvídatelnost vychází z principu seskupování příbuzných konceptů u sebe (podtřídy) – tedy v zobrazení informací v kontextu.

Nezbytná příslušnost prvku třídě – každý pojem musí náležet pod nějakou třídu/ podtřídu.

Vzájemná exkluzivita – každý pojem může náležet vždy pouze do jedné třídy.

Hierarchické klasifikace mají mezi systémy pro pořádání informací zcela jedinečné místo díky některým vlastnostem, které tento způsob pořádání přibližují lidskému myšlení [KWASNIK 1999]:

1. Klasifikace je úplná a vyčerpávající

Rozsah klasifikace musí být známý předem, neboť z něj vychází i hloubka zpracování předmětů a vzájemné vztahy mezi zpracovávanými koncepty. V rámci této klasifikace jsou vzájemné významové vztahy dovedeny velmi důsledně až k předmětům v nejnižší hierarchické úrovni.

2. Dědičnost v zápisu notací

Notace – (alfa)numerický kód označující polohu konceptu v hierarchii se tvoří posloupností notace nejvyšší třídy a tříd nižších. Například třída ve třetí hierarchii, označená číslem „3“, bude mít na dvou předchozích pozicích (dvě předchozí hierarchie) dva další znaky. Úplná notace tak má tvar „1.2.3“. Na základě tohoto kódu je možné vyjádřit obsah dané kategorie (např. typy motorů). Tyto notace jsou velmi praktickým a stručným prostředkem navigace, neboť jeho další úrovně se tvoří přidáváním dalších znaků, které reprezentují další podtřídy (jednotlivé typy motorů – elektrický, dieselový, benzinový...). Část informace je tak vyjadřována samotnou hierarchickou strukturou.

3. Skutečné definice konceptů

Hierarchické klasifikace vyjadřují skutečné významy konceptů[19] jejich zařazením do určitých kategorií. Pojem „matka“ je tak vyjádřen skupinou pojmu „žena“ a podskupinou „vychovávající dítě“. Výklad tohoto pojmu tak v klasifikaci můžeme číst jako „žena, která vychovává dítě“.

4. Vizualizace, globální náhled

Patrně největším přínosem hierarchických klasifikací je zvýraznění kontextu. Pomocí stromové struktury je možné zobrazit vztahy mezi třídami, zjistit příbuzné kategorie, či jejich pozici těchto kategorií ve vztahu k výchozí skupině. Vzájemný vztah konceptů je vyjádřen i vzdáleností, kterou jsou od sebe odděleny v hierarchickém stromu. Vždy je možné přikročit od celku k jednotlivosti a naopak, což tyto klasifikace předurčuje i k jinému použití (vyjadřování společenského významu, hodnoty apod.).

Na druhé straně jsou hierarchické klasifikace spojeny s řadou problémů při jejich vytváření, údržbě a využívání. Mezi největší nedostatky hierarchických klasifikací patří:

1. Tvorba a aktualizace klasifikace

[19] Konceptem jsou označována tříděná slova a pojmy, reprezentující objekty v reálném světě.

Vzhledem ke komplexnosti je tvorba klasifikace velmi pomalá a nákladná. Tyto systémy jsou kvůli svému rozsahu velmi obtížně udržovatelné a aktualizovatelné. Svět se vyvíjí a tento typ klasifikace reaguje na jeho změny vždy se značným zpožděním.

2. Flexibilita

Jak uvádí Catherine Bertolucci [2003] na příkladu Linného hierarchické klasifikace rostlinných druhů: *„...protože Linného taxonomie musí být vědecky přesná, musí být zároveň flexibilní."* Při tvorbě hierarchické klasifikace se zpravidla nepočítá s jejími změnami v dalších letech. Z toho důvodu je klasifikace dosti rigidní, a je velkým problémem přiřadit do hierarchie další třídy – např. průnik dvou oborů reprezentovaný novým – tzv. hraničním oborem, který se váže k oběma třídám.

3. Několikanásobná hierarchie

Aristotelův názor na jediné správné vnímání světa je sice ve filozofii překonán, ale v hierarchické klasifikaci nadále zůstává. Barbara Kwasnik [1999] k tomu dodává: *„Většina fenoménů je chápána tak, že má několik, pravděpodobně se překrývajících, přesto však samostatných skupin vlastností a vztahů, v závislosti na kontextu a cíli jejich reprezentace."* Každý koncept můžeme vnímat individuálně na základě vlastních zkušeností – například les můžeme zařadit do kontextu přírody nebo rekreace, aniž by bylo jeho zařazení chybné. Pevné definování významu pojmu tak omezuje možné asociace a objevování dalších vazeb mezi koncepty.

4. Granularita tříd

Některé klasifikace – např. Deweyho desetinné třídění (DDC) nebo Mezinárodní desetinné třídění (MDT) se snaží ve všech úrovních hierarchie zachovat stejnou granularitu údajů – pod hlavní třídou se podtřídy dělí vždy na dalších deset. Tento systém je sice vhodný pro zápis notací, ale je velmi omezující k vyjádření skutečných vztahů mezi koncepty. Jestliže je granularita v jedné úrovni hierarchie větší, dochází ke zkreslování skutečných vztahů a vazeb mezi koncepty, a hierarchie tak nemůže odrážet reálný stav světa.

Hierarchické klasifikace jsou velkým přínosem v oblastech, kde je znám rozsah problematiky a vzájemné vazby mezi koncepty. V tom případě systém slouží nejen jako prostředek navigace, ale zároveň i jako systém pro pořádání (tj. poznávání světa). Patrně díky tomu jsou hierarchické klasifikace během 2000 let vývoje stále nejvýznamnějším typem klasifikační soustavy pro pořádání informací. Budoucnost těchto systémů se v současné době nedá přesně odhadnout. V době dynamických změn technologií je otázkou, zda a nakolik bude hierarchická klasifikace schopna dostát požadovaným změnám v technologiích. Zatím se však zdá, že hierarchické klasifikace si svůj význam udrží i nadále.

5.3 APLIKACE UNIVERZÁLNÍCH HIERARCHICKÝCH KLASIFIKAČNÍCH SYSTÉMŮ V PROSTŘEDÍ INTERNETU

Klasifikační systémy mají za sebou více než tisíciletou historii. Současně užívané soustavy vznikaly postupně od 19. století a byly vyvíjeny v rámci mezinárodní spolupráce. Jedná se především o Deweyho desetinné třídění (DDC), Mezinárodní desetinné třídění (MDT) a Třídění Kongresové knihovny Spojených států (LCC). Kromě těchto třídění jsou v menší míře rozšířeny třídění Henryho E. Blisse a Ranganatanovo dvojtečkové třídění (Colon classification). Vzhledem k tomuto dlouhému vývoji jsou především tři výše zmíněné systémy pokládány za standardní způsob pořádání informací v knihovnách a proto je zřejmá snaha tyto systémy transformovat pro použití v rámci internetu. To je motivováno snahou vyjít vstříc uživatelům, kteří jsou již na určitý způsob třídění zvyklí a předpokladem, že se v něm budou orientovat daleko lépe než v jiné soustavě. V současné době existuje řada projektů, které se snaží tyto klasifikační soustavy alespoň částečně aplikovat.

K hlavním výhodám aplikace tradičních klasifikačních systémů do prostředí internetu patří [KOCH, 1998]:

- podpora prohlížení jako možnosti navigace
- automatické generování struktury během prohlížení (orientace ve struktuře)
- schéma může být zobrazeno různým způsobem jako navigační pomůcka
- znalost systému mezi uživateli a knihovníky
- podpora více jazyků při pořádání a přístupu ke kolekci
- interoperabilita mezi různými informačními systémy
- podporuje rozšíření či zúžení dotazu
- užití klasifikace dodává kontext hledaným termínům a eliminuje špatné výsledky,
- které by se mohly objevit kvůli užití přirozeného jazyka (např. homonymie)
- mohou doplnit kontext a strukturu, pokud by byly použity pro zobrazení výsledků
- při hledání pomocí klíčových slov (např. výsledkem z vyhledávače by byl odkaz

do struktury kategorií v katalogu Yahoo!)

- klasifikace jsou stabilní a kontinuální (při nových vydáních klasifikace nehrozí
- překlasifikování stávajících dokumentů).

5.4 DEWEYHO DESETINNÉ TŘÍDĚNÍ (DDC)

Desetinné třídění Melvila Deweyho je vyvíjeno od roku 1876, kdy vyšla jeho první verze. Od roku 1988 je toto třídění dále vyvíjeno v rámci společnosti OCLC. Ke konci roku 2004 je dostupné již 22. vydání tohoto třídění. Současně existuje pod názvem *WebDewey* i v elektronické podobě, přístupné on-line.[20] Deweyho třídění je nej-rozšířenější systematickou soustavou, užívanou pro pořádání v knihovnách. Jeho popularita se proto odráží i v počtu projektů, které se tento systém snaží aplikovat do prostředí internetu, především ke klasifikaci webových stránek.

Projekty, které jsou zde uvedeny, nejsou a ani nemají být konečným výčtem. Jde pouze o některé ukázky aplikace DDC do prostředí internetu. Obsáhlejší, ale starší seznam projektů je dostupný online v rámci studie o roli klasifikace pro zdroje na internetu [KOCH ET. AL., 1997].

5.4.1 Blue Web'n

http://www.kn.pacbell.com/wired/bluewebn/index.cfm

Blue Web'n je kolekce více než 1800 odkazů z oblasti vzdělávání a je vytvářena společností SBC v rámci jejího většího projektu **„The Knowledge Network Explorer (KNE)**“. Kolekce je tříděna podle formátu (referenční zdroje, projekty...) a věkového zaměření a stejně jako u příbuzných projektů se snaží i o kvalitativní kontrolu odkazů, která je u odkazů zobrazena formou hvězdiček [BLUE WEB´N, 1995].

Záznamy obsahují jméno stránky, její URL, obsáhlejší abstrakt, datum přidání do systému a datum revize odkazu. Základní formou navigace je prohlížení kategorií, dostupné je i vyhledávací rozhraní, kde lze hledání omezit podle věku nebo tematické oblasti.

Všechny odkazy jsou tříděny podle Deweyho desetinného třídění – pro třídění jsou zvoleny pouze některé části jako „Umění“, „Vzdělávání“ nebo „Ekonomika“. Toto třídění je vidět pouze ve formě hierarchie samotné a u odkazů není zobrazeno. Jednotlivé notace z Deweyho třídění jsou dostupné také v nabídkové liště pro vyhledávání. Počet notací v nabídce je vzhledem k velikosti kolekce poměrně veliký – celkem 72. Většina z nich je ze třetí úrovně třídění (trojmístný kód) a jsou zastoupeny výběrově. Poměr notací z jednotlivých tříd není rovnoměrný, a tak jsou některé třídy zastoupeny třinácti odkazy (třída 3xx) a jiné pouze třemi (třída 1xx).

Projekt je velmi pečlivě udržován a každý týden aktualizován – zájemci o novinky mohou každý týden obdržet email s informacemi o nově přidaných odkazech. Tento vyšší standard ve srovnání s jinými projekty je zjevný – na rozdíl od projektů jednotlivců má společnost SBC kapacity pro údržbu projektu.

[20] Dostupné z http://www.oclc.org/dewey/about/default.htm.

5.4.2 CyberDewey

http://www.anthus.com/CyberDewey/CyberDewey.html

Pozadí tohoto projektu je poněkud nejasné, neboť k projektu není dostupná dostatečná dokumentace. Podle informací na stránkách projektu je CyberDewey pravděpodobně soukromou aktivitou pana Davida Mundieho, který začal tento adresář tvořit v roce 1995.

V adresáři je možné procházet jednotlivé kategorie ze základního hierarchického stromu nebo využít abecední seznam vybraných kategorií, které jsou doplněny kódy DDC. I když se na první pohled zdá, že je zde nabízena celá hierarchická struktura Deweyho desetinného třídění, není tomu tak. Od druhé hierarchické úrovně jsou zastoupena pouze vybraná témata, která jsou ještě navíc často přesměrována na odpovídající kategorie některých webových katalogů – nejčastěji na *WWlib* a na *Yahoo!*.

Celý adresář pravděpodobně není pravidelně aktualizován a z tohoto důvodu nejsou všechny jeho odkazy spolehlivé.

5.4.3 Dewey Browse

http://www.deweybrowse.org/

Podobně jako předchozí projekt, i tato stránka je soukromým projektem jednotlivce – v tomto případě paní Gail Shea Grainger, která pracuje jako školní knihovnice na Chesterfield School, a podle její specializace je katalog zaměřen na témata vhodná pro základní školy (Grades K-12).

Adresář je prostým souborem několika stránek, tříděných do skupin podle základní hierarchie Deweyho třídění na deset skupin od třídy 000-099 po třídu 900-999. Samostatně existuje ještě třída 92 – biografie. Na druhé úrovni adresářů jsou prosté tabulky s přehledem vybraných tříd, pod které jsou řazeny jednotlivé odkazy.

Přestože se opět jedná o projekt jednotlivce, je celý katalog velmi dobře udržován. Celý projekt není ovšem katalogem v pravém slova smyslu, spíše jej můžeme označit jako jednoduchý rozcestník odkazů.

5.4.4 BUBL Link

http://www.bubl.ac.uk/link

Služba BUBL Link, spuštěná v roce 1997, vznikla jako součást informačního servisu pro vyšší vzdělávání a komunitu knihovníků a informačních pracovníků. Projekt katalogu odkazů LINK (LIbraries of Networked Knowledge) je rozšířením původní elektronické nástěnky pro knihovny, ze které pochází i akronym názvu – **BU**lletin **B**oard for **L**ibraries [DAWSON, 1997].

Tato služba byla od svého vzniku v roce 1990 založena na dobrovolné spolupráci, od roku 1995 získala finanční zdroje od **Joint Information Systems Committee** (**JISC**) a v lednu 1995 se stala národní informační službou, kterou garantuje Centre for Digital Library Research (CLDR) na University of Strathclyde.

Na službě je vidět výrazný kvalitativní rozdíl oproti jiným obdobným projektům. Systém nabízí široké možnosti vyhledávání – prohlížení hierarchického stromu, vyhledávání, abecední seznam předmětů nebo seznamy dokumentů podle typů či zemí. Hierarchický strom Deweyho třídění je zde veden až do třetí úrovně, ve které je dostupných přibližně

12 000 odkazů na zdroje pro akademickou sféru. Jednotlivé odkazy jsou pravidelně kontrolovány a každý týden k nim přibývají nové. BUBL LINK tyto odkazy pečlivě vybírá a jeho úlohou je tak i kvalitativní kontrola obsahu adresáře.

Deweyho třídění ve zmíněných třech úrovních není úplné – některé kategorie jsou sloučeny a jiné naopak více specifikovány (např. ve třídě 400 „Language" je dostupná Švédština pod kódem 439.7). Pravděpodobným důvodem těchto drobných odlišností jsou dokumenty, které je na internetu možné nalézt, a které jsou zatříděny do tohoto adresáře. I když BUBL LINK neobsahuje kompletní hierarchickou strukturu DDC, jedná se o velmi zdařilou aplikaci tohoto třídění pro dokumenty v prostředí internetu.

5.5 LIBRARY OF CONGRESS CLASSIFICATION (LCC)

Třídění Kongresové knihovny Spojených Států (LOC) vychází z tzv. *expanzivního třídění* Charlese Ammise Cuttera. Systém byl mezi léty 1899 až 1940 přizpůsoben potřebám Kongresové knihovny a upraven také s ohledem na tematiku fondu této knihovny. Třídění je postaveno na alfanumerickém základě, obsahuje 21 tříd (A-Z) s několika výjimkami (I,O,W,Y,X).

To, že je třídění Kongresové knihovny tolik rozšířené, není ani tak zásluhou kvality třídění, jako spíše celkového vlivu této knihovny na zpracování a katalogizaci knih. U veškerých zpracovaných záznamů knih a dalších dokumentů uvádí LOC notaci svého třídění společně s notací Deweyho desetinné klasifikace. Vzhledem k potenciálním problémům s převodem klasifikační notace je proto pro mnoho organizací výhodnější přejmout kompletní bibliografický záznam včetně věcného zpracování[UKOLN, 1997c].

Je zajímavé, že zpracování dokumentů na internetu za pomoci třídění LOC je řešeno také v klasických knihovnách,[21] které v katalozích registrují i webové stránky. Tyto záznamy jsou však zpravidla ve stejné kategorii jako záznamy elektronických zdrojů a jejich rozsah je značně omezený.

5.5.1 Library of Congress Subject Headings (LCSH)

http://www.loc.gov/cds/classweb/

Předmětová hesla Kongresové knihovny Spojených států jsou nejrozšířenějším řízeným slovníkem s univerzálním záběrem[22]. Slouží jako doplněk hierarchické klasifikaci Library of Congress Classification (LCC), kde upřesňují předmětovými hesly tématiku dokumentu, zařazeného do kategorie LCC. Tato předmětová hesla jsou používána jako hlavní či doplňková třídící soustava v několika projektech – např. INFOMINE, FAST nebo BUBL LINK.

Při srovnání rozšíření LSCH v projektech s rokem 1997, kdy vyšla podrobná studie rozšíření tradičních klasifikačních schémat v prostředí internetu [UKOLN, 1997c], se zdá, že počet projektů, užívajících LCSH jako hlavní klasifikační schéma se výrazně zmenšil. Problémem může být počet předmětových hesel, který se pohybuje v řádu několika desítek tisíc a také licenční podmínky pro jejich užití. LCSH je produktem Kongresové knihovny, která použití tohoto třídění licencuje. Některé projekty mohou získat časově omezenou výzkumnou licenci, po jejímž vypršení už toto třídění nesmí být dále v projektu používáno.

5.5.2 Cyberstacks (Project Aristotle)

http://www.public.iastate.edu/~CYBERSTACKS/homepage.html

---

21 Cardinal Strich College Library (http://library.stritch.edu/)

Seattle Pacific University Library (http://www.spu.edu/depts/library/).

22 Přehled tříd je dostupný např. na stránce: http://fantasia.cs.msstate.edu/lcshdb/index.cgi.

Projekt CyberStacks, provozovaný na Iowa State University, je sbírkou dokumentů, nalezených na webových stránkách a tříděných za pomoci klasifikace Kongresové knihovny. Dokumenty jsou řazeny pod jednu nebo více kategorií tohoto třídění a je k nim přiřazena i stručná anotace. Většinu záznamů tvoří monografie, seriály, samostatné soubory nebo i záznamy vyhledávacích služeb. Projekt se zároveň zaměřuje na kvalitní a ověřené dokumenty, a tak veškeré záznamy odkazují na plné texty, pocházející především z akademického prostředí [McKiernan, 1999a].

Autor projektu – Garry McKiernan se zároveň snaží o zapojení součástí Iowa State University do procesu recenzování jednotlivých zdrojů, případně i do dalšího rozvoje Cyberstacks. V současné době se rozvíjí spolupráce s **Center for Indigenous Know-ledge for Agriculture and Rural Development** (CIKARD) a s **International Institute of Theoretical and Applied Physics** (IITAP), dvěma mezinárodními výzkumnými centry, které jsou umístěny na Iowa State University.

Pro třídění se používá zkrácená notace Kongresové knihovny, na jejímž základě je možné vybrat relevantní kategorii. Cyberstacks nezahrnuje kompletní třídění, ale omezuje se zatím na kategorie vědy a technologie. Uživatel se může v hierarchii volně pohybovat prohlížením jednotlivých tříd a podtříd klasifikace. Vzhledem k dostupným materiálům jsou kategorie šířeji definované a nemají za úkol naprosto přesně popsat obsah odkazovaných stránek, ale přibližně je zařadit do správné skupiny a usnadnit tak uživatelům orientaci. Pro přesnější vyhledávání pak lze využít další možnosti – „Cross-classification index“ (abecedně seřazený seznam všech dostupných kategorií) a „Title index“ (abecedně seřazený seznam názvů dokumentů, které jsou v Cyberstacks registrovány).

Záznam stránky nebo jiného dokumentu obsahuje klasifikační notaci, název kategorie v třídění, název a anotaci. Samostatně stojí za zmínku údaj „To search“, který informuje o možnostech vyhledávání, případně další navigace na stránkách.

Podobně jako ostatní projekty, počet záznamů v Cyberstacks je pouze v řádu několika tisíc. Projekt je pravděpodobně spravován opět jedním člověkem – Garry McKiernanem a aktualizace záznamů v databázi není jistá. Jediný časový údaj na stránkách projektu – říjen 1998 se vztahuje k poslední aktualizaci hlavní stránky. Není proto jisté, zda tento projekt i nadále pokračuje.

5.5.3 Electronic Reference Collection (E-Ref)

http://icrc.bloomu.edu/

Kolekce odkazů na referenční materiál na webových stránkách Harvey A. Andruss Library vznikla původně pod názvem jako „Internet Collegiate Reference Collection“ v roce 2002. Původně se jednalo pouze o malý seznam odkazů na referenční materiály (v říjnu 2004 obsahoval tento projekt pouhých 122 záznamů), který se později rozrostl až na 883 záznamům (stav k únoru 2007), který k listopadu 2008 klesl na 860 záznamů.

Pro třídění je používáno třídění Kongresové knihovny (LCC), které bylo původně použito pouze do druhé úrovně. Klasifikace další – třetí úrovně začala v roce 2003, kdy se stávající rozsah tříd ukázal jako příliš obecný. Nyní jsou všechny záznamy rozděleny do 68 tříd LCC. Z důvodu slučování některých tříd (např. kategorie zahrnující E200-E999) a vynechání tříd jiných se tak opět nejedná o úplnou aplikaci LCC.

Jak je patrné z celkového počtu záznamů, jednotlivé třídy obsahují velmi málo záznamů. Každý z nich obsahuje informace o názvu stránky a jejím tvůrci a abstrakt. Záznamy jsou doplněny třídou LCC, klíčovým slovem LCSH a názvem kategorie referenčního materiálu (seznamy, encyklopedie, časopisy apod.). Rozsahem ani kvalitou se tento projekt neliší od jiných a spadá spíše do kategorie průměrných.

5.6 MEZINÁRODNÍ DESETINNÉ TŘÍDĚNÍ (MDT)

Mezinárodní desetinné třídění začali vyvíjet Paul Otlet a Henri La Fontaine po roce 1895, kdy v rámci velkolepého projektu na registraci světového vědění založili Mezinárodní ústav pro bibliografii (dnešní **International Federation for Information and Documentation – FID)**. První verze tohoto třídění byla vydána mezi léty 1904 a 1907.

Přestože jsou kombinační schopnosti tohoto systému ve srovnání s Deweyho tříděním (DDC) a tříděním Kongresové knihovny (LCC) lepší, není tento systém rozšířen do takové míry jako jeho konkurenti. Důvodem je jeho menší podpora

ve srovnání s DDC (podporuje ho OCLC, která ho používá pro klasifikaci svých záznamů) nebo LCC (pro klasifikaci ho prosazuje Kongresová knihovna). Studie organizace UKOLN [1997] hodnotí toto třídění následovně: mezi jeho silné stránky počítá jeho rozšířenost, flexibilitu pro potřeby různých organizací a jednoduchost. Mezi jeho slabiny naopak uvádí pomalejší aktualizaci třídění, jeho komplexnost (tj. není vhodné pro klasifikaci pouze jedné problematiky) a slabší pokrytí některých tematických skupin (např. životní prostředí, medicína).

Nad další budoucností tohoto třídění visí otazník, neboť organizace FID de facto zaniká a slučuje se s **IFLA (International Federation of Library Association)**. O další vývoj MDT se bude dále starat nově založená organizace **Universal Decimal Classification Consortium**.[23]

Na trend vzájemného propojování a upravování klasifikací mezi sebou tak, aby byly navzájem kompatibilní (jednalo se především o snahy propojit MDT a LCC) upozornila počátkem 90. let Ingetraut Dahlberg [DAHLBERG, 1995], která kladla otázku, zda to není počátek unifikace velkých hierarchických klasifikací.

Projektů, které využívají Mezinárodní desetinné třídění je ve srovnání s jeho konkurenty poskrovnu. Konsorcium MDT uvádí na svých stránkách celkem osm projektů[24]. Ve skutečnosti ale dosud existuje (je v provozu) pouze polovina ze všech uvedených. Nejvýznamnějším z nich je GERHARD, který využívá MDT jako strukturu pro automatickou katalogizaci. Zbývající projekty často pouze využívají MDT jako jednu z navigačních možností. Projekt **The Social Science Information Gateway (SOSIG)** je spíše než univerzálním rozcestníkem oborovým portálem, který užívá části MDT jako hlavního nástroje pro prohlížení. Podobně funguje i projekt **Ibiblio**, které užívá MDT jako jedné ze dvou možností zobrazení hlavních kategorií.

Zajímavým poznatkem je odklon některých projektů od tohoto třídění. V minulosti ho využívali například **BUBL** (viz. kapitola 5.4.4), který dnes využívá třídění DDC nebo medicínský portál **OMNI**, který se přeorientoval na specializovanou soustavu třídění pro medicínu – na řízený slovník **MESH**.

5.6.1 Intute: Social Sciences (dříve SOSIG)

http://www.intute.ac.uk/socialsciences/

Služba Intute: Social Sciences vznikla původně pod jménem SOSIG jako informační brána pro společensko-vědní obory. Byla založena výzkumnými organizacemi **The Economic and Research Council (ESRC)** a **The Joint Information System Committe (JISC)**[25] za podpory Evropské unie v rámci telematického programu DESIRE[26]. Služba SOSIG se v roce 1999 stala součástí sítě Resource Discovery Network (RDN). Název **Intute: Social Sciences** vznikl po sloučení původního portálu **SOSIG** se zaměřením na společenské vědy s projektem **Altis** z University of Birmingham zaměřeným na oblast sportu a rekreace.

Podobně jako jiné informační brány chce i Intute informovat o kvalitních dokumentech na internetu, které jsou volně přístupné. Za zmínku stojí velmi přísná výběrová kritéria pro nové zdroje (např. úroveň odbornosti, absence reklamy, úroveň popisu zdroje). Jako vedlejší produkt těchto kritérií vznikl i manuál pro posuzování kvality dokumentu v rámci internetu, který je pod názvem *„Internet Detective"* volně dostupný[27].

Klasifikace pomocí MDT je dostupná formou hierarchie pro prohlížení kategorií. Hierarchie původního portálu SOSIG obsahovala 162 předmětových kategorií (subject headings), které byly rozděleny do šestnácti hlavních kategorií. Portál nepřevzal kompletní hierarchii, ale zaměřil se pouze na oblasti, které se vztahují k tematice tohoto portálu. Jednotlivé zde používané kategorie byly:

Economics, Developement

---

23 http://www.udcc.org/

24 http://www.udcc.org/internet.htm

25 Stejná organizace založila i projekt BUBL.

26 http://www.desire.org

27 http://sosig.ac.uk/desire/internet-detective.html

Feminism

Education

Environmental Issues

Ethnology, Social Anthropology

Geography

Government, Military Science

Law

Management, Accountancy, Business

Philosophy

Politics – International Relations

Psychology

Social Science General – Methodology

Social Welfare – Community, Disability

Sociology

Statistics – Demography.

Z tohoto výčtu je zjevné, že jednotlivé kategorie pocházely z různých úrovní hierarchie MDT a jejich vzájemné spojení je čistě pragmatické (Military 344, Government 35 apod.). Tento přístup byl pro klasifikaci kolekce dokumentů SOSIG (k září 2004 celkem 28 178 záznamů s odkazy) zcela postačující.

Sami autoři nepokládají třídění MDT za dokonalé a vytýkají mu i některé zásadní nedostatky – především pomalou aktualizaci třídící soustavy a nedostatečnou aktualizaci odborné terminologie, používané v třídění [UKOLN, 1997B]. Aplikace MDT pro potřeby SOSIG byla zvolena ze dvou hlavních důvodů. Prvním je široké užití tohoto třídění po celém světě, především pak v Evropě, kam je projekt zaměřen (součást kulturního a vědeckého rozvoje v rámci EU). Druhým důvodem je absence specializovaného oborového třídění pro tento soubor témat. Místo vyvíjení vlastního systému proto autoři raději volili nedokonalé klasické třídění a zaměřili se na hlavní poslání projektu – služby informační brány.

Radikální změna v oblasti třídění nastala po sloučení s projektem **Altis.** Záznamy již nejsou zařazovány do třídění MDT a jsou nyní tříděny především do odborných tezaurů General social science (HASSET), Government, politics and anthropology (IBSS), Social work and welfare (SCIE),Research methods & tools (SRM) a CAB thesaurus (CABI). V pokročilém vyhledávání je možné procházet jednotlivá hesla výše uvedených tezaurů a zobrazit záznamy, které jsou k nim přiřazeny.

Projekt Inute (dříve SOSIG) se snaží uplatňovat neformální standardy a proto převzal standard **ROADS (Resource Organisation and Discovery in Subject-based Services)**, který slouží pro vzájemnou interoperabilitu mezi informačními bránami. Další snahy směřují i do oblasti přidávání metadat k záznamům – používaným standardem je Dublin Core. SOSIG se patrně v blízké době přesune spíše k projektům automatizované klasifikace – uvažuje se o automatickém přidávání metadat k záznamům a o automatizovaném hledání a vyhodnocování nových zdrojů (technologie zvaná harvesting – viz kapitola 4.124.124.12s. 43). Od roku 1998 spolupracuje s projektem automatizované analýzy stránek COMBINE. Projekt SOSIG ale nikdy nebude zcela odkázán na automatizované zpracování záznamů. Autoři sami uvádějí, že pravděpodobně nikdy nebude technologie natolik dokonalá, aby byla schopna analyzovat stránky stejně kvalitně jako člověk. Nejdůležitějším kritériem pro zpracování záznamů je jejich kvalita, a proto se i do budoucna počítá s kombinací lidského a strojového zpracování.

5.6.2 Ibiblio

http://www.ibiblio.org

Ibiblio je hybridním řešením mezi aplikací klasického třídění a webovým katalogem. Projekt je provozován za spolupráce **Center for the Public Domain** a **MetaLab, University of North Carolina – Chapel Hill**, která dříve existovala pod názvem **SunSITE**. Podle vlastního vyjádření jde o *„jednu z největších kolekcí volně dostupných informací“* [IBIBLIO, 2001].

Tento projekt se skutečně zaměřuje na informace v různých podobách a tak zde najdeme záznamy textových dokumentů, hudebních souborů nebo softwaru. Knihovna má také zajímavé partnery, mezi které patří VA Linux's SourceForge a společnost IBM, která pro celý projekt dodává hardware.

Hlavním cílem je propagace a podpora volně šiřitelného softwaru (tzv. Open Source) a podpora jejich komunit, hostování vývojových projektů a publikování informací s touto tematikou.

Pro třídění a navigaci je kromě vyhledávání použito také MDT, jako alternativní prostředí pro prohlížení odkazů, které je možné přepnout z výchozího nastavení vlastních deseti kategorií.[28] Tato klasifikace je zde aplikována pouze v první úrovni, druhá úroveň je již tvořena volnými předmětovými hesly, které charakterizují skupiny nabízených odkazů (např. Cartoons, Architecture). Některé odkazy vedou z druhé úrovně přímo na cizí stránky, jiné odkazují do vlastního FTP archivu. Na projektu je patrné, že sbírka dokumentů je pouze jednou z nabízených služeb a proto nelze očekávat pokročilé řešení pořádání. Přesto je projekt velmi zajímavý pro svůj přístup ke klasifikaci různých typů materiálů dohromady. Nezodpovězenou otázkou zůstává proč tvůrci tohoto projektu upřednostnili MDT namísto pro USA obvyklejšího Deweyeho desetinného třídění.

5.6.3 Katalog OKO

http://www.zrc-sazu.si/oko/

Slovinský projekt OKO je dílem inženýrky chemie Nadi Celia, která sama tento projekt spravuje [CELIA, 2002]. Projekt se zaměřuje na záznamy kvalitních dokumentů ze všech tematických oblastí MDT. Mezi registrovanými formáty jsou například archivy, databáze, seznamy, odkazy i audiovizuální dokumenty.

Jde o aplikaci základních devíti tříd MDT, které dále vedou do druhé až páté úrovně hierarchie (liší se v závislosti na třídě). V poslední dostupné úrovni je seznam záznamů o dokumentech, přístupných na internetu. Ty mají dvojjazyčné popisky ve slovinštině a v angličtině, avšak obsah hesel a abstrakt je pouze slovinsky.

Na rozdíl od mnoha podobných projektů je OKO projektem stále aktivním a pravidelně aktualizovaným, i když je, co se počtu odkazů týče, projektem velmi malého rozsahu.

5.7 OSTATNÍ UNIVERZÁLNÍ HIERARCHICKÉ KLASIFIKAČNÍ SYSTÉMY

Prostředí internetu neodráží plně množství hierarchických třídění, které jsou používány pro třídění v knihovnách a dalších institucích. Některé existující tradiční klasifikační systémy mají komplikovanou strukturu, kterou je obtížné převést do prostředí internetu [ELLIS, 1999], jiné mají pouze lokální význam.

Na internetu je zpřístupněno i několik menších hierarchických třídění. Některá z těchto třídění jsou zpřístupněna pouze v podobě knihovního katalogu nebo přehledu tříd, jiná jsou zakomponována do různých projektů, které klasifikují dokumenty, přístupné na internetu.

5.7.1 Regensburger Verbundklassifikation (RVK)

http://www.bibliothek.uni-regensburg.de/Systematik/systemat.html

Třídění regensburské knihovny je jedno z nejmladších hierarchických třídění - vzniklo v roce 1965. Toto třídění je alfanumerické – první hierarchická úroveň je označena písmeny abecedy A- ZY a rozvíjí se do dalších úrovní. Další třídy jsou definovány buď prostou dvojicí písmen nebo jako rozsah těchto kódů – například „CL – CZ“ pro třídu Psychologie. Třetí – poslední úroveň hierarchie je v podobě alfanumerického kódu – například „CM 4000“ pro oblast statistiky v psychologii. Tento kód je vlastně složen ze dvou částí – jde o sloučení dvou tříd: třídy CL – CZ Psychologie (definice tématu) a CM Allgemeines. Geschichte und Methodik (definice typu dokumentu).

[28] Skupiny MDT, které jsou řazeny abecedně za sebou od první úrovně.

Třídění je přístupné na webových stránkách zatím pouze jako kompletní struktura k prohlížení, pomocí které je možné vyhledávat v knihovním katalogu regensburské univerzity. Toto třídění je svou jednoduchostí a kombinačními možnostmi velmi zajímavé a budoucnost ukáže, zda je možné jeho další rozšíření mimo univerzitu v Regensburgu.

5.7.2 Dutch Basic Classification (BC): Nederlandse Basisclassificatie (BC), projekt Dutch Electronic Subject Service (DutchESS)

http://www.kb.nl/dutchess

Dutch Basic Classification (BC) je národním nizozemským klasifikačním schématem, který byl vytvořen pro sdílenou klasifikaci v systému PICA, většinou užívaným nizozemskými veřejnými a akademickými knihovnami. Klasifikace byla vytvořena Nizozemskou národní knihovnou v Kodani mezi léty 1996-1998. Hlavními úkoly této klasifikace je poskytovat stejné klasifikační schéma pro všechny knihovny, zapojené do systému PICA a koordinovat nákup a vytváření knihovního fondu v účastnických knihovnách.

Dutch Basic Classification (BC) existuje již ve své třetí verzi, která je z roku 1998. Třídění se skládá ze 48 hlavních tříd, označenými čísly od 01 do 89. Každá tato třída obsahuje podtřídy, které jsou od hlavního kódu odděleny tečkou a označeny dvojciferným číslem od 00 do 99. Výsledný klasifikační kód tak vypadá následovně:

18.56 Czechoslovakian language and/or literature

Podobně jako další projekty se i DutchESS zaměřuje na klasifikaci kvalitních, volně přístupných dokumentů, které jsou dostupné na webových stránkách. Projekt používá třídění Dutch Basic Classification (BC), které je aplikováno do druhé hierarchické úrovně, třetí úrovní jsou odkazy na dokumenty v kategoriích. Záznamy obsahují základní bibliografická data (autor, název, popis, typ dokumentu, BC kód) a přímý odkaz na dokument.
Webové stránky projekt DutchESS byly na počátku roku 2007 staženy a celý projekt byl pravděpodobně z důvodů chybějících financí ukončen.

5.8 ZHODNOCENÍ APLIKACE UNIVERZÁLNÍCH TŘÍDĚNÍ NA INTERNETU

Univerzální třídění byla vytvořena s ambicemi univerzálního záběru celosvětového poznání, což je i při nejlepší snaze úkol nesplnitelný. Implementace univerzálních třídění do prostředí internetu je většinou pouze částečná a tak nedochází k úplnému převodu všech tříd a podtříd.

Hanne Albrechtsen [ALBRECHTSEN, 1998] upozorňuje v souvislosti s univerzálními tříděními na zajímavý paradox: „Klasifikační systémy byly často standardizovány aby podpořily meziknihovní spolupráci. To mělo za následek, že se výzkum klasifikačních systémů soustředil na vývoj univerzálních třídění, které by mohly být zavedeny v centrálních agenturách za účelem řízení organizace znalostí v knihovnách. Jako výsledek takové standardizace se klasifikace stala „neviditelnou prací“, která **je vykonávána bez ohledu na potřeby místní komunity uživatelů.“**

Implementace těchto systémů do prostředí internetu tak jen zvýrazňuje jejich omezení, mezi která patří především:

Údržba a rychlost změn

Údržba složitých soustav a pravidelná aktualizace dokumentů představují pro aplikaci klasických třídění největší překážku. Proto se některé projekty snaží aplikovat tradiční klasifikační soustavu, s využitím automatizovaných procesů – především automatickou klasifikací. Klasifikace přichází vždy pozdě, nicméně časový odstup mezi změnou a jejím zachycením v klasifikačním systému se dá zkrátit. I pro tištěné dokumenty jsou však tyto změny velmi pomalé a v prostředí internetu se problém stává ještě zřetelnějším.

Komplexnost, která je na překážku dynamické změně situace na internetu

Jak je patrné z většiny projektů, úplná aplikace celého univerzálního klasifikačního systému je příliš náročná na to, aby mohla být některým projektem dokončena. Většina z nich se pokouší převzít pouze malé části klasifikace, nicméně i ty je náročné udržovat bez pomoci automatizovaného systému aktuální. V případě automatizovaných systémů klasifikace (viz kapitola 7.4, s. 92) je aktualizace lepší, ale ani ony zatím nejsou schopny převzít kompletní údržbu klasifikační struktury se všemi třídami a podtřídami.

Rozsah práce a náklady potřebné k aplikaci celých univerzálních systémů pro internet jsou příliš velké vzhledem k předpokládanému výsledku a využití těchto třídění. Zatím jediný celý univerzální klasifikační systém – DDC – převádí do elektronické podoby, vhodné pro automatickou klasifikaci, pouze společnost OCLC, jejíž součástí je vydavatel a majitel autorských práv k tomuto třídění, společnost Forest Press.

Dynamika změn ve výzkumu a v terminologii

Výzkum přináší velmi rychle nové poznatky, které je potřeba co nejdříve zachytit v klasifikační soustavě. Univerzální klasifikační systémy jsou příliš komplexní a složité na to, aby tyto změny probíhaly pravidelně. Oborová třídění jsou na změny v terminologii daleko lépe připravena po stránce soustavy třídění (jednoduché soustavy, zpravidla tezaury) i po stránce odborné. To se týká i zachycení interdisciplinárních témat, která vznikají nově (např. biotechnologie, medicínská informatika).

Výčet těchto problémů není vyčerpávající, Traugott Koch [KOCH, 1998] uvádí i další problematická témata spojená s tradičními klasifikacemi – často se objevuje nelogické nebo nesystematické řazení kategorií a notace třídění jsou zbytečně složité.

Steve Steinberg se domnívá, že současné univerzální klasifikační systémy prožívají krizi:„Dokonce i knihovníci připouští, že současně užívaná schémata jsou zastaralá a neadekvátní; z fráze „klasifikace v krizi“ se stalo v knihovnické komunitě klišé.“ [STEINBERG, 1996] Jeho názor je vůči těmto systémům velmi kritický, a možná i ne úplně oprávněný. Po několika desítkách let jsou tyto systémy velmi precizně vypracovány a pravděpodobně neexistuje jiná forma pořádání informací, která by zachycovala tak široké spektrum témat společně s jejich komplexními vazbami.

Univerzální klasifikační systémy přežily mnohé změny, ale jejich význam se mění a jako přístup k pořádání informací v prostředí internetu jednoznačně nebudou mít rozhodující vliv. Po ztrátě vůdčího postavení pro pořádání informací v oblasti tradičních dokumentů, kde je zatlačila do pozadí oborová pořádání, budou muset tyto systémy hledat svoji novou roli i v dynamickém prostředí internetu.

5.9 TEZAURY A ŘÍZENÉ SLOVNÍKY

Tezaury a řízené slovníky jsou významnou formou organizace informací především pro dokumenty v elektronické podobě. K jejich velkému rozšíření došlo současně s budováním elektronických oborových databází (např. MeSH, Compendex), kde začaly být užívány jako hlavní pořádací schéma. Řízené slovníky a tezaury se používají pro pořádání informací různými způsoby. Některé mohou být zamýšleny k prohlížení hierarchické struktury, zatímco jiné obsahují pouze prostý abecední seznam povolených termínů [RUSSELL, 2001]. Podle formy jejich užití je můžeme rozdělit do tří skupin:

- hesláře
- řízené slovníky
- tezaury

První skupina – **hesláře** jsou jednoduchými seznamy předmětů/popisných termínů, které jsou zpravidla seřazeny abecedně nebo numericky. Je to nejjednodušší způsob omezení množství možných popisných termínů pouze na definovanou skupinu, která může mít i formu jednoduché hierarchie – první úroveň definuje tématickou skupinu, úroveň druhá je seznamem odkazů. Tento způsob bývá často používán u různých oborových bran a rozcestníků, kde jsou takto definovány širší kategorie podle svého obsahu.

Řízené slovníky obsahují seznam termínů, které mohou být použity pro popis dokumentu. Mohou být vyčerpávajícím seznamem popisných termínů (například databáze EBSCOhost) nebo mohou být uváděny v hierarchii. Řízený slovník je přechodem mezi nejjednodušší formou předmětového pořádání – seznamy předmětů a tezaurem jako formou pořádání nejpokročilejšího.

Pro řízené slovníky je typická tzv. **prekoordinace**. Jako prekoordinovaný se označuje systém indexování, kde jsou termíny a fráze předem určeny, a kde hledání probíhá pouze pomocí těchto výrazů, tj. pokud není použita přesná fráze, kterou je dokument označen, nebude nalezen.

Tezaury jsou typem řízeného slovníku, který ve svém seznamu termínů a výrazů, povolených pro popis dokumentů umožňuje **definovat vztahy mezi těmito termíny a frázemi**. Tyto vztahy mohou být definovány ve formě hierarchie (nadřízený, podřízený termín) nebo ve formě odkazů (preferovaný termín, viz. též apod.). Tezaury jsou velmi využívanou formou pořádání pro tyto vlastnosti – na rozdíl od hierarchických klasifikací jsou velmi flexibilní a umožňují rychlou změnu (rozšíření, vyřazení termínu).

Tezaury jsou typem **postkoordinovaného pořádání,** tj. takové, kde se vyhledává na základě termínů z rotačního slovníku a je nalezen jakýkoli dokument, obsahující všechny hledané termíny. Nemusí se jednat o přesnou frázi, budou nalezeny jakékoli dokumenty, které obsahují ve svých popisných termínech tato slova v jakémkoli kontextu. Oproti řízenému slovníku mohou být termíny v tezauru zařazeny do více hierarchických větví (tzv. možnost polyhierarchie), definujících preferované výrazy (tj. ty, které mají být použity pro popis) a asociované termíny (termíny související s popisovaným termínem).

5.9.1 Výhody a omezení

Hlavní výhodou užití tezaurů a řízených slovníků je omezení šíře popisných termínů, tj. je možné zvolit pouze ty, které jsou uvedeny ve slovníku. Nevýhodou je obecnost termínů – proto, aby byla zajištěna jednotnost při obsahové klasifikaci je záměrně potlačen specifický popis dokumentů a některé významné charakteristiky obsahu tak nejsou uvedeny.

Na rozdíl od hierarchických soustav poskytují předmětová třídění pouze omezený kontext dalších předmětů a oborů. Znamená to, že tato třídění se soustřeďují na menší, zpravidla tematicky ohraničené oblasti, a vztahy k dalším oborům se nezabývají. Možnost znázornění vzájemných vztahů mezi předměty formou vizualizace je velmi omezená a nelze ji srovnávat s potenciálem velkých hierarchických systémů.

Karl Fast [FAST, 2002] definuje tento přístup k pořádání následovně: „Řízený slovník je způsobem jak vložit interpretační vrstvu sémantiky mezi termín vložený uživatelem a databází tak, aby lépe vystihoval záměr uživatele.” Řízené slovníky a tezaury jsou přechodem mezi strukturovanou klasifikací (přístup hierarchických klasifikací) a hledáním ve volném textu (přístup vyhledávačů). Jak je uvedeno v předchozí citaci – řízené slovníky jsou jakousi sémantickou vrstvou, která dodává hledání slov význam. Tento přístup je efektivní a zároveň jednoduchý a adaptabilní. Oproti vyhledávání v plném textu umožňují řízené slovníky nalézt skutečně hledané dokumenty, aniž by výsledky hledání ovlivnila slova, která se vyskytují nahodile u sebe. To jsou také důvody velkého rozšíření těchto forem pořádání informací, které jsou i v prostředí internetu rozšířenější než hierarchické pořádací systémy.

Řízené slovníky mají pro svou flexibilitu velikou budoucnost, zároveň však pokračuje vývoj sofistikovanějších systémů, které by mohly lépe interpretovat procesy lidského poznávání a porozumění konceptů, které se za slovy skrývají. Projekty jako například *WordNet* [WORDNET, 2004] se tak snaží současný systém předmětových třídění posunout ke komplexnímu porozumění procesu lidského poznávání a vyvinout systémy, které by pracovaly na základě stejných principů jako lidský mozek. V budoucnu lze v předmětových tříděních očekávat ještě větší vliv z poznatků disciplín kognitivní psychologie a studia umělé inteligence, které by tyto přístupy k pořádání informací posunuly k vyšší přesnosti a uživatelské ergonomii.

5.9.2 Hesláře

5.9.2.1 Medical biochemistry subject list

http://web.indstate.edu/thcme/mwking/subjects.html

Seznam Medical biochemistry můžeme označit za kombinaci výkladového slovníku s oborovým rozcestníkem. Hlavní seznam nabízí odkazy ve formě přibližně padesáti biochemických témat. Tento seznam vede do druhé úrovně, kde je nabídka plných textů o tématu společně s odkazy na příbuzné stránky. Texty k hlavním tématům jsou rozděleny na části, na které je možné se dostat pomocí odkazů z horní části stránky. Hlavním účelem stránky je informovat o definovaných tématech, což se projevuje i na počtu odkazů na jiné stránky, kterých je v porovnání s jinými rozcestníky velmi málo.

5.9.2.2 U.S. Census Bureau subject index

http://www.census.gov/main/www/subjects.html

Statistický úřad USA využívá tento abecední seznam témat jako navigační pomůcku pro hledání v seznamu dokumentů na svých stránkách. Seznam je členěn podle písmen abecedy a pod hlavním heslem odkazuje na podrobnější témata (např. Decennial Census: – 1990 Census). Zároveň jsou v tomto seznamu využívány i jednoduché odkazy na další příbuzné tematické oblasti, které jsou označeny jako „see also“.

5.9.2.3 PINAKES Subject list

http://www.hw.ac.uk/libWWW/irn/pinakes/pinakes2.html

Projekt PINAKES, který je pojmenován podle katalogu, ve kterém básník Kalimachos sepsal knihy v Alexandrijské knihovně, je seznamem oborových rozcestníků a předmětových bran. Nabízí dvě rozhraní – abecední a tematický seznam. Většina odkazovaných rozcestníků a bran byla vytvořena v rámci rozvojových programů Evropské unie (**Information Society Technologies Programme**) a grantového programu Velké Británie **eLib**. Kromě těchto projektů seznam odkazuje i na další multioborové předmětové brány, které jsou provozovány v rámci akademických nebo komerčních projektů.

5.9.2.4 Federal Information Subject List

http://docs.lib.duke.edu/federal/guides/fed_sub.html

Projekt Perkins Library na Duke University (USA) nabízí jednoduchý seznam oficiálních stránek vládních úřadů USA, tematicky řazených podle jejich funkce a náplně práce. Kromě těchto zdrojů jsou v seznamu uvedeny i neoficiální a komerční stránky, pokud poskytují přínosné informace. Jednotlivé skupiny jsou široce definovány (např. Business, Economics and Labor, Education, Environment) z důvodu množství témat a specializací.

5.9.3 Řízené slovníky

5.9.3.1 GEM (Gateway to Educational Materials): educational resources

http://thegateway.org/

GEM je konsorciálním projektem, který si klade za cíl zpřístupnit zdroje pro vzdělávání ze státních, univerzitních i komerčních stránek. Materiály, přístupné prostřednictvím GEM, poskytuje řada organizací, podílejících se na projektu. V březnu 2007 bylo zpřístupněno na 35 000 různých materiálů pro vzdělávání, které poskytuje přes 600 členů konsorcia.

K prohlížení existuje více slovníků: subject (předmětová hesla), type (typ materiálu), level (věková skupina, pro kterou je materiál určen), keywords (klíčová slova), mediator (zprostředkovatel dokumentu – např. dokument je určen pro učitele, který z něj bude učit), beneficiary (skupina konečných příjemců, pro kterou je materiál určen – např. rodiče nebo studenti z menšinových skupin), price code (informace o případných poplatcích za využití materiálu). Slovníky zde slouží pro zobrazení a vyhledání všech uvedených údajů. Záznamy dokumentů jsou strukturovány do tzv. polí podle typu údaje stejným způsobem, jako je tomu u většiny databází. Veškeré výrazy, obsažené v některém ze slovníků, jsou odkazem, který vede k vyhledávání dalších záznamů stejného typu.

5.9.3.2 DeCS Health science descriptors

http://decs.bvs.br/I/decswebi.htm

DeCS – Health Sciences Descriptors je strukturovaným slovníkem z oblasti medicíny a zdravotní péče, zpracovaným ve třech jazycích (anglicky, portugalsky, španělsky). Tento slovník byl vytvořen na základě nejznámějšího zdravotnického strukturovaného slovníku MeSH s cílem poskytnout jednotné vyhledávací rozhraní a terminologii ve třech jazycích. DeCS slouží k indexaci článků ze všech typů vědeckých publikací a k vyhledávání například v databázích LILACS a MEDLINE.

K březnu 2007 obsahuje DeCS celkem 26 851 hlavních deskriptorů, ze kterých je 3 656 z oblasti veřejného zdravotnictví a 1950 z oblasti homeopatie. Slovník je vytvořen ve stromové struktuře která má v první úrovni sedmnáct hlavních hesel. Tato struktura se dále rozvíjí podle tematického záběru přes několik dalších úrovní až k seznamu deskriptorů. Každý deskriptor je ve všech třech jazycích a obsahuje stručný abstrakt. Kromě abstraktu je každý deskriptor označen třemi identifikačními údaji: kódem větve stromové struktury (Tree number), pořadovým číslem záznamu a unikátním identifikátorem záznamu.

5.9.3.3 Biosis

http://scientific.thomson.com/support/products/previews/bcv/

Strukturovaný slovník Biosis je používán pro indexaci dokumentů v databázích zaměřených na přírodní vědy. Tento slovník obsahuje termíny, databází BIOSIS Previews, Biological Abstracts and Biological Abstracts/RRM používané od roku 1993. Veřejně dostupný je seznam hesel ve dvou úrovních.

5.9.3.4 Geo-guide subject catalog

http://www.sub.uni-goettingen.de/cgi-bin/ssgfi/navigator.pl?db=geo&type=subj

Projekt Univerzity v Göttingenu (Německo) je slovníkem, zaměřeným na oblast geologie, geografie, hydrologie a příbuzných oborů. Základní úroveň prohlížení je rozdělena na osmnáct hlavních kategorií, které vedou k podrobněji specifikovaným kategoriím ve druhé úrovni. I když se jedná o strukturovaný slovník, některá další třídění jsou zde využívána jako doplňková. Je zde použito univerzálního hierarchického třídění **Göttinger Online Klassifikation (GOK)** jako hlavního doplňkového navigačního systému. Toto třídění je v rejstříku uváděno v převodních tabulkách s ostatními tříděními – Mezinárodní desetinné třídění (MDT), Deweyho desetinné třídění (DDC), Basis-Klassifikation des GBV (BK), pro usnadnění orientace mezi třídami a jejich slovním popisem. Pro odborníky je přínosem i seznam zkratek a akronymů organizací, který je uváděn jako další doplňkový rejstřík.

5.9.3.5 MedWeb @ Emory University

http://www.medweb.emory.edu/MedWeb/default.htm

MedWeb je katalogem stránek z oborů medicíny, biomedicíny a zdravotnictví, spravovaným knihovnou Robert W. Woodruff Health Sciences Center Library na Emory University (USA, stát Georgia). Katalog je primárně určen pro akademickou komunitu samotné univerzity a tomu odpovídá i členění hesláře. Ten je rozdělen do

121 kategorií, které mohou obsahovat druhou úroveň nebo již zobrazují vlastní odkazy na webové stránky.

5.9.4 Tezaury[29]

5.9.4.1 UNESCO thesaurus

http://www.ulcc.ac.uk/unesco/

Tezaurus UNESCO byl vytvořen pro indexaci dokumentů v archivech a databázích stejnojmenné mezinárodní organizace. Jeho první verze pochází z roku 1977, aktuální druhá verze byla publikována v roce 1995. Tezaurus je v angličtině, ale obsahuje i francouzské a španělské ekvivalenty anglických termínů.

UNESCO tezaurus se skládá z menších tezaurů (tzv. mikrotezaurů), které se tematicky věnují určené problematice. Těchto mikrotezaurů je sedm (education, science, culture, social and human science, information and communication, countries and country groupings) a k jejich vzniku vedla snaha o přehlednou orientaci v tematice.

Tento tezaurus používá pro definování vztahů mezi předměty všechny výrazové prostředky (tj. vztahy hierarchické, odkazy na preferované termíny a odkazy na synonyma). Celkem obsahuje přibližně 6600 termínů (deskriptorů a nedeskriptorů).

5.9.4.2 Humanities And Social Science Electronic Thesaurus (HASSET)

http://www.data-archive.ac.uk/search/hassetSearch.asp

HASSET byl vytvořen pro klasifikaci v katalogu UK Data Archive – mezinárodního centra, zaměřeného na uchovávání a šíření digitálních dokumentů z oblasti společenských věd. Toto centrum se zaměřuje především na archivaci oficiálních dokumentů, které ve svém nabídkovém systému zájemcům prodává.

HASSET sestává z 21 předmětových kategorií, vedoucím přímo k záznamům dokumentů a jejich objednávce. Přestože je HASSET označován jako tezaurus, vztahy mezi

předměty a skupinou jsou definovány pouze jednoduchou dvouúrovňovou hierarchií. Vzhledem k absenci jakýchkoli výrazových prostředků tezaurů je proto HASSET spíše jednoduchým řízeným slovníkem s jednoduchou hierarchií.

5.9.4.3 High-level thesaurus (HILT)

http://hilt.cdlr.strath.ac.uk/

Projekt HILT má v této oblasti zvláštní postavení; nejedná se o tezaurus jako takový, ale o výzkumný projekt, který se zabývá propojováním tezaurů a dalších klasifikačních schémat mezi sebou. Podobně jako řada podobných projektů je HILT financován z prostředků rozvojové agentury Velké Británie **JISC (Joint Information Systems Committee)**. Tento projekt se zaměřil především na vzájemnou převoditelnost různých tezaurů a řízených slovníků tak, aby bylo možné převádět odbornou terminologii do různých forem a typů klasifikací.

Projekt HILT probíhal ve dvou fázích. První fáze měla jako hlavní cíl zmapování používaných třídění a metodickou přípravu pro konkrétní technické řešení. V rámci druhé fáze projektu, označené jako **HILT Phase II**, byl vytvořen server pro převod terminologií mezi různými typy klasifikace. Server by v první fázi měl rozpoznat terminologii DDC, LCSH a tezauru UNESCO. Jejich převod je řešen technologií „Machine 2 Machine“(M2M), která zajišťuje tento převod automaticky. Problémem, který se při převodu objevuje, je rozdílná granularita třídění mezi sebou. Převod je vždy pouze určením nejbližšího příbuzného termínu a není možné předpokládat, že pokaždé bude tento převod úplně výstižný.

Projekt HILT skončil k listopadu 2004 a prokázal možnost převodu předmětových i hierarchických třídění (DDC) mezi sebou. V rámci projektu bylo vytvořeno koncepční i technické řešení těchto převodů. Tyto převody různých typů třídění mezi sebou, jak autoři sami dodávají [HILT, 2004], musí mít především pro institucionální uživatele vyšší přínos než náklady, a tak jsou ekonomická hlediska často rozhodujícím faktorem pro volbu typu třídění.

[29] V textu uvedené příklady tezaurů jsou pouze ilustrativní. Podrobnější seznam dostupných tezaurů a oborových klasifikací je dostupný na: http://sky.fit.qut.edu.au/~middletm/cont_voc.html

5.9.4.4 Finnish Virtual Library (Jyväskylä Virtual Library)

http://www.jyu.fi/library/virtuaalikirjasto/engroads.htm

http://www.linkkitalo.fi/

Tento projekt nabízí několik tematicky zaměřených tezaurů, které jsou zde použity jako třídící soustava pro kvalitní dokumenty, dostupné v prostředí webových stránek. Tyto tezaury jsou dostupné v angličtině a finštině. Každý záznam stránky obsahuje odkaz na stránku a seznam nadřazených a podřazených termínů. Svým rozsahem nejsou tyto tezaury nijak veliké – například tezaurus z oboru „Sport science" obsahuje 184 záznamů, tezaurus „Hydrobiology and limnology" pouhých 35. Od roku 2005 je projekt dostupný pouze v rámci projektu **„Science Linkhouse"** (http://www.linkkitalo.fi/).

5.10 OBOROVÉ KLASIFIKACE

Pořádání informací pomocí taxonomií není pouze doménou knihovníků a informačních specialistů. Význam taxonomie, případně návazných kódů je uznáván i v jiných oborech, které tímto způsobem řeší otázku navigace ve svých dokumentech se specializovanou tématikou.

Tyto klasifikace se zaměřují pouze na vlastní obor, a proto nejsou tak rozsáhlé a tematicky komplexní jako třídění knihovnická. Ukazuje se zde pragmatická povaha těchto systémů – roztřídit dokumenty do jednotlivých tříd bez složité hierarchie. Pro komplikovanější třídění (např. chemie) jsou namísto taxonomické klasifikace používány systémy kódů, které zajišťují dohledatelnost informací a přitom nekladou tak vysoké administrativní nároky na inovaci a údržbu (zařazení nových témat apod.) ve srovnání s hierarchickou klasifikací. Pro třídění jsou také využívány tezaury, které jsou dostatečně flexibilní pro zachycení změn v poznání a zároveň ukazují základní hierarchické vztahy mezi předměty.

Oborové klasifikace vznikají pod vedením **profesních organizací** (např. American Mathematical Society), **oborových portálů** (EEIS), **časopisů** (Journal of Economic Literature) nebo přímo jako **iniciativa mezinárodní** (ekonomické kódy NAICS). Rozšíření těchto klasifikačních schémat je omezeno vlastním oborem, nicméně tam se tyto klasifikace stávají respektovaným standardem, který je pro přesné zatřídění dokumentu významnější než komplexní knihovnické klasifikace. Především přírodní vědy (medicína, chemie, matematika...) takto vytváří standardy komunit, které zefektivňují vzájemnou odbornou komunikaci.

Vůdčí postavení v tomto vývoji mají oborové organizace v USA. Jde hlavně o oborová výzkumná centra jako *National Library of Medicine* (NLM) nebo *National Library of Agriculture* a také o americká oborová sdružení, která jsou svým významem i mezinárodní – *American Chemical Association* (ACM) nebo *American Mathematical Society* (AMS). Některé z těchto organizací také produkují specializované databáze, pro které dále používají svá třídění.

V rámci internetu (především webových stránek) jsou tyto systémy využívány v oborových digitálních knihovnách a volně přístupných i placených databázích. Využití těchto třídění se ani v prostředí internetu nijak neliší od prostředí tištěných dokumentů, a tak se zpravidla jedná o úplnou aplikaci třídění bez jakýchkoli změn. Vzhledem k jednoduchosti těchto soustav se pro jejich prohledávání používají plnotextové vyhledávače a automatizované zpracování dokumentů pro ně zatím není aktuální. Možným důvodem je nižší počet publikovaných prací na webových stránkách ve srovnání s obory společensko-vědními a humanitními. V oblasti přírodních věd jsou zpravidla vyšší nároky na kvalitu publikovaných prací a tak s výjimkou oblasti výpočetní techniky, většina autorů dává přednost vydání článku v časopise tištěném. Tak je většina článků popsána v databázích, a popis dokumentů, původně zveřejněných na webových stránkách, tvoří pouze menší část práce.

5.10.1 Přírodní vědy

5.10.1.1 Matematika

Třídění: **The Mathematics Subject Classification (MSC)**

Původce: American Mathematical Society (AMS)

Projekt: Mathematics on the Web

http://www.ams.org/mathweb/index.htm

Třídění MSC je užíváno pro kategorizaci dokumentů především pro databáze Mathematical Reviews (MR) a Zentralblatt MATH (Zbl). Současná verze třídění „2000 Mathematics Subject Classification (MSC2000)“ je revizí původní verze, které vznikla v roce 1991, kdy ji vytvořili editoři výše zmíněných databází.

MSC je rozvedeno do přibližně 5000 alfanumerických kódů, které jsou tvořeny dvou až pětimístnými kódy. Celkem má MSC 65 hlavních tříd, které začínají dvěma čísly od 00-xx po 97-xx. Čísla a písmena se v kódu střídají, takže konečný kód je složen ze čtyř čísel a jednoho písmene – například kvantová fyzika má kód 81V80 [AMS, 2000].

Konkrétní aplikací tohoto třídění pro dokumenty z prostředí internetu je stránka AMS „Mathematics on the Web“[30], kde jsou do základních tříd řazeny prosté odkazy na konkrétní dokumenty, řazené pod titulky jako jsou např. „Časopisy“ nebo „Preprinty“. Na rozdíl od humanitních oborů se velká část dokumentů a dalších materiálů nachází v FTP archívech a tak zde nejsou pouze odkazy na webové stránky. Materiál, který je v tomto katalogu pořádán, je velmi pestrý – od odkazů na stránky odborných časopisů přes preprinty a soubory bibliografií až po specializovaný software. Záznamy tvoří pouze odkaz a titulek – v některých případech je připojeno i jméno autora, případně organizace. Celý katalog tak odráží pragmatický přístup který ponechává pouze nejdůležitější informace v zájmu menší pracnosti a zrychlení komunikace.

30 http://www.ams.org/mathweb/index.html

5.10.1.2 Technické obory

Třídění: **Engineering Index Thesaurus (EI)**

Původce: Engineering Information (Velká Británie)

Projekt: Engineering E-Library, Sweden" (EELS)

http://eels.lub.lu.se/ae/

Projekt *Engineering E-Library, Sweden* (EELS) je oborovým rozcestníkem, který se zaměřuje na technické obory. Projekt vznikl v roce 1996 za spolupráce knihoven Royal Institute of Technology a univerzitní knihovny v Lundu. K dispozici jsou dvě služby:

elektronická knihovna **„Engineering E-Library, Sweden“** (EELS), která obsahuje více než 1460 vybraných dokumentů, které prošly procesem kvalitativního hodnocení a podrobného popisu včetně klasifikace kódy EI.

Služba **„All Engineering“**. Jedná se o programově generovaný seznam „všech“ stránek se zaměřením na techniku, který je možné prohlížet na základě názvu, země původu a počtu citací na dokument.

Pro digitální knihovnu byly dokumenty vybírány a indexovány ručně, pro službu „All Engineering“ probíhala indexace za pomoci specializovaného programu (harvesting software). Tak je v současnosti indexováno kolem 253 000 stránek a další přibližně 1,5 miliónu odkazů a citací na tyto stránky.

Pro klasifikaci stránek elektronické knihovny je použito standardní klasifikace technických oborů Engineering Index Thesaurus, který je používán v nejvýznamnější databázi pro technické obory – **COMPENDEX**.

5.10.1.3 Medicína

Třídění: **Medical Subject Headings (MeSH)**

Původce: National Library of Medicine (USA)

Projekt: původně OMNI (Organising Medical Networked Information), součást portálu BIOME, nyní **Intute: health & life services**

http://www.intute.ac.uk/healthandlifesciences/medicine/

Vznik hesláře MeSH je úzce svázán s rokem 1960, kdy National Library of Medicine (NLM) začala vytvářet bibliografickou databázi pro oblast medicíny Index Medicus, která byla později přejmenována na Medline.

Od počátku byl MeSH zamýšlen jako dynamický seznam hesel, pomocí kterých by bylo možné předmětově popsat materiál, který byl indexován v databázi. Nejedná se přímo o tezaurus, protože se striktně nedrží formálních pravidel, ale spíše o volný heslář. Ten umožňuje rychle reagovat na dynamické změny v medicíně (jen v edici 2003 bylo přidáno 1251 deskriptorů) [NLM, 2003] a přitom uvádí tyto deskriptory v kontextu nadřazených, podřazených a příbuzných termínů.

OMNI je předmětově orientovaným rozcestníkem kvalitních zdrojů v prostředí internetu se specializací na medicínu a biomedicínu, který vznikal podobně jako podobné projekty ve Velké Británii (např. EEVL pro techniku nebo SOSIG pro společenské vědy) v rámci rozvojového programu **eLib** rozvojové agentury JISC's (Joint Information Systems Committee). Rozcestník OMNI byl spuštěn na univerzitě v Nottinghamu. V roce 1998 vyhlásila agentura JISC program **Resource Discovery Network (RDN)**, který měl vytvořit z existujících rozcestníků koordinovanou síť pro vyšší vzdělávání ve Velké Británii [OMNI, 2004].

Na základě tohoto záměru začala univerzita v Nottinghamu spolupracovat s dalšími profesními organizacemi, především **The Royal College of Surgeons of England** a **The German National Library of Medicine**, na vzniku širšího oborového rozcestníku, který by oborově pokrýval oblast zdraví a přírodních věd. Takto se zrodil projekt BIOME – extenze stávajících služeb rozcestníku, která integruje již existující rozcestníky do jednotného vyhledávacího rozhraní. OMNI se tak stalo jednou ze součástí BIOME, která v současnosti zastřešuje pět tematických rozcestníků:

AgriFor – rozcestník zaměřený na zemědělství, potravinářství a lesnictví

OMNI – pro oblast zdraví a medicíny

NMAP – pro tematiku spojenou s profesemi ve zdravotnictví a ošetřovatelství

VetGate – pro veterinární disciplíny

BioResearch – pro výzkum v oblasti biologie a biomedicíny

Natural Selection – pro zemědělství, lesnictví a životní prostředí.

V jednotlivých rozcestnících jsou uvedeny odkazy především na dokumenty (zpravidla www stránky) odborných organizací s podrobnými anotacemi. Rozcestníky z oboru medicíny nejsou příliš rozsáhlé – k listopadu 2004 obsahoval OMNI 8522 záznamů, NMAP 3547 záznamů a BIORES 2601 záznamů [LIPSCOMB, 2000]. Přestože tyto počty nejsou velké, jsou oborové rozcestníky velmi oceňovány odbornou veřejností jako alternativní sekundární zdroje pro výzkum.

Na konci roku 2006 byly uvedené rozcestníky sloučeny pod novým názvem **INTUTE.** Toto sloučení umožnilo také několikanásobný nárůst počtu odkazů - ve čtyřech tematických skupinách (Science, Engineering and Technology, Arts and Humanities, Social Sciences, Health and Life Services) obsahuje portál INTUTE k březnu 2007 více než 100 000 ověřených webových zdrojů.

Třídění: Medical Subject Headings (MeSH)

Původce: National Library of Medicine (USA)

Projekt: The NLM Gateway

http://gateway.nlm.nih.gov/

Dalším projektem, kde se používá MeSH, je portál Národní lékařské knihovny USA (NLM) – NLM gateway. Jde o další projekt, který se snaží z jednoho místa zpřístupnit řadu informačních zdrojů NLM s jednotným vyhledávacím rozhraním. Systém automaticky rozešle dotaz do každé z dostupných databází a sjednotí výstup z každé z nich do jednoho seznamu. Nabízené bibliografické informace doplňuje služba *PubMed*, která zpřístupňuje některé články v plných textech [NLM, 2004].

Portál NLM gateway v současnosti nabízí 12 databází, všechny z produkce NLM. Nejznámější z nich je databáze zaměřená na široké spektrum oborů medicíny – MEDLINE. Další databáze jsou již úzce profilovány jako např. TOXLINE (bibliografické informace pro toxikologii) HSDB (informace o nebezpečných sloučeninách) nebo DIRLINE (adresář zdravotnických organizací).Heslář MeSH je hlavním třídícím systémem u většiny databází, u některých je doplněno i třídění americké chemické asociace pro chemické látky a sloučeniny – CAS registry numbers.

5.10.1.4 Zemědělství

Třídění: **AGRICOLA Subject Category Codes (SCC)**

Původce: National Agricultural Library (USA)

http://agricola.nal.usda.gov/

Agricola Subject Category Codes (SCC) je systémem předmětových kódů, užívaných National Agricultural Library (NAL), USA pro indexaci bibliografických záznamů pro databázi AGRICOLA. Toto třídění je ve formě čtyřmístných alfanumerických kódů, kde je první znak písmenem (A-X) a další tři znaky jsou číselným označením třídy.

Třídění má pouze jednu úroveň, rozdělenou do 214 tříd, které jsou blíže upřesněny slovním vyjádřením např. E550 Rural developement. Některé ze tříd sdílejí stejné téma, které je blíže upřesněno závorkou např. F120 Plant production (field crops) a F130 Plant production (range and pasture crops), což je zde způsob, jak nahradit hierarchické vztahy mezi třídami.

Agricola Subject Category Codes (SCC) je využíván kromě databáze AGRICOLA také na portálu NAL nazvaném Agriculture Network Information Center (AGNIC).

Třídění: AGROVOC

Původce: Food and agriculture organization of the United Nations (FAO)

http://www.fao.org/agrovoc

Vícejazyčný tezaurus AGROVOC byl vytvořen ve spolupráci FAO a Komisí Evropského společenství v 80. letech jako hlavní indexační a třídící systém pro databáze, které FAO produkuje – AGRIS a CARIS. Cílem vytvoření tohoto tezauru bylo pokrytí terminologie v tematických oblastech, které jsou v databázích sledovány, a zároveň umožňovat práci ve více jazycích.

Poslední – třetí verze tohoto tezauru je z roku 1997 a obsahuje 16 105 základních popisných termínů v angličtině a 9 480 anglických, 8 693 francouzských a 12 086 španělských synonym[AGROVOC, 2004]. Kromě těchto zmíněných jazyků jsou termíny dostupné také v čínštině, arabštině, portugalštině a také v češtině.

AGROVOC je tezaurem, a tak jsou i jeho vazby komplexnější než u systému izolovaných kódů. Každý termín má s dalšími termíny vzájemné vztahy, které mohou být definovány jako vztah nadřízený/podřízený, vztah ekvivalence a příbuznost termínu. Použití tohoto tezauru není omezeno pouze na databáze FAO, ale užívá se jej i jako doplňkové třídící soustavy v dalších oborových projektech.

Třídění: **CAB Thesaurus**

Původce:CABI Publishing

Projekt: INTUTE

http://www.intute.ac.uk/

CAB tezaurus je úzce spojen s databází z oblasti zemědělství a příbuzných disciplín CAB. Záběr této databáze, zaměřené na aplikované přírodní vědy, je velmi široký – od veterinární vědy přes entomologii, parazitologii, rozvoj venkova a agroturistiku.

Z tohoto důvodu její vydavatel – společnosti CABI Publishing vytvořil zvláštní třídění, které v současnosti obsahuje přibližně 59 000 popisných termínů, a je tak největším tezaurem pro oblast zemědělství a příbuzných disciplín [NLM, 2004].

Tezaurus CAB kromě databází CAB používají pro třídění záznamů elektronických zdrojů mimo jiné také dvě databáze v projektu **Intute - VetGate** (veterinární disciplíny) a **AgriFor** (zemědělství, potravinářství a lesnictví).

Třídění: Aquatic Sciences and Fisheries Thesaurus (ASFA Thesaurus)

Původce: Food and agriculture organization of the United Nations (FAO)

Projekt: http://www4.fao.org/asfa/asfa.htm

Tezaurus ASFA vytváří oddělení Fisheries and Aquaculture organizace FAO a je zaměřena na problematiku mořské biologie, rybářství a příbuzných oborů. Stejně jako drtivá většina databází FAO věnuje speciální pozornost aktuálním tématům rozvojových zemí. Databáze je budována od roku 1971 a k březnu 2007 obsahuje přibližně 1,144 milionu záznamů.

Třídění: National Agricultural Library Thesaurus (NALT)

Původce: National Agricultural Library, US department of Agriculture (NAL)

Projekt: http://agclass.nal.usda.gov/agt/agt.shtml

Tezaurus NALT byl poprvé zveřejněn v roce 2002 jako pomůcka pro rešeršní službu United States Departement of Agriculture (USDA).

Tezaurus je primárně určen pro indexování informací zemědělského charakteru a návazné vylepšování procesu vyhledávání těchto informací. V současnosti je užíván především jako indexační slovník databáze AGRICOLA a interních informačních systémů USDA.

Tezaurus tematicky zahrnuje problematiku biologických, přírodních a sociálních věd včetně jejich terminologie a je řazen do 17 předmětových kategorií.

5.10.1.5 Ekonomika

Třídění: **North American Industry Classification System (NAICS)**

Původce: US Economic Classification Policy Committee, Statistics Canada,

Mexico's Instituto Nacional de Estadistica, Geografia e Informatica.

Projekt: U.S.Census Bureau

http://www.census.gov/epcd/www/naics.html

Dlouhá léta se pro klasifikaci ekonomických odvětví, činností a produktů používala klasifikace Standard Industry Classification (SIC). Tato klasifikace je od roku 1997 nahrazována novým tříděním NAICS, které původně vzniklo jako iniciativa států Severní Ameriky (Spojené státy americké, Kanada, Mexiko).

Jedná se o soustavu numerických kódů, které se hierarchicky větví na maximálně pět úrovní. Každý kód tak může být nejvýše šestimístný (první hierarchie je označena kódem o dvou číslicích). Kódy NAICS se již staly standardním způsobem třídění ekonomické produkce a jsou hojně užívány v různých ekonomických databázích i oficiálních vládních dokumentech.

Prohlížení a hledání kódů je přístupné na stránkách statistického úřadu Spojených států [31.]

Třídění: **Journal of Economic Literature Classification system (JEL)**

Původce: American Economic Association (Journal of Economic Literature)

Projekt: AEAweb

http://www.aeaweb.org

Třídění Americké ekonomické asociace (AEA) podobně jako ve výše zmíněných případech souvisí se vznikem vlastní bibliografické databáze. Databáze AEA – EconLit vznikla v roce 1969 a je považovaná za hlavní informační zdroj pro ekonomii a příbuzné disciplíny.

Třídění JEL je v první úrovni rozděleno na devatenáct hlavních kategorií, které jsou označeny velkými písmeny abecedy (A-Z). Úroveň druhá blíže definuje podtřídy – např. „B1 –History of Economic Thought through 1925“ a úroveň třetí je seznamem nejpodrobnějších kategorií. Podtřídy druhé úrovně obsahují poměrně malý počet kategorií;jejich počet se liší podle tématu od jedné do devíti kategorií. Konečný kód je taktřímístný –např. „H21 Efficiency; Optimal taxation“ [Journal of Economic Literature, 2004] .

Třídění: World Bank Thesaurus

Původce: The World Bank

http://www.multites.com/wb/

31 http://www.census.gov/epcd/www/naics.html

Světová banka používá tento tezaurus jako pomocné třídění k hlavnímu třídícimu systému - Library of Congress Subject Headings. Tezaurus pro světovou banku vytvořila společnosti Multisystems, která jej také provozuje na vlastních webových stránkách.

Tezaurus je členěn do 357 hlavních hesel. Každé heslo je spojeno s numerickým kódem, který blíže definuje tematické zařazení hesla – např. 28.05.00 Hydro Power Vocabulary.

Třídění: National Statistics Socio-economic Classification (NS-SEC)

Původce: Office for National Statistics, Velká Británie

http://www.statistics.gov.uk/about/search.asp

Národní statistický úřad ve Velké Británii vyvinul vlastní třídění pro oficiální průzkumy a statistiky Toto třídění je zaměřeno na klasifikaci povolání dospělé populace a oficiálně je používáno od roku 2001.

Celý systém je založen na alfanumerických kódech, které jsou členěny na osm analytických skupin. Tyto skupiny jsou pouze pomůckou pro rozdělení skupin pro pozdější statistické vyhodnocení; samotné klasifikační schéma je tvořeno dvouúrovňovou hierarchií. První dva znaky kódu – např. L4 označují hlavní třídu, číslo oddělené tečkou pak blíže určuje kategorii.

Příklad:

L4 Lower professional and higher technical occupations

L4.2 'New' employees

L4.3 'Traditional' self-employed lower professionals and higher technical

Zdroj: [National statistics, 2002]

Tento klasifikační systém je klasickým zástupcem oborového třídění, obsahuje pouze několik desítek tříd, zato natolik specifických, že postačují pro zamýšlený účel.

5.10.2 Zhodnocení oborových třídění

Z výše uvedených příkladů jsou zřejmé výhody oborových klasifikací, mezi které můžeme zahrnout:

- pragmatičnost systému

Většina oborových systémů minimalizuje počet hierarchických úrovní (existují-li vůbec) a omezuje systém na alfabetický/numerický systém tříd a kódů.

- minimální syntaxe (existuje-li vůbec)

Notace, které jsou tvořeny nejčastěji numerickými nebo alfanumerickými kódy, slouží k popisu samy o sobě; další kombinační možnosti nejsou zapotřebí.

- snadná údržba

Vzhledem k rozsahu systémů je velmi jednoduché a rychlé upravit třídění podle aktuálního vývoje poznání. Oborová komunita je navíc natolik adaptabilní, že je schopna dohodnout se na změnách velmi rychle.

Mezi slabiny oborových třídění naopak můžeme uvést:

- definice kategorií – chybí specifické kategorie

Je zde patrné pravidlo o obecnosti a specifičnosti – každý obor si hledá hranici, kdy třída již nebude svým popisem dostačující a bude nutné vytvořit třídu novou. Všechna třídění se spíše uchylují k definování obecnějších kategorií, sloužících pouze jako rámcové třídění, které je doplněno vyhledáváním klíčových slov.

- chybí komplexní náhled na problematiku

Oborová třídění jdou vždy pouze do hloubky (tj. podrobně rozebírají problematiku jednoho oboru). Na druhou stranu již nemají širší záběr, který by zařadil témata do kontextu jiných oborů či náhledů. Za cenu vysoké funkčnosti tak oborová třídění poskytují minimální či žádné doplňkové informace, na rozdíl od systematických univerzálních třídění. Předpokládá se, že uživatelé pochází z komunity, která se v problematice orientuje a proto by jakákoli další kontextová informace byla zbytečná.

- zachycení skutečných vztahů mezi koncepty

Zpravidla není v oborovém třídění možné zachytit vztah mezi více skupinami, kódy. Dokument může patřit do jedné či více skupin, ale není možné přesně definovat různé aspekty (typ dokumentu, země, jazyk...). Důvodem je i velmi stručná notace a syntax, která postrádá jakékoli kombinační možnosti.

- interoperabilita třídění

Otázka, zda se skutečně jedná o slabiny těchto systémů, je velmi diskutabilní. Již z povahy těchto třídění vyplývá jejich úzké zaměření, a nelze proto očekávat rozsah tříd srovnatelný s univerzálním tříděním. Pokud jsou oborová třídění považována uživatelskou komunitou za standard, jakákoli převoditelnost již není nutná; uživatelů mimo oborovou komunitu není tolik, aby bylo potřeba tyto úpravy podnikat.

Oborová třídění jsou příkladem, kdy pořádací systém vzniká na základě konkrétních informačních potřeb. Teorie se plně podřizuje praxi, což je také důvodem bezproblémového přejímání těchto systémů oborovými komunitami. Minimální vizualizace (hierarchie apod.) odstraňuje i zdroj potenciálních konfliktů (pořadí témat, filozofický náhled) a striktně podřizuje třídění jedinému cíli – nalézt hledanou informaci.

Tato třídění se vyvíjejí společně s oborem, a pokud nenastane výrazná změna ve vývoji poznání nebo ve stavu komunity, budou nadále sloužit svému poslání. Síla oborových třídění, která spočívá ve vysoké adaptabilitě a jednoduchosti, pravděpodobně převáží ve srovnání s komplexními systémy a můžeme tak jen očekávat další rozvoj oborových třídění na úkor komplexních klasifikací.

6 WEBOVÉ KATALOGY

> Tato "internetová škola" vidí klasifikaci jako jednoduchou hierarchickou interpretaci současných potřeb uživatele spíše než akademický pohled na poznání. Přispívá tak ke změně chápání reprezentace znalostí v prostředí elektronických zdrojů, která bere v potaz větší dynamiku a menší životnost informací.
>
> Allan Wheatley [2002]

6.1 PŘEDMĚTOVÉ STROMY JAKO FORMA KLASIFIKACE

Webové katalogy, označované také jako katalogy stránek nebo předmětové stromy (angl. subject-trees), jsou seznamy klasifikovaných odkazů na stránky uspořádané podle tematiky v hierarchicky děleném stromě. Tento strom má hierarchickou strukturu, ale jeho vlastnosti a pravidla pro pořádání předmětů se liší od přísně definovaných pravidel pro

hierarchické klasifikace. V předmětovém stromě se dělí koncepty do různých větví – podtříd, ale na rozdíl od hierarchií nemají některé vlastnosti – například dědičnost třídy. Ve stromě tak entity mají systematické ale už ne generické (rodové) vztahy.
Barbara Kwasnik k předmětovým stromům jako typu klasifikace uvádí tyto charakteristiky [Kwasnik, 1999]:

1. Kompletní a vyčerpávající informace

Každý předmětový strom je rozdělen hierarchicky od nejvyšší úrovně (univerzum, obor) po nejnižší hierarchii, kde jsou popsány jednotlivé dokumenty. Nejdůležitější prvky jsou ve vyšší hierarchii, méně důležité prvky v nižší. Předmětový strom tak roste především přidáváním dalších prvků na nejnižších úrovních tohoto stromu.

2. Systematická a jasná pravidla pro rozlišení předmětů

Struktura navigačního stromu je dána vztahy mezi jeho prvky. Ty mohou být určeny vztahem nadřízený – podřízený prvek (celek a jeho část) anebo odkazem na příbuzné kategorie.

3. Vzdálenost předmětů je také vyjádřením jejich příbuznosti

Předmětový strom určuje také fyzickou vzdálenost mezi předměty a tak určuje míru příbuznosti mezi nimi. Pokud existuje výčet odkazů (entit), ty mohou být řazeny podle významu. Například ve skupině Univerzita Karlova mohou být prvky řazeny tak, aby byl prvním odkazem „Rektorát“ a dalšími odkazy „Fakulta přírodovědecká“ „Fakulta filozofická“ atd. To znamená, že tímto způsobem je zobrazen vztah mezi částmi univerzity – od centrálního orgánu (rektorát) po jeho podřízené části (fakulty).

4. Relativní zastoupení předmětů a témat

Témata jsou v předmětových stromech zastoupena podle množství existujících stránek s daným tématem, tj. neexistuje zde poměrné zastoupení témat a není nutné dělit skupinu na stejný počet podskupin jako je tomu například u univerzálních hierarchických třídění MDT nebo DDC. Podskupiny se vytvářejí dynamicky – jakmile je skupina přeplněna odkazy, začne se dělit na další tematické podskupiny. Stejně tak je možné podskupiny slučovat, je-li jejich počet velmi nízký.

5. Polyhierarchie

Jakýkoli prvek je možné zařadit do více větví předmětového stromu – tj. neexistuje zde exkluzivita tříd tak jako u hierarchií.

Předmětové stromy jsou hierarchickým systémům pořádání velmi příbuzné i v jejich nedostatcích, které můžeme uvést v následujících bodech [Kwasnik, 1999]:

1. Pevná struktura stromu

Předmětový strom je budován od první – tj. nejvyšší úrovně směrem k úrovním nižším. I když je možné přidávat další témata a odkazy do existujících skupin, jeho struktura je pevná a nelze ji dynamicky měnit. Pokud by existovalo téma, které by nešlo do existující struktury zařadit, musela by být kvůli jeho zařazení změněna celá struktura stromu. Tento nedostatek je u katalogů na internetu částečně řešen polyhierarchií ve formě odkazů – téma může být zařazeno v jiné skupině a v jiné skupině je na něj odkazováno.

2. Jednostranný tok informací

V hierarchii je tok informací ve dvou směrech – ve vertikálním mezi hlavními třídami, nadřazenými a podřazenými, a mezi vedlejšími větvemi, které sdílí stejnou nadřazenou třídu (lateral – sibling classes). Předmětový strom je založen na vertikálním pohybu mezi úrovněmi; vedlejší větve klasifikace od sebe nejsou rozlišeny a mohlo by tak dojít k záměně kontextové informace. Například při hledání geografického názvu Ithaca je možné najít město v USA (stát New York) nebo ostrov v Řecku. Pokud tyto dva pojmy nejsou zařazeny do kontextu (stát New York, Řecko) může dojít k jejich záměně [Kwasnik, 1999].

3. Zvolená perspektiva vztahů

Předmětový strom nemůže vždy přesně zdůraznit zajímané aspekty vztahů mezi různými entitami. Většina vztahů je charakterizována aspektem příslušnosti k tematické skupině (např. chirurgie je disciplínou medicíny). Již není možné uvést všechny její vztahy k dalším přírodním vědám nebo její tematické přesahy a použití v jiných oborech (bio-mechanika, veterinární medicína apod.). Každé téma tak může mít zajímavé asociace, které nemusí být v předmětovém stromu viditelné.

6.2 SPECIFIKA WEBOVÝCH KATALOGŮ

Použití předmětových stromů pro webové katalogy je velmi specifické, protože nedostatky této koncepce pořádání informací lze do značné míry eliminovat využitím vlastností hypertextu – především relativním odkazováním.

Polyhierarchie je u katalogů na internetu realizována především ve formě vzájemných odkazů; téma nebo celá skupina má své pevné místo ve větvi předmětového stromu a z jiných kategorií, kde by mohl uživatel toto téma hledat, je odkazován do této sekce. Tato vlastnost je velmi zajímavá pro uživatele, neboť struktura stromu je tak relativní a existuje více cest, jak hledanou kategorii ve stromu najít. Katalogy označují tyto odkazy různými způsoby – například znakem „@“.

Stejným způsobem může být řešena i perspektiva vztahů. U sekcí katalogu jsou kromě podřízených skupin často také odkazy na příbuzná témata, která se k této sekci vztahují, ale jejich vztah nemůže být vyjádřen prostou hierarchií podřízené-nadřízené téma. Katalog tak uživatelům nabízí to, co si přejí – volnost pohybu a více navigačních cest, kterými mohou nalézt žádanou informaci.

Webové katalogy se také pokouší odstranit nedostatek kontextových informací v kategoriích tzv. *navigační lištou* (navigation bar), která uživatelům ukazuje adresářovou cestu od nejvyšší úrovně až po kategorii, kterou si v současnosti prohlíží. Informace, zobrazené v této navigační liště by pro příklad, uvedený v předchozí části, vypadaly takto:

Directory > Regional > U.S. States > New York > Cities > Ithaca

Directory > Regional > Countries > Greece > Prefectures > Kefalonia (Kefallinia) > Islands > Itháki (Ithaca)

Zdroj: Katalog Yahoo!

Tyto katalogy umožňují volnosti pohybu v navigaci, a mají také další jedinečné vlastnosti, které je činí velmi populárními pro uživatele:

Jsou zpracované a popsané lidmi

Lidé hledají odkazy a tvoří jejich anotace. To znamená, že vystihnou skutečný obsah dokumentů daleko lépe než při strojovém zpracování (přístup vyhledávačů).

Kontrola kvality odkazů

Vedlejším produktem lidského zpracování je kontrola kvality odkazů. Do katalogu jsou řazeny pouze takové odkazy, které odpovídají kategorii. Je zde malá pravděpodobnost, že by do katalogu byl zařazen odkaz, který se netýká této tematiky, což je u vyhledávačů běžné.

1. Navigace je intuitivní

Struktura katalogu je přehledná a pomocí hierarchických odkazů se každý dostane k tématu, které ho zajímá.

Informace jsou uváděny v dalších souvislostech
Informace v katalogu není izolovaná, je zařazena do kontextu tematických kategorií a tak může poskytnout zajímavý náhled na celou problematiku. Katalogy jsou vhodné pro vyhledávání obecných témat a pro uživatele, kteří se potřebují zorientovat v problematice.

Pro předmětový strom je nejdůležitější definování jeho první úrovně. Jak píše Barbara Kwasnik [Kwasnik 1999], první úroveň ovlivňuje vypovídací schopnost celého stromu – tj. dělení na skupiny a podskupiny a vztahy jejich prvků navzájem. Počet skupin v první úrovni tak nesmí být malý, neboť pak je jejich dělení příliš obecné, ani velký, aby třídění zůstalo přehledné.

U vybraných významných webových katalogů je počet kategorií mezi devíti až patnácti kategoriemi, jak to ukazuje přehledová tabulka 6.

Projekt	*Open Directory*	*Yahoo!*	*LookSmart*	*Infomine*	*Lii.org*
Adresa	*www.dmoz.org*	*www.yahoo.com*	*www.looksmart.com*	*infomine.ucr.edu*	*www.lii..org*
Počet kategorií 1. úrovně	15	14	12	9	15

Tabulka 5: Počty kategorií první úrovně u vybraných webových katalogů

6.3 ROZDĚLENÍ KATALOGŮ PODLE TYPŮ

Univerzální katalogy mají obecný záběr a slouží jako základní orientace mezi dostupnými zdroji na internetu. Tyto katalogy jsou obsahově nejrozsáhlejší, ale tematická šíře je na druhé straně kompenzována menší podrobností jednotlivých témat. Univerzální katalogy mohou odkazovat na informace v jakémkoli jazyce nebo mohou být zaměřené pouze na jeden jazyk nebo zemi. Těchto národních katalogů je obrovské množství, které se odhaduje v řádech tisíců.

Specializované katalogy se zaměřují pouze na určitou tematickou oblast, která je rozdělena podle charakteru tříděných témat. Charakteristické jsou například webové katalogy zaměřené na software nebo na sortiment internetového obchodu.

Webové katalogy můžeme rozdělit také podle role, kterou v rámci služeb nabízených na stránkách katalog hraje. Katalogy existují ve dvou základních formách – katalogy samostatné a katalogy v rámci portálů.

Prvním typem jsou **samostatné webové katalogy**. Zpravidla jde o akademické projekty, které jsou financovány z různých grantů a jejich vývoj je svázán s komunitou uživatelů, případně s podporou větších firem (např. Open Directory Project). Tyto projekty mají velmi nejistou budoucnost – podpora z grantů je časově omezená a pak je otázkou, zda se podaří pro další provoz zajistit finanční zdroje. Jak uvádí [JANES, 2003]: *„nikdo nemá odpovědnost za zakládání veřejné knihovny pro internet. Jiné knihovny mají komunity nebo instituce, kterým slouží; na oplátku za služby které tím poskytují jsou také finančně zajištěny“*. Toto je přirozeným důsledkem bezplatně poskytovaných služeb – buď je možné zajistit financování z jiných zdrojů (např. lii.org) nebo udržovat projekt za podpory komunity dobrovolníků (Open Directory, Vlib).

Provozovat webový katalog je finančně velmi náročné. Z tohoto důvodu je většina katalogů **dostupná jen jako jedna z nabízených služeb** v rámci tzv. *portálů*. **Portálem** označujeme webovou stránku, která nabízí uživatelům řadu služeb; nejedná se o úzce profilovanou službu, ale o komplex návazných služeb, které mají uživateli nabídnout co nejvíce možností. Pokud se takovému portálu daří uživatele na svých stránkách udržet co nejdéle, úměrně tomu stoupá i jejich atraktivita pro reklamu a také z nich plynoucí zisk. Tak je možné nabízet tuto službu uživatelům bez toho, aby pro ně byla zpoplatněna. Tímto způsobem je provozován druhý největší webový katalog – Yahoo!.

Negativem tohoto přístupu je podřízení nabízeného obsahu reklamě, která je často obtížně odlišitelná od skutečného obsahu (reklamní odkazy apod.). Přehled různých typů webových katalogů je uveden například na adrese *Internet Search Engine Database*[32.]

6.4 POROVNÁNÍ WEBOVÝCH KATALOGŮ A UNIVERZÁLNÍCH KLASIFIKAČNÍCH SYSTÉMŮ

Webové katalogy jsou klasickým zástupcem pragmatického třídění, které Allan Wheatley [Wheatley, 2000, s. 120] popisuje jako: „Tato „internetová“ škola vidí klasifikaci spíše jako jednoduchou hierarchickou interpretaci současných uživatelských potřeb než jako akademický pohled na stávající poznání a těží z rychlého přizpůsobování na změny elektronických zdrojů“.

Vytváření webového katalogu je výrazně levnější než tvorba univerzální klasifikace. Podle studie [Wheatley, 2000, s. 126] jsou k tomu tyto důvody:

1. Slovník termínů, které jsou užívány k popisu termínů je jednodušší a stručnější. Zároveň je snáze přizpůsobitelný pro popis zvláštních případů.
2. Klasifikace webových katalogů je jednodušší, protože není nutné najít jedinečné umístění ve struktuře klasifikace. Klasické univerzální třídění se oproti tomu snaží sloužit i jako orientace pro fyzické nalezení knihy v polici knihovny a tak je jedinečné umístění nezbytné.
3. Neexistuje zde žádný index termínů či notace pro třídy.
4. Zařazení nových položek je jednoduché, protože se jedná o přidání dalších položek do abecedně řazeného seznamu.
5. Očekávaná životnost elektronických dokumentů je mnohem menší než u dokumentů tištěných. Uživatel klade důraz především na nové materiály a tak by vyčerpávající nebo příliš podrobná klasifikační osnova mohla být na překážku rychlé aktualizace dokumentů.

Cílem těchto katalogů není zmapování všech existujících dokumentů podle jejich obsahu, ale pouze dokumentů, které jsou v prostředí internetu a byly nalezeny editorskými týmy (případně za pomoci automatizované klasifikace) nebo zaregistrovány jejich tvůrci. Tomu odpovídá i struktura jejich kategorií. V tabulce č.7 jsou srovnány kategorie první úrovně u dvou nejvýznamnějších webových katalogů (*Yahoo!* a *Open Directory*) s klasifikačními schématy Deweyho desetinného třídění (DDC) a Mezinárodního desetinného třídění (MDT).

32 http://www.isedb.com/

Yahoo!	Open Directory	DDC	MDT
Business & Economy	Arts	Generalities	Generalities
Computers & Internet	Business	Philosophy & psychology	Philosophy, psychology
News & Media	Computers	Religion	Religion, theology
Entertainment	Games	Social sciences	Social sciences
Recreation & Sports	Health	Language	Natural sciences
Health	Home	Natural sciences & mathematics	Technology
Government	Kids and Teens	Technology (Applied sciences)	The arts
Regional	News	The arts	Language, linguistics, literature
Society & Culture	Recreation	Literature & rhetoric	Geography, biography, history
Education	Reference	Geography & history	
Arts & Humanities	Regional		
Science	Science		
Social Science	Shopping		
Reference	Society		
	Sports		

Tabulka 6: Porovnání první úrovně webových katalogů a knihovnických třídění

Při srovnání kompletní struktury zjistíme, že pokrytí témat mezi katalogy a tradičním tříděním je podobné. Jak se ukazuje i ve studii [SAEED, 2001], kde autoři porovnávali webový katalog Yahoo! a Deweyho desetinné třídění (DDC), webový katalog může velmi dobře nahradit kategorie klasického univerzálního třídění. Hlavní rozdíly jsou ve strukturování témat do úrovní. Webové katalogy do první úrovně řadí kategorie, které obsahují uživatelsky nejzajímavější témata. Podobná situace je i u vyhledávačů, kde je většina dotazů z oblasti zábavy a volného času [SULLIVAN, 2002B].

Vzhledem k výše uvedeným vlastnostem a také díky menším nákladům vytlačují webové katalogy v prostředí internetu aplikace univerzálních klasifikačních systémů. Přesto mají mnoho problémů, které v budoucnu mohou ovlivnit jejich další rozšíření.

Michelle Hudon uvádí mezi problémy **nedostatek standardizace, chybí jednota v definici a řazení kategorií, položky často nejsou dostatečně specifické a jsou zde i problémy s definicí vztahů mezi strukturami**. *„Jinými slovy, základní principy, které ovlivňují podobu tradičních klasifikačních schémat se v tříděních zdrojů na internetu zřídka vyskytují“* [GODBY, 2000]. Jak autorka dále dodává, je velmi obtížné předvídat, jak se tyto systémy budou vyvíjet v budoucnu, kdy množství jejich odkazů několikanásobně vzroste.

Webové katalogy jsou příkladem intuitivní navigace, která se snaží nezatěžovat uživatele s jakoukoli speciální přípravou. To se ukazuje jako velká přednost při porovnání s nejvýznamnějším způsobem vyhledávání informací na internetu – vyhledávači. Je pravděpodobné, že v budoucnosti budou vznikat především projekty kombinující vyhledávač s webovým katalogem. Touto kombinací lze některé nedostatky odstranit či alespoň eliminovat na menší míru.

6.5 NEJVÝZNAMNĚJŠÍ UNIVERZÁLNÍ KATALOGY

6.5.1 Open Directory Project (ODP)

http://www.dmoz.org

Projekt *Open Directory*, známý též jako DMOZ (*Directory Mozilla*) je největším lidmi tvořeným katalogem v prostředí internetu. K listopadu 2008 obsahuje více než 4,5 miliónů stránek ve více než 590 000 kategoriích. Rozsahu tohoto projektu bylo dosaženo zapojením komunity editorů, kteří spravují části a sekce tohoto katalogu; těchto editorů je v současnosti kolem 75 000. Myšlenka Open Directory Project je založena na hnutí *Open Source*, které se snaží vytvářet především software, který je volně k dispozici všem zájemcům. Licence těchto produktů (tedy i Open Directory) předpokládá, že každé zlepšení a úprava tohoto softwaru bude opět zdarma k dispozici komunitě uživatelů. Tímto způsobem je také katalog využíván různými vyhledávači, které ho na svých portálech nabízí jako doplňkový způsob navigace. Open Directory tak nalezneme například u vyhledávačů Google, Altavista nebo Lycos [NETSCAPE, 2002A].

Projekt tohoto typu a rozsahu se neobejde bez náročné administrace. Ta je zajišťována společně s hostováním projektu společností Netscape Communication Corporation, kde malá skupina pracovníků dohlíží na ediční politiku, řízení projektu a zajišťuje technickou podporu projektu [NETSCAPE, 2004].

Celý projekt je v angličtině, ale jeho atraktivita je v možnosti úprav do národních jazyků. Jestliže se tedy najdou editoři, je možné přeložit nebo vytvořit nové kategorie v národním jazyce. Tyto části nemusí být stejné ve všech dostupných jazycích; nejvíce odkazů je v angličtině, v méně rozšířených jazycích jako je afrikánština, kurdština nebo jazyk fársí může být k dispozici pouze několik desítek odkazů. Kromě odkazů a anotací v národním jazyce je možné také zobrazit kompletní rozhraní katalogu v některém z nabízených jazyků (také v češtině). Pro přístup ke katalogu v národním jazyce je vedle nabídky kategorií první úrovně také odkaz „Word“, který vede ke všem kategoriím které jsou v národních jazycích. Zároveň je tento obsah v cizích jazycích odkazován u každé kategorie v angličtině (pokud existuje).

Myšlenka Open Source se projevuje i v ediční politice. Jsou registrovány pouze stránky, které odpovídají kvalitou a volným přístupem. Stránka nemá být „zrcadlem“ již existující stránky, má mít kvalitní obsah, a musí být kompletní. Naopak, z katalogu jsou vyřazovány reklamní prezentace, obchodní stránky, stránky, které obsahují pouze odkazy a stránky různých obchodních systémů nebo katalogů (např. Multi-level marketing). Stejně tak jsou zakázány stránky s ilegálním obsahem nebo stránky sestávající z výsledků vyhledávačů [NETSCAPE, 2002B]. Open Directory také důsledně používá standard metadat PICS (viz kapitola 4.10.1, s. 32) pro odlišení kategorií se stránkami, které nejsou vhodné pro děti.

Systém přidávání nových odkazů je založen na nezávislé správě kategorií. Většina kategorií má svého dobrovolného editora, který ve svém volném čase edituje stávající odkazy a přidává nové. Veškerá práce se řídí závaznými pravidly, kterými je možné udržovat obsah katalogu kvalitní. Editoři nemusí všechny stránky vyhledávat a přidávat sami, ale jejich práce je spíše v hodnocení návrhů nových stránek, které jim poslali uživatelé katalogu. Pravidla pro editory jsou poměrně podrobná a definují oblasti jako je případný konflikt zájmů editora (každý vyplňuje tyto údaje do databáze), komunikace s hlavními editory nebo úpravy sekcí.

Samostatným tématem jsou pravidla pro tvorbu taxonomií (tj. hierarchické struktury kategorie). Obecná pravidla stanovují, že kategorie může být rozdělena pokud počet odkazů, které obsahuje, překročí počet 20 [NETSCAPE, 1998]. Toto pravidlo je spíše doporučením, protože často může být výhodnější ponechat větší kategorii s více odkazy. Nejčastěji se tvoří tematické podkategorie (blíže definují obecné téma), ale vytváří se také podkategorie jazykové a regionální (například Česká republika – Praha). Jako poslední varianta třídění je možné uspořádat odkazy abecedně podle nabídkové lišty. Tento způsob je ale doporučován jako poslední řešení jak uspořádat rozsáhlý soubor odkazů. Pro pojmenování kategorií a podkategorií je určen heslář *„Preferred Terms“*[33], který obsahuje některé vlastnosti tezaurů. Tento heslář je určen pro sjednocení terminologie ale také pro zachování konzistence myšlenkového konceptu při organizaci všech kategorií tohoto katalogu. V hesláři je dostupných 27 tematických okruhů, které určují preferovaný termín, jeho definici a seznam synonym, pro které má být použit. Dodržování termínů z tohoto hesláře sice není povinné, nicméně je z výše uvedených důvodů (konzistence kategorií) doporučováno.

Příklad tematického okruhu *„Directories“*.

> **Directories**
> Scope: Use for sites containing alphabetical or classified lists of resources covering a particular subject area.

33 http://dmoz.org/preferredterms.htm

Directories is used in place of Guides for sites which provide a straight list of sites, sometimes with a brief description and not, generally, including additional information. See also Resources and Search Engines.

Use For

Christian Portals
Database of Falun Dafa Websites
Directories and Links Collections
Directories and Lists

(část vynechána)

Jak se ukazuje, proces tvorby katalogu Open Directory zahrnuje metody pořádání informací, které v kombinaci s vlastnostmi hypertextu posunují tradiční formu taxonomické klasifikace výrazně vpřed. Zároveň tento projekt potvrzuje, že dobrovolná komunita je schopna společným úsilím vytvořit katalog, který je rozsáhlý a zároveň dostatečně specifický aby jej mohli využívat i odborníci. Neznamená to, že by webové katalogy mohly nahradit oborová třídění, ale spíše to poukazuje na výhody tohoto způsobu pořádání informací ve srovnání s implementací projektů tradičních knihovnických třídění do prostředí internetu (viz kapitola 5, s. 48).

6.5.2 Yahoo!

http://www.yahoo.com/

Katalog *Yahoo!* byl založen v roce 1994 dvěma postgraduálními studenty Stanford University - Davidem Filo a Jerry Yangem jako zájmový projekt, kterému věnovali svůj volný čas. Nyní se jedná o jeden z největších portálů, jehož služby využívá několik set miliónů uživatelů (jen registrovaných uživatelů je 237 miliónů) ve 25 zemích a ve 13 jazykových mutacích [YAHOO, 2003A]. Hlavní předností portálu *Yahoo!* ve srovnání s jeho konkurenty je nabídka komplexního řešení hledání informací. Není zde k dispozici pouze katalog webových stránek, ale současně i vyhledávač a další služby orientované odborně (finanční zpravodajství, mapy, adresáře firem) i pro volný čas (e-mail zdarma, zpravodajství, zábava). Společnost *Yahoo!* se stala také jedním z vůdců trhu vyhledávačů poté co se sloučila s firmou *Inktomi*, která provozuje například vyhledávače *Alta Vista* nebo *AlltheWeb* (viz kapitola 9, s. 117).

V první úrovni má tento katalog čtrnáct hlavních skupin. Podle studie [WHEATLEY, 2002] má katalog až devět úrovní; většina odkazů je umístěna do páté až osmé úrovně. Počet odkazů, který Yahoo! obsahuje není přesně znám. Při absenci oficiálních údajů tak musíme vystačit s odhady. Podle uvedených aktualizací[34] je do katalogu denně přidáváno průměrně 200-250 nových záznamů[35] a tak může být přibližný roční nárůst až 90 000 záznamů. Greg Notess [NOTESS, 2003B] odhaduje počet odkazů tohoto katalogu k roku 2003 na tři milióny. Při přepočtu podle výše uvedených údajů tak velikost katalogu k 1.3.2007 může být přibližně 3,3 milionu. Aktuálnější údaje o velikosti z nezávislých zdrojů bohužel zatím nejsou k dispozici

Integrovaný vyhledávač může hledat v internetu nebo pouze v katalogu, což je výhodné pro rychlé nalezení kategorií, případně samotných odkazů. Původně celý portál používal licencovaný vyhledávač společnosti Google. Od roku 2004 používá Yahoo! pro vyhledávání vlastní technologii. Tento krok souvisí patrně s tím, že se Yahoo! v roce 2003 spojilo s významným dodavatelem vyhledávacích technologií společností Inktomi, která provozovala několik významných vyhledávačů [YAHOO, 2003B].

Další významnou službou je *MyYahoo*, která umožňuje osobní nastavení některých služeb tohoto portálu. Podle [DELANEY, 2004] se firma na oblast personalizace soustřeďuje a v budoucnu by se mohly objevit služby jako nastavení

34 http://dir.yahoo.com/new/

35 Počítáno průměrem z nových záznamů mezi 1 – 14.12.2004.

preferencí hledání, kdy by se na nejvyšších místech zobrazily výsledky, které se vztahují k bydlišti uživatele nebo zpravodajství ze zdrojů, které uživatel předem označí. V polovině roku 2004 se také na adrese http://mysearch.yahoo.com objevila služba personalizovaného vyhledávání, která uživateli po registraci umožní třídit nalezené odkazy podle nastavení, blokovat nevyhovující stránky a ukládat si historii vyhledávání.

Yahoo! se snaží poskytovat také informace lokálního charakteru. Proto od roku 2004 funguje služba *Yahoo! Local*, která poskytuje informace o dění ve vybraných deseti městech USA (například New York, Los Angeles nebo San Francisco) [WASSERMAN, 2004].

Nejnovější reklamní kampaň Yahoo! se snaží přesvědčit potenciální i stávající zákazníky o tom, že se svými integrovanými službami (e-mail, finanční zpravodajství, finanční zpravodajství apod.) se může stát jediným informačním zdrojem, který budou potřebovat [GARDNER, 2004]. V první řadě je ale nutné zákazníky přesvědčit o tom, že kombinované vyhledávací schopnosti Yahoo! se vyrovnají zatím nejpopulárnějšímu způsobu hledání informací na internetu – vyhledávači Google, který má zatím největší podíl v oblasti vyhledávacích služeb na internetu. V této souvislosti nabízí Yahoo! řadu služeb pro práci i volný čas; od specializovaných stránek pro zájmové skupiny (děti, koníčky apod.) po software a komerční služby spolupracujících firem (např. Verizon, AT&T).

Yahoo! se snaží neustále zvyšovat svůj podíl na trhu internetové reklamy, který přináší vyhledavačům a zábavním portálům největší část zisku. Ve srovnání se svým největším rivalem – společností Google však pořád citelně zaostává.

6.5.3 The WWW Virtual Library

http://vlib.org/

Projekt *„The WWW Virtual Library"* (VLIB) je jedním z prvních katalogů na webových stránkách. Zajímavostí je, že byl zahájen (dnes již sirem) Timem Berners-Lee, tvůrcem jazyka HTML a konceptu prostředí World Wide Web (www). Od svého počátku v roce 1993 je tento projekt založen na spolupráci dobrovolníků, kteří udržují jednotlivé sekce tohoto katalogu. Celý projekt je také od ledna 2000 řízen komisí, která hlavně koordinuje spolupráci částí projektu mezi sebou. I když VLIB je poměrně malým katalogem webových stránek, patří mezi historicky zajímavé projekty, které ukazují snahu organizovat obsah webových stránek od zrodu této služby. Hlavním cílem projektu je poskytovat přehled o stránkách s přehledným, aktuálním a vyváženým obsahem, který je možné snadno využít. I když je Virtual Library webovým katalogem, je možné k odkazům doplnit i stručný úvod k odkazované problematice. Zajímavostí je i organizace samotného katalogu. Ten fakticky propojuje 263 samostatných katalogů na různých serverech po celém světě.

Celý systém je založen na podobném principu jako projekt Open Directory (viz kapitola 6.5.1, s.78) kde dobrovolníci spravují tematické sekce, o které se sami zajímají a udržují katalog aktuální. Registrační proces pro editory je zde velmi jednoduchý. Potenciální editor si pouze vybere oblast, kterou chce spravovat, zaregistruje se do databáze a může začít pracovat. Projekt Virtual Library má pro své editory také pravidla pro práci, která jsou spíše obecná a nelze je srovnávat s podrobně rozpracovanými pravidly Open Directory. Stránky, které jsou přidávané do katalogu by měly nést logo projektu, na stránkách musí být možnost zpětné vazby pro uživatele a editor se musí zaregistrovat v centrální databázi a následně odebírat pravidelné informace o projektu [MANNING, 2002].

Projekt Virtual Library se liší ve způsobu práce editorů – každý z nich si sám vytváří stránku s odkazy, na kterou umístí logo projektu a adresu odešle hlavnímu administrátorovi projektu. Stránka je poté po dobu jednoho měsíce ve zkušebním provozu, kdy ji ostatní editoři mohou komentovat a navrhovat její úpravy. Po této době je stránka zařazena do katalogu a je považována za plnohodnotnou stránku projektu.

Centrální katalog, který má i tři zrcadla (Velká Británie, Švýcarsko, Argentina), tak obsahuje pouze dvě úrovně hierarchie, které odkazují na stránky editorů, kde je třetí (a často závěrečná) úroveň hierarchie a záznamy dokumentů. Touto distribucí stránek a kategorií mezi své editory je tento projekt jedinečný. Přes malý objem odkazů a nejistý provoz projektu v současnosti (nejsou k dispozici aktuální zprávy o jeho vývoji), je tento projekt historicky pozoruhodným pokusem o třídění webových stránek.

6.5.4 Looksmart

looksmart

http://www.looksmart.com

Pozn. Jde již o zaniklou službu, uvedení služby slouží k dokreslení předchozí nabídky katalogů.

Looksmart, založený roku 1995, byl typickým představitelem komerčního portálu, který nabízel vyhledávač, webový katalog a nabídku vyhledávání novinových článků, která byla poskytována ve spolupráci se společností Thomson – Gale. Je částí celé sítě služeb, pod kterou například patří vyhledávač Lycos nebo informační portál věnovaný výpočetní technice – CNET.

Katalog Looksmart bylrozdělen na dvě části – první je určena pro registraci komerčních stránek (placená služba) a druhá pro registraci stránek nekomerčních. První část pod označením **„Looklinstings“** vyžadovala registraci a po zaplacení stránku uvedla pod sponzorovanými odkazy, které slouží jako reklama typu „pay-per-click“ (poplatek se účtoval teprve poté, co uživatel na uvedený odkaz klikl a dostal se tak na inzerovanou stránku). Katalog Looksmart jako takový zanikl a jeho mateřská společnost se nyní soustřeďuje na poradenství v oblasti vyhledávání a reklamy

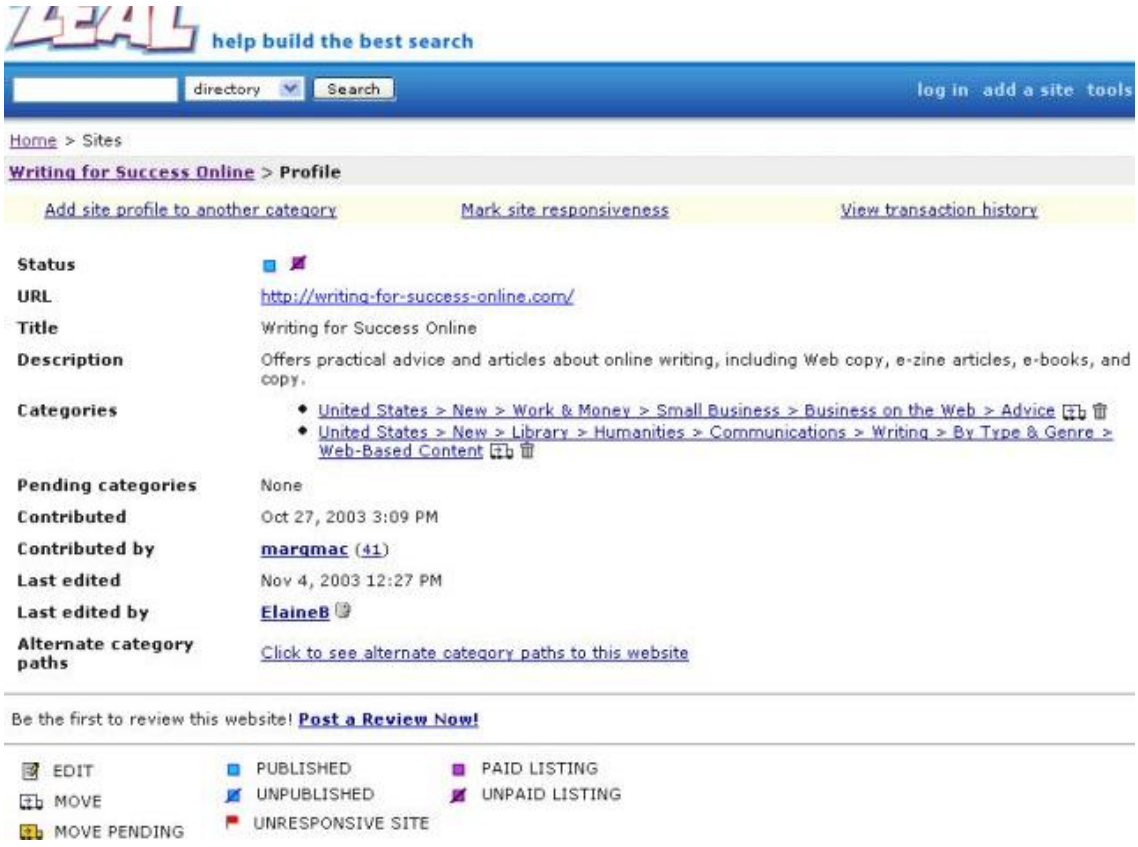
Druhou částí - komunitně tvořený **katalog Zeal**[36,] který byl součástí sítě Looksmart a tvořil vlastní jádro katalogu, nabízeného na stránkách looksmart.com Oba zmíněné katalogy byly dostupné z kategorií na stránkách Looksmart.com. Placené odkazy jsou zde přímo uvedené, zatímco katalog Zeal.com byl pouze odkazován.

Služba katalogu **Zeal.com byla bohužel na začátku roku 2007 zrušena**, oficiálně zdůvodněná „změnou obchodní strategie“ portálu Looksmart. Povahu změny obchodní strategie společnosti Looksmart dokládá nová služba *Find Articles*, která prohledává recenze knih a směřuje čtenáře k jejich zakoupení. Katalog se tak změnou strategie stal spíše obchodním portálem a následkem zmizelo zajímavé řešení pro organizaci informací na internetu i komunita, která tento katalog pomáhala vytvářet.

Katalog Zeal.com měl deset kategorií nejvyšší úrovně, které měly až devět úrovní. Každá stránka katalogu měla u svých odkazů zajímavou grafickou legendu, která formou ikon u každého odkazu podává základní informace o zdroji (typ a obsah stránky apod.). Zajímavou funkcí bylo také hodnocení odkazů uživateli – každý může oznámkovat kvalitu odkazu na stupnici 1-10. Toto hodnocení má vliv na zařazení editora do vyšší kategorie s dalšími pravomocemi.

Zeal registroval všechny stránky, které nabízely zajímavé informace nebo službu, zdarma dostupné všem uživatelům. Přidávat stránky do tohoto katalogu mohli pouze zaregistrovaní členové [WALL, 2004]. Podle údajů na stránkách měla komunita Zeal.com k 18.12.2004 celkem 205 862 členů. Pro přidání stránky do tohoto katalogu bylo nutné dodržovat předepsaná kritéria, která určovala například možný formát stránek, zakázaný a povolený obsah, psaní titulků a popisků pro stránky nebo pravidla pro správu kategorie. Každý editor mohl přidávat stránky, které byly následně ještě vyhodnoceny kmenovými zaměstnanci sítě Looksmart a teprve poté přidány do katalogu. Jako pomůcka pro tyto editory existovalo dvacet kontrolních otázek, které pomáhaly zhodnotit, zda je stránka vhodná pro přidání.

36 http://www.zeal.com



Obrázek 14: Záznam z katalogu Zeal.com

Za přidávání stránek a další práci v katalogu dostávají dobrovolní editoři body, které slouží jako hodnocení jejich zapojení do komunity. S přibývajícími body se mohou posunout do kategorie uživatelů s vyššími pravomocemi až do kategorie *„Zealot“*, s nejvyššími pravomocemi pro správu odkazů. Tato kategorie umožňuje mimo jiné samostatně spravovat zvolenou kategorii a vkládat do ní odkazy přímo bez nutnosti jejich schválení editory společnosti Looksmart [LOOKSMART, 2004]. Tímto způsobem motivace editorů je možné zařazovat velké množství odkazů s minimálními náklady pro firmu.

Podle tiskového prohlášení z roku 2003 obsahoval katalog zeal.com přes 250 000 odkazů [LOOKSMART, 2003]. Počet komerčně zařazených odkazů není přesně znám. Dohromady tak může jít řádově o několik set tisíc odkazů, což je několikrát méně než největší katalogy Open Directory a Yahoo!.

Jako náhradu zaniklé komunity Zeal.com vytvořil Looksmart službu **Furl.net.** Tvůrci je služba označována jako *„osobní online organizační složku“*, která umožňuje archivovat, třídit a sdílet informace, dostupné v prostředí webu. Základní vlastností je ukládání plných textů článků, nikoli pouze odkazů na ně. Tím dochází k vytváření nové uživatelské komunity, které se soustřeďuje okolo zajímavých témat a zdrojů. Služba Furl.net funguje zatím přibližně jeden rok a tak je na její zhodnocení příliš brzy. Rozhodně ale představuje potenciál pro řešení individuálního pořádání informací v prostředí world wide webu.

6.6 SPECIALIZOVANÉ KATALOGY

6.6.1 About.com

http://www.about.com

V roce 1997 založil Scott Kurnit společnost The Mining Company, o dva roky později přejmenovanou na About.com s vizí nového typu navigační služby. Základní myšlenkou celého projektu je, že člověk je nejlepším průvodcem k informacím, které jsou v prostředí internetu dostupné. Vznikl tak specializovaný portál, zaměřený na různá témata, která spravují

odborní editoři, jejichž počet k listopadu roku 2008 dosahuje počtu 750.

About.com je komerčním projektem a tak je zajímavé porovnat model správy těchto stránek. Na rozdíl od sítě dobrovolníků jsou editoři odborníky ve svém oboru, kteří musí nejprve projít výběrovým řízením, ve kterém se hodnotí jejich odborná způsobilost k tématu. Po úspěšném řízení každý editor začíná spravovat svého tématického průvodce, ke kterému přidává nejen odkazy na ostatní stránky, ale také sám píše články. Za tuto práci jsou editoři odměňováni – výše jejich honoráře závisí na počtu zobrazení jejich článků, společnost About.com garantuje minimální výši 725 dolarů, pokud budou přibývat počty zobrazení stránek [About.com, 2007].

Portál About.com je tak hybridním projektem, který se pohybuje mezi encyklopedií, naučným slovníkem, webovým katalogem a tematickým portálem. Nabízená témata jsou značně široká – od vaření a koníčků přes filozofii, průvodce zdravím a duševními poruchami, biologii, programování až po cestovní průvodce. Každé téma má svá podtémata rozdělená podle obsahu na několik dalších. Jejich počet a obsah se liší u každé úrovně průvodce. Ve druhé úrovní je nabídka tematických podkategorií (např. u kategorie „Domácí úkoly“ podtémata „Umění“, „Jazyky“, „Literatura“ nebo „Věda“). Ve třetí, zpravidla poslední úrovni, se témata dělí na kategorie „Essentials“ (úvodní informace k problematice), „Articles & Resources“ (články a odkazy na jiné stránky) a „Buyer´s Guide“ (doporučení pro nákupy, případně nabídky zboží v on-line obchodech). U některých kategorií je i nabídka časopisů a knih společnosti PRIMEDIA Inc., která od roku 2000 tvůrce tohoto projektu – firmu About, Inc. vlastní.

Celý projekt odráží svůj komerční původ ve všudypřítomných reklamních panelech a odkazech, nicméně rozsah a kvalita informací, které nabízí, jsou v prostředí internetu zcela jedinečné.

6.6.2 Download.com

http://www.download.com/

Portál Download.com je součástí mediální sítě CNET, která se specializuje na služby a zpravodajství o výpočetní a audiovizuální technice. Download.com je katalogem softwaru z kategorie shareware a freeware, který je zdarma nabízen ke stažení. Odkazy jsou řazeny do hierarchie o třech úrovních – první určuje hlavní téma (např. Internet), druhá blíže specifikuje kategorii softwaru (např. Browsers) a třetí úroveň zobrazuje seznam dostupných programů s odkazy k jejich stažení. Tento projekt má jako hlavní cíl zpřístupnění softwaru svým uživatelům a tomuto účelu se podřizuje celá filozofie organizace odkazů. Proto je zde celá organizace pouze nezbytnou pomůckou pro dosažení potřebných informací.

6.6.3 eBay.com

http://hub.ebay.com/buy?ssPageName=h:h:cat:US

Nejznámější on-line aukční síť – *eBay* byla založena v roce 1995 jako obchodní platforma, kde je možné koupit a prodat téměř cokoli. V současnosti má přes sto miliónů zaregistrovaných uživatelů, kteří mohou nabízet své zboží. V některých zemích (např. Austrálie, Německo nebo Indie) existují i národní stránky této aukční sítě, které svým uživatelům nabízí rozhraní v národním jazyce.

Hlavní nabídka vede k hierarchickému stromu, který je určen pro nakupující. Pod 31 kategoriemi je v závislosti na druhu zboží nabídka dalších podkategorií, a celý strom má celkem tři až čtyři hierarchické úrovně. První úroveň je seznamem tematických skupin, druhá nebo třetí úroveň upřesňuje typ zboží a v poslední úrovni je nabídka konkrétního druhu a typu zboží, které se může lišit cenou a kvalitou. Nabídka zboží je často velmi nekonvenční a setkat se tak můžeme například s prodejem rezervovaných (a zaplacených) zájezdů nebo pohřebních uren.

6.6.4 Bizrate.com

http://www.bizrate.com/

Dalším obchodním portálem, který vytvořil vlastní webový katalog je *Bizrate*. Jeho obchodní filozofií je nabídka širokého spektra sortimentu od výpočetní techniky ke klenotům a následné porovnání cen v různých on-line obchodech, které zvolené zboží nabízí. Celkem je podle informací portálu indexováno 30 miliónů produktů z více než 40 000 obchodů. Portál registruje pomocí svého vyhledávače *ShopRank* veškeré dostupné obchody a jejich nabídku zboží. Zároveň nabízí obchodníkům možnost placené reklamy, která uvede jejich zboží na prvních místech mezi výsledky hledání. Pokud si zákazník porovná ceny, je odkázán přímo do on-line obchodu zvoleného prodejce. Uživatelé mají možnost hodnotit svou spokojenost se službami obchodníků a tak tento portál slouží zároveň jako uživatelské fórum.

Veškeré nabízené zboží je v hierarchickém stromu, který má tři úrovně. První úroveň nabízí dvacet tematických okruhů (např. Computers & Software), druhá úroveň dělí zboží podle typů (např. Hardware) a třetí nabízí různé typy produktů. Ty jsou zpravidla doplněny fotografií zboží, recenzí a cenovým rozpětím v registrovaných obchodech. Uživatel si může produkt označit a porovnat jeho parametry s jiným, nebo přímo porovnat ceny a termíny dodání v různých obchodech.

Třetí úroveň umožňuje také vyhledávat podle vlastností zboží. U všech kategorií lze vyhledat zboží podle výše ceny, výrobce, produktové řady a dále podle dalších vlastností nabízeného zboží (např. u digitálních fotoaparátů je možné vyhledávat podle typu záznamového média nebo podle optického rozlišení). Toto doplňkové vyhledávání funguje na stejném principu jako knihovnické třídění S. Ranganatana, kde jsou předměty tříděny podle svých vlastností a náhledů – tzv. *faset*. Celý systém tvorby kategorií a „fasetového vyhledávání“ funguje na základě automatizované kategorizace používaného vyhledávacího systému *ShopRank* bez lidské účasti.

7 AUTOMATIZOVANÁ KLASIFIKACE A KATEGORIZACE

Vynález pozdější zničí vynález předchozí
a vytlačí jej z lidské obliby.
Lucretius Carus Titus

Klasifikace prováděná lidmi je pro zpracování velkého množství dokumentů nevhodná. I když je relativně přesná, její nevýhody – cena, konzistence klasifikace a časová náročnost zpracování převažují. Od počátků automatizace se proto rozvíjejí myšlenky o tom, jak vytvořit automatizovaný systém, který by byl schopen třídit a klasifikovat dokumenty s minimálním lidským podílem práce, či úplně bez lidské práce.
Dynamický rozvoj internetu tuto potřebu ještě více zvýraznil a je zřejmé, že klasifikace prováděná lidmi nemůže stačit ani pro popsání menší části tohoto prostoru. Z přehledu lidmi tvořených předmětových katalogů a adaptací hierarchických klasifikací vyplývá, že i při distribuci práce mezi velké množství lidí existuje hranice, kterou stávající metodikou práce nelze překročit.[37]

7.1 ÚVOD DO AUTOMATIZOVANÉ KLASIFIKACE

Automatizovaná klasifikace je proces, při kterém jsou dokumenty shromažďovány a analyzovány specializovaným softwarem a na základě této analýzy přiřazeny do existující struktury kategorií. Jako struktura pro tyto systémy slouží zpravidla univerzální či oborové třídění, do kterého systém řadí další dokumenty, které vyhodnotil pro zmíněnou kategorii jako relevantní.
Jeden z autorů Nordic Metadata Project – Traugott Koch [1998] k tomuto spojení uvádí, že tradiční třídění může být užito jako rámce pro třídění ve znalostní bázi, kam systém ukládá informace o dokumentech. Tradiční třídění zde slouží jako třídící struktura, do které jsou ukládány záznamy o dokumentech na základě automatizovaných procesů. Tyto projekty řadí dokumenty k jednotlivým třídám klasifikace na základě komplexních lingvistických a dalších modulů, které již částečně spadají do oblasti umělé inteligence a nemohou zatím v přesnosti nahradit analýzu dokumentu člověkem. Zároveň zde platí i argument, zmíněný u projektu SOSIG – vývoj nové klasifikace je časově i finančně velmi nákladný a proto je někdy vhodnější převzít nedokonalý, ale již existující systém.
Některé z projektů jsou uvedeny ve výzkumné zprávě Petera Gietze[38] [2001], ale i v této oblasti je vývoj velmi dynamický a velká část projektů ze seznamu již neexistuje.

7.2 AUTOMATICKÁ KATEGORIZACE

Automatická kategorizace, označovaná také jako shlukování nebo klastrování (z anglického „clustering“), je proces, při kterém jsou dokumenty tříděny do skupin podle sdílených témat (charakteristik). Tyto skupiny nejsou předem dané a vytvářejí se automaticky na základě analýzy nalezených dokumentů. **Vytvářením nových skupin, založených na vlastnostech dokumentů se automatizovaná kategorizace liší od automatizované klasifikace,** která na základě analýzy rozřazuje dokumenty do předem stanovené struktury kategorií, často univerzálního hierarchického třídění.

Tvorba systémů automatické kategorizace je v současné době jedním z nejrychleji rostoucích odvětví informačního průmyslu, protože nabízí relativně spolehlivé řešení třídění dokumentů za přijatelnou cenu.

---

37 Např. projekt Open Directory uvádí přibližně 4 milióny katalogizovaných stránek.

38 http://www.daasi.de/reports/Report-automatic-classification.html

Hlavním problémem dokumentů, dostupných na internetu či intranetu, je jejich nestrukturovanost tj. dokument nemá jakékoli doplňkové informace (metadata, klíčová slova), ani není uložen v hierarchickém systému, který by prozradil alespoň kontextové informace (vztah k nejbližším tématům apod.). Systémy automatické kategorizace dokáží dokumenty třídit bez jakýchkoli dalších informací, nicméně jsou-li dokumenty zařazeny v hierarchii (adresáře, složky) nebo pokud obsahují metadata, přesnost kategorizace systému se výrazně zvýší.

Proces kategorizace se tak skládá ze dvou fází:

- nastavení a fáze „tréningu“ systému
- vlastní kategorizace.

Průběh tréningové fáze má na správnou funkci systému automatizované kategorizace největší vliv, především u systémů fungujících na principu porovnávání vzorků. V této části je potřeba naplnit systém menší skupinou dokumentů stejného typu, jaký bude tímto systémem klasifikován i v budoucnu. Tyto systémy jsou tzv. samoučící, což znamená, že jsou schopny na základě algoritmů zlepšovat svou schopnost analyzovat a zařazovat dokument do správné kategorie. Vložené dokumenty – tzv. tréningová skupina by měly co nejpřesněji odrážet typ a problematiku dokumentů, které budou zpracovány v budoucnu, a také odrážet jejich poměrové zastoupení.

Tréningová skupina je systémem nejprve automaticky oklasifikována a následně je tato klasifikace zkontrolována lidmi, kteří mohou špatně zařazené dokumenty přeřadit do vhodnější kategorie. V této fázi se analyzují hlavní chyby při zpracování testovacích dokumentů, zjišťují se jejich příčiny (terminologie, gramatika...) a následně se upraví systém. Podle reprezentativnosti dokumentů v tréningové skupině a délky testování je možné výrazně zvýšit přesnost kategorizace. Systém se na základě této výchozí skupiny bude učit kategorizovat další dokumenty, kde ale vzhledem k jejich objemu bude už mnohem menší možnost zjištění chyb a jejich nápravy.

Vlastní kategorizace se skládá z několika kroků. V první fázi je určen okruh dokumentů pro zpracování, který může být relativně fixní (firemní archiv) nebo dynamický (došlá e-mailová korespondence, firemní intranet). Systémy automatizované kategorizace se v současné době používají hlavně pro řešení informačních potřeb firem, a proto se jedná o omezený počet dokumentů, i když podle [Zorn, 1999] se u velkých firem může jednat i o terabyte elektronických informací za jeden den. Jde spíše o firemní řešení jako archivy, firemní intranet nebo dynamickou klasifikaci došlých e-mailů. Druhým krokem je úprava dokumentů pro vlastní kategorizaci, která zahrnuje především ošetření jazykových výrazů (stop slova, synonyma) a gramatických jevů (skloňování, časování, parafrázování apod.). Po tomto zpracování jsou dokumenty analyzovány statistickými a lingvistickými metodami, a výsledky těchto analýz jsou dále zpracovány místa výskytu,a četnosti klíčových slov, frází a jejich gramatik se vyhodnocují jejich hlavní témata. Ve třetí části se vyhodnocují společná témata dokumentů a na jejich základě se automaticky vytváří kategorie, kam jsou dokumenty, vyhodnocené jako příbuzné zařazeny. Některé ze systémů automatické kategorizace umožňují i generování seznamu těchto kategorií, který může být i vizualizován do hierarchické podoby.

7.3 METODY AUTOMATICKÉ KLASIFIKACE A KATEGORIZACE

Systémy automatizované klasifikace a kategorizace fungují na základě přístupů:

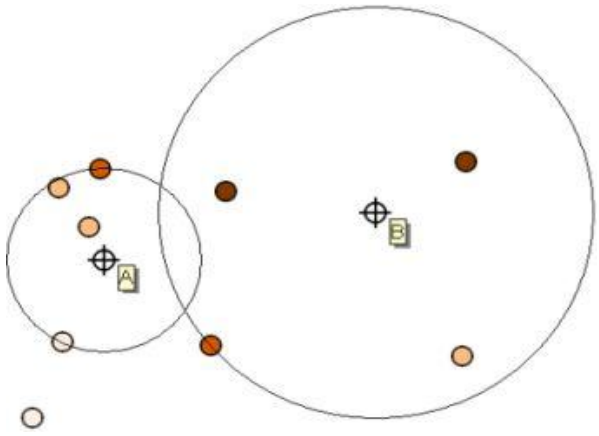
- porovnávání vzorků (pattern matching)
- definice pravidel (rule-based)
- kombinace výše uvedených přístupů

7.3.1 Porovnávání vzorů

Přístup, založený na porovnávání vzorů spočívá ve srovnávání dokumentů v interní (tzv. tréningové) bázi s dokumenty kategorizovanými. Podle podobnosti jsou vyhodnoceny příbuzné dokumenty, které jsou již řazeny stejně jako pokusný vzor.
Testování a „tréningová“ fáze jsou pro systémy založené na tomto přístupu zcela zásadní. Na úspěšnosti definice tréningového vzoru a jejím upřesnění záleží reálná přesnost celého systému v praxi. Nejrozšířenějšími metodami, které používají porovnávání vzorků jsou [Lubbes, 2003]:

- Metoda vzdálenosti mezi dokumenty
- Support Vector Machine (SVM)
- Bayesovské modelování
- Neuronové sítě

7.3.1.1 Metoda vzdálenosti mezi dokumenty



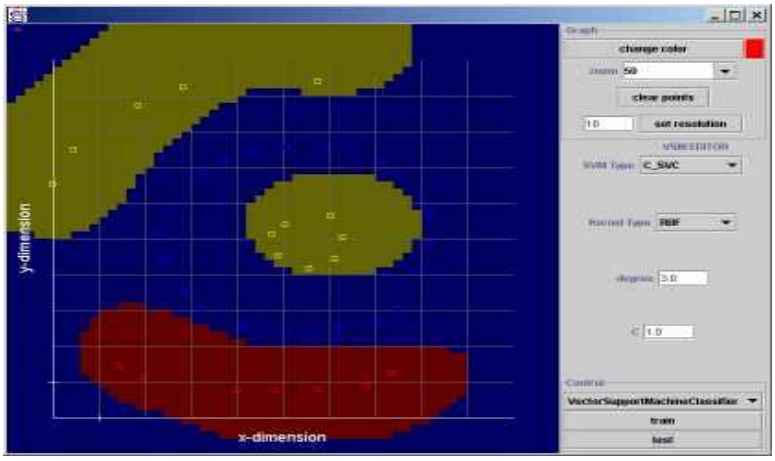
Obrázek 15: Model metody vzdálenosti mezi dokumenty Zdroj: http://www.quantdec.com/SYSEN597/GTKAV/section9/chapter_29b.htm

Tato metoda je založena na grafickém znázornění témat dokumentů v třídimenzionálním grafu, umožňuje tak vizualizovat příbuznosti témat podle jejich vzdálenosti od sebe. Každé téma je zobrazeno jako pravidelná kruhová množina (zóna) se

svým středovým bodem – tzv. centroidem, který je grafickým zachycením hlavních předmětových hesel tématiky. Hranice množiny okolo tohoto centroidu určují maximální možnou vzdálenost dokumentu, který může být ještě považován za relevantní pro tuto kategorii. Každý dokument může obsahovat více témat, a je tak znázorněný jako bod (body), umístěný nejblíže kategoriím, vystihujícím jeho obsah.

7.3.1.2 Support Vector Machine (SVM)

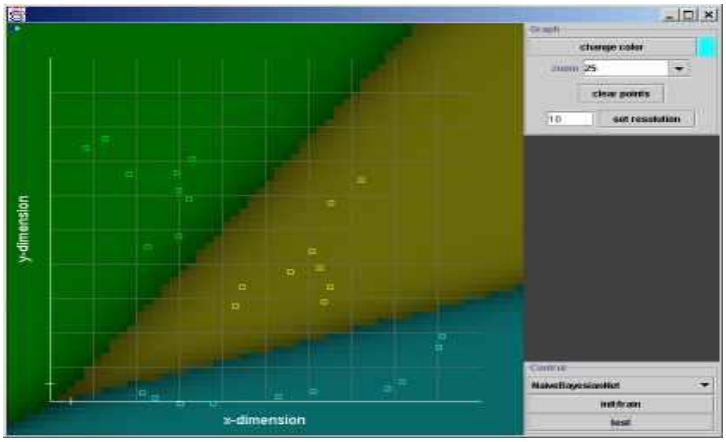
Kirk Lubbes [Lubbes, 2003] popisuje SVM jako formu rozšíření metody vzdálenosti mezi dokumenty. Ta předpokládá, že pro zobrazení hranice tematických zón je nejvýstižnější jejich zobrazení formou pravidelného kruhového tvaru. Metoda SVM předpokládá, že pravidelný tvar nemusí věrně reprezentovat skutečné hranice mezi tématy, a proto je zobrazuje ve formě tvarově nepravidelných zón. Zóny se mohou částečně překrývat, ale nepravidelný tvar dokáže mnohem lépe oddělovat témata od sebe.



***Obrázek 16**: Modelování množin SVM* *Zdroj: http://www.dreier.cc/index.php?topic=alnai&sub=svm*

7.3.1.3 Bayesovské modelování

Bayesovské modelování je založeno na teorii pravděpodobnosti. Předpokladem je výrok, že podle míry podobnosti dokumentů mezi sebou (tj. v testovací databázi a v ostrých datech) se dá určit i příbuznost tématická. Pokud je dokument z testovací skupiny správně zařazen do odpovídající kategorie (vektoru), existuje pravděpodobnost, že další podobný dokument je tematicky příbuzný. Problémem tohoto modelu je způsob práce se slovy a frázemi. Model předpokládá, že slova a fráze jsou nezávislá, a tak může dojít k významovému posunu při vyhodnocení tématu (např. rozdíl mezi „records manager“ a samostatnými slovy „record“ a „manager“) [Lubbes, 2003].



***Obrázek 17**: Bayesovské modelování Zdroj: http://www.dreier.cc/index.php?topic=alnai&sub=bayes*

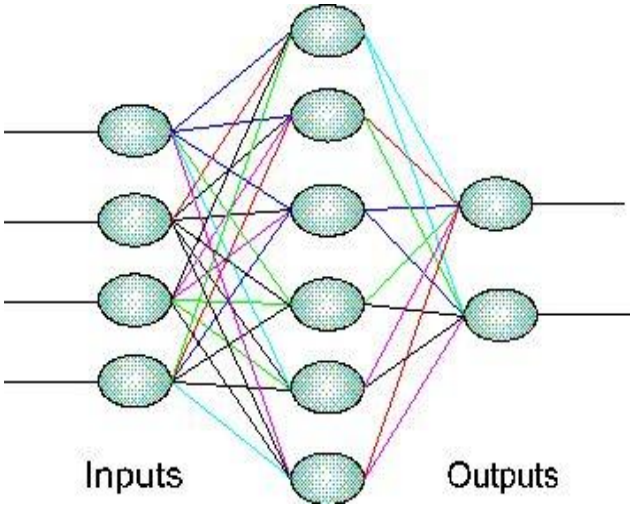
Neuronové sítě

Neuronové nebo přesněji „umělé neurální sítě"[Lubbes, 2003] jsou komplexní metodou pro zhodnocení témat dokumentu z různých pohledů. Neuronové sítě jsou určitým pokusem o simulaci funkce lidského mozku. Ten umí propojovat různé koncepty na základě asociací, a následně je podle jejich charakteristik zařazovat do stávající struktury poznatků. Podobně funguje neuronová síť – tj. na základě mnoha propojených částí (neuronů), ve kterých se paralelně vyhledává řešení specifického problému. Při jeho řešení jsou takto objevovány různé nové asociace nebo přístupy k jeho úspěšnému vyřešení.

Podle způsobu „učení" můžeme neuronové sítě rozdělit na dva typy:

Sítě, které se „učí s učitelem" (např. perceptronové), jsou založeny na tom, že neuronová síť **srovnává svůj výstup s výstupem svého učitele** a nastavováním vah synapsí upravuje výstup (hodnot v matici) tak, aby se snížil rozdíl mezi skutečným a požadovaným výstupem [MYSLÍK, 1996]. Do systému jsou tak ve dvojicích vkládána vstupní i výstupní data. Znamená to, že se neuronová síť na základě vstupních dat učí, jaký výstup je požadován. Množina dat, použitá pro vyladění funkce systému (data „učitele") se nazývá *„tréningovou množinou"* a na výběru dokumentů, které jsou v ní obsažené, závisí i budoucí přesnost hodnocení.

Samoorganizující sítě nemají žádný srovnávací vzorek k určení správnosti (jde tedy o proces učení bez účasti učitele – angl. unsupervised learning). *„Algoritmus těchto sítí je navržen tak, aby ve vstupních datech hledal vzorky s určitými vlastnostmi, tj. podle závislosti. Tak je třeba možno analyzovat, jaký vliv má roční období na burzu, počet myší na úrodu apod."* [MYSLÍK, 1996]. Na základě nalezených zákonitostí jsou sítě schopny vstupní data organizovat.



***Obrázek 18**: Jednoduchý model neuronové sítě Zdroj: http://www.e-*

Neuronové sítě mohou odvodit vzory nebo analyzovat trendy i v komplikovaných nebo nepřesných souborech dat. Podle studie [Stergiou – Siganos, 1996] patří mezi výhody používání neuronových sítí:

1. Adaptivní učení
   schopnost učit se na základě vložených tréningových dat
2. Sebeorganizace (self-organization)
   vytvoření vlastní organizace a reprezentace dat, která přijímá v procesu učení
3. Operace v reálném čase
   procesy řešení problémů probíhají paralelně, tj. souběžně je řešeno několik kroků či variant
4. Chybová tolerance prostřednictvím kódování redundantních informací síť může fungovat i při částečném poškození; snižuje se ovšem její výkonnost.

Neuronová síť funguje podobným způsobem jako lidský mozek, tj. na základě mnoha propojených částí (neuronů), ve kterých se paralelně vyhledává řešení specifického problému. Při jeho řešení jsou takto objevovány různé nové asociace nebo přístupy k jeho úspěšnému vyřešení.

Úspěšnost neuronové sítě závisí především na tréningovém vzorku, který musí obsahovat dokumenty s jednoznačným významem (odstranění polytematických dokumentů apod.). Nejsou-li dokumenty v tréningovém vzorku dobře zvoleny, může to vést k chybné funkci celého systému.

7.3.2 Definování pravidel

Systémy založené na definování pravidel třídí dokumenty podle uživatelských nastavení na základě výskytu určitých slov, frází nebo komplexních výrazů. Veškeré podmínky jsou definovány ve formě logického výrazu IF (podmínka) THEN (požadovaná operace). Pravidel může být definováno i více a v systému je možné nastavit i prioritu jejich zpracování, případně pravidla seřadit do rozhodovacího diagramu. Systém je založen na aplikaci stávajících znalostí uživatele do podoby pravidel (aplikované znalosti), kterými se určí způsob zpracování dokumentu.

U těchto systémů platí, že čím větší je objem zpracovávaných dokumentů a jejich témat, tím větší počet podmínek je nutné nastavit. Tyto podmínky musí být navíc vzájemně kompatibilní nebo musí být určeno pořadí jejich zpracování.

Oblíbenost tohoto přístupu u uživatelů vysvětluje Fabrizio Sebastiani slovy: *„Systémy, založené na definování pravidel, jsou populární, neboť uživatelé těchto systémů mohou přesně definovat kritéria, podle kterých je dokument klasifikován. Tyto systémy mohou podporovat komplexní operace a rozhodovací stromy a přinášet tak velmi přesné výsledky“* [LUBBES, 2003]. Pomocí vyladěného rozhodovacího stromu lze u těchto systémů dosáhnout částečného porozumění sémantice dokumentů a ve spolupráci s tezaurem nebo jiným slovníkem odborné terminologie zpřesnit zařazení dokumentů do příslušných skupin.

7.3.3 Vizualizace

Vizualizace je kognitivní proces poznávání člověka, při kterém si formuje mentální obraz určité oblasti, která ho obklopuje (doprava, příroda, konkrétní vědecký obor apod.). Tento mentální obraz se v oblasti pořádání informací objevuje buď v podobě klasifikačního systému hierarchického typu, nebo ve formě tzv. *vědecké vizualizace* s pomocí některého z přístupů automatické kategorizace (nejčastěji je k tomuto účelu používána metoda neuronových sítí). V informatice a informační vědě znamená termín „vizualizace“ nebo „vědecká vizualizace“ následující: *„vizuální reprezentace určité oblasti pomocí grafiky, obrázků, animovaných sekvencí stejně jako zvukového záznamu pro ztvárnění obsahu, struktury a dynamického chování velkých datových skupin, které zastupují systémy, události, procesy a objekty.”* [WILLIAMS, 1995, S. 163].

Grafická forma umožňuje zobrazit asociace, které by v jiné formě zobrazení nebyly patrné a i ve velké skupině dat pomůže analyzovat vzory, podle nichž jsou data seskupena. Vizualizace je přínosná především jako grafická navigační pomůcka, která umožňuje uživateli stejný systém orientace, na jaký je ve své kultuře zvyklý [WILLIAMS, 1995].

Pro účely klasifikace může mít vizualizační aplikace podobu map, schémat nebo grafů ve dvou nebo i třech dimenzích zobrazení. Konkrétní podobou vizualizace může být vytváření grafických map, animace, sdružování prvků k sobě nebo transformace dat. Výrazovými prostředky pro vizualizované aplikace jsou velikosti grafických oblastí a především barva a světlo. Předpokládá se, že vizualizace se v budoucnu zaměří i na další smysly a bude přinášet i zvuk, případně další, nevizuální vjemy.

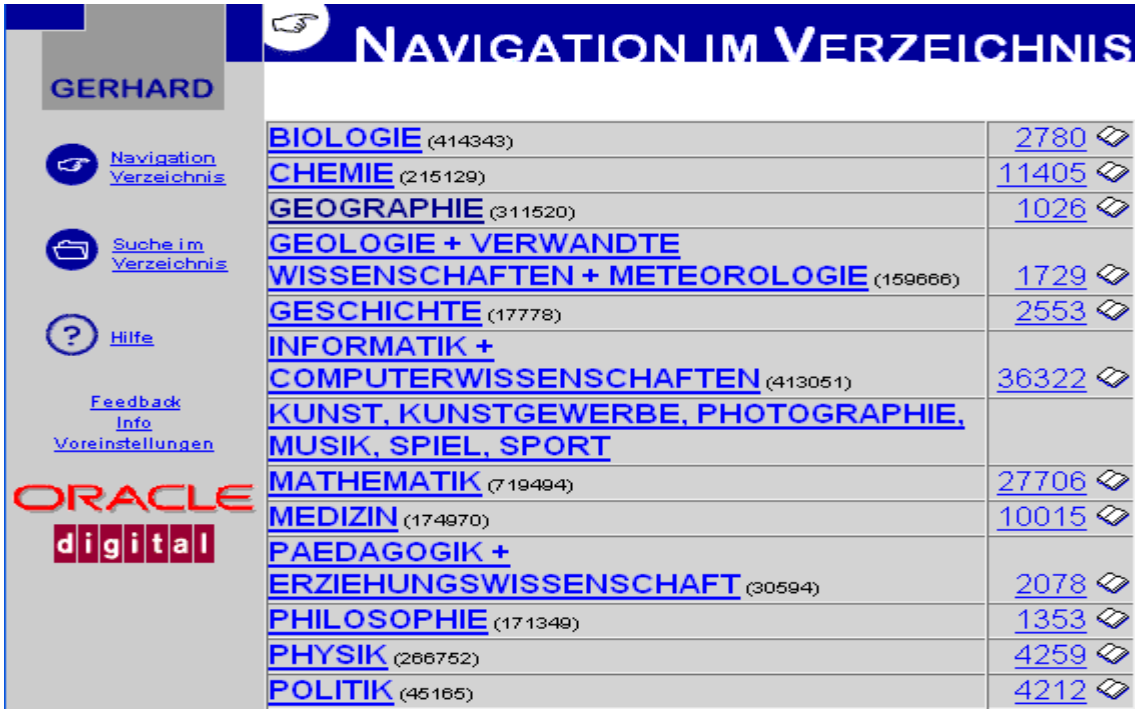
Řešení klasifikačních systémů pomocí vizualizační technologie není dokonalé, hlavní problém je v prezentaci dat, kde platí stejná omezení jako u statistik tj. prezentace dat může zásadně ovlivnit pochopení celé komunikované informace. Z tohoto důvodu je nutné vizualizační systémy ověřovat a získat zpětnou vazbu od jeho uživatelů.

7.4 PROJEKTY AUTOMATIZOVANÉ KLASIFIKACE

7.4.1 German Harvest Automated Retrieval and Directory (GERHARD)

http://www.gerhard.de

GERHARD je specializovanou službou, která pomocí automatizovaných procesů sbírá informace o webových stránkách v německém jazyce, které následně klasifikuje do struktury Mezinárodního desetinného třídění (MDT). Projekt vznikl v roce 1996 spoluprací v knihovně univerzity Oldenburg za podpory **Deutsche Forschunggemeinschaft (DFG)**. GERHARD má navíc velmi dobrou podporu i od komerčních firem – techniku pro projekt dodala společnost Digital a databázový software pro změnu firma Oracle. Hlavním cílem projektu bylo spojení dvou přístupů pro vyhledávání – prohlížení (kontextové navigace) i vyhledávání vědeckých a odborných dokumentů v prostoru německých webových stránek.



***Obrázek 19**: Rozhraní projektu GERHARD* *Zdroj: http://www.gerhard.de*

Z důvodu obtížného zpracovávání různých formátů dokumentů jsou sledovány pouze dokumenty ve formátu HTML. Podle autorů [WATJEN, 1998] tvořil formát HTML v pilotním projektu, který zkoumal dostupné formáty dokumentů na stránkách univerzity v Oldenburgu, pouze cca 15% celkového datového objemu souborů, ale zároveň 70% všech textových dokumentů. Toto pokrytí se jeví jako dostatečné a pokrytí dalších formátů (PDF, LaTeX, PS...) bylo v době vzniku formátu velmi problematické. Další problémy byly zjištěny i v rámci přístupu k dokumentům prostřednictvím jiných protokolů – např. FTP, což bylo jen dalším argumentem pro omezení rámce hledání pouze na protokol HTTP.

Pro indexaci dokumentů byla stanovena německá jazyková oblast (stránky v němčině a angličtině) a indexace je zaměřena na dokumenty na serverech univerzit a výzkumných ústavů, státních organizací a vědeckých společností. Program pro indexaci stránek – tzv. harvestor je pokročilejší verzí softwaru, který byl použit v jednom z předešlých evropských projektů – **Nordic WAIS-WWW Project**. Výkonnost harvesteru je přibližně 170 000 indexovaných stránek za měsíc. K říjnu 2004 tento projekt na svých stránkách uvádí registraci 1 284 819 dokumentů a 6 182 891 přiřazení dokumentů k některé ze 70 000 tříd MDT.

Mezinárodní desetinné třídění (MDT) bylo zvoleno jako klasifikační rámec pro uspořádání stránek, které jsou automaticky indexovány. Pro potřeby automatické indexace bylo do databáze vloženo přibližně 27 megabytů dat, které zastupovaly 70 000 tříd MDT a jejich notací. Proces klasifikace spočívá v extrakci frází z HTML stránek, jejich srovnání s popisným textem MDT tříd, a v následném přidělení notace stránce.

Použitý klasifikační systém pracuje v následujícím krocích:

úprava třídění

úprava HTML textu (kódu)

analýza notací

První část – úprava notací spočívá v úpravě přirozeného jazyka klasifikace do podoby, ve které je tento jazyk strojově zpracovatelný. Jedná se především o vynechání různých komentářů a vysvětlivek, vložených mezi třídami klasifikace, a o úpravu názvů tříd odstraněním diakritiky a přehlásek. Po těchto krocích jsou z názvů tříd a podtříd připraveny fráze, které by se mohly ve zpracovávaných dokumentech vyskytovat. Tyto fráze jsou připraveny i s ohledem na přirozený jazyk, takže existují různé varianty slov a synonyma, která mají pomoci při zpracování různých gramatických a lexikálních jevů. Pro další zpracování je nezbytné zachovat komplexní záznam, který bude obsahovat notaci, popis a synonymum.

Příklad:

Notace	Popis	Synonymum
369,5.000.504	ženy a životní prostředí	životní prostředí a ženy

V některých případech bylo nutné ošetřit i různé morfologické jevy německého jazyka tak, aby bylo možné najít tyto fráze ve zpracovávaném dokumentu.

Příklad:

Termín MDT	Textová sekvence
umwelt und frauen	umwelt und die frauen
klassisches griechisch	aspekte des klassischen griechischs

Tento postup vyžaduje odstranění tzv. stop slov a omezení dalších tvarů slov prostřednictvím skloňování a časování. Pro tento účel zpracoval Institut für Semantische Informationsverarbeitung seznam stop slov v angličtině a němčině, který se skládal z různých slovních druhů (předložky, příslovce) a také ze sloves, často užívaných jako pomocná. Zároveň pro potřeby tohoto zpracování existuje pomocný morfologický strom, který ukazuje varianty rozšíření kmene slova (např. frau-frauen). Každé slovo s touto vlastností, je pak označeno speciálním symbolem, který umožňuje snížit rozepsané počty variant slova.

V druhé části je zpracován text stránky. Z formátu HTML je převeden do formátu prostý text – ASCII a jsou z něj odstraněna veškerá diakritická znaménka. Dále je text analyzován na stop slova, která jsou odstraněna.

Ve třetím kroku – analýze jsou srovnávána slova a fráze ze slovníku MDT s textem zdrojového dokumentu. Pokud slova odpovídají nějakým záznamům, je zvolen ten, který má nejdelší popisný text. Například text dokumentu *„klassischen grieches“* by mohl být přiřazen záznamu s textem *„aspekte des klassischen griechischs“***.** Možnost zpracování víceslovných výrazů je ale omezena. Jako částečné řešení je možné přiřadit ke hledaným slovům symbol pro rozšíření, který vyhledá jakýkoli tvar slova, které bude na začátku obsahovat shodný kořen.

Každému dokumentu je možné přiřadit více notací. V další fázi je dokument a jemu přiřazené notace zpracován statistickou analýzou. Ta využívá hierarchickou strukturu třídění a na jejím základě srovnává podobnost přiřazených notací podle jejich vzdálenosti

v hierarchickém stromu. Čím více notací má dokument přiřazeno, tím přednější je i jeho zařazení do tematické skupiny MDT. Zároveň platí, že čím delší popisný text je pro dokument zvolen, tím je jeho zařazení ke třídě specifičtější [WATJEN, 1998].

Financování projektu GERHARD skončilo v roce 1998. Vzhledem k tomu, že se projekt ukázal jako přínosný, bylo rozhodnuto o pokračování projektu ve druhé fázi pod názvem **GERHARD II**. Návazný projekt byl spuštěn až po zajištění finančních zdrojů v roce 2001. Hlavní cíle návazného projektu jsou následující:

- vylepšení statistického softwaru pro kategorizaci
- vylepšení shromažďování a výběru dokumentů (filtrování dokumentů s nesmyslným obsahem)
- vylepšení klasifikace (rozšíření stávajícího počtu tříd MDT, aktualizace lexikonu a lepší rozlišení notací MDT)
- přidání automatického rozlišení jazyka a typu dokumentu
- vytvoření profilových služeb pro uživatele (profily pro hledání apod.)

[Diekmann, 2002].

Projekt GERHARD je v oblasti automatizované klasifikace dokumentů na internetu jedním z nejpokročilejších. Jako důkaz kvality projektu můžeme vnímat i sponzoring projektu ze strany firem IBM a Digital. Pokud se podaří zajistit financování a další vývoj, může tento projekt přinést řadu zajímavých řešení pro automatickou klasifikaci.

7.4.2 DESIRE

http://www.desire.org/

Development of a **E**uropean **S**ervice for **I**nformation on **R**esearch and **E**ducation (DESIRE) byl rozvojovým projektem Evropské Unie, který probíhal ve dvou fázích (1996-1998 a 1998-2000), za spolupráce deseti institucí z Nizozemí, Norska, Švédska a Velké Británie. Jedním z cílů tohoto rozsáhlého projektu byla práce na automatizované klasifikaci, kterou zajišťovala laboratoř NetLab univerzitní knihovny v Lundu (Švédsko) pod vedením Traugotta Kocha. V rámci projektu se porovnávaly možnosti automatizované a manuální klasifikace za účelem jejich optimální kombinace.

V první části projektu DESIRE byla vybudována síť oborových předmětových bran, jako jsou EELS,[39] SOSIG[40] nebo DutchESS[41]. Zpětná vazba od uživatelů ukázala, že sice oceňují katalogizaci kvalitních oborových zdrojů v těchto branách, nicméně chtějí hledat v takovém množství zdrojů, který by byl srovnatelný se současnými vyhledávači. To bylo podnětem pro druhou část projektu, která měla zjistit možnosti pro automatizaci části těchto předmětových bran. Dílčí úkoly byly stanoveny následovně:

---

39 http://eels.lub.lu.se

40 http://www.sosig.ac.uk

41 http://www.kb.nl/dutchess

- vylepšení metod shromažďování dokumentů pomocí automatizovaného předmětového indexu
- testování různých metod automatizované klasifikace na dokumentech z prostředí www (dokumenty ze stejného oboru)
- vývoj vylepšené kombinace manuálně tvořeného katalogu a automaticky generovaného předmětového seznamu zdrojů, který by umožňoval prohlížení formou hierarchické klasifikace a zároveň prohledávání příbuzných oborových bran pomocí protokolu Z39.50, indexačního softwaru Zebra[42] a metadat standardu Dublin Core.

Pro testování automatizované klasifikace byla vybrána oblast technických informací a jako rámec pro třídění byl zvolen tezaurus, používaný v databázi INSPEC – Ei thesaurus. Pro testovací účely byla vytvořena databáze, obsahující 155 611 webových stránek ze serverů, které poskytují informace pro tuto oblast.

Ei thesaurus byl před použitím pro automatickou klasifikaci upraven vyřazením tzv. stop slov, převodem popisných termínů z velkých na malá písmena, vyřazením geografických jmen (což nebylo hlavním úkolem klasifikace), vyřazením jednoho či dvou znakových výrazů pro snížení redundance a úpravou slov softwarem Porters pro potenciální pravostranné rozšíření při hledání odpovídajících výrazů.

Na základě záznamů ve zkušební databázi byly extrahovány veškeré informace o metadatech, nadpisech a textu dokumentu. Poté byla srovnána veškerá slova z tezauru s extrahovaným textem a pokud byla nalezena shoda, k záznamu byl přiřazen odpovídající klasifikační kód. Tyto kódy byly ohodnoceny tzv. skóre, což byl údaj, který ukazoval míru shody na základě komplexnosti termínu (jednoslovný výraz, fráze), umístění termínu (nadpis, metadata, text) a typu klasifikace (hlavní klasifikace, doplňková klasifikace).

Každý víceslovný výraz byl také vyhodnocen na úplnost podle typu umístění v textu a na tomto základě mu byla přidělena váha podle níže uvedené tabulky (tabulka 8).

Toto hodnocení více umožňuje rozlišit mezi hlavním popisným výrazem a doplňkovými klíčovými slovy v obou typech klasifikace. V hodnocení byl při vážení významu více znevýhodněn termín, nalezený pomocí booleovské algebry a jednoslovný výraz, protože jsou hlavními zdroji chyb (nalezení homonym apod.). Vyhodnocení tohoto experimentu s automatizovanou klasifikací bylo rozděleno na několik částí, které měly vyhodnotit třídění z různých aspektů.

Typ termínu

klasifikace	přesná fráze	výraz nalezený ve větě boolovskou algebrou	jedno slovo z výrazu
OC	4	2	1
MC	8	3	2
MC - main classification and/or OC - optional classification			

***Tabulka 7**: Vážení významu frází v projektu DESIRE*

Automatizované zařazení do úrovní třídění přibližně odpovídalo zařazení ve zdrojovém tezauru, nicméně rozdělení záznamů do tematických tříd nebylo rovnoměrné. Důvodem byla patrně tematika dokumentů, které jsou v prostředí webových stránek dostupné; tematicky neodpovídala všem třídám tezauru a některá témata (obecná mechanika) jsou tak zastoupena více než jiná (mechanizace pro těžbu). Autoři [Ardö, 1999] proto poznamenávají, že větší strukturovanost hierarchie může velmi zpřesnit výsledky automatizované klasifikace.

Srovnání upravené podoby tezauru pro automatizované zpracování s originální verzí ukázalo, že tyto úpravy (možnost pravostranného rozšíření apod.) odpovídají originálnímu Ei thesauru na 57-66%. Na přesnost má vliv především počet slov, kde platí přímá úměra (jednoslovný výraz odpovídá přesněji než víceslovný).

[42] http://www.indexdata.dk/zebra/

Pro zjištění míry přesnosti automatizované klasifikace byl proveden ještě další test, ve kterém experti z oblasti techniky a technologie zkoumali míru relevance klasifikace na náhodně vybraných třídách. Míra přesnosti ve třech zvolených třídách kolísala od 37% po 75,5%. Míra celkové správnosti tak je přibližně 59%, což odpovídá intervalu přesnosti při srovnávání upravené podoby tezauru s originální verzí. Rozdíly mezi výsledky v různých třídách však neumožňují tento údaj chápat jako míru přesnosti metody automatizované klasifikace.

Hlavní zdroje chyb v klasifikaci byly v následujících oblastech:

1. termíny v tezauru se vyskytují i v jiných kontextech, než se předpokládalo

Na základě vyhodnocení pak byly některé dokumenty zařazeny do více tříd, z nichž některé neodpovídaly jejich obsahu. Například termín „drives" byl řazen do třídy 602.1 (mechanical drives) a 705 (electric generators and motors) stejně jako do třídy 632.4 (pneumatic equipment). Tento problém může být řešen větším rozlišením termínů doplňkovou klasifikací, a tak řadit termíny pouze do odpovídajících tříd.

2. náhodné přiřazení termínu tezauru nevýznamovému slovu

V některých případech došlo k chybnému přiřazení homonyma v plném textu stránky k termínu tezauru. Autoři jako příklad uvádí termín „natural language", který je v textu poměrně četný. Tento problém by mohl být odstraněn analýzou gramatiky a použitím slovníku významových frází, který by pomohl blíže určit význam výrazu.

3. zařazení dokumentu do příbuzné/hraniční kategorie

Některé dokumenty byly zařazeny pod třídu, která také odpovídá z hlediska obsahu, ale není hlavním tématem dokumentu. Problematickým se ukázalo rozlišení hlavních a vedlejších témat. To souvisí s přiřazením odpovídajících klíčových slov dokumentu a také s definicí dostatečně specifických kategorií v třídění. Tento problém by byl pravděpodobně vyřešen použitím univerzálního třídění, které má dostatek podrobně definovaných tříd, nicméně v menších oborových tříděních jako je Ei thesaurus bude i nadále přetrvávat.

Projekt DESIRE a především jeho část o automatizované klasifikaci přinesl řadu zajímavých zjištění a experimentů, z kterých mohou vycházet další spolupracující (GERHARD, SCORPION) či jiné, návazné projekty. Experimenty, provedené v rámci tohoto programu ukázaly, že výzkum automatizované klasifikace může dospět k výsledkům, které se již dají označit za prakticky využitelné, a jediným limitujícím faktorem tak zůstává finanční zabezpečení těchto výzkumů.

7.4.3 Project Renardus

http://www.renardus.org

Renardus je jedním z projektů, které byly realizovány ve spolupráci grantového programu UK **Electronic Libraries (eLib)** a širšího výzkumného programu Evropské unie pro rozvoj informační společnosti (Information Society Technologies Programme[43]).

Projekt vznikl v roce 2000 se dvěma hlavními cíli:

vytvořit rámec pro spolupráci Evropských oborových rozcestníků (subject-gateways), který by vedl k novým, nadstavbovým službám pro uživatele

- výzkum možností standardizace a technických řešení v oblasti sdílení metadat [Heery, 2001].

Renardus není přímo projektem, který by samostatně používal automatizovanou klasifikaci pro popis dokumentů na webových stránkách, ale slouží k automatizované reklasifikaci mezi dokumenty, které již byly zpracovány.
Renardus se měl zaměřit především na existující oborové rozcestníky, které většinou také vznikly za finanční podpory EU – např. SOSIG, EEVL nebo OMNI. Každý z těchto rozcestníků shromažďuje, kvalitativně posuzuje a třídí odborné zdroje na internetu, které následně zveřejňuje ve formě podrobných záznamů s anotacemi. Přínosem tohoto přístupu je poskytování velmi kvalitních informací oborové komunitě společně s možností jejich vyhledávání podle zadaných kritérií.

Problémem rozcestníků byla jejich izolovanost – každý z nich fungoval samostatně, bez dalších vazeb na ostatní, byť hraniční zdroje. Projekt Renardus si kladl za cíl vylepšit tuto spolupráci a pomocí automatizované klasifikace a mapování

[43] http://cordis.europa.eu/ist/

metadat v jednotlivých projektech poskytnout alternativní přístup k nabízeným informacím ve všech zapojených rozcestnících.

Přístup k záznamům jednotlivých rozcestníků poskytuje Renardus prostřednictvím tzv. **mapování klasifikace rozcestníků,** tj. jejich převodem do univerzální klasifikace. Smyslem tohoto převodu je distribuovat dotaz (třída univerzálního třídění) do specializovaných klasifikací, a tak poskytnout uživatelům možnost vyhledávat z jednoho místa ve více nesourodých systémech.

Jako univerzální schéma pro převod bylo zvoleno Deweyho desetinné třídění (DDC). Třídy DDC jsou mapovány do jednotlivých třídění (např. Mathematics Subject Classification nebo Nederlandse Basiclassificatie). K tomuto mapování je používán software, vytvořený jako součást německého projektu CARMEN.

Protože není možné, aby veškeré termíny měly ekvivalent v DDC, byly stanoveny tzv. **úrovně relevance**. Ty definují, kde je možné zvolit podrobnější nebo obecnější termíny pro popis nebo míru překrytí mezi termíny. Na základě tohoto mapování Renardus odkazuje přímo do hierarchie konkrétního rozcestníku současně s uvedením jeho jména.

Jedna z autorek projektu Rachel Heery [2001] upozorňuje na skutečnost, že toto mapování klasifikací je prozatím velmi experimentální, nicméně se může stát velmi zajímavou metodou pro slučování navigačních systémů mezi sebou.

Kromě tohoto mapování projekt Renardus zkoumá i možnost sdílení a interoperability metadat mezi rozcestníky. Jedná se především o standardizaci používaných schémat, sdílení metadat (tj. možnost vyhledávání v metadatech i pro cizí vyhledávací služby) a existují i plány na automatizovaný sběr metadat (tzv. harvesting). Zdá se, že kombinace hierarchické klasifikace společně s využitím metadat může být velmi zajímavým přístupem pro další pořádání informací. Spojuje totiž jak požadavky na popis zdroje pro vyhledávače (metadata), tak vizualizaci třídění pro člověka (hierarchická klasifikace).

V roce 2002 skončilo financování projektu z prostředků programu EU. Na stránkách je sice informace o konsorciu, které vzniklo pro další rozvíjení projektu, ale od roku 2002 nejsou stránky změněny. Z tohoto důvodu se zdá, že projekt byl v tomto roce také zastaven a zatím nejsou jakékoli známky o tom, že by mohl být obnoven [HEERY, 2001].

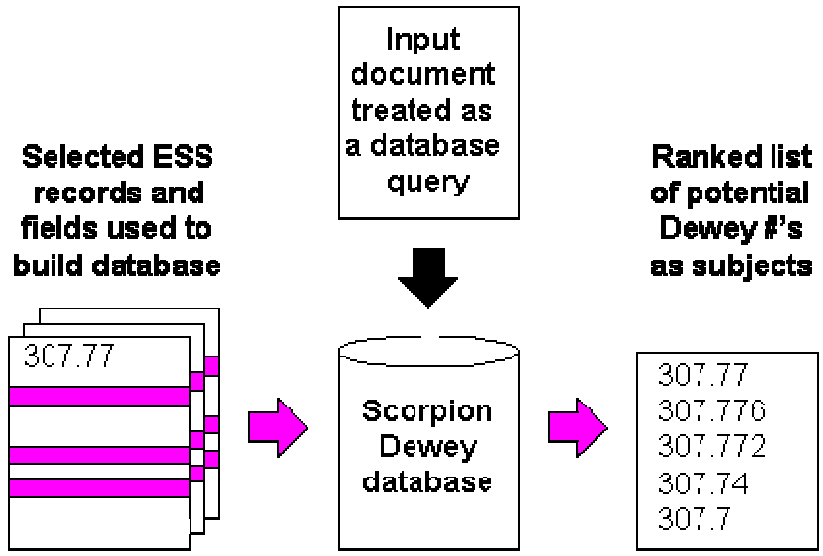
7.4.4 Scorpion

http://www.oclc.org/research/software/scorpion/default.htm

Projekt Scorpion je jedním z experimentů společnosti **OCLC** pro nový přístup ke katalogizaci a třídění dokumentů na internetu. Projekt vychází ze základního předpokladu, že lidské zpracování nikdy nemůže obsáhnout rozsah dokumentů, které jsou na internetu dostupné, a proto je jedinou možnou cestou pro zpracování těchto dokumentů automatická katalogizace a klasifikace.

Jak je patrné z obrázku 21, proces automatické klasifikace probíhá pomocí databázového systému, do kterého jsou vkládány webové stránky (jejich URL adresy). V databázi je celá stránka tematicky vyhodnocena specializovaným softwarem (ranking system) a následně je kontaktován elektronický systém Editorial Support System (ESS), který obsahuje Deweyho klasifikaci. Na základě tematického vyhodnocení jsou ke stránce přiřazeny třídy Deweyho klasifikace (v některých případech také klasifikace Kongresové knihovny), které jsou pak zobrazeny uživateli.

Jak upozorňují tvůrci [SHAFER, 1997], celý systém je experimentální, a ukazuje se, že je potřeba každý dokument pro



***Obrázek 20**: Klasifikace v projektu SCORPION Zdroj: http://orc.rsch.oclc.org:6109/b-asis.html*

automatickou klasifikaci „předzpracovat“. To znamená rozdělit dokument na několik samostatných tematických částí, a ty posléze klasifikovat. Tento přístup znamená také úzce spolupracovat s lingvistickým softwarem a „učit“ ho vyhodnocovat obsah dokumentu na požadované úrovni podrobnosti.

Hlavním výstupem projektu je tzv. **„Automatic Subject Assignment“** – formulář, pomocí kterého je možné vyhodnotit stránku na základě zadané URL adresy. Systém je vskutku experimentální, což je vidět i na výsledcích – vyhodnocení je velmi hrubé, nehledě na jazyk zadané stránky. Tento problém je zčásti řešitelný již zmíněným „předzpracováním“ stránky nebo zadáním stránky takové, která je tematicky jednotná – tj. neobsahuje pro vyhodnocovací systém matoucí informace, jako jsou různé odkazy na obecnější témata, stránky typu rozcestníku a podobně.

Tento projekt probíhal od roku 1996 a v roce 2000 byl ukončen, nicméně hodnotící formulář je pořád v provozu.

7.4.5 Wolverhampton Web Library – WWLib

http://www.scit.wlv.ac.uk/wwlib/newclass.html

Pozn. Projekt byl ukončen v roce 2006 a jeho stránky již nejsou funkční.

Katalog Wolverhampton Web Library (WWLib) je provozován School of Computing & Information Technology University of Wolverhampton. Na stránkách projektu je ale upozornění, že tento katalog není přímo oficiálním projektem University of Wolverhampton, ale spíše doplňkovou činností (spare time activity). Záběr tohoto katalogu je užší než u jeho konkurentů – projekt je zaměřen pouze na stránky, které se nacházejí v doméně Velké Británie (uk).

Záměrem autorů je kombinovat automatizované zpracování spolu s tradiční klasifikací tak, aby vznikl vyhledávač, který bude poskytovat intuitivní přístup k přesným informacím [WWWLIB, 1997].

Jako klasifikační schéma bylo zvoleno Deweyho desetinné třídění – verze 20. Hlavní důvody pro jeho výběr shrnují autoři do tří hlavních bodů – DDC má univerzální záběr, vícejazyčný záběr pro katalogizaci dokumentů a uživatelé knihovny jsou na systém zvyklí. V katalogu je celkem 4874 záznamů – odkazů s velmi stručnou anotací, které jsou rozděleny do deseti tříd DDC. V rámci jednotlivých tříd je ke každému záznamu přiřazena co nejpřesnější notace, která by vystihla tematiku dokumentu. Z tohoto přístupu vyplývá, že v jednotlivých třídách je možné najít spíše soubor jednotlivých odkazů, kde jsou obsaženy pouze takové notace, které se k nim vztahují.

Automatická klasifikace je řešena pomocí programu **The Taxonomy And Path Enhanced Retrieval (TAPER)**, který klasifikuje dokumenty porovnáváním jejich klíčových slov s klíčovými slovy klasifikační kategorie. Celý program má šest částí:

„pavouk" (spider) – automaticky prochází a dokumenty v prostředí WWW

indexační část – přebírá stránky od „pavouka", ukládá jejich místní kopii, přiřazuje jim formulář pro metadata a identifikační číslo v archivu

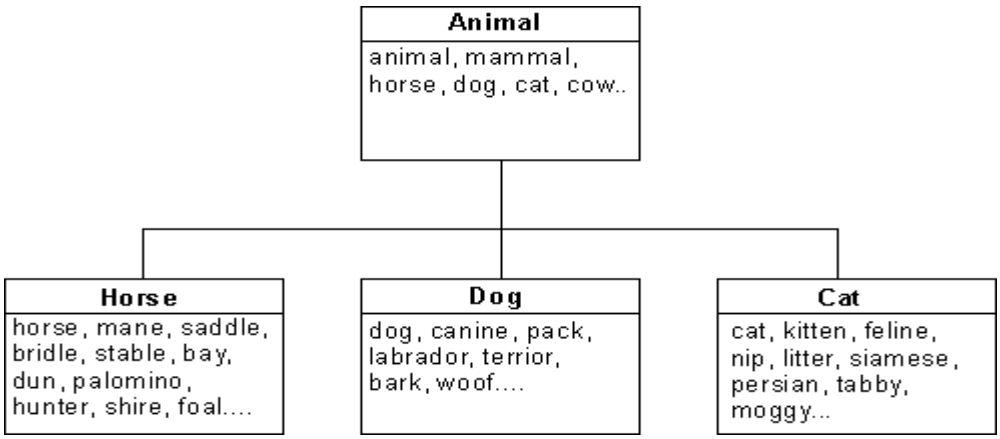
analytická část – analyzuje stránky včetně jejich odkazů na jiné stránky, a pokud odkazují na stránky domény .uk, předá jejich adresu „pavoukovi"

klasifikační část – zpracuje výsledek z indexační části a přiřadí třídy DDC

tzv. „builder" - na základě výsledku zpracování v indexační části přiřadí dokumentu do formuláře metadata a vytvoří jejich index, aby na základě dotazu bylo možné rychle přiřadit klíčová slova k identifikačnímu číslu dokumentu v archivu

vyhledávací modul – přijímá dotaz od uživatele, porovná ho s indexem, který vytvořil „builder", vyhledá identifikační čísla dokumentů, která tato slova obsahují, načte jejich lokální kopie pro případné zobrazení a generuje detailní přehled výsledků hledání.

Klasifikační část – část 4 probíhá formou porovnávání slov z originálního dokumentu se skupinou slov, která reprezentuje část klasifikační hierarchie (např. jedna třída DDC). Každá taková část má speciální seznam tzv. stop slov, která signalizují,



***Obrázek 21**: WWlib – model přiřazení dokumentu ke třídě klasifikace Zdroj: http://www.scit.wlv.ac.uk/~ex1253/classifier/*

že klíčová slova v dokumentu jsou užší než současná klasifikační úroveň. V takovém případě zpracovávaný dokument označí třídou, kde byl zpracován a pošle ke zpracování na nižší úroveň – podtřídu.

Schéma na obrázku č. 22 ukazuje příklad kategorie „animal". Pokud klíčová slova odpovídají slovům, uvedeným v hlavní skupině, mohou být odeslána ke zpracování do nižších kategorií. Zde se na základě matematického vzorce vypočítá vzájemná podobnost klíčových slov a pokud je výsledný koeficient vyšší než 0,5, je dokument považován za odpovídající a je zařazen do příslušné podkategorie. Autoři projektu zatím uvádějí přesnost tohoto zpracování přibližně 40%.

Projekt WWLib vznikl v roce 1995 a nezdá se, že za dobu své existence prodělal větší designové (a patrně ani softwarové) změny. Technická dokumentace je pravděpodobně z roku 1997, ale žádné novější informace o jeho vývoji nejsou k dispozici. Je otázkou, zda se bude tento projekt dále rozvíjet, nicméně jeho řešení, které se technicky velmi podobá modelu, používanému komerčními vyhledávači, ukazuje, že se jedná o slibnou technologii do budoucnosti.

7.4.6 INFOMINE

http://infomine.ucr.edu

Infomine je projektem University of California a slouží jako specializovaný katalog informačních zdrojů pro akademickou sféru, především pak pro přírodovědné a společensko-vědní obory. Tento katalog obsahuje přibližně 23 000 záznamů různých dokumentů a informačních zdrojů, dostupných prostřednictvím internetu – webových rozcestníků, odborných archivů nebo také volně přístupných a placených databází. Tímto se tento projekt výrazně liší od ostatních – místo jednotlivých dokumentů spíše eviduje informační zdroje, navíc úzce zaměřené na vědeckou komunitu. Katalog obsahuje devět tematických skupin, které jsou od sebe barevně odlišeny.

Jednotlivé záznamy obsahují kromě názvu a adresy delší abstrakt a také informaci o tom, zda se jedná o zdroj veřejně přístupný, či zda se jedná o zdroj placený. Navigace je možná buď prohlížením tematických skupin, prohlížením kategorií klasifikace Kongresové knihovny (LCC) nebo vyhledáváním pomocí klíčových slov.

INFOMINE není projektem, kde by byla klasifikace Kongresové knihovny omezena pouze na předmětové skupiny. Původně byly záznamy klasifikovány a doplňovány lidmi. Později byla klasifikační soustava Kongresové knihovny doplněna soustavou předmětových hesel – **Library of Congress Subject Headings (LCSH)**. Oba systémy jsou na sobě nezávislé, avšak lze je úspěšně kombinovat.

Na základě této kombinace vznikl v rámci projektu INFOMINE další výzkum, zaměřený na automatizovanou analýzu dokumentu a jeho klasifikaci. Tento výzkum, označený jako - **INFOMINE LCSH classification research project** vede dr. Steve Jones (část analýzy dokumentů do LCSH) a dr. Eibe Frank (převod LCSH do LCC) z University of Waikato, což ukazuje, že projekt INFOMINE svým významem přesáhl svou zakládající organizaci. Cílem výzkumu je automatizované zpracování klíčových slov z originálního dokumentu, jejich analýza a následné přiřazení předmětových hesel LCSH [JONES, 1994]. Dále jsou analyzována předmětová hesla LCSH a na základě jejich rozboru je dokument zařazen do hierarchického stromu pod jednu nebo několik tříd.

Zpracování zdrojových dokumentů je dosaženo pomocí software, který je schopný se učit na základě předchozích analýz. Tento software je tak „učen“ na skupinách dokumentů, vybraných z INFOMINE, MARC záznamů, které obsahují LCSH (katalog University of California) a dokumentů, nalezených vyhledávačem Google. Tato skupina dokumentů je velmi obsáhlá (jen knihovních záznamů je jeden milión) a poskytuje dostatek nových termínů.

Dosavadní výsledky ukázaly, že toto řešení může být úspěšné, nicméně zatím se potýká s několika problémy. Prvním z nich je velikost souboru předmětových hesel – předmětových hesel LCSH je několik set tisíc, což znamená obrovské nároky na softwarovou analýzu. Zároveň se objevuje veliká roztříštěnost v užití předmětových hesel. Autoři uvádějí, že při analýze v největší pokusné skupině – v dokumentech nalezených vyhledávačem Google, se 87% předmětových hesel vyskytlo pouze jednou [JONES, 1994]. Z toho vyplývá, že tato testovací skupina byla malá, a proto autoři pro další fázi projektu plánují použít jako testovací skupinu katalog MELVYL University of California, který obsahuje 24 miliónů záznamů.

Počet přesně přiřazených klasifikací pro dokumenty je zatím velmi nízký. Ukázalo se ale, že řada hesel LCSH, která byla přiřazena automaticky, je velmi blízko heslům, která dokument popisovala. Zajímavé také je, že automatizovaný systém spíše přiřadí dokumentu přesnější hesla na rozdíl od lidí, kteří volí hesla spíše obecnější (například hesla v INFOMINE) [JONES, 1994].

Ve druhé části projektu – přiřazení předmětových hesel LCSH třídám LCC autoři vyvinuli aplikaci *„LCSHtoLCC“* v programovacím jazyce Java, která je také schopná se učit [FRANK, 1994]. Tento program se učí přiřazovat třídy LCC k předmětovým heslům LCSH na základě testovací skupiny záznamů dokumentů, kde jsou oba údaje uvedeny. Touto testovací skupinou je momentálně katalog SCOTTY University of California, který obsahuje přibližně 800 000 záznamů.

Zatím byl program LCSHtoLCC vyzkoušen pro klasifikaci 50 000 záznamů MARC, při které **přesně zařadil 58% záznamů**. Na základě těchto výsledků autoři vyvíjejí i novou metodologii pro hodnocení přiřazených klasifikací, které jsou velmi blízko, ale neodpovídají přesnému zatřídění. Takto byla vyhodnocena další část výsledků - 4% jako příliš úzce definované a 3% naopak velmi obecné [FRANK, 1994].

V budoucnu by mohlo být automatizované přiřazování klasifikací dokumentům jednodušší – v plánu je vytvoření zjednodušené hierarchie pro prohlížení, která bude vycházet z LCC, užití novější verze LCC pro podrobnější rozřazení dokumentů do hierarchie a konečně přiřazení třídy LCC dokumentu přímo, bez předchozího zpracování do hesel LCSH [FRANK, 1994].

Tento výzkum je dalším dokladem toho, že automatizované procesy mohou klasické bibliografické třídění upravit takovým způsobem, že je možné minimalizovat lidskou práci při procesu zpracování. Tuto tezi je ale nutné v budoucnost potvrdit zpřesněním výsledků specializovaných programů.

7.5 PROJEKTY AUTOMATIZOVANÉ KATEGORIZACE

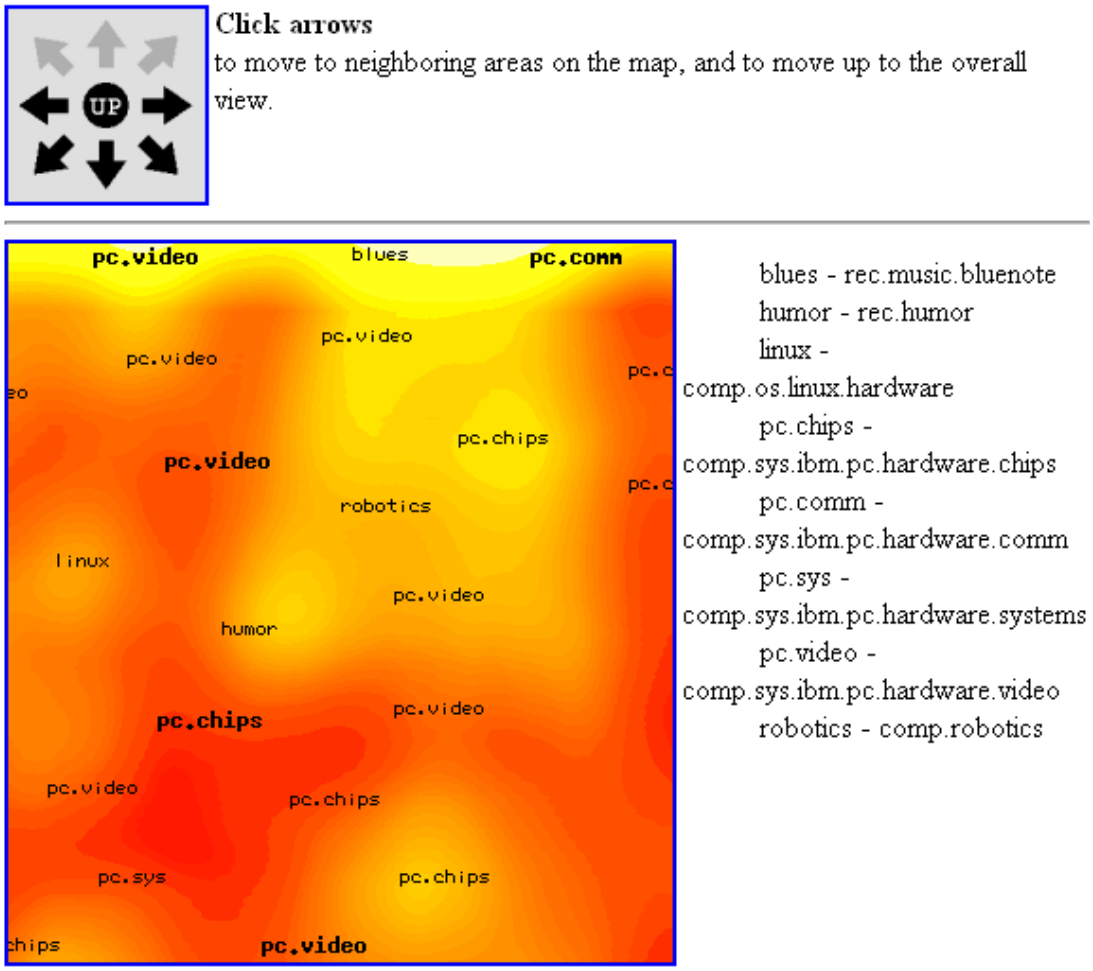
Automatizovaná kategorizace je využívána především pro komerční dokumentová a archivační řešení různých firem. U těchto systémů je využíváno často více metod kategorizace společně s dalšími metodami třídění jako jsou řízené slovníky nebo speciální oborové taxonomie, které pomáhají zpřesnit zařazení a opětovné vyhledání dokumentů.

Projekty automatizované kategorizace, které jsou dostupné na internetu, mají především výzkumný charakter. Jedná se zpravidla o experimentální projekty s omezenou životností. Některé projekty zanikají nebo jsou konzervovány bez dalšího vývoje, jiné se transformují do komerčních produktů (např. WebGlimpse[44]). Přehled některých projektů automatizované kategorizace je uveden v projektu Aristotle [MCKIERNAN, 1999B], nicméně tento přehled je již značně zastaralý.

7.5.1 WEBSOM

http://websom.hut.fi/websom/

[44] http://www.webglimpse.net



***Obrázek 22**: Rozhraní projektu WEBSOM* *Zdroj: http://websom.hut.fi/websom*

WEBSOM je projekt **Neural Networks Research Centre**, Helsinki university of Technology. Záměrem pro tento projekt byla organizace webových stránek s různou tematikou do smysluplných kategorií pomocí tzv. **SOM (Self-organizing maps)** algoritmu, který vyvinul **Teuvo Kohonen,** vedoucí tohoto projektu. Algoritmus je založen na technologii neuronových sítí a zanáší textové dokumenty do třídimenzionální mapy, která poskytuje grafický přehled o příbuznosti dokumentů mezi sebou.

Protože technologie SOM bývá založena na mapě o třech dimenzích, vede omezení projektu na dvě dimenze ke ztrátě některých informací a následně k nižší přesnosti kategorizace stránek. Toto omezení však bylo akceptováno za cenu možnosti vizualizovat tyto předměty prostřednictvím softwaru SOM_PAK do hierarchie, dostupné na webových stránkách.

Projekt zpřístupňuje dvě vizuální mapy – vizuální mapu témat 80 diskuzních skupin tzv. Usenetu v angličtině a mapu finských bulletinů ve finštině. Autoři zároveň vytvořili SOM mapu, obsahující přibližně 7 miliónů patentových abstraktů, tato mapa však není z důvodu autorských práv zpřístupněna veřejnosti.

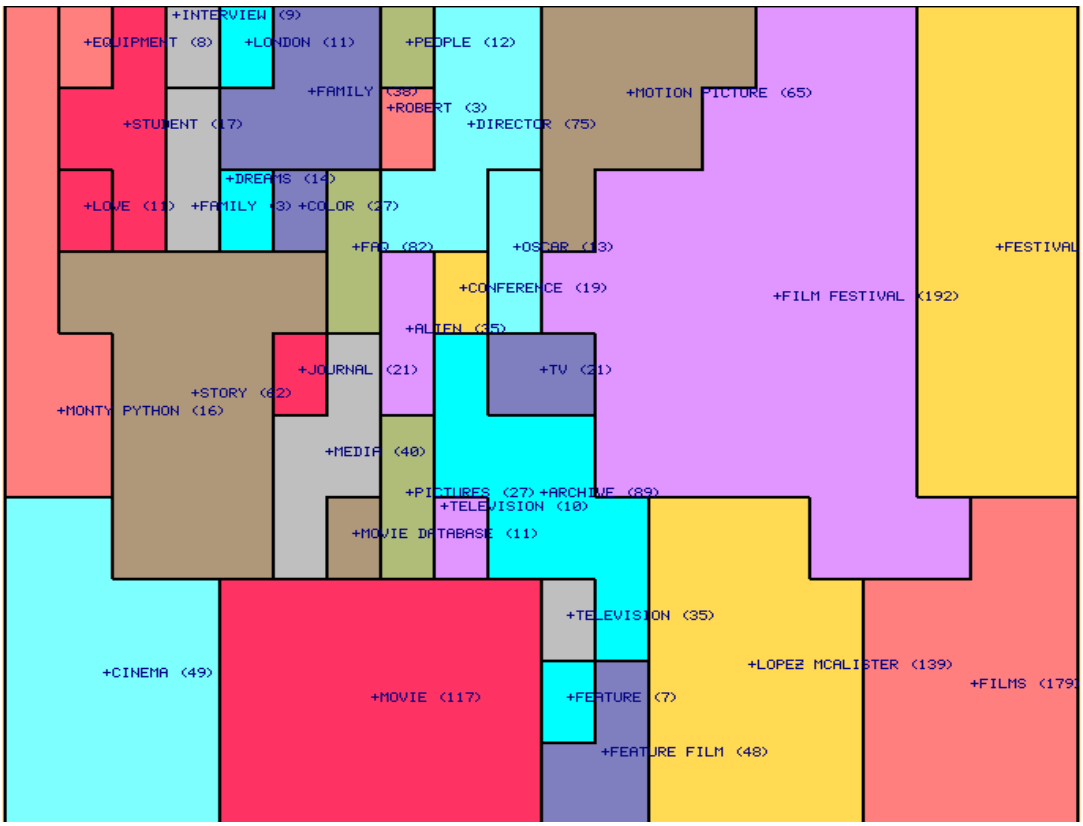
Mapa osmdesáti Usenet skupin obsahuje přes milion oklasifikovaných textových souborů (stránek, textových dokumentů), rozdělených do skupin podle vzájemné příbuznosti. Navigace v těchto souborech je řešena formou tzv. klikací mapy, kde uživatel klikne na obrazovou mapu, která funguje jako odkaz. Mapy jsou rozděleny podle úrovně podrobností do tří částí – první mapa přináší přehled základních skupin, mapa třetí nejpodrobněji definuje poslední úroveň skupin. Poslední úrovní zobrazení jsou konkrétní texty, rozdělené podle diskuzních skupin, do nichž patří. Navigaci v této úrovni je možné upřesnit pomocí šipek, které umožňují plynule přecházet mezi nejbližšími body (skupinami) neuronové sítě (tzv. nody) [WEBSOM, 1999].

Projekt vznikl již v roce 1996 a poslední příspěvky o něm jsou z roku 2000. Není proto jasné, zda vývoj projektu pokračuje i nadále. Jak je z projektu patrné, neuronové sítě mají velikou budoucnost pro automatickou klasifikaci a mohou se stát jedním z hlavních způsobů organizace informací nejen pro dokumenty dostupné prostřednictvím internetu.

7.5.2 Entertainment map (ET-Map)

http://ai.eller.arizona.edu/research/dl/etmapdemo.htm

ET-Map je prototyp kategorizačního systému, vyvinutého Artifical Intelligence Lab of the University of Arizona. Projekt byl financován na základě grantu National Science Foundation (NSF) „Intelligent Internet Categorization and Search" na léta 1995-1998 a grantu NSF/ARPA/NASA *„Illinois Digital Library Initiative project"* na léta 1994-1998 [ET-MAP, 1998].



***Obrázek 23**: Rozhraní projektu ET-map*

Zdroj: http://ai3.eller.arizona.edu/ent/entertain1/

Cílem ET-Map bylo demonstrovat škálovatelný automatizovaný přístup ke kategorizaci a vyhledávání webových stránek. Tento projekt měl poskytovat granularitu údajů vyhledávačů stejně jako zachovat vícevrstvou hierarchii předmětových katalogů. Pro praktickou demonstraci bylo vybráno přes 110 000 webových stránek, orientovaných na zábavu, ze kterých byla vytvořena několikavrstvá SOM mapa, rozdělená na různé tematické oblasti. Mapa graficky zobrazuje i poměrné zastoupení jednotlivých oblastí na celkovém objemu všech kategorizovaných stránek. Klasifikované stránky jsou kategorizovány do skupin po maximálně 100 odkazech a skupiny jsou rozděleny do dvou až tří grafických map. Podobně jako u projektu WEBSOM jsou kategorie podle příbuznosti témat graficky znázorněny vedle sebe.

Tento projekt má již spíše historický význam neboť značná část uváděných odkazů již není funkční. Přesto potvrzuje domněnku, že je možné v systému pro pořádání informací zachovat jak přístup prohlížení, tak vyhledávání.

7.5.3 Open Architecture Server for Information Search and Delivery (OASIS)

Projekt OASIS byl v rámci výzkumného programu INCO Copernicus za spolupráce vědců z Ruska, Ukrajiny, Německa a Irska zahájen s cílem vyvinout inteligentní vyhledávací službu, založenou na metodách neuronových sítí. Hlavním úkolem tohoto projektu bylo vyvinout nový systém k hledání informací na webových stránkách, který by podporoval nové typy dotazů. Systém OASIS podporuje dotazy ve formě příkladů stránek a zpětné vazby od uživatele. Na základě této zpětné vazby systém vytvoří z relevantních stránek seznam klíčových slov, kde bude uvedena i jejich relativní váha, která má odrážet význam daného slova v dotazu. Tato technika slouží zároveň pro zpřesnění vnitřních vyhledávacích algoritmů systému.

Systém OASIS je definován jako obecné řešení, které je možné přizpůsobit současným technologiím – tj. systém je možné kombinovat se stávajícími vyhledávacími službami.

Servery, využívající technologii projektu OASIS, jsou tak součástí řady komerčních systémů pro automatizovanou kategorizaci – za všechny je možné uvést společnost Insuma [INSUMA, 2002][45].

7.5.4 People Helping One Another Know Stuff (PHOAKS)

http://www.cs.indiana.edu/~sithakur/l542_p3/#

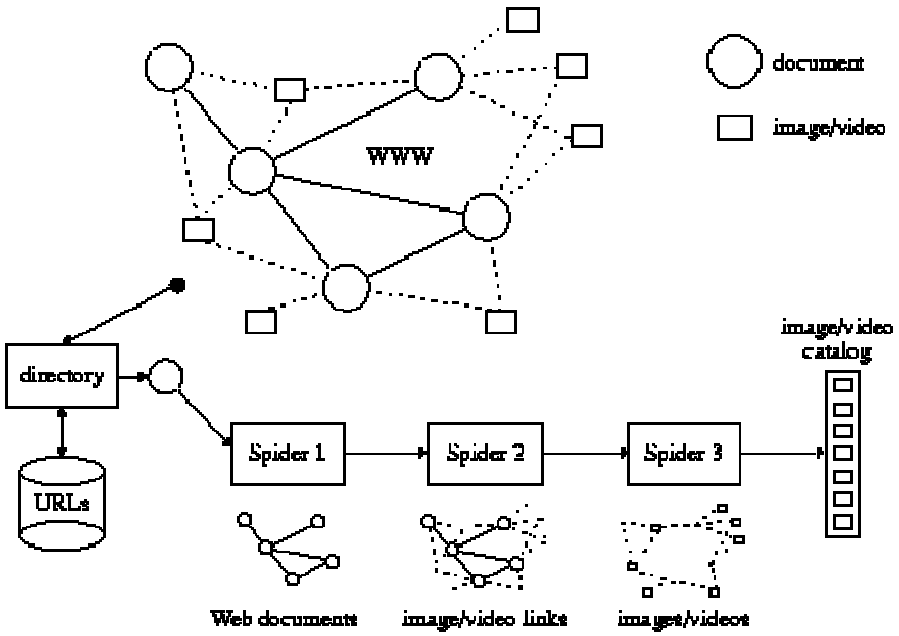
Systém PHOAKS se zaměřuje na automatickou klasifikaci stránek, odkazovaných v příspěvcích uživatelských skupin sítě USENET. Příspěvky do skupin Usenetu často obsahují zajímavé odkazy na specializované stránky a tyto odkazy jsou automaticky vyhodnoceny a tematicky zařazeny do příslušné kategorie. Projekt, který vznikl v roce 1996, dnes již není plně funkční a pravděpodobně byla realizována pouze jeho pilotní fáze.

7.5.5 Web Seek

WebSEEk je katalog a vyhledávač pro obrazové a video soubory na webových stránkách. Tento systém sbírá prostřednictvím softwarových „agentů“ obrázky a video soubory, které následně automaticky analyzuje, indexuje a přiřadí do odpovídající kategorie. Proces vyhledávání a katalogizace probíhá zpracováním textových souborů, odkazů na obrázky a obrázků samotných. Proces zpracování ukazuje obrázek č. 6.10.

Tento postup se snaží zkombinovat dostupné textové informace (popisky, titulky) s obrázky, a na základě sémantických analýz přiřadit související informace a obrázky k sobě (např. popis a odkaz na obrázek s obrázkem samotným).Poté, co systém stáhne nalezené obrázky do vlastní databáze, následuje extrakce termínů a výrazů z názvů a popisků obrázků a tyto výrazy jsou poté srovnány s řízeným slovníkem. Ten vyřadí slova, nepoužitelná jako deskriptory (např. picture, image) a porovnáním dalších termínů se slovníkem přiřadí obrázek k tematické skupině.

Tematické skupiny jsou zařazeny do hierarchického stromu, který určuje vztahy nadřízenosti a podřízenosti mezi celkem 941 skupinami.



Obrázek 24: *Web Seek – proces vyhledávání a zpracování obrázků a video souborů*
Zdroj: http://persia.cs.columbia.edu:8008/paper/

[45]

Celková přesnost systému se blíží 92%, i když přesnost katalogizace může být v závislosti na tematice dokumentu nižší [SMITH, 1996].

K listopadu 2004 je v projektu zkatalogizováno 665 115 obrázků a video souborů. V budoucnosti chtějí autoři zlepšit možnosti zpracování obrázků na základě jejich tvaru, textury a prostorové úpravy. Zároveň je plánováno rozlišování obličejů na obrázcích a také extrakce textu, který je na obrázku či videu zachycen.

7.5.6 PARser for Content Extraction and Layout Structure (PARCELs)

http://parcels.sourceforge.net

Klasifikace webových stránek jako celku je velmi obtížná. Jako předzpracování se proto užívá klasifikace tematických bloků stránky, které je možné lépe analyzovat. Systém PARCELs takto stránky rozděluje na bloky, které následně zpracovává automatizovanými algoritmy. Systém analyzuje stránky z hlediska strukturních a lexikálních informací. Každý typ informace může dodat smysl stránce a jejich odděleným zpracováním je tak možné zpřesnit klasifikaci stránky.

Pro testovací provoz systému autoři zvolili zpravodajské (News) servery z důvodu jejich komplexního designu, různých úrovní podrobnosti, a také pro srovnání výkonu s podobnými projekty, které klasifikovaly podobné stránky [LEE, 2004]. Pro strukturní zpracování byly analyzovány tři typy struktury kódu HTML (lineární, tabulární a s formátovací vrstvou – CSS), které mají specifické nároky na automatizované zpracování. U textových prvků se analýza soustředila na prvky vyšší úrovně (high-level features) jako tzv. parts-of-speeach (POS), typy odkazů, počty obrázků a rozsah textu, a na prvky nižší úrovně (low-level features), jako slova samotná a jejich statistickou četnost.

Celý systém byl naplněn testovacími daty, u kterých byla bez odladění chybovost 70,6%. Na základě experimentů autoři přistoupili k analýze o třech fázích, kdy se postupně bloky zpracovávaly od nejméně významných – reklam, dekorací stránek (cca 66% textu) přes obsahově významnější bloky – příbuzná témata, navigace stránek (cca 24%), po nejvýznamnější části jako titulky a hlavní obsah, které obsahovaly přibližně 10% celkového obsahu. Po vyladění těchto úrovní zpracování vykazuje systém chybovost přibližně 17,6%, což vzhledem k prvním výsledkům představuje významný posun. Přesto se tyto výsledky nijak významně neodlišují od podobných automatizovaných systémů. V budoucnu plánují autoři výrazně zpřesnit klasifikační možnosti systému a klasifikovat prvky do speciálních polí, jako například autor, datum nebo místo. Systém PARCELs je na výše uvedené adrese volně dostupný pod licencí open-source.

7.6 PERSPEKTIVY AUTOMATIZOVANÉ KLASIFIKACE A KATEGORIZACE

Automatizovaná klasifikace a především kategorizace jsou v současné době nejpokročilejšími metodami pořádání informací, dostupných na internetu a prostřednictvím internetu, a zdá se, že tyto technologie budou jedním z nejprogresivnějších přístupů i v budoucnu. Automatizovaná kategorizace nemůže nikdy zcela nahradit ostatní formy pořádání informací. I když je dosahována přesnost pořádání přes 60-75%, což už činí tuto technologii zajímavou i z hlediska komerčního využití, (ve studii společnosti Microsoft [Lubbes, 2003]) bylo dosaženo přesnosti 80% při zařazení dokumentu do kategorie první úrovně). Tato metoda se bude i nadále potýkat s nedostatkem pochopení významu a vyhodnocení hlavního tématu dokumentu. Z tohoto důvodu většina komerčních produktů pro automatizovanou kategorizaci tuto metodu kombinuje s dalšími přístupy (tezaury, řízené slovníky, taxonomie), které mohou pomoci blíže určit téma dokumentu a jeho vztah k tématům ostatním.

Přesnost těchto systémů bude i nadále značně závislá na jejich nastavení a vyladění, především na implementaci a cvičení na tzv. tréningové bázi dokumentů, která musí poměrově reprezentovat témata, která budou kategorizována v budoucnu [LUBBES, 2003]. Tuto podmínku je obtížné naplnit, a tak i přesnost kategorizace závisí na kvalitě trénování systému.

Výzkum a rozvoj těchto systémů je velmi nákladný a neobejde se bez finančního zajištění v rámci projektů a grantů. Zároveň je tato technologie pro zmíněnou přesnost velmi zajímavá i z hlediska komerčního využití, což dokládá množství softwarových systémů. Ve srovnání s cenou lidské práce jsou náklady na automatizovanou kategorizaci a klasifikaci nesrovnatelně nižší, a proto je přesnost těchto systémů s ohledem na cenu přijatelná.

8 NEKONVENČNÍ A INOVAČNÍ PŘÍSTUPY

Nic není objeveno a zároveň hned dokonalé. Cicero Marcus Tullius

Implementace univerzálních a oborových třídění, nebo automatizovaná kategorizace a klasifikace představují přístup, založený především na analýze obsahu, který je ve formě textu. Kromě tohoto přístupu se objevily i jiné přístupy, zaměřené více na povahu nebo účel informace. Klasifikace není vázána na informační obsah dokumentu (především z hlediska textové informace), ale spíše je zaměřena na rozdělení elektronických dokumentů podle typů a povahy dokumentu.

Formy těchto klasifikací jsou různé – od systému pro uspořádání e-mailů přes analýzu stránek podle textového či grafického formátu po rozdělení stránek podle geografického výskytu. Tyto projekty často využívají metod umělé inteligence – jmenovitě kategorizaci nebo vizualizaci, nicméně všechny tyto prostředky jsou pouze nástrojem pro zpracování. Tuto skutečnost komentuje Wallace Koehler, autor jednoho z projektů, slovy: „Automatizace není konečným produktem; je to krok k rozšíření stávajících pomůcek, není zde proto, aby nahradila katalogizaci webu a proces vyhledávání dokumentů; věřím, že spíše než knihovník a počítačový specialista, kteří zakládají webové katalogy, bude o vhodnosti metod rozhodovat informační vědec, který určí kdy je vhodné „vyhodit robota“, kdy je třeba užít tradiční přístupy a kdy mají být oba přístupy zkombinovány.“ [Kohler, 1998].

Většina projektů je experimentálních a jejich hlavním cílem je nabídnout alternativu stávajícímu přístupu klasifikace, a umožnit tak zpracování několikanásobně většího množství dokumentů, než v současné době. Řada z nich je také ve formě doplňků a rozšíření stávajících vyhledávačů a snaží se zpřesnit výsledky hledání podle svého záměru (např. NEXAS – hledání lidí nebo GeoViser – zobrazení geografické lokace nalezených výsledků).

8.1.1 Klasifikace stránek na základě jejich metriky a charakteristiky URL (W. Koehler)

Autor projektu – Wallace Koehler zahájil tento projekt, aby prokázal nutnost vytvoření nového typu klasifikace pro zdroje na internetu, pro něž podle jeho názoru již tradiční třídění nestačí. Jako jeden z argumentů uvádí nevhodnost klasifikace pro nové, specifické typy dokumentů, které se v tomto prostředí vyskytují, a které jsou obtížně zpracovatelné po obsahové stránce (např. seznamy a rozcestníky).

Tento projekt se proto soustředil na analýzy bibliografických informací, které mohou být získány přímo z URL zdrojů, a na kvantitativní indikátory souborů stránek. Na základě této analýzy autor navrhl šest skupin, do nichž lze zařadit stránku podle typu objektů, které na ní převládají:

Average (bez dominantního objektu)

Wordsworth (dominantním objektem je text)

Coffee-Table (dominantní je grafika)

Mogul (dominantní jsou multimédia)

Retriever (dominantní služba ftp/gopher)

Post Office (dominantní e-mail) [KOHLER, 1998].

Toto rozdělení poměrně výstižně definuje různé typy dokumentů v nejpoužívanějších službách sítě internet, i když se zde projevuje i doba vzniku této studie – rok 1997 a zastoupení služby gopher, která je dnes již jen historickým předchůdcem dnešní služby www.

Celý experiment s mapováním typů objektů na stránkách probíhal ve dvou fázích – v roce 1996 a 1998. Porovnáním obou let bylo zjištěno, že dominantním objektem na více než polovině stránek je text, ale i to, že 30% stránek, zkoumaných v předchozím roce již neexistovalo. Na základě těchto zjištění je zřejmé, že internet nemůže být vnímán jako stabilní prostředí, v němž je možné klasifikovat pouze nově přidané dokumenty, ale jedná se o dynamické prostředí, které je neustále nutné zpracovávat jako celek a přehodnocovat tak i aktualitu a strukturu klasifikace. Wallace Koehler z tohoto důvodu předkládá model zpracování dokumentů na internetu, skládajících se z následujících třech kroků:

popis na základě URL (zjištění služby, domény a její úrovně apod.)

kvantifikační analýza stránek („fyzické parametry“) (datová velikost, typ objektů, hloubka úrovní částí dokumentu/stránek, hustota hypertextových odkazů)

analýza změny a životnosti stránek (stabilita dokumentu).

Tento model je ve své podstatě přenesením stejných principů, jaké platí při popisu a klasifikaci tištěných dokumentů; výjimkou je třetí krok, který při zpracování tištěných dokumentů neexistuje. Z důvodu stále rostoucího množství nových dokumentů a možnosti zpracování alespoň jejich části se však tento krok jeví jako nezbytný.

Experiment Wallace Koehlera je nutné vnímat jako určité „předzpracování“ množiny dokumentů pro klasifikaci. Uvedené metody tak mohou být účelné pouze ve spolupráci s jinou metodou klasifikace.

8.1.2 Klasifikace e-mailových zpráv

Jako jeden z mála se tento projekt věnuje pořádání zpráv elektronické pošty – e-mailu. U této služby je pořádání zpráv velmi specifické – systém pro organizaci si totiž každý uživatel vytváří sám. Zároveň jsou e-mailové zprávy problematické i pro automatizovanou klasifikaci nebo kategorizaci – obsahují většinou velmi málo textu, a tak je obtížné zjistit přesný obsah a kontext zprávy.
Maureen Mackenzie [2000] ve své studii zkoumala přístupy uživatelů k pořádání e-mailových zpráv. Pomocí dotazníku a osobních rozhovorů s patnácti manažery podniků se snažila zjistit, podle jakých kritérií pořádají svou elektronickou poštu. Systémy pro organizaci e-mailů vycházely ze dvou modelů – subjektivní kategorizace zpráv buď podle témat, nebo podle činnosti, kterou bylo nutné na základě této zprávy vykonat. Uživatelé v první kategorii vytvářeli složky a podsložky, odrážející jejich aktuální projekty a funkce, které ještě mohli rozdělit podle stáří zpráv na měsíce nebo jiné časové období. Druhá kategorie uživatelů třídila své zprávy rámcově na tři kategorie:

- co musím vykonat
- co dělají ostatní a mohlo by mě to zajímat
- co zatím nepotřebuji, ale mohlo by mě to zajímat v budoucnu.

Tento výzkum je výjimečný i pro povahu a význam tohoto typu dokumentu. Jak uvádí ve svém výzkumu španělští odborníci: *„Klasifikace e-mailu se objevila jako anomálie z důvodu velmi osobního vztahu mezi indexátorem a rešeršérem, ale také z důvodu malé potřeby vzít v úvahu [informační] potřeby ostatních. Poznávací proces, který je podkladem a „hlavní aktivitou v dokumentačním procesu“ klasifikace, musí zůstávat jediným cílem budoucího výzkumu.“*[MARCO, 1993]. Oblast klasifikace e-mailových zpráv tak patrně i v budoucnu zůstane omezena na zpracování a archivaci zpráv v podnikových systémech. E-mailové zprávy, které mohou mít význam pro více lidí jsou zasílány v rámci služby USENET, která už má webové rozhraní, a tak zde publikované zprávy spíše spadají do klasifikace webových stránek (např. projekt PHOAKS).

8.1.3 Named Entity eXtraction and Association Search (NEXAS)

V prostředí internetu je často problémem nalézt osoby, které jsou považované v reálném světě za autority – ať již ve vědecké komunitě či jako známou osobnost ve společnosti. Vyhledávač často nalezne množství osob, mezi nimiž je obtížné nalézt člověka, aktivního v oblasti, která nás zajímá. Výzkumný tým z Nippon Telegraph and Telephone Corporation vyvinul systém Named Entity eXtraction and Association Search (NEXAS),který jako extenze běžného vyhledávače slouží k automatizovanému zpracování a přiřazení výsledků hledání k autoritativním záznamům osob v databázi systému.

Systém NEXAS používá k vyhledávání jmen komerční vyhledávače[46], na nalezených stránkách identifikuje tvary jmen a vypočítá relevanci tohoto výskytu v porovnání s ostatními stránkami. V dalším kroku systém vyexportuje slovníkový tvar jména osob společně se seznamem stránek, na kterých se toto jméno vyskytuje. Tento seznam je řazen stejně jako u vyhledávačů – tj. podle relevance stránek.

V pilotním projektu autoři analyzovali přibližně 52 miliónů webových stránek na serverech s japonskou doménou .jp.

[46] V projektu byl použit Google http://www.google.com.

Následně pomocí volně šiřitelného morfologického analyzátoru MeCab[47] na stránkách identifikovali jména osob ve tvaru příjmení – osobní jméno[48]. Následně byl z těchto stránek a ze slovníků osobností sestaven seznam, obsahující přibližně šedesát tisíc osobních jmen i příjmení. Na jedné stránce se vyskytovalo průměrně 2,4 jména a jméno jedné osoby se v průměru objevilo devatenáctkrát. Pomocí slovníku autoři sestavili pořadí nejčastěji užívaných jmen ve výzkumném vzorku a zjistili, že pořadí výskytu jmen nemusí odpovídat vždy zájmu veřejnosti o tyto osobnosti, ale může být způsobeno automaticky generovanými odkazy (např. hráči basebalového týmu jsou odkazováni ze sportovních stránek) nebo pomocí tzv. search engine (index) spammingu – techniky, která má oklamat hodnotící systém vyhledávačů a nerelevantní stránky v hodnocení co nejvíce zviditelnit mezi výsledky.

NEXAS ukázal, že jeho využívání nemusí být omezeno na autoritativní kontrolu a analýzu jmen osob, ale lze jej použít i k dalším účelům. Autoři uvádějí, že pokud by byl jejich systém propojen s knihovními katalogy, mohl by automaticky nabídnout seznam knih, který hledaný autor napsal nebo nalézt knihu na základě ISBN. Jiným využitím by mohla být identifikace společnosti podle kódu, pod kterým se její akcie prodávají na burzách [HARADA, 2004]. Systém NEXAS je jedním z prvních experimentů se zaváděním kontroly autorit v prostředí webových stránek. Význam tohoto přístupu se bude neustále zvyšovat s rostoucím počtem nově publikovaných stránek a je vysoce pravděpodobné, že podobné experimenty budou ve spojení s dalšími metodami pořádání informací (řízené slovníky, tezaury, automatizovaná klasifikace) pokračovat i nadále.

8.1.4 Knowledge Class

http://www.uky.edu/~xlin/kclass.html

Systém Knowledge Class je dalším z projektů inteligentního rozhraní k vyhledávačům. Celý projekt vychází z předpokladu, že každý uživatel potřebuje vlastní systém pro organizaci stránek. Systém je koncipován jako vrstva mezi vyhledávačem a uživatelem, který pomůže zpřesnit dotaz na základě předchozích hledání, uložených v tomto rozhraní. Při použití systému Knowledge Class si uživatel zvolí tematickou oblast, ve které si z tezauru vybere podrobnější téma. Předměty jsou organizovány do hierarchických struktur, jež mají dvě úrovně. Pro další hledání si vybere z nabídky vyhledávač, který chce pro vyhledání použít, a zvolí podrobnější téma z tezauru. Poté se provede analýza dotazu a v rámu se objeví výsledky hledání zvoleného vyhledávače.

Knowledge Class je zajímavým pokusem o kombinaci tezauru s vyhledávačem. Pro zjištění, jak je toto spojení vhodné by bylo zapotřebí definovat hlubší úrovně tezauru, což se zatím v tomto projektu nestalo.

8.1.5 MeURLin

Od počátku existence modelu adresování na internetu jsou pokusy využít adresy URL jako zdroje pro analýzu obsahu stránek. Tyto experimenty se pokouší uvést do praxe systém, který by nahradil zatím nefungující protokol identifikátoru URI.

Záměrem projektu MeURLin bylo zjistit, zda je URL vhodný jako údaj pro klasifikaci webových stránek a odpovědět mimo jiné na otázky, jak přesný bude systém založený pouze na analýze URL a zda lze jeho přesnost zlepšit analýzou dalších údajů (odkazy a jejich texty na jiných stránkách).
V testovací fázi byly za pomoci matematických a lingvistických metod rozděleny URL adresy 4 167 stránek z archivačního projektu WebKB na segmenty podle protokolu URI: schéma :// host / cesta / dokument . extenze. Tím byly izolovány části jako doména, adresáře a název dokumentu. Následně byly tyto části uloženy do databáze a oklasifikovány podle samoučícího se automatizovaného systému SVMlight.

Výsledky testů ukázaly, že samotná analýza URL je přibližně třikrát efektivnější než při použití doplňkových informací – odkazů na stránky a jejich popisných textů. Tato metoda přináší užitečné informace a zároveň je velmi rychlá. Autoři hodnotí výsledky testů pozitivně a plánují tento přístup použít v dalším experimentu, který bude zaměřen na klasifikaci stránek, obsažených v katalogu Open Directory Project [KAN, 2004].

---

47 http://packages.debian.org/unstable/misc/mecab.html

48 Toto pořadí je běžné v psané podobě japonštiny.

8.1.6 GeoViser

Projekt GeoViser je zajímavou kombinací vyhledávače s vizualizačním rozhraním, která je schopna výsledky hledání rozřadit podle jejich geografického umístění v USA. Informace o umístění zdroje je často velmi užitečnou součástí odpovědi, a tak autoři projektu zkombinovali vyhledávač Infoseek s vlastním systémem pro zpracování a vizualizaci.

Systém má několik etap zpracování a vyhodnocení dotazu. V první části je pomocí uživatelského rozhraní systému zaslán vyhledávači dotaz, který vrátí seznam výsledků. Výsledky jsou v dalších krocích vyhodnoceny podle relevance, jsou vyřazeny duplicity, a odkazy jsou sdruženy do skupin podle svých klíčových slov a stránky, ze které pocházejí.

V dalším zpracování jsou identifikovány IP adresy stránek, na kterých byly nalezeny výsledky, a následně jsou vyhledány další informace o těchto stránkách – především stát unie a město, kde je adresa registrována. Pro tyto informace je použita služba Národního statistického úřadu USA *Topologically Integrated Geographic Encoding and Referencing (TIGER)*[49]. Tato služba zobrazuje detailní mapy Spojených států a umožňuje přímé odkazování z jiných webových stránek nebo systémů. Na základě informací o umístění serverů jsou pak na mapách, zpřístupněných službou TIGER, zobrazeny výsledky dotazů podle polohy serveru, na kterém se nacházejí. GeoViser je systémem ve vývoji, a jeho autoři předpokládají u dalších verzí vylepšení vizualizačního mechanismu a spolupráci s metadaty na stránkách [GOVINDARAJAN, 1998].

8.2 TEMATICKÉ MAPY

8.2.1 Koncept tematických map

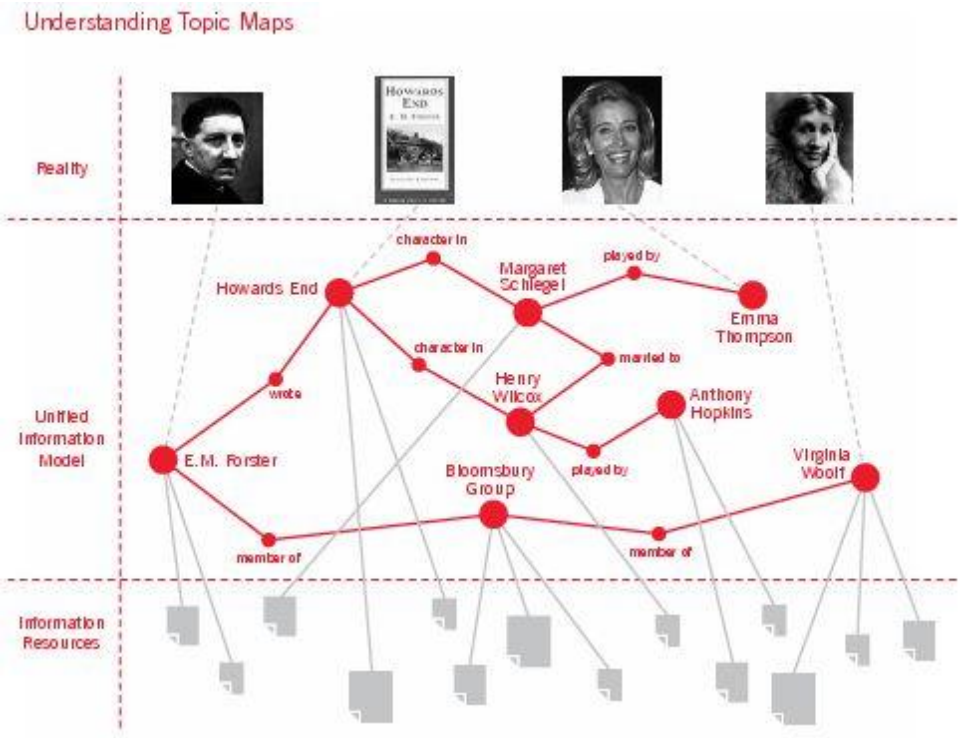
Jednou z nových metod pro pořádání informací jsou tematické mapy (topic maps). Spojením značkovacích jazyků s vizualizačními možnostmi nabízí optické pořádání informací, které je pro uživatele přehledné a tak i lépe pochopitelné.

Tematické mapy – XTM (XML Topic Maps) jsou od roku 2000 ISO standardem (ISO 13250), který definuje vytváření a zobrazení navigačních struktur pomocí jazyka XML. Svou koncepcí vychází z jednoduchých rejstříků které jsou seznamy různých prvků (např. jmen autorů, předmětů, projektů) společně s odkazem na stranu, kde se heslo vyskytuje. Na rozdíl od rejstříků rozlišují tyto mapy tři druhy prvků, které společně tvoří význam:

Topic (téma)
Association (asociace)
Occurrence (výskyt)

Téma je označení pro jakýkoli koncept nebo nehmotnou entitu, která zastupuje různé předměty, osoby nebo jiné hmotné objekty v reálném světě. Tématem může být například „Jan Werich“, které zastupuje konkrétní osobu, která žila v letech 1905 – 1980. Témata mohou být sdružována podle svého typu na základě tzv. tematických typů. Jan Werich tak může být tematickým typem „spisovatel“, pohádka tematickým typem „literární útvar“, a West Pocket Revue tematickým typem „divadelní hra“.
Téma je spojeno s konkrétním dokumentem, který o něm pojednává. Například téma „Jan Werich“ je spojeno s gramofonovou deskou Pěst na oko, kde Jan Werich hraje. Tato gramofonová deska je výskytem (occurrence) tématu „Jan Werich“. Každé téma je zpravidla spojeno s více výskyty, v podobě různých knih, článků, audiovizuálního záznamu apod.

49 http://tiger.census.gov



Obrázek 25: *Model reprezentace tematických map* *Zdroj: http://www.ontopia.net/solutions/brochures/Only_Connect.pdf*

Asociace popisuje vztahy mezi tématy. Tematické mapy rozdělují asociace do dvou vrstev na asociace témat a na jejich výskyty. Tak je možné odlišit významové vazby od konkrétních výskytů. Stejně jako témata se asociace dělí podle typu. Typy asociací tak mohou být „napsáno“ nebo „narozen v“.

Problémem řady systémů pro pořádání informací je nedostatečné definování vztahů mezi tříděnými entitami. Například tezaury určují význam mezi dvěma prvky vztahem nadřízený – podřízený, odkazem na jiný termín nebo asociací termínu. Významové vazby tak předpokládají předchozí znalosti problematiky a inteligenci uživatele, nezbytnou pro odvození přesného význam těchto vazeb (např. chirurgie je lékařskou disciplínou, která může úzce spolupracovat s interní medicínou). Jak píše Steve Pepper „znalost se od informace značně liší: je rozdíl mezi tím znát věc a mít o ní informaci“ [Pepper, 2002].

Tento problém tematické mapy odstraňují přesným definováním asociační vazby mezi dvěma prvky. Pro přesné definování vztahu je možné určit tzv. asociační role, které přesně popisuje vztah dvou prvků mezi sebou. Například u asociace „ovlivněn“ mezi dvěma skladateli (Verdi, Puccini) je důležité vědět kdo ovlivnil koho [Pepper, 2002]. Tímto mechanismem je možné vytvořit vazbu, která přesně definuje význam, aniž by bylo potřeba tuto informaci jakkoli interpretovat nebo doplnit předchozí znalostí.

Syntaxe tematických map vychází z XML a dalších standardů (např. pro odkazy se používá hypermediální strukturovaní jazyk HyTime[50]). Celá struktura je viditelná z níže uvedeného příkladu:

```
 <topic id="authorship">
<baseName>
<baseNameString>Authorship</baseNameString>
```

50 ISO/IEC 10744:1997 Hypermedia/Time-based Structuring Language.

```
 </baseName>
</topic>

<topic id="author">
 <baseName>
<baseNameString>Author</baseNameString>
 </baseName>
</topic>

<topic id="work">
 <baseName>
<baseNameString>Work</baseNameString>
 </baseName>
</topic>

<association>
 <instanceOf>
<topicRef xlink:href="#authorship"/>
 </instanceOf>

 <member>
<roleSpec>
 <topicRef xlink:href="#author"/>
</roleSpec>
<topicRef xlink:href="#tim-bray"/>
 </member>

 <member>
<roleSpec>
 <topicRef xlink:href="#work"/>
</roleSpec>
<topicRef xlink:href="#xml-rec"/>
 </member>
</association>
```

Zdroj: Garshol, 2002

Uvedený příklad definuje téma „autorship“, „author“, „work“ a definuje vztahy mezi nimi. Zároveň odkazuje na autora (tim-bray), který je uveden jako téma (topic) celé mapy[51].

Na první pohled je viditelná podobnost s metadatovými schématy, jmenovitě s RDF (viz kapitola 4.6.1, s. 23). Při podrobnějším zkoumání ale zjistíme, že vzájemná podobnost existuje, i když je velmi malá. Obě technologie vznikly pro reprezentaci znalostí, obě definují abstraktní model a syntaxi pro výměnu dat, založenou na XML [Pepper, 2002b].

Rozdíly mezi tematickými mapami a RDF popsal ve svém článku Steve Pepper [2002b]. Mezi hlavními rozdíly mezi RDF a tematickými mapami jmenuje:

51 Tato část zdrojového textu je vynechána.

1. Jiné kořeny a perspektiva

Tematické mapy vycházejí svou koncepcí z knižních rejstříků, seznamů a tezaurů. RDF vychází z formální logiky a matematické teorie grafů. Tematické mapy jsou zamýšleny pro organizaci informací z lidské perspektivy, RDF na organizaci informací z perspektivy strojového zpracování.

2. Odlišná koncepce

Tematické mapy se zaměřují na předmět nebo téma. K tomuto předmětu/tématu doplňuje vztahy s dalšími předměty/tématy a konkrétní zdroje. RDF je zaměřeno na konkrétní zdroj, ke kterému doplňuje další informace.

3. Jiná úroveň sémantiky

V tematických mapách je mnohem složitější sémantika. Témata mají různé druhy charakteristik (např. druhy, jména, výskyty, role). Sémantika v RDF je založena na modelu „vlastnost – hodnota“, který neumožňuje zachycení složitějších vazeb mezi tématy.

4. Role a směry

Výrok v RDF má určený směr – např. „Tom namazal chleba máslem“ nemusí znamenat, že „chleba byl Tomem namazán máslem“ i když se jedná o stejnou asociaci [Pepper, 2002b]. V RDF to vede k vytváření duplicitních asociací, které zachycují oba směry, i když se jedná o asociaci jedinou. Tematické mapy tento problém nemají; není možné vložit asociaci „Tom namazal chleba máslem“ aniž by zároveň znamenala, že „chleba byl Tomem namazán máslem“ protože jde o jedinou asociaci [Pepper, 2002b].

5. Množství asociací mezi prvky

Výrok v RDF a jeho asociace jsou vždy binární a vyjadřují vztah mezi podnětem (objektem) a přísudkem (slovesem – činností) jako v přirozeném jazyce. Tematické mapy počet asociací, spojených s výrokem, neomezují; každá asociace může mít řadu rolí, které vyjadřují komplexní vztahy.

6. Záběr, kontext a vícejazyčná podpora

Tematické mapy řeší problém platnosti kontextu pomocí ohraničení rozsahu tématu. Kontext tématu může být snadno vyjádřen definováním rozsahu platnosti výroku. Například téma „vaření“ může být omezeno svým rozsahem na geografickou oblast „Mexiko“. V rozsahu platnosti výroku tak jsou témata jako „tacos“, „fajita“, „spanish rice“ (mexická jídla), ale už ne „pizza“, která se vztahuje k italské kuchyni.

Tematické mapy jsou zároveň ideální pro podporu více jazyků. Důležitou vlastností konceptu záběru tématu je označit každé téma více jmény v různých jazycích. Každý uživatel si pak může zvolit jazykovou nebo i terminologickou preferenci, se kterou bude pracovat.

8.2.2 Srovnání tematických map s ostatními metodami pořádání

Tematické mapy koncepčně vychází z tezaurů, knižních rejstříků a seznamů, a je zajímavé srovnat jejich vlastnosti s nejužívanějšími metodami pro pořádání informací. Některé vlastnosti těchto systémů přejímají (především definice významu), některé nedostatky naopak odstraňují. Při srovnání s těmito systémy je možné vidět, že tematické mapy jsou velmi zajímavým konceptem, který by mohl fungovat jako samostatná metoda pro pořádání informací, ale mohou také částečně nahradit a částečně spolupracovat se stávajícími metodami pro pořádání informací. V každém případě tato metoda přináší celkový posun v chápání pořádání informací a jejich reprezentace.

8.2.2.1 Metadata

O metadatech podrobně pojednává kapitola 4 (s. 17). V krátkosti můžeme definovat metadata jako „soubor výroků o informačním zdroji“ [Garshol, 2004], které se vztahují ke konkrétnímu dokumentu. Hlavní rolí metadat je identifikace předmětů pro hledání (tj. podpora hledání). Hlavním obsahem metadat jsou klíčová slova, popisující téma dokumentu a další informace použitelné pro identifikaci (jméno autora, datum vytvoření dokumentu, technické informace o formátu dokumentu apod.).

Hlavní nedostatky metadat ve srovnání s tematickými mapami jsou [Garshol, 2004]:

- Zachycení vztahů k ostatním tématům

Rozsah informací poskytovaných metadaty je omezen pouze na jediný, izolovaný dokument. Metadata nijak nedefinují jakékoli vazby k okolním dokumentům nebo tématům.

- Nedostatečná úroveň kontextuálních informací

Metadatový záznam obsahuje řadu izolovaných výroků, aniž by rozlišoval jejich význam (označení nejdůležitějšího klíčového slova v uvedeném seznamu), případně jinak definoval vztahy mezi nimi (např. jazyk XML souvisí s jazyky HTML a SGML).

- Metadata mohou usnadnit hledání pouze za určitých podmínek

Úspěšnost při hledání dokumentu závisí na použití stejného termínu, který je v záznamu. Při použití synonyma nebo jiné varianty názvu již hledání není úspěšné. Použití metadat pro vyhledávání tak má stejné problémy jako vyhledávání pomocí vyhledávačů (viz kapitola 9, s. 117). Vyhledávání pouze omezí množinu dokumentů, kterou je potřeba dále zpracovat, ale nemusí znamenat nalezení zamýšleného obsahu.

Metadata mohou být užívána společně s tematickými mapami pro podporu hledání, ale pro jejich funkci nejsou nezbytné. Tematické mapy podporují kontextuální vyhledávání, a identifikace témat a jejich vztahů je řešena pomocí syntaxe (použití interního identifikátoru).

8.2.2.2 Řízené slovníky

Nejjednodušší metodou pro pořádání informací je definování tzv. řízeného slovníku. Jde o seznam klíčových slov, které mohou být použity k popisu konkrétního tématu. Hlavním cílem je omezení možných synonymních termínů a sjednocení terminologie pro popisovanou skupinu dokumentů. Na podobném základě fungují různé soubory tzv. *autorit* (seznamů jmen, názvů apod.), ve kterých je definována preferovaná forma, určená pro popis.

Tato metoda je zaměřena na řešení problémů přirozeného jazyka pro obsahový popis dokumentů, ale **nezabývá se otázkou popisu konceptu a jeho vztahů**. Z tohoto důvodu jsou řízené slovníky používány především jako pomůcka pro popis dokumentů, nikoli jako systém pro pořádání informací. Pro tento účel se využívá spíše tezaurů (viz. kapitola 5.9, s. 61), které kombinují přednosti řízených slovníků s taxonomickými systémy (viz kapitola 5, s. 48).

8.2.2.3 Taxonomická klasifikační schémata

Taxonomická klasifikačními schémata jsou založena na definici terminologie předmětů, definování jejich vzájemných vztahů na základě nadřízenosti a podřízenosti a zachycení těchto vztahů do podoby hierarchického systému. Bližší popis těchto systémů je uveden v kapitole 5.2, s. 49. Tyto systémy jsou jedním z nejstarších způsobů kontextuální navigace, která je založena na prohlížení předmětů. Zařazením předmětu do kategorie tvůrce pomáhá uživateli pochopit základní kontext zobrazením vztahu nadřízený celek – část (např. stomatologie je částí medicíny), a příbuznosti (např. chirurgie je příbuzná se stomatologií, protože jsou zařazeny pod stejnou kategorií vedle sebe). Navigační schémata, založená na taxonomii jsou důležitá pro zachycení konceptů, ale selhávají v jiných oblastech. Hlavními problémy jsou:

Rozlišení druhu vztahů mezi tématy

Základním zachyceným vztahem mezi tématy je podřízený – nadřízený. Další vztahy mohou vyplývat z vytvořených skupin (témata v jedné kategorii jsou si vzájemně příbuzná), ale druh vzájemných vztahů (např. souvisí s, sdílí zdroje, má společný základ...) není nijak systémem definován a musí být interpretován uživatelem.

Definice rozdílů mezi pojmy
Hierarchická soustava nijak neumožňuje rozlišit jazykové a významové jevy jako je synonymie, homonymie nebo vztah

ekvivalence. Existuje-li téma, které nemá ustálenou terminologii, podstatná část obsahu může být zcela vynechána nebo duplicitně zařazena (např. ekvivalentní termíny „topic maps“ a „topic navigational maps“)[Garshol, 2004].

Odkazy mezi tématy
Vzhledem k historickému propojení těchto systémů s fyzickým uspořádáním knih a dalších dokumentů v knihovně není možné odkazovat z jednoho tématu na jiné, případně řadit témata na více míst. Tento problém řeší až webové katalogy (viz kapitola 6, s. 73), které zjednodušily komplikovaná pravidla (např. dědění vlastností) a odkazují na témata mezi kategoriemi.

8.2.2.4 Tezaury

Tezaury jsou pravděpodobně nejpokročilejší formou univerzálního pořádacího systému pro zachycení vztahů mezi předměty. Kromě vztahu nadřízený – podřízený definují i vztahy ekvivalence a systém vzájemných odkazů. To již umožňuje poměrně přesně určit význam tématu ve vztahu k ostatním, a minimalizovat možnost jiných významových interpretací. Tezaury jsou jedním z výchozích konceptů pro tematické mapy, které se snaží vylepšit sémantické možnosti a zavést systém jednoznačně definovaných vztahů.

8.2.3 Výhody použití tematických map

Hlavní přínosy tematických map jsou shrnuty v následujících bodech:

Definice pravidel pro sémantiku

V mapě lze definovat vztahy na základě aplikování znalostí do formy podmínek (např. e-mailovou adresu může mít pouze osoba nebo organizace). Tím lze omezit vyhledávání nerelevantních informací.

Podpora hledání

Systém nalezne nejen záznam, který nejlépe odpovídá dotazu, ale také popisné informace. Nalezené téma navíc může sloužit jako výchozí bod pro prohlížení příbuzných témat, vztahů nebo konkrétních dokumentů.

Podpora komplexních dotazů

Při hledání je možné odlišit typ hledané informace (např. pouze osoby, pouze dramatická díla tohoto spisovatele) a zadávat i komplikované dotazy, které je jinak obtížné zpracovat bez lidské analýzy (např. Kdy se v Praze narodil významný spisovatel?)

Orientace na téma, nikoli na zdroj

Zásadní změnou konceptu oproti starým systémům pro pořádání informací je odlišení tématu od konkrétních dokumentů, které se k němu vztahují. To umožňuje lepší ztvárnění vztahů mezi tématem a dokumentem, který o tématu přináší informaci.

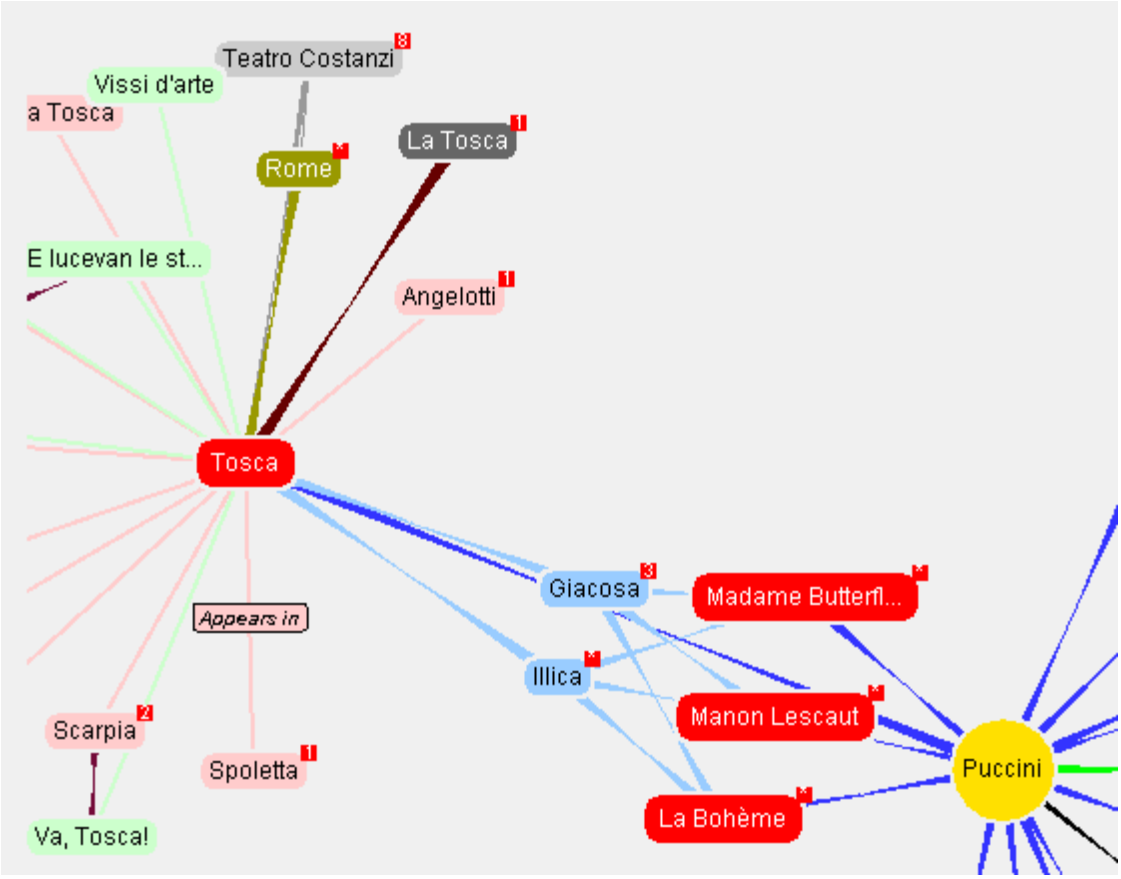
Zachycení vztahů mezi tématy

Tematické mapy mohou přesně definovat a následně rozlišit vztahy, které jsou mezi tématy navzájem, případně i mezi tématy a jejich konkrétním výskytem (tj. dokumentem).

Vizualizace tematických map
Vytvořené tematické mapy je možné vizualizovat pomocí speciálních prohlížečů. Nejznámějším z nich je volně šiřitelný prohlížeč Omnigator[52] společnosti Ontopia, který mapy vizualizuje v programovacím jazyce Java.

---

[52] http://www.ontopia.net/omnigator/models/index.jsp



***Obrázek 26**: Aplikace Omnigator pro vizualizaci tematických map*

Technologie tematických map je vzhledem k výše uvedeným přednostem (především zachycení konceptů, vyhledávání a vizualizace objektů) velmi perspektivní. Její vývoj je ale zatím v počátcích a tak je pravděpodobné, že se zde, stejně jako u ostatních technologií, časem ukáží nedostatky této koncepce.

9 VYHLEDÁVAČE

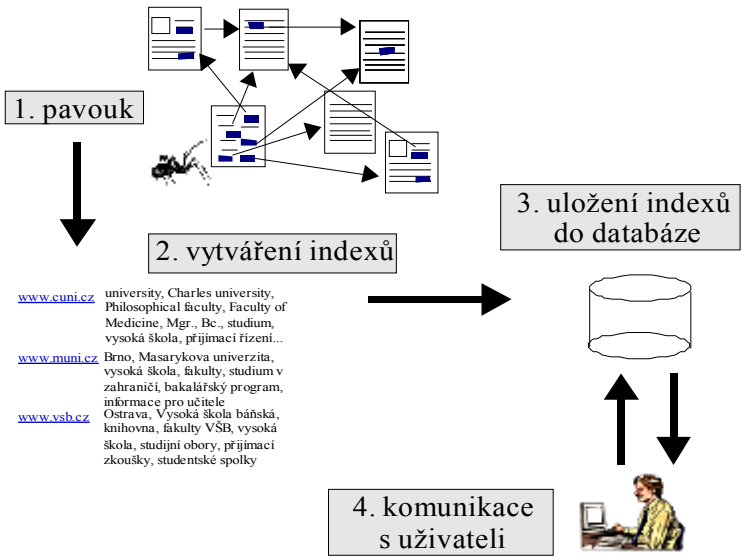
Domníváme se, že vyhledávání informací
[na internetu] je objektivní. Není tomu tak.
Stephen Arnold [2003]

Nejrozšířenější technikou vyhledávání informací na internetu je v současnosti používání tzv. **vyhledávačů** (search-engines), které na základě zadaných klíčových slov předloží uživateli seznam stránek, které by ho mohly potenciálně zajímat. Technologie vyhledávačů je založena na používání metod informatiky, umělé inteligence a počítačové lingvistiky, a snaží se pracovat s co nejmenším podílem práce člověka, anebo zcela bez jeho zapojení.

9.1 PRINCIP FUNKCE VYHLEDÁVAČŮ

Vyhledávač se zpravidla skládá ze tří hlavních částí (viz obrázek 8.1). První částí je tzv. **pavouk** (spider, crawler), což je označení pro specializovaný software, který prochází webové stránky nebo i jiné zdroje (především FTP archivy nebo jiné služby), čte a ukládá data, která na nich najde, a následuje všechny odkazy, které ze zkoumané stránky vedou. Toto zpracování stránek probíhá pravidelně v intervalu několika dní až týdnů, aby software mohl zjistit případné změny v obsahu.

Druhou částí je tzv. **index** neboli **katalog**. Tato část zpracovává veškeré informace od části první – od „pavouka" a vytváří seznam (index) všech nalezených slov, obrázků, případně dalších formátů dokumentů společně s místem jejich výskytu. S každou aktualizací dat se tento seznam mění a k dispozici zůstává pouze aktuální verze. Tato část – **index** je pro vyhledávací software vždy výchozí množinou dat pro hledání.



Obrázek 27: Schéma funkce vyhledávače

Vlastní **vyhledávací software** je třetí částí vyhledávače. Tento software prohledává data ve druhé části – v indexu, na základě speciálních algoritmů nalezne odpovídající dokumenty, a seřadí je podle míry, jakou odpovídají zadanému dotazu (tzv. relevance). Určování této relevance je klíčem úspěchu vyhledávače, protože se největší měrou podílí na

výsledném dojmu pro uživatele [SULLIVAN, 2002]. Vyhledávací software je pro uživatele viditelný v podobě ovládacího rozhraní. Kromě relevance je pro uživatele důležitý i vzhled a jednoduchost ovládání a přehlednost tohoto rozhraní.

9.2 RELEVANCE

9.2.1 Problematika určení významu informace

Otázka, co je pro uživatele přínosnou informací, je diskutována již od vzniku elektronických vyhledávacích systémů. Steffano Mizzaro [MIZZARO, 1997] sestavil vyčerpávající bibliografii různých studií relevance, v nichž je začátek těchto diskuzí datován již do 30. let 20. století. Pojem relevance je často používán bez toho, aniž bychom rozuměli jeho významu. Obecně můžeme relevanci definovat jako stav, kdy jsou v informačním systému nalezeny informace, které odpovídají zadanému dotazu. Vedle relevance se také objevuje termín **pertinence**, který popisuje stav, kdy jsou v informačním systému) nalezeny informace, které **odpovídají informační potřebě uživatele** (tj. to, co uživatel ve skutečnosti potřeboval).

Tyto definice však plně nevystihují tuto problematiku a tak je vhodné přiblížit relevanci na komplexním modelu. V modelu S. Mizzara [1997] je relevance popsána jako *„vztah mezi dvěma entitami ze dvou skupin“*. Autor v modelu do první skupiny řadí **dokument** (fyzickou entitu, kterou uživatel obdrží), **zástupce** (reprezentaci dokumentu ve formě bibliografického záznamu), a **informaci** (to, co uživatel při čtení dokumentu získá).

Ve druhé skupině jsou zařazeny: **problém** (to, co uživatel potřebuje vyřešit), **informační potřeba** (reprezentace problému v mysli uživatele), **požadavek** (reprezentace informační potřeby v přirozeném jazyce uživatele) a **dotaz** (ztvárnění informační potřeby v systémovém jazyce informačního systému; tj. v databázi či vyhledávači). Relevanci pak můžeme vnímat jako vztah mezi dvěma prvky z první a druhé skupiny: například relevanci mezi dokumentem a požadavkem, relevanci mezi informací, kterou uživatel získá a svou informační potřebou a podobně [MIZZARO, 1997].

Relevanci tak můžeme znázornit jako bod ve čtyř dimenzionálním prostoru, kde jsou **prvky ze dvou výše uvedených skupin** (dokument, zástupce informace a problém, informační potřeba, požadavek a dotaz), **téma nebo kontext** a **časová osa**, která vystihuje situace od vzniku problému po jeho úspěšné vyřešení. Tuto situaci zjednodušeně přibližuje obrázek č. 8.2, kde je vynechána časová osa a kontext. Každá zobrazená přímka, která spojuje dva body, představuje relevanci (pro zvýraznění je zdůrazněna kolečkem).

Tento model vystihuje podstatu tohoto problému – přiřazení správné informace informační potřebě člověka. Podstatnou roli hraje také přiřazení významu hledané informaci, tj. posuzování relevance. To je zcela individuální a každý ho posuzuje podle informační potřeby, kterou často neumí vyjádřit a kontextu dosavadních informací k hledanému tématu.

V případě vyhledávačů je otázka relevance úzce svázána se schopností uživatele formulovat svou informační potřebu (co potřebuji udělat, jakou informaci hledám) do podoby dotazu ve vyhledávači (kombinace klíčových slov, operátorů a nabídkových polí). Jak je zřejmé z části 9.6, problémem je malá schopnost uživatelů svoji informační potřebu takto zpracovat. Problém je přenášen na vyhledávače, které se snaží použít různé techniky (statistické sledování často kladených dotazů, zpětná vazba, tezaury, kladení dotazů ve formou přirozeného jazyka apod.) pro nalezení dokumentů, které by mohly uživatele zajímat. Přes veškerou snahu vyhledávače nemohou nikdy plně nahradit lidskou schopnost porozumění přirozenému jazyku, což bude vždy problémem jak pro relevanci ve vztahu dotaz uživatele a nalezený dokument, tak ve vztahu zpracování a vyhodnocení dokumentu vyhledávačem.

9.2.2 Určování relevance dokumentů vyhledávači

Proč mají vyhledávače určovat relevanci dokumentů? Vzhledem k počtu nalezených odkazů, který se pohybuje v řádu několika set tisíc, je to jediná možnost, jak uživateli alespoň částečně pomoci v orientaci. Určování této relevance má také své nevýhody. Phil Bradley [2000] je shrnuje do dvou bodů:

Kvalita vyhodnocení a tedy i relevance, přímo závisí na prohledávaných datech. Pokud tato data (v případě vyhledávačů především webové stránky) jasně nedeklarují svůj hlavní obsah, vyhledávač to není schopen určit sám.

Uživatelé jsou odkázáni na schopnost autorů psát stránky, které budou dobře hodnoceny vyhledávači. I když jsou metody určování relevance zaměřeny na vyloučení nekvalitních stránek, je možné, že v průběhu zpracování budou hodnotícím mechanismem vyloučeny stránky, které by uživatel považoval za relevantní.

Základem pro porozumění informaci je lidská schopnost formulovat myšlenky do přirozeného jazyka, případně je převést opačně (tj. ze slov a vět do myšlenkových konceptů). Tato lidská schopnost se zatím nedá nahradit automatizovanými systémy, a tak se metody porozumění obsahu dokumentů soustřeďují především na statistické zpracování textu a na analýzu zdrojového kódu dokumentu, který je zpravidla založen na HTML a dalších značkovacích jazycích (XHTML, XML...). Konkrétní metody určování relevance vyhledávači jsou přísně střeženým obchodním tajemstvím jejich provozovatelů. Obecně se však předpokládá, že se pro hodnocení relevance používají kombinace následujících metod:

Četnost hledaných slov

Jedna z nejstarších metod pokládala za nejrelevantnější dokument s největším počtem hledaných slov. Tvůrci stránek tuto metodu brzy odhalili a tak začali uvádět ve svých dokumentech množství slov, které nesouvisela s tématem stránek a tak tuto metodu zcela znehodnotili.

Hustota slov

Pokud má stránka 100 slov a klíčové slovo je zde pětkrát, znamená to hustotu 5% (5 klíčových slov na 100 slov celkem). Vyhledávač tak hodnotí význam klíčových slov v poměru k dalším tématům na stránkách.

Vzdálenost termínů od sebe

Vyhledávače neumí rozeznat význam a tak se pokouší odvodit jej alespoň přibližně podle vzdálenosti klíčových slov od sebe. Klíčová slova, která jsou u sebe, hodnotí výše než jiná, která mohou být v dokumentu bez vzájemného vztahu.

Strukturní značky jazyka HTML

Jazyk HTML a jeho nástupci jsou jazyky strukturovanými, a tak je možné přesně určit význam jednotlivých prvků v jejich zdrojovém kódu. Párová značka (tzv. *tag*) „<TITLE></TITLE>“ je označením titulku stránky, značka <H1></H1> označením nadpisu první úrovně a tak dále. Tímto způsobem je možné hodnotit váhu klíčových slov podle jejich umístění do těchto značek. Předpokládá se, že podobné zpracování probíhá i u dokumentů v jiných formátech (např. PDF, soubory MS Office).

5. Značky „meta“

První primitivní forma metadat se objevila v jazyku HTML verze 2.0 (viz kapitola 4.8.1,). Tato metadata se stala jednou ze základních možností, jak popsat dokument. Obecný model pro značky META vypadá takto:

<META name="definice údajů" content="obsah">

Do obecného modelu, který je zde uveden, se doplňují dva typy údajů: **„name“** pro definici typu údaje (např. autor, obsah, určení kódová stránky) a **„content“,** který doplňuje uvedený typ údaje konkrétním obsahem (konkrétní jméno autora, popis obsahu stránky, klíčová slova apod.). Vysoký význam pro hodnocení, který těmto značkám původně vyhledávače přikládaly, se stal důvodem k tzv. *index spammingu* a tak byl následně jejich význam v hodnocení rapidně snížen.

6. Odkazy

Vyhledávač Google jako jeden z prvních zavedl systém hodnocení, který zohledňuje odkazy z jiných stránek na stránku hodnocenou. Tento systém předpokládá, že pokud je stránka vysoce odkazována, znamená to, že ji uživatelé považují za relevantní.

7. Autorita stránek

Greg Notess [1999] uvádí, že například Google doplňuje hodnocení relevance o určení autority stránek. Stránky univerzit nebo vládních organizací jsou hodnoceny výše, než stránky osobní nebo takové, u kterých se nedá ověřit původ jejich informací.

Další metody hodnocení mohou mít souvislost se zpětnou vazbou uživatelů, kdy se vyhodnocují statistické záznamy o hledaných tématech a kladených dotazech. Většina vyhledávačů tak nabízí funkce typu *„More like this“* (hledání stránek podle vzoru nalezené stránky) nebo *„Did you mean?“* (návrhy často hledaných témat a kontrola překlepů). Podle dokumentovaného chování tak mohou vznikat tematické tezaury, které při zadání dotazu zohlední podle klíčového slova tématiku, která je pravděpodobně hledána.

Hodnocení relevance vyhledávačem nikdy nemůže být dokonalé. Příčinou nezdaru vyhledávání jsou často uživatelé sami, ale jak komentuje Phil Bradley, také různé způsoby hodnocení relevance ve vyhledávačích: *„Příčinou toho, když hledáte na jednom vyhledávači a zdá se, že nic nemůžete najít … může být jenom to, že proces hodnocení neodráží vaše představy o relevanci a tak často stačí změnit vyhledávač a se stejným dotazem budete mít štěstí jinde!“* [Bradely, 2000]. Kvalita výsledků hledání přesto není nijak zázračná. Vyhledávače najdou značnou část, ale nikdy vše, co je k problematice dostupné. Chovají se tak podle pravidla Alexe de Toqueville, *„pokud většina hlasovala, výsledkem je průměr“* [Arnold, 2003][53], nebo-li jestliže bylo nalezeno alespoň něco k zadanému dotazu, považujme to za úspěch.

9.2.3 Index spamming

„Všechny standardní techniky pro hodnocení relevance selhaly kvůli nečekanému aspektu velmi dynamické povahy Webu. Nebo možná přesněji kvůli lidské povaze. Od doby, kdy začaly být vyhledávače používány pro hledání informací, se tvůrci webových stránek neustále snaží zvýšit hodnocení svých stránek v těchto vyhledávačích.“ [Notess, 1999].

Význam co nejvyššího umístění stránek v seznamech odkazů vyhledávačů dal vzniknout novému odvětví, které se nazývá *Search Engine Optimalization* (tzv. SEO). To se zaměřuje na optimalizaci stránek pro vyhledávače různými úpravami stránky (značky META, úprava nadpisů různých úrovní, klíčová slova v názvech apod.). V následné soutěži za co nejvyšším počtem návštěvníků stránek mnoho tvůrců přestalo respektovat veškeré zásady slušnosti a pravdivosti a pro získání lepší pozice v hodnocení svých stránek používají podvodné metody, které se souhrnně označují jako **index spamming**.


Diagnostics | Articles | Authors | OS-2 | Directories | Amiga | Networking | Associations | Unix | DOS | Macintosh | Windows | Submitting Services | Web Rings | Resources
Printing | Calendars and Planners | Database | Editors | Editors | .. | CE | Desktop and System | List Management | Calendars | Viewers and Cataloguing | Digital Video | Clients | Email | Microsoft PhotoDraw | NT | Copyright and Protection | Plug-ins | Password Protection | Screen Capture | Programming | Editors | WWW | Food and Beverages | Business and Finance | Browsers | Editors | Clocks and Timers | FTP | Optimization | Delphi | Cache Tools | Mathematical Graphing | Encryption | Add-Ons | Bookmark Managers | Anti-Virus | Keylogger | Offline Browsing Tools | Utilities | Hexadecimal | Financial, Insurance and Home Inventory | Password Recovery | Mouse and Keyboard | MAPI | Collection | FastCAD | Extractors | Download Managers | Editing

Obrázek 28: Ukázka index spammingu

Principem těchto metod je zmást hodnotící algoritmus vyhledávačů tak, aby bylo možné vytvořenou stránku v hodnocení posunout co nejvýše. Časté praktiky jsou například:

- uvádění množství slov, které nemají souvislost se stránkami v metadatech (značky META) nebo přímo v textu stránky
- seznam klíčových slov na stránce ve stejné barvě jako pozadí (neviditelný text)

[53] When the majority votes, the result is mediocrity.

- množství odkazů na jiné stránky
- tzv. *„farmy odkazů“* – stránky, které neobsahují nic jiného než odkazy na jiné stránky, které mají podporovat. Tento systém byl vytvořen za účelem oklamat hodnotící mechanismus vyhledávače Google.

Z určování relevance se stal boj mezi ctižádostivými autory a tvůrci vyhledávačů, kteří musí zajistit určitou relevanci nalezených odkazů, aby obstáli v náročné konkurenci tohoto trhu. Tento boj se neustále vyvíjí – na nový způsob hodnocení relevance reagují autoři dalšími triky. Řešení tohoto problému bohužel zatím není v dohledu. Jedná se čistě o etický problém, který se v prostředí internetu, kde schází centrální řídící autorita, prakticky nedá vyřešit.

9.3 PROBLÉMY SOUČASNÝCH VYHLEDÁVAČŮ

Pokud akceptujeme koncepci vyhledávačů jako hledání klíčových slov, doplněné dalšími nabídkami k usnadnění hledání, musíme si uvědomit omezení a možné problémy, které jsou s nimi spojeny. Kromě index spammingu jsou hlavní problémy vyhledávačů:

- velikost indexu
- aktualizace indexu
- indexace speciálních formátů dokumentů dynamických stránek.

Velikost indexu označuje množství dokumentů, které vyhledávač nalezne a v kterých hledá. Tato velikost je přibližným měřítkem množství informací, které je v těchto vyhledávačích možné najít. Žádný z vyhledávačů neindexuje všechny dostupné stránky či jiné dokumenty. Zároveň studie [Lawrence-Giles, 1999] poukázala na to, že pokrytí webových stránek vyhledávači mezi léty 1997 a 1999 pokleslo, a vzhledem k dynamickému nárůstu počtu nových stránek se bude tento trend prohlubovat. Dále tato studie dochází k závěru, že vzájemné překrytí indexů mezi různými vyhledávači je velmi nízké. To znamená, že neustále bude existovat obsah, který nebude zachycen ani jedním vyhledávačem, a zároveň vzroste množství unikátních odkazů u každého z vyhledávačů.

Aktualizace indexu je pro vyhledávání klíčová. Uživatel vždy komunikuje s databází vyhledávače a tak nutně dochází ke zpoždění mezi časem indexace informace vyhledávačem a její prezentací uživateli. Podle statistik Grega Notesse z roku 2003 [Notess, 2003] je průměrnou dobou pro obnovení indexu přibližně jeden měsíc. Tuto dobu potvrzuje i francouzská studie [How do search tools work, 2003] z října stejného roku. Společnost Inktomi podle [Metadent, 2004] tvrdí, že její „pavouk“ jménem Slurp dokáže indexovat 10 miliónů stránek za den a veškeré změny se tak v indexu projeví do dvou dnů. I když je to oproti roku 2003 zřetelný posun, riziko hledání ve starých informacích přetrvává i nadále.

V prostředí internetu a speciálně v prostředí služby word wide web se objevují i další formáty dokumentů, které kladou zvýšené nároky na zpracování. Jde především o soubory ve formátu PDF, který se stal standardem pro oficiální a akademické publikace a o soubory ve formátech MS Office, které jsou známkou malé schopnosti (či lenosti)uživatelů publikovat informace ve formátu HTML. Tyto dokumenty jsou ve vyhledávačích zpracovávány doplňkovými moduly (např. pro čtení souborů PDF). Přesto jejich indexace není bezproblémová, a tak jejich podíl v celkovém objemu indexu vyhledávače bude vždy nižší než počet stránek založených na jazyku HTML.
Problémem jsou také různé dynamické stránky, které není možné indexovat, neboť se vytvářejí takříkajíc „na míru“ podle požadavku uživatele. Tak dochází k tomu, že řada materiálů je obtížně dohledatelná. Oblast těchto špatně vyhledatelných informací je označována jako „invisible web“ a objevují se první specializované vyhledávače, které se na tuto oblast zaměřují, a jsou ve vyhledávání těchto typů dokumentů úspěšné.

Vyhledávače mají před sebou celou řadu velkých výzev. Stephen Arnold [Arnold, 2003] je rozděluje na následující okruhy:

- Zpracování přirozeného jazyka.

  Porozumění definice termínu a jeho propojení s konceptem je a bude v automatizovaném zpracování vždy problémem. Jevy jako metafory, synonymie nebo homonymie jsou pro vyhledávací systém těžko řešitelné bez asistence člověka. Podskupinou těchto problémů bude vyřešení jazyků se znakovými sadami, které jsou jiné než latinka. Jazyky jako čínština, korejština nebo arabština budou pro zpracování a vyhledávání těžkým oříškem.

- Lidé nemají představu o správné odpovědi na svůj dotaz.

Lidé při hledání často nemají představu o tom co hledají, a snaží se své myšlenky převést do podoby dotazu (blíže viz kapitola o relevanci – s.120). Vyhledávače musí vyvíjet další pomůcky, které v tomto procesu uživateli pomohou. Už nyní se objevují nástroje pro návrhy dotazů – např. služba Google Suggest[54], která navrhne uživateli klíčové slovo a zároveň uvede počet odkazů, které na ně má ve své databázi.

- Se změnou demografie populace se musí měnit i typy vyhledávacích systémů

Popularita systémů pro sdílení souborů peer-to-peer ukázala, že mohou existovat i jiné modely informačních systémů. Přístup k naplnění informační potřeby tak může existovat například v modelu „položte otázku a rozešlete ji všem uživatelům systému" [Arnold, 2003]. Zárodky takových „informačních komunit" můžeme vidět u některých obchodních systémů jako eBay nebo Bizsearch, kde je uživatelské fórum důležitým informačním zdrojem pro rozhodování. Řadu let ale fungují různá uživatelská fóra (v prostředí www nebo služby USENET), která jsou pro řadu technických oborů (především výpočetní technika) důležitějším informačním zdrojem než vyhledávače.

- Personalizace vyhledávacích služeb

Tendence personalizace je v prostředí internetu stále zřetelnější. Vidíme ji v podobě úpravy prostředí vyhledávacích portálů (např. Yahoo!), ve vytváření zvláštních nastavení pro e-mailové účty nebo doporučení produktů, které by nakupujícího podle jeho zájmu mohly zajímat (např. Amazon). Idea nového „sémantického" webu by měla celý systém uvést do života prostřednictvím osobních vyhledávačů, které by na základě nastavení uživatele automaticky zjistili požadovanou informaci.

- Ověřování informací

Uživatelé chtějí přesná, dohledatelná data. Především v případě vědeckých, zdravotnických a spotřebitelských informací je důležité zajistit jejich pravdivost. Ne vždy jde o životně důležitou informaci jako u popisu nežádoucích účinků nového léku, ale pokud je naše jednání těmito informacemi ovlivněno, stává se pro nás jejich pravdivost nejdůležitějším faktorem.

Problémy vyhledávačů nebudou pravděpodobně nikdy úplně vyřešeny. Dokonalost výpočetní techniky se neslučuje s lidskou nedokonalostí ve formulování myšlenek, a tak člověk množství chyb vytváří tím, že nepochopil koncept a problémy tohoto systému vyhledávání informací. Hlavní oblastí, která by se měla zlepšit, je tak osvěta uživatelů (viz část 9.6).

9.4 TRH VYHLEDÁVAČŮ

Trh vyhledávačů je velmi náročným ekonomickým prostředím, které jednotlivé firmy neustále žene do technologických a marketingových inovací jen aby si udržely svou pozici na trhu, případně aby svůj tržní podíl rozšířily. Změny, které nastávají na žebříčcích hodnocení ukazují, jaký význam firmy tomuto prostředí přikládají, a jak jsou tyto faktory hnacím motorem dalšího vývoje.

Tržní podíly jednotlivých firem se od zrodu této technologie významně změnily. Přehled nejlépe hodnocených vyhledávačů (tj. zároveň vyhledávačů s největším tržním podílem) je uveden v příloze č. 1. Danny Sullivan ve svém článku [Sullivan, 2005] nazývá ostré soupeření mezi vyhledávači „válkami" a rozdělil je do čtyř období: 1. válka (září 1997 – červenec 1999), 2. válka (září 1999 – červen 2000), 3. válka (červen 2002 – prosinec 2002) a 4. válka (od srpna 2003).

V těchto „válkách" jde především o zlepšení relevance vyhledávače a zvláště o zvýšení deklarované velikosti indexu

Vyhledávač	Deklarovaná velikost indexu	Limit pro indexaci stránek (v kilobytech)
Google	8.1 miliard	101 K
MSN	5.0 miliard	150 K
Yahoo	4.2 miliard (odhad)	500 K
Ask Jeeves	2.5 miliard	101 K+

54

***Tabulka 8**: Deklarované velikosti indexů k lednu 2005 Zdroj: SULLIVAN, 2004A*

vyhledávače. Tato hodnota je totiž hlavním marketingovým údajem, který je mezi uživateli podvědomě chápán jako určitý ukazatel kvality. Firmy se tak předháněly a předhání v publikování tiskových zpráv o velikosti svých indexů. Situace k lednu 2005 je uvedena v tabulce 9[55].

Deklarované velikosti indexů není možné objektivně porovnat, navíc se mohou lišit i definice, co je měřitelnou jednotkou (webová stránka včetně grafiky, počítání textu a grafiky jako samostatných dokumentů apod.). Odhady velikosti indexů vyhledávačů na serverech jako searchenginewatch.com nebo searchengineshowdown.com se tak provádějí zkušebním vyhledáním série dotazů na každém z nich a následném porovnání počtu odkazů. Na konci roku 2004 se objevil i specializovaný software Thumbshots Ranking[56], který umožňuje porovnání velikostí indexů dvou vyhledávačů podle zadaného slova.
Tabulka 10 ukazuje, že se v těchto válkách postupně mění soupeři. Méně konkurenceschopní odpadávají z tohoto boje, nové firmy a technologie se objevují a také dochází ke slučování některých firem mezi sebou (např. firma Inktomi a Yahoo!). Pro srovnání je možné uvést statistiky vyhledávačů s největším indexem mezi léty 1997-2002, které jsou v tabulce 10.

březen 2002	Google, WiseNut, AllTheWeb
duben 2001	Google, Fast, MSN (Inktomi)
duben 2000	Fast, Alta Vista, Northern Light
březen 1999	Northern Light, Alta Vista, HotBot
květen 1998	Alta Vista, HotBot, Northern Light
únor 1998	HotBot, Alta Vista, Northern Light
červen 1997	HotBot, Alta Vista, Infoseek

Tabulka 9: Vyhledávače s největším indexem 1996 – 2002 Zdroj: Notess, 2002

Tento trh výstižně popisuje Stephen Arnold [2003] slovy: „Darwinistická podstata vyhledávacího průmyslu dovoluje aby se malé specializované společnosti na trhu rychle objevily a často zmizely stejně rychle... Doufejme, že některé [z nových společnosti] přežijí a budou prosperovat. Je ale sporné, zda v blízké budoucnosti bude tyto nové firmy výzvou pro dominantní společnosti na trhu“. Všechny uvedené firmy nabízí své vyhledávací služby zdarma[57]. Zisk, který je podmínkou jejich přežití tak musí získat jinak – z reklamy (včetně placených odkazů ve výsledcích), nebo z licencování své technologie dalším vyhledávačům. Tento trh je velmi perspektivní a výnosný. Jen za rok 2002 uvádí [Arnold, 2003] zisk firem Overture 500 miliónů dolarů a Google 300 miliónů dolarů. Vzhledem k těmto tržbám se nabízí otázka, zda není objektivita ve výsledcích těchto vyhledávačů omezována na úkor placených odkazů. V tomto prostředí má každý uživatel na výběr – používat vyhledávač zdarma nebo využít placených vyhledávačů, které budou zcela bez reklam. Vývoj a provoz každého vyhledávače musí být zaplacen jakoukoli formou. Přirozeně se tak na trhu nejlépe drží firmy, které jsou již dostatečně finančně zajištěné – např. MSN (provozuje firma Microsoft) nebo Google (nyní již veřejně obchodovaná akciová společnost). Zároveň tyto firmy mají dostatečný podíl na trhu reklamy na internetu a jsou ziskové. Celosvětové odhady tržních podílů vyhledávačů jsou velmi obtížně zjistitelné, takže musíme pro ilustraci zvolit pouze Spojené státy americké.

Podíl dotazů v různých vyhledávačích v květnu 2004 a září 2007 je znázorněn v tabulce 11. I když jde pouze o vyhledávače z USA, zdá se, že několik velkých firem začíná dominovat celému trhu (Google, MSN, Yahoo!). Tato dominance neplatí

[55] Novější údaje nebyly v době uzávěrky dat pro práci k dispozici.
56 http://ranking.thumbshots.com/

57 Existují také firmy jako např. *Verity* nebo *Northern Light*, které se zaměřují na firemní sektor a nabízejí placené vyhledávače nebo komplexní firemní řešení pro vyhledávání.

v celém světě bezvýhradně. Některé národní vyhledávače[58], které mají zvláštní podporu jazyka (skloňování, časování) a indexují hlavně domácí zdroje, jsou oblíbenější, protože z pohledu uživatelů přinášejí často relevantnější informace. Průzkum můžeme ale vnímat jako podíly na celosvětovém trhu, kterému společnosti z USA dominují. Prvních pět firem s největším indexem je popsáno v další části této kapitoly.

58 V České republice například http://www.jyxo.cz.

Vyhledavač	Podíl (v % všech dotazů)	
	Květen 2004 (Zdroj: Sullivan, 2004b)	Září 2007 (Zdroj:Nielsen NetRankings 2007)
Google	36,8	54
Yahoo	26,6	19,5
MSN Search	14,5	12
AOL	12,8	6

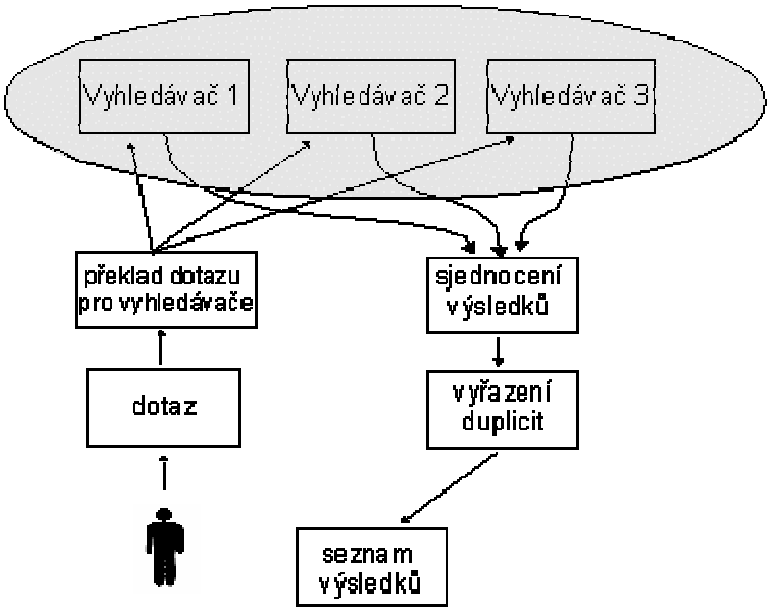
Tabulka 10: Distribuce dotazů na vyhledavače 2004 – 2007

9.5 SPECIALIZOVANÉ VYHLEDÁVAČE

9.5.1 Meta vyhledávače

Vzhledem k nabídce řady konkurenčních vyhledávačů se brzy zrodila myšlenka na jejich vzájemné propojení, které by umožnilo využít předností každého z nich. Tak se objevila koncepce meta vyhledávače, který tvoří specifické vyhledávací rozhraní mezi uživateli a vlastními vyhledávači.
Na obrázku 32 je zobrazená funkce meta vyhledávačů. Uživatel klade dotaz pouze v jednom vyhledávacím rozhraní meta vyhledávače, který dotaz uživatele přeloží do dotazovacích jazyků jednotlivých vyhledávačů a následně jim je odešle. Poté meta vyhledávač vyhodnotí jejich odpovědi, vyřadí z nich duplicity a jeden unikátní seznam všech výsledků se zobrazí uživateli.

V případě meta vyhledávačů se tak nejedná o originální řešení vyhledávání, ale pouze o sjednocení výsledků cizích vyhledávačů. Meta vyhledávače jsou oblíbené u uživatelů, kteří chtějí prohledat co největší objem dokumentů a přitom nechtějí ztrácet čas hledáním na každém vyhledávači zvlášť. Mezi oblíbené meta vyhledávače patří například Ixquick, Profusion nebo Mamma.[59]

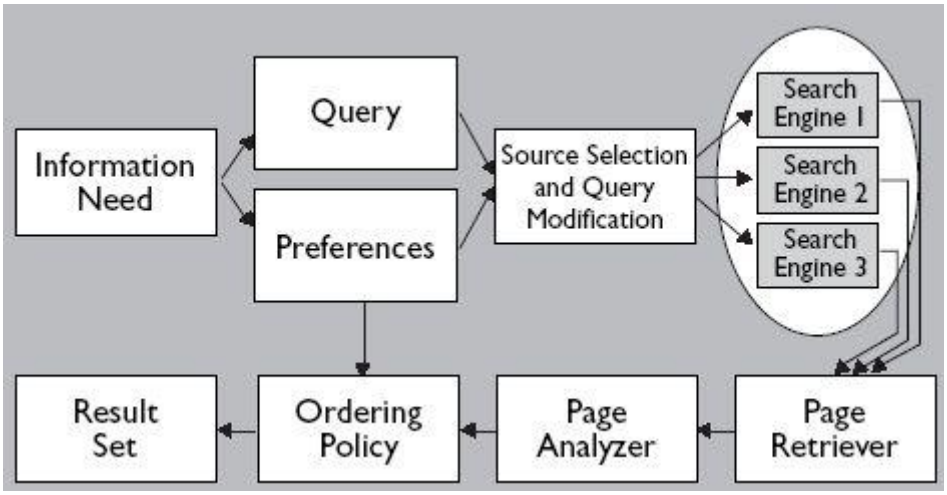


Obrázek 29: Funkce meta vyhledávače

[59] Ixquick (http://ixquick.com), Profusion (http://profusion.com), Mamma (http://mamma.com).

Obrázek 30: Vyhledávač Vivisímo Zdroj: http://vivisimo.com

Meta vyhledávačem, s rozšířenou funkčností, neomezujícím se na pouhé sjednocování cizích seznamů výsledků je



***Obrázek 31**: Architektura Inquirus 2* *Zdroj: GLOVER, 2001*

Inquirus[60]. Jde o výzkumný projekt, testující nové metody vyhodnocení relevance odkazů. Autoři [Glover, 2001], poukazují na to, že tradiční meta vyhledávač vyhodnocuje relevanci stránek a jejich pořadí v seznamu výsledků pouze podle převzatých výsledků, tj. názvu, abstraktu a URL zdroje. Tento projekt se snaží ponechat možnost řazení výsledků na uživateli:„Spíše než koncept relevance, používáme koncept hodnoty. Hodnota dokumentu je subjektivní... řešení pro organizaci, které zohledňuje koncept hodnoty a není založeno pouze na vyhodnocení klíčových slov, je velmi žádoucí.“ [Glover, 2001, s.4]. Inquirus 2 nabízí několik způsobů hodnocení stránek, k nimž se pojí zvláštní seznam zdrojů a pravidla

60 http://inquirus.nj.nec.com/i2/inq2.pl

pro vyhodnocení relevance. Uživatel má možnost zvolit si způsob, který mu vyhovuje a na jeho základě bude mít seřazené výsledky.
Architektura Inquirus 2, popsaná na obrázku 34, je tak rozšířením tradičního modelu meta vyhledávače. Uživatel formuje svou informační potřebu v dotaz, a na základě jeho preferencí jsou zvoleny zdroje pro hledání. Po nalezení výsledků je jejich seznam řazen podle předem zvolených kritérií.

9.5.2 Klastrovací vyhledávače (seskupování odkazů)

Některé vyhledávače a meta vyhledávače využívají metod automatizované kategorizace pro odlišení nalezených témat. Na základě vloženého dotazu jsou vyhledány stránky, které jsou analyzovány, a následně sdruženy do skupin podle příbuznosti svých charakteristik. Například dotaz „Havel“ na vyhledávači *Vivisímo* rozdělí nalezené odkazy do skupin:

Vaclav Havel (bývalý prezident ČR)

Brandenburg (odkazy k městu Brandenburg an der Havel)

Werder (řeka Havel)

Art (malířka Julia Havel).

Tímto způsobem je velmi dobře vyřešena problematika odlišení konceptů slov mezi sebou. Zároveň je možné jednoduše zúžit vyhledávané téma a upravit dotaz. Technologii klastrování využívá například *Ask Jeeves* (pravděpodobně v kombinaci se svou databází lidmi zpracovaných odpovědí na otázky), *SurfWax*, *ez2find* nebo *Dogpile*.[61]

Jednoznačným lídrem tohoto segmentu je společnost **Vivísimo, Inc.** se svými vyhledávači *Vivisímo* a *Clusty*.[62] Metoda třídění odkazů je zde založena na textové a lingvistické analýze, podle které jsou rozdělovány odkazy do kategorií a podkategorií. Zdrojem výsledků jsou vyhledávače (např. *Gigablast*, *Overture*, nebo *MSN Search*), zpravodajské servery (např. *CNN, NY Times, USA Today* nebo *Reuters*), ale také vládní a odborné portály (např. *FirstGov*, *Business.com* nebo *PubMed*).

Vyhledávač Clusty spolupracuje také se specializovanými vyhledávači BizRate (hledání zboží) a PicSearch (vyhledání obrázků a multimédií), a nabízí tak seskupování do speciálních kategorií. Zboží může být například řazeno podle druhu (např. knihy) a dále podle podskupin (žánr knihy), podle prodejce, nebo podle své ceny.

Svou technologii licencuje Vivisímo Inc. řadě velkých firem i vládních organizací, mezi kterými můžeme uvést Boeing, Novell, Symatec, Pfizer, Sun nebo NASA.

9.5.3 Vizualizační vyhledávače

Metoda vizualizace je používána i u některých vyhledávačů. Proces vizualizace je zpravidla řešen v programovacím jazyce Java nebo Flash, který umožňuje dynamicky zobrazit jednotlivé koncepty a jejich propojení.

Pravděpodobně nejpokročilejší možnosti nabízí vyhledávač KartOO[63], využívající zobrazení mapy technologií Flash, který umožňuje vizuální mapy ukládat, upravovat nebo tisknout. Je zde dokonce podporováno i vyhledávání v přirozeném jazyce.

Druhým zástupcem vizualizačního vyhledávače je WebBrain[64]. Umožňuje zadávat dotazy ve formě klíčových slov a frází,

---

61 SurfWax (http://surfwax.com), ez2find (http://ez2find.com), Dogpile http://www.dogpile.com).

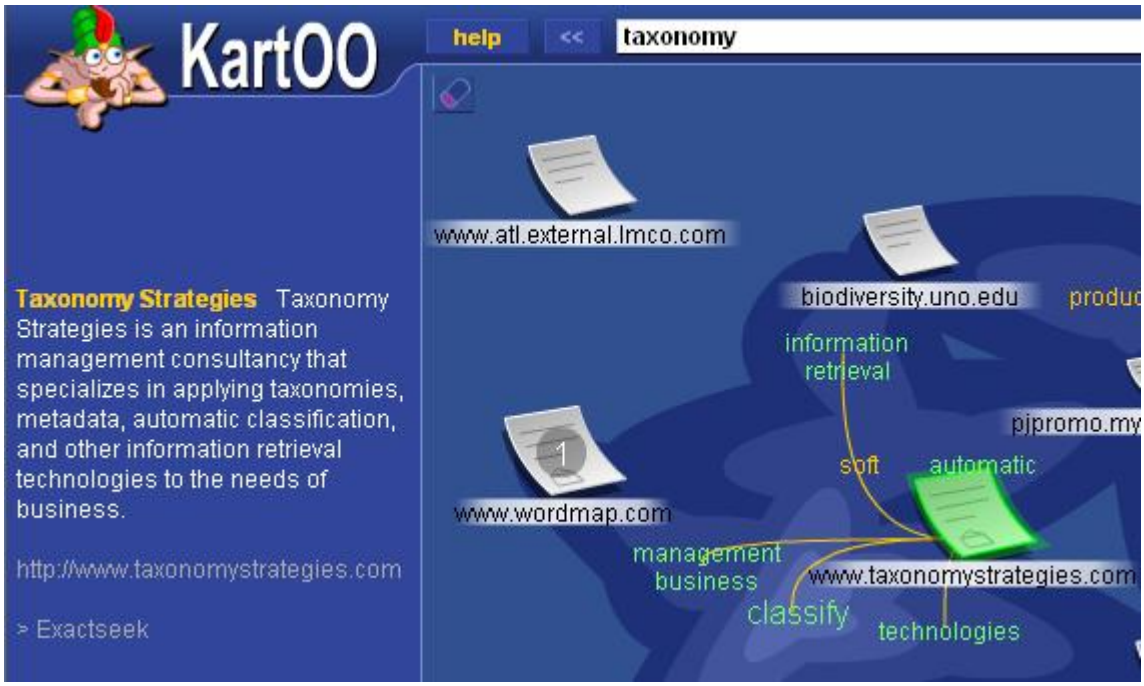
62 http://vivisimo.com, http://clusty.com.

63 http://kartoo.com

64 http://www.webbrain.com

které jsou hledány ve webovém katalogu Open Directory. Vyhledávač nemá žádné pokročilé rozhraní, a zdá se, že jej jeho provozovatel, firma TheBrain Technologies Corporation, vytvořila jako podporu pro své další produkty PersonalBrain, BrainEKP. Jedná se o specializovaný vizualizační software pro pořádání informací, umožňující třídit dokumenty na počítači stejným vizuálním způsobem, jako ve vyhledávači WebBrain.

Metodu vizualizace pro pořádání informací využívají i jiné aplikace, které se ale této práce dotýkají pouze okrajově.



Obrázek 32: Vyhledávač KartOO *Zdroj: http://kartoo.com*

Příkladem může být výkladový jazykový slovník Thinkmap Visual Thesaurus[65].

9.5.4 Vyhledávací agenti

Vyhledávače jsou nesmírně populární a tak se nejen v rámci marketingových úvah objevila myšlenka na vytvoření personalizovaných vyhledávačů. Ty jsou dostupné jako speciální programy, jejímž prostřednictvím je možné vyhledávat na internetu. Jde o různě komplikované systémy, které mohou pracovat na principu meta vyhledávače nebo i s podílem umělé inteligence. Podle své komplexnosti jsou tyto programy určeny pro jednotlivé uživatele (jednodušší systémy, založené na meta vyhledávání), nebo pro náročnější zákazníky a firmy (samoučící se systémy, založené na umělé inteligenci).

Většina dostupných programů z první skupiny slouží jako vyhledávač pro internet stejně jako pro hledání na počítači. Přínos těchto programů je především ve větším uživatelském komfortu při hledání; z hlediska technického řešení **jde pouze o vylepšení konceptu meta vyhledávačů**. Nejznámějšími zástupci těchto programů jsou *Copernic*, *WebFerret*, nebo *x1*[66]

K této skupině vyhledávacích agentů se přidávají také specializované programy dalších producentů, především firem, provozujících vyhledávače, které tak přenášejí boj o zákazníka i do této oblasti (viz. kapitola 0, s.134).

65 http://www.visualthesaurus.com/

66 http://www.copernic.com, http://www.ferretsoft.com/, http://www.x1.com/.

Druhou skupinou „agentů“ jsou programy, schopné samostatné činnosti na základě vložených dat a instrukcí (tzv. **inteligentní agenti**), nebo programy, které se samy mohou přenášet v počítačové síti (například internetu) a na každé stránce znovu vykonat zadanou úlohu (tzv. **mobilní agenti**)[67].

Vývoj těchto agentů je podle [Kotz et. al., 1999] úzce svázán s množstvím různých uživatelů a jejich informačních potřeb. To přináší požadavek personalizace hledání, kterou lze vhodně zajistit prostřednictvím těchto agentů.

Inteligentní i mobilní agenti jsou dostupní jako hotové produkty (např. Deep Query Manager[68], Botbox.com[69]), nebo jsou vytvářeny na míru zákazníkům (např. BotQL[70]). S vyhledávacími agenty (především s mobilními agenty) je úzce spojena také budoucnost konceptu sémantického webu, kde by software společně s ontologiemi (slovníky popisných termínů) a metadaty na stránkách tvořit základní rámec.

9.6 VYHLEDÁVAČE A HLEDANÁ TÉMATA: VÝZKUM ZÁJMŮ A DOTAZŮ JEJICH UŽIVATELŮ

9.6.1 Schopnosti uživatelů vyhledávat

Nejužívanější technologie současnosti, kterými vyhledávače jsou, ukazuje i další informace o uživatelích a jejich zájmech. Tři velké studie uživatelského chování ve vyhledávačích – [JANSEN, 1998], [JANSEN, 2000] a [SPINK ET.AL., 2001] poukazují na společné charakteristiky uživatelského chování při vyhledávání, které charakterizuje především neznalost nebo neschopnost uživatelů využít všech možností, které jim vyhledávače nabízí. Tento stav vystihuje studie [SPINK ET.AL., 2001] slovy: *„Lidé tráví stále více času vytvářením, hledáním a využíváním elektronických informací. Jejich interakce s webovými vyhledávači je ale krátká a limitovaná...“*

Všechny tři studie ukazují, že uživatelé zatím nejsou připraveni ochotni strávit vyhledáváním a analýzou výsledků delší čas. Proto vyhledávání zpravidla sestává ze zadání jednoslovného výrazu do vyhledávacího pole a ve výsledcích kontroly několika málo uvedených odkazů, uvedených na prvních pozicích. Je zajímavé, že tento trend potvrzuje již výzkumná studie [POLLOCK, 1997], kde autorky v závěru uvádí, že: *„Vyhledávače by měly jasně vysvětlit koncept hledání na internetu jako proces spíše než jednorázovou událost.“*

67 Přehled různých projektů mobilních agentů je dostupný na http://www.agentlink.org.

68 http://brightplanet.com/products/dqm.asp

69 http://www.botox.com

70 http://www.botql.com

	Spink, et.al. 2001	Jansen, et.al., 2000	Jansen, et.al., 1998
počet dotazů na osobu(během jednoho vyhledávání)			
jeden	48.4%	77.6%	67.0%
dva	20.8%	13.5%	19.0%
tři a více dotazů	31.0%	4.4%	7.0%
počet slov v jednom dotazu			
jedno	26.6%	25.8%	31.0%
dvě	31.5%	26.0%	31.0%
tři	18.2%	15.0%	18.0%
použití pokročilých vyhledávacích technik*	25.0%	20.4%	10.0%
počet stránek, zobrazených ve výsledcích			
pouze první stránka	28.6%	85.2%	58.0%
pouze první dvě stránky	19.0%	7.5%	19.0%

* Jakékoli přidání dalších vyhledávacích operátorů (Booleovské operátory, operátor rozšíření apod.).

Tabulka 11: Využití funkcí vyhledávačů

Přehledová tabulka č. 12 ukazuje trendy ve využívání vyhledávačů na základě dat z výše uvedených studií. Výsledky těchto studií můžeme vzhledem k objemu dat zkoumaných vzorků dotazů a vzhledem k počtu uživatelů (SPINK, ET. AL. (2001) – 1 025 910 dotazů, 200 000 uživatelů; JANSEN (2000) – 153 645 050 dotazů; JANSEN (1998) – 51 473 dotazů) chápat jako reprezentativní.

Výsledky ukazují, že drtivá většina uživatelů se nesnaží klást více dotazů, při hledání nepoužívají jakékoli vyhledávací operátory (Booleova algebra, operátory rozšíření, omezení vyhledávání apod.), a co je nejhorší, věnují velmi malou pozornost zpracování výsledků a zpětné vazbě pro vyhledávač. Více než polovina uživatelů nemusí najít výsledek proto, že se podívají pouze na první dva zobrazené výsledky. Podle těchto studií není využívána ani funkce pro zpětnou vazbu typu *„More like this"*, která umožňuje vyhledávat podle vzoru již nalezené stránky. Ve všech studiích jsou výsledky využití této zpětné vazby, která výrazně zlepšuje výsledky, velmi nízké. Studie Spink, et.al. uvádí **9.7%**, zbývající dvě **5%**, respektive **11%**.

Chyby, kterých se uživatelé při hledání dopouštějí charakterizovala Annabel Pollock [1997] do několika skupin:

- uživatelé očekávají, že počítač bude rozumět dotazu v přirozeném jazyce[71]
- snaha vyjádřit několik dotazů najednou
- definování příliš obecných nebo příliš specifických termínů
- překlepy a gramatické chyby.

71 Toto je příčinou obliby vyhledávače *Ask Jeeves*, který umožňuje klást jednoduché dotazy v přirozeném jazyce.

Důležitým zjištěním této studie je i to, že uživatelé často postrádají nezbytný obecný přehled, který jim umožní porozumět nalezeným výsledkům. Je zde uveden příklad člověka, který hledal informace o cestování v USA, a nepokládal stránky společnosti *Lonely Planet* za relevantní, protože nevěděl, že se jedná o vydavatele cestovních průvodců.

Výsledky všech studií potvrzují předpoklady pro využívání vyhledávačů – technické řešení je na uspokojivé úrovni a největším nedostatkem stále zůstává příprava uživatelů. Snížení chybovosti ve vyhledávání by mohly přinést i asistenční nástroje, jaké popisuje Amanda Spink [2002]: *„Potřebujeme novou generaci vyhledávacích nástrojů na internetu, založených na větším porozumění lidských informačních potřeb. Tyto nástroje by pomáhaly s tvorbou a úpravou dotazu, opravou překlepů a dalšími analytickými problémy, které omezují schopnost uživatelů najít informaci, kterou potřebují."*

9.6.2 Vyhledávaná tématika

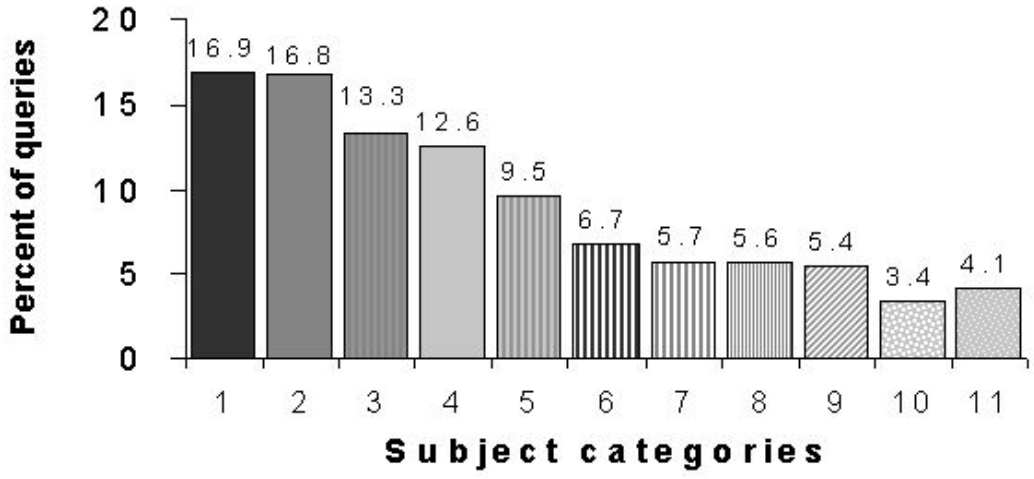
Technická data, která vyhledávače shromažďují (tzv. logy), umožňují také analýzu hledaných témat. Údaje o vyhledávané tématice jsou zajímavým zdrojem informací o vývoji uživatelské komunity, jejích zájmů a informačních potřeb, ale také o charakteru využívání internetu jako informačního zdroje.

Studie [SPINK ET.AL., 2001] analyzovala rozložení témat 1 025 910 dotazů z roku 1997. Témata byla analyzována a sdružena do jedenácti kategorií (viz. obrázek 37).

Nejhledanější témata jsou z oblasti zábavy a volného času (16.9%). Těsně za ní následuje kategorie sex a pornografie (16.8%). Autoři studie upozorňují, že ne všechny dotazy v této kategorii se nutně týkají pornografie, ale mnoho z nich se týká jiných – především zdravotních aspektů sexuality. Studie [SPINK ET.AL., 2001] dále upozorňuje na to, že i když se na první pohled může zdát, že pornografie je hlavním tématem dotazů ve vyhledávačích, sexuální tématika se objevuje pouze v každém šestém dotazu a, jak je výše uvedeno, jedná se i o stránky se zdravotní tématikou.

Komerčně orientované stránky jsou předmětem cca 13% dotazů, přibližně 10% dotazů je z oblasti zdraví a přírodních věd. Informace o lidech a společnosti tvoří pouze 12,4%, což může potvrzovat změnu v charakteru využívání internetu od komunikace k zábavě. V komentáři autoři této studie připouští, že i v tomto rozdělení je určitá míra zobecnění, nicméně přesto by mělo být reprezentativní.

Je zajímavé, že rozložení uživatelských dotazů neodpovídá tematickému rozložení informací na internetu podle studie [LAWRENCE-GILES, 1999]. Tato studie zjistila, že 83% stránek má obsah komerčního charakteru, 6% stránek je zaměřeno na vzdělávání, vědu a výzkum, 3% stránek obsahují informace o zdraví, 2% jsou osobní stránky a pouze 1% stránek bylo zaměřeno na pornografii. V porovnání s touto studií se výsledky studie [SPINK ET.AL., 2001] se výsledky značně liší, ale rozdíly je možné vysvětlit změnami v obsahu stránek či v uživatelských preferencích, ke kterým došlo mezi srpnem 1997 (shromážděny podklady pro studii [SPINK ET.AL., 2001] a únorem 1999 (studie [LAWRENCE-GILES, 1999]).



***Obrázek 34**: Rozložení témat dotazů podle výzkumu Amandy Spink* *Zdroj: SPINK ET.AL., 2001*

Legenda pro tematické kategorie: 1. Zábava, volný čas, 2. Sex, pornografie, 3. Obchod, cestování, ekonomika, 4. Počítače a internet, 5. Zdraví, věda, 6. Lidé, místa, věci, 7. Společnost, kultura, etnika, náboženství, 8. Vzdělávání, společensko-vědní obory, 9. Umění, 10. Vláda a státní správa, 11. Neznámé, nelze analyzovat.

Názor o změně uživatelských preferencí podporuje i studie [SPINK ET. AL., 2002], která porovnávala uživatelské dotazy v letech 1997, 1999 a 2001 ve vyhledávači *Excite*. Ve výsledné tabulce (č. 13) je zřetelný posun v tematickém zastoupení dotazů. Významnou změnou je také zvýšení podílu dotazů v jiných jazycích než v angličtině, které naznačuje změnu v poměru počtu uživatelů internetu z dalších zemí. Vyhledávaná tématika se lokálně liší a je ovlivněna společenským a kulturním děním. Jako důkaz tohoto tvrzení může sloužit služba *Google Zeitgeist*[72], která analyzuje dotazy uživatelů v 16 velkých zemích světa (např. USA, Francie, Německo, Rusko nebo Japonsko) v týdenních a měsíčních intervalech.

Srovnání s nejvýznamnějšími událostmi roku 2004 a statistikou dotazů ukazuje, že na prvních příčkách v těchto přehledech byly vyhledávány nejvýznamnější zahraniční či vnitřní politické a kulturní události v roce 2004, (např. prezidentské volby v USA, válka v Iráku, teroristický útok v Beslanu). Ve Velké Británii tak byl nejčastějším dotazem v roce 2004 „bbc news", který patrně souvisí s událostmi kolem obvinění premiéra Tonnyho Blaira zpravodajskou společností BBC z falšování důkazů o iráckých zbraních hromadného ničení.

rank	1997 Excite data set (2,414 queries)	1999 Excite data set (2,539 queries)	2001 Excite data set (2,453 queries)
1	19.9% Entertainment or recreation	24.5% Commerce, travel, employment	24.7% Commerce, travel, employment
2	16.8% Sex and pornography	20.3% People, places, or things	19.7% People, places, or things
3	13.3% Commerce, travel, employment	10.9% Computers or Internet	11.3% Non-English or unknown
4	12.5% Computers or Internet	7.8% Health or sciences	9.6% Computers or Internet
5	9.5% Health or sciences	7.5% Sex and pornography	8.5% Sex and pornography
6	6.7% People, places, or things	7.5% Entertainment or recreation	7.5% Health or science
7	5.7% Society, culture, ethnicity, or religion	6.8% Non-English or unknown	6.6% Entertainment or recreation
8	5.6% Education or humanities	5.3% Education or humanities	4.5% Education or humanities
9	5.4% Performing or fine arts	4.2% Society, culture, ethnicity, or religion	3.9% Society, culture, ethnicity, or religion
10	4.1% Non-English or unknown	1.6% Government	2.0% Government
11	3.4% Government	1.1% Performing or fine arts	1.1% Performing or fine arts

Tabulka 12: Distribuce dotazů do tematických kategorií

Zdroj: SPINK ET. AL., 2002

Zpravodajství přesto není hlavním vyhledávaným tématem, jak to dokládá druhý nejhledanější výraz hledaný ve Velké Británii – „big brother". Toto slovní spojení označuje populární televizní soutěž a podtrhuje změnu ve využívání internetu jako zábavního média. V celkovém hodnocení dotazů na vyhledávači Google za rok 2004 z celého světa převládají celebrity (Pamela Anderson) a zábava v různých formách (filmy Harry Potter nebo Pán prstenů/The Lord of the Rings, počítačové hry, hudba ve formátu mp3 apod.).

[72] Viz http://www.google.com/press/zeitgeist.html. Název by měl znamenat *„Duch doby"*.

9.6.3 Společenské sítě (Social networks) a linkovací systémy

Přes veškerou snahu zatím není funkční koncept sémantického webu, který předpokládá rapidní zlepšení v komunikaci mezi lidmi a stroji. Vyhledavače mají stále své nedostatky, a tak v poslední době roste význam tzv. „společenských sítí“ (social networks), které umožňují lidem vytvářet nové vztahy na základě sdílených hodnot a zájmů. Ve své podstatě se jedná o návrat k jednomu z nejstarších záměrů internetu – kontaktovat lidi mezi sebou a na základě osobní komunikace/vztahu efektivně vyměňovat informace mezi sebou. Tento model po uživateli nevyžaduje znalost dotazovacích jazyků ani speciálních rešeršních postupů a zároveň umožňuje účelnou výměnu často jiným způsobem nedostupných informací.

Společenské sítě (social network) se zároveň posunují od čistě zábavných sítí (např. www.friendster.com) do obchodních a pracovních sítí (např. www.xing.com), který jen potvrzuje vzrůstající komerční potenciál internetu. Obchodní a pracovní sítě, jako zmíněný příklad, umožňují lidem registraci podle jejich oboru, firmy a pozice a následně - na základě těchto údajů vyhledávání pro všechny uživatele systému. Potenciál těchto sítí je obrovský vzhledem k tomu, že tento typ údajů (a jejich objem) lze vyhledávači nalézt mnohem obtížněji a pomaleji.

Varianta společenských systémů jsou tzv. „linkovací systémy“ (např. www.digg.com, www.reedit.com, www.linkuj.cz). Ty umožňují uživatelům vzájemně sdílet odkazy na zajímavé webové stránky a hodnotit jejich kvalitu. Tímto procesem je zaručena vyšší relevance než u vyhledávačů, které mohou hodnocení uživatelů odvozovat pouze nepřímo podle návštěvnosti stránek a frekvence odkazů na ně (např. Google). Uživatelská komunita pečlivě hlídá kvalitu odkazů a není divu, že relevance odkazů je řádově vyšší než je tomu zpravidla v případě vyhledavačů. Rozsahem odkazů linkovací systémy sice nemohou vyhledavačům konkurovat, přesto se jedná o zajímavý doplněk, který uživatelům vyhovuje.

9.7 DALŠÍ VÝVOJ A TRENDY V OBLASTI VYHLEDÁVAČŮ

Je pravděpodobné, že vyhledávače si i v budoucnu udrží pozici vůdčí technologie pro vyhledávání informací na internetu. Tato technologie, kterou lze označit za obdobnou, jež používají systémy automatizované klasifikace a kategorizace, bude nadále prostředkem pro zpracování alespoň částí stále se zvětšujícího prostoru internetu. Vedle vývoje vyhledávací technologie můžeme pozorovat další vývojové trendy, které lze rozdělit do několika skupin:

Personalizace vyhledávače

Vyhledávač již není pouhým nástrojem pro vyhledávání informací, ale spíše výchozím místem pro jakoukoli práci na internetu. Tím, že si vyhledávač mnozí nastavují jako výchozí – domovskou stránku, získávají vyhledávače na významu. Firmy, které tyto vyhledávače provozují, si tuto pozici uvědomují a snaží se proto vylepšit dostupné vlastnosti a spektrum služeb tak, aby uživatel našel vše, co potřebuje na jejich serveru a nemusel hledat jinde. Tyto nabídky zahrnují i personalizaci vyhledávacího rozhraní, nastavení speciálních funkcí (např. integrace plánovacího kalendáře k poštovnímu účtu na serveru *Yahoo!*) nebo informace, které zohledňují fyzické bydliště uživatele (např. obchodní porovnávací systém *Bizrate* dokáže na základě zadané adresy v USA vypočítat poštovné u zboží, které zákazníka zajímá).

Ze stejného principu vychází i jazykové mutace těchto vyhledávačů, které se snaží více orientovat na lokální zdroje, a konkurovat tak národním vyhledávačům. Ze stránek vyhledávačů se stávají spíše univerzální portály se širokou nabídkou služeb, jak to ukazují příklady *Google* nebo *Yahoo!*.

Vyhledávací lišty pro použití v prohlížeči (toolbars)

Řada vyhledávačů se snaží o to, aby uživatel využíval jejich služeb co nejvíce. Jednou z možností, které uživatelům nabízí jsou proto tzv. „vyhledávací lišty” (toolbars). Jde o malý program, který se integruje do okna prohlížeče a ze kterého lze přímo vyhledávat. Tyto lišty nabízí v rámci konkurenčního boje řada velkých (např. *MSN Toolbar suite*, *Google Toolbar*), ale i menších vyhledávačů (např. *Dogpile Search Toolbar, A9 Toolbar*). Uživatel, který má takovou lištu nainstalovanou, bude pravděpodobně vyhledávat především pomocí této lišty, což pro vyhledávač znamená udržení zákazníka a budování vlastní pozice na trhu. Tyto lišty dnes nabízí řadu doplňkových vlastností od vyhledávání ve „zlatých stránkách” a seznamech firem v USA, přes vyhledávání kurzů akcií až po doplňkový software pro blokování samovolně se otevírajících oken prohlížeče (tzv. pop-up window).

Tento trend vývoje může je odpovědí na praxi společnosti Microsoft a jejího (v současnosti pravděpodobně nejpoužívanějšího) operačního systému Windows a prohlížeče Internet Explorer, který obsahuje také integrovanou funkci vyhledávání na serveru MSN, který tato společnost vlastní.

Doplňkové vyhledávací programy (personal computer search)

Vyhledávací lišty byly prvním krokem k přemístění vyhledávacího rozhraní z webových stránek na počítač uživatele. Hlavním cílem společností, které vyhledávače vlastní, je snaha o globální řešení všech informačních potřeb uživatele svojí technologií. Na konci roku 2004 se proto začaly objevovat aplikace pro personalizované vyhledávání na disku pevného počítače uživatele (tzv. desktop search). Tyto aplikace (např. Copernic Desktop Search, Google Deskbar, od ledna 2005 také Yahoo! Desktop Search)kombinují vyhledávání souborů na pevném disku, osob v adresáři nebo emailů s vyhledáváním webových stránek. Technologie jedné firmy tak může zajistit kompletní servis pro vše, co uživatel potřebuje. V této oblasti můžeme očekávat velmi tvrdý boj, protože tyto programy mohou ovlivnit pozici současných vyhledávačů na trhu [Delaney, 2004]. Především ze strany vyhledávače Google to lze považovat za protiútok vůči společnosti Microsoft, která se snaží integrací svého vyhledávače s operačním systémem Windows ohrozit jeho vůdčí postavení na tomto poli.

Důvodem souboje největších vyhledávačů o počítač uživatele je také možný příjem z reklamy. Pokud se budou indexovat dotazy, které tento uživatel klade, je možné připravit cílenou reklamu právě pro tohoto uživatele. Je možné, že tyto programy budou dostupné jako tzv. *adware* tj. software zdarma, v jehož liště se objevuje reklama. Jestliže se tato technologie uchytí, je možné očekávat i znásobení zisků z reklamy (viz prohlížeč Opera).

Jak ukazuje tento vývoj, konkurenční boj mezi vyhledávači se netýká pouze vlastní vyhledávací technologie, ale ve velké míře také marketingu. Nadstandardní služby uživatelům (doplňkové služby e-mailu zdarma apod.), personalizace vyhledávače nebo nabídka doplňkového softwaru jsou cestou, jak zvýšit podíl na trhu a následně i vydělat. Komerční prostředí se tak v případě vyhledávání stává hnacím motorem pro vývoj, který je pro oblast pořádání informací nesmírně důležitý.

Zatím nelze vyhledávače považovat za optimální řešení pro pořádání informací na internetu. Jak uvádí Steve Steinberg [1996]: „dokonce ani lidé nejsou schopni rozhodnout jaká informace je relevantní prozadanou otázku. Pokoušet se, aby to dělal počítač je skoro nemožné.”

10 ZÁVĚR

Matka: Co vy víte, co je pořádek!

Kornet: Dát věci tam, kde byly.

Petr: Dát věci, tam kde mají být.

Matka: Ba ne. Dát věci tam, kde je jim dobře, ale tomu nerozumíte.

Karel Čapek: Matka

Problematika pořádání informací je složitá a neexistuje jednoduché řešení, které by mohlo jednou provždy diskuze o tomto tématu uzavřít. Důvodem je fakt, že pořádání informací je proces, nikoli jednorázově řešitelný problém. Obecné problémy pořádání informací zůstávají pořád stejné. Teoretickým řešením je vyvážit všechny prvky, které ovlivňují tento proces, zejména:

- zajištění přístupu k informacím
- stejná úroveň a kvalita znalostí uživatelů
- řešení individuálních informačních potřeb univerzálním systémem
- ekonomická rentabilita nákladů, investovaných do pořádání informací.

K těmto faktorům se v případě internetu přidávají ještě další, zmíněné v kapitole 2, například velmi malá životnost informací nebo exponenciální nárůst nových informací. Fenoménem, který se ukazuje v současné době je **změna v komunikační roli internetu**. Původní představy o internetu jako celosvětové knihovně znalostí, které byly mnohokrát zmíněny v různých koncepcích (např. Al Gore) se ukázaly jako nepřesné. Internet do značné míry knihovnou je, nicméně funguje na základě zcela odlišných principů než se očekávalo.

Poskytování informací se nestalo hlavní využívanou službou internetu. Tou se stala komunikace. Tento trend popisuje Graham Spencer, jeden ze zakladatelů firmy Excite, již v roce 1996 slovy: „[internet] je o lidech, hledajících lidi, ne o lidech, hledajících informace“ [Steinberg, 1996]. Změnu dokládají různé statistiky využívání různých služeb internetu, které publikují různé renomované agentury.

Podle grafů společnosti Nielsen/Netrankings (viz obrázek 1.4) jsou nejvyužívanějšími stránkami e-mailové portály (přístup k e-mailu přes webové rozhraní) a stránky zaměřené na obchod či na zábavu. Internet nadále zůstává především komunikačním kanálem, ale zároveň se stává obchodním prostředím. Idea univerzální knihovny poznatků se tak neuskutečnila a zatím není ani důvod domnívat se, že k tomu dojde v budoucnosti.

10.1 PŘÍSTUP K INFORMACÍM A JEHO CENA

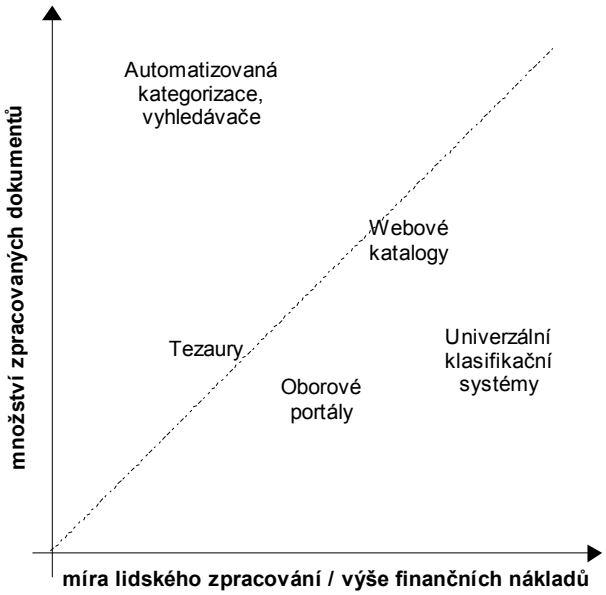
Hlavním problémem bude vždy poskytování kvalitního obsahu – jak u primárních informací (stránky, soubory samotné), tak v systémech pro jejich třídění. Vše je a bude ovlivňováno zajištěním financování pro tyto projekty. Na obrázku 9.1 jsou pro přibližné srovnání uvedeny v grafu finanční náklady (množství lidské práce) některých populárních technologií pro pořádání informací v prostředí internetu ve srovnání s množstvím zpracovaných dokumentů. Podle způsobu financování můžeme projekty rozdělit do několika skupin:

specializované projekty pro určitou skupinu

komunitní projekty, založené na dobrovolné spolupráci

komerční projekty s nepřímým financováním

komerční projekty, poskytující informace za přímou úhradu



Obrázek 35: množství zpracovaných dokumentů ve srovnání s náklady na zpracování

Specializované projekty pro určitou skupinu jsou financovány z vlastních prostředků organizace nebo prostřednictvím různých grantových programů. Tyto projekty slouží zpravidla pouze oborové komunitě, která je nadále financuje (např. Engineering Village společnosti Engineering Information).

Jako překvapivě životaschopné se zatím ukazují komunitní projekty, založené na dobrovolné spolupráci. Webové katalogy jako Open Directory, ZEAL nebo encyklopedie Wikipedia jsou příkladem, že taková spolupráce může přinést mnohem kvalitnější produkt než ten, který je založen na komerční bázi, ať už s přímou úhradou od uživatele (viz. encyklopedie Encarta) nebo financovaný nepřímo (viz webový katalog Yahoo!).

Komerční projekty s nepřímým financováním jsou nejpoužívanějším modelem pro zpřístupňování obsahu na současného internetu. Práce a náklady projektu jsou zaplaceny nejčastěji ve formě reklamy (např. sponzorované odkazy ve vyhledávačích Google, MSN, Yahoo!), nebo placenou nabídkou dalšího zboží a služeb k zakoupení (např. About.com, BizRate.com). Je pravděpodobné, že i nadále bude tento typ projektů nejpočetnější – uživatel je motivován kvalitní službou, za kterou nemusí přímo nic platit, provozovatel služby má zajištěný zisk jiným způsobem.

Komerční projekty, poskytující informace za přímou úhradu jsou nejstarším používaným modelem financování. Toto financování se objevuje se u specifických služeb (např. marketingové informace – Nielsen/Netrankings), nebo při poskytování strategických informací, především z oblasti ekonomiky a financí (Dow Jones, Bloomberg apod.). Společným znakem těchto služeb je jejich obchodní potenciál a věrohodnost informace – uživatel získává ověřenou informaci, na jejímž základě může odpovědně rozhodnout. Tuto záruku věrohodnosti v projektech s jiným modelem zpřístupňování informací zpravidla nezíská.

Andy Powell[73] z centra UKOLN upozorňuje na zajímavý paradox: náklady na projekty předurčují jejich budoucnost. Financující organizace se ptají jaký je konkrétní přínos těchto projektů, tedy jaký je poměr mezi investovanými prostředky a získanou hodnotou. Projekty různých oborových portálů jsou velmi pracné a nákladné z hlediska financí i lidské práce. Přesto obsahují pouze malý zlomek kvalitních dokumentů – i když velmi kvalitně popsaných a zatříděných. Klasifikace a výběr dokumentů, které jsou v těchto projektech popisovány, je stále procesem individuálního přiřazování hodnoty těmto informacím.

Uživatelé zatím nejsou schopni ocenit tuto kvalitu, namísto toho ocení spíše široký záběr dokumentů, ze kterých jsou sami schopni vybrat ty, které jsou relevantní – chtějí definovat vlastní kritéria pro kvalitu. Celý problém se tak přesouvá do oblasti nalezení kompromisu mezi kvalitou zpracování a jejími náklady a zároveň mezi individuálním a objektivním vnímáním hodnoty dokumentu ve vztahu k informační potřebě.

Jak uvádí [JANES, 2003] „nikdo nemá odpovědnost za zakládání veřejné knihovny pro internet. Jiné knihovny mají komunity nebo instituce, kterým slouží; na oplátku za služby které tím poskytují jsou také finančně zajištěny“. Vystihuje tím i základní orientaci pro pořádání informací v budoucnosti: důležité jsou dílčí projekty pořádání informací v malých tematických komunitách; obecné třídění má význam pouze jako obecný přehled, který má tato třídění propojovat pouze rámcově.

10.2 INDIVIDUÁLNÍ INFORMAČNÍ POTŘEBY

Pořádání informací musí alespoň částečně odhlédnout od dosud vytvořených systémů a akceptovat pragmatický přístup k pořádání informací. Uživatele nezajímá technologie ani teoretické podklady třídění, má vždy pouze jediný cíl – najít požadovanou informaci.

Tuto skutečnost dokumentuje Hanne Albrechtsen [1998] na příkladu Ballerup Public Library, kde používají upravenou verzi třídění pro děti. V tomto třídění jsou kategorie pojaty následovně:

1. počítače
2. astronomie, příroda a životní prostředí, zvířata
3. první lásky, horoskopy, problémy mladých
4. koně
5. zábava, dobrodružství, humor
6. fantasy, science fiction
7. knihy, které se snadno čtou.

Jak autorka sama uvádí „Z pohledu disciplín nebo ze sémantického pohledu je oddělení předmětů jako zvířata a koně nesprávné nebo nelogické. Dětem nicméně toto pořádání vyhovuje.“ [Albrechtsen, 1998]. Zdejší knihovna pouze pragmaticky zhodnotila zájmy dětí a přizpůsobila organizaci celé kolekce jejich vnímání a jejich potřebám.

Problémem mnoha klasifikačních systémů je, že se snaží zpracovat svoji tématiku systematicky. Uživatel přitom chce pouze takový systém, ve kterém se bude dobře orientovat, což vždy nemusí být totéž (např. struktura kategorií Yahoo!).

Dosud nezodpovězenou otázkou zůstává, zda jsou pro klasifikaci těchto dokumentů skutečně odpovídající staré systémy třídění jako DDC, LCC nebo MDT. Automatizovaná klasifikace prokázala funkčnost takového modelu, nicméně nutně neprokazuje kvalitu třídícího systému. Z tohoto hlediska by proto bylo zajímavé stejný model aplikovat na nové systémy třídění (webové katalogy) a porovnat výsledky.

Odpověď na otázku vhodnosti třídící soustavy nakonec zodpoví sami uživatelé. Zatím se zdá, že nové systémy třídění dokumentů již zaujaly neotřesitelnou pozici a klasické knihovnické třídění je pro oblast internetu téměř mrtvým konceptem – „není překvapením, že [stará klasifikační schémata] jsou slabá v klasifikaci znalostí v „nově“ založených disciplínách, jako jsou genetika nebo elektrotechnika. Nejdůležitější je, že knihovnické klasifikační systémy jsou svázány omezeními, která v digitálním světě neexistují. Fyzicky může být kniha umístěna pouze na jednu polici, dokument

[73] Osobní rozhovor 1.10.2004 Bath: UKOLN

v digitální podobě může být umístěn do několika kategorií za cenu pouhých několika byte“ [Steinberg, 1996]. To neznamená, že principy a teoretické podklady klasických knihovních třídění jsou bezcenné. Jsou zkušenostmi a podkladem pro další práci. Komplexní systém čistě hierarchického systému je zastaralý, a je potřeba jej inovovat.

10.3 JE POŘÁDÁNÍ INFORMACÍ POTŘEBNÉ I DO BUDOUCNOSTI?

Vývoj informačních technologií je velmi rychlý a možnosti, které se tím nabízejí vedou také k domněnkám o tom, že problematika pořádání informací může být vyřešena touto technologií bez potřeby teoretického zázemí. Traugott Koch [1998] například jako alternativu k univerzálním klasifikačním schématům uvádí nahrazení plnotextovým vyhledáváním a indexací slov. Jakkoli se tento závěr může zdát správný, nezohledňuje minimálně následující:

- nároky na analýzu významu textu v přirozeném jazyce
- schopnost uživatelů klást dotazy

V předchozích kapitolách (viz kapitola 7 – „Automatizovaná klasifikace a kategorizace“) jsou zmíněny problémy, spojené s automatizovaným zpracováním přirozeného jazyka. Obsahový popis dokumentu bude pravděpodobně vždy závislý na zpracování člověkem, který je schopný analyzovat jeho hlavní téma. Strojové zpracování zatím není schopné určit nosné téma a odlišit jej od témat vedlejších. Je otázkou, zda strojové zpracování k tomuto stádiu někdy dospěje, neboť tato problematika je úzce spojena se sémantikou – přiřazování významu slovům.

Myšlenka indexování slov a následného plnotextového vyhledávání problém porozumění jazyku neřeší, ale přesunuje ho na vyhledávací systém a na uživatele. Protože vyhledávací systém nebude nikdy dokonalý, tento koncept vyžaduje schopnost uživatelů sestavovat dotazy. Jde tedy o náročnější formu navigace než při prohlížení hierarchie (viz kapitola 6, s. 73).

Pro srovnání rozdílů mezi konceptem vyhledávání klíčových slov (technologický přístup) a prohlížením předmětů ve struktuře (přístup pořádání informací), by mohla být přínosná analýza úspěšnosti uživatelů při hledání oběma způsoby. Jedině tak by bylo možné prokázat, že koncept vyhledávání v plném textu je rovnocennou náhradou systémům pro pořádání informací pomocí různých struktur (např. webové katalogy, nebo univerzální klasifikace).

Pravděpodobnější variantou pro budoucí vývoj je proto kombinace automatizovaného zpracování s dalšími metodami, která by usnadnila přiřazení významu indexovaným slovům. Různé slovníky, rejstříky a ontologie tak významně usnadní pochopení významu a kontextu informace. Jak různé projekty ukazují, kombinace obou zmíněných způsobů pro pořádání informací je velmi perspektivní (např. metadata a specializované vyhledávače – harvestory viz kapitola 4.12).

10.4 TRENDY PRO BUDOUCÍ VÝVOJ POŘÁDÁNÍ INFORMACÍ

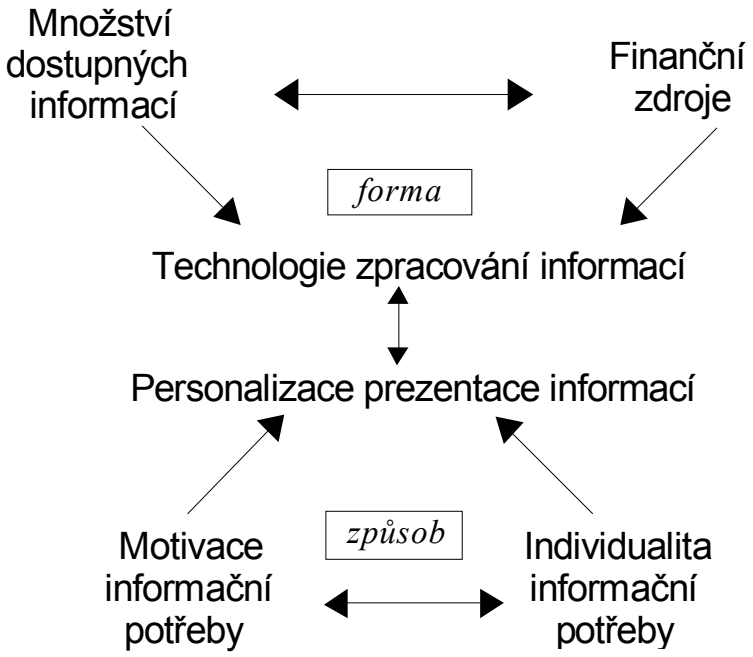
Vzhledem ke stále rostoucí publikaci dokumentů (viz kapitola 1) je zpracování člověkem stále nákladnější a přináší menší efekt. Budoucnost proto musí kombinovat automatizované metody pro zpracování dokumentů společně s prezentací jejich výsledků do podoby, která je pro lidské vnímání rychle pochopitelná.

Organizace informací by se tak měla zaměřit na větší porozumění procesů lidského poznání a vnímání informací, které zůstávají relativně neměnné, a více se soustředit také na kombinaci prezentace výsledků při automatizovaného zpracování dokumentů (např. automatizovaná kategorizace). Vzrůstá tak význam vizualizace a dalších pomocných technik prezentace, které zlepšují vnímání a pochopení informací

Na obrázku 39 jsou znázorněny vztahy mezi vybranými faktory, které ovlivňují pořádání informací. S ohledem na dosavadní vývoj informačních technologií včetně zpracování přirozeného jazyka se zdá, že první část schématu (forma – technologie zpracování informací) bude stále více spíše doménou metod výpočetní techniky a umělé inteligence. Záběrem oboru pořádání informací se bude více stávat část druhá (způsob – personalizace prezentace informací), pro kterou se budou využívat všechny dostupné možnosti, které zpřehlední jejich prezentaci a přesný přenos znalostí.

Problematika pořádání informací se přesunuje od konceptu fyzického umístění jednotky ke koncepci uložení a opětovného vyhledání. Problémy spojené například s tvorbou notací se tak odsunují do pozadí. Místo nich vznikají nové úkoly a výzvy které souvisí především s identifikací dokumentu a jeho obsahu, technologií přesného vyhledání a již zmíněné prezentace výsledků.

V budoucnu je pravděpodobné, že další technologie se budou snažit zohlednit trend personalizace vyhledávání a



Obrázek 36: Faktory, ovlivňující pořádání informací

pořádání informací. Lze tedy očekávat přechod od globálních řešení pořádání (např. vyhledávače) ke specializovaným či přímo osobním řešením.

Problematika pořádání informací je stará jako lidstvo samo. Jejímu řešení bylo věnováno mnoho úsilí, ale doposud se nikomu nepodařilo nalézt definitivní řešení. Klíčem k pochopení problematiky je, že se nejedná o řešení pracovního problému, ale o proces řešení. Z historie je jasné, že se tato problematika bude vyvíjet i nadále, dokud bude člověk pro svou práci využívat informace.

11 POUŽITÉ ZDROJE

11.1 KAPITOLA 1: SPECIFIKA INTERNETU PRO POŘÁDÁNÍ INFORMACÍ

Bergman, Michael K. *The Deep Web: Surfacing the hidden value* [online]. 2001 [cit. 2006-10-21]. Dostupné z URL: http://www.brightplanet.com/images/stories/pdf/deepwebwhitepaper.pdf.

CNN. 2003. Dell saying bye to floppy disk drives. *CNN* [online]. Feb 07 2003, [cit. 2004-07-12]. Dostupné z URL: http://www.cnn.com/2003/TECH/ptech/02/07/dell.floppydisks.reut/index.html.

ClickZ Stats staff. 2005. *Population Explosion!* [online]. 2005 [cit. 2005-02-08]. Dostupné z URL: http://www.clickz.com/-stats/sectors/geographics/article.php/5911_151151.

Computer Industry Almanac. Worldwide Internet Users Top 1 Billion in 2005 [online]. January 4, 2006 [cit.2006-11-04]. Dostupné z URL: http://www.c-i-a.com/pr0106.htm.

Computer Industry Almanac Inc. 2001. *U.S. has 33% share of internet users worldwide zear-end 2000* [online]. 2001 [cit. 2005-02-08]. Dostupné z URL: http://www.c-i-a.com/pr0401.htm.

Dahlberg, Ingetraut. 1974. *Grundlagen universeller Wissensordnung Probleme und Möglichkeiten eines universalen Klassifikationssystemes des Wissens.* Pullach bei München:Verlag Dokumentation, 1974. xviii, 366 s. 3-7940-3623-9. Kapitola 7.1.3 Die ''Zwei-Systeme Theorie der Klassifikation'', s. 275-277.

Graphic, Visualization, & Usability Center (GVU). *GVU's 10th WWW user survey* [online]. 1998 [cit. 2005-02-08]. Dostupné z URL: http://www.cc.gatech.edu/gvu/user_surveys/survey-1998-10/.

Grey, Matthew. 1996. *Internet statistics: Growth and usage of the web and the internet* [online]. 1996 [cit. 2005-02-08]. Dostupné z URL: http://www.mit.edu:8001/people/mkgray/net/index.html.

Guilli, A 1 - Signorini, A. The indexable web is more than 11,5 bilion pages [online]. 2005 [cit. 2006-10-21]. Dostupné z URL: www.cs.uiowa.edu/~asignori/web-size/size-indexable-web.pdf.

International Telecommunication Union (ITU). Internet indicators: subscribers, users and broadband subscribers [online]. 2008 [cit.2008-11-02]. D Dostupné z URL: http://www.itu.int/ITU-D/ icteye/Reporting/ ShowReportFrame.aspx? ReportName=/ WTI/InformationTechnologyPublic& RP_intYear=2007&RP_intLanguageID=1

Jaenecke, Peter. 1994. To what end knowledge organization? Knowledge Organization, 1994, vol. 21, no.1, s. 3-11.

Kiel, Ewald. 1994. Knowledge organization needs epistemological openess. A reply to Peter Jaenecke. Knowledge Organization, 1994, vol. 21, no.3, s. 148-152.

Lavoie, Brian F. - O'Neill, Edward T. - Bennett, Rick. Trends in the Evolution of the Public Web 1998 - 2002. D-Lib Magazine [online]. April 2003 roč. 9, č. 4, Dostupné z URL: http://www.dlib.org/dlib/april03/lavoie/04lavoie.html.

Lawrence, Steve – Giles, C. Lee. 1999. Accessibility of information on the web. Nature, 1999, vol. 400, no. 8 July 1999, s. 107-109.

Madden, Mary. Internet Penetration and Impact [online]. 4/26/2006 [cit.2006-11-04]. Dostupné z URL: http://www.pewinternet.org/PPF/r/182/report_display.asp.

Nielsen/Netrankings. 2001. Lowest income surfers are the fastest growing group on the web [online]. 13 March 2001 [cit.2005-01-26]. Dostupné z URL: http://direct.www.nielsen-netratings.com/pr/pr_010313.pdf.

Nielsen/Netrankings. 2004. China takes prize for world's second largest at home internet population as numbers reach 56.6 million [online]. 03/24/04 [cit. 2004-06-24]. Dostupné z URL: http://www.nielsen-netratings.com/pr/pr_020422_hk.pdf.

11.2 KAPITOLA 2: PŘÍSTUPY K POŘÁDÁNÍ INFORMACÍ NA INTERNETU

Barker, Joe. 2004. *Glossary of internet & web jargon* [online]. 2004 [cit.2005-01-04]. Dostupné z URL: http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Glossary.html.

Dood, David G. 1996. Grass-root cataloging and classification: Food for thought from World wide web subject-oriented hierarchical list. *Library Resources and Technical Services ,* 1996, vol. 40, no.3, s. 275-286.

11.3 KAPITOLA 3: METADATA

Baker, Thomas – Dekkers, Makx. 2003. Identifying metadata elements with URIs. *D-Lib Magazine* [online]. July/August 2003 [cit. 2004-07-30], vol.9, no. 7/8. Dostupné z URL: http://www.dlib.org/dlib/july03/baker/07baker.html.

Baker, Thomas – Powell, Andy. 2004. *Summary of DC Usage board meeting in Shanghai, October 2004* [online]. 2004 [cit. 2004-12-23]. Dostupné z URL: http://dublincore.org/usage/meetings/2004/10/Meeting-summary.shtml.

Berners-Lee, Tim. 1993. *Naming and addressing: URIs, URLs* [online]. W3 Consortium, 1993 [cit. 2004-01-26]. Dostupné z URL: http://www.w3.org/Addressing/.

Burnard, Lou M. – Sperberg-McQueen, M. 2002. *TEI Lite: An Introduction to text encoding for interchange* [online]. 2002 [cit. 2004-12-23]. Dostupné z URL: http://www.tei-c.org/Lite/teiu5_en.html.

Burnard, Lou M. 1995. *Text encoding initiative. 1. introduction* [online]. 1995 [cit. 2004-12-22]. Dostupné z URL: http://www.tei-c.org.uk/Lite/U5-Intro.html.

Christian, Eliot. 2003. *GILS: Overview - ideas behind the GILS approach* [online]. 2003 [cit. 2004-07-06]. Dostupné z URL: http://www.gils.net/overview.html.

DCMI. 1995. *About the initiative* [online]. 1995 [cit.2004-06-30]. Dostupné z URL: http://dublincore.org/about/.

DCMI. 1998. *Dublin Core metadata for resource discovery: Request for comments: 2413.* 1998 [cit. 2004-06-30]. Dostupné z URL: http://www.ietf.org/rfc/rfc2413.txt.

DCMI. 2003. Dublin Core metadata element set, Version 1.1: Reference description [online]. 2003 [cit. 2004-06-30]. Dostupné z URL: http://dublincore.org/documents/dces/.

Feldman, Susan. 2003. *Special IDC report portals magazine* [online]. 2003 [cit. 2004-07-15]. Dostupné z URL: http://www.portalsmag.com/print/default.asp?ArticleID=5286 .

GILS. 1997. *Global information locator service* [online]. 1997 [cit. 2004-07-07]. Dostupné z URL: http://www.g7.fed.us/gils.html.

Getty Research Institute. 2000. Introduction to metadata: Pathways to digital information. Metadata standard crosswalks [online]. 2000 [cit. 2004-06-17]. Dostupné z URL: http://www.getty.edu/research/conducting_research/standards/intrometadata/3_crosswalks/.

Hakala, Juha et. al. 1998. The Nordic metadata project: Final report [online]. Helsinki University Library, 1998 [cit. 2004-06-30]. Dostupné z URL: http://www.lib.helsinki.fi/meta/nmfinal.htm.

Hodge, Gail. 2001. *Metadata made Simpler: A guide for libraries* [online]. 2001 [cit. 2004-06-18]. Dostupné z URL: http://www.niso.org/news/Metadata_simpler.pdf.

ISOC. 1998. RFC2396: Uniform Resource Identifiers (URI): Generic syntax [online]. The Internet Society, 1998 [cit. 2004-01-26]. Dostupné z URL: http://ftp.ics.uci.edu/pub/ietf/uri/rfc2396.txt.

IPT. 2004. *Introduction to NewsML* [online]. International Press and Telecommunications Council, 2004 [cit. 2004-07-16]. Dostupné z URL: http://www.newsml.org/pages/intro_main.php.

Lawrence, Steve – Giles, C. Lee. 1999. Accessibility of information on the web. *Nature,* 1999, vol. 400, no. 8 July 1999, s. 107-109.

Library of Congress. 2003a. *METS: An overview & tutorial* [online]. Library of Congress help desk, 2003 [cit.2004-07-06]. Dostupné z URL: http://www.loc.gov/standards/mets/METSOverview.v2.html.

Library of Congress. 2003b. *Development of the Encoded archival description DTD* [online]. Library of Congress, 2003 [cit. 2004-07-08]. Dostupné z: URL: http://www.loc.gov/ead/eaddev.html.

Lynch, Clifford. 1998. Identifiers and their role in networked information applications. *Bulletin of the American Society for Information Science and Technology,* Dec 1997/Jan 1998, vol. 24, no.2, s. 17-21.

Miller, Eric. 1998. An introduction to the Resource Description Framework. *D-Lib Magazine* [online]. 1998, [cit. 2004-06-12]. Dostupné z URL: http://www.dlib.org/dlib/may98/miller/05miller.html.

Miller, Ken – Matthews, Brian. 2004. Having the right connections: the LIMBER project. 2004-07-15, vol. 1, issue 8, article no. 37,

Moen, William.1997. Application profile for the Government Information Locator Service (GILS) [online]. 1997 [cit. 2004-12-23]. Dostupné z URL: http://www.gils.net/prof_v2.html.

OCLC.2004. *xISBN* [online]. 2004 [cit. 2004-07-20]. Dostupné z URL: http://www.oclc.org/-research/projects/xisbn/default.htm#.

Open Archives Initiative.2004. *The Open Archives Initiative Protocol for Metadata Harvesting* [online]. 2004 [cit. 2005-01-24]. Dostupné z URL: http://www.openarchives.org/OAI/openarchivesprotocol.html.

PICS. 1996. PICS label distribution label syntax and communication protocols 1996 [cit. 2004-07-18]. Dostupné z URL: http://www.w3.org/TR/REC-PICS-labels#Embedding.

PRISM. 2003. *PRISM: Publishing requirements for industry standard metadata* [online]. 2003 [cit. 2004-07-15]. Dostupné z URL: http://www.prismstandard.org/Pam_1.0/PRISM_1.2h.pdf.

Pitti, Daniel. 1999. Encoded archival description. *D-Lib Magazine* [online]. 1999, roč. 5, no. 11, [cit. 2004-11-27]. Dostupné z URL: http://www.dlib.org/dlib/november99/11pitti.html.

Svoboda, Arnošt. 1997. SGML. *Zpravodaj ÚVT MU* [online]. 1997, roč. VIII, [cit. 2004-07-12]. Dostupné zURL: http://www.ics.muni.cz/cgi-bin/toISO-8859-2/bulletin/issues/vol08num01/svoboda/svoboda.html

Tomaiuolo, Nicholas – Packer, Joan. 1996. An analysis of Internet search engines: Assessment of over 200 search queries. *Computers in libraries,* 1996, vol. 16, no.6, s. 58-62.

Turner, James – Moal, Veronique. 2003. *Welcome to MetaMap* [online]. 2003 [cit. 2004-06-18]. Dostupné z URL: http://mapageweb.umontreal.ca/turner/meta/english/metadata.html.

Vellucci, Sherry L. 1998. Metadata. *Annual review of information science and technology,* 1998, vol. 33, s. 187-222.

W3C. 1996. Request for comments: 1945: Hypertext transfer protocol -- HTTP/1.0 [online]. 1996 [cit. 2004-07-15]. Dostupné z URL: http://www.w3.org/Protocols/rfc1945/rfc1945 .

W3C. 1998. *Metadata Activity Statement* [online]. 1998 [cit. 2004-06-18]. Dostupné z URL: http://www.w3.org/Metadata/Activity.html.

W3C.1999. *HTML 4.01 Specification - W3C recommendation* [online]. 1999 [cit. 2004-07-15]. Dostupné z URL: http://www.w3.org/TR/html4/.

W3C. 2000. *XHTML Basic - W3C recommendation* [online].2000 [cit. 2004-07-15]. Dostupné z URL: http://www.w3.org/TR/xhtml-basic/.

W3C. 2001. URIs, URLs, and URNs: Clarifications and recommendations 1.0: Report from the joint W3C/IETF URI planning interest group [online]. 2001 [cit. 2004-01-26]. Dostupné z URL: http://www.w3.org/TR/uri-clarification/.

W3C. 2004. *PICS statement of principles* [online]. W3 Consortium,2004 [cit. 2004-07-15]. Dostupné z: http://www.w3.org/PICS/principles.html.

Weibel, Stuart.1997. *The 4th Dublin Core metadata workshop report* [online]. 1997 [cit. 2004-06-28]. Dostupné z URL: http://www.dlib.org/dlib/june97/metadata/06weibel.html.

Weibel, Stuart – Jul, Erik – Shafer, Keith. 1995. *PURLs: Persistent Uniform Resource Locators* [online]. 1995 [cit. 2004-07-08]. Dostupné z URL: http://purl.oclc.org/docs/new_purl_summary.html.

XMI. 2002. *XML metadata interchange (XMI)* [online]. 2002 [cit. 2004-07-14]. Dostupné z URL: http://www.oasis-open.org/cover/xmi.html.

11.4 KAPITOLA 4: KLASIFIKAČNÍ SCHÉMATA

AGROVOC thesaurus [online]. 2004 [cit. 2004-11-03]. Dostupné z URL: http://www.icpa.ro/ AgroWeb-/AIC/RACC/Agrovoc.htm.

Albrechtsen, Hanne – Jacob, Elin K. 1998. The dynamics of classification systems as boundary objects for cooperation in the electronic library. *Library Trends,* 1998, vol. 47, no.2, s. 293-303.

American Mathematical Society. *2000 Mathematics Subject Classification* [online]. 2004 [cit. 2004-10-08]. Dostupné z URL: http://www.ams.org/msc/index.html.

Bertolucci, Katherine. Hapiness is taxonomy: Four structures for Snoopy. *Information Outlook,* 2003, vol. 7, no.3, s. 36-45.

Celia, Nada. 2002. *Katalog OKO: Catalogue Oko* [online]. 2002 [cit. 2004-11-13]. Dostupné z URL: http://www.zrc-sazu.si/ko/Katalog_OKO.htm.

Dahlberg, Ingetraut. 1995. Tendencias actuales en Organizacion del Conocimiento. Current trends in knowledge organization.*Proceedings of the first ISKO-Spain conference, Madrid 4-5 Nov 93. Edited by Fco. Javier Garcia Marco. Saragossa, Spain 1995,* s. 7-25.

Dawson, Alan – Simpson, Jan. 1997. How BUBL benefits academic librarians. *Ariadne* [online]. 1997, [cit. 2004-09-01]. Issue 10. Dostupné z URL: http://www.ariadne.ac.uk/issue10/bubl/.

Ellis, David – Vasconcelos Ana. 1999. Ranganatan and the Net: Using facet analysis to search and organise the World Wide Web. *Aslib Proceedings,* 1999, vol. 51, no.1, s. 3-11.

Fast, Karl, et al. 2002. *What Is A Controlled Vocabulary?* [online]. 2002 [cit. 2004-11-17]. Dostupné z URL: http://www.-boxesandarrows.com/archives/what_is_a_controlled_vocabulary.php.

Ibiblio. 2001. *About ibiblio* [online]. 2001 [cit. 2004-10-08]. Dostupné z URL: http://www.ibiblio.org/about.

Journal of Economic Literature Classification system [online]. 2004 [cit.2004-11-01]. Dostupné z URL: http://www.aeaweb.org/journal/elclasjn.html.

Koch, Traugott. 1998. *Possible advantages of using traditional library classification systems in Internet services* [online]. 1998 [cit.2004-09-10]. Dostupné z URL: http://www.lub.lu.se/tk/demos/mex9808b.html.

Koch, Traugott, et al. 1997. *The role of classification in Internet resource description and discovery*. Bath (Great Britain): University of Bath. UKOLN Metadata group, 14 May 1997. Dostupné z URL: http://www.ub.lu.se/desire/radar/reports/-D3.2.3/class_v10.html.

Kwasnik, Barbara H. 1999. The Role of Classification in Knowledge Representation and Discovery. *Library Trends,* 1999, vol. 48, no.1, s. 22-50.

Lipscomb, Carolyn. 2000. *Medical Subject Headings (MeSH)* [online]. 2000 [cit.2004-12-02]. Dostupné z URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?action=stream&blobtype=pdf&artid=35238.

McKiernan, Garry. 1999. Points of view: Conventional and neo-conventional access and navigation in digital collections. *Journal of Internet Cataloging,* 1999, vol. 2, no.1, s. 23-41.

National statistics: History, origins and conceptual basis [online]. 2002 [cit.2004-11-01]. Dostupné z URL: http://www.statistics.gov.uk/methods_quality/ns_sec/history_origin_concept.asp.

NLM. 2004. *NLM gateway fact sheet* [online]. 2004 [cit.2004-11-01]. Dostupné z URL: http://www.nlm.nih.gov/-pubs/factsheets/gateway.html.

NLM. *Summary of changes in 2003 MeSH* [online]. 2003 [cit.2004-10-22]. Dostupné z URL: http://www.nlm.nih.gov/mesh/summ2003.html.

OMNI: Background [online]. 2004 [cit.2004-10-24]. Dostupné z URL: http://omni.ac.uk/about/background.html.

Outcomes of HILT consultation: Draft interim report [online]. 2004 [cit.2004-11-22]. Dostupné z URL: http://hilt.cdlr.strath.ac.uk/Reports/Consultation.html.

Russell, Rosemary – Day, Michael. 2001. HILT: High-level thesaurus. Automated and manual approaches to the provision of thesauri and subject vocabularies [online]. 2001 [cit.2004-11-20]. Dostupné z URL: http://hilt.cdlr.strath.ac.uk/Repo-rts/ Documents/hilt-interface-10.doc.

SBC knowledge network explorer: Blue Web'n homepage [online]. 1995 [cit.2004-11-13]. Dostupné z URL: http://www.kn.pacbell.com/wired/bluewebn/index.cfm.

Steinberg, Steve. 1996. Seek and ye shall find (maybe). *Wired,* 1996, vol. 4, no.5, s. 108-114.

The role of classification in Internet resource description and discovery. Bath (Velká Británie): UKOLN Metadata group. 14 May 1997. Dostupné z URL: http://www.ukoln.ac.uk/metadata/desire/classification/class_3.htm.

The role of classification schemes in Internet resource description and discovery. 2.3. Library of Congress Classification (LCC). Bath (Velká Británie): UKOLN Metadata group, 14 May 1997. Dostupné z URL: http://www.ukoln.ac.uk/metadata/desire/classification/class_4.htm.

WordNet [online]. 2004 [cit.2004-11-26]. Dostupné z URL: http://www.cogsci.princeton.edu/~wn/w3wn.html.

11.5 KAPITOLA 5: WEBOVÉ KATALOGY

About.com. 2007. *Guide Compensatio* [online]. 2007 [cit. 2007-04-01]. Dostupné z URL: http://beaguide.about.com/compensation.htm.

Delaney, Kevin. 2004. Yahoo lets users fine-tune web wearches. *Wall Street Journal,*2004, October 5, s. D. 9.

Gardner, Rachel. Jul 16, 2004. Yahoo! places emphasis on successful searching. *Campaign,* July 16, 2004, s. 10.

Godby, Jean – Vizine-Goetz, Diane. 2000. ISKO participants discuss ways librarianship can improve responsiveness of the Web. *OCLC Newsletter,* 2000, vol. 247,s. 22-25.

Janes, Joseph. 2003. Internet Librarian. *American Libraries,* March 2003, vol. 34, no.3, s. 86-87.

Kwasnik, Barbara H. 1999. The Role of Classification in Knowledge Representation and Discovery. *Library Trends,* 1999, vol. 48, no.1, s. 22-50.

Looksmart, Inc. 2003. *LookSmart announces community additions to directory surpass 250,000 listings mark worldwide* [online]. 2003 [cit. 2004-12-18]. Dostupné z URL: http://zeal.com/about/press_room/press_releases/mar_18_03.jhtml.

Looksmart, Inc. 2004. *Guidelines Overview* [online]. 2004 [cit. 2004-12-18]. Dostupné z URL: http://-zeal.com/guidelines/overview.jhtml.

Manning, Gerard. 2002. *About the Virtual library* [online]. 2002 [cit. 2004-12-17].Dostupné z URL: http://vlib.org/AboutVL.html.

Netscape Communications Corporation. 2004. *Who we are and what we do* [online]. 2004 [cit. 2004-12-10]. Dostupné z URL: http://www.dmoz.org/help/geninfo.html.

Netscape Communications Corporation. 2002a. *About the Open directory project* [online]. 2002 [cit. 2004-12-10]. Dostupné z URL: http://www.dmoz.org/about.html.

Netscape Communications Corporation.2002b. *Open Directory Editorial Guidelines* [online]. 2002 [cit. 2004-12-16]. Dostupné z URL: http://dmoz.org/guidelines/include.html.

Netscape Communications Corporation. 1998. *Open Directory editorial guidelines. Subcategories* [online]. 1998 [cit. 2004-12-16]. Dostupné z URL: http://dmoz.org/guidelines/subcategories.html.

Saeed, H. – Chaudry, A. S. 2001. Potential of bibliographic tools to organize knowledge on the Internet. *Knowledge Organization,* 2001, vol. 28, no.1, s. 17-26.

Sullivan, Dany. 2002. *What people search for - most popular keywords* [online]. 2002 [cit.2004-12-15]. Dostupné z URL: http://searchenginewatch.com/facts/article.php/2156041.

Wall, Aaron. 2004. *LookSmart & Zeal Directories* [online]. 2004 [cit.2004-12-18]. Dostupné z URL: http://www.search-marketing.info/directories/looksmart.htm.

Wasserman, Todd. 2004. Latest Yahoo! search: Local web surfers. *Brandweek,* 2004, vol. 45, no.35, s. 11-12.

Wheatley, A. 2000. Subject trees on the internet. *Journal of Internet Cataloging,* 2000, vol. 2, no.3/4, s. 115-41.

Yahoo! Inc. 2003a. *Yahoo media relations: Company overview* [online]. 2003 [cit. 2004-12-11]. Dostupné z URL: http://docs.yahoo.com/info/misc/overview.html.

Yahoo! Inc. 2003b. *Yahoo! and Inktomi announce completion of acquisition. Inktomi a Wholly-Owned Subsidiary of Yahoo* [online]. 2003 [cit. 2004-12-15]. Dostupné z URL: http://docs.yahoo.com/docs/pr/release1071.html.

11.6 KAPITOLA 6: AUTOMATIZOVANÁ KLASIFIKACE A KATEGORIZACE

Ardö, Anders – Koch, Traugott. 1999. *Automatic classification demonstration page (DESIRE II)* [online]. 1999 [cit.1999-07-02]. Dostupné z URL: http://www.lub.lu.se/desire/demonstration.html.

Diekmann, Bernd. 2002. *Projektantrag GERHARD II* [online]. 2002 [cit.2004-11-06]. Dostupné z URL: http://www.gerhard.de/info/gerhard2.html.

ET-map [online]. 1998 [cit. 2004-11-06]. Dostupné z URL: http://ai2.bpa.arizona.edu/ent/.

Frank, Eibe – Paynter, Gordon. 1994. *Eibe Frank and Gordon Paynter´s LCSH to LCC research* [online]. 1994 [cit. 2004-11-13]. Dostupné z URL: http://infomine.ucr.edu/?view=projects/lcc_classification/.

Gietz, Peter. 2001. *Report on automatic classification system* [online]. 2001 [cit. 2001-06-19]. Dostupné z URL: http://www.daasi.de/reports/Report-automatic-classification.html.

Heery, Rachel – Carpenter, Leona - Day, Michael. 2001. Renardus project developements and the wider digital library context. *D-Lib Magazine* [online]. 2001 roč. 7, č. 4. Dostupné z URL: http://www.d-lib.org/dlib/april01/heery/04heery.html.

Insuma distributed search engine [online]. 2002 [cit. 2004-11-22]. Dostupné z URL: http://www.in-suma.de/insuma/download/insuma-whitepaper.pdf.

Jones, Steve. 1994. *Steve Jone´s LCSH classification research* [online]. 1994 [cit. 2004-11-13]. Dostupné z URL: http://infomine.ucr.edu/?view=projects/lcsh_classification.

Koch, Traugott. 1998. *Possible advantages of using traditional library classification systems in Internet services* [online]. 1998 [cit. 2004-09-10]. Dostupné z URL: http://www.lub.lu.se/tk/demos/mex9808b.html.

Lee, How-Chen – Kan, Min-Yen - Lai Sandra. 2004. *Stylistic and lexical co-training for web block classification.* [online]. Washington: WIDM´04, November 12-13, 2004 [cit. 2004-12-10]. Dostupné z URL: http://www.comp.nus.edu.sg/~kanmy/papers/p53-how.pdf.

Lubbes, R. Kirk. 2003. So you want to implement automatic categorization? *Information Management Journal,* 2003, vol. 37, no.2, s. 60-69.

McKiernan, Gerry. 1999. *Project Aristotle(sm): Automated categorization of web resources* [online]. 1999 [cit. 2004-11-26]. Dostupné z URL: http://www.public.iastate.edu/~CYBERSTACKS/Aristotle.htm.

Myslík, Vladimír. 1996. *Neuronové sítì.* 1996 [cit. 2005-04-05]. Dostupné z URL: http://aldebaran.feld.-cvut.cz/~xmyslik/www/neural.html.

Shafer, Keith. 1997. *Scorpion helps catalog the Web* [online]. 1997 [cit. 2004-11-10]. Dostupné z URL: http://orc.rsch.oclc.org:6109/b-asis.html.

Smith, John – Chan, Shin-Fu. 1996. *Searching for images and videos on the world-wide web* [online]. 1996 [cit. 2004-11-26]. Dostupné z URL: http://persia.ee.columbia.edu:8008/paper/.

Stergiou, Christos – Siganos, Dimitrios. 1996. Neural networks. *SURPRISE 96 Journal* [online]. vol. 2, no. 4 (Final reports) [cit. 2005-02-18], Dostupné z URL: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.

Watjen, Hans Joachim et. al. 1998. *Bericht zum DFG Projekt: GERHARD* [online]. Biblioteks- und Informationssystem (BIS) der Carl von Ossietzky Universitat Oldenburg, 1998 [cit. 2004-10-22]. Dostupné z URL: http://www.gerhard.de/info/do-kumente/dokumentation/gerhard/bericht.pdf.

WEBSOM - self-organizing maps for internet exploration [online]. 1999 [cit. 2004-11-06]. Dostupné z URL: http://websom.hut.fi/websom/.

Williams, James - Sochats, Kenneth - Morse, Emile. 1995. Visualization. *Annual review of information science and technology,* 1995, vol. 30, no.1, s. 161-207.

WWWLib. 1997. *Automatic classification of web resources using Java and Dewey Decimal Classification.* [online]. 1997 [cit. 2004-09-11]. Dostupné z URL: http://www.scit.wlv.ac.uk/~ex1253/classifier/.

Zorn, Peggy et al. 1999. Mining meets the web. *Online,* 1999, vol. 27, no.17, s. 17-28.

11.7 KAPITOLA 7: NEKONVENČNÍ A INOVAČNÍ PŘÍSTUPY

Garshol, Lars Marius. 2002. *What are topic maps?* [online]. 2002 [cit.2005-02-18]. Dostupné z URL: http://www.xml.com/lpt/a/2002/09/11/topicmaps.html.

Garshol, Lars Marius. 2004. *Metadata? Thesauri? Taxonomies? Topic maps!: Making sense of it all* [online]. 2004 [cit.2005-02-018]. Dostupné z URL: http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html.

Govindarajan, Jayesh – Ward, Matthew. 1998. *GeoViser: Geo-spatial clustering and visualization of* search engine results. Worchester (Massechusets): Worchester Polytechnic Institute. Computer Science Departement, 1998. Dostupné z URL: http://citeseer.ist.psu.edu/cis?q=Geoviser&cs=1.

Harada, Masanori – Sato, Shin-ya – Kazama, Kazuhiro. 2004. Finding authoritative people from the web.*Joint Conference on Digital Libraries (JCDL 2004).* Tucson, June 7-11, 2004. Dostupné z URL: http://www.ingrid.org/~harada/publications/-JCDL2004/p064-harada.pdf.

Kan, Min-Yen. 2004. Web page categorization without the web page.*WWW2004.* New York, May 17-22, 2004. Dostupné z URL: http://www.comp.nus.edu.sg/~kanmy/papers/www04.pdf.

Koehler, Wallace. 1999. Classifying web sites and web pages: the use of metrics and URL characteristics as markers. *Journal of Librarianship and Information Science,* 1999, vol. 31, no.1, s. 297-307.

Mackenzie, Maureen. 2000. The personal organization of electronic mail messages in a business environment: an exploratory study. *Library & Information Science Research,* 2000, vol. 22, no.4, s. 405-426.

Marco, Francisco – Navarro, Miguel. 1993. On some contributions of the cognitive sciences and epistemiology to a theory of classification. *Knowledge Organization,* 1993, vol. 20, no.3, s. 126-132.

Pepper, Steve. 2002b. *Ten theses on Topic maps and RDF* [online]. 2002 [cit.2005-02-19]. Dostupné z URL: http://www.ontopia.net/topicmaps/materials/rdf.html.

Pepper, Steve. 2002a. *The TAO of Topic maps* [online]. 2002 [cit.2005-02-18]. Dostupné z URL: http://www.ontopia.net/topicmaps/materials/tao.html.

11.8 KAPITOLA 8: VYHLEDÁVAČE

Arnold, Stephen. 2003. In search of the good search: The invisible elephant. *Searcher,* 2003, vol. 11, no.3, s. 40-56.

Bradley, Phil. 2000. The relevance of underpants to searching the Web. *Ariadne* [online]. 2000, roč. , č. 24, [cit. 2005-01-12]. Dostupné z URL: http://www.ariadne.ac.uk/ issue24/search-engines/.

Delaney, Kevin. 2004. Yahoo lets users fine-tune web wearches. *Wall Street Journal,*2004, October 5, s. D. 9.

Glover, Eric J., et al. 2001. Web Search - Your Way. *Communications of the ACM,* 2001, vol. 44, no. 12, s. 97-102.

How do search tools work: the freshness. The average freshness for Google, AltaVista, Alltheweb and Inktomi [online]. 2003 [cit.2005-01-14]. Dostupné z URL: http://www.revue-referencement.com/ENGLISH/search_tools_freshness.htm

Jansen, B.J.Spink, A. – Saracevic, T. 2000. A study of users queries on the Web. *Information Processing and Management,* 2000, vol. 36, no.2.

Jansen, B.J., et al. 1998. Real life information retrieval: A study of user queries on the Web. *SIGIR Forum,* 1998, vol. 33, no.1, s. s. 5-17.

Kotz, David – Gray, Robert S. Mobile agents and the future of the internet. *ACM Operating Systems Review,* 1999, vol. 33, no.3, s. 7-13.

Lawrence, Steve - Giles, C. Lee. 1999. Accessibility of information on the web. *Nature,* 1999, vol. 400, no. 8 July 1999, s. 107-109.

Metadent. 2004. *Search engine software company Inktomi* [online]. 2004 [cit.2005-01-14]. Dostupné z URL: http://www.metamend.com/inktomi.html.

Mizzaro, Stefano. 1997. Relevance: The whole story. *Journal of the American Society for Information Science and Technology,* 1997, vol. 48, no.9, s. 810-832.

Nielsen NetRankings 2007. *Nielsen online announces september U.S. search share rankings* [online].Nielsen NetRankings, October 19,2007[cit. 2007-10-22]. Dostupné z URL: http://www.netrankings.com/pr/pr_071019.pdf.

Notess, Greg. 2003. *Search engine statistics: Freshness showdown* [online]. May 17, 2003 [cit. 2005-01-14]. Dostupné z URL: http://www.searchengineshowdown.com/stats/freshness.shtml.

Notess, Greg. 2002. *Search engine statistics: Relative size showdown* [online]. 2002 [cit.2004-12-22]. Dostupné z URL: http://www.searchengineshowdown.com/stats/size.shtml.

Notess, Greg. 1999. On the net: Rising relevance in search engines. *Online,* 1999, vol. 23, no.3.

Pollock, Annabel – Hockley, Andrew. 1997. What´s wrong with internet searching. *D-Lib Magazine* [online]. 1997, [cit. 2005-01-15]. Dostupné z URL: http://www.dlib.org/ dlib/march 97/bt/03pollock.html

Spink, Amanda, et al. 2002. From e-sex to e-commerce: Web search changes. *Computer ,* 2002, vol. 35, no.3, s. 107-109.

Spink, Amanda, et al. 2001. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology,* 2001, vol. 52, no.12, s. 1073-1075.

Steinberg, Steve. 1996. Seek and ye shall find (maybe). *Wired,* 1996, vol. 4, no.5, s. 108-114.

Sullivan, Danny. *Search engine sizes* [online]. 2005 [cit. 2005-01-28]. Dostupné z URL: http://searchenginewatch.com/reports/article.php/2156481.

Sullivan, Danny. 2004a. *Search engine size wars V erupts* [online]. 2004 [cit. 2004-11-11]. Dostupné z URL: http://blog.searchenginewatch.com/blog/041111-084221.

Sullivan, Dany. 2004b. *comScore media metrix search engine ratings* [online]. 2004b [cit.2005-01-13]. Dostupné z URL: http://searchenginewatch.com/reports/article.php/2156451.

Sullivan, Dany. 2002. *How search engines work* [online]. 2002 [cit.2004-12-22]. Dostupné z URL: http://searchenginewatch.com/webmasters/article.php/2168031.

11.9 KAPITOLA 9: ZÁVĚR

Albrechtsen, Hanne – Jacob, Elin K.1998. The Dynamics of Classification Systems as Boundary Objects for Cooperation in the Electronic Library. *Library Trends,* 1998, vol. 47, no.2, s. 293-303.

Janes, Joseph. 2003. Internet Librarian. *American Libraries,* March 2003, vol. 34, no.3, s. 86-87.

Koch, Traugott. 1998. *Possible advantages of using traditional library classification systems in Internet services* [online]. 1998 [cit.2004-09-10]. Dostupné z URL: http://www.lub.lu.se/tk/demos/mex9808b.html.

Steinberg, Steve. 1996. Seek and ye shall find (maybe). *Wired,* 1996, vol. 4, no.5, s. 108-114.

SEZNAM TABULEK

SEZNAM OBRÁZKŮ