Citace odborné literatury jako nástroj rozvoje služeb a integrace digitálních knihoven

Eva Bratková *
brt@cuni.cz

Tento text je elektronickým postprintem dříve publikovaného konferenčního příspěvku:

BRATKOVÁ, E. Citace odborné literatury jako nástroj rozvoje služeb a integrace digitálních knihoven. In *AKP 2001 : Automatizace knihovnických procesů – 8 : Sborník z 8. ročníku semináře pořádaného ve dnech 24.-25. dubna 2001 v Liberci.* Praha : ČVUT, Výpočetní a informační centrum 2001, s. 109-120. ISBN 80-01-02-366-4.

Abstrakt: Příspěvek je věnován úloze, významu a možnostem využití citací při budování a rozvíjení služeb digitálních knihoven dostupných na WWW v oblasti vědy. Pozornost je věnována zejména volně dostupným systémům (zatím zahraničním a z vybraných oborů) a směrům či trendům jejich vnitřního i vzájemného propojování a integrace. Specifikovány jsou vybrané problémové okruhy a procesy uplatňování citací při budování archivů a knihoven digitálních dokumentů (automatické zjišťování digitálních dokumentů v prostoru WWW, automatické rozpoznávání a rozbor citací, citační propojování, tvorba „novodobých" citačních indexů, vyhledávání informací a zajišťování nadstavbových služeb, včetně informetrických analýz). Komentář zahrnuje i částečné porovnávání služeb komerčních citačních rejstříků se službami novodobých volně dostupných citačních indexů, jež vyrůstají nad současnými systémy digitálních informací.

Klíčová slova: vědecké informace, citace, automatická indexace citací, citační indexy, citační propojování digitálních dokumentů, digitální knihovny, elektronické archivy.

1 Úvodem

Tvorba citací (a způsoby citování zvláště) je procesem, který „trápí" nejen řadu knihovníků a informačních pracovníků v ČR, ale, jak dokládají četné dotazy, také řadu odborníků a studentů z nejrůznějších oborů. S nástupem Internetu a WWW se situace pro „našince" ještě více zkomplikovala, protože identifikace internetového zdroje a vytvoření té „správné" a jednotné formy bibliografického odkazu na něj je na úrovni tradičních zápisů otázkou poměrně dosti složitou. Jasné instrukce nevneslo do situace ani publikování převzaté mezinárodní normy ISO 690. Její dvě části si v řadě stejných situací odporují, v případě elektronických dokumentů na WWW uživatel mnohdy váhá, jak správně použít to či ono pravidlo. V řadě oborů profesionálové uplatňují raději své vlastní styly citování převzaté většinou z prostředí komerčních domácích a zahraničních vydavatelů.

Nicméně, následující příspěvek není věnován problematice tvorby citací a způsobů citování. Naopak, mohl by být jistou inspirací a možná i odpovědí na otázku, jestli tradičním a zejména pak novodobým informačním systémům, které citace zpracovávají, vadí či nevadí, že autoři vědecké literatury nestejně a někdy i velmi odlišně uvádějí a strukturují citaci na jedno a totéž dílo. A pokud nevadí, jaké jsou užívány metody k sjednocování takových citací ve prospěch budování kvalitních služeb daného systému?

Evidentní v tuto chvíli je skutečnost, že trendem jsou již také zcela nové způsoby citování, jejichž základem jsou jednoznačné identifikátory publikací, zejména pak elektronických. Možná se již brzy dočkáme i u nás sofistikovaných programů a systémů, které snadným

* Ústav informačních studií a knihovnictví, Filozofická fakulta UK v Praze, U Kříže 8, 158 00 Praha 5

způsobem vytvoří za autora citaci použitého pramene a třeba rovnou s hypertextovým propojením. Jak snadné by v takovém případě mohlo být automatické vytváření citačních rejstříků a návazné zkoumání užívání literatury ve vědě, rozvoje vědy, jejích trendů atd.

2 Trendem je propojování informací a informačních zdrojů

Vyhledávání a přístupy uživatelů k potřebným informacím všeobecně - a v oblasti vědy a výzkumu zvláště, se od nástupu Internetu proměnily do podoby, o které jsme před 2-3 desítkami let v podstatě jenom snili, pokud vůbec. Vedle možností online vyhledávání z bibliografických a citačních databází je možné pohodlné získávání plných textů přímo z databází, repozitářů digitálních knihoven či archivů elektronických tisků nebo jednoduše z prostoru volného webu. Neexistuje-li dokument v jiné než papírové formě, nastoupí již i u nás v řadě oborů dobře zavedená služba jejich elektronického dodávání nebo alespoň dodání papírové kopie přes službu MVS.

Využívání elektronických zdrojů a služeb jednotlivě po sobě však již dnes uživatelům také přestává stačit. Jejich nároky na kvalitnější služby s rozvojem technologií přirozeně rostou. Požadavkem a trendem doby je účelné **propojování informací** v podobě hyperodkazů, které by umožnilo jejich co nejrychlejší a nejefektivnější užití. Uživatel požaduje, aby na základě bibliografické informace, kterou si vyhledá online v bibliografické nebo citační databázi nebo kterou zjistí v dokumentu ve formě citace, měl možnost pomocí hyperodkazu přímo text dokumentu získat a to nejlépe v digitální formě.

Řada odborných týmů již proto několik let zkoumá, vytváří a rozvíjí nové technologie a systémy zajišťující automatické a dynamické propojování informací. Jde například o propojení bibliografického záznamu s abstraktem z oborové databáze na plný text dokumentu lokalizovaný v digitální knihovně komerčního nakladatele, o propojení katalogizačního záznamu knihy z knihovnické databáze na záznam knihy v databázi internetového knihkupectví nebo citace z časopiseckého článku na záznam v oborové bibliografické databázi nebo na záznam časopisu v katalogu knihovny [10]. Jednou z často komentovaných technologií, jež zajišťuje propojování různých informačních zdrojů, je technologie označovaná zkratkou **„SFX“** (Special Effects), která byla připravena ve spolupráci odborníků Gentské univerzity (Belgie) a knihovny Národní laboratoře v Los Alamos (USA) a aplikována v rámci jejich informačních fondů [9] za podpory řady komerčních nakladatelů a poskytovatelů informací.

Velmi angažovanou organizací v projektech, které řeší propojování informačních zdrojů v elektronickém prostředí na komerční bázi, je pochopitelně také americký Institut vědeckých informací **ISI**. Jeho několik desítek let budovaná prestižní databáze **Science Citation Index (SCI)** je v podstatě jediným kandidátem, který přichází do úvahy, pokud systém, který směřuje k budování integrovaného informačního celku, míří také k aplikaci **citačních vazeb**, které, jak známo, jsou v oblasti vědy a výzkumu vazbami nejcennějšími. Zdůrazněme, že v takovém případě půjde o systémy hybridní povahy a budou budované na komerční bázi.

Strategie samotného ISI, který uživatelům na celém světě připravil zatím svůj nejdokonalejší výrobek - webovské rozhraní **„The Web of Science“** pro přístup k citačním rejstříkům (s řadou nových prvků reprezentujících především vnitřních vazby mezi bibliografickými záznamy), směřuje také k budování automatických hypertextových vazeb na vybrané a žádané externí zdroje plných textů [2]. Budování propojovaných systémů, v nichž bude integrována báze SCI jako celek nebo nějaká její část, představuje jeden ze základních směrů v rozvoji komplexních informačních systémů, v nichž jsou aplikovány citační vazby.

Text příspěvku v dalších částech tento směr dále nekomentuje. Jeho pozornost je zaměřena na nové další směry aplikace citačních vazeb, tentokrát v prostoru plnotextových digitálních fondů. Vybrány jsou dva nejzajímavější současné modely uplatňování citačních vazeb a budování novodobých citačních rejstříků a jejich služeb.

3 Citace jako nástroj integrace stávajících archivů elektronických tisků

První významný směr využívání a uplatňování citací k propojování digitálních dokumentů reprezentují dnes již velmi dobře známé systémy tzv. **archivů elektronických tisků (Eprint archives)**. Jde o novodobé systémy elektronického publikování dokumentů s výraznou depozitní funkcí, které se soustavně rozvíjejí již od počátku 90. let 20. století. Staly se pozoruhodnými veřejně dostupnými digitálními fondy aktuální vědecké literatury (preprintů, ale i dalších typů tzv. šedé literatury) ve vybraných oborech, jako jsou fyzika, matematika, počítačová věda, medicína, ale třeba i vědy kognitivní. Elektronické archivy jsou odborníky považovány za významný doplněk velké řady komerčně zpřístupňovaných zdrojů vědecké literatury nabízených předními světovými vydavateli (Elsevier Science, Springer-Verlag, John Wiley & Sons aj.). Je třeba zdůraznit, že jde (zatím) o oborově profilované systémy, které však díky citačnímu propojování směřují k vybudování systémů polytematického charakteru.

Mezi nejvýznamnější elektronické archivy patří archiv **LANL** (*http://arxiv.org/*) provozovaný v rámci Národní laboratoře v Los Alamos (Nové Mexiko, USA). Jeho repozitář obsahuje již více než 130 000 dokumentů s nárůstem 25 000 dokumentů ročně a uživatelskou základnou čítající cca 50 000 uživatelů denně. Jeho součástí je také archiv CoRR (Computing Research Repository) dostupný buď přímo prostřednictvím archivu LANL nebo přes integrované rozhraní virtuální digitální knihovny NCSTRL (podrobnější charakteristiku těchto systémů viz v příspěvku sympozia INFOS 2000 [3]).

Jedním z klíčů ke zdokonalení funkčnosti těchto systémů je, jak dokladují výzkumné záměry [4], **citační propojování** archivovaných digitálních dokumentů. Propojování se vstřícně nabízí zejména proto, že jejich fondy plných textů jsou volně přístupné přes WWW. Archiv LANL již jako první v tuto chvíli demonstruje automatické propojování vybraných citací uvedených v dokumentu na záznamy jiného dokumentu uloženého ve stejném archivu s následným propojením na jeho plný text. Propojení je dáno jednak užívaným zdrojovým formátem (HyperTeX), jednak existencí **unikátního identifikátoru** (např. „hep-ph/9912313“), který je v systému přidělován automaticky každému nově ukládanému dokumentu. Identifikátory jsou modelovány dle principů daných americkým systémem „The Handle System“ rozvíjeného organizací CNRI [1]. Mezi samotnými autory, kteří své dokumenty do archivu ukládají sami, se rozšiřuje zajímavý způsob citování výlučně přes identifikátor dokumentu (jako hyperodkazy jsou často vidět i v anotacích záznamů reprezentujících uložené dokumenty). Na základě toho je také možné odkazy v systému dynamicky aktualizovat (např. při ukládání více verzí jednoho dokumentu). Protože archiv LANL nemá zatím přístup ke všem dokumentům, které autoři citují, a protože systém archivu neindexuje plné texty dokumentů (vyhledávání je realizováno na bázi metadat), je na základě těchto identifikátorů jiným externím systémem (databází SPIRES-HEP Stanfordské univerzity) zajišťována speciální služba, která připravuje pro každý uložený dokument komplexní seznam záznamů dokumentů, které daný dokument cituje (prohlížení záznamů směrem zpět v časové rovině), a seznam záznamů dokumentů, které daný dokument citují (prohlížení záznamů směrem dopředu v časové rovině). Příslušné seznamy, které je možné zobrazovat ze záznamů dokumentů pomocí hyperodkazů **„refers to“** a **„cited by“** tedy v danou chvíli představují jakési první provizorní citační seznamy, jež zajišťují alespoň základní statistiku citovanosti (užití) dokumentů.

Otázky vnitřního, ale i vnějšího propojování archivů elektronických tisků prostřednictvím citačních vazeb v současné době rozvíjí společný americko-britský projekt **„OpCit" (The Open Citation Project)** (*http://opcit.eprints.org/*), který byl veden jako dílčí úkol mezinárodního programu „International Digital Libraries Collaborative Research" (*http://www.dli2.nsf.gov/intl.html*[†]). Významnou roli při řešení technologických a jiných otázek hraje také nově založená iniciativa otevřených archivů **„The Open Archives Initiative (OAI)"** (*http://www.openarchives.org*), v rámci níž jsou rozvíjeny i zkušenosti získané z provozu kooperativního systému pro oblast počítačové vědy „CoRR" (*http://www.acm.org/repository*) a zkušenosti z experimentálního projektu online otevřeného časopisu „The Open Journal Project" (*http://journals.ecs.soton.ac.uk*) [5].

Základním cílem projektu „OpCit" je dosáhnout v budoucnosti v maximální možné míře situace, kdy uživatel bude moci v režimu online pohodlně, efektivně a rychle získávat celý komplex potřebných hypertextově propojených dokumentů a popřípadě i dalších informací v digitální formě pro řešení svých úloh. Reálné a účinné citační propojování lze, jak se domnívají někteří odborníci, uskutečnit jedině v prostředí volně dostupného digitálního fondu [4, s. 631]. Pokud by nějaká část potřebných dokumentů nebyla uživateli k dispozici (získá-li např. jenom informaci bibliografickou či abstrakt), bude to znamenat ztrátu, v lepším případě pak nastoupení namáhavé a někdy komplikované cesty získávání primárních dokumentů (takový způsob je nám konečně mnohým velmi důvěrně znám stále i dnes). Neméně důležitým cílem je také zajišťování komplexních bibliometrických či informetrických analýz, které pomohou lépe porozumět tomu, jak je vědecká literatura užívána v tvůrčím procesu.

Projekt „OpCit" zahrnuje celkem 8 dílčích komponent, z nichž některé se paralelně již řeší. Jde o následující okruhy:

- První okruh je věnován podstatným otázkám **nového a univerzálněji pojatého depozitního systému pro autory** včetně jeho rozhraní a infrastruktury deponovaných digitálních textů, který by měl vyhovovat i dalším vědeckým komunitám, jež doposud podobné archivy nemají (sociální nebo humanitní obory). Aktuálně se na základě doporučení iniciativy OAI řeší problematika tagování a sdílení metadat, jež by měla být plně v souladu s nově doporučenými standardy (včetně požadavků jazyka XML, struktury RDF či sémantiky Dublin Core). Návazně se řeší problematika **unifikace formátu pro citace** (v této chvíli autoři do existujících archivů přispívající používají velmi různorodé formáty), protože bez ní by bylo dynamické citační propojování dokumentů značně ztížené (zde srovnej s procesem automatické indexace citací v systému ResearchIndex komentované v kapitole 4). Výsledky řešení otázky unifikace citačních záznamů v tomto projektu budou určitě zajímavé i pro domácí odborníky, protože variabilita užívaných formátů bibliografických odkazů a hlavně způsobů citování v ČR je značná. Zájem ze strany odborníků z různých oborů sjednocovat formáty citací existuje a bude to možná právě prostředí digitálních systémů, které tomuto procesu napomůže. Na základě výsledků projektu „OpCit" by bylo možné připravit jistá doporučení nebo i metodiky pro domácí autory, zejména ty, kteří již připravují své dokumenty pro časopisy s databázemi archivů nebo i pro již nově zakládané digitální archivy či knihovny (viz nově založená digitální knihovna ETRDL, *http://dienst.muni.cz/*[‡]).
- Druhý okruh je zaměřen na řešení problematiky **nového designu uživatelského rozhraní** zejména s ohledem na požadavky pohodlné navigace mezi digitálními dokumenty v prostředí otevřených archivů elektronických tisků (v současné chvíli jsou např. výsledné

[†] Zdroj již není dostupný.

[‡] Zdroj již není dostupný.

dokumenty zobrazovány ve formátech TeX, Postscript nebo PDF, které poskytují pouze malé nebo žádné možnosti navigace). V rámci tohoto okruhu byl připraven prototyp rozhraní pro vyhledávání nad více elektronickými archivy najednou, který byl označen zkratkou UPS (The Universal Preprint Service) [8]. Jeho reálná podoba je k dispozici v rámci experimentálního provozu pod označením „ARC“ (*http://arc.cs.odu.edu/*[§]).

- Okruh třetí až pátý řeší klíčovou problematiku **extrakce údajů citací** ze všech textů uložených v archivu v takové formě, aby mohla být použita **pro automatické hypertextové propojování**. Konkrétně se předpokládá **generování hypertextových vazeb** pro všechny citace zahrnuté v archivu a z druhé strany **automatické doplňování hypertextových vazeb** do dokumentů v archivu již uložených. Uvedené okruhy tvoří jádro celého projektu.
- Okruh šestý až sedmý bude po ukončení úloh 3-5 a úlohy 1 věnován definitivnímu řešení ukládacího (depozitního) a vyhledávacího procesu v elektronických archivech.
- Poslední okruh řeší podstatné otázky **informetrických analýz citací a zdrojových dokumentů**, jež bude možné realizovat na základě uložených dat. Bude možné připravovat řadu statistik, jež v konvenčních citačních systémech, založených pouze na bibliografických záznamech, nebylo možné uskutečňovat (zde také srovnej se současnými výstupy systému ResearchIndex popsanými v kapitole 4).

Projekt „OpCit“ využívá, jak bylo uvedeno výše, některých výsledků již ukončeného britského projektu „The Open Journal Project“ (OJ), který řešil, kromě jiného, zejména otázky automatického citačního propojování článků lokalizovaných ve volně dostupných elektronických časopisech za využití bibliografických dat z vybraných komerčních bází, zejména pak části báze citačního rejstříku SCI [5]. Experiment umožnil prověřit online navigaci z článku na článek via připravené citační vazby. Automatické doplňování citačních vazeb bylo realizováno na základě technologie **DLS** (Distributed Link Service). Některé z modulů tohoto software budou adaptovány a rozvinuty i současném projektu.

V rámci optimalizace počítá projekt „OpCit“ také s možnou realizací vazeb na dokumenty či další typy informací z externích zdrojů. Může jít (v závislosti na dohodách s příslušnými partnery) například o:

- vazby dokumentů na databáze či soubory se jmény autorů (soubory autorit), s hesly z řízených slovníků, s deskriptory tezaurů apod.
- vazby na komentáře k uloženým preprintům, resp. i odpovědi jejich autorům
- vazby na elektronické archivy časopisů s publikovanými články, které vzešly z preprintů
- vazby na recenze publikovaných článků
- vazby na externí komerční bibliografické databáze (např. INSPEC, INIS aj.)
- vazby na jiné veřejně dostupné archivy nebo digitální knihovny
- vazby na server autora dokumentu uloženého v archivu apod.

Systémy archivů elektronických textů zatím své citační propojování dokumentů a návazné služby a výstupy zatím připravují. Podaří-li se dořešit zejména jednotlivé technologické, softwarové, organizační a jiné otázky, budou připraveny kvalitněji a efektivněji než doposud sloužit vědeckým komunitám i dalším uživatelům sítě Internet. Budou popřípadě též připraveny k integraci vyššího řádu s dalšími systémy, a to i z oblasti komerčních zdrojů.

[§] Zdroj již není dostupný.

4 Automatické budování citačních rejstříků v rámci digitálních knihoven

Druhý a velmi zajímavý směr využívání a uplatňování citací v prostředí digitálních fondů je zastoupen systémem, který experimentuje v oblasti tvorby nekonvenčních automaticky vytvářených citačních rejstříků, je systém **„ResearchIndex (nově také Citeseer)“** (*http://citeseer.ist.psu.edu/*). Hlavní tvůrci systému [6-7] jsou zaměstnanci Výzkumného ústavu společnosti NEC (Princeton, N.J., USA). Specifikem systému je citační rejstřík budovaný na bázi digitální knihovny, která zahrnuje volně dostupné dokumenty z oboru počítačová věda v celosvětovém záběru. Jde tedy o oborově profilovaný systém. Zdrojové dokumenty pocházejí zejména z primárního výzkumu (preprinty, výzkumné zprávy, disertace nebo i časopisecké a konferenční články). V současné době jde pouze o dokumenty ve formátech, jež nejsou běžně indexovány vyhledávacími systémy Internetu (Postscript a PDF). Knihovna, jejímž cílem je zlepšit vyhledávání a rozšiřování vědecké literatury uložené distribuovaně na nejrůznějších místech Internetu, popř. i mimo něj, je volně dostupná 24 hodin denně. Cílem systému je také zlepšení jeho dalších parametrů týkajících se nákladů, ceny na jeho provoz, dostupnosti primárních digitálních dokumentů, úplnosti registrace, včasnosti a zejména efektivnosti dalšího využívání dat v oblasti informetrie či scientometrie.

Dalším specifikem systému je unikátní technologie zajišťující **autonomní indexaci citací (ACI, Autonomous Citation Indexing)**. Pomocí ní se ze zdrojových dokumentů v konečné fázi automaticky vytváří citační index (vedle indexu reprezentujícího zdrojové dokumenty). Program ACI umožňuje detekovat a extrahovat z dokumentů citace, identifikovat citace reprezentující stejný dokument, byť se vyskytují v různých formách (viz obr. 1), a také identifikovat kontexty citací ve zdrojových dokumentech. Databáze zahrnuje v současné době cca 300 000 plných textů dokumentů a cca 4 000 000 citací. Uživatelům systém poskytuje zajímavé výstupy a služby, z nichž řada je zcela nového charakteru, budeme-li tento systém v určitých parametrech porovnávat se známým komerčním systémem SCI/ISI.

4.1 Zjišťování, získávání a zpracování zdrojových dokumentů

Systém ResearchIndex sice zahrnuje vlastního crawlera pro zjišťování a stahování dokumentů z WWW, zatím jej však nevyužívá. V systému se praktikuje jejich rychlejší vyhledávání pomocí známých vyhledávačů (Alta Vista, HotBot, Excite aj.). Využívá se strategie volby kombinace účinných selekčních údajů (jde např. o anglická slova „publications“, „conferences“, „proceedings“, „papers“, „postscript“ aj.). Monitorovány jsou pravidelně také elektronické diskusní skupiny. Využívá se i přímého nahlašování dokumentů k indexaci. Možností pro budoucnost je také propojení na vydavatele elektronických publikací. Specifickou funkcí systému je automatické zjišťování a vyřazování **duplicitních dokumentů**.

Stažené dokumenty jsou konvertovány do ASCII formátu pomocí programu PreScript, který také používá známá novozélandská digitální knihovna NZDL (*http://www.nzdl.org*). Automaticky se ověřuje, zda dokument má či nemá vědeckou či výzkumnou povahu (testuje se přítomnost citací nebo jejich sekcí). Významnou etapu ve zpracování zdrojových dokumentů představuje jejich indexace a tvorba běžného invertovaného souboru pro potřeby vyhledávání (vedle standardních se využívá i některých specifických technik s ohledem na funkce systému). Systém nevyužívá slovník stop slov z důvodu zvýšení přesnosti při vyhledávání (u autorů je např. často nutné v dotazu uvést iniciály jejich křestního jména).

4.2 Analýza a zpracování citací

Program ACI zajišťuje v systému tvorbu citačního indexu, který se v základních parametrech podobá komerčnímu citačnímu indexu SCI/ISI. Ve zdrojových dokumentech (oproti SCI jsou zahrnuty nejen články z časopisů) se zcela automaticky (nikoliv manuálně) na základě

analýzy textu zjišťují seznamy bibliografických odkazů (vodítkem může být identifikace hlavičky seznamu nebo seznam samotný). Následuje **extrakce jednotlivých citací**. Jsou identifikovány buď na základě jednoduchých **citačních identifikátorů** nebo jenom na základě **vertikálního mezerování** či **odsazování** záznamů. Extrahovány jsou údaje jako autor a název citovaného dokumentu, název zdrojového dokumentu, rok publikování, čísla stránek a zejména již zmíněný **citační identifikátor** (označení typu **„[12]“**, **„[Boll99]“, „(Cameron 1997)“ aj.**), pomocí něhož lze poté lokalizovat citaci v textu. Na základě citace je možné v textu zdrojového dokumentu automaticky extrahovat i její kontext, tj. množinu slov v okolí citačního identifikátoru (viz obr. 4). Systém je schopen zpracovat také různé varianty citačního identifikátoru v případě, že v citaci jsou uvedeni všichni autoři nebo jenom první z nich. Využívá se pravidelně se opakujících výrazů v citaci. K identifikaci údajů citací se využívá i doplňkové databáze jmen autorů, názvů časopisů apod.

Bibliografické odkazy na stejný dokument mohou být různě strukturovány a mohou zahrnovat různá množství údajů a také různá uspořádání těchto údajů (viz obr. 1). Význam programu ACI spočívá především ve schopnosti rozpoznat, že citace odkazují na stejný dokument. Využito je heuristického postupu založeného **na automatickém rozkladu citace**. Stálé výrazy, tj. údaje, které mají relativně stejnou syntax, pozici a složení, se zpracovávají jako první. Může se např. zjistit, že citační identifikátor vždy předchází celému záznamu a že zůstává nezměněn ve všech ostatních záznamech. Jakmile program zjistí pravidelné vlastnosti u nějaké citace, využije trendů v syntaktických vazbách mezi údaji pro předvídání místa, kde požadovaný údaj existuje, existuje-li vůbec. Např. informace o autorovi většinou vždy předchází názvu citované práce apod. Na základě této schopnosti je systém ResearchIndex schopen generovat seznamy citací (citovaných prací) a také statistiky citační frekvence [7].

[DGS93] S. Das & C.L. Giles & G.-Z. Sun, *Using Prior Knowledge in an NNPDA to Learn Context-Free Languages*, in C.L. Giles, S.J. Hanson & J.D. Cowan (Eds.), Advances in Neural Information Processing Systems 5, pp. 65-72, Morgan Kaufmann, 1993.

Das, S., Giles, C. L. and Sun, G. Z. (1993). Using prior knowledge in an NNPDA to learn context-free languages.

Das, S, Giles, C.L., and Sun, G. (1992). Using prior knowledge in an nnpda to learn context-free languages. Advances in Neural Information Processing Systems, volume 5.

DAS, S. (1993) Using prior knowledge in a NNPDA to learn context-free languages. In Hanson, S. J., Cowan, J. D., and Giles, C. L. (eds), Advances in Neural Information Processing Systems 5, 65--72. San Mateo: Morgan Kaufmann.

Obr. 1

Jak je vidět na obr. 1, rozpoznávání údajů citací a jejich skupin není vůbec jednoduché. Všechny údaje obvykle zahrnují nějaké chyby. Například čárka je často užívána k oddělení jednotlivých bloků údajů, ale je také užívána k oddělování několika jmen autorů nebo se často vyskytuje v názvech dokumentů. V jiných citacích odděluje bloky údajů tečka, která se ale také vyskytuje ve zkratkách. Jindy se v záznamech nevyskytuje vůbec žádná interpunkce. Tvůrci systému užívají několik metod pro identifikaci a seskupování citací k identickým dokumentům. Jde o měření vzdálenosti mezi řetězci symbolů (časté je například měření editační vzdálenosti), měření frekvence termů v textu (v jednom dokumentu i celém fondu), informace o údajích nebo jejich uspořádání v celkové struktuře záznamu a také o pravděpodobnostní modely, v rámci nichž se využívá známých bibliografických informací z dostupných databází k identifikaci údajů obsažených ve struktuře analyzovaných citací.

Uplatňovaný algoritmus je po mnoha testech považován v současnosti za dostatečný pro praktické použití. V r. 1999 byla u prototypu CiteSeer zjištěna **chybovost** jenom cca 5 % [7].

Uživatel i dnes náhodně najde chyby (při pořádání záznamů dle citovanosti aj.). Možnosti dalšího zlepšování však existují, výzkum pokračuje zejména v oblasti aplikace pravděpodobnostních modelů. Systému nabízí i možnost oprav chyb ze strany uživatelů.

Systém ACI umí rovněž identifikovat **bibliografické údaje zdrojových dokumentů** (i tato metoda je využívaná ve výše zmíněné knihovně NZDL). K identifikaci názvů a autorů z hlavičky dokumentů se úspěšně využívá informací o fontech a mezerování. Identifikace zdrojových dokumentů umožňuje analýzu **citovanosti**. Systém je na základě porovnání jmen autorů citovaných a citujících dokumentů schopen určit **autocitace**, čehož pak využívá při řazení záznamů dokumentů. Zjišťovat je možné také dokumenty, které citují mnoho vysoce citovaných dokumentů (angl. tzv. **„hubs“**), nebo také dokumenty, které jsou vysoce citovány (angl. **„authorities“**). Pro uživatele je např. zajímavé řazení rešerše podle dokumentů, které citují mnoho vysoce citovaných dokumentů (jde zpravidla o kvalitní přehledové práce, výukové materiály apod.). Systém dále zpřesnil algoritmus řazení podle citovanosti tím, že byla zavedena normalizace celkového počtu citací vzhledem k roku publikování dokumentu (zejména pro nové dokumenty se vypočítává tzv. **očekávaná citovanost**). Další aplikace analýz citovanosti jsou ve stadiu výzkumu (identifikování vědeckých komunit, expertů oboru, vztahů mezi dokumenty a vývojem literatury oboru aj.).

4.3 Vyhledávání informací a služby digitální knihovny

Systém ResearchIndex nabízí dva základní typy vyhledávání informací: **prohlížení** seznamů záznamů nebo vyhledávání pomocí **přímé formulace dotazu**. Oba způsoby ve fázi zobrazování výsledků umožňují netradiční **hypertextovou navigaci** mezi záznamy citovaných a citujících dokumentů včetně **přístupu k plnému textu** zdrojových dokumentů, tedy službu, kterou není možné uvidět v systému SCI/ISI (třeba v rámci nejnovějšího rozhraní „The Web of Science“). Jde o unikátní modelovou situaci poskytování služeb, které budou běžné (snad?) v budoucnosti.

V režimu **prohlížení záznamů** systém nabízí 5 automaticky generovaných rejstříků (seznamů), z nichž některé jsou typické pouze pro prostředí digitální knihovny (v komerčním systému SCI nemohou být realizovány). Jde o :

- Předmětový rejstřík (Computer Science Directory) k záznamům zdrojových dokumentů (třídění zahrnuje celkem 17 základních kategorií); standardní řazení je podle citovanosti dokumentů (Authorities), nabízí se ale i řazení podle stupně citovanosti vysoce citovaných dokumentů (Hubs/tutorials), podle data publikování (Date) a podle očekávané citovanosti (Expected authorities); názvy dokumentů jsou hypertextově propojeny s komplexním záznamem, jenž zahrnuje kromě základních údajů i různé citační podrobnosti a statistiky.
- Rejstřík nejnavštěvovanějších zdrojových dokumentů (Most Accessed Documents); standardně jsou zahrnuty záznamy z celé databáze (All-time most accessed documents) volit lze záznamy dokumentů v aktuálním roce (Most Recently Accessed Documents).
- Rejstřík nejcitovanějších zdrojových dokumentů (Most cited source document); standardní seznam zahrnuje všechny dokumenty publikované od roku 1990, volit lze však i množinu záznamů z vybraného roku.
- Citační rejstřík (Most Cited Citations, Most cited articles in Computer Science); standardně jsou řazeny záznamy všech citovaných prací generovaných k aktuálnímu roku, volit lze také seznamy za jednotlivé roky. Záznamy jsou propojeny na stránku se seznamem citujících dokumentů včetně kontextů citací (odkaz „Doc“ může být hypertextově propojen na plný text).

- Rejstřík nejcitovanějších autorů v oboru počítačová věda (Most Cited Authors) je specifickým seznamem systému (k roku 2001 zahrnuje cca 474 188 autorů); jména autorů jsou propojena s jejich personální stránkou, existuje-li, a to na základě služby HomePage Search (*http://hpsearch.uni-trier.de/*[**]) Univerzity v Trieru v SRN.

Přímé vyhledávání podle klíčového slova je realizováno buď v indexu citačním nebo v indexu zdrojových dokumentů. Systém ResearchIndex podporuje plně booleovský typ vyhledávání, vyhledávání frází a také vyhledávání proximitní. V případě nevhodně zapsané formulace dotazu systém nabízí její alternativu sám. Prozatím není zajišťována žádná podpora při řízení variantních jmen autorů, a proto je nutné při vyhledávání citací prací určitého autora dávat pozor při zápisu jeho jména.

Obr. 2 je ilustrací výstupu z vyhledávání v citačním indexu. Zobrazuje první 4 záznamy citací k tématu „metadata“ (k 19. 1. 2001). Hledání lze výhodně omezit (Restrict to) pouze na pole Autor nebo Název. Standardní řazení záznamů je podle očekávaného počtu citujících dokumentů (Expected citations) vzhledem k roku (čísla uváděná za hodnotou očekávané citovanosti znamenají celkový počet dokumentů citujících daný článek a počet autocitací v kulaté závorce). Volit lze i řazení podle celkového počtu citujících dokumentů (Citations) nebo podle roku vydání citovaného dokumentu (Date). Záznam citace předchází odkaz, který vede k unikátnímu zobrazení kontextu citací (viz kontext citace díla C. Lagoze aj. na obr. 4). Zkratka „Doc“ může být hypertextově propojena se záznamem zdrojového dokumentu.

930 citations found. Retrieving citations... **Order: citations weighted by the expected number for a given year.** Restrict to: Author field Title field Order by: Citations Date Hits: 50 100

First 100 articles Next 100

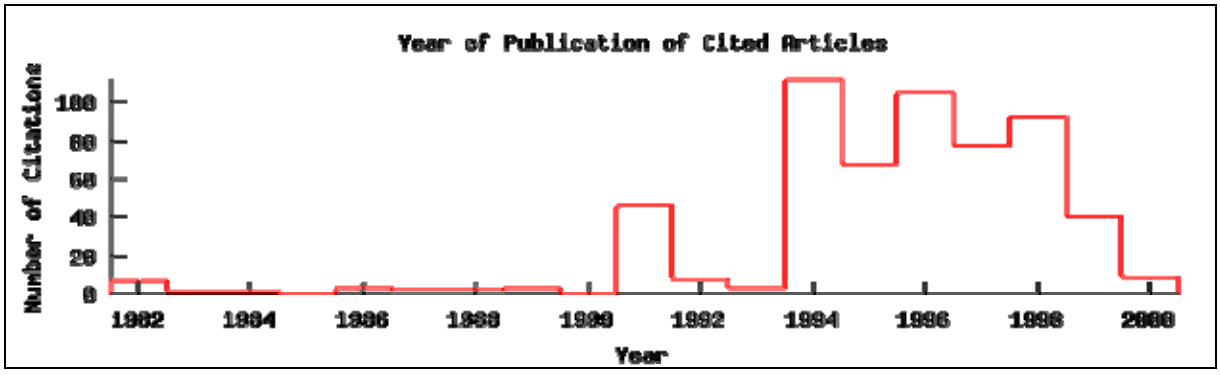
Context Doc ***35.2*** *24 (3):* Carl Lagoze, Clifford A. Lynch, and Ron Daniel, Jr. *The Warwick Framework: A container architecture for aggregating sets of* ***metadata***. Technical Report TR96-1593, Cornell University, Computer Science Dept., June 1996.

Context Doc ***27.9*** *22 (0):* R. Jain and A. Hampapur. ***Metadata*** *in video database*. ACM SIGMOD RECORD, 23(4):27---33, Dec. 1994.

Context Doc ***25.9*** *24 (2):* Siegel, M.; Madnick, S.: *A* ***Metadata*** *Approach to Resolving Semantic Conflicts*, Proc. 17th VLDB Conf., Barcelona, Spain, 1991

Context Doc ***24.1*** *19 (4):* G. Ganger and Y. Patt. ***Metadata*** *update performance in file systems*. In Proceedings of the First Symposium on Operating Systems Design and Implementation, pages 49--60, November 1994.

Obr. 2



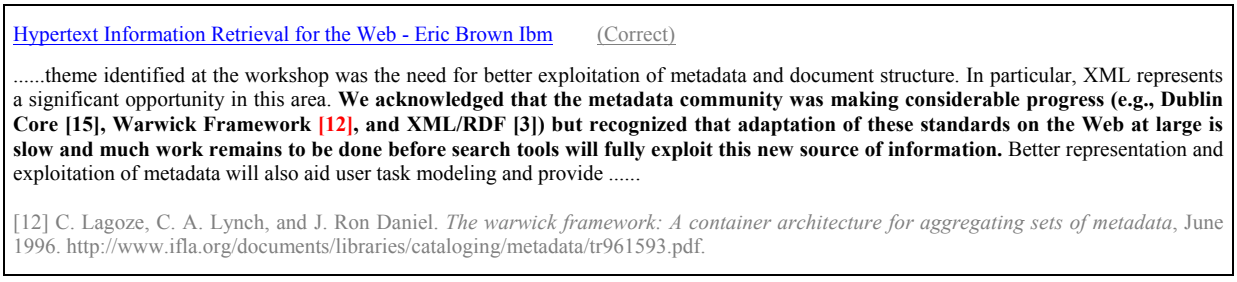
Obr. 3

Zvláštností zobrazování seznamů citací je **graf** (viz obr. 3), který ukazuje vztah počtu citovaných dokumentů (citovanosti) k roku jejich publikování (autocitace jsou z výpočtu vyloučeny). Ukázka dokládá, že tematika metadat se vyskytuje již od roku 1982, ovšem

[**] Zdroj již není dostupný.

významný nárůst nastal až v 90. letech 20. století, zejména pak po roce 1994 s rozvojem WWW (údaj je potřebné pokládat za hypotetický s ohledem na záběr digitálního fondu systému ResearchIndex).

Pomocí speciálního hyperodkazu „Context“ (viz obr. 2) lze zobrazit seznam názvů všech dokumentů, které daný dokument citují, včetně kontextu citací. Na obr. 4 je ukázka kontextu vybrané citace práce C. Lagoze aj. (cituje ji E. Brown). Název citujícího dokumentu je ve formě hyperodkazu veden na stránku s jeho komplexním záznamem. Množina slov, která je uvedena zvýrazněným písmem, pochází z citujícího dokumentu. Součástí množiny slov je i příslušný citační identifikátor (zobrazuje se barevně). Za množinou slov následuje originální bibliografický odkaz (jeho forma se zpravidla liší od formy ve výchozím seznamu citací). V citaci připojená adresa URL není v současné době taktéž systém zpracovávána.



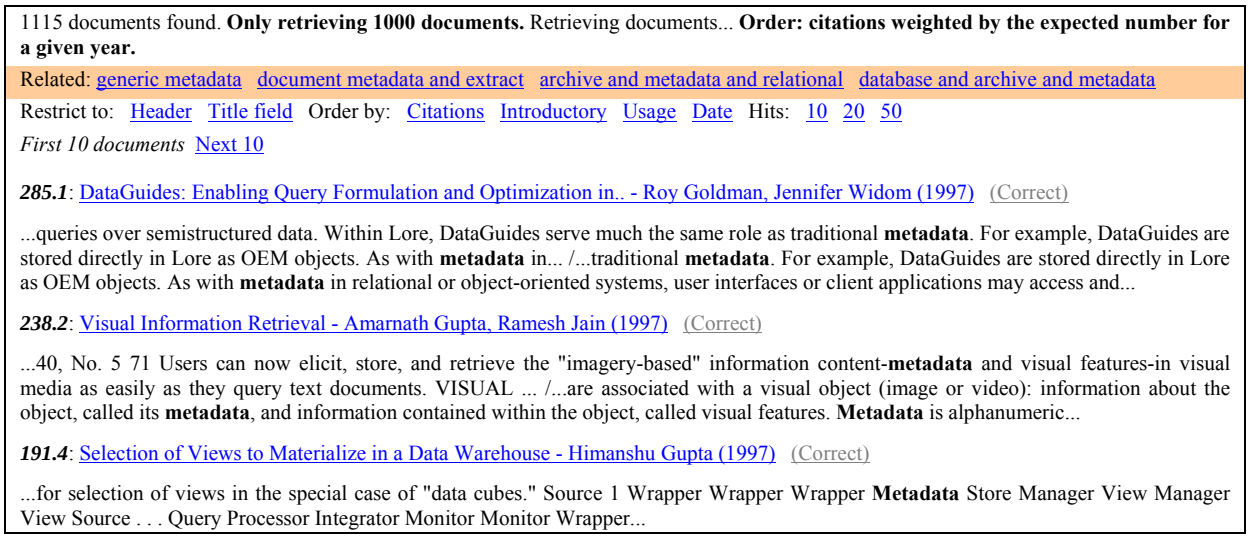
Hypertext Information Retrieval for the Web - Eric Brown Ibm (Correct)

.....theme identified at the workshop was the need for better exploitation of metadata and document structure. In particular, XML represents a significant opportunity in this area. **We acknowledged that the metadata community was making considerable progress (e.g., Dublin Core [15], Warwick Framework [12], and XML/RDF [3]) but recognized that adaptation of these standards on the Web at large is slow and much work remains to be done before search tools will fully exploit this new source of information.** Better representation and exploitation of metadata will also aid user task modeling and provide .....

[12] C. Lagoze, C. A. Lynch, and J. Ron Daniel. *The warwick framework: A container architecture for aggregating sets of metadata*, June 1996. http://www.ifla.org/documents/libraries/cataloging/metadata/tr961593.pdf.

Obr. 4

Vyhledávání v **indexu zdrojových dokumentů** vede k zobrazení seznamu jejich zkrácených záznamů (viz obr. 5). Standardně systém řadí podle očekávané citovanosti dokumentu vzhledem k datu vydání (Expected citations), volit lze i řazení podle celkové citovanosti (Citations), podle počtu vysoce citovaných dokumentů (výstižně uvedeného výrazem Introductory), podle využívání dokumentu v systému (Usage) a data (Date). Vyhledávání daného klíčového slova lze omezit pouze na pole Hlavičky dokumentu nebo na pole Název.



1115 documents found. **Only retrieving 1000 documents.** Retrieving documents... **Order: citations weighted by the expected number for a given year.**

Related: generic metadata | document metadata and extract | archive and metadata and relational | database and archive and metadata

Restrict to: Header | Title field | Order by: Citations | Introductory | Usage | Date | Hits: 10 | 20 | 50

First 10 documents Next 10

285.1: DataGuides: Enabling Query Formulation and Optimization in.. - Roy Goldman, Jennifer Widom (1997) (Correct)

...queries over semistructured data. Within Lore, DataGuides serve much the same role as traditional **metadata**. For example, DataGuides are stored directly in Lore as OEM objects. As with **metadata** in... /...traditional **metadata**. For example, DataGuides are stored directly in Lore as OEM objects. As with **metadata** in relational or object-oriented systems, user interfaces or client applications may access and...

238.2: Visual Information Retrieval - Amarnath Gupta, Ramesh Jain (1997) (Correct)

...40, No. 5 71 Users can now elicit, store, and retrieve the "imagery-based" information content-**metadata** and visual features-in visual media as easily as they query text documents. VISUAL ... /...are associated with a visual object (image or video): information about the object, called its **metadata**, and information contained within the object, called visual features. **Metadata** is alphanumeric...

191.4: Selection of Views to Materialize in a Data Warehouse - Himanshu Gupta (1997) (Correct)

...for selection of views in the special case of "data cubes." Source 1 Wrapper Wrapper Wrapper **Metadata** Store Manager View Manager View Source . . . Query Processor Integrator Monitor Monitor Wrapper...

Obr. 5

Zobrazené záznamy zahrnují základní údaje o dokumentu zjištěné v průběhu automatické indexace dokumentu. Resumé (Query-Sensitive Summary), které je významným prvkem v zobrazovaných záznamech, zahrnuje věty, v nichž se vyskytují hledaná slova. Na základě resumé mohou uživatelé posuzovat blíže relevanci dokumentu (její algoritmicky vypočtená

hodnota předchází záznam). Identifikační údaje jsou propojeny na komplexní záznam. Od r. 2001 systém umožňuje také vyhledávání tematicky příbuzných rešerší k rešerši výchozí přímo z obrazovky seznamu záznamů zdrojových dokumentů (viz nabídka „Related“ na obr. 5). Příbuzné dokumenty jsou vyhledávány na základě několika algoritmů (uplatňuje se již výše zmíněná metoda výpočtu frekvence termů, výpočtu podobnosti slov na bázi vektorového modelu a také metoda výpočtu frekvence společných citací v dokumentech z celé databáze).

Komplexní záznamy jsou dalším výrazným znakem systému. Kromě podrobných formálních údajů včetně abstraktu a odkazů vedoucích k zobrazení plných textů dokumentů v různých formátech, zahrnují řadu dílčích seznamů generovaných systémem. K základním patří:

- seznam příbuzných dokumentů (Active bibliography (related documents))
- seznam podobných dokumentů na úrovni věty (Similar documents (at the sentence level))
- seznam dokumentů, které si v rámci systému také prohlíželi uživatelé daného dokumentu (Users who viewed this document also viewed)
- seznam dokumentů, které citují daný dokument (Cited by) - jde o unikátní zobrazování záznamů směrem dopředu v časové rovině
- seznam kontextů citujících dokumentů (Context of citations to this paper)
- seznam citací uvedených v dokumentu (Citations made in this document) - jde o zobrazování záznamů směrem zpět v časové rovině
- seznam dalších příbuzných dokumentů na bázi společných citací (Related documents from co-citation).

Systém „modelové“ digitální knihovny ResearchIndex se úspěšně i nadále rozvíjí. Existuje mnoho cest k zlepšování a rozšiřování přístupu k vědecké literatuře na WWW. Jednou z nich je i cesta digitalizace dokumentů, které jsou prozatím jenom v tištěné formě. Prioritně by mohly být digitalizovány dokumenty, které tento systém vyhodnocuje jako nejcitovanější. Slibně se rozvíjejí i další služby, jako např. **průběžné informování** uživatelů o nových přírůstcích ve zdrojových dokumentech nebo citacích (uživatelé si mohou zadat svůj **uživatelský profil**, na základě něhož jim jsou informace rozesílány pomocí e-mailu).

Připravit lze také **vazby** digitální knihovny na **externí diskusní skupiny vědců, na externí komerční databáze,** jako je databáze SCI/ISI aj. Je v zájmu uživatelů zajišťovat plné texty dokumentů, které nejsou součástí digitální knihovny a které z důvodů ochrany autorských práv zpřístupňují pouze systémy komerčních vydavatelů. Digitální knihovnu ReserachIndex lze vlastně už v současné chvíli pokládat za významný doplněk komerčně dostupné databáze SCI/ISI v oboru počítačová věda. Má totiž některé nesporné přednosti:

- je budována zcela automaticky pomocí programu ACI (databáze SCI vyžaduje práci člověka, a to v několika fázích zpracování); náklady na zpracování dat se tak výrazně snižují
- nezahrnuje žádná limitační hlediska pro výběr vědecké literatury (to však může být z jistého pohledu chápáno i negativně); ISI pracuje na bázi přísně vymezené akviziční politiky
- zahrnuje všechny typy vědecké literatury (SCI zahrnuje jenom články z časopisů)
- zajišťuje volný přístup k plným textům dokumentů s možností unikátní oboustranné navigace mezi dokumenty (SCI zpřístupňuje plné texty v digitalizované podobě pouze na základě služby EDD a na komerční bázi)
- pomocí automatické indexace může vytvářet netradiční nadstavbové služby (kontexty citací, sledování příbuzných dokumentů, analýza a vyhodnocování trendů ve vědě apod.).

Digitální knihovna ResearchIndex představuje zajímavý experiment v současném rozvoji informačních systémů. Reprezentuje modelovou situaci výrazného trendu propojování informací a integrace systémů. Přispívá k celkovému zlepšení organizace, zpracování, vyhledávání, rozšiřování a zpřístupňování vědecké literatury v rámci sítě Internet.

5 Závěr

Rozvoj informačních technologií v posledních deseti letech silně působí na formování nejen integrovaných informačních komplexů, v rámci nichž dochází, zpravidla na bázi komerční, k propojování nejrůznějších bibliografických, knihovnických, nakladatelských aj. zdrojů, ale také na rozvoj a rozšiřování novodobých systémů fondů digitálních informací, z nichž řada je volně dostupných na WWW. I tyto novodobé systémy se začínají navzájem propojovat a jednou z podstatných vazeb, která je v současných projektech předmětem velkého zájmu, zejména v oblasti vědy, je vazba citační. Příspěvek poukázal na dva zajímavé modely a trendy uplatnění citací a citačních vazeb v prostředí volně dostupných textových digitálních informací, které je svojí povahou optimální pro takové uplatnění s možností budování následných netradičních služeb, které jsme doposud v konvenčních systémech nemohli z objektivních důvodů vidět. Vývoj ukáže, který z uvedených směrů bude pokračovat, který zanikne či se promění ve směr jiný. Více než pravděpodobné je také další propojování nových systémů s volně dostupnými fondy digitálních informací se systémy provozovanými na komerční bázi - to vše ve prospěch uživatelů informací. Co nejdokonalejší uplatňování citačních vazeb mezi informačními entitami je oním faktorem, který tomu bude napomáhat.

Použitá literatura a WWW odkazy

1. ARMS, W.Y. Key Concepts in the Architecture of the Digital Library. *D-Lib Magazine* [online]. July 1995 [cit. 2001-03-23]. Dostupný také z WWW: <http://www.dlib.org/dlib/July95/07arms.html>. ISSN 1082-9873.
2. ATKINS, Helen. The ISI® Web of Science® - Links and Electronic Journals : How links work today in the Web of Science, and the challenges posed by electronic journals. *D-Lib Magazine* [online]. September 1999, vol. 5, no. 9 [cit. 2001-03-23]. Dostupný také z WWW: <http://www.dlib.org/dlib/september99/atkins/09atkins.html>. ISSN 1082-9873.
3. BRATKOVÁ, E. Elektronické archivy vědecké literatury a jejich integrace. In *INFOS 2000 : zborník z 30. medzinárodného informatického sympózia , ktoré sa konalo v dňoch 3.-6. apríla 2000 v Starej Lesnej*. Bratislava : Spolok slovenských knihovníkov, 2000, s. 187-200. Dostupný také z WWW: <http://www.aib.sk/infos/infos2000/19.htm>. ISBN 80-81565-80-5.
4. HARNAD, S.; CARR, L. Integrating, navigating, and analysing open Eprint archives through open citation linking (the OpCit project). *Current science*. September 2000, vol. 79, no. 5, s. 629-638. Dostupný také z WWW: <http://www.iisc.ernet.in/~currsci/sep102000/629.pdf>. ISSN 0011-3891.
5. HITCHCOCK, S. aj. Citation Linking : Improving Access to Online Journals. In *Proceedings of the second ACM International Conference on Digital Libraries : ACM Digital Libraries '97, Philadelphia, PA, July 23-26, 1997*. New York : ACM, 1997, s. 123. Dostupný také z WWW: <http://journals.ecs.soton.ac.uk/acmdl97.htm>.
6. LAWRENCE, S.; BOLLACKER, K.; GILES, C.L. Indexing and Retrieval of Scientific Literature. In *Proceedings of the Eighth International Conference on Information Knowledge Management : CIKM '99, November 2-6, 1999, Kansas City, Missouri*. New York : ACM Press, 1999, s. 139-146. Dostupný z WWW: <http://clgiles.ist.psu.edu/papers/CIKM-1999-indexing-retrieval-sci-lit.pdf>. ISBN 1-5811-3146-1.

7. LAWRENCE, S.; GILES, C.L.; BOLLACKER, K. Digital Libraries and Autonomous Citation Indexing. *Computer (IEEE)*. 1999, vol. 32, no. 6, s. 67-71. Dostupný také z WWW: <http://clgiles.ist.psu.edu/papers/IEEE.Computer.DL-ACI.pdf>. ISSN 0018-9162
8. Van de SOMPEL, H. aj. The UPS Prototype : An Experimental End-User Service across E-Print Archives. *D-Lib Magazine* [online]. 2000, vol. 6, no. 2 [cit. 2001-03-23]. Dostupný z WWW: <http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html>. ISSN 1082-9873.
9. Van de SOMPEL, H.; HOCHSTENBACH, P. Reference Linking in a Hybrid Library Environment. Part 1-3. *D-Lib Magazine* [online]. 2000, vol 5, no. 4, no. 10 [cit. 2001-03-23]. Dostupný z WWW: <http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html>, <http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt2.html>, <http://www.dlib.org/dlib/october99/van_de_sompel/10van_de_sompel.html>. ISSN 1082-9873.
10. Van de SOMPEL, H.; BEIT-ARIE, O. Open Linking in the Scholarly Information Environment Using the OpenURL Framework. *D-Lib Magazine* [online]. 2001, vol. 7, no. 3 [cit. 2001-03-23]. Dostupný z WWW: <http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html>. ISSN 1082-9873.