# Technologie digitálních knihoven

**Miroslav BARTOŠEK**
Ústav výpočetní techniky, Masarykova univerzita, Brno
bartosek@ics.muni.cz

INFORUM 2006: 12. konference o profesionálních informačních zdrojích
Praha, 23. - 25.5. 2006

***Abstrakt.*** *Digitální knihovny mají za sebou zhruba 15 let rozvoje. Cílem příspěvku je poskytnout přehled o technologiích tvořících základní kameny soudobé infrastruktury digitálních knihoven, a to v oblastech budování sbírek, metadat, identifikace a interoperability. Zmíněny jsou i nejvyužívanější volně dostupné systémy na podporu realizace digitálních knihoven. V závěru příspěvku je připojeno krátké zamyšlení nad novým vymezením digitálních knihoven v éře googlizace.*

## Úvod

Citát na úvod: „Jedním z diskutovaných témat na letošní konferenci *Computers in Libraries* (Washington, 22.-24.3.2006) byla potřeba změn a inovace knihoven, aby byly schopny vyhovět požadavkům nové generace uživatelů – označovaných jako „Milenialisté" (Millennials, uživatelé narození po roce 1982 a přicházející do dospělosti na začátku nového tisíciletí) – kteří očekávají, že jakákoliv hledaná informace bude okamžitě k dispozici *kdekoliv, kdykoliv a nejlépe prostřednictvím mobilního zařízení*. Jeden z řečníků vyslovil obavu, že pokud se knihovny potřebám Milenialistů nepřizpůsobí již teď, ztratí tuto potenciální uživatelskou skupinu navždy". (Bonita Wilson, editoriál dubnového čísla D-Lib Magazine 2006 – viz [1]).

Za jednou z odpovědí na požadavek okamžitého a časem/prostorem neomezeného přístupu k informacím jsou pokládány *digitální knihovny*. Impulsem a nutnou podmínkou pro rozvoj digitálních knihoven bylo dosažení určité technologické úrovně – zejména v oblasti uchovávání velkých objemů digitálních dat (vysokokapacitní a přitom levná záznamová média), komunikací (rychlá, všeobecně dostupná a snadno použitelná počítačová síť) a výpočetních zařízení (dostupné osobní počítače na jedné straně a výkonné servery na straně druhé). Potřebné technologické úrovně bylo ve vyspělých zemích dosaženo počátkem 90. let minulého století, a zhruba od téže doby se datuje i rozvoj digitálních knihoven. Dnes se tedy digitální knihovny nachází ve věku dospívání [2]. Každé dospívání (puberta?) nese s sebou své první úspěchy, ale i řadu problémů. K úspěchům digitálních knihoven patří technologie jako je Google (jeho počátky byly iniciovány výzkumem v oblasti digitálních knihoven), standardy Dublin Core, METS, OAI-PMH, OpenURL, DOI, systémy Greenstone, DSpace či EPrints, řada rozsáhlých a dobře fungujících digitálních knihoven, a mnohé další. Problémy a obavy pak vyplývají především ze skutečnosti, že v řadě oblastí, které již na úsvitu digitálních knihoven byly identifikovány jako *nezbytné* součásti infrastruktury digitálních knihoven, stále nejsou k dispozici praktická řešení vhodná pro široké nasazení v praxi. Jako příklady lze uvést chybějící globální jednotný persistentní identifikační systém, malé pokroky v řešení sémantické interoperability, problém kvalitních metadat a optimálního metadatového schématu, či všeobecně nerozvinutou bezpečnostní infrastrukturu.

Hlavní pozornost ve výzkumu (a částečně i praxi) digitálních knihoven se v posledních letech přesunula od budování izolovaných systémů digitálních knihoven k hledání a vytváření potřebné infrastruktury, která by umožnila jednak efektivnější realizaci vlastních digitálních knihoven, jednak jejich užší vzájemnou spolupráci. Podobně jako klasické knihovny, které jsou vytvářeny a provozovány jako samostatné instituce, nicméně fungují v rámci kooperativního systému, v němž každá z knihoven plní určitou roli, měly by i digitální knihovny být schopny vytvářet rozsáhlé propojené systémy podle potřeb uživatelů[1]. Přestože v současnosti není ještě žádná ucelená globální infrastruktura digitálních knihoven k dispozici, existuje pro ni již řada dobře usazených základních stavebních kamenů. Cílem příspěvku je tyto základní kameny identifikovat a stručně je přiblížit knihovníkům se zájmem o problematiku digitálních knihoven. Nejde o technický popis standardů či technologií, na to zde není ani prostor ani publikum, nýbrž o zasazení věci do širších souvislostí. Připojeny jsou i odkazy na vybrané digitální knihovny a zdroje podrobnějších informací o zmiňovaných technologiích.

[1] Připomeňme vizi digitální knihovny předestřenou C.Lynchem a H.García-Molinou již v roce 1995 v [3]: „The common vision is one of tens of thousands of repositories of digital information that are autonomously managed yet integrated into what users view as a coherent digital library system."

## Budování sbírek

Existují tři základní strategie pro vytváření obsahu digitální knihovny: (a) digitalizace informačních zdrojů existujících v analogové podobě, (b) začleňování zdrojů vzniklých přímo v digitální podobě (born-digital), (c) sklízení informačních zdrojů z webu (harvesting).

(a) Samotný proces *digitalizace* je již dobře zvládnutelná a poměrně rutinní záležitost. Existuje řada standardů, dlouholetých zkušeností a doporučení (Best Practices) pro digitalizaci jednotlivých typů dokumentů, médií či formátů v různých aplikačních oblastech. Ty lze vhodně využít a vyhnout se tak znovu-objevování již vynalezených věcí. Vznikly nové grafické formáty vhodné pro efektivní prezentaci obrazů naskenovaných dokumentů na Internetu. Jako příklad lze uvést formáty DjVu [4] (vhodný zejména pro bitonální textové dokumenty) či MrSID [5] (pro mapy s vysokým rozlišením či jiné rozsáhlé grafické soubory dat), které nabízí velmi úsporné uložení obrázků při zachování vysoké kvality zobrazení. Podrobnější informace k oběma formátům byly již prezentovány i na konferenci Inforum, např. v přednášce [6].

Významný impuls pro rozvoj digitalizačních technologií v posledních letech poskytly masové digitalizační projekty jako například *American Memory* [7] (Library of Congress), *JSTOR* [8] či *The Million Book Project* [9] (Carnegie Mellon University s partnery z Číny a Indie), nověji pak především *Google Book Search* [10] (Google) a jeho přímý konkurent *Open Content Alliance* [11] (Internet Archive, Yahoo! a další instituce). Významně se na poli digitalizace angažují i přední světoví nakladatelé odborné a vědecké literatury (Elsevier, Springer, odborné společnosti typu ACM, IEEE a další), kteří zpětně digitalizují často svou kompletní produkci a nabízí ji jakou součást vlastních (komerčních) digitálních knihoven nebo samostatně v podobě historických archivů (umožňujících nakladatelům znovu prodat – tentokrát v digitálním kabátě – svou starší produkci). Díky masovým digitalizačním aktivitám byly v poslední době výrazně zdokonaleny jak technické prostředky (vysoce výkonné skenery zvládající stovky až tisíce stran textu za hodinu) tak digitalizační postupy. To se projevilo na dramatickém poklesu nákladů na skenování (cena za naskenovanou stránku textu poklesla řádově; u velkých projektů se v současnosti pohybuje již v řádu centů) a na rychlém růstu dostupného digitálního obsahu.

(b) Zejména v akademickém prostředí hrají stále důležitější roli digitální repozitáře, které slouží pro ukládání a zpřístupňování odborné a vědecké produkce samotnými autory. Velká část těchto zdrojů je provozována v režimu open access – jsou volně k dispozici. Významného postavení dosáhly zejména oborově specializované pre-printové či post-printové archivy, jakým je v oblasti fyziky, astrofyziky, matematiky a informatiky systém *Arxiv.org* [12] (vytvořený již v roce 1991!, v současnosti provozovaný pod patronací Cornellovy univerzity a nabízející více jak 370.000 eprintů s měsíčním přírůstkem 4-5000 článků), v oblasti ekonomie *Research Papers in Economics* (RePEc) [13] nebo v oblasti astronomie a astrofyziky systém *Astronomics Data System* (ADS) [14] (sponzorovaný NASA). V souvislosti s rozvojem obecných programových systémů na podporu tvorby repozitářů (DSpace, EPrints, Fedora) vznikla a dále vzniká řada institucionálních repozitářů, shromažďujících a vystavujících na Internetu intelektuální produkci příslušné instituce. Mezi největší z nich patří instalace *DSpace na Cambridgské univerzitě* [15] (repozitář obsahoval v době psaní článku přes 133.000 položek). Jako příklad systému EPrints uveďme „mateřskou“ instalaci *University of Southampton EPrints Repository* [16], pro knihovníky může být zajímavá instalace *E-LIS* [17], což je volně dostupný mezinárodní archiv dokumentů z oblasti knihovní a informační vědy.

(c) V oblasti vytváření sbírek prostřednictvím automatizovaného *sklízení dokumentů* z Internetu patří k nejznámějším a nejrozsáhlejším aktivitám projekt *Internet Archive* [18], který se již od roku 1996 snaží průběžně archivovat obsah celosvětového webu. Objem dosud sklizených a archivovaných dokumentů přesahuje již 1 Petabyte (10**15 Bytů, neboli miliardu Megabytů) a měsíčně v současnosti přibývá do archivu více jak 20 Terabytů dat. Paralelně (a v poslední době i ve spolupráci) s touto aktivitou provádí řada národních knihoven po celém světě své národní sklizně webu, jako součást programů zaměřených na uchovávání kulturního dědictví – viz například český projekt *WebArchiv* [19]. V rámci daných aktivit byla vyvinuta řada pokročilých nástrojů pro sklízení a archivaci webu (např. sklízecí roboti Heritrix, NEDLIB harvester či z architektury Apache Lucene vycházející Nutch, nástroje pro zpřístupnění archivu sklizených dokumentů WERA, NutchWAX, a další). Automatizované sklízení lze používat nejen plošně, je možné využít ho i pro cílené selektivní sklízení. Tento přístup je využíván například pro budování specializovaných automatizovaných digitálních knihoven, jako je např. citační digitální knihovna z oblasti computer science *CiteSeer* [20], dříve známá pod názvem ResearchIndex (Penn State University).

## Metadata

Metadata jsou srdcem každé (digitální) knihovny. Popisují informační zdroje, jejich vnitřní strukturu, vztahy a souvislosti mezi zdroji navzájem, podmínky a způsob zpřístupnění, atd. Knihovny mají staletou praxi ve vytváření metadat (bibliografických záznamů, rejstříků) a značně pokročily ve standardizaci zejména popisných metadat – viz třeba standardy MARC. Nicméně klasické knihovní metadatové standardy nejsou v digitálním prostředí vždy dobře použitelné – souvisí to zejména s jejich složitostí (předpokládají školeného katalogizátora) a primární orientací na tištěné typy dokumentů.

Prvním pokusem o vytvoření univerzálního digitálního metadatového formátu byl standard *Dublin Core* (DC) [21]. Jeho základ vznikl roku 1995 a průběžně je mezinárodní komunitou uživatelů dále rozvíjen. Je tvořen 15 metadatovými prvky vybranými tak, aby charakterizovaly základní společné atributy většiny typů informačních zdrojů (název, tvůrce, datum, popis, identifikátor, typ, formát, atd.). Celý formát je navržen s důrazem na maximální jednoduchost, aby byl i nezaškoleným uživatelem snadno použitelný pro autorskou samokatalogizaci. Po počátečním nadšení se ukázalo, že představy o tom, že DC se stane univerzálním formátem, který umožní zbavit se dosavadní nepřehledné změti velice různorodých a nekompatibilních digitálních metadatových schémat, nejsou reálné. Dublin Core má své velmi důležité místo v oblasti jednoduchého popisu, jako převodník mezi různými formáty a jako metadatový základ obecných interoperabilních technologií, například OAI (viz níže část věnovaná interoperabilitě). Není to však všelék, pro řadu aplikací je příliš jednoduchý a málo flexibilní.

S druhým pokusem o metadatový „svatý grál“ přišla v roce v roce 2002 Kongresová knihovna v podobě metadatového standardu *MODS* [22]. Vychází ze standardu MARC21, který je však zjednodušen, přizpůsoben popisu digitálních zdrojů a převlečen do moderního kabátu v podobě značkovacího jazyka XML. MODS nabízí 19 základních prvků (titleInfo, name, typeOfResource, genre, originInfo, physicalDescription, atd.), které mohou být dle potřeby zjemňovány dalšími 64 sub-elementy (title, subTitle, partName, partNumber, ...). Uživatel si tak může volit míru detailnosti jednotlivých prvků – od jednoduchého popisu alá Dublin Core až po podrobný popis na úrovni knihovnických marcových záznamů. Standard vychází z knihovnického prostředí a obecně vypadá velmi zajímavě. Je však poměrně ještě hodně „čerstvý“, takže se zatím nedá odhadnout, nakolik ho široká Internetovská komunita přijme i pro aplikace mimo okruh digitálních knihoven[2].

Třetím stavebním kamenem digitálních knihoven v oblasti metadat je *METS* – *Metadata Encoding and Transmission Standard* [23]. Jde o metadatový standard zcela jiného druhu než DC či MODS. Nejedná se v tomto případě o popisná metadata, nýbrž o jakýsi „kontejner“ ve formě XML dokumentu, který slouží k propojení všech možných metadatových záznamů (popisných, strukturálních, technických, administrativních) a všech zdrojových souborů tvořících jeden digitální objekt v digitální knihovně. Složitější digitální objekty – například kniha či časopis – jsou obvykle tvořeny velkým počtem souborů (obsahujících např. obrázky či texty jednotlivých stran) a množstvím nejrůznějších metadatových záznamů, které je třeba všechny správně propojit a „zabalit do jednoho digitálního balíku“. A to je právě to, co umí METS. Tento standard vznikl přímo pro potřeby digitálních knihoven, zobecněním řešení a zkušeností z projektu *The Making of America II* [24] pod patronátem asociace *Digital Library Federation* [25] (o údržbu a rozvoj standardu se stará Kongresová knihovna).

## Identifikace

Digitální knihovny a jimi zpřístupňované informační digitální objekty se nacházejí v prostředí *celosvětové* webové sítě. Aby mohly být tyto objekty napříč celým světem propojovány a odkazovány (např. z bibliografického záznamu v rámci jedné knihovny vede odkaz na plný text daného informačního zdroje nacházející se v jiné digitální knihovně), musí jim být přiděleny *identifikátory*. Aby propojení objektů – a služby na propojení objektů závisející – byly spolehlivé, musí použité identifikátory splňovat řadu důležitých podmínek. Mezi ty hlavní patří *globální jednoznačnost* (dva různé objekty na webu nesmí mít přidělenu stejnou hodnotu identifikátoru), *persistence* (identifikátor musí být trvale platný a funkční), *nezávislost na lokaci* (přesuneme-li objekt na jiné místo webu, identifikátor objektu by měl zůstat nezměněn) a *směrovatelnost* (zadáme-li identifikátor do webového prohlížeče, infrastruktura webu by nás měla automaticky nasměrovat na příslušný objekt, ať se nachází v rámci globální sítě kdekoliv).

---

[2] Podobně jako je MODS modernizovaným ekvivalentem standardu MARC pro bibliografické záznamy, vytvořila Kongresová knihovna i standard MADS, který je obdobným ekvivalentem pro marcovské *autoritní* záznamy.

Hlavním identifikačním systémem používaným v současnosti na webu je URL (Uniform Resource Lokator). Problém s URL spočívá v tom, že nesplňuje všechny výše uvedené podmínky; je sice globálně jednoznačné a směrovatelné, není ale persistentní a nezávislé na lokaci (důsledkem toho je množství neplatných odkazů s nimiž se v praxi potýká každý z nás). Vývojáři Internetu si potřebu persistentního globálního identifikátoru – identifikující objekt samotný a nikoliv jeho umístění – uvědomili již dávno a navrhli (teoretické) řešení v podobě identifikátoru URN, Uniform Resource Name. Z řady důvodů (složitost, nákladnost, koncepční spory) k praktickému nasazení univerzálního mechanismu URN dosud nedošlo. Výzkumníci a vývojáři přišli proto s řadou vlastních řešení, které sice nejsou univerzálním řešením pro web jako takový, mohou však dobře posloužit při implementaci systémů digitálních knihoven. Mezi nejvýznamnější a nejpoužívanější globální digitální identifikační systémy současnosti patří PURL, Handle a DOI.

Systém *PURL* (Persistent URL) [26] využívá pro identifikaci „staré známé“ URL; snaží se ale překonat jeho problémy s trvanlivostí a závislostí na lokaci, a to prostřednictvím nepřímé adresace. Jako identifikátor je zdroji přiděleno zvolené PURL, což je URL na PURL serveru, např. http://purl.oclc/bartosek/1. Toto PURL odkazuje na adresu (lokaci *bartosek/1* na PURL serveru). Na ní je teprve umístěno skutečné aktuální URL zdroje, a na toto URL je uživatel automaticky přesměrováván. Uživatelé zdroje znají a používají jen jeho PURL. Pokud se změní URL zdroje, stačí aby správce zdroje provedl příslušnou změnu na dané PURL adrese a uživatelé jsou automaticky přesměrovávání na nové URL. Systém PURL má výhodu v tom, že je jednoduchý a navíc směrovatelný bez nutnosti používat jakýkoliv speciální software (je to vlastně stále jen URL).

Větší funkcionalitu a širší možnosti využití nabízí *systém Handle* [27], vyvinutý a udržovaný americkou organizací CNRI (Corporation for National Research Initiatives, nevýdělečná organizace podporující rozvoj národní informační infrastruktury). Jde vlastně o kompletní – na systému URL nezávislou – identifikační infrastrukturu pro přidělování, správu i směrování identifikátorů na Internetu. Tato infrastruktura je tvořena systémem pro distribuované přidělování identifikátorů, sadou vlastních protokolů, rozšíření do www-prohlížečů a sítí serverů pro směrování identifikátorů. Celý systém je velmi propracovaný a efektivní, využívá ho řada dalších nadstavbových technologií (např. systém DOI, viz níže), softwarů na podporu digitálních knihoven i nezávislé implementace velkých digitálních knihoven. Jedinou vadou na kráse je skutečnost, že nebyl začleněný do obecné infrastruktury Internetu, stojí tak nějak vedle ní.

Zřejmě nejúspěšnějším a nejpropracovanějším systémem digitálních identifikátorů současnosti je *Digital Object Identifier – DOI* [28]. Vznikl na zakázku Asociace amerických nakladatelů s cílem vytvořit spolehlivý komerčně použitelný systém digitální identifikace objektů (jakéhokoliv typu) pro potřeby obchodování na Internetu (skutečné využití je ale mnohem širší). Používání identifikátorů DOI je bezplatné, nicméně jejich přidělování je obvykle zpoplatněno (záleží na obchodním modelu té-které registrační agentury). Celý systém je po ekonomické stránce nastaven tak, aby si vydělal na svůj běžný provoz i další rozvoj, a nebyl tak závislý na grantech či dostatku dobrovolníků. Po technologické stránce je systém DOI postaven jako aplikace nad systémem handle, který mu poskytuje potřebnou technickou infrastrukturu. Jednou z nejúspěšnějších registračních agentur DOI je *Crossref* [29], jejímž cílem je zajistit pomocí identifikátorů DOI spolehlivé propojování (klikatelné citování) vědecké a odborné literatury na webu od různých nakladatelů – především článků v časopisech a sbornících. Aktuálně je v systému CrossRef zapojeno přes 1600 nakladatelů a odborných společností, kteří již přidělili na 20 miliónů identifikátorů pokrývající přes 14.000 titulů periodik. Měsíčně je novým objektům (knihám, článkům, obrázkům, recenzím, aj.) přidělováno na čtvrt miliónu identifikátorů.

## Interoperabilita

Interoperabilitou rozumíme schopnost spolupráce mezi různorodými (technologicky, provozně, organizačně) digitálními knihovnami. Klasické automatizované knihovní systémy využívají pro svou interoperabilitu *protokol Z39.50* [30], který umožňuje vyhledávat, získávat, případně i modifikovat bibliografické záznamy napříč mezi různorodými knihovními systémy. Pro potřeby digitálních knihoven není ale tento protokol moc použitelný. Příčin je několik. Z39.50 je natolik rozsáhlý a složitý, že jeho implementace je příliš nákladná; vyplatí se jen u velmi rozsáhlých a drahých systémů. Byl navržen v době před vznikem webu, takže nevyužívá dnešních webových technologií, na nichž je většina digitálních knihoven založena. Kvůli jeho složitosti je obtížné přizpůsobovat ho novým potřebám. Navíc současný trend otevřených systémů dává přednost jednoduchým standardům a technologiím, ze kterých je možné vyskládat složitější řešení podle potřeby – podobně jako lze z jednoduchých kostek LEGA skládat rozmanité stavby různé složitosti. Podle schopností, potřeb a invence každého tvůrce.

Jaká řešení mají tedy tvůrci soudobých digitálních knihoven k dispozici? Není jejich mnoho, ale některé přece jen existují. I když jde většinou o technologie spíše jednoduché, se značně omezenou funkcionalitou[3].

Stoupenci protokolu Z39.50 přišli s výrazně odlehčenou verzí protokolu označovaného *SRU/SRW* [31] (Search/Retrieve via URL resp. Search/Retrieve Web Service) postaveného plně na webových technologiích. Spolu s vyhledávacím jazykem *CQL* (Common Query Language) vytváří základ infrastruktury pro implementaci globálních *meta-vyhledávacích služeb*[4]. Tyto protokoly (označované také jako „Z39.50 nové generace") jsou však zatím příliš čerstvé, takže s nimi ještě není dostatek praktických zkušeností.

Větší zkušenosti jsou s další „lehkotonážní" technologií – *Open Archive Initative* (OAI) [32], která je založena na metodě sklízení metadat. Základní princip je jednoduchý: oddělme roli poskytovatele obsahu (ten ať se stará pouze o svá data a není zatěžován starostmi kolem jejich zpřístupnění) a roli poskytovatele služeb (ten sklízí metadata z vybraných digitálních knihoven, ukládá je do své lokální databáze a poskytuje nad nimi služby uživatelům). K tomu, aby mohl být tento princip realizován, bylo třeba standardizovat dvě věci: formát metadat (tím je Dublin Core) a protokol pro sklízení metadat (tím je protokol OAI-PMH – Protocol for Metadata Harvesting, jednoduchá nadstavba nad webovým protokolem HTTP). Protože jak Dublin Core tak OAI-PMH jsou velmi jednoduché, je pro poskytovatele dat snadné implementovat je do jakéhokoliv informačního systému na webu. Pak již záleží jen na kreativitě poskytovatele služeb, co vše bude z webu sklízet a co vše dokáže nad sklizenými metadaty nabídnout.
Jako příklad možného využití technologie OAI uveďme hypotetický návrh realizace národního registru elektronických vysokoškolských kvalifikačních prací eVŠKP: Jednotlivé vysoké školy implementují své vlastní repozitáře eVŠKP (libovolnou technologií) a metadata o jednotlivých pracích ve formátu Dublin Core vystaví pro sběr protokolem OAI-PMH. Pověřená instituce bude tyto metadatové záznamy průběžně sklízet pomocí některého z volně dostupných harvesterů OAI-PMH a ukládat je do své lokální databáze. Nad touto databází vytvoří rozhraní pro webové vyhledávání. Záznamy eVŠKP vyhledaných v národním registru budou obsahovat URL odkazující na plný text práce v repozitáři příslušné vysoké školy. Celý takový systém lze díky jednoduchosti technologie OAI implementovat velmi rychle a s minimálními náklady.

Třetí technologií spadající do oblasti interoperability je standard *OpenURL* [33]. Tento standard je základem pro realizaci *kontextově citlivých* vazeb mezi zdroji na webu[5]. O co jde: standardní odkaz na webu směřuje vždy na jedno a to samé místo. V některých případech by však bylo vhodné, aby cílové místo odkazu bylo různé – podle toho, *kdo* na daný odkaz klikl. Například v bázi Web of Science je citace článku obsahující odkaz (URL) na plný text článku. Klikne-li na odkaz uživatel z Masarykovy univerzity, měl by být nasměrován na plný text článku v databázi zakoupené Masarykovou univerzitou. Klikne-li na ten samý odkaz uživatel z Karlovy univerzity, měl by být nasměrován do fulltextové databáze zakoupené Karlovou univerzitou (což může být úplně jiná databáze než v případě MU). Jak ale takovou „schizofrenní" vazbu na webu realizovat? Jednoduše – odkaz ve Web of Science nesmí směřovat na plný text článku, nýbrž vede na servisní službu, která má informace o předplacených zdrojích jednotlivých zákazníků a provádí automatické přesměrování původního „kliknutí" na správné místo. Jak se ale servisní služba dozví, na co vlastně uživatel klikl, jinými slovy – který článek vlastně požadoval? A tady právě vstupuje do hry OpenURL: podle tohoto standardu jsou totiž metadata požadovaného článku (název, autor, časopis, ročník, číslo, ...) zakódována přímo do výchozí vazby ve Web of Science. Čili namísto odkazu v podobě URL směrujícího na konkrétní plný text článku, je ve Web of Science umístěna vazba v podobě OpenURL směrující na servisní službu a nesoucí v sobě zakódovaná metadata o daném článku. Po kliknutí dostane servisní služba OpenURL i s metadaty článku, zjistí původce kliknutí, ověří, kam má daný uživatel přístup, a tam ho přesměruje na plný text příslušného článku.

---

[3] Příklad webu ukazuje, že jedině jednoduchá řešení mají v praxi šanci na masové rozšíření. Navíc často lze i s jednoduchými technologiemi dosáhnout díky hromadném nasazení překvapivě průlomová řešení.

[4] Na rozdíl od tzv. *federativního vyhledávání*, kdy data pro vyhledávání jsou předem předzpracována (viz OAI či Google, který nejdříve sesbírá data z webu do své lokální databáze a vyhledávání pak probíhá nad touto lokální databází), realizují meta-vyhledávací technologie vyhledávání nad zadanými zdroji přímo, až po zadání příslušného dotazu. Mohou tak získávat aktuálnější výsledky a vyhledávání není omezeno pouze na statické webové zdroje.

[5] Jde též o tzv. „problém správné kopie" (Appropriate Copy Problem).

Velcí producenti komerčních zdrojů si velmi rychle uvědomili potenciál skrytý v technologii OpenURL a své zdroje tomuto standardu záhy přizpůsobili (učinili je tzv. OpenURL-enabled). Díky tomu je dnes možné efektivně propojovat informační zdroje a digitální knihovny od různých producentů nikoliv fixně, ale přesně podle potřeb a požadavků jednotlivých zákazníků. Technologie OpenURL má i další zajímavá využití a patří v současnosti mezi významné stavební kameny digitálních knihoven.

## Další technologie

I v dalších oblastech tvorby a nasazení digitálních technologií existují využitelné standardy či přístupy. Například zajímavé alternativy pro citlivou oblast autorských práv přináší iniciativa *Creative Commons* [34], která nabízí autorům volbu řady méně restriktivních forem ochrany než klasický copyright (namísto tradičního „všechna práva vyhrazena“ nabízí licence typu „některá práva vyhrazena“, pomocí nichž si autor sám stanovuje míru ochrany svých práv, tudíž i míru volnosti použití jeho díla). Reakcí na stále rostoucí ceny vědeckých časopisů produkovaných a šířených tradičními cestami komerčních nakladatelů je alternativa v podobě *otevřeného přístupu*, Open Access. Z této oblasti zmiňme alespoň *Budapešťskou iniciativu* (Budapest Open Access Initiative) [35], která se snaží podporovat volný přístup k vědecké a odborné literatuře prostřednictvím technologií pro samoarchivaci (self-archiving) a časopisů s otevřeným přístupem (open-access journals). V oblasti dlouhodobé archivace rozsáhlých digitálních sbírek je užitečným vodítkem referenční model *OAIS* (Open Archival Information System) [36] zavádějící ucelený komplexní systém opatření a nástrojů pro správu a dlouhodobé udržování digitálních dat.

Mezi stavební kameny infrastruktury digitálních knihoven patří samozřejmě i mnohé obecné technologie vzniklé mimo komunitu vývojářů digitálních knihoven, nejčastěji pak v oblasti gridů, sémantického webu a e-commerce. Jako příklad zmiňme alespoň tři z nich: (a) *webové služby*, Web Services [37], umožňující vytvářet distribuované systémy prostřednictvím vzdáleného volání operací na webu, (b) autentizační systém *Shiboleth* [38] pro distribuovanou webovou autentifikaci napříč různými organizacemi, (c) technologii *peer-to-peer sítí* [39], [40], v níž spolu komunikují přímo klienti bez nutnosti využívat služeb centra, jak je tomu v případě klient-server architektury.

## Systémy pro tvorbu digitálních knihoven

Instituce stojící před úkolem realizovat svou vlastní digitální knihovnu má obvykle dilema – navrhnout a vytvořit vlastní systém, nebo použít některý z volně dostupných nebo komerčních systémů? První varianta je lákavá, slibuje možnost nechat si „ušít“ systém přesně na míru podle potřeb a požadavků instituce. Pokud se ale nejedná jen o vystavení nevelkého množství dokumentů na web, pak implementace komplexního systému digitální knihovny není úkol zrovna snadný. I když je možné a žádoucí použít technologické stavební kameny zmíněné v předchozích částech příspěvku, obvykle zatím není těchto stavebních kamenů ani zdaleka dost na to, aby z nich bylo možné seskládat celou stavbu digitální knihovny. Vývoj vlastní digitální knihovny je investice velmi náročná na finanční a zejména specializované lidské zdroje.

Varianta komerčního řešení snižuje nároky na množství vlastních specialistů, zejména programátorů, nikoliv informačních pracovníků. Existuje velké množství produktů od různých firem nabízených pod různými nálepkami: digitální knihovna, systém pro správu digitálního obsahu či systém pro správu aktiv (content management system, assets management system), atd. Problémem všech těchto systémů může být jejich nedostatečná univerzalita, flexibilita a zejména vysoká cena. Pokud instituce nemá ještě dostatečné zkušenosti v oblasti digitálních knihoven a tudíž nemá ani jasné představy o tom, na co chce daný systém použit, může být přístup – „koupit drahý systém a pak teprve zkoušet, k čemu by mohl vlastně sloužit“ – silně kontraproduktivní. V těchto případech může být rozumnou alternativou volba některého z volně dostupných systémů. I těchto systému existuje velká škála a ani ony pochopitelně nemusí vždy zaručovat požadovanou univerzálnost, flexibilitu či funkcionalitu. Ideální řešení obvykle v praxi neexistují. Nicméně mohou poskytnout vhodný startovací základ, minimálně pro získání potřebných zkušeností a pořízení dostatečně velkého objemu digitálního obsahu, aby bylo později možné zodpovědně uvažovat o dalších variantách.

Ze široké škály volně dostupných softwarových produktů pro tvorbu a provoz digitálních knihoven zmiňme alespoň čtyři: systém *Greenstone* [41] – vyvíjený již delší dobu na University of Waikato na Novém Zélandu (a využívaný i pro potřeby UNESCO), systém *DSpace* [42] – navržený ve spolupráci HP Labs s knihovnami na Massachusetts Institute of Technology jako institucionální repozitář, nyní dále rozvíjený mezinárodní komunitou uživatelů, systém *EPrints* [43] – vyvinutý na University of Southampton jako nástroj pro tvorbu e-printových archivů, a konečně systém *FEDORA* [44] – flexibilní

univerzální digitální repozitář, vyvíjený s podporou Mellonovy nadace týmem na University of Virginia a Cornell University. Podrobněji pojednává o těchto systémech příspěvek V. Krejčíře na této konferenci.

## Závěr: Googlizace a nový pohled na digitální knihovny

Obrovský úspěch vyhledávače Google podpořený následně jeho dravou expanzí do mnoha dalších informačních oblastí (digitalizace, obrázky, mapy, ...) charakterizují někteří odborníci jako tzv. „googlizaci" digitálních knihoven a informací všeobecně. Na jednu stranu přináší tento fenomén nepochybně razantní pokrok a impuls pro rozvoj celého odvětví. Na druhou stranu vzbuzuje často krátkozraké představy o tom, že Google vyřeší všechny problémy za nás – že nemá cenu něco rozvíjet, stačí si prostě jen počkat, až to Google udělá. Navíc, vedle Googlu jsou zde i další hráči. Zatímco v minulosti hrály knihovny pro informační zázemí společnosti klíčovou roli, dnes jsou některé jejich funkce přinejmenším duplikovány dalšími subjekty – vedle zmiňovaného Googlu jsou to třeba komerční poskytovatelé informací a informačních služeb.

Googlizace podněcuje k znovu-zamyšlení nad tím, co vlastně jsou a k čemu by měly sloužit digitální knihovny. Podle [2] nejsou – či neměly by být – jen o vyhledávání („lze to najit?") a přístupu („lze to získat?"). Tyto funkce sice jsou a zůstanou důležité, jsou ale jen částí mnohem širšího informačního prostředí. Klasické knihovny jsou také mnohem více než jen pouhá skladiště knih, map a časopisů. Jsou i místem, kde se lidé setkávají, sdílí a vyměňují si poznatky.

Jaké by tedy měly být digitální knihovny (blízké) budoucnosti? Citátem jsme celý náš příspěvek začali, jiným citátem si dovolím i zakončit:

„Digitální knihovny by se měly vyrovnat tradičním knihovnám a dál dramaticky rozšířit jejich možnosti. Měly by být mnohem víc než jen vyhledávací portál. Podobně jako tradiční knihovny, měly by poskytovat kvalitní *výběr* informačních zdrojů zohledňující profil knihovny, a měly by poskytovat *služby*, které cílové komunitě napomáhají v efektivním využívání těchto zdrojů. Protože však nejsou omezovány fyzickým prostorem, časem a médiem, měly by být mnohem přizpůsobivější a vstřícnější ke svým komunitám uživatelů, než jsou tradiční knihovny. Měly by být *kolaborativní* – umožňovat uživatelům přispívat do knihovny svými znalostmi (ať již aktivně prostřednictvím anotací, recenzí, komentářů apod. nebo pasivně prostřednictvím vzorců chování při využívání zdrojů). Měly by být také *kontextové* – podchycením rozšiřující se pavučiny vztahů a znalostních vrstev mezi vybranými primárními zdroji. V tomto pojetí by jádrem digitální knihovny měla být vyvíjející se informační báze, snoubící profesionální výběr s moudrostí davu (wisdom of crowds)." [2]

## *Literatura:*


[1] WILSON, B. Change and the Need for Innovation. Editorial. *D-Lib Magazine* [online]. April 2006, roč. 12, č. 4. Dostupný z WWW: <http://www.dlib.org/dlib/april06/04editorial.html>.

[2] LAGOZE, C. et al. What is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL. *D-Lib Magazine* [online]. November 2005, roč.11, č.11.
Dostupný z WWW: <http://www.dlib.org/dlib/november05/lagoze/11lagoze.html>.

[3] LYNCH, C., GARCÍA-MOLINA, H. *Interoperability, Scaling, and the Digital Libraries Research Agenda* [online]. IITA Digital Libraries Workshop, 1995.
Dostupný z WWW: <http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.htm>.

[4] DjVu Zone. http://www.djvuzone.org/

[5] MrSID. LizardTech Home Page. http://www.lizardtech.com/

[6] VOJTÁŠEK, F. Využití grafických formátů JPEG a DjVu v digitalizaci. Sborník konference INFORUM 2006. Dostupný z WWW:
<http://www.inforum.cz/inforum2000/prednasky/vyuzitigrafick.htm>.

[7] American Memory. http://memory.loc.gov/

[8] JSTOR. http://www.jstor.org/

[9] The Universal Library – Million Book Project. http://www.ulib.org/

[10] Google Book Search. http://books.google.com/

[11] Open Content Alliance. http://www.opencontentalliance.org/

[12] ArXiv.org. http://arxiv.org/

[13] Research Papers in Economics (RePEc). http://repec.org/

[14] Astronomics Data System (ADS). http://adswww.harvard.edu/

[15] DSpace@Cambridge. http://www.dspace.cam.ac.uk/

[16] University of Southampton EPrints Repository. http://eprints.ecs.soton.ac.uk/

[17] E-LIS. The open archive for Library and Information Science. http://eprints.rclis.org/

[18] Internet Archive. http://www.archive.org/

[19] WebArchiv. http://www.webarchiv.cz/

[20] CiteSeer.IST. http://citeseer.ist.psu.edu/ , http://citeseer.csail.mit.edu/

[21] Dublin Core Metadata Initiative. http://dublincore.org/

[22] MODS. Metadata Object Description Schema. http://www.loc.gov/standards/mods/

[23] METS. Metadata Encoding & Transmission Standard. http://www.loc.gov/standards/mets/

[24] The Making of America II. http://sunsite.berkeley.edu/MOA2/

[25] DLF. Digital Library Federation Home Page. http://www.diglib.org/

[26] PURL. Persistent URL Home Page. http://purl.oclc.org/

[27] The Handle System. http://www.handle.net/

[28] DOI. The Digital Object Identifier System. http://www.doi.org/

[29] Crossref.org Home Page. http://www.crossref.org/

[30] Z39.50 Maintenance Agency Page. http://www.loc.gov/z3950/agency/

[31] SRU: Search/Retrieve via URL. http://www.loc.gov/standards/sru/

[32] OAI. Open Archives Initiative. http://www.openarchives.org/

[33] The OpenURL Framework for Context-Sensitive Services. NISO.
http://www.niso.org/committees/committee_ax.html ,
http://www.niso.org/standards/standard_detail.cfm?std_id=783

[34] Creative Commons. http://creativecommons.org/

[35] Budapest Open Access Initiative. http://www.soros.org/openaccess/

[36] OAIS. Reference Model for an Open Archival Information System.
http://public.ccsds.org/publications/archive/650x0b1.pdf

[37] Web Services.

[38] Shibboleth. http://shibboleth.internet2.edu/

[39] Peer to Peer sítě. Wikipedia. http://cs.wikipedia.org/wiki/P2P

[40] Digital Library Architecture. DELOS Web Site. http://ii.umit.at/research/delos_website/

[41] Greenstone Digital Library Software. http://www.greenstone.org/

[42] Dspace Federation. http://www.dspace.org/

[43] Eprints.org http://www.eprints.org/

[44] FEDORA Digital Repository System. http://www.fedora.info/