Digitální knihovny

Miroslav Bartošek

Ústav výpočetní techniky, Masarykova universita
Botanická 68a, 602 00 Brno
bartosek@ics.muni.cz

Abstrakt. K problematice digitálních knihoven lze přistupovat z mnoha různých pohledů. Jedním z nich může být technické hledisko preferující popis standardů, multimediálních formátů a praktických postupů při digitalizaci a konstrukci digitálních repozitářů. Na opačné straně spektra může stát přístup zkoumající digitální knihovny z pohledu informační vědy a případně i z hlediska jejich nasazení v širším společensko-právním kontextu. Další možnou variantou je popis co největšího počtu projektů a úspěšných realizací. Náš přístup je jiný; snaží se o obecný a pokud možno systémový popis základních oblastí ve výzkumu a praxi digitálních knihoven z pohledu počítačové vědy – obecné architektury, globální identifikační infrastruktury, metadat, interoperability a distribuovaného vyhledávání informačních zdrojů. V každé z výše uvedených oblastí jsou kromě charakteristiky obecných přístupů uvedeny i příklady praktických řešení patřících mezi základní stavební kameny současné infrastruktury digitálních knihoven. Zmíněny jsou nejvýznamnější programy na podporu rozvoje digitálních knihoven a zařazena je také podrobná bibliografie pro zájemce o hlubší studium popisované problematiky.

Klíčová slova: digitální knihovny, definice-DL, architektura, metadata, interoperabilita, globální vyhledávání zdrojů, identifikace informačních objektů.

1 Úvod

1.1 Co je to digitální knihovna ?

Přestože pojem *digitální knihovna* (DL – digital library) patří v posledních letech k těm nejfrekventovanějším, panuje řada nejasností, co vlastně tento termín obnáší. Jednou z příčin je to, že obsah pojmu digitální knihovna se průběžně vyvíjí – tak, jak se vyvíjí jeho technologická základna, výpočetní technika. Jiný důvod souvisí s tím, že problematikou digitálních knihoven se zabývá mnoho různorodých odborných komunit, z nichž každá si vytváří vlastní náplň tohoto pojmu v souladu se svým zaměřením: z pohledu databázového specialisty představuje digitální knihovna informační systém využívající architekturu federativních databází, pro odborníky zabývající se hypertextem a šířením informací je to jen jedna z aplikací nad Webem, knihovník vidí v DL další krok v automatizaci na cestě od knihovny analogové, papírové, přes automatizovanou či hybridní (fyzické sbírky s automatizovaným katalogem) až po digitální (většina či veškeré informace a služby knihovny jsou

digitální). V neposlední řadě přispívá ke zmatení pojmů i skutečnost, že pojmem digitální knihovna jsou někdy označovány systémy, které – přinejmenším z pohledu informačního specialisty – jsou o něčem úplně jiném: soubory algoritmů a procedur, systémy na správu dokumentů, apod.

S masovým rozšířením Internetu po nástupu webových technologií se objevily i představy, že celý Internet resp. Web jsou vlastně jednou digitální knihovnou. S tím však odborníci z oblasti informační vědy nesouhlasí. Clyford Lynch, jeden z předních amerických informačních specialistů, připomíná, že Web nebyl pro podporu organizovaného publikování a vyhledávání informací vůbec navržen. Výstižně to charakterizoval Carl Lagoze v [47]: "*Although the Internet provides access to an enormous amount of information, the current state-of-the-art falls far short of what is commonly viewed as a* library service *– that is, relatively easy navigation of and access to a set of documents that are part of a* collection*. The notion of a collection is important in that it implies that the set of documents was not selected haphazardly, but by some trusted intermediary. Current users of the Internet confront an information space where the quality of documents is far from reliable, facilities for locating documents are primitive, and access to a specific document frequently means wading through a Tower of Babel of architecture dependencies and file formats*."

Přestože od zveřejnění tohoto názoru uplynulo již několik let a vývoj například v oblasti tzv. sémantického Webu dosáhl od té doby pozoruhodných výsledků, má výše uvedená charakteristika stále svou platnost.

Takže, co jsou to vlastně ty digitální knihovny ? Z mnoha desítek existujících „definic" uveďme alespoň dvě. První z nich je velmi obecná a pochází z počítačového prostředí [4]:

- *Digitální knihovna je spravovaná sbírka informací spolu s odpovídajícími službami, přičemž informace jsou uloženy v digitální podobě a jsou dostupné prostřednictvím sítě.*

Klíčovými slovy v definici jsou: spravovaná sbírka informací (collection) – služby – informace v digitální podobě – přístup přes síť. To, že jde o sbírku informací, která je nějakým systematickým způsobem spravována, řízena, má v definici zásadní význam. Proud dat zasílaný družicí na Zemi není knihovnou. Avšak tatáž data, jakmile jsou systematicky uspořádána, stávají se sbírkou v digitální knihovně. Podobně málokdo bude považovat za digitální knihovnu databázi obsahující finanční záznamy jedné společnosti; ale soubor takových záznamů z mnoha společností již může být částí nějaké digitální knihovny.

Druhá charakteristika, převzatá ze [80], pochází z prostředí knihoven a naznačuje, že digitální knihovna v jejím chápání je především knihovnou; vychází z tradičních knihovních funkcí jako je výběr, zpřístupnění a uchovávání materiálu a zdůrazňuje, že digitální knihovny budou vždy budovány tak, aby sloužily konkrétní komunitě uživatelů (představa všeobjímající univerzální DL není v praxi reálná):

- *Digitální knihovny jsou organizace, které poskytují zdroje (včetně specializovaného personálu) umožňující provádět výběr, strukturování a zpřístupnění sbírek digitálních prací, tyto práce dále distribuovat, udržovat jejich integritu a dlouhodobě uchovávat – a to vše s ohledem na snadné a ekonomické využití určitou komunitou nebo množinou komunit uživatelů.*

Z mnoha definic a projektů vyplývají určité společné základní znaky digitálních knihoven:

- pro DL není klíčovou otázkou digitalizace fyzického materiálu, nýbrž organizace elektronické sbírky za účelem lepšího přístupu
- DL obvykle není jedna uzavřená entita (pro zdůraznění tohoto aspektu mnozí autoři používají zásadně a výhradně množné číslo – digitální knihovny)
- informační zdroje tvořící DL jsou *heterogenní* (způsobem uložení-organizací-správou objektů a použitými platformami), *dynamické* (začleňováním a vyřazováním komponent do/ze struktury DL) a *multimediální* (povahou dat)
- realizace DL vyžaduje technologie pro propojení různých (autonomně spravovaných) informačních komponent
- toto propojení musí být pro uživatele transparentní
- cílem je zajistit uživateli jednotný (koherentní) přístup k relevantním digitálním informacím bez ohledu na jejich formu, formát, způsob a místo uložení.

Na vývoji a nasazení digitálních knihoven v praxi se podílí zejména dvě skupiny odborníků. První z nich jsou informační profesionálové (včetně knihovníků, nakladatelů a široké skupiny poskytovatelů informací, jako jsou například indexační a abstrakční služby). Druhou skupinu tvoří počítačoví specialisté a vývojáři Internetu.

1.2 Krátce z historie

Vize digitálních knihoven provází v různých podobách větší část historie výpočetní techniky. Podstatný pokrok však nastal v této oblasti až počátkem 90.let minulého století, kdy prudký rozvoj informačních a komunikačních technologií umožnil začít v praxi realizovat představy teoretiků a efektivně zpřístupňovat první slibné výsledky širokému okruhu uživatelů.

Vraťme se krátce do historie. V literatuře jsou nejčastěji uváděni dva průkopníci, kteří nejvíce inspirovali generace výzkumníků a propagátorů digitálních knihoven. Prvním z nich je Vannevar Bush, profesor MIT a ředitel Národního úřadu pro vědecký výzkum a vývoj USA v období druhé světové války. Ve svém vizionářském článku *As We May Think* publikovaném v roce 1945 [13] se zabýval problémem efektivnějšího „automatizovaného" zpracování odborných informací („*our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their porpose*."). Analyzoval potenciální možnosti, které pro získávání, ukládání, vyhledávání a získávání informací nabízelo využití soudobých technologií a nastínil vizi systému využívajícího fotografické postupy a kompresi dat pomocí mikrofilmů. (Bushem navržený systém Memex koncepčně odpovídá dnešnímu osobnímu počítači, v němž jsou informace provázány asociativními vazbami – předchůdce hypertextu a koncepce dnešního webu).

Druhou často citovanou osobností je J.C.R.Licklider, který v 60. letech minulého století studoval na MIT možnosti transformace knihoven s využitím digitálních počítačů (na rozdíl od Bushe, který – ačkoliv již číslicové počítače znal – vycházel ještě z analogových technologií). V roce 1965 publikoval knihu *Libraries of the Future*, v níž identifikoval výzkum a vývoj potřebný k realizaci skutečně použitelné digitální knihovny a nastínil vizi digitální knihovny po 30 letech – tedy v roce 1994. V obecné rovině jsou jeho předpovědi pozoruhodně přesné a mnohé z nich se vyplnily, i když ne vždy v jim očekávané podobě; celkově výrazně podcenil to, čeho

všeho se dá dosáhnout využitím hrubé výpočetní síly a naopak přecenil pokroky založené na rozvoji umělé inteligence a počítačových metod zpracování přirozeného jazyka.

V šedesátých letech se také objevují první významné praktické výsledky v nasazení výpočetní techniky pro zpracování informací v knihovnách, mezi které bezesporu patřil jednak vývoj formátu MARC (Machine-Readable Cataloguing) v Kongresové knihovně USA (Library of Congress) standardizujícího strukturu bibliografického záznamu v elektronické podobě a využití tohoto formátu pro sdílenou katalogizaci knihoven v systému OCLC, jednak rozvoj online knihovních katalogů (knihovníky označovaných termínem OPAC, Online Public Access Catalogue). Navzdory všem překážkám vyplývajícím z tehdejších technických omezení podnítily tyto první výsledky řadu optimistických předpovědí. Jeden příklad za všechny: A.L.Samuel předpovídal v [69] v roce 1964, že do 20 let papírové knihovny zaniknou. Důvody, proč se většina předpovědí ze 60.let nenaplnila, byly samozřejmě různé; často však mezi ty hlavní patřily důvody finanční. Pro vyplnění Samuelovy vize by bylo třeba zdigitalizovat zhruba 100 miliónů různých knih, přičemž údaje z amerického prostředí [14] uvádí cenu digitalizace v rozmezí 2-6 USD za stránku; ještě mnohem větší část nákladů by bylo ovšem třeba na kompenzace autorských práv.

Počátkem 90. let začíná v oblasti digitálních knihoven skutečný boom. Zásluhu na tom měla skutečnost, že technologický pokrok ve všech třech pro DL kritických oblastech zahrnujících

- *computing* (výpočetní a krátkodobá i dlouhodobá paměťová kapacita)
- *communications* (globální síť a přenosová kapacita)
- *content* (množství informace v digitální podobě)

dosáhl dostatečně vysokého stupně při rozumně nízké jednotkové ceně a široké všeobecné dostupnosti, což umožnilo začít realizovat projekty reagující na reálné potřeby uživatelů. To vše odstartovalo prudký rozvoj v oblasti digitalizace, elektronického publikování a šíření informací, což přineslo i nový impuls pro výzkum a vývoj v oblasti digitálních knihoven (dalšími výraznými podněty bylo masové celosvětové rozšíření webových technologií a všeobecná potřeba efektivnějšího sdílení vědeckých poznatků). Vyspělé země podpořily tento trend zřízením štědře dotovaných programů na podporu výzkumu a vývoje (nejvýznamnějším z nich byl americký program DLI-1, Digital Library Initiative Phase 1 a na něj v současnosti navazující DLI-2), ale i prakticky orientovaných projektů (například britský program eLIB). Podrobněji se o nich zmíníme v závěru příspěvku.

1.3 Proč digitální knihovny

Počáteční představa digitální knihovny vycházela z koncepce klasické knihovny a byla orientována především na digitalizaci existujících sbírek jako nástroje pro zlepšení klasických knihovních služeb, zejména v následujících oblastech:

- vzdálený a nepřetržitý přístupu k informacím
- efektivnější metody vyhledávání (např. fulltextové)
- lepší využití fondu (souběžný přístup k jednomu a témuž dokumentu)
- sdílení informací mezi různými knihovnami

- dokonalejší ochrana fondu (nahrazení zranitelných fyzických objektů digitálními).

Záhy se však ukázalo, že potencionální možnosti digitálních knihoven jdou nad rámec možností klasických knihoven s fyzickými dokumenty, a projevují se například možnostmi neomezené globální integrace digitálních repozitářů v celosvětovém měřítku, novými formami a formáty informací, možností permanentní aktualizace informace uložené v digitální knihovně, nebo zcela novými typy služeb (přeformátováváním dokumentů on-fly do různých formátů či dokonce jazykových verzí, vytvářením složených děl, vyjednáváním autorských a přístupových práv, aj.). Přes tyto odlišnosti a přestože provozně ani organizačně nemusí mít digitální a klasické knihovny vůbec nic společného, principiálně mají řadu shodných rysů:

1. systematicky budovanou sbírku *datových objektů*
2. obsahovou analýzu datových objektů ve sbírkách a z ní vyplývající soubory *metadatových struktur* (katalogy, rejstříky, indexy, tezaury)
3. množinu *služeb* (přístupové metody, správa dat, akvizice, vyhodnocování, referenční služby, SDI)
4. tématické zaměření
5. sledování kvality
6. dlouhodobé uchovávání materiálu.

Metody a postupy klasických knihoven jsou za mnoha staletí svého vývoje dobře propracovány a tvoří ucelený efektivně fungující systém. Digitální knihovny však přináší nové výzvy a problémy, pro jejichž řešení nelze často klasické postupy použít vůbec nebo jen ve velmi omezené míře. Po počátečním optimismu z první poloviny 90.let se ukázalo, že problém budování funkčních digitálních knihoven je mnohem složitější, než se zdálo. Principiálním problémem a základem všech obtíží je nedostatečně propracovaná technologie na straně jedné a nepřipravené společenské prostředí zahrnující složitý komplex navzájem provázaných problémů z oblasti ekonomické, právní, sociální a etické na straně druhé. To, na co měly klasické knihovny dlouhá staletí, musí digitální knihovny řešit za pochodu a během několika málo let.

1.4 Aktuální stav, hlavní současné aktivity a zdroje informací

V oblasti digitálních knihoven probíhá v současnosti velké množství aktivit jak z oblasti *základního a aplikovaného výzkumu* (vůdčí roli v tomto směru hrají zejména Spojené státy s množstvím nejrůznějších odborných aktivit především na univerzitách a ve velkých výzkumných knihovnách v čele s Kongresovou knihovnou), tak i *praxe*, kde existují stovky velmi rozsáhlých a ambiciózních projektů zaměřených ať již na digitalizaci či budování konkrétních DL poskytujících cenné informace a služby příslušným komunitám, nebo na implementaci nových prototypů ověřujících v praxi nové přístupy, potřeby a chování uživatelů. Oproti situaci z konce minulého století je u těchto projektů znát posun od „experimentování“ k budování infrastruktury. Ačkoliv dosud neexistuje žádné ucelené, univerzální a všeobecně přijaté řešení DL a v mnoha směrech chybí potřebná globální infrastruktura, která by umožnila škálovat a propojovat znalostní sítě reprezentované jednotlivými DL obdobně, jako je tomu dnes u komunikačních sítí reprezentovaných Internetem a Webem, je již k dispozici řada

základních technologických kamenů v podobě standardů (Z39.50, Dienst, Dublin Core, Handles) a volně dostupných nástrojů pro implementaci základních funkcí digitálních knihoven (za všechny uveďme balík Greenstone software z University of Waikato na Novém Zélandu [33]). Velmi často se výzkum v oblasti digitálních knihoven překrývá s jinými oblastmi jako jsou E-commerce (metadata, interoperabilita, bezpečnost) nebo Distributed Knowledge Environment.

Dění v oblasti digitálních knihoven mapuje řada časopisů, konferencí, specializovaných workshopů a také courseware – kursů na vysokých školách [31].

a) časopisy:

- *D-Lib Magazine* [16] – elektronický časopis zaměřený hlavně na prakticky orientovaný výzkum v oblasti DL; vychází měsíčně od roku 1995 v CNRI s podporou DARPA. Je volně dostupný na Webu a v současnosti je patrně nejprestižnějším zdrojem odborných informací o dění v oblasti DL
- *International Journal on Digital Libraries* – klasický tištěný časopis z nakladatelství Springer-Verlag. Specializuje se spíše na teoretický výzkum, vychází od roku 1997, bohužel však s nepravidelnou periodicitou.
- *Ariadne* [2] – elektronický časopis pro informační specialisty zejména z UK. Informuje o aktivitách z oblasti DL, vydává ho čtvrtletně UKOLN (UK Office for Library and Information Networking); volně dostupný na Webu
- *RLG DigiNews* [68] – elektronický časopis zaměřený na oblast digitalizace a uchovávání digitální informace. Je vydáván Cornellovou univerzitou ve spolupráci s organizací RLG – Research Libraries Group sdružující kolem 160 výzkumně zaměřených knihoven, archivů a dalších paměťových institucí převážně z USA. Vychází dvakrát měsíčně, je volně dostupný na Webu.

Problematice DL byla věnována také některá speciální čísla časopisů přehledově zaměřených na informační technologie, jako *Communications of the ACM* (vrací se k digitálním knihovnám vždy pravidelně po 3 letech, viz čísla z dubna 1995, z dubna 1998 a z května 2001) nebo *IEEE Computer*.

b) konference:

Z těch nejvýznamnějších je třeba uvést pravidelně každoročně od roku 1996 pořádané konference ADL – Advances in Digital Libraries (IEEE) a ACM Conference on Digital Libraries (od roku 2001 pořádané společně pod názvem JCDL – Joint Conference on Digital Libraries) a evropskou konferenci ECDL – European Conference on Research and Advanced Technology for Digital Libraries.

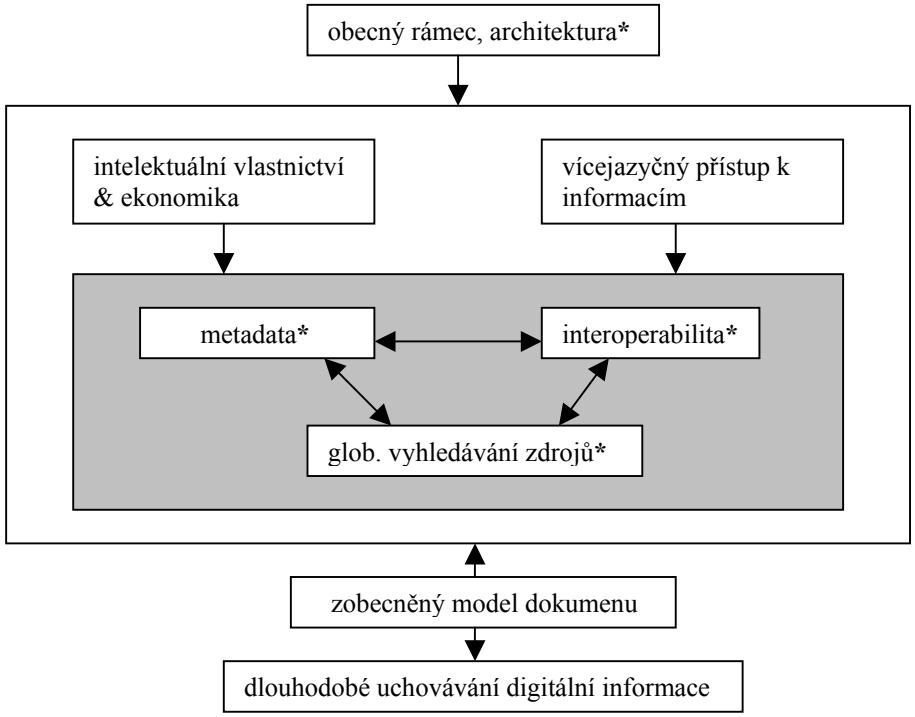
c) metaseznamy:

- IFLA Electronic Collections [39]
- Berkeley Digital Library Sunsite [11]
- ARL Digital Initiatives Database [3]

Vyšlo i několik monografií, z toho dvě přehledové: práce [48] je prakticky zaměřená práce a popisuje spíše technické aspekty DL, nejnovější kniha o DL [4] má povahu obecné přehledové encyklopedie přes celou oblast. Řada dalších monografií se věnuje již konkrétním dílčím aspektům DL, jako jeden příklad za mnohé uveďme [45] se zaměřením na problematiku digitalizace obrazové informace.

2 Klíčové oblasti výzkumu a praxe digitálních knihoven

Termín „digitální knihovny“ je typicky zastřešující pojem. Problematika digitálních knihoven a aspekty jejich realizace jsou totiž natolik široké, že se s trochou nadsázky dá říci, že pod tento pojem lze schovat „téměř cokoliv“ z mnoha oblastí počítačové vědy (databáze, informační systémy, umělá inteligence, počítačové sítě, bezpečnost), ale navíc i mnoho aspektů z řady společenských věd (z knihovní a informační vědy, práva, ekonomie, sociologie, psychologie, lingvistiky). Takové bezbřehé pojetí nám však příliš nepomůže. Zaměříme-li se na oblasti, které jsou pro digitální knihovny skutečně klíčové, dostaneme následující obrázek (adaptováno dle [71]); oblasti popisované v další části příspěvku jsou označeny hvězdičkou:



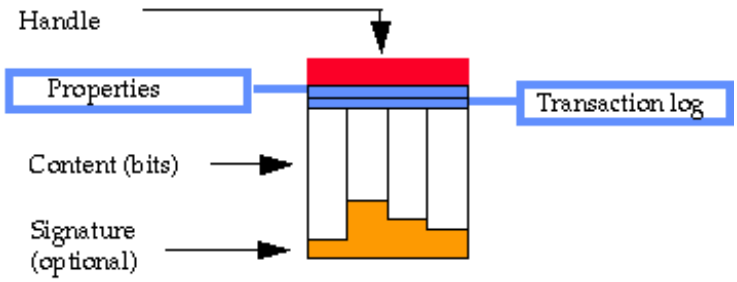
3 Obecný rámec a architektura digitální knihovny

K tomu, aby mohl vzniknout globální systém kooperujících digitálních knihoven, je třeba vytvořit a uvést do života potřebnou globální informační infrastrukturu. Obecná architektura digitální knihovny navržená na dostatečně vysoké úrovni abstrakce umožňuje formalizovat představy o funkcích a fungování digitálních knihoven a současně identifikovat „middleware“ Internetu potřebný pro realizaci distribuovaných digitálních informačních služeb (budoucnost DL je dnes s budoucností Internetu pevně spojena).

3.1 Kahn-Wilenského architektura

Nejpropracovanější obecnou architekturu digitálních knihoven podali Kahn a Wilensky v [44]; experimentální systém vycházející z této architektury byl pak realizován v rámci National Digital Library Project v Kongresové knihovně [6].

Základním prvkem architektury je *digitální objekt*, datová struktura pro základní samostatně použitelnou informační jednotku tvořená dvěmi základními částmi: obsahem (content) a klíčovými metadaty (tvořenými globálním jednoznačným identifikátorem digitálního objektu, označovaným jako *handle*, a dalšími blíže nespecifikovanými neměnnými meta-údaji, například 'autor'). Obsahem digitálního objektu může být buď sekvence bitů reprezentující konkrétní digitální materiál (může být zahrnut i ve vícero formách), množina jiných datových objektů (složený objekt), množina identifikátorů objektů (meta-objekt), případně jiné datové typy – poskytuje tak dostatečnou flexibilitu pro reprezentaci libovolně složitých informačních objektů a vztahů mezi nimi. Digitální objekty mohou být buď proměnlivé (obsah objektu lze měnit i po jeho uložení do repozitáře – ať již jde o nárazové změny nebo přímo dynamické informační objekty) nebo fixní. Schéma jednoduchého digitálního objektu ukazuje následující obrázek:



Podle typu materiálu mohou být digitální objekty rozděleny do *kategorií* (např. text v SGML, počítačový program, digitalizovaný zvuk) a pro každou kategorii mohou být stanovena pravidla pro převod materiálu do jednotlivých typů digitálních objektů, struktura metadat apod. Tak je tomu např. v realizovaném systému [6]; obecná architektura ovšem úmyslně s žádnými specifickými typy materiálu nepracuje, aby udržela co nejvyšší míru obecnosti, neomezovala/nepředjímala budoucí technologický vývoj a ponechávala dostatečnou míru flexibility pro konkrétní implementace.

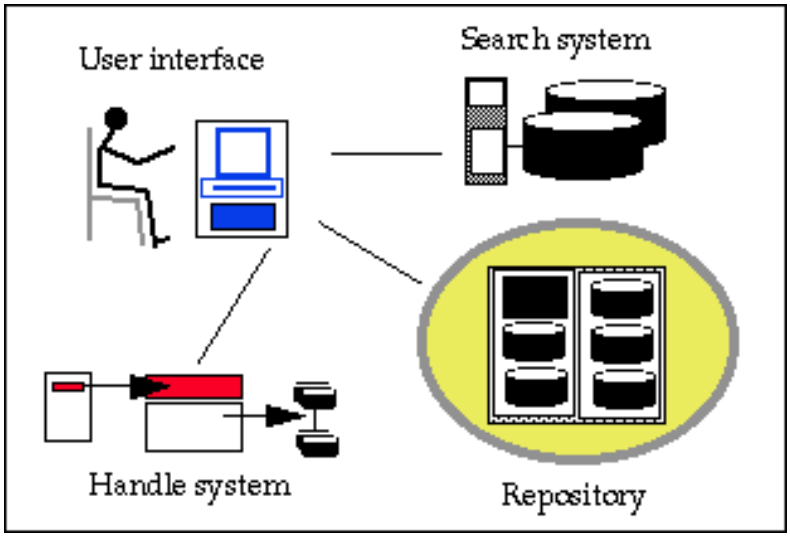
Digitální objekty jsou uloženy v *repozitářích,* které mají přiřazeno jednoznačné globální jméno. Repozitář umožňuje řízený přístup k digitálním objektům v něm uloženým. Pro každý digitální objekt udržuje dva typy metadatových záznamů. Prvním z nich je *záznam vlastností objektu* (properties record) obsahující údaje například o autorských právech a podmínkách pro zpřístupnění digitálního objektu, technické vlastnosti jako formáty a přístupové protokoly, bibliografické údaje, nebo administrativní data (např. datum/čas uložení objektu do repozitáře). Druhým je *transakční záznam* (transaction log) zaznamenávající veškeré transakce repozitáře týkající se daného digitálního objektu. Spolu s neměnnými a na repozitáři nezávislými

klíčovými metadaty tvoří tyto dva záznamy základní metadatový popis digitálního objektu.

Každý repozitář komunikuje s okolím prostřednictvím jednoduchého *repozitářového přístupového protokolu* RAP (Repository Access Protocol) umožňujícího ukládání digitálních objektů, zpřístupnění digitálních objektů, případně další operace – to vše při zajištění odpovídajícího zabezpečení. Digitální knihovna může sestávat z mnoha repozitářů různých typů.

Další komponentou architektury je tzv. *handle-system* sloužící jako globální distribuovaný resoluční mechanismus, který pro digitální objekt identifikovaný svým identifikátorem vrátí seznam repozitářů, které tento objekt udržují. Handle-system byl v praxi úspěšně realizován v Corporation for National Research Initiatives (CNRI) a patří dnes mezi jedny z nejpropracovanějších a v DL-projektech nejužívanějších systémů pro globální identifikaci a resoluci informačních objektů na Internetu [36]. Podrobnější popis je uveden níže v kapitole o identifikátorech.

Schéma kooperace jednotlivých komponent digitální knihovny je naznačeno na dalším obrázku:



1. **search:** uživatel specifikuje svůj požadavek přes high-level uživatelský interface; ten ho přeformuluje na dotaz pro globální vyhledávací systém, který vrátí seznam informačních zdrojů splňujících požadavek uživatele
2. **select:** uživatel vybere ze seznamu informační zdroj, který ho zajímá
3. **retrieve1:** uživatelský interface předá identifikaci digitálního objektu odpovídajícího zvolenému informačnímu zdroji globálnímu resolučnímu systému a získá identifikaci příslušného repozitáře
4. **retrieve2:** přes příslušný RAP si uživatel vyžádá potřebný digitální objekt z repozitáře
5. **display:** získaný objekt zobrazí uživateli.

Z diskusí nad výše naznačenou obecnou architekturou vyplynula řada podmínek klíčových pro její realizaci v praxi [7]. V dalším textu rozebereme alespoň tři z nich: (a) technický návrh musí být začleněn do konkrétního právního a sociálního prostředí; (b) uživatelé požadují intelektuální díla a ne digitální objekty; (c) klíčovými stavebními bloky informační infrastruktury jsou jména a identifikátory

3.2 Začlenění do právního a sociálního prostředí

První informační systémy na Internetu byly vytvořeny konkrétními odbornými a vědeckými komunitami s ohledem na jejich vlastní potřeby; primárním cílem byla rychlá efektivní a zejména *bezplatná* výměna informací. Mnohé z těchto systémů jsou velmi úspěšné dodnes a dále se rozvíjí (jako dva příklady z oblasti elektronického publikování připomeňme systém RFC používaný IETF a pre-printový server E-Print Archives [8] v Los Alamos sloužící celosvětové komunitě fyziků). Situace u obecných digitálních knihoven je ale mnohem složitější, neboť aby byly v praxi použitelné, musí respektovat mnohem širší ekonomický, sociální a právní kontext. Typický příklad: hudební díla představují živobytí jak skladatelů tak hudebníků, kteří vyžadují poplatky za jejich používání (stejně tak i nahrávací studia). Taková díla se mohou stát součástí digitální knihovny pouze a jen tehdy, pokud bude digitální knihovna podporovat jejich zájmy (viz kauza Napster v posledních letech).

Legislativní rámec upravující způsob klasického vytváření, publikování a využívání intelektuálních výtvorů, který zahrnuje množství složitě provázaných konceptů od copyright, provedení díla, fair-use, soukromí, ochrany osobnosti, až po komunikační zákony, daně, národní bezpečnost aj. – se utvářel velmi dlouhou dobu tak, aby vybalancoval zájmy všech subjektů (zájem autorů tvořit, vydavatelů vydávat, uživatelů díla využívat, společnosti ochraňovat produktivní prostředí a zajišťovat bezpečnost). Digitální publikování a šíření informací, nebude-li právně vhodně ošetřeno, představuje pro tento složitě vybalancovaný systém hrozbu s mnoha potenciálně obrovskými dopady (dle [4] představuje informační a zábavní průmysl asi 5% ekonomiky USA; radikální změna technologie může znamenat obrovské ekonomické změny vedoucí až k zániku velkých firem a celých odvětví, a to i s příslušnými sociálními dopady. Nejde přitom jen o komerční sektor, který je samozřejmě rozhodující; ale i taková Kongresová knihovna má 4.500 zaměstnanců...). Právní situace kolem digitálních knihoven je o to složitější, že vzhledem k jejich globálnímu charakteru nestačí upravit a aplikovat národní legislativu, ale je třeba dojednat, vytvořit a uvést do života odpovídající legislativu na mezinárodní úrovni.

Hlavní problém spočívá v tom, že vytvoření potřebného legislativně-společensko-ekonomického prostředí není již problém technologický, ale společenský a jako takový je mnohem složitější a časově mnohem náročnější. Zkušenost také ukazuje, že ne vše se dá předvídat, že nejprve musí ve společnosti vzniknout určité vzorce chování a ty lze teprve právně kodifikovat.

3.3 Hierarchická abstrakce intelektuálního díla (IFLA model)

Digitální objekty jsou sice základními stavebními kameny obecné architektury DL, uživatelé ale obvykle potřebují odkazovat na informační zdroje na vyšší úrovni abstrakce. Navržená architektura umožňuje reprezentovat libovolně složité objekty a vztahy mezi nimi, pro její naplnění je však třeba aplikovat nějakou všeobecně

přijatelnou a dostatečně obecnou abstrakci informačních objektů. Nejrozšířenější a nejvyužívanější kategorizace informačních objektů pochází ze studie *Functional Requirements for Bibliographic Records* [41] organizace IFLA (International Federation of Library Associations and Institutions) z roku 1998 a rozlišuje čtyři úrovně:

- *dílo (work)* – intelektuální či umělecký výtvor jako abstraktní koncept (Homérova Illiada, Beethovena 5.symfonie, operační systém Unix, apod.)
- *vyjádření (expression)* – konkrétní realizace, časoprostorová fixace, daného díla (například Illiada byla nejprve realizována ústním podáním, poté formou písemného zápisu; symfonie může být realizována jako zápis partitury nebo jako některé z mnoha jejich hudebních provedení)
- *projev (manifestation)* – fyzické „ztělesnění" nějakého vyjádření daného díla (například text Illiady může být „projeven" v několika různých rukopisech nebo v různých knižních vydáních; určité provedení symfonie může být zaznamenáno na hudební CD nebo na videokazetě jako záznam televizního přenosu)
- *jednotka (item)* – jeden (z potenciálně mnoha) exemplářů daného projevu, kopie (například jeden výtisk daného vydání knihy, konkrétní výtisk podepsaný autorem, kopie souboru, apod.).

Stručně vyjádřeno: *dílo* je realizováno prostřednictvím jednoho či více *vyjádření*, to je materializováno v jednom či několika různých *projevech*, ty jsou pak rozmnoženy v jedné či mnoha *jednotkách*. Praktické zkušenosti ukazují, že tento 4-úrovňový model je schopen plně postihnout všechny aspekty uživatelských zájmů o informační objekty nejen v oblasti digitálních knihoven, ale i v ostatních oblastech (např. e-commerce). Navíc, na rozdíl od jiných modelů v jiných komunitách, nechybí knihovnám jasně propracovaný systém jeho uplatňování (například přesný překlad díla je považován za jeho nové vyjádření, zatímco volný překlad může být novým dílem; podobně je za nové dílo považována změna žánru – například dramatizace nějaké novely).

Přesné rozlišování různých abstrakcí informačních objektů je užitečné a v některých případech dokonce zcela zásadní nejen v oblasti digitálních knihoven (pro vyhledávání, odkazování, správu vlastnických a autorských práv); oblast e-commerce je toho typickým příkladem (viz např. [38]).

4 Jména a identifikátory

Čím více se při výměně informací (resp. obchodu) snižuje potřeba fyzického kontaktu mezi uživatelem informace a jejím poskytovatelem (resp. mezi kupujícím a prodávajícím), tím víc roste potřeba být schopen věci jednoznačně pojmenovávat a identifikovat. Schopnost jednoznačně globálně identifikovat informační objekty (a tuto identifikaci dynamicky jednoznačně propojit s informačním objektem nacházejícím se kdekoliv v globální síti) je *zcela zásadní* pro nasazení jakéhokoliv distribuovaného globálního informačního systému.

4.1 Koncept URN

Stávající Internet zatím nenabízí dostatečně univerzální, všeobecně podporovaný a rozšířený identifikační systém pro informační objekty, který by splňoval základní požadavky zformulované již počátkem 90.let pro koncepci URN, Uniform Resource Name [76]:

- *globální rozsah* : dané jméno je celosvětově jednoznačné a nezávislé na lokaci
- *persistence* : přidělené jméno trvá „na věky“, i po zaniknutí popisované entity
- *škálovatelnost* : jméno musí být přidělitelné pro jakýkoliv představitelný typ entity
- *legacy support* : systém musí podporovat existující identifikační systémy
- *rozšiřitelnost* : musí umožnit budoucí rozšíření identifikačního schématu

Stávající URL tyto požadavky nesplňuje, neboť identifikuje lokaci, nikoliv entitu (intelektuální obsah). Smyslem návrhu URN je naopak identifikovat entitu bez ohledu na její momentální umístění; musí ovšem existovat tzv. *resoluční mechanismus*, který pro zadané URN zjistí aktuální umístění entity tímto URN identifikované. Syntaxe URN je následující:

URN:nid:nss

kde *nid* (Namespace Identifier) je identifikátor určitého identifikačního systému (třeba DOI, viz níže), a *nss* (Namespace-specific String) je konkrétní identifikátor v daném systému. Jak je vidět, URN „nespoléhá“ na jediný identifikační systém, ale naopak poskytuje zastřešení pro neomezený počet schémat splňujících stanovené podmínky (zahrnujících i popis technik pro realizaci resolučního mechanismu - viz RFC-2611).
Přestože obecná idea URN je jasná a také návrh jeho infrastruktury byl nedávno již dokončen, implementace globálních řešení na Internetu jsou zatím omezené. Příčiny:

- globální resoluční systém pro URN (na bázi DNS) ještě není po celém Internetu rozšířený
- stávající www-prohlížeče zatím nepodporují URN tak, jak podporují URL
- panují nejasnosti, kdy a kterým entitám URN přidělovat (problém verzí, různých formátů, úrovní)
- problém jednoduché (single-point) nebo násobné (multiple) resoluce: resoluční mechanismy vycházející ze současné www-technologie vrací jediné URL, zatímco služby založené na URN obecně mohou vyžadovat identifikovat více instancí entity či více k ní odpovídajících služeb
- v neposlední řadě je to otázka finanční: identifikátory URN jsou sice zdarma, ale budování a udržování resolučních služeb nikoliv – náklady jsou obrovské a někdo to musí zaplatit; žádní dobrovolníci zatím na obzoru nejsou...

Knihovny se problémem identifikace fyzických informačních zdrojů zabývají již dlouho; popišme současný stav a co z toho je k dispozici i digitálním knihovnám [34].

4.2 Klasické identifikátory: ISBN, ISSN, SICI/BICI, ISTC

Již zhruba 30 let používají knihovny a nakladatelé identifikaci ISBN (International Standard Book Number), ISSN (International Standard Series Number) a další identifikátory k identifikaci tištěných publikací, neboli projevu díla (v terminologii IFLA modelu). Digitální knihovny a e-publikování však vyžadují komplexnější vícevrstvou identifikaci, počínaje samotnými autory a konče těmi nejmenšími

jednotkami informací, s nimiž lze v Internetu samostatně manipulovat a prodávat je, jako jsou např. články v časopisech. Současný stav v oblasti standardů je následující:

a) autoři:
ISADN (International Standard Authority Data Number) – umožňuje jednoznačně identifikovat každého autora. Zatím v praxi nerealizováno (již několik let se diskutuje otázka jeho skutečné potřebnosti a technické realizovatelnosti; sílí pozitivní názor)

b) dílo:
ISTC (International Standard Textual Work Code) – pro textová díla
ISAN (International Standard Audiovisual Number) – pro audiovizuální díla
ISWC (International Standard Musical Work Code) – pro hudební díla.
Na všech třech standardech ISO v současnosti intenzivně pracuje; nejdále jsou práce na ISTC, na podzim 2001 je plánována publikace jeho prvního veřejného návrhu. Připravuje se též zahájení prací na normě ISXX pro obrázky.

c) projev:
ISBN, ISSN, ISMN, a další... to, co je k dispozici a již dlouhou dobu se využívá

d) komponenta:
SICI (Serial Item and Contribution Identifier) – pro články
BICI (Book Item and Component Identifier) – pro kapitoly v knize, obrázky, poznámky apod. Norma SICI existuje jako ANSI/NISO standard již od roku 1991, ale zatím se v praxi moc nevyužívá; má však budoucnost. BICI je v pozici pracovního návrhu standardu do ledna 2002.

Popis jednotlivých standardů dobře charakterizuje rozdílné přístupy ke koncepci identifikátorů:

ISBN – International Standard Book Number

Čísla ISBN přidělují nakladatelé. Podle normy ISO 2108-1978 začíná číslo ISBN vždy zkratkou ISBN následovanou 10-místným číslem rozděleným do čtyř bloků proměnlivé délky oddělených spojovníkem, např. ISBN 80-00-01978-9. První blok přiděluje mezinárodní agentura ISBN a identifikuje zemi, v níž nakladatel působí (0 a 1 anglická oblast, 80 ČR a SR). Druhý blok přiděluje národní agentura ISBN a identifikuje nakladatele. Může mít délku 2-6 číslic; čím větší nakladatel, tím kratší jeho číslo. Třetí blok přiděluje nakladatel a určuje konkrétní vydání knihy či její formy; délka je volena tak, aby celková délka čísla ISBN byla deset znaků. Poslední blok tvoří kontrolní číslice, která se vypočítává z předchozích devíti cifer podle modulu 11 s využitím váhových koeficientů. Národní agentura shromažďuje informace o všech přidělených ISBN v dané zemi.

ISBN je příkladem tzv. *inteligentního* či složeného identifikátoru, který kromě vlastní identifikace nese ještě další explicitní informaci (země, nakladatel). Systém je u klasických fyzických informačních zdrojů velmi úspěšný, ale v digitálním světě má problémy:

1. na Internetu může být nakladatelem kdokoliv – což vede k exponenciálnímu nárůstu požadavků na nakladatelská čísla; částečně se to dá řešit vyčleněním

vyhrazených identifikátorů pro knihy publikované jednotlivci, ale obecně není tento systém pro potřeby Webu dostatečně flexibilní;

2. dramatický nárůst publikací po vzniku elektronického publikování a také to, že ISBN je často přidělován i pro menší informační jednotky než kniha, vede k tomu, že ve velmi krátké době (odhaduje se do roku 2010) se prostor ISBN čísel vyčerpá!

Z výše uvedených důvodů vyžaduje systém ISBN důkladnou revizi, a to nejpozději do roku 2006. Přitom jakákoliv změna dosavadního systému bude mít obrovské dopady s velkými náklady na knihovní a informační sektor, srovnatelnými s náklady na řešení problému Y2K. Jedním z navrhovaných řešení (které není sice koncepční, ale umožní získat čas) je rozšířit ISBN na 13 cifer tím, že na jeho začátek bude přidán kód EAN „978“ (knihy), používaný celosvětově v obchodu a e-commerce. Po dohodě s EAN by pak totiž bylo možné využít pro ISBN i druhý prefix „979“ (hudebniny) a číselný prostor ISBN tak zdvojnásobit

ISSN – International Standard Series Number

Na rozdíl od ISBN je číslo ISSN (ISO norma 3297-1998) tzv. *hloupý* či jednoduchý identifikátor, který v sobě nenese žádnou sémantiku. Má tvar 8 cifer rozdělených do dvou bloků po čtyřech cifrách oddělených spojovníkem, např. ISSN 0885-2308. Posledním znakem je kontrolní znak (obdobně jako u ISBN). Všechna čísla ISSN jsou přidělována a centrálně spravována Mezinárodním centrem pro ISSN; na jaře 2001 obsahovala jeho databáze kolem jednoho miliónu záznamů (celková kapacita je deset miliónů). Spolu s každým přiděleným ISSN je v databázi uložen metadatový záznam o příslušném periodiku či seriálové publikaci.

Elektronické časopisy zatím kapacitu ISSN vážněji neohrožují (ročně se zatím přiděluje kolem 50.000 čísel), problém je však s velmi krátkým poločasem jejich rozpadu. Navíc elektronické časopisy nemusí být vydávány v ročnících, svazcích a jednotlivých číslech, takže podle posledních aktualizací katalogizačních pravidel může být za kandidáta na přidělení čísla ISSN považována každá webovská stránka, pod níž jsou shromažďovány nové dokumenty.

SICI – Serial Item and Contribution Identifier

I když je SICI americkou normou již od počátku 90.let (ANSI/NISO Z39.56, viz [59]), zatím se příliš nevyužívá. Důvodem může být jak neexistence jeho podoby v mezinárodní ISO normě, tak skutečnost, že mezi nakladateli je zatím o něm poměrně malé povědomí a může jim připadat poměrně složitý. Chybí také mezinárodní centrum, které by využívání tohoto standardu dostatečně propagovalo.

Příklad: článek Marka Needlemana „Computing resources for an online catalog – 10 years later“ publikovaný v časopise Information technology and libraries, svazek 11, číslo 2 (červen 1992), str. 168-, bude mít SICI:

0730-9295(199206)11:2<168:CRFAOC>2.0.TX;2-#

Identifikátor je tvořen ISSN číslem časopisu následovaným údaji o čísle, údaji o článku (první písmena slov z názvu) a kontrolní částí (verze standardu 2.0, typ zdroje je tištěný text TX). SICI je příkladem identifikátoru, který může být plně „vypočítán“,

tj. automaticky vygenerován přímo z článku nebo jeho metadat. Identifikátor BICI vypadá podobně, jeho standardizace ale ještě není dokončena.

ISTC – International Standard Textual Work Code

Podle pracovní verze návrhu standardu z února 2001 tvoří číslo ISTC šestnáct hexadecimálních cifer rozdělených do čtyř bloků, např.

ISTC 0A9-2002-12B4A105-6

První blok představuje identifikátor registrační agentury, kterých může být až 4096. Každá z nich může přidělit až miliardu čísel každým rokem až do roku 9999. Jedním z požadavků na agenturu je schopnost vytvářet a udržovat metadata pro díla; přirozenými kandidáty se tak stávají například národní knihovny (množství všech existujících textových intelektuálních výtvorů zahrnujících mj. i články je obrovské a jejich kompletní katalogizace je nepředstavitelně náročný úkol). Druhý blok představuje rok, třetím je identifikátor díla a posledním kontrolní číslice.

Kromě dosud uvedených identifikačních schémat existují mnohá další, ať již ve formě oficiálních standardů nebo standardů de-facto (jakým je také PII – Publisher Item Identifier), viz například [64]. Pro všechny výše uvedené příklady identifikátorů lze na Internetu implementovat příslušný globální resoluční systém s využitím koncepce URN. U „hloupých" identifikátorů typu ISSN (identifikátor nenese sám o sobě žádnou informaci o tom, kde hledat informaci o informačním objektu) je k tomu zapotřebí globální centrální databáze; naproti tomu resoluci ISBN lze postavit na distribuovaném systému například národních bibliografií (databází bibliografických záznamů při národních knihovnách mapujících knižní produkci daného národa).

Vedle výše uvedených identifikátorů vycházejících primárně ze světa klasických dokumentů existuje několik systémů vytvořených již přímo pro zdroje na Internetu. Zmíníme stručně tři z nich: PURL, Handle a DOI.

4.3 PURL – persistentní URL

Tento systém [66] realizovaný organizací OCLC byl jedním z prvních pragmatických řešení vyvinutých pro knihovnicko-informační komunitu s cílem využít to, co již současný Internet nabízí (http a URL) a přitom co nejjednodušším způsobem odstranit základní problém s identifikací přes URL – závislost identifikace zdroje na jeho umístění. PURL je URL poskytující *nepřímou adresaci* zdroje. Princip je velmi jednoduchý: Informační zdroj na Internetu dostane přidělený identifikátor PURL například ve tvaru *http://purl.oclc.org/catalog/item1* a teprve na této „adrese" je uloženo skutečné URL zdroje. Funkčně je tedy PURL normálním URL, které však neodkazuje přímo na umístění zdroje, nýbrž na zprostředkující resoluční službu. Ta propojí identifikátor PURL se skutečným URL a vrátí ho klientovi. Klient pak dokončí URL transakci standardním způsobem (přes http příkaz redirect). Pokud se změní umístění zdroje, změní jeho správce hodnotu uloženou na adrese http://purl.oclc.org/catalog/item1, ale samotné PURL (externí jméno) se nikdy nemění. Uživatelé se mohou volně zaregistrovat na PURL serveru a poté si vytvářet vlastní identifikátory PURL, mohou dokonce volně stahovat příslušný software a instalovat vlastní resoluční PURL server.

4.4 Systém "handles"

Technologii známou pod názvem „handles" [36] vyvinula CNRI jako součást obecné architektury DL navržené v [44] pro jednoznačnou identifikaci digitálních objektů. Ačkoliv byl tento systém vyvinut nezávisle na konceptu URN, je s ním kompatibilní a lze ho považovat za vůbec první systém URN použitý v oblasti digitálních knihoven. Současná verze systému je založena na protokolu HTTP s identifikátorem vloženým do dokumentu ve formě hypertextové vazby odkazující na resoluční server systému handle. Identifikátor handle má následující tvar:

hdl:cnri.dlib/magazine

kde první část (prefix cnri.dlib) je tzv. pojmenovávací autorita, která je přidělována hierarchicky (nejvyšší úroveň *cnri* je přidělována centrální autoritou, zbytek již lokálně). Část za lomítkem je libovolný řetěz znaků jedinečný v rámci dané pojmenovávací autority. Architektura systému handle je dvojúrovňová – jeden globální registr a neomezený počet lokálních serverů; z důvodů lepší výkonnosti a lepší dostupnosti služeb je implementována jako distribuovaný systém s decentralizovanou administrací (globální registr identifikátorů tak není centralizován fyzicky nýbrž virtuálně). Každá z jeho komponent může být rozprostřena mezi více počítačů a data mohou být automaticky replikována, k dispozici je řada cachovacích služeb. Pro maximální využití a přímou resoluci identifikátorů (včetně násobné resoluce) je třeba doinstalovat do uživatelského www-prohlížeče příslušný software ve formě plug-in (jsou volně k dispozici); prostřednictvím proxy serverů lze systém používat i s neadaptovanými prohlížeči, avšak již ne s plnou funkčností.

Systém je velmi propracovaný; jeho slabá stránka však spočívá v tom, že se jej nepodařilo prosadit jako Internetovský standard (patrně hlavně proto, že IETF nechtělo připustit rozmnožování různých koncepcí resolučních služeb a podporuje pouze vlastní koncept v podobě URN). Nicméně řada velmi úspěšných současných systémů je na handles založena – jmenujme alespoň NDLP program Kongresové knihovny [51], NCSTRL – distribuovanou digitální knihovnu technických zpráv z oblasti computer science [57] a iniciativu amerických nakladatelů DOI.

4.5 DOI – Digital Object Identifier

V roce 1996 vznikla z popudu Asociace amerických nakladatelů iniciativa DOI [23], jejímž cílem bylo vytvořit systém pro identifikaci digitálních objektů (prací chráněných copyrightem) pro potřeby komerčních vydavatelů. Vznikl systém, který je od roku 1998 dále rozvíjen mezinárodní nadací International DOI Foundation (IDF). Jako resoluční mechanismus identifikátorů DOI je využíván systém handle popsaný výše. Syntaxe DOI byla specifikována normou ANSI/NISO Z39.84-2000. Příklad:

doi:10.1006/123456

Prefix 10.1006 sestává z konstanty 10 (slouží k odlišení DOI od ostatních implementací systému handle), za níž po tečce následuje numerický identifikační kód registrující organizace (1006 je například kód Academic Press). Sufix za lomítkem obsahuje identifikátor digitálního objektu a může jím být cokoliv za předpokladu, že v rámci dané registrující organizace bude jednoznačný. To dává registrující organizaci možnost použít volně libovolné identifikační systémy – jak globální, např.

doi:10.1000/ISBN1-900512-44-0, tak i lokální (to je zásadní rozdíl oproti koncepci URN, která použití každého identifikačního systému umožňuje pouze tehdy, pokud bylo stanoveným postupem standardizováno v Internetovské komunitě). Číslo DOI identifikuje dílo, nikoliv projev díla (viz IFLA model výše), takže tištěná verze článku a jeho digitální kopie mají totéž číslo. Systém DOI je silně centralizovaný; každá registrující organizace musí všechna jí vydaná DOI čísla registrovat u (zatím jediné) registrační agentury, resoluce probíhá přes tuto centrální databázi (http://dx.doi.org/10.1000/ISBN-1-900512-44-0). Zajímavým rysem je, že povinnou součástí registrace čísla DOI (kromě stavových dat specifikujících umístění objektu) je také předání DOI-metadat popisující daný objekt; ta pak mohou být vrácena jako výsledek procesu resoluce, když není možno zpřístupnit objekt samotný – například z licenčních důvodů.

V současné verzi poskytuje DOI pouze persistentní identifikátory (čili v zásadě totéž co mnohem jednodušší PURL), ale IDF má ambice rozvinout ho do kompletního systému na podporu řízení správy vlastnických a autorských práv. Všeobecně se zatím nepředpokládá, že by se DOI někdy vyvinul v obecně použitelné řešení pro identifikaci všech typů dokumentů na Internetu, a to jednak z důvodů technických (viz neúspěch při pokusu o standardizaci systému handle), ale zejména z důvodů ekonomických (platí se nemalé částky jak za registraci registrující organizace, tak i za každé zaregistrované číslo DOI). Koncem roku 2000 bylo v systému zaregistrováno 150 registrujících organizací – nakladatelů. V systému CrossRef [15], který využívá DOI pro vytváření citačních vazeb v oblasti vědeckých publikací (citation-linking), byla v té době aktivní zhruba polovina nakladatelů registrovaných v DOI a databáze systému obsahovala záznamy asi 3 miliónů článků. Od června 2001 jsou v DOI k dispozici nástroje umožňující realizovat vícenásobnou resoluci čísel DOI.

5 Metadata

Obecně jsou metadata *informace o informacích*; v kontextu digitálních knihoven je lze charakterizovat jako počítačově zpracovatelné strukturované informační objekty popisující vlastnosti jiných informačních objektů. Protože metadata jsou klíčovou komponentou pro obrovskou škálu velice různorodých služeb (vyhledávání informačních zdrojů a jejich výběr, autentizaci, interoperabilitu, správu vlastnických práv, dlouhodobou archivaci a řadu dalších), existuje velmi mnoho různých metadatových schémat.

5.1 Úvod a stručný přehled

Klasické knihovny jsou od počátku své existence postaveny na vytváření a využívání metadat (bibliografických záznamů) a totéž platí i pro knihovny digitální. Avšak mezi bibliografickými metadaty v klasických knihovnách a síťovými metadaty pro digitální síťové prostředí je jeden základní koncepční rozdíl: zatímco bibliografický záznam usiluje o kompletní popis zdroje, síťová metadata jsou *specializovaná* – pokrývají vždy jen určitou část, jen některé aspekty zdroje. Tento rozdíl je dán dvěma faktory: za prvé organizačním modelem používaným při tvorbě metadat (u klasických knihoven je to jedna centrální autorita, například národní knihovna, zatímco u DL jde o řadu různých komunit pracujících nezávisle na sobě a podle svých specifických

potřeb). Za druhé je zde jiný model přístupu k samotnému zdroji: protože v klasické knihovně neměli uživatelé metadat obvykle přímý přístup k informačnímu zdroji, museli být schopni učinit své rozhodnutí o užitečnosti zdroje výhradně na základě znalosti jeho metadat. V digitálním síťovém prostředí jsou naproti tomu často zdroje dostupné přímo, uživatel je může bezprostředně konzultovat, což eliminuje potřebu komplexního popisu. Stejně tak je možné snadno zpřístupňovat různá metadata daného zdroje a nové technologie na bázi RDF nabízí možnost je vzájemně propojovat, kombinovat a vytvářet tak složitější popisy dynamicky podle potřeby.

Metadata lze klasifikovat podle různých hledisek. Z hlediska jejich obecného použití se obvykle dělí na metadata *popisná* (slouží k obecnému popisu zdroje za účelem jeho vyhledání, identifikace a selekce), *strukturální* (zachycují formát a strukturu zdroje za účelem jeho správného uložení a zobrazování) a *administrativní* (slouží ke správě zdroje, včetně řízeného přístupu a archivace). Jiná typologie může rozčleňovat metadatová schémata podle bohatosti jejich struktury a míry detailnosti popisu: od jednoduchých často proprietárních schémat s jednoduchými nepříliš strukturovanými formáty (například automaticky generovaná metadata Internetovských vyhledávačů) až po velké propracované mezinárodní standardy typu MARC nebo značkovací systémy typu TEI [74]. Zatímco popisná metadata bývají často uložena v katalozích a indexech udržovaných vně repozitářů s digitálními objekty, strukturální a administrativní data bývají naopak často vložena přímo do digitálního objektu.

V tomto příspěvku se zmíníme o některých přístupech z oblasti popisných metadat. Zájemce o problematiku metadat pro účely dlouhodobé archivace odkazujeme na [22], příkladem z oblasti správy vlastnických a autorských práv je DOI [24], domovská stránka iniciativy INDECS – INteroperability of Data in E-Commerce Systems [38] zase poskytuje dobrý vstupní bod pro studium problematiky metadat v oblasti e-commerce. Ze zástupců metadat pro netextové dokumenty zmiňme alespoň standard MPEG-7, Multimedia Content Description Interface [56]. Podrobný přehled popisných metadat lze nalézt v [18] a [72], rozsáhlý seznam internetovských zdrojů na téma metadat s odkazy na různá metadatová schémata je udržován na [40] a analýzu výzkumných témat v oblasti metadat podává [28].

Hlavním účelem síťových popisných metadat je zlepšit přesnost vyhledávání a výběru digitálních informačních zdrojů oproti tomu, co dnes nabízí Internetovské vyhledávače (velký ‘ohlas’, malá ‘přesnost’). Prvním pokusem v tomto směru byl návrh RFC-1807 (Bibliographical Format for Technical Reports) využitý například v DIENST [19], protokolu a implementaci systému distribuovaných DL-serverů použitém v řadě DL-projektů. Nejznámějším a patrně nejperspektivnějším formátem v tomto směru je však standard Dublin Core.

5.2 DC – Dublin Core

Klasická bibliografická metadata vycházející například ze standardu MARC jsou příliš složitá a pravidla pro jejich použití (nejčastěji Angloamerická katalogizační pravidla AACR2) příliš komplikovaná na to, aby je byl schopen používat i někdo jiný než jen profesionální katalogizátoři. Naproti tomu metadata generovaná automaticky Internetovskými vyhledávači či ad-hoc doplňovaná do záhlaví html dokumentů jsou příliš primitivní na to, aby mohla podstatněji ovlivnit přesnost vyhledávání. Dublin

Core [25] je pokus o kompromis: vytvořit metadatový standard rozumně jednoduchý na to, aby ho mohli využívat i nezaškolení autoři publikací na Webu, na druhou stranu však dostatečně propracovaný, univerzální a flexibilní i pro netriviální aplikace v co nejširším spektru oborů a oblastí. Ambice DC jdou ještě dál: stát se všeobecně rozšířeným a používaným standardem, který může díky své jednoduchosti posloužit i jako základna pro sémantickou interoperabilitu mezi jinými složitějšími formáty. Podrobnější popis motivace a historie DC lze nalézt v [10].

Dublin Core lze používat dvěmi způsoby; prvním je tzv. *nekvalifikovaný Dublin Core* (tj. DC bez kvalifikátorů), kdy uživatel má k dispozici 15 základních metadatových prvků popisujících obsah (název, předmět, popis, pokrytí, typ, zdroj, vztah), intelektuální vlastnictví (tvůrce, přispěvatel, vydavatel, práva) a instanci síťového zdroje (identifikátor, datum, jazyk, formát). Každý prvek je nepovinný, opakovatelný a na pořadí prvků nezáleží. Implementátor konkrétní aplikace může dokonce přidávat své vlastní specifické prvky, ty však samozřejmě nebudou globálními aplikacemi využitelné. Druhý způsob představuje tzv. *kvalifikovaný Dublin Core*, kdy pro zpřesnění popisu zdroje lze jednotlivé prvky DC dopřesnit pomocí dvou typů kvalifikátorů: kvalifikátor prvku (zužuje sémantiku prvku – např. Autor.Ilustrátor) a kvalifikátor hodnoty (specifikuje způsob interpretace zadané hodnoty prvku – například Datum = "1999-04-12" : ISO8601). Existuje seznam standardizovaných kvalifikátorů pro jednotlivé prvky DC, implementátoři mohou opět doplňovat vlastní kvalifikátory. Všechny kvalifikátory musí však splňovat tzv. princip "dumb-down", což znamená, že kvalifikovaný DC-záznam musí být korektně zpracovatelný i aplikací navrženou pro nekvalifikovaný DC (aplikace jednoduše ignoruje kvalifikátory a hodnoty všech prvků musí i po této restrikci odpovídat sémantice základních prvků nekvalifikovaného DC).

Uveďme příklad kvalifikovaného záznamu Dublin Core:

IDENTIFIER=http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf : URL
TITLE = Guidance on expressing the Dublin Core within the RDF
TITLE = Dublin Core in RDF: Eine Anleitung
CREATOR = Eric Miller
CREATOR = Paul Miller
CREATOR.Illustrator = Dan Brickley
DESCRIPTION.Abstract = This work describes work carried out by …
SUBJECT.Keywords = Dublin Core; DC; Resource Description Framework; RDF; XML
PUBLISHER = Dublin Core Metadata Initiative
CONTRIBUTOR = Dublin Core Data Model Working Group
DATE.Created = 1999-07-01 : ISO8601
DATE.Revised = 1999-11-10 : ISO8601
LANGUAGE = en : RFC1766
TYPE = Working Draft
FORMAT.Medium = text/html : IMT
MYELEMENT.Checksum = 123456 : XYZ

Tvůrci Dublin Core (mezinárodní komunita informačních specialistů, knihovníků a vydavatelů koordinovaná organizací OCLC) se soustředili pouze na precizní specifikaci sémantiky a syntaxi záměrně ponechávali volnou (nepředjímalo se, kde a jak budou metadata Dublin Core ukládána). Časem se však ukázalo, že je vhodné vypracovat základní návody na způsoby zápisu alespoň v základních formátech,

jakými jsou HTML, XML a RDF. Dublin Core se stal Internetovským doporučením již ve druhé polovině 90.let (RFC-2413) a koncem minulého století byl již v pozici de-facto standardu přeloženého do desítek jazyků (česká verze byla vytvořena a je spravována na Masarykově univerzitě v Brně [26]). V červnu 2001 byl organizací NISO schválen jako návrh standardu Z39.85 a postoupen instituci ANSI k dokončení americké standardizační procedury.

5.3 Metadata Kongresové knihovny

Tento „firemní-standard“ metadat [50] byl vypracovaný pro potřeby DL-projektů zastřešených velmi ambiciózním programem *National Digital Library program* [51] vedeným Kongresovou knihovnou. Na rozdíl od Dublin Core se tedy nejedná o univerzální standard, nýbrž o příklad schématu navrženého (byť dostatečně obecně) pro potřeby jednoho konkrétního praktického programu. Toto metadatové schéma zahrnuje současně metadata strukturální, administrativní a částečně i popisná (plná popisná data jsou uložena v katalozích), a to pro podporu všech funkcí digitální knihovny: řízení přístupu, vyhledávání, prezentaci, administraci, persistentní identifikaci a dlouhodobé uchovávání digitálních objektů. Metadatové záznamy mají pět úrovní odpovídajících 5-úrovňové hierarchické struktuře objektů. Nejvyšší úroveň, tzv. *set*, je jedna digitální sbírka. Sbírka je tvořena jedním či více *agregáty*, tj. skupinami digitálních objektů stejného typu (text, video) či stejné správy. Agregát sdružuje *primární objekty*, koherentní jednotky odpovídající fyzické jednotce (knize, nahrávce). Primární objekt může sestávat z několika komponent respektive pohledů – *meziobjektů* (například kniha může mít dva meziobjekty: první z nich je obrazová prezentace knihy, sestávající z obrázků vzniklých naskenováním jejích stran; druhý meziobjekt zastupuje textovou komponentu knihy, soubor textů vzniklých např. převodem z obrázků pomocí OCR a sloužící pro fulltextové vyhledávání). Konečně poslední úroveň tvoří *terminální objekty*, což jsou jednotlivé soubory s digitální informací.

Metadatové schéma má celkem 77 prvků, ne všechny musí být v metadatovém záznamu vyplněny; některé jsou specifické pro konkrétní typ média (jiné pro digitalizovanou zvukovou nahrávku, jiné pro digitalizovaný obrázek), další zase pro úroveň objektu (například metadata pro terminální objekty jsou především strukturální).

5.4 Standardy vycházející ze struktury MARC

Předchozí dva příklady popisovaly metadatová schémata vytvořená přímo pro digitální/síťové zdroje. Aby byl náš obrázek úplnější, musíme zmínit i opačný pól: národní a mezinárodní bibliografické standardy založené na struktuře MARC (Machine-Readable Cataloguing). Tyto byly vytvořeny původně pro popis fyzických informačních objektů v knihovnách a až posléze byly upravovány i pro potřeby digitálního světa. Na základě společné koncepce vyvíjené od poloviny 60.let v Kongresové knihovně vznikla postupně celá rodina více či méně kompatibilních standardů zohledňující národní specifika – např. USMARC (USA), UKMARC (UK), CANMARC (Kanada), a v roce 1977 pod záštitou IFLA i mezinárodní standard UNIMARC, který pak mnohé země – „samozřejmě“ kromě těch výše uvedených – převzaly za svůj národní formát (což je případ i českého standardu UNIMARC-CZ).

Pro tyto standardy je charakteristická velmi bohatá struktura polí a podpolí, velice jemná granularita popisu a detailní propracovanost návazných pravidel pro tvorbu záznamů. Pro představu uveďme příklad zjednodušeného UNIMARCovského záznamu v tzv. řádkové struktuře :

```
001    CASLIN0000001
005    19960312
010    $a80-7050-237-1
100    $a19960305d1996####k##y0czey0103####ba
101 0# $acze
102    $aCZ
200 1# $aZáznam pro souborný katalog$eUNIMARC$iTištěné
        monografie$fPracovní skupina CASLIN pro standardizaci
        a jmenné zpracování ... [et al.]
205    $a1. vyd.
210    $aPraha$cNárodní knihovna České republiky$d1996
215    $a 31 s.
225 1# $aStandardizace$vč. 4
675    $a025.3$9v
711 02 $aCASLIN$bPracovní skupina pro standardizaci a ...
801 #0 $aCZ$bABA001$c19960312$gAACR2$91
801 #3 $aCZ$bABA001$c19960515
910    $aABA001
```

Aniž bychom zabíhali do podrobností, alespoň základní vysvětlení: řádek začíná vždy třípísmenným názvem pole, za kterým mohou nebo nemusí následovat až dva poziční indikátory upřesňující obecné charakteristiky pole. Hodnota pole je strukturována do podpolí označených znakem „$“ následovaným jednoznakovým názvem podpole (má-li pole jen základní dále nestrukturovanou hodnotu, je uložena v podpoli $a). Například pole 200 obsahuje názvové údaje, které jsou v našem příkladě tvořeny hlavním názvem `Záznam pro souborný katalog` v podpoli $a, další názvová informace `UNIMARC` v podpoli $e, název části `Tištěné monografie` v podpoli $i, a údaji o odpovědnosti v podpoli $f. Standard definuje pro každé pole seznam všech jeho možných podpolí, jejich charakteristiky, vzájemné vazby a závislosti.

MARCovské formáty umožňují popsat (fyzický) informační zdroj do těch nejjemnějších nuancí způsobem vhodným pro počítačové zpracování. Na druhé straně jsou však příliš složité pro použití mimo oblast knihoven vybavených specializovanými informačními pracovníky.

5.5 XML a RDF

Nadefinovat strukturu a obsah metadat je jedna věc, další otázkou je jak metadata zapsat a kam je uložit. Protože v současnosti realizované digitální knihovny jsou pevně svázány s Webem, je přirozené využít k tomu standardních webových technologií. Jednoduchá síťová metadata jsou často vkládána přímo do digitálního informačního zdroje, například pomocí META-značky do záhlaví HTML dokumentů. Jedním z klíčových důvodů tak úspěšného prosazení Webu byla i jednoduchost jazyka HTML. V současnosti je však tato jednoduchost i jeho největší slabinou; HTML je orientován na prezentaci dokumentu a nikoliv na zachycení jeho struktury.

Značkovací jazyk XML [81], jako nástupce HTML, je pokusem o kompromis mezi jednoduchostí HTML a sílou SGML (který je naopak pro široké použití příliš flexibilní a komplikovaný) a dokáže dobře zapsat potřebné struktury. Navíc byl navržen s ohledem jednak na snadné vytváření programů pro manipulaci s XML dokumenty, jednak na relativně „bezbolestný" přechod z HTML. V současnosti je XML nejčastější forma zápisu síťových metadat všeho typu a proniká i do takových „bašt" jako je standard MARC, viz projekt XMLMARC [82].

Každé metadatové schéma má tři aspekty – *sémantiku, syntaxi a strukturu*. Sémantika definuje interpretaci jednotlivých prvků, jejich význam (CO chceme říci o informačním zdroji). Dublin Core je o sémantice – přesně specifikuje význam každého z jeho 15 prvků. Syntaxe je naopak o tom, jak metadata formálním a přesným způsobem zapsat (čili JAK svá tvrzení o informačním zdroji vyjádříme). Nástrojem pro zápis Dublin Core je například jazyk XML. Struktura definuje vztahy mezi metadatovými prvky, v ideálním případě i mezi prvky různých metadatových schémat. Nástrojem pro vyjádření a zápis struktury metadat je *RDF – Resource Description Framework* [67], který jako svůj vyjadřovací jazyk používá XML. Strukturální model RDF sestává ze zdrojů, atributů a hodnot a představuje vlastně způsob zápisu orientovaných grafů představujících vztahy mezi zdroji. Uvažujme například jednoduchou větu „Shakespeare je autorem hry Hamlet".

V Dublin Core je zdrojem popisovaný dokument a atributem (vlastností) příslušný prvek DC; zachytit v něm lze pouze jednoduchá tvrzení typu „dokument – prvek – hodnota", takže výše uvedenou větu zaznamenáme jako:

zdroj		*atribut*		*hodnota*
Hamlet	->	CREATOR	->	Shakespeare
Hamlet	->	TYPE	->	hra

Jiná metadatová schémata mohou také obsahovat prvek pro autora, ale třeba jinak pojmenovaný (nikoliv CREATOR, ale např. AUTHOR) a naopak prvek pojmenovaný TYPE mohou používat s úplně jinou sémantikou. Proto musí RDF explicitně vyjádřit, že prvky ‚creator' a ‚type' mají v tomto případě takový význam, jaký jim dává standard Dublin Core – k tomu slouží mechanismus jmenného prostoru xmlns, viz příklad níže. Předpokládejme, že Hamlet je reprezentován webovým zdrojem http://hamlet.org; pak naši větu můžeme zapsat v RDF následovně:

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:DC="http://purl.org/dc/elements/1.1/">
    <rdf:description  rdf:about = "http://hamlet.org/">
      <DC.creator>Shakespeare</DC.creator>
      <DC.type>hra</DC.type>
    </rdf:description>
</rdf:RDF>
```

Ze zápisu je zřejmé, že atributy 'description' a 'about' jsou definovány ve schématu RDF, zatímco 'creator' a 'type' ve schématu Dublin Core. Pokud bychom chtěli doplnit popis Hamleta o atribut z nějakého jiného metadatového schématu než Dublin Core, stačilo by doplnit do záznamu jmenný prostor s odkazem na definici tohoto

schématu <xmlns:NM="http://www.abc.cz/new-metadata/"> a přidat do popisu nový atribut s příslušným prefixem <NM.novy-atribut>xyz</NM.novy-atribut>.

Strukturální model RDF umožňuje zachytit i mnohem složitější vztahy, například umožňuje, aby atribut zdroje odkazoval na jiný zdroj. Předpokládejme, že v nějaké databázi existuje záznam o Shakespearovi obsahující jeho metadata, kde a kdy žil, co napsal, apod. V našem příkladě by atribut DC.creator odkázal na tento záznam následovně:

<DC.creator rdf:about = "http://people.net/WS/">Shakespeare</DC.creator>

Popsaným způsobem lze v RDF propojovat metadata s digitálními objekty, vyměňovat metadata z různých metadatových schémat a skládat z jednoduchých komponent libovolně složité metadatové popisy. Jednou z oblastí, v níž je velmi často potřebné zachytit i hodně komplikované vztahy, je oblast vlastnických a autorských práv; uveďme typický příklad:

„Kyoto Records declares that it has acquired from the Japanese Copyright Society acting as agent for the Harry Fox Agency acting as agent for Diva Songs. Inc, the mechanical reproduction right in respect of the song „Rising Tide" by Joan Quincy for inclusion in the CD „Californian Sunsets" for distribution only in Japan."

6 Interoperabilita

Jak již bylo řečeno v úvodu, obecná představa vychází z toho, že digitální knihovna není nějaký monolitický produkt, ale naopak systém dynamicky propojovaných spolupracujících komponent, které samy o sobě mohou být autonomní a nezávisle spravované. „*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*" říká se v [54]. Termínem *interoperabilita* bývá označována schopnost spolupráce mezi technicky různorodými a organizačně nezávislými komponentami při řešení určitého úkolu. Někdy se s mírnou nadsázkou tvrdí, že všechny technické problémy/výzvy digitálních knihoven nejsou nic jiného než jen různé aspekty interoperability.

6.1 Úvod a stručný přehled

Existuje velice široké spektrum pohledů na interoperabilitu: na jednom konci lze pohlížet na interoperabilitu jen jako na použití společných nástrojů a rozhraní pro vytvoření *povrchní jednoty* pro přístup a navigaci, na opačném konci je pak vysoce ambiciózní *hluboká sémantická interoperabilita*, kdy inteligentní technologie dokáží poskytnout koherentní pohled na různorodý informační obsah a služby digitálních knihoven (zatím jde o hudbu budoucnosti). Někde mezi těmito dvěmi extrémy je primárně *syntaktická interoperabilita*, kdy výměna metadat a použití protokolů pro přenos digitálních objektů a formátů založených na těchto metadatech umožňují poskytnout omezenou koherenci obsahu, která pak musí být ještě doplněna lidskou interpretací.

Při zkoumání interoperability ukazuje [4] závislost mezi funkcionalitou a cenou. Většina v současnosti používaných metod pro interoperabilitu (například webové standardy HTTP, HTML, URL) dosahují jen průměrné funkcionality, ale zase za nízkou cenu a s velmi širokým uplatněním (příklad webovských vyhledávačů).

Naopak většina high-end služeb (založených například na využití standardů Z39.50 či SGML) dosahuje vysoké funkcionality, ale za vysokou cenu, která často brání jejich širšímu využití. Většina výzkumu v oblasti DL je pak vedena snahou najít ten správný "zlatý střed".

Systematický pohled na interoperabilitu a přístupy k jejímu dosažení shrnuje [63]. Uvádí, že problém interoperability se bezprostředně dotýká všech pěti základních funkcí digitálních knihoven – správy informací (ukládání, organizace a získávání informace), prezentace informací uživatelům, komunikace mezi částmi systému, řízení systému, a ochrany informačních zdrojů a uživatelů včetně jejich práv. Ačkoliv porovnávání úspěšnosti jednotlivých řešení je v oblasti interoperability velmi obtížné (různé přístupy vychází z různých předpokladů a mají různé často protikladné cíle), navrhuje šest základních kritérií, které přece jen poskytují určité vodítko:

- vysoký stupeň autonomie komponent
- nízká cena infrastruktury
- snadnost přidání nové komponenty
- snadnost používání komponenty
- šíře celkové složitosti
- škálovatelnost v počtu komponent.

V některých případech se může stát, že rozhodnutí optimalizující jedno z těchto kritérií mohou negativně ovlivnit jiné (například systém, který minimalizuje cenu infrastruktury, může být použitelný jen pro jednoduché úkoly, nebo ho může být obtížné používat vůbec).

Existuje řada velmi rozdílných přístupu k dosažení požadovaného stupně interoperability; práce [63] popisuje pět základních tříd přístupů:

1. *silné standardy* : nejstarší přístup založený na tom, že heterogenní komponenty se shodnou na standardu, který zajistí určitou omezenou míru homogenity mezi nimi. Příkladem jsou standardy Z39.50, HTML/HTTP.

2. *rodiny standardů* : v tomto případě má implementátor komponenty k dispozici ne jeden standard, ale celou rodinu standardů, z nichž může volně vybírat a dosáhnout tak vyššího stupně autonomie než v předchozím případě. Příkladem je e-commerce, kdy při implementaci platebního modulu může uživatel volit mezi řadou standardizovaných platebních schémat (DigiCash, First Virtual, některá z mnoha platebních karet).

3. *vnější zprostředkování* : je cesta k dosažení velmi vysokého stupně autonomity komponent. Mechanismus pro zajištění interoperability je umístěn mimo spolupracující komponenty v podobě samostatných zprostředkujících modulů nazývaných „wrappers" nebo „proxies", které překládají datové formáty a komunikační protokoly komponent do/z interního standardu systému (provádí mapování mezi globálními a lokálními schématy na úrovni komponent). Příkladem z oblasti propojování sítí jsou gateways. Nevýhodou tohoto řešení může být vyšší cena přidání nové komponenty zahrnující i vytvoření příslušné zprostředkující mezikomponenty.

4. *interakce založená na specifikacích* : cílem je umožnit používat komponenty bez pomoci speciálních předběžných opatření a bez prostředníků. Pro každou

komponentu existuje přesný formální popis sémantiky a struktury všech jejích dat a operací; komponenty pak mohou mezi sebou interagovat díky tomu, že jsou schopny zjistit si specifikace spolupracovníků a zařídit se podle nich. Příkladem nástrojů pro implementaci tohoto typu přístupů je knowledge-sharing nástroj pro softwarové agenty – jazyk Agent Communication Language a jeho Knowledge Interchange Format (KIF) nebo jazyky SETL a PAISLey pro software-reuse umožňující popsat sémantiku funkcionality dané komponenty čistě deklarativním způsobem. Tento přístup přináší vysokou míru autonomity, avšak současně vysokou náročnost přidání nové komponenty (popsat dostatečně podrobně komponentu může být velmi složité a někdy v praxi i nemožné).

5. *mobilní funkcionalita* : je založena na mobilních softwarových agentech, kteří cestují sítí na místa, kde zpřístupní potřebné služby. Z novějších technologií umožňují například Javovské applety přístupy pro doručení nových funkcionalit klientským komponentám až v době běhu (takovou novou funkcionalitou může být třeba schopnost komunikovat s jinou komponentou systému). Tento přístup je velmi lákavý a efektivní zejména z pohledu snadnosti přidání nové komponenty. Na druhou stranu jeho implementace může být nákladná zejména z hlediska komunikačního (pokud na straně klienta neexistuje dlouhodobá vyrovnávací paměť typu *cache*, může ten samý programový kód cestovat po síti opakovaně stále dokola), ale i z hlediska bezpečnostního (kontroly autenticity a bezpečnosti kódu na každém přijímacím místě sítě). Tento přístup je také silně závislý na existenci silného standardu (v daném případě širokém rozšíření prohlížečů podporujících příslušný Javovský standard).

Jiné obecné pohledy na interoperabilitu přináší [29]. Rozlišuje několik abstraktních úrovní interoperability, od obecné transportní vrstvy a na aplikační oblasti nezávislého middleware (Z39.50, distribuované objekty např. s technologií CORBA) až po úrovně specifické pro digitální knihovny – vrstva informačního modelu, správy informací/dokumentů, správy vlastnických a autorských práv, a nejvyšší vrstva týkající se sociálních souvislostí. Pokorný v [65] zmiňuje 4 úrovně praktické kooperace digitálních knihoven:

úroveň	**způsob kooperace**
federace	striktní použití standardů (syntaktické, sémantické, obchodní) příklad: MARC, Z39.50
sklízení metadat	DL nabízejí základní metadata; jednoduchý protokol a registrace příklad: otevřené archívy, např. iniciativa OAI
shromažďování dat	DL nekooperují; informace se musí hledat explicitně pomocí služeb příklad: Internetovské vyhledávače
vyhledávací middleware	zdroje vybavené metadaty jsou volně zapojovány do kooperace příklad: Z39.50, XML, RDF, SDLIP [62].

Z obrovského množství nástrojů, přístupů a projektů z oblasti interoperability uveďme jen několik málo vybraných zástupců z těch všeobecně nejznámějších, které charakterizují typické nebo nové perspektivní přístupy: vyhledávací protokol Z39.50, přístup přes sklízení metadat v Open Archives Initiative (OAI), Stanfordský projekt InfoBUS, a technologii OpenURL a SFX pro otevřené kontextově-citlivé propojování zdrojů (reference linking).

6.2 Protokol Z39.50

Z39.50 je mezinárodním standardem pro komunikaci mezi počítači, který umožňuje jednomu počítači (klient, origin) vyhledávat a získávat informaci na jiném počítači (databázový server, target) a to v heterogenním prostředí, nezávisle na operačních systémech, databázích a dotazovacích jazycích. I když koncepčně není vázán na žádný konkrétní druh informací ani typ databází, největší jeho současnou aplikační oblastí jsou bibliografická data a knihovní katalogy. Základ standardu vznikl v roce 1984 jako výsledek projektu Linked Systems Project předních amerických knihoven a od té doby prošel třemi ANSI/NISO verzemi: 1988 (v1), 1992 (v2) a 1995 (v3). Verze 1 není s ostatními kompatibilní; verze 3 je nadmnožinou verze 2 a byla přijata jako mezinárodní standard ISO 23950. Podrobný přehled historie a motivací ve vývoji Z39.50 lze nalézt v [53].

Z39.50 je založen na abstrakci databázového vyhledávání, která je obecnější než například u SQL. Server provozuje jednu či více databází obsahujících záznamy; s každou databází je spojena množina přístupových bodů (indexů), které mohou být použity pro vyhledávání. Protokol je stavový (na rozdíl od bezstavového HTTP) a relačně orientovaný, interakce mezi klientem a serverem je založena na koncepci seance (session): klient otevře spojení se serverem, provede sekvenci interakcí a uzavře spojení. Během sezení si klient i server pamatují stav jejich interakce. Zdůrazněme, že Z39.50 je protokol mezi dvěmi počítači, nijak nespecifikuje uživatelské rozhraní, kterým bude ke klientskému počítači přistupovat uživatel.

Typická seance začíná tím, že klient naváže spojení se serverem a vyvolá inicializační službu *init*, během které si obě strany vyjednají podrobnější detaily spolupráce (kterou verzi protokolu podporují, používanou množinu znaků a jazyk, maximální délku záznamu předávaného ze serveru, požadavek na autentikaci uživatele, apod.). Poté může klient pomocí služby *explain* zjistit detaily o serveru a jím nabízených službách: databáze dostupné pro prohledávání a jejich přístupové body (indexy), podporovaná syntaktická schémata a datové formáty, třídicí možnosti, ale také obecné charakteristiky jako popis serveru, provozní doba, případná omezení a cena za použití. Po těchto úvodních operacích může klient vyslat vyhledávací dotaz pomocí služby *search*; standard specifikuje 6 typů vyhledávání od booleovského, přes standard ISO 8777 Commands for Interactive Text Searching, ANSI standard Common Command Language (CCL) až po SQL – běžně však bývá plně implementováno jen booleovské vyhledávání. Dotaz tedy může mít následující význam:

Najdi v databázi 'Knihy' všechny záznamy, pro které přístupový bod 'title' obsahuje hodnotu 'sen' a přístupový bod 'author' obsahuje hodnotu 'shakespeare'.

Server provede hledání, vytvoří výsledkovou množinu, tzv. *result set* a uloží si ji, takže klient se na ni může následně v dalších příkazech odvolávat – zmenšit velkou výsledkovou množinu upřesňujícím hledáním, setřídit ji, vymazat, apod. V závislosti na parametrech příkazu hledání vrátí server klientovi buď jen počet vyhledaných záznamů, nebo přímo jeden či více záznamů z výsledkové množiny. Jakmile je hledání dokončeno, vyšle klient službu *present*, v níž serveru specifikuje, které záznamy z výsledkové množiny a v jakém formátu mu mají být zaslány (standardně se používá buď textový formát nebo formát MARC, ale možné jsou i jiné varianty).

Kromě dosud popsaných služeb nabízí protokol ještě řadu dalších – pro procházení indexů, řízení přístupu (možnost serveru vyslat žádost o autentizaci

uživatele, informovat o postupu dlouhého vyhledávání), možnost účtování, ukončení seance, a také tzv. *rozšířené služby*, což je v zásadě mechanismus pro asynchronní vzdálené volání procedur, pomocí nichž lze realizovat např. další operace nad výsledkovou množinou – jako její uchovávání mezi seancemi, zařazení do fronty pro zaslání Emailem nebo tisk, pro zaznamenání dotazů, které mohou být na serveru prováděny opakovaně v určenou dobu (SDI), a další. Ve verzi Z39.50-1995 je možné provádět ze strany klienta také aktualizaci záznamů v databázi na serveru.

Protokol Z39.50 je na jednu stranu velmi mocný a flexibilní, na druhou stranu hodně rozsáhlý (jeho úplná specifikace má kolem 160 stran) a náročný na implementaci i správné nastavení pro bezchybnou funkci v dané doméně (potřeba společného *profilu* specifikujícího kterých vlastností a nastavení protokolu bude při komunikaci využíváno). Jak již bylo řečeno, jeho hlavní oblastí nasazení jsou bibliografické knihovní databáze, ale existují i profily pro využití standardu v oblasti vládních informačních systémů, vědecko-technických databází, geografických informační systémů, v muzeích a DL-sbírkách.

Má-li knihovnický systém zabudován Z-klienta, lze použít protokol Z39.50 jako meziplatformní standard pro interoperabilitu při vyhledávání následujícím způsobem: uživatel zformuluje dotaz v jazyce svého knihovního systému a vybere pro vyhledávání cizí vzdálený katalog se Z-serverem. Dotaz je přeformulován do Z39.50 a zaslán Z-serveru cizího katalogu; ten přeloží dotaz do vyhledávacího jazyka cílové databáze a příjme výsledek vyhledávání. Výsledek pošle Z-klientovi, který ho předá knihovnickému systému pro zobrazení v jeho standardním uživatelském rozhraní. Z-klient může být implementován také tak, aby vyhledávací dotaz rozeslal paralelně více specifikovaným Z-serverům, což například umožňuje realizovat virtuální souborné katalogy.

Existuje několik volně dostupných samostatných Z-klientů, které si lze instalovat a využívat pro prohledávání informačních zdrojů podporujících protokol Z39.50, například BookWhere, Z-navigator a další. Další rozvoj protokolu Z39.50 řídí mezinárodní skupina Z39.50 Implementors Group (ZIG) pod patronací Kongresové knihovny, která zodpovídá za Z39.50 v roli Agentury pro jeho údržbu a rozvoj [52]. V létě 2001 má být dokončena a předložena k hlasování pracovní verze návrhu nové verze standardu Z39.50-2001, paralelně probíhá iniciativa Z39.50 Next Generation (ZNG), jejímž cílem je přiblížit protokol směrem k webovým přístupům a technologiím a snížit náročnost jeho implementace.

6.3 Open Archives Initiative (OAI)

Za vznikem Open Archives Initiative [60] koncem roku 1999 je rostoucí nespokojenost vědců s tradičním modelem vědeckého publikování (dlouhá doba od nabídnutí příspěvku k jeho zveřejnění a stále rostoucí cena předplatného časopisů) spolu s pozitivními zkušenostmi s novými modely publikování v podobě on-line repozitářů typu e-print (viz e-Print archive [8], NCSTRL [57]). OAI je zaměřena na podporu rozvoje tohoto typu publikování tím, že nabízí technický mechanismus a organizační struktury pro podporu interoperability mezi *otevřenými archívy* (pojem „otevřený“ je zde ve smyslu architektury systému, nikoliv nutně ve smyslu bezplatného či neomezeného přístupu; pojem „archív“ je chápán volně jako jakýkoliv repozitář pro ukládání informací na webu). Jako metoda pro dosažení potřebné low-

barrier interoperability bylo zvoleno tzv. *sklízení metadat*, kdy *poskytovatelé dat* (archívy) mají k dispozici relativně snadno implementovatelný mechanismus pro externí zviditelnění informací (metadat) o obsahu archívu, což umožňuje třetí straně – *poskytovatelům služeb* – tyto informace z mnoha archívů automatizovaným způsobem shromažďovat a budovat nad nimi různé nadstavbové služby. Technický aspekt tohoto řešení zahrnuje tři komponenty:

- *společný metadatový standard* – Open Archives Metadata Set (OAMS) : povinnou součástí metadat je nekvalifikovaný Dublin Core, různé odborné komunity mohou volitelně doplnit další metadata v jejich specifickém schématu. Metadata jsou zabalena do XML-záznamu, který obsahuje záhlaví (jednoznačný identifikátor, datum vytvoření či změny záznamu), metadata, a popis metadat. Záznamy jsou uloženy u poskytovatele dat v repozitáři, který musí podporovat OAI sklízecí protokol, a mohou obsahovat odkaz na vlastní dokument, který může nebo nemusí být volně dostupný.
- *jednotné identifikační schéma* : musí být jednoznačné a má následující tvar oai:arXiv:hep-th01 . První část tvoří konstantní řetězec „oai“, za ním je jednoznačný identifikátor repozitáře (archív ho obdrží při registraci u OAI), poslední částí je libovolný identifikátor jednoznačný uvnitř daného repozitáře. Resoluce identifikátorů bude probíhat přes centrální OAI resoluční službu s podporou OpenURL (viz níže)
- *protokol pro sklízení metadat* : původní návrh počítal s využitím protokolu Dienst, ale pro zjednodušení implementace byl nakonec vytvořen samostatný OAI protokol na bázi HTTP obsahující šest jednoduchých příkazů.

V současné době se OAI nachází v etapě rozšiřování a experimentování s technickou infrastrukturou a funkcionalitou celého systému (plánováno asi do poloviny roku 2002). Potřebné organizační podporu, stabilitu a zdroje dodávají iniciativě DLF (Digital Library Federation) a CNI (Coalition for Networked Information).

6.4 Stanfordský InfoBus

Jedním z nejobsáhlejších prakticky realizovaných řešení interoperability byl projekt *The Stanford Integrated Digital Library Project* realizovaný na Stanfordské univerzitě v 2.polovině devadesátých let v rámci amerického programu DLI-1. Projekt byl zaměřen na vývoj technologií pro integraci širokého spektra existujících i budoucích heterogenních sbírek a informačních zdrojů do virtuální digitální knihovny s jednotným přístupem ke všem jejím komponentám. Výsledky výzkumu byly realizovány v systému *InfoBus* [73] (název vychází z analogie s hardwarovou sběrnicí propojující různé hardwarové komponenty do jednoho funkčního celku) využívajícího technologii distribuovaných objektů na bázi systému CORBA (Common Object Request Broker Architecture).[1]

[1] Jedním ze základních komponent architektury CORBA je Object Request Broker (ORB), který po přidání k aplikačnímu programu realizuje vztah klient-server mezi distribuovanými objekty. Pomocí ORB může klient transparentně volat požadovanou operaci (metodu) nějakého serverovského objektu, který se může nacházet na stejném počítači nebo kdekoliv v síti. ORB najde objekt, který může realizovat požadovanou operaci, předá mu parametry, vyvolá jeho příslušnou metodu a vrátí výsledek. Klient samotný nemá žádné povědomí o

Namísto pokusu adaptovat existující informační systémy je InfoBus ponechává tak jak jsou. Pro každý z nich je zkonstruován zprostředkující 'wrapper', což je CORBA objekt reprezentující příslušnou online službu. Tyto zprostředkující objekty (proxies) komunikují s existujícími systémy v jejich „mateřském" komunikačním jazyku a transformují zprávy do/z interního standardního rozhraní, kterým je protokol DLIOP (DL InterOperability Protocol) podporující distribuované objekty. Například nějaký klient s vyhledávacím rozhraním Z39.50 chce vyhledávat v nějaké online informační službě, jakou je například Dialog. K tomu je zapotřebí dvou zprostředkujících objektů, jeden pro překlad mezi Z39.50 a DLIOP, druhý pro překlad mezi Dialogem a DLIOP. Ve Stanfordu vyvinuli řadu takových zprostředkujících objektů umožňujících prostřednictvím InfoBusu komunikovat libovolným Z39.50 klientem s velkou škálou informačních služeb, které Z39.50 nepodporují (souběžně byla na Universitě v Michiganu implementována proxy, která zase zprostředkovává přes InfoBus Z39.50 zdroje). Dále byly vyvinuty proxy pro HTTP, webovské vyhledávače a řadu dalších služeb.

Architektura InfoBusu obsahuje řadu dalších komponent potřebných pro realizaci komplexního systému: SMA – standardní metadatová architektura pro unifikovaný popis informačních služeb a jejich zdrojů pro podporu vyhledávání, STARTS (STAnford protocol proposal for internet ReTrieval and Search) – vrstva sloužící k organizaci vyhledávání na bázi metasearching, včetně výběru zdroje, vyhodnocení dotazů a slučování výsledků hledání, UPAI (Universal Payment Application Interface) – řešící mechanismus placení za poskytnuté služby, FIRM (Framework for Interoperable Rights Management) – řada propracovaných technik a přístupu pro řízený přístup ke zdrojům s ohledem na zajištění dodržování konkrétních podmínek vlastnických práv.

6.5 OpenURL a SFX

Problematika otevřeného kontextově-citlivého propojování zdrojů (open and context-sensitive linking) patří v posledních dvou letech k jedné z nejživějších oblastí DL. Představme si následující situaci: dnešní typická digitální knihovna nějaké instituce se skládá z řady heterogenních informačních zdrojů, ať již vlastních (knihovní katalog, digitalizované sbírky) nebo cizích (licencované fulltextové časopisy v elektronické podobě, informační databáze, abstrakční a citační služby, volně přístupné zdroje na Internetu), které jsou dostupné buď externě v repozitářích příslušných producentů či zprostředkovatelů nebo lokálně v podobě zrcadlených zdrojů či dle místních potřeb upravených systémů. Provozovatel a uživatelé takové digitální knihovny mají zájem na tom, aby informace z jednotlivých zdrojů byly co nejvíce integrovány, například *provázány* pomocí klikatelných hypertextových vazeb jdoucích *napříč těmito zdroji*: z citace v komerční citační databázi na záznam publikace v lokálním katalogu, ze záznamu v katalogu nebo z citace v seznamu referencí nějakého článku na plný text článku v elektronickém časopise příslušného nakladatele, ze slov v názvu článku nebo jeho předmětového hesla na relevantní informace v příslušném Internetovském vyhledávači, apod. Navíc by tyto vazby měly být „inteligentní" v tom smyslu, aby

---

tom, kde se tento objekt nalézá, jak a v jakém programovacím jazyku byl implementován, ani pod kterým operačním systémem je spouštěn. ORB tak umožňuje realizovat interoperabilitu mezi aplikacemi na různých počítačích v heterogenním distribuovaném prostředí.

zohledňovaly konkrétního uživatele a odkázaly ho vždy na zdroj odpovídající jeho statutu (např. na plný text licencovaného článku v případě zaměstnance instituce, na volně dostupný abstrakt pokud je uživatelem cizí osoba). Standardní „linkovací" řešení nabízená v posledních letech komerčními producenty informačních zdrojů jsou omezená (mají dosah jen v rámci informačního prostoru daného producenta), kontextově necitlivá (odkazují vždy na stejný cíl bez ohledu na to, který uživatel a s jakými právy je používá) a uzavřená (nedovolují třetí straně – například knihovně – nastavovat tyto vazby podle svých vlastních potřeb).

Řešení, které umožňuje překonat omezení dosavadních přístupů a realizovat představy z úvodu tohoto odstavce, nabízí nově se rodící standard OpenURL a nad ním postavený aplikační rámec SFX (od Special Effects) vycházející z výsledků výzkumu konce 90.let na universitě v belgickém Ghentu [78], [79]. Podstatou řešení je, že na rozdíl od klasických vazebních referencí, kdy *výchozí zdroj* (např. citace článku) odkazuje hypertextovou vazbou přímo na *cílový zdroj* (plný text článku), se oddělí popis zdroje (citace s odkazem) od poskytování vazeb, takže obecné vazební schéma pak vypadá následovně:

výchozí zdroj odkazuje na *servisní službu* (service component), která teprve odkazuje na správný *cílový zdroj*.

Implementace tohoto schématu v kontextu SFX je založena na několika principech:

1. servisních služeb existuje více, uživatel je i se svými právy registrován u některé z nich (servisní službu může implementovat třeba jeho knihovna nebo nějaká třetí strana)

2. aby servisní služba mohla určit (dynamicky) správné cílové zdroje (nemusí být jeden) pro daný výchozí zdroj a daného uživatele, potřebuje znát podrobnosti o výchozím zdroji – jeho metadata

3. tato *metadata nese v sobě přímo URL výchozího zdroje*, na který uživatel klikl, a to zakódována v podobě OpenURL. Například výchozím zdrojem nechť je citace článku v databázi Medline nakladatele Ebsco Publishing: *Moll, JR. Attractive electrostatic interactions. J Biol Chem. 2000 Nov 3, 275(44):34826-32. doi:10.1074/jbc.M004545200*
Nakladatel doplní k této citaci OpenURL, které může mít následující tvar:
http://sfx1.exlibris.com/demo?sid=ebsco:medline&aulast=Moll&auinit=JR&date=2000-11-03&stitle=J%20Biol%20Chem&volume=275&issue=44&spage=34826
První částí OpenURL je adresa servisní služby, za ní následuje identifikátor zdroje, v němž uživatel klikl na OpenURL, a poslední částí jsou metadata a identifikátory výchozího zdroje zakódována dle specifikace OpenURL [61] (NISO již zahájilo tzv. zrychlené řízení pro přijetí OpenURL jako ANSI standardu)

4. protože OpenURL jsou ve výchozím zdroji vytvářena dynamicky, je možné a potřebné v nich adresu servisní služby měnit tak, aby odpovídala té správné servisní službě příslušného uživatele. K propojení uživatele a jeho servisní služby nenabízí současná infrastruktura Webu žádný dostatečně obecný solidní mechanismus. Nicméně existuje několik pragmatických řešení (například mechanismus CookiePusher).

5. koncepce předpokládá spolupráci producentů informačních zdrojů ve smyslu doplnění OpenURL-odkazů do jejich zdrojů, a třetích stran při implementaci

servisních služeb. Překvapivě během velmi krátké doby od zveřejnění specifikace OpenURL ohlásila řada světově významných producentů informací dostupnost svých „OpenURL enabled“ zdrojů, a izraelská firma Exlibris (producent knihovního systému Aleph používaného v Národní knihovně ČR a v některých dalších velkých státních vědeckých knihovnách) získala licenci na SFX [70] a uvedla na trh první komerční implementaci servisní služby SFX-server a komplexní řešení pro integraci heterogenních digitálních zdrojů (zahrnujících i SFX server) pod názvem Metalib.

Souhrnný scénář práce v prostředí SFX vypadá následovně:

- uživatel přes standardní www-prohlížeč vyhledá v informačním zdroji (např. v citační databázi Web of Science) výchozí zdroj (citaci článku) a klikne na jeho OpenURL
- OpenURL zdroje odkazuje na servisní službu uživatele; ta je aktivována a z obdrženého OpenURL si vyzvedne metadata výchozího zdroje
- servisní služba vyhodnotí metadata výchozího zdroje (například provede vyhledání informací o výchozím zdroji v různých databázích, k nimž má uživatel oprávnění)
- vrátí uživateli klikací seznam příslušných cílových zdrojů (appropriate extended service links), který může zahrnovat: plný text zdroje, odkaz na záznam v lokálním online katalogu s uvedením kde v knihovně se nachází fyzická verze dokumentu, odkazy na další práce autora výchozího zdroje vyhledané Internetovským vyhledávačem, atd.

V [77] je uveden příklad jednoho z dalších možných využití technologie OpenURL a SFX v kombinaci se systémem DOI, který umožňuje aplikovat výše uvedený scénář i na informační zdroje, které nepodporují OpenURL.

Technologie OpenURL a SFX otevírá nové low-barier možnosti pro širokou integraci (interoperabilitu) heterogenních informačních zdrojů v současných digitálních a heterogenních knihovnách.

7 Globální vyhledávání zdrojů

Stejně tak jako jsou navzájem provázány předchozí dvě probírané oblasti digitálních knihoven (metadata pro interoperabilitu a interoperabilita metadat), tak i oblast globálního vyhledávání zdrojů v distribuovaném prostředí DL souvisí velmi těsně jak s metadaty tak s interoperabilitou – a naopak.

7.1 Úvod a stručný přehled

Detailní rozbor všech aspektů této problematiky lze nalézt v [30]; stručně je lze shrnout do pěti podoblastí: organizace, systémy, digitální obsah, rozhraní a metriky.

Organizace: v oblasti distribuovaného vyhledávání má každé řešení svůj organizační aspekt. Mezi heterogenními distribuovanými nezávisle spravovanými systémy musí vždy existovat určitá forma koordinace, má-li být vyhledávání zdrojů dostatečně efektivní. Jak již bylo naznačeno u interoperability, tato koordinace může mít velmi rozdílné formy – od širokého rozšíření silných standardů a komunikačních

protokolů až po velmi volnou kooperací založenou jen na použití stejných základních technologií (shromažďování dat z www-serverů Internetovskými vyhledávači). Strategie pro organizaci distribuovaných komponent DL musí brát do úvahy různorodost zainteresovaných institucí, jejich rozdílné priority, potřeby, cíle – ale také třeba bezpečnostní a cenové přístupy.

Systémy: existuje silná potřeba vyvinout systémovou infrastrukturu podporující vyhledávání, navigaci, zprostředkovávání a získávání informací v záplavě různorodých dat dostupných online. Součástí této infrastruktury musí být nástroje pro výběr informačních bází na systémové úrovni (routing dotazů ke správným fyzickým serverům), interakci informačních bází s překonáním jejich heterogenity (mezirepozitářové protokoly, distribuované vyhledávací protokoly, mechanismy pro zajištění bezpečnosti, soukromí, kooperativní autentifikace, placení) a zajištění konzistence ve složitém distribuovaném systému.

Obsah: množství a variabilita forem digitálního obsahu vyžaduje být schopen řešit efektivně problémy jako je optimální výběr informačních bází na logické úrovni (za použití metadat pro popis celých informačních bází zahrnujících na jedné straně obsah a jeho kvalitu, ale na druhé straně též výkonnostní, cenové a další přístupové parametry), dotazovací jazyky pro netextové informační zdroje (multimediální a dynamické dokumenty), nástroje pro ohodnocování vyhledaných informačních zdrojů (ratings) a efektivní filtraci informací, a konečně také mechanismy pro překonání sémantické heterogenity mezi informačními bázemi umožňující přechod od vyhledávání explicitních informací k získávání implicitních poznatků (knowledge discovery).

Rozhraní: oblast HCI (human-computer interaction) lze z pohledu digitálních knihoven rozdělit zhruba do čtyř rovin; první dvě se tradičně týkají vstupu a výstupu (mechanismy konstrukce a zadávání dotazů na vstupu, prezentace či vizualizace výsledků při výstupu), další dvě mají co do činění s pokusy o strojové porozumění tomu, co uživatel zamýšlí provádět (task understanding) a naopak pochopení procesů realizovaných systémem ze strany uživatele (process exposure) – zatímco někteří uživatelé jsou mnohem produktivnější, když rozumí tomu, jak jejich nástroj pracuje, jiní mohou být větším množstvím detailů zmatení a preferují black-box přístup. Řešením může být podpora pro široký individualizovaný přístup.

Metriky: pro vyhodnocování efektivity různých řešení a přístupů jsou vytvářeny nejrůznější taxonomie pro různé třídy uživatelů a vzorce jejich chování, pro dotazovací mechanismy, prezentaci výsledků apod., které je nutné testovat na reálných datech a reálných uživatelích. Silně je pociťována potřeba odpovídajících rozsáhlých ověřovacích prototypových řešení (testbeds), které by zahrnovaly velké množství distribuovaných informačních bází, široké spektrum médií a formátů a diversifikovanou informaci z pohledu kvality, časových charakteristik a cílových tříd uživatelů – to vše spolu s distribuovanou sdílenou kolekcí služeb a vyhledávacích a navigačních nástrojů.

Dosavadní praxe ve sféře globálního distribuovaného vyhledávání zdrojů potvrzuje řadu poznatků z historie v tom smyslu, že hrubá výpočetní síla zatím vítězí nad přístupy založenými na umělé (a někdy i přirozené) inteligenci. Arms v [5] popisuje oblasti, v nichž využití hrubé síly přineslo v posledních letech překvapivě dobré výsledky: vyhledávání informací (webovské vyhledávače), rozhodování nakolik vyhledaný dokument odpovídá zadanému dotazu (přístupy z oblasti sémantiky

dokumentů, viz např. DL projekt University v Illinois [75]), vyhodnocování důležitosti dokumentů (řadicí algoritmus systému Google [32]), archivace digitálního dědictví (automatizovaný přístup v Internet Archive [43] nebo švédském programu Kulturarw3 [46]), citační analýza (ResearchIndex [12]), kontextové propojování informačních zdrojů (SFX viz výše), automatická extrakce metadat z multimediálních digitálních objektů (Informedia Digital Video Library na univerzitě Carnegie Mellon [42]) nebo pokusy o vytvoření automatického referenčního knihovníka (projekt na univerzitě ve Washingtonu [9]).

7.2 DL a Internetovské vyhledávače

Informační exploze na Internetu vyvolala potřebu okamžitého pragmatického řešení problému jak v chaotickém moři informací vyhledávat a zprostředkovávat přístup k požadované informaci. Odpovědí byly Internetovské vyhledávací služby – vyhledávače (search engines) a adresáře (directories). Při srovnání vyhledávačů s přístupy klasických knihoven jsou rozdíly markantní; stručně a výstižně to charakterizuje [5]: *„Prakticky všechno co je nejlepší v knihovních katalozích, je mizerné u webovských vyhledávačů – a naopak"*. V tomto duchu byly až donedávna i digitální knihovny a Internetovské vyhledávače považovány obecně za dvě naprosto nezávislá paradigmata využívající webovského prostředí k vytváření informačních repozitářů. Práce [35] ukazuje, že ve skutečnosti mají obě hodně společného a je třeba je chápat nikoliv jako konkurenční nýbrž alternativní doplňující se přístupy (vyhledávače pro rychlou první odpověď, DL pro vysoce kvalitní cílenou informaci). Digitální knihovny jsou teoreticky dobře podložené, perspektivní, nabízí či slibují širší a v mnoha aspektech lepší služby; prakticky jsou však zatím stále ještě ne dostatečně zvládnuté a v globálním měřítku nerealizované. Webovské vyhledávače jsou naopak prakticky realizované a široce dostupné, avšak jejich vyhledávání je obecně málo přesné, zaměřené pouze na oblast volně dostupných zdrojů na tzv. povrchovém webu (pro vyhledávače nedostupný „hluboký" web je údajně až 500 krát rozsáhlejší [17]) a řadu dalších služeb nad rámec vyhledávání nerealizují vůbec.

Ve své krátké historii prošly oba přístupy třemi etapami s mnoha podobnými charakteristikami:

Vyhledávače: první generace (základní vyhledávače) je představována relativně jednoduchými přístupy založenými na jednoduchých metadatových strukturách, plnotextových indexech. Existují v podobě vyhledávačů buď univerzálních (AltaVista, Lycos) nebo specializovaných (MedHunt, TravelFinder). Druhá generace (meta-, multi-vyhledávače) klade důraz na snazší metody pro lokalizaci zdrojů, redukci nasbíraných výsledků, jednoduché metody jejich ohodnocování a kombinaci více různých základních vyhledávačů (MetaCrowler, SavvySearch). Třetí generace (popularity-, parallel-, portal-vyhledávače) spojuje vyhledávače a adresářové služby a nabízí pokročilejší techniky pro vyšší kvalitu služeb (lepší ohodnocování, kontextové techniky pro identifikaci relevantních vazeb), zohlednění uživatelských potřeb (uživatelská zpětná vazba a individualizace) a rychlejší vyhledávání. Příklad: Google, FAST, DirectHit, FizziLab.

Digitální knihovny: první generaci (stand-alone DL) představovaly víceméně klasické plně digitalizované a izolované digitální knihovny s lokálně ohraničeným a centralizovaným digitálním materiálem. Existovaly buď jako univerzálněji zaměřené

(digitální knihovna Kongresové knihovny, projekt Alexandria) nebo specializované (Making of America na Michiganské univerzitě, digitální knihovna ACM). Druhá generace (federalizované DL) byla nejčastěji organizována jako federace několika nezávislých samostatných DL organizovaných kolem společného tématu a nabízející jednotné uživatelské rozhraní pro transparentní přístup k heterogenním komponentám (viz Networked CompSci Technical Reference Library [57]). Třetí generace (sklízené DL) je představována virtuálními digitálními knihovnami poskytujícími sumarizovaný přístup k relevantnímu materiálu rozmístěném po globální síti. Obsahem takové knihovny bývají pouze metadata získávaná s využitím automatizovaných technik sklízení (harvesting) na základě definic informačního prostoru vytvářených informačními specialisty a při kontrole potřebné kvality (SourceBank, ArticleCentral.com!).

V [35] se předpovídá postupně konvergující vývoj obou přístupů: přes inteligentní vyhledávače a inteligentní digitální knihovny více využívající technik umělé inteligence a správy znalostí, až po společný Mega/Meta Portál poskytující unifikovaný přístup a deklarativní vyhledávání ke všem datovým repozitářům vytvořeným libovolnými technikami z obou přístupů.

8 Programy a projekty

V současnosti existují tisíce dokončených nebo probíhajících projektů digitálních knihoven po celém světě. Popsat stručně jen malou část z nich by vyžadovalo samostatnou rozsáhlou přednášku. V textu již byly zmíněny či odkázány některé projekty podílející se na vývoji vybraných klíčových komponent současné infrastruktury DL. Doplňme proto jen ty celosvětově nejdůležitější programy, které podporují jak výzkum tak i praktický vývoj a budování konkrétních digitálních knihoven, a které přinášejí nejvýznamnější podněty pro celou oblast.

8.1 Digital Library Initiative – Phase 1

Od počátku 90.let probíhala v odborných kruzích ve Spojených státech široká diskuse o potřebě zásadní pomoci výzkumu na podporu vlny nově vznikajících projektů z oblasti digitálních knihoven a jeho začlenění do programu národní informační infrastruktury. Pod koordinací National Science Foundation (NSF) a za spolupráce s agenturou DARPA (Defense Advanced Research Project Agency) a kosmickou agenturou NASA vznikl pětiletý program Digital Library Initiative, Phase 1 (DLI-1) [20] pro období 1994-1998, jehož cílem bylo: *dosáhnout zásadního technologického pokroku při sběru, ukládání a organizaci digitálních informací a jejich uživatelsky přívětivého zpřístupnění v globální síti*. Jako prostředek k dosažení tohoto cíle byla zvolena masivní finanční podpora jen velmi omezenému počtu špičkových výzkumných projektů z různých oblastí DL, které měly šanci na dosažení zásadního průlomu v poznání nových technologií a jejich ověření prostřednictvím rozsáhlých prototypových řešení (testbeds). Celkem bylo vybráno 6 projektů předních amerických univerzit (každá z nich vytvořila k řešení projektu výzkumnou alianci zahrnující řadu dalších subjektů, včetně významných komerčních firem), z nichž

každý dostal podpůrný grant ve výši 4 miliónů USD (včetně dalších zdrojů dosáhly celkové náklady na řešení těchto projektů 75 miliónů USD):

University of Michigan DL Project: zaměřený na vytváření rozsáhlé multimediální DL z oblasti věd o zemi a výzkumu vesmíru tvořené velkým množstvím informačních repozitářů a systematickým způsobem zpřístupňující na Internetu velké množství informací z mnoha různých tématických oblastí.

University of Illinois – Building the Interspace: DL Infrastructure for a University Engineering Community: zaměřený na integraci přístupu k textovým dokumentům ve formě (různě označkovaných) elektronických verzí článků v SGML z odborných technicky zaměřených časopisů od různých producentů. Součástí bylo i zkoumání algoritmů využívajících statistických technik pro analýzu sémantiky dokumentů.

University of California-Berkeley – The Environmental Electronic Library: A Prototype of a Scalable, Intelligent, Distributed Electronic Library: zaměřený na vývoj technologií pro inteligentní přístup k obrovským distribuovaným databázím obsahujícím fotografie, satelitní snímky, mapy, videozáznamy, plné texty a další typy dokumentů s cílem zpřístupnit rozsáhlé množství veřejně přístupných dat z oblasti životního prostředí.

Carnegie Mellon University – Informedia: Integrated Speech, Image and Language Understanding for Creation and Exploration of Digital Video Libraries: využití integrovaných technologií z oblastí rozpoznávání řeči, porozumění přirozenému jazyku a zpracování obrazu/videosekvencí pro obsahově založené vyhledávání v terabytové digitální video-knihovně.

Stanford University Integrated Digital Library Project: vývoj technologií pro integraci širokého spektra existujících i budoucích heterogenních sbírek a informačních zdrojů do virtuální digitální knihovny s jednotným přístupem ke všem jejím komponentám. Vyvíjené technologie byly ověřovány na prototypu InfoBus.

University of California-Santa Barbara – The Alexandria Project: Towards a Distributed DL with Comprehensive Services for Images and Spatially Referenced Information: digitální knihovna pro snadný přístup k rozsáhlým a různorodým sbírkám map, obrázků, leteckých snímků z Kalifornské oblasti s využitím nástrojů z geografických informačních systémů.

8.2 Digital Library Initiative – Phase 2

Na program DLI-1 bezprostředně navázal jeho následník DLI-2 pro období 1998-2002. Není již zaměřen jen na výzkum, ale také na budování digitálních sbírek a rozšíření sféry působnosti i do nových oblastí, především do medicínských a humanitních oborů. Jeho moto je: *zajistit vedoucí roli ve výzkumu klíčovém pro vývoj nové generace digitálních knihoven, zvýšit využívání a použitelnost globálních distribuovaných síťových informačních zdrojů a povzbudit stávající i nové komunity v zaměření na nové inovativní aplikační oblasti DL.* Stručně by se dal program charakterizovat podle následujících hesel: lépe využívat to co již existuje a zjistit, co ještě chybí – komunikovat a spolupracovat – učinit technologii (pro uživatele) neviditelnou.

Ke třem vyhlašovatelům DLI-1 se přidaly další instituce (Kongresová knihovna, Národní lékařská knihovna a další), zvýšil se objem finanční podpory na 15 miliónů

USD ročně po dobu pěti let a program se stal otevřeným (průběžně vypisovaná nová kola výběrového řízení, širší zaměření projektů a různá délka řešení, možná účast zahraničních partnerů). Počátkem roku 2001 získalo grant celkem již 47 projektů, z toho 11 mezinárodních, pokrývajících velmi široké spektrum výzkumných a aplikačních oblastí; mezi nimi jsou i projekty navazující na oněch šest již uzavřených projektů z DLI-1. Podrobnější informace o programu a jednotlivých projektech lze získat na [21].

8.3 Electronic Library Programme (eLIB)

Britský program eLIB [27], the Electronic Library Programme, probíhal ve třech etapách v letech 1994-2000 a na rozdíl od převážně výzkumně zaměřeného programu DLI byl orientován ryze prakticky s cílem pokrýt co nejširší oblast středoškolského a vysokoškolského sektoru – při řešení celkem 80 projektů se do něj zapojilo více než stovka vzdělávacích institucí. Mezi podporované oblasti v prvních fázích programu patřily e-publishing a digitalizace, přístup k elektronickým zdrojům a elektronické dodávání dokumentů, vzdělávání a výuka; v závěrečné fázi se podpora soustředila na hybridní knihovny, dlouhodobé uchovávání digitálního materiálu, realizace souborných virtuálních katalogů s využitím technologie Z39.50 a zejména transformaci řešení a služeb vytvořených v prvních fázích projektu do podoby trvale provozovatelných služeb. Program se stal katalyzátorem pro široký rozvoj elektronických informačních služeb a digitálních knihoven a získávání teoretických i praktických zkušeností z oblasti DL na britských vzdělávacích institucích.

8.4 National Digital Library Program (NDLP)

Kongresová knihovna získala první rozsáhlejší zkušenosti s velkoplošnou digitalizací a zpřístupňováním digitálního obsahu v pilotním projektu *American Memory* (1990-1995). V návaznosti na něj pak vyhlásila pětiletý program National Digital Library Program (NDLP), který James O'Donnell v předmluvě k [49] označil za knihovní "Apollo Project". Cílem tohoto programu [51] bylo ve velmi krátké době zdigitalizovat a zpřístupnit na síti 5 miliónů artefaktů ze sbírek Kongresové knihovny týkajících se Americké historie (jedinečné fotografie, rukopisy, vzácné knihy, mapy, zvukové nahrávky, filmy), zejména pro potřeby výuky na všech typech škol, od mateřských až po univerzity (hlavní cílovou skupinou jsou žáci základních a středních škol). Výsledky programu, na kterém spolupracuje s Kongresovou knihovnou řada dalších významných knihoven, škol i komerčních organizací, jsou soustředěny do více jak stovky digitálních multimediálních sbírek sdružených pod American Memory Historical Collections [1]; v době psaní tohoto příspěvku obsahovaly sbírky přes 7 miliónů digitálních položek. Jen pro představu: v polovině roku 1999 zaměstnával program NDLP více než 100 osob a měl roční rozpočet 12 miliónů USD. Program vyvinul vlastní standardy, digitalizační postupy a doporučení, metody integrace heterogenních digitálních sbírek a prezentační metody; velká pozornost je věnována problematice dlouhodobého uchovávání digitální informace.

8.5 Ostatní

Evropská unie zatím nemá žádný samostatný program zaměřený výlučně na digitální knihovny. Nicméně v rámci tématického programu IST 5.rámcového programu existuje v rámci skupiny Multimedia Content and Tools jako jedna z hlavních oblastí *Digital Heritage and Cultural Content*, která každoročně vyhlašuje několik témat úzce s problematikou DL souvisejících (např. téma Next Generation Digital Collections vyhlášené pro rok 2001). Kromě výzkumně zaměřeného programu IST existuje paralelní aplikační program eEurope Initiative na léta 2001-2004, v jehož rámci byla vyhlášena aktivita *eContent* zaměřená na vytváření a zpřístupňování evropských digitálních sbírek.

Z dalších zemí, které jsou na poli DL velmi aktivní, zmiňme především Německo, Francii (a její projekt Bibliotheca Universalis, který se postupně rozvinul v mezinárodní kooperativní digitalizační program *Paměť světa* po záštitou UNESCO [55]), a z mimoevropských zemí především Austrálii a Kanadu. Široce pojaté a štědře dotované programy na podporu rozvoje digitálních knihoven nejsou však jen doménou státních rozpočtů, jak o tom svědčí například ambiciózní Library Digital Initiative program Harvardovy univerzity [37].

Literatura

1. American Memory: Historical Collections for the National Digital Library. http://memory.loc.gov/
2. Ariadne Magazine. http://www.ariadne.ac.uk/
3. ARL Digital Initiatives Database. http://www.arl.org/did/
4. Arms, W.Y.: *Digital Libraries*. MIT Press, Cambridge, 2000. ISBN 0-262-01880-8.
5. Arms, W.Y.: *Open Access to Digital Libraries: Must Research Libraries Be Expensive?* Invited Talk to European Conference on DL 2000, Lisbon, 2000. http://www.ibl.pt/org/agenda/ecdl2000/arms.htm
6. Arms, W.Y., Blanchi, C., Overly E.A.: An Architecture for Information in Digital Libraries. *D-Lib Magazine*, February 1997. http://www.dlib.org/
7. Arms, W.Y.: Key Concepts in the Architecture of the Digital Library. *D-Lib Magazine*, July 1995. http://www.dlib.org/
8. arXiv.org e-Print archive. http://www.arxiv.org/
9. Automatic Reference Librarian Project. University of Washington. http://www.cs.washington.edu/research/diglib/
10. Bartošek, M.: Vyhledávání v Internetu a DUBLIN CORE. *Zpravodaj ÚVT MU*. ISSN 1212-0901, (1999) roč.9, č.4, s.1-4 . http://www.ics.muni.cz/bulletin/issues/vol09num04/bartosek/bartosek.html
11. Berkeley Digital Library SunSITE. http://sunsite.berkeley.edu/
12. Bratková, E.: Citace odborné literatury jako nástroj rozvoje služeb a integrace digitálních knihoven. In: *Automatizace knihovnických procesů 8*, Barbora Ramajzlová (Ed.), ČVUT Praha (2001), 109-120. Též na: http://knihovny.cvut.cz/akp/clanky/12.pdf
13. Bush, V.: As We May Think. *Atlantic Monthly*. July (1945) 101-108. Též na: http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm

14. Cleveland, G.: Digital *Libraries: Definitions, Issues and Chalenges*. UDT Occasional Paper #8. International Federation of Library Associations and Institutions. 1998. http://www.ifla.org/VI/5/op/udtop8/udtop8.htm
15. CrossRef. http://www.crossref.org/
16. D-Lib Magazine. http://www.dlib.org
17. Deep Web. CompletePlanet. http://www.completeplanet.com/tutorials/deepweb/index.asp
18. Dempsey, L., Heery, R.: A Review of Metadata: *A Survey of Current Resource Descriptive Formats*. DESIRE RE-1004 Project, Deliverable D3.2 (1). EU Telematics Applications Programme. March 1997. http://www.ukoln.ac.uk/metadata/desire/overview/
19. Dienst. http://www.cs.cornell.edu/cdlrg/dienst/DienstOverview.htm
20. Digital Library Initiative, Phase 1 (DLI-1). http://www.dli2.nsf.gov/dlione/
21. Digital Library Initiative, Phase 2 (DLI-2). http://www.dli2.nsf.gov/
22. Digital Preservation. OCLC/RLG Working Group. http://www.oclc.org/digitalpreservation/
23. DOI – Digital Object Identifier. http://www.doi.org/
24. *The DOI Handbook*. International DOI Foundation, 2001. http://dx.doi.org/10.1000/182
25. Dublin Core Metadata Initiative. http://dublincore.org/
26. Dublin Core Czech. http://www.ics.muni.cz/dublin_core/DC-czech-1.1.html
27. Electronic Library Programme (eLIB). http://www.ukoln.ac.uk/services/elib/
28. EU-NSF Working Group. *Metadata for Digital Libraries: A Research Agenda*. Report. ERCIM-DELOS, 1999. http://www.iei.pi.cnr.it/DELOS/NSF/metadata.html
29. *EU-NSF Digital Library Working Group on Interoperability between Digital Libraries*. Position Paper. ERCIM-DELOS, 1999. http://www.iei.pi.cnr.it/DELOS/NSF/interop.htm
30. EU-NSF Working Group. *Resource Discovery in a Globally-Distributed Digital Library*. Report. ERCIM-DELOS, 1999. http://galileo.iei.pi.cnr.it/DELOS/REPORTS/resourcediscovery.htm
31. Fox, E.: Digital Libraries – Virginia Tech Courseware. http://ei.cs.vt.edu/~dlib/
32. Google. http://www.google.com/
33. Greenstone. New Zealand Digital Library. http://nzdl.org/
34. Hakala, J.: Document Description and Access – New Chalenges. In: *CASLIN 2001*, pracovní materiály ke konferenci. Knihovna Akademie věd ČR, Praha, (2001), 33-46.
35. Hanani, U., Ariel, J.F.: *The Parallel Evolution of Search Engines and Digital Libraries: Their Convergence to the Mega-Portal*. In Proc. of Kyoto International Conference on Digital Libraries, 2000, 269-276.
36. Handle System. CNRI. http://www.handle.net/
37. Harvard University Library Digital Initiative. http://hul.harvard.edu/ldi/
38. INDECS: Interoperability of Data in E-commerce Systems. http://www.indecs.org/
39. IFLANET – Digital Libraries: Resources and Projects. IFLA. http://www.ifla.org/II/diglib.htm

40. IFLANET – Digital Libraries: Metadata Resources. IFLA. http://www.ifla.org/II/metadata.htm
41. IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records : final report*. IFLA, 1998. Též na http://www.ifla.org/VII/s13/frbr/frbr.htm
42. Informedia Project. Carnegie Mellon University. http://www.informedia.cs.cmu.edu/
43. Internet Archive. http://www.archive.org/
44. Kahn, R., Wilensky, R.: *A Framework for Distributed Digital Object Services*. Technical Report hdl:cnri.dlib/tn95-01, CNRI, May 1995. http://www.cnri.reston.va.us/home/cstr/arch/k-w.html
45. Kenney, A.R., Oya, Y.R.: *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. RLG, 2000. ISBN 0-9700225-0-6
46. Kulturarw3 Project. The Royal Library, National Library of Sweden. http://kulturarw3.kb.se/html/kulturarw3.eng.html
47. Lagoze, C., Shaw, E., Davis, J.R., Krafft, D.B.: *Dienst: Implementation Reference Manual*. Cornell University, May 1995.
48. Lesk, M.: *Practical Digital Libraries. Books, Bytes, and Bucks*. Morgan Kaufmann Publishers, Inc, San Francisco, 1997. ISBN 1-55860-459-6.
49. *LC21: A Digital Strategy for the Library of Congress*. Committee on an Information Technology Strategy for the Library of Congress, Computer Science and Telecommunications Board, National Research Council. 288 stran. 2001. ISBN 0-309-07144-5. Též na http://books.nap.edu/html/lc21/
50. Library of Congress – Core Metadata Elements. http://lcweb.loc.gov/standards/metadata.html
51. Library of Congress – National Digital Library Program. http://memory.loc.gov/ammem/dli2/html/lcndlp.html
52. Library of Congress – Z39.50 Maintenance Agency. http://lcweb.loc.gov/z3950/agency/
53. Lynch, C.: The Z39.50 Information Retrieval Standard. Part 1: A Strategic View of Its Past, Present and Future. *D-Lib Magazine*, April 1997
54. Lynch, C., García-Molina, H.: *Interoperability, Scaling, and the Digital Libraries Research Agenda*. IITA Digital Libraries Workshop Report, 1995. http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.htm
55. Memoriae Mundi Series Bohemica (Paměť světa). Národní knihovna ČR. http://digit.nkp.cz/
56. MPEG-7. http://www.cselt.it/mpeg/
57. NCSTRL – Networked Computer Science Technical Reference Library. http://www.ncstrl.org/
58. Nikolaou, C., Marazakis, M.: System Infrastructure for Digital Libraries: A Survey and Outlook. In: *SOFSEM'98*, Rovan, B. (Ed.), Springer-Verlag, (1998), 186-203.
59. NISO Standards and Technical Reports. Techstreet. http://www.techstreet.com/nisogate.html
60. Open Archives Initiative. http://www.openarchives.org/
61. OpenURL Syntax Description. http://www.sfxit.com/OpenURL/openurl.html

62. Paepcke, A., et al: Search Middleware and the Simple Digital Library Interoperability Protocol. *D-Lib Magazine*, March 2000. http://www.dlib.org
63. Paepcke, A., Chang, C.K., García-Molina, H., Winograd, T.: Interoperability for Digital Libraries Worldwide. *Communication of the ACM*, 41(4) (1998), 33-43.
64. Paskin, N.: Information Identifiers. *Learned Publishing*, 10(2) (1997), 135-156. Též na http://www.elsevier.nl/homepage/about/infoident/
65. Pokorný, J.: Digitálni knihovny: Principy a problémy. In: *Automatizace knihovnických procesů 8*, Barbora Ramajzlová (Ed.), ČVUT Praha (2001), 27-38. Též na http://knihovny.cvut.cz/akp/clanky/03.pdf
66. PURL – Persistent URL. OCLC. http://purl.oclc.org/
67. RDF. W3C Consortium. http://www.w3.org/RDF/
68. RLG DigiNews. http://www.rlg.org/preserv/diginews/
69. Samuel, A.L.: The Banishment of Paperwork. *New Scientist* 21 (1964) 529-530.
70. SFX. ExLibris. http://www.sfxit.com/
71. Schauble, P., Smeaton, A.F. (Eds): A Research Agenda for Digital Libraries. *Summary Report of the Series of Joint NSF-EU Working Groups on Future Directions For Digital Libraries Research*. Brussels, October 1998. Též na http://www.iei.pi.cnr.it/DELOS/NSF/Brussrep.htm
72. Snijder, R.: *Metadata Standards and Information Analysis: A Survey of Current Metadata Standards and the Underlying Models*. http://www.geocities.com/ronaldsnijder/index_files/index.html
73. Stanford University Digital Libraries Project. http://www-diglib.stanford.edu/diglib/pub/userinfo.html
74. Text Encoding Initiative. http://www.tei-c.org/
75. University of Illinois DL. http://dli.grainger.uiuc.edu/idli/idli.htm
76. Uniform Resource Names. http://www.ietf.org/html.charters/urn-charter.html
77. Van de Sompel, H., Beit-Arie, O.: Open Linking in the Scholarly Information Environment Using the OpenURL Framework. *D-Lib Magazine*, March 2001.
78. Van de Sompel, H., Hochstenbach, P.: Reference Linking in Hybrid Library Environment. Part 1, Part 2. *D-Lib Magazine*, April 1999
79. Van de Sompel, H., Hochstenbach, P.: Reference Linking in Hybrid Library Environment. Part 3. *D-Lib Magazine*, October 1999
80. Waters, D.J.: What are digital libraries? *CLIR Issues*, July/August 1998. http://www.clir.org/pubs/issues/issues04.html
81. XML. W3C Consortium. http://www.w3.org/XML/
82. XMLMARC Project. http://xmlmarc.stanford.edu/

Annotation:

Digital Libraries

The paper provides description of main fields in research and practice of Digital Libraries from the computer science perspective – general architecture, global information identifier schemas, metadata, interoperability, and distributed information discovery. Overviews of general approaches in each of the fields are complemented with examples of practical solutions representing building blocks of current DL infrastructure. Short excursion into DL history, description of main programmes supporting development in the field of digital libraries as well as the extensive bibliography for further study are also included.