DIGITÁLNÍ KNIHOVNY, JEJICH ORGANIZACE A SLUŽBY

Otakar Pinkas, Vysoká škola ekonomická v Praze

Cíl

Cílem příspěvku je představit organizaci a služby digitálních knihoven na dvou základních typech, kterými jsou jednoduchý plnotextový archiv a distribuovaný síťový informační systém.

1. Úvod

Rozvoj služby WWW znamenal průlom do způsobu zveřejňování a distribuování dokumentů. Internetové prostředí se stalo univerzálním a pružným informačním médiem. Kromě navigování v dokumentech WWW serverů se zavedením indexovacích a vyhledávacích strojů stalo velmi populárním vyhledávání informací. Ukázalo se však, že určité skupiny uživatelů potřebují přesnější vyhledávací nástroje a garantovaný přístup k pečlivě vybraným a dobře popsaným zdrojům.

V r. 1993 byla v USA vyhlášena Iniciativa v oblasti digitálních knihoven (Digital Libraries Initiative), jejímž cílem byl výzkum a vývoj digitálních knihoven. K americké výzvě se přidala řada vyspělých států, jejichž odborníci čerpali z amerických zkušeností. Tento příspěvek referuje hlavně o německých knihovnách,. protože americké digitální knihovny nebývají široce otevřené a používají často speciální klientské programy.

2. Organizace digitálních knihoven

Existuje mnoho pokusů o definici, vymezení nebo charakteristiku pojmu digitální knihovna. Je to pochopitelné, neboť se jedná o nový fenomén. Pro potřeby tohoto příspěvku lze vycházet z volnějšího vymezení: digitální knihovna je organizovaná sbírka digitálních dokumentů a soubor služeb určených k využívání určitou skupinou uživatelů v síťovém informačním prostředí Internetu.

Je snad oprávněné vymezit dva krajní typy digitálních knihoven: jednoduché plnotextové digitální archivy (např. vysoké školy) a distribuované síťové informační systémy v regionálním, národním a mezinárodním měřítku. Dále se pokusíme charakterizovat tyto typy a popíšeme jejich reálné příklady.

2.1 Jednoduchý plnotextový digitální archiv univerzity

Tento typ digitální knihovny obvykle obsahuje publikace učitelů a studentů školy. Mezi běžné druhy evidovaných dokumentů patří vědecké, diplomové a disertační práce, ale i projekty, noviny a časopisy, texty přednášek, aj.

Počet dokumentů se pohybuje od několika desítek do několika tisíců. Digitální archiv zajišťuje dlouhodobé uložení dokumentů a jejich spolehlivou citovatelnost (konstantní adresace). Dokumenty jsou trvale online přístupné a jejich obsah se nemění ani neruší. Za obsah ručí autor, který je také vlastníkem autorských práv. Úkolem autora je vyplnění archivní přihlášky a umístění příslušných souborů v dohodnuté pracovní oblasti serveru.

Provozovatel archivu podporuje vznik elektronické publikace, určuje archivní pravidla a zajišťuje vlastní archivaci. Obvykle nabízí celou řadu přístupových metod k archivovaným dokumentům, včetně uživatelského rozhraní ve formě WWW. Bývá zvykem udržovat vlastní systém ukládání, indexování, vyhledávání a prezentace publikací archivního fondu.

Provozovatelem archivu jsou zpravidla výpočetní střediska univerzit spolu s jejich knihovnami. V SRN je obecně přijata zásada katalogizovat vysokoškolské práce, které jsou pak dostupné i v tištěné formě. Knihovní katalogizační záznamy odkazují na umístění digitálních textů.

Datové soubory archivu jsou ve většině případů soustředěny na jednom serveru. Softwarové prostředky pocházejí ze sféry volného šíření a jsou upravovány a doplňovány o různé druhy CGI skriptů. Správci archivních serverů umožňují robotům vyhledávacích strojů přístup do archivů: vyskytují se i případy, kdy je nutný souhlas správce nebo kdy je přístup robotům odepřen. Uživatelé pracující s prohlížeči WWW se většinou nemusí registrovat ani přihlašovat.

2.2 Distribuované síťové informační systémy

Mnozí odborníci chápou digitální knihovny právě jako zvláštní případ distribuovaných síťových informačních systémů. Distribuované digitální knihovny mají některé charakteristické vlastnosti: distribuovanost, heterogenita fondů a služeb, interoperabilita, federalizace a přizpůsobitelnost (škálovatelnost). Distribuovanost může být různého druhu: zde stačí říci, že distribuovanost znamená rozdělení fondů a služeb do prostorově oddělených míst. Interoperabilita je schopnost autonomních a různorodých systémů vzájemně spolupracovat. Federalizace znamená propojení digitálních sbírek do jednoho celku, který lze oslovovat v jednom společném jazyce. Přizpůsobitelnost je schopnost systému reagovat na růst počtu dokumentů a poskytovatelů, uživatelů a jejich požadavků, bez nutnosti zásadní změny struktury celého systému.

Uživateli se jeví distribuovaný systém digitální knihovny jako jeden logický celek s vymezenými částmi, se kterým komunikuje v jednotném informačním jazyce a z něhož dostává informační odpovědi na své požadavky. Podle svých potřeb a znalostí určuje, zda bude volit některé nebo všechny poskytovatele informačních obsahů, kteří jsou součástí systému.

Distribuované digitálních knihovny mají celou řadu zajímavých a specifických úloh, z nichž vybíráme jednotné dotazování a automatické směrování dotazů.

Specifické úlohy distribuovaných digitálních knihoven

V distribuovaném systému s různými poskytovateli existují různé soubory atributů (polí) pro popis dokumentů nebo bibliografických záznamů

v různých databázových systémech. Existují i různé způsoby interakce s automatizovaným systémem: jeden si udržuje informace o spojení s uživatelem, jiný nikoli.

Jednotný dotazovací jazyk

Je-li v distribuované digitální knihovně zaveden společný dotazovací jazyk, musí systém být schopen efektivního překladu do dotazovacího jazyka specifické databáze. Správného překladu se dosahuje tím, že se uvnitř systému využívá mechanismu hierarchizace atributů. Definujeme-li atribut „původce", můžeme považovat za jeho specializaci atributy „autor" a „editor". Cesta od „autora" k „původci" je nazývána generalizací. Problém je samozřejmě složitější, protože kromě jednotlivých atributů musíme brát v úvahu i další prvky dotazovacího jazyka jako jsou booleovské a proximitní operátory, operátory levostranného a pravostranného rozšíření, atp. Některé vyhledávací systémy vůbec nepracují s booleovskými operátory a používají vážení termínů (vektorové vyhledávání).

Automatický výběr poskytovatelů dokumentů

V systémech více poskytovatelů je vhodné mít k dispozici komponentu digitální knihovny, která automaticky směruje dotaz na nadějná místa. Taková komponenta musí mít trvale k dispozici popisy obsahu určitého souboru databází, aby mohla provést správný výpočet a směrování dotazu. V systému ROADS (VB) se vytvářejí centroidy, což jsou speciální indexy určité databáze jako celku. Centroid je soubor atributů a k nim příslušných unikátních hodnot, které se v databázi vyskytují. Jestliže se v nějakém centroidu nevyskytne v seznamu hodnot atributu „autor" jméno Maleček, je zřejmé, že dotaz nebude do odpovídající databáze zaslán. Směrování dotazu se řídí nejen obsahovými hledisky, ale i dalšími, jako jsou cena za komunikaci, momentální zatížení serveru, atp.

3. Příklady reálných digitálních knihoven

3.1 Plnotextový archiv MONARCH Tech. univerzity v Chemnitz

Názorným příkladem digitálního archivu je archiv MONARCH. Lze ho charakterizovat obecnými vlastnostmi digitálního archivu, které jsou uvedeny výše. Skládá se ze dvou funkčních částí: archivační a přístupové.

Postup archivace

Autor vyplní WWW archivní formulář zahrnující údaje o autorovi (jméno a příjmení, fakulta), údaje o publikaci (název, druh publikace: diplomová práce, disertace, atp., jazyk, zdroj: souborový adresář a jména souborů), údaje o obsahu publikace (klíčová slova, abstrakt), údaje o archivaci (archivační lhůta, vyplňovatel přihlášky a jeho e-mail). Přípustné ukládací a prezentační formáty dokumentů jsou postscript, DVI, HTML, ASCII. Kontrolní program ověří způsob vyplnění, uloží dokument do pracovní oblasti a později jej v dávce přesune na konečné cílové místo. Dávka dokumentů se potom indexuje pomocí programu GLIMPSE, který slouží rovněž k plnotextovému vyhledávání.

Přístup do archivu

Uživatel pracuje se standardním WWW prohlížečem, jehož funkční možnosti je dobré rozšířit o zobrazování dalších druhů dokumentů. Adresa archivu je: http: //archiv.tu-chemnitz.de/pub/rok/číslo/ – rok a číslo jsou čtyřmístné.

Je-li uživateli znám rok a číslo publikace, může vstupovat do archivu přímo. Může také listovat v seznamu publikací uspořádaném podle let. Vyhledávat může podle bibliografických údajů extrahovaných z archivačního formuláře a pomocí slov z plného textu dokumentu. V rešeršním dotazu lze používat booleovské operátory a maskování argumentů. Záznamy lze filtrovat podle druhu nebo formátu dokumentů. Lze prohledávat jen vybranou část archivu, např. soubory disertací. Rešeršní formulář existuje v jednoduché a detailní formě. Přehled rešeršních výsledků obsahuje dotaz s tabulkovým přehledem prvků dotazovacího jazyka (argumenty, pole, logické operátory, filtry) a očíslovaný seznam zkrácených záznamů (název, autor, druh publikace, MIME typ a URL). U každého záznamu je uvedeno, v kterém poli došlo ke shodě argumentu a hodnoty pole. Od názvu vede hypertextový odkaz k úplnému záznamu (obsahuje navíc např. délku souboru). Kliknutím na název se vyvolá přenos úplného textu dokumentu.

Každá stránka plného textu obsahuje základní navigační tlačítka: skok na první/poslední stránku, přechod na předchozí/následující stránku. Způsob citování: Autor, název, http://archiv.tu-chemnitz.de/pub/rok/číslo/ – obě čísla čtyřmístně.

3.2 Systém MeDOC (Multimediale Elektronische Dokumente)

Koncepce a architektura

Posláním systému MeDOC (SRN) je zabezpečovat efektivní výměnu vědeckých a odborných informací a zjednodušovat obstarávání literatury v oblasti informatiky. Během let 1995–1998 byl uveden do praxe jako první systém digitální knihovny v SRN. Tento automatizovaný distribuovaný informační systém podporuje vytváření, ukládání, zpětné vyhledávání a výstup plnotextových informací v síťovém informačním prostředí Internetu. Obsahuje různorodé dokumenty od jednoduchých sdělení až po hierchicky uspořádané jednotky, kterými jsou např. vědecké a odborné časopisy. Dokumenty jsou rozmístěny v různých sbírkách na různých místech pod správou různých databází nebo informačních rešeršních systémů. Dokumenty jsou většinou volně přistupné, ale část je dostupná jen po splnění licenčních podmínek. Mezi informační objekty patří také celé systémy, např. automatizované bibliografické systémy.

Vzhledem ke složitosti a velké distribuovanosti má systém pět funkčních vrstev: uživatelská, uživatelských agentů, zprostředkovatelů, agentů poskytovatelů, poskytovatelů. V nejobecnějším pohledu jde o tříčlenný vztah: uživatelé – prostředníci – poskytovatelé. Vrstva uživatelských agentů umožňuje napojení do systému pomocí WWW prohlížeče. Vrstva agentů poskytovatelů přemosťuje různorodost poskytovatelských systémů a způsobů

interakce s nimi, čímž je dosaženo unifikovaného přístupu. Zprostředkovatelská vrstva umožňuje např. automatické směrování dotazů.

Z technického hlediska jsou vrstvy realizovány komunikujícími komponentami, které se nazývají agenti. Komunikace komponent probíhá podle protokolu MeDOC (na bázi HTTP). Požadavky na komponenty a vracené výsledky mají jednotné globální schema.

Uživatelský agent spravuje vždy určitou skupinu uživatelů (existuje více uživatelských agentů). Registruje a kontroluje uživatele a jejich přístupová práva, přijímá a odesílá jejich požadavky a shromažďuje, ukládá a eviduje výsledky obdržené od jiných komponent. Např. umožňuje registrovanému uživateli archivaci rešeršních dotazů a výsledků po několik týdnů.

Každému poskytovatelskému systému přísluší vždy jeden jeho agent. Ten transformuje jednotné požadavky MeDOC (např. rešeršní dotazy) do specifické formy poskytovatelského systému a naopak. Komunikuje s uživatelským agentem a brokerem, což je název komponenty zprostředkovatelské vrstvy. Specifickým poskytovatelem dokumentů je plnotextový archiv MeDOC, který existuje na více než pěti místech v Německu a obsahuje plné texty dokumentů z informatiky. Prostřednictvím agentů uživatele a poskytovatele lze v archivu vyhledávat, navigovat v celém fondu a listovat v plných textech dokumentů. Dokumenty lze elektronicky objednávat a expedovat. Zpoplatňování existuje v několika variacích.

Služby systému

Uživatel se může v systému zaregistrovat, přihlásit se k zahájení relace, klást informační dotazy, navigovat ve fondu, listovat v plných textech a objednávat dokumenty. Registrace je aktivní nebo pasivní. V prvním případě stačí do WWW formuláře zapsat jméno a heslo, avšak přístupová práva jsou omezena na volné dokumenty. V druhém případě uživatel kontaktuje určitou osobu. Při přihlášení se zapisuje elektronická adresa a heslo. Relace je zabezpečena jednorázovým klíčem. Informační dotaz (jednoduchý nebo detailní) může být zaslán vybraným, doporučeným (broker) nebo všem poskytovatelům. Přijaté výsledky agent uživatele roztřídí podle místa původu. Zvláštností je asynchronní reakce na dotaz. Nečeká se na shromáždění výsledků od všech poskytovatelů, což v praxi znamená reaktivaci dotazu (reload).

Navigace v plnotextovém archivu je implicitní nebo explicitní. Při implicitní je možné klikat na podtržená slova krátkého záznamu (autor, vydavatel, aj.), čímž se dostáváme k uspořádaným seznamům (např. knihy určitého vydavatele). Explicitní navigace nabízí volbu navigačních kritérií. Listování v plném textu dokumentu se odehrává přímo v archivu. K dispozici jsou nadstandardní možnosti. Objednávání a zasílání dokumentů jsme neměli možnost v praxi ověřit. Složitější dotazy zvládá systém s obtížemi a ne vždy správně přeloží informační dotaz do cílového tvaru, což vede k chybovým hlášením poskytovatele. V každém případě je potěšením jedním dotazem oslovit v síti Internet několik desítek poskytovatelů najednou.

Literatura:
1. Griffin, S.: NSF/DARPA/NASA digital libraries initiative. A program manager's perspective. D-Lib Magazine, July/August 1998. URL: http://www.dlib.org/dlib/july/98/07/griffin.html.
2. Boles, D. – Dreger, M. – Grossjohann, K. – Lohrum, S. – Menke, G.: MeDOC. Architektur und Funktionalitaet des MeDOC-Dienstes. Technischer Bericht, MeDOC 1996. URL: http://www.inf.tu- beURLin.de/ medoc3/ together/ivsges.ps.gz.
3. Fuhr, N.: Object-oriented and database concepts for the design of networked information retrieval systems. URL: http://ls6.informatik.uni-dortmund.de./ir/reports/96/Fuhr-96.html
4. MONARCH. Multimedia online archiv Chemnitz. Hilfe. URL: http://archiv.tu-chemnitz.de/hilfe.html
5. Glimpse – a UNIX search engine. URL:http://glimpse.cs.arizona.edu/