Profil Z39.50 JIB

Profil pro jednoduché vyhledávání a stahování záznamů

Jan Pokorný
Univerzita Karlova v Praze
Ústav výpočetní techniky
jan.pokorny@ruk.cuni.cz

INFORUM 2004: 10. konference o profesionálních informačních zdrojích
Praha, 25. - 27.5. 2004

1. PROTOKOL Z39.50 V ČR

Zpřístupňování on-line zdrojů pomocí protokolu Z39.50 je v České republice druhým nejběžnějším způsobem (po webovém přístupu), zejména díky integračním projektům jako je Jednotná informační brána (JIB) a podpůrným finančním programům jako je VISK8. Dnes existuje poměrně hustá síť knihoven a informačních institucí, které mohou pomocí protokolu Z39.50 vyměňovat své záznamy nebo dokumenty.

Samotná implementace protokolu Z39.50 však nestačí k optimálnímu vyhledávání. Protokol Z39.50 totiž představuje poměrně volný standard, který vyžaduje přesnější specifikaci v rámci skupiny, která ho využívá. Bylo proto nutné vytvořit takovou specifikaci na národní úrovni, aby všichni účastníci mohli pomocí Z39.50 komunikovat dohodnutým způsobem - aby komunikované dotazy a odpovědi interpretovali všechny zapojené Z39.50 servery a klienti stejným způsobem.

V České republice se téměř bez výjimek používá k Z39.50 komunikaci sada atributů 1 s názvem **Bib-1**, která je nezávislá na struktuře formátu MARC. Přestože se z velké části vyhledávání týká záznamů ve formátu MARC a zdálo by se logické využívat spíše sadu atributů 17, která respektuje strukturu MARC přímo, sada 1 umožňuje využití i jiných datových struktur než je MARC, což je velkou výhodou zvláště do budoucna, kdy se počítá s využíváním digitálních sbírek apod. Nutnost přesné specifikace protokolu Z39.50 v českých podmínkách byla proto ale o to větší. Od specifikace se totiž v první řadě očekával popis mapování atributů Bib-1 na jednotlivá pole/podpole MARC. Vzhledem k tomu, že využívání standardu MARC je v současné době v České republice duální, (část knihoven ještě využívá dlouho využívaný standard UNIMARC a část knihoven již přešla na nově podporovaný MARC 21), bylo třeba definovat mapování do obou standardů.

Specifikaci nastavení protokolu Z39.50 definuje tzv. **profil Z39.50**. Profilů Z39.50 existuje celosvětově celá řada a každý z nich plní svoji normativní roli v určité oblasti. V České republice se první snahy o vytvoření národního profilu objevily z iniciativy pracovni skupiny implementátorů Z39.50 (ZIG), u jejíhož zrodu stála Státní technická knihovna, a výsledkem byl návrh **Profilu CZ** z listopadu 2002, který koncepčně vycházel v profilu Bath. Přibližně ve stejné době začala nabízet své služby JIB, která ke komunikaci se zdroji využívá převážně právě protokol Z39.50. V krátké době se podařilo v JIB zajistit širokou nabídku českých on-line zdrojů, převážně katalogů, a umožnit je paralelně prohledávat. JIB se tak stala prvním systémem, kde se začaly prakticky projevovat všechny možné problémy nejednotné implementace Z39.50 v jednotlivých zdrojích.

2. DOTAZOVÁNÍ PŘES Z39.50 NA PŘÍKLADU

K tomu, abychom problémy Z39.50 pochopili, musíme dobře chápat datové toky celého procesu vyhledávání. Komunikace Z39.50 se skládá typicky ze 4 členů:

a) **vyhledávací klient**, který převezme od uživatele dotaz a předá ho klientovi Z39.50
b) **klient Z39.50**, který vysílá dotaz příslušnému serveru Z39.50

c) **server Z39.50**, který dotaz přijímá a předává ho vyhledávacímu stroji
d) **vyhledávací stroj** s datovým úložištěm, který provede vyhledávání

Typické je, že tyto členy mezi sebou nepoužívají na všech úrovních stejný komunikační protokol, dotazovací jazyk ani formát dat. Často při předávání dat dalšímu členovi komunikace provádějí konverze protokolu, dotazu i formátu. Pokud neexistují přesné definice těchto konverzí, dochází ke zkreslení původního dotazu a vrácené výsledky mají sníženou relevanci - vyhledávání funguje nekorektně.

Nejlépe si celý proces komunikace dotazu můžeme ukázat na příkladu vyhledávání z katalogizačního klienta:

Vyhledávací klient

Katalogizátor si ve svém editoru spustí vyhledávání záznamu na vzdáleném serveru Z39.50 na základě uvedení autora dokumentu, např. "Jiří Zikmund". Protože daná knihovna používá formát MARC 21, vygeneruje katalogizační klient dotaz, který můžeme zapsat např takto:

FIND 100$a="Zikmund, Jiří"

Klient Z39.50

Katalogizační klient předá dotaz klientovi Z39.50, při čemž provede konverzi dotazu do syntaxe atributů Z39.50. K takové konverzi musí použít mapování polí a podpolí MARC 21 do sady atributů Bib-1. Výsledný dotaz může vypadat takto:

FIND 1=1003 2=3 3=1 4=1 5=100 6=1 "Zikmund Jiří"

Pozn.: V tomto modelovém příkladu je jméno autora interpretováno jako fráze, protože cílový server indexuje jména jako fráze. Z fráze byla odstraněna oddělovací čárka.

Server Z39.50

Dotaz je zaslán vzdálenému serveru Z39.50 (klient Z39.50 i server Z39.50 musí podporovat stejné atributy, jinak dojde ke ztrátě informace v dotazu). Aplikace musí na serveru provést konverzi ze syntaxe atributů Z39.50 do dotazovacího jazyka vyhledávacího stroje. Vyhledávací stroje pracují na velice rozdílných principech a s různými datovými strukturami. Často se stává, že neumí přesně interpretovat význam dotazu, není schopen takový dotaz realizovat nebo není schopen vyhledávání provést v požadovaném čase. Pokud by vyhledávací stroj pracoval nad relační dotabází, mohl by výsledný dotaz SQL s ohledem na strukturu databáze vypadat např. takto:

SELECT COUNT(id) FROM knihy WHERE hl_autor = "Zikmund Jiří"

Vyhledávací stroj

Vyhledávací stroj spustí dotaz. Databáze je organizována tak, že bibliografické záznamy v MARC 21 jsou uloženy jako celý plný text do jednoho pole a aplikace z nich vytváří podle použitých pravidel uměle rejstříky na principu indexů. Ty poté využívá k vyhledávání. V indexech se ztrácí informace o pozici údajů v poli/podpoli, indexy jsou navíc společné pro více polí podle obsahu (např. autoři ze všech relevantních polí).

Vyhledávací stroj zjistí, kolik záznamů vyhovuje dotazu. Následně vyhledá

záznamy a množinu odešle prostřednictvím serveru Z39.50 zpět klientovi.

Z příkladu jasně vyplývá, že vyhledávání přes Z39.50 má celkem 3 kritické momenty, které se vždy týkají konverzí:

První konverzi musí provést už klient, které dotaz odesílá. V případě bibliografických databází v UNIMARC nebo MARC 21 se tato konverze v ČR většinou vždy týká převodu polí a podpolí MARC na Bib-1. Klient Z39.50, který dotaz posílá, a server Z39.50, který dotaz přijímá, musí podporovat stejné atributy Bib-1. Konverzi a nastavení atributů může z velké části pozitivně ovlivnit právě **profil Z39.50**.

Druhou konverzi provádí serverová aplikace, která musí dotaz z rozhraní Z39.50 v Bib-1 převést na dotazovací jazyk vyhledávacího stroje. Zde je situace velmi různorodá a těžko se dá zobecňovat. Téměř každý výrobce AKS používá odlišný databázový systém a zpravidla data ukládá do zcela proprietárních datových struktur. Ukládání dat je navíc často záměrně utajovaný údaj, který producenti AKS neradi zveřejňují. Vlastnosti vyhledávání nejčastěji ovlivňují indexy, jakožto struktury, kde jsou data formalizována, a datové typy. Chování druhé konverze můžeme ovlivnit pouze pomocí **doporučení na indexování jednotlivých polí**.

Třetí konverze vzniká již při vzniku dat v databázi. Zůstaneme-li u bibliografických dat, data vznikají v UNIMARC nebo MARC 21 a poté jsou ukládána do databáze. Relační struktury jsou pro uložení struktur MARC nevhodné, a proto většina producentů záznamy ukládá plnotextově a k nim vytváří pomocné vyhledávací tabulky a indexy. Způsob, jakým jsou data do pomocných tabulek segmentována, rozhoduje o kvalitě celého AKS a projevuje se samozřejmě nejvíce při vyhledávaní. Tato konverze je zcela věcí producenta AKS.

Kromě těchto tří momentů situaci ještě komplikují dvě skutečnosti: v ČR se dnes podporují 2 formáty MARC, a to UNIMARC a MARC 21, mezi nimiž se musí provádět konverze, a používá se několik znakových sad, nejčastěji CP1250 (Windows), UTF-8 (Unicode) a ISO-8859-2 (ISO Latin 2), mezi kterými se musí též provádět převod. Řešení pro konverze UNIMARC/MARC 21 i pro konverze znakových sad nabízí v nové podobě **funkce pro stahování záznamů JIB**, o které pojednává jiný příspěvek v tomto sborníku.

3. PROFIL Z39.50 JIB

Z předchozího příkladu vyplývá, že úloha profilů Z39.50 je tedy zejm. v definici využívaných atributů a mapování atributů na vnější struktury dat, v našem případě na UNIMARC a MARC 21.

Při provozu JIB se projevilo, že Profil CZ, který byl v návrhu, nebyl v jednotlivých úrovních ve většině zdrojů implementován a hlavně se pro účely jednoduchého vyhledávání a stahování záznamu jevil jako zbytečně náročný. Bylo třeba vytvořit profil, které by definoval pouze minimální požadavky na implementaci Z39.50, aby se na jedné straně dosáhlo standardizace dotazování a na druhé straně se tohoto stavu dosáhlo rychle díky technické nenáročnosti. Členové pracovní skupiny JIB proto na základě zkušeností s vyhledáváním ve více než 60 zdrojích navrhli profil určený pro české prostředí s názvem **Profil Z39.50 JIB**. Profil se snaží vytvořit takové prostředí, aby bylo možno zdroje, splňující níže uvedené požadavky, dotazovat stejným dotazem a obdržet srovnatelné výsledky.

České katalogy a databáze jsou značně různorodé, a proto nebylo snadné najít společný základ pro definici Z39.50 atributů. Byla proto zvolena cestu definice skutečně minimálních požadavků:

a) definice určuje nastavení pouze 7 základních polí, podle kterých se nejčastěji vyhledává
b) definice řeší pouze vyhledávání (SEARCH, PRESENT), nikoli funkce zápisu (UPDATE)

Toto minimum by mělo být podmínkou kooperace (v oblasti vyhledávání) s dalšími systémy na národní úrovni.

V únoru 2004 se v STK v Praze sešla k projednání návrhu pracovní skupina ZIG, která návrh přijala a přítomní producenti serverů Z39.50 působící na českém trhu přislíbili implementaci profilu během letošního roku.

Text Profilu Z39.50 JIB je zveřejněn na http://www.stk.cz/ZIG/ProfilJIB.rtf.