Z39.50 a (versus) XML

LETTERSOUP

Martin Svoboda
Státní technická knihovna, Praha

Lettersoup neboli písmenková polévka nás bude obtěžovat asi už napořád. Věci jsou složité a jejich pojmenování mnohaslovná, a tak se k nám dostávají jen v podobě zkratek, převážně anglických termínů. A jen zřídkakdy zkratka dává alespoň tušit, co se za ní skrývá. Před krátkým časem byly tajemné zkratky ISO 2709, UNIMARC, AACR2, jen si na ně knihovníci zvykli, už jsou tu další, o nic srozumitelnější. Potřeba otevřít souborný katalog CASLIN také jako zdroj pro sdílenou katalogizaci vyvolala v nově zřízené konferenci *lib-auto@publib.upol.cz* diskusi, v níž zkratky poletovaly sem i tam, až na výjimky velmi seriózně a odborně, bohužel příliš často pro větší část knihovnické obce asi dost nesrozumitelně. Diskutovalo se o několika souvisejících, avšak nesrovnatelných věcech: pokusil jsem se přispět svou troškou do mlýna se snahou trochu vykolíkovat terén a vyznačit v něm záchytné body. A nyní jsem byl vyzván, abych napsal úvod k článkům trojice autorů.

Dnes asi nikdo nepochybuje, že bez standardizace by fungování naší společnosti dost drhlo. Benediktinské klášterní knihovny v zamčených skříních, či v lepším případě s knihami přikovanými k policím, žádnou standardizaci nepotřebovaly. Bez pomoci knihovníka byl čtenář tak jako tak ztracen – a někdy i s ní, viz Jméno růže. Jak knihovny rostly a knihovníků tolik nepřibývalo, vznikaly nástroje umožňující nejprve knihovníkům a posléze i čtenářům samotným najít předmět svého zájmu. Vznikaly soupisy, rejstříky, věcná pořádání a začaly se objevovat skříně katalogů s mnoha šuplíky. Dokud byly katalogy „šity na míru“ knihovně a čtenář navštěvoval jen tu svoji, nehrály rozdíly v místních zvyklostech velkou roli. Vznik souborných katalogů a snad i ohled na nebohého čtenáře přiměl knihovníky vypracovat obecně platná pravidla a dodržovat je. Převedení katalogů do elektronické podoby a dostupnost mnoha různých katalogů v Síti nutnost všeobecně akceptovaných standardů jen podtrhují.

Předmětem standardizace v knihovnách jsou dokumenty, k nim příslušná metadata a procedury sloužící (v konečném cíli) k vyhledání a dopravení dokumentu čtenáři. Nebudeme se zabývat standardy fyzické reprezentace dokumentů, pouze jejich podoby elektronické. V každém případě je třeba si uvědomit, že standardizace každého a tím spíše složitého objektu může mít řadu vrstev, tak jako třeba vojenská blůza je šitá ze standardní látky, standardními nitěmi, zapíná se na standardní knoflíky a lze na ni připnout standardní výložky. Funkčnost standardu vyšší vrstvy může, ale nemusí, být ovlivněna změnou standardu vrstvy nižší: vyměním-li khaki knoflíky za zlaté stejné velikosti, bundu půjde stále zapnout, nahradím-li knoflíky sice khaki, ale jiné velikosti, funkčnost se pokazí. Při úvaze o záměně na jedné úrovni je tedy třeba dobře zvážit, zda změna neovlivní vrstvy vyšší.

Dokumenty

Dnes je převážná většina tištěných dokumentů připravována v elektronické podobě. Mezi formáty rozličných textových editorů vyniká jeden nesmírně mocný standard, či spíše nástroj pro tvorbu standardů. Tím je SGML (Standard Generalised Mark-up Language – standardní zobecněný značkovací jazyk) neboli ISO 98840, pocházející z konce 60. let, původně od firmy IBM a používaný velkými firmami a polygrafickými společnostmi. Jeho dosud nevyčerpaná síla spočívá v tom, že je naprosto otevřený, dovoluje vytvářet standardy pro popis dokumentů libovolného druhu a nejrůznější míry složitosti, jeho značky jsou běžně čitelné při zobrazení textu v tom nejtriviálnějším editoru, a co je nejdůležitější, jeho součástí je i jazyk pro popis těchto definic. Popis systému značek se jmenuje DTD (Document Type Description) a může být „potravou“ pro syntaktický analyzátor, který bez dalšího vstupu je pak schopen zkontrolovat formální správnost každého objektu vytvořeného podle tohoto DTD. Rozhodnutí použít SGML jako základ pro značkování elektronických hypertextových dokumentů nepochybně usnadnilo obrovský rozmach Sítě během několika let.

Při značkování dat (a nemusí jít jen o jazyky založené na SGML, znalci Knuthova T_EXu – užívaného v univerzitním prostředí pro sazbu zejména matematických, ale i běžných textů [Knuth, 1990] – o tom vědí své) je možno vyjít ze dvou principů: značkovací jazyk navrhnout jako významový – značky říkají, **co** to je mezi nimi (např. <autor1> </autor1>) nebo prezentační – značky říkají, **jak** se mají data uzavřená mezi nimi zobrazit (např.). HTML je kočkopes, některé značky (tagy) jsou významové <body>, <h1>, atd., jiné prezentační <background …>, <font …>, atd. V době pouze znakových prohlížečů (kdo si dnes vzpomene na Lynx?) tato koncepční nekonzistence tolik nevadila, s nástupem prohlížečů grafických vystoupila jasně. Pokusem o odstranění prezentačních slabin HTML jsou CSS, Cascading Style Sheets, to však byl jen půlkrok, byť správným směrem. Teprve CSS přenesené do prostředí XML – tady se jmenují XSL (eXtended Stylesheet Language) – jsou to pravé: XML se stará o markup významový a nezabývá se tím, **jak** se data zobrazí; XSL se stará o to, jak se zobrazí obsah příslušného XML tagu, aniž dbá o to, **co** je obsahem. Prohlížeče schopné korektně zobrazovat XML+XSL se začínají objevovat.

Metadata

Tohle kouzelné slůvko neoznačuje nic jiného, než to, co knihovníci dávno dobře znají, totiž data o datech, přeloženo do knihovnického jazyka „záznam dokumentu“.

Tak jako snem katalogizátora je, aby každý dokument v sobě obsahoval CIP (Cataloguing in Publication), možnosti kvalitně indexovat web – a tedy účinnosti Altavisty, Yahoo!, atd. – by nesmírně pomohla přítomnost kvalitních „katalogizačních údajů" – tedy metadat – v každém dokumentu na Síti. Nedostatek kvalitních významových tagů HTML se pokusil odstranit kompromis mezi knihovníky a internetisty, tj. Dublin Core, dublinské jádro (jeho symbolem je ohryzek a také tomu odpovídá, bibliografický záznam je ohlodán až na kost pouhých patnácti údajů). Bohužel podporu rozšířených vyhledávacích strojů Dublin Core zatím nezískal, uplatňuje se v kooperačních systémech.

Donedávna nejdiskutovanějším předmětem standardizace v knihovnách jsou bibliografické záznamy. Těm se nyní věnovat nebudeme, lze jen konstatovat, že SGML může konstruovat popis záznamu dokumentu stejně dobře jako popis dokumentu samotného. Z tohoto hlediska poskytuje XML elegantní, soudobou a kvalitnější náhradu více než 35 let starého konceptu MARC/ISO 2709[1)] – umožňuje obnošenou blůzu MARCu zapnout zlatými knoflíky. Ale nejen to, stejný princip lze použít i na jiné „kousky" dat, které je třeba při komunikaci mezi systémy vyměňovat (jak uvidíme dále). Drobnou nevýhodou je to, že MARC ve struktuře ISO2709 dovede dnes číst každý slušný knihovnický systém (a také Z39.50 klient), kdežto klientů schopných číst a prezentovat XML je zatím málo, ale to čas jistě rychle změní.

Protokoly

Diplomatický protokol kanonizuje pravidla, podle nichž probíhá komunikace mezi dvěma suverénními partnery tak, aby každá strana správně rozuměla tomu, co jí druhá strana říká. Podobně protokoly na Síti jsou pravidla, jak si spolu mohou dvě běžící aplikace vyměňovat data tak, aby jim obě strany přikládaly pokud možno naprosto identický význam. Protokoly mohou rozeznávat a pamatovat si stav, v němž se právě proces komunikace nachází, to znamená, že existuje jistými příkazy ohraničené období komunikace mezi dvěma aplikacemi, tzv. session, ve kterém platí parametry stanovené na začátku v jakémsi prologu; jednodušší protokoly berou každou výměnu jako nezávislou.

Takový je i dnes na Síti nejběžnější protokol http: vše, co komunikující aplikace potřebují vědět o tom, v jakém stavu se nacházely v minulém kroku, musí při každé výměně opakovat (obvykle prostřednictvím URL předávaného mezi klientem a serverem). Takovým protokolům se říká **bezstavové**, zde neexistuje žádná něčím vymezená session. Pravidla omezující chování klienta i serveru jsou jen velmi volná, a dovolují tak nesmírnou tvárnost. Také pravidla pro formát vyměňovaných dat jsou jen velmi volná a umožňují tak obrovskou flexibilitu, ovšem za cenu toho, že dobře rozumí jen ten, komu je odpověď serveru obvykle určena, tj. člověk (a to ještě ne vždycky).

Naproti tomu existuje řada, tzv. **stavových** protokolů (ftp, telnet, Z39.50), které se vyznačují zahajovacím a zakončovacím dialogem, mezi nimiž probíhají vlastní užitečné interakce. Protokol Z39.50 slouží vyhledání, získání a v poslední době opět i zpětnému předání údajů. Dnešní protokol Z39.50 vznikl ze dvou zdrojů: z ideálního evropského, původně SRU (Search, Retrieve, Update – tedy najdi, podej, oprav) protokolu ISO 10162/163 navrženého na základě OSI (Open Systems Interconnection) modelu a z americké pragmatické původně jednodušší verze Z39.50, omezené pouze na SR (Search, Retrieve). Protokol Z39.50 definuje řadu **služeb** (např. Init, Search, Scan, Sort, Present, podrobněji viz např. [Holm, 1998]), tj. interakcí mezi klientem a serverem (v Z39.50 nazývanými target a origin), včetně jejich možných chování, požadavků a odpovědí. S některými službami, které definují, **co** se má **dělat** (vyhledat záznamy odpovídající dotazu, ukázat rejstřík, setřídit výsledky, zobrazit je, atd.), jsou spojená **data**, **s nimiž** se to má dělat (dotaz na **co**, prohlížet rejstřík **od**, atd. až po záznamy k zobrazení, downloadu či uploadu). Data mohou být v podobě záznamu v některé z protokolem uznávaných syntaxí (ISO2709, SUTRS[2)] a možná i jiné, brzy zřejmě XML) a v některém uznávaném formátu (UNIMARC, MARC21 a řada jiných MARCů)[3)]. Podobně dotaz může být formulován v různých syntaxích (RPN – reverzní polská notace, ISO 8777 – Commands for Interactive Text Searching, dříve označovaný jako CCL – Common Command Language a jiné). Tato variabilita je na jedné straně příjemná, na druhé straně je na překážku snadné domluvě. Proto vznikly specifikace, vymezující jakých syntaxí a formátů je dovoleno používat; těm se říká **profil**.

Dnes má Z39.50 za sebou řadu let vývoje, v jehož průběhu získal značnou sílu a obecnost, tím bohužel také značnou složitost. Přesto, nebo právě proto, je to dnes stále se šířící princip budování informačních systémů, zahrnujících knihovnické zdroje (katalogy, abstraktové/indexové databáze i primární data, ba dokonce i systémy elektronického dodávání dat) [Lynch, 1999]. Pokud chceme, aby systémy spolupracovaly v heterogenním prostředí, máme v zásadě dvě možnosti: buď každý systém „rozumí" všem ostatním, což je nepraktické a při rostoucím počtu různých systémů nezvládnutelné (byť i toto řešení v Česku existuje: MVS ing. Písečného), anebo každý systém umí komunikovat v dohodnutém protokolu a s dohodnutou strukturou dat. Od zcela primitivního ATS (Author, Title, Subject) profilu se Z39.50 dopracoval k tzv. Bath profile, který je na nejlepší cestě stát se NISO standardem (http://www.ukoln.ac.uk/interop-focus/bath/). Z toho, jak různorodé firmy se vývojem aplikací Z39.50 zabývají i na základě sledování konference vývojářů Z39.50 soudím, že zánik protokolu Z39.50 nehrozí. Říkat, že Z39.50 bude vytlačen XML, je matení různých věci dohromady. XML možná nahradí ISO2709 pro zápis struktury předávaného bibliografického záznamu, možná i pro jiná data, možná bude použit i pro zápis řídicích dat protokolu. To však stále z XML nedělá protokol. K tomu XML není určen, a tak protokol samotný nahradit nemůže, sám protokolem není.

Shrnutí

Napsal jsem toho hodně, nevím, zda to přispěje k lepší orientaci, na závěr zkusím shrnout: věnovat se rozvoji XML je jistě dobré, myslím, že se lze vcelku spolehnout

na to, že hlavní díl práce na vývoji klientů schopných XML+XSL hezky interpretovat za nás odvedou vývojáři webovských nástrojů – potřebují je. Naproti tomu soudím, že máme-li do budoucna zajistit možnosti kvalitního hledání v databázích jemně strukturovaných dat, nezbývá nám, než se věnovat Z39.50, tam nelze příliš očekávat podporu ze strany komerčních uživatelů Sítě, kteří dnes pohánějí její rozvoj. Nepochybně je rozumné vývoj ostře sledovat a jakmile se objeví standard, který bude mít naději na dlouhodobou podporu „zábavního průmyslu", vážně se jím zabývat.

Poznámky:

1) Skutečně, formát ISO 2709 dává jasně najevo, že pochází z prediluviálních dob IBM 1410, kdy byty ještě neexistovaly a každý řetězec znaků musel být ukončen speciálním separátorem. Na druhé straně jeho výhradně znaková podoba mu umožnila přežít tak vážnou konkurenci, jakou byl v roce 1974 formát zavedený ACS pro službu Chemical Abstracts. ACS tehdy produkovala víc bibliografických záznamů než všichni ostatní producenti dohromady, její formát měl všechny číselné informace v binární podobě a byl tak daleko prostorově výhodnější – tehdy na tom ještě záleželo. Přesto, nezávislost normy ISO 2709 na použitém počítačovém systému jí umožnila přežít dodnes.
2) SUTRS – Simple Unstructured Text Record Syntax je definována v Z39.50.
3) Zcela pomíjím běžně zanedbávaný fakt, že ani týž formát nezaručí, že ve stejných „krabičkách" (označkovaných tagy některého MARCu) nebudou jiná než očekávaná data, to mohou zaručit jen tatáž a navíc stejně interpretovaná katalogizační pravidla.

Literatura:

HOLM, Liv A. The ONE project. In: HABERMANN, Heinz. *ELAG 20th Seminar : Quality of Electronic Services, Staatsbibliothek zu Berlin – Preussischer Kulturbesitz.* (DBI) 1998. Dostupný z: < http://www.kbr.be/elag/20seminar/papers/one.htm >

KNUTH, Donald E. *The TeX book*. Addison-Wesley : Reading, 1990.

LYNCH, Clifford. Souborné katalogy. In: *Souborné katalogy: organizace a služby.* Praha : Národní knihovna ČR, 1999. Dostupný z : <http://www.caslin.cz:7777/caslin99/c3.html>

■ ■ ■ ■ ■