XML a vědecké informace

Miloslav NIČ
Laboratoř informatiky a chemie, VŠCHT Praha
Miloslav.Nic@vscht.cz

INFORUM 2005: 11. konference o profesionálních informačních zdrojích
Praha, 24. - 26.5. 2005

Abstract. *XML technologie zásadním způsobem změnily svět počítačů a v současné době pomáhají přetvářet standardní publikační postupy. To je však pouze začátek a v blízké budoucnosti bude XML hrát klíčovou úlohu v informační infrastruktuře vědy a dalších odvětvích lidské činnosti. Přednáška prezentuje několik příkladů uplatnění XML při zpracování a vyhodnocení vědeckých informací a identifikuje slibné oblasti s potenciálem pro další rozvoj.*

Úvod

Znalost XML (eXtensible Markup Language) je běžně vyžadována od uchazečů o zaměstnání v softwarových firmách a o XML se zajímá i řada laiků. Přesto je historie tohoto jazyka a počítačových technologií, které z něj vycházejí, velmi krátká. Tento jazyk byl navržen v průběhu 90. let minulého století a první oficiální standard byl odsouhlasen konsorciem W3C[1], jedním z hlavních tvůrců standardů Internetu, 10. února 1998[2]. XML je nyní podporováno mnohými softwarovými produkty a řada výrobců používá tuto podporu i při marketingovém prosazování výrobku, ačkoli skutečný význam této podpory dokáže ocenit jen málokterý uživatel.

Je logické, že XML našlo svoji cestu i do oblasti vědy a techniky. V tomto případě lze dokonce hovořit o tom, že věda nenásleduje cestu vytyčenou jinými, ale naopak problémy vědy a vědeckých publikací byly jedním z impulzů, které k rozvoji jazyka XML vedly.

Účelem tohoto článku je zhodnocení významu XML pro vědeckou komunikaci a identifikace předností XML-technologí, které může ocenit i ten, kdo není programátorem. Článek operuje s některými pojmy, pro jejichž pochopení je výhodná základní znalost jazyka HTML a dalších publikačních technologií, tato znalost však není nezbytnou podmínkou pro pochopení základních myšlenek.

Základy XML

Na Internetu[3-5] i v knižní podobě[6] je nyní dostupné nepřeberné množství úvodů do XML, a proto je zbytečné přidávat další, nicméně považuji za důležité zdůraznit některé hlavní charakteristiky, které učinily z XML tak všestranně uživaný nástroj.

Znalost základů XML je velmi užitečná pro vyspělejší uživatele počítačů, a to nejen pro jejich programátory, protože s vynaložením minimální námahy umožňuje výrazně urychlit spolupráci mezi počítačem a jeho uživatelem. Praktickým základům jazyka je možné se naučit během několika minut, i když samozřejmě i v něm existují záludnosti, které dokáží na dlouhou dobu zaměstnat zkušeného experta.

Jazyk XML je vlastně obyčejný text, který je možné rozšiřovat o doplňkové informace týkající se celého nebo částí dokumentu. Tak:

<tvrzení>Znalost XML je velmi důležitá pro informační pracovníky**</tvrzení>**

představuje korektní příklad užití jazyka XML. Vlastní text je zapsán stejně, jako v každém jiném dokumentu, jediný rozdíl spočívá v informaci o tom, že se jedná o "tvrzení". Co si představit pod konceptem "tvrzení" již XML neřeší, záleží na čtenáři nebo programátorovi, aby si vytvořil představu (nebo napsal program).

Účelem XML je stanovit jednoznačná formální pravidla tak, aby daný text šlo co nejlépe a nejjednodušeji rozdělit na logické útvary a poté zkontrolovat, že všechna formální pravidla byla dodržována. Pravidla XML jsou velmi jednoduchá, vše umístěné v závorkách "<..>" je instrukce, vše ostatní vlastní text (Zvídavý čtenář se může pozastavit nad problémem, jak zapsat tyto závorky v samotném textu, například při vyjádření faktu že jedna je méně než dva. Tento problém XML samozřejmě jednoduše řeší, a toto řešení je popsáno v každém začátečnickém tutoriálu.)

Kontrola dodržování pravidel je velmi silnou stránkou XML. S vynaložením minimální námahy je možné provádět řadu kontrol, které jsou v jiných způsobech zápisu velmi problematické. Pokud je však kontrola obtížná, tak se v praxi často neprovádí vůbec.

Základním prvkem kontroly je splnění podmínky tzv. "well-formedness". Na této úrovni kontroly je ověřeno, zda dokument splňuje všechny formální podmínky, které jsou od XML dokumentu očekávány. Pokud program, určený pro nahrávání XML-dokumentu do paměti počítače (tzv. parser), zjistí jakýkoliv prohřešek proti pravidlům, je povinen ukončit svoji normální činnost a ohlásit neopravitelnou chybu. Tento drakonický požadavek působí občas problémy začínajícím autorům

XML, kteří byli dosud zvyklí tvořit HTML dokumenty pro dnešní generace internetových prohlížečů, které nějakým způsobem zobrazily téměř cokoli. Zdůrazňuji výraz "nějakým způsobem", u mnoha chybných HTML dokumentů není možné jednoznačně určit, co vlastně autor od počítače očekával, a tedy záleží na každé aplikaci, jak si s touto nejednoznačností poradí.

XML takovouto možnost "odhadování významu" na této úrovni nepřipouští. Pokud autor dokumentu nesplní všechny od něj očekávané podmínky, tak dokument bude odmítnut. Autor tím získává jistotu, že nepřehlédl nějakou maličost, která může změnit význam informací, a programátor aplikace, která bude dokument zpracovávat, je zbaven nutnosti kontrolovat, zda zdrojový text neobsahuje triviální chyby (tato kontrola a automatiké opravování chyb v prohlížečích Internetu představuje většinu programového kódu a je zdrojem řady problémů).

Samozřejmě i dokument, který projde prvním krokem této kontroly "well-formedness", ještě může být zcela nesmyslný. Představme si např. dokument, ve kterém uchováváme adresy svých známých. Pokud v takovémto textu zapomeneme uvést jméno a příjmení našeho známého, tak celý záznam ztrácí svoji hodnotu.

V rámci XML však máme možnost zvýšit požadavky a zkontrolavat, zda je dokument "valid", tedy, zda vedle formálních pravidel syntaxe rovněž obsahuje další očekávané informace a zda jsou tyto informace uvedeny na očekávaném místě.

Způsobů validace je celá řada, k běžně používaným nástrojům z této oblasti patří DTD[7], XML Schema[8], Relax[9] či Schematron[10] validace. S jejich pomocí je možno vyjádřit, v jakém pořadí jsou data podávána (např. nejdříve jméno, poté příjmení), ale rovněž i kontrolovat možnosti typických překlepů (číslice v příjmení). Pokud množina hodnot je přesně dána (např. člověk je muž či žena), lze tímto způsobem zajistit pouze tyto hodnoty, některé nástroje dokonce umožní kontrolu, zda koncovka "ová" není souběžně uvedena s údajem "muž".

Výhodou těchto technologií je, že řadu kontrol může vytvářet i člověk, který neumí programovat, a také programátor může s jejich pomocí pracovat mnohem efektivněji než s použitím jiných programovacích prostředků. V dnešní době již existuje software, který umožnňuje vytvářet kontroly s pomocí grafických nástrojů, při jejichž používání ani uživatel nemusí tušit, že pracuje s XML a jeho kontrolními standardy. Jelikož však výstupy těchto nástrojů jsou standardizovány, je možné kombinovat programy různých výrobců, což v jiných případech bývá značně problematické.

Tento fakt představuje velmi důležitou hodnotu, kterou XML přináší. Jelikož řada XML-technologií je standardizována a má obecné využití, zvyšuje se šance, že různé softwarové produkty budou schopny mezi sebou komunikovat. Nežijeme v ideálním světě, a tedy toto očekávání tzv. interoperability není plně uspokojeno, nicméně přechod k technologiím XML výrazně vylepšil komunikaci mezi rozdílnými softwarovými komponenty.

XML se rozšířilo do mnoha sfér lidské činnosti a od programátorů je běžně očekávána znalost potřebných technik z této oblasti. Díky tomu můžete mnohem efektivněji využívat velmi drahý čas programátorů. Zatímco dříve se s každým novým projektem musel programátor seznamovat s novými způsoby zápisu informací vhodných pro počítač, díky standardizaci na úrovni XMLtato seznamovací doba odpadá nebo je alespoň výrazně zkrácena.

Není sporu, že technologie XML představují významný krok v rozvoji informačních technologií a nikoliv slepou uličku. Znamená to, že se již není třeba bát investic (finančních i časových) do jejich zvládnutí a do vytváření XML- aplikací. XML tak nachází cestu do velkého množství oborů a věda patří k oblastem, kde nasazení XML přináší bohaté výsledky.

Matematika a MathML

Matematika je základní vědeckou disciplinou a možnost sdělování matematických údajů bezchybným a kontrolovatelným způsobem je velmi důležitá.

Není tedy divu, že jednou z vůbec prvních aplikací XML bylo vytvoření jazyka MathML (Mathematical Markup Language)[11], který umožňuje zápis matematických vztahů ve formátu XML. V matematice je nutná jednoznačnost, běžně používané grafické zápisy jsou však srozumitelné pouze člověku, neboť ten si dokáže řadu nejednoznačností vysvětlit díky znalosti konvencí a porozumění obsahu sdělení. Naprogramovat počítač tak, aby byl schopen vyložit všechny záludnosti notace je však velmi obtížné.

MathML je proto rozděleno na dvě části. Prezentační část umožňuje zobrazit matematické výrazy tak, jak je zvykem při publikaci matematických děl. Pokud ovšem potřebujete zaručit, že zápis je jednoznačný pro počítač, můžete použít obsahovou notaci, která přesně vystihuje matematický smysl.

Samotný zápis může být poměrně složitý, to však nemusí běžné uživatele zajímat, neboť při tvorbě vzorců je samozřejmostí využívat specializované editory, které mohou být integrovány do běžných nástrojů pro tvorbu dokumentů. Dá se očekávat, že MathML se v blízké budoucnosti stane univerzální matematickou notací, přenositelnou mezi různým software. Znalci mezi čtenáři vědí, že ještě existují oblasti, v nichž je výrazová síla MathML nedostačující, ale pro běžné použití je již naprostá většina problémů vyřešena.

Obrázky, grafy a SVG

Vědecké práce by často byly bez ilustrací zcela nepochopitelné. Počítače si musí informace o grafických objektech předávat v mnoha různých situacích, a každý, kdo se zpracováním obrazů v textu někdy zabýval dobře ví, jak složité je vyznat se v záplavě různých formátů. Volba správného formátu je přitom velmi důležitá, protože chyba ve volbě může mít fatální následky v následujících krocích.

Pro vytváření grafických objektů je velmi často výhodné používat vektorovou grafiku. To znamná, že obrázek je rozložen na velké množství přímek, kružnic a dalších křivek. Toto rozložení jednak výrazně omezuje nároky na paměť počítače a navíc umožňuje mnohem snazší úpravy v obrázku, a to manuální i softwarové.

Pro zápis vektorové grafiky nabízí XML řadu možností a není tedy divu, že SVG (Scalable Vector Graphic - XML zápis pro vektorovou grafiku)[12] se stalo jedním z nejrozšířenějších aplikací XML. Zatímco SVG je využíváno v mnoha oblastech, jeho nasazení v technických a vědeckých oborech je obzvláště významné. Pokud si uvědomíme, že mezi typické grafické objekty, které jsou vhodné pro vektorové zpracování patří mapy, záznamy spekter nebo grafy závislostí, je jejich použití zřejmé. Obrázky založené na vektorech lze rovněž mnohem snáze programaticky vyhodnocovat, tedy např. třídit nebo identifikovat klíčové prvky.

XML a Chemie

V předchozích odstavcích jsem psal o grafice. Typickou vědou, založenou na obrazových informacích, je chemie. Struktury organických sloučenin lze textem vyjádřit jen velmi obtížně, a proto jsou chemici zvyklí komunikovat kresbou.

Pro záznam chemických struktur je samozřejmě možné použít SVG, jakožto obecného grafického formátu. Obecnost SVG však v sobě skrývá nevýhody. Symboly v chemických strukturách mají svůj přesně daný význam (jména atomů, řády vazeb, náboje atd.) a tato informace se v SVG ztrácí. Pokusy programaticky získat tuto informaci z SVG-zápisu jsou většinou velmi náročné a často neúspěšné.

Není tedy divu, že chemičtí informatici hledají způsoby zápisu, které by s chemickými vlastnostmi počítaly. Prvním dotaženým pokusem o zápis chemie ve formátu XML je CML (Chemical Markup Language)[13] profesorů Rusta a Rzepy. Tento jazyk sehrál velmi významnou úlohu při rozvoji chemické notace i samotného XML, trpí však některými problémy, které mu zabraňují splnit veškeré nároky kladené na chemickou notaci, a proto výzkum chemické notace je oblast, která je stále otevřená a v níž působí i naše skupina.

Velmi zajímavá je možnost kombinace grafiky SVG s chemickým zápisem dat, která řeší mnoho problémů spojených s prezentací a automatickým vyhodnocením chemických dat. Vývoj takového přístupu patří k hlavním výzkumným cílům naší skupiny, Dr. Košata je autorem molekulového editoru BKChem[14], který takovéto spojení umožňuje. Příkladem nasazení této technologie je zpracování XML-verze knihy "IUPAC Compendium of Chemical Terminology - Gold Book"[15].

XML a bibliografické formáty

Software pro správu bibliografických citací se stává nezbytnou pomůckou pro vědeckou práci. V této oblasti jsou nejvíce používané programy Reference Manager[16], ProCite[17] a EndNote[18], všechny v současnosti patřící společnosti Thomson[19].

Všechny tyto programy podporují jako jednu z možností export a import v XML formátu. Mohlo by se zdát, že koncového uživatele nemusí zajímat, v jakém formátu jsou data sdílena. Z krátkodobého pohledu je to zřejmě pravda, s dlouhodobější perspektivy však tomu tak není. Uzamknutí dat v binárním nebo komplikovaném textovém formátu může vytvořit významnou bariéru při přechodu na jiné produkty. Spolehlivý převod dat z formátů XML je mnohem snazší a spolehlivější a pokud se tedy v budoucnosti objeví významný konkurent narůstající hegemonii společnosti Thomson, bude možné zajistit spolehlivý převod do konkurenčního nástroje.

Ale i v případě, že uživatel se rozhodne setrvat u svého současného nástroje, dostupnost zdrojů XML pro něj znamená velkou výhodu. Pokud se uživatel rozhodne zpracovat svá data způsobem, který program nenabízí, nebude mít problém nalézt kvalifikovaného pracovníka, který mu zpracování provede.

XML a databáze

Jedním z hlavních cílů XML při svém vzniku bylo poskytnou prostředky pro sdílení dat mezi software. Je tedy přirozené, že v této oblasti nalezlo XML široké uplatnění ve vědecké oblasti.

Za příklad je možné uvést databázi molekulových interakcí IntAct[20], která nabízí data o jednotlivých interakcích ve formátu XML. Tato databáze je jedním z příkladů nasazení XML-technologí při studiu bílkovin a jejich funkcí, ve kterém je třeba spolehlivě zaznamenat velká množství dat a výběr formátů je zde velmi důležitý. Obdobné nasazení je rovněž možné zaznamenat u genomických databází[21].

Používání formátů XML přináší velkou výhodu v tom, že je možné těžit z vývoje v jiných než vědeckých oblastech. Na vývoji XML-databází jsou zainteresovány velké světové softwarové firmy, neboť finančně silné skupiny mají eminentní zájem na rychlém prohledávání komplexních dat. Revoluční změny v této oblasti lze očekávat od nástupu technologií založených na specializovaném jazyku XML pro prohledávání XML- databází - XQuery[22].

Závěr

Je nepochybné, že s XML a s ním spojenými technologiemi se budeme setkávat stále častěji. Počet dokumentů a dat dostupných ve formátu XML stále stoupá, hledání v Google navrací přes 7 milionů dokumentů s koncovkou ".xml"[23] a další milion s koncovkou ".xhtml"[24]. Informace zpracované ve formátu XML výrazně usnadňují automatické zpracování obsahu informace a tím otevírají cestu k webu nové generace, občas označované jako Semantic Web[25].

Literatura:

1. http://www.w3.org/
2. http://www.w3.org/TR/1998/REC-xml-19980210
3. http://www.zvon.org
4. http://www.w3schools.com
5. http://www.topxml.com/
6. http://www.amazon.com
7. http://www.w3.org/TR/REC-xml/
8. http://www.w3.org/TR/xmlschema-0/
9. http://relaxng.org/
10. http://xml.ascc.net/resource/schematron/schematron.html
11. http://www.w3.org/Math/
12. http://www.w3.org/TR/SVG/
13. http://www.xml-cml.org
14. http://bkchem.zirael.org
15. http://gold.zvon.org/
16. http://www.refman.com/
17. http://www.procite.com/
18. http://www.endnote.com/
19. http://www.isiresearchsoft.com
20. http://www.ebi.ac.uk/intact/index.jsp
21. http://snp.ims.u-tokyo.ac.jp/XML.html
22. http://www.w3.org/TR/xquery/
23. http://www.google.com/search?as_qdr=all&q=+xml+filetype%3Axml&btnG=Search
24. http://www.google.com/search?as_qdr=all&q=+xhtml+filetype%3Axhtml&btnG=Search
25. http://www.w3.org/2001/sw/