Ukládání geodat do XML nativních databází

Albrechtová Zdeňka
Geomatika
Západočeská univerzita v Plzni
Univerzitní 22
306 14 Plzeň
E-mail: ZAlbrechtova@t-email.cz

Abstract

The basic idea of my work is possibility testing to use native XML databases for saving geographical data. After the brief acquaintance with the basics of XML (XML, Xpath, Xquery, XSLT, ...), native XML databases (kinds, basic charackteristics, ...) and XML formats (GML, cGML, ...) of geodata, it can be directly seen in six geodata files of three chosen database systems (4Suite, Berkeley DB XML, eXist) whether it is possible to use this kind of databases in geoinformatics for solutions of concrete projects.

Abstrakt

Základní myšlenkou této práce je otestování možností využití nativních XML databází pro ukládání geografických dat. Po stručném seznámení se základy XML (XML, XPath, XQuery, XSLT, …), s nativními XML databázemi (druhy, základní charakteristiky, …) a s XML formáty geodat (GML, cGML, …) je na vybraných třech databázových systémech (4Suite, Berkeley DB XML, eXist) za pomoci šesti souborů s geodaty přímo ukázáno, zda by bylo v geoinformatice možné tento druh databází využít k řešení konkrétních úkolů.

Úvod

V současné době se v oblasti geověd stále více uplatňují různé formáty značkovacího jazyka XML (např. GML, cGML, G-XML), a to především z důvodu jejich nenáročnosti na software. S rostoucím využitím těchto formátů se dá předpokládat, že data uložená tímto způsobem musí být někde uskladněna a spravována.

Jako i v jiných oblastech i zde se nabízí využití databázových systémů. Pro ukládání a spravování XML dat existuje speciální druh databází. Jedná se o nativní XML databáze, jejichž hlavní předností je přímé (nativní) ukládání XML dokumentů.

Mimo vlastní existenci těchto databází přispěl k jejich využití při testování pro ukládání geodat zejména rozvoj dotazovacích jazyků v oblasti XML technologií (např. zdokonalení jazyka XPath ve standardizované verzi 2.0, vznik jazyka XQuery). Rozvoj dotazovacích jazyků, zejména pak vznik jazyk XQuery, je pro geodata obzvláště důležitý, protože při dotazování nad těmito druhy dat je nutné vytvářet složitější dotazy, což právě jazyk XQuery umožňuje. Pomocí tohoto dotazovacího jazyka je tedy možné získávat velmi podobné výsledky jako při dotazování v běžných GIS aplikacích.

Základním úkolem této práce je na vybraných nekomerčních nativních XML databázových systémech a pokud možno reálných geodatech ukázat, zda by bylo možné těchto databází v praxi využít.

Postup realizace projektu

- výběr zástupců nekomerčních nativních XML databází
- získání geodat ve formátu XML
- testování vybraných databází při práci s geodaty

Výběr zástupců nekomerčních nativních XML databází

Při řešení problematiky ukládání geodat do XML nativních databází bylo nutné v úplném počátku vybrat zástupce nekomerčních XML nativních databází, poněvadž nekomerčních databázových systémů je poměrně velký počet (viz tabulka 1) a zaobírat se každou databází by pro názornost nemělo velký smysl.

Název databáze	Typ databáze	Vývojář/Tým vývojářů
4Suite	objektově orientovaná	FourThought
Berkeley DB XML	klíč - hodnota	Sleepycat Software
DBDOM	relační	K. Ari Krupnikov
dbXML	vlastní model	dbXML Group
eXist	vlastní model	Wolfgang Meier
myXMLDB	MySQL	Mladen Adamovic
Ozone	objektově orientovaná	ozone-db.org
Sedna XML DBMS	vlastní model	ISP RAS MODIS
Timber	Shore, Berkeley DB	University of Michigan
XDBM	vlastní model	Matthew Perry, Paul Sokolovsky
Xindice	vlastní model	Apache Software Foundation
XpSQL	relační	Makato Yui

Tabulka 1: Nekomerční nativní XML databáze podle [1]

Protože se jednotlivé databázové systémy liší především typem, byla tato charakteristika zvolena jako zásadní, a tak na základě typu došlo k výběru kandidátů. V dalším kroku byl brán zřetel na rozšířenost a známost každého jednotlivého databázového systému, přičemž se samozřejmě upřednostnily známější a rozšířenější aplikace. V konečném výsledku byly vybrány čtyři databázové systémy, ze kterých ale pouze tři mohly být použity:

- 4Suite jako zástupce objektově-orientovaných databází,
- Berkeley DB XML jako zástupce databáze typu klíč-hodnota,
- eXist jako zástupce databází s vlastním modelem, jež v rámci nekomerčních XML nativních databází tvoří nejobsáhlejší skupinu.

Čtvrtým vybraným databázovým systémem, a tedy nakonec nepoužitým, byl DBDOM jako zástupce relačních databází. Při práci s tímto databázovým systémem se vyskytly neodstranitelné problémy, pro které ho nebylo možné využít.

DBDOM samozřejmě není jediným relačním databázovým systémem, dalším, a zároveň také posledním, možným kandidátem této skupiny je systém XpSQL. Nad operačním systémem Windows se však tato aplikace nepodařila nainstalovat, a tak nakonec je skupina relačních nekomerčních XML nativních databází bohužel nezastoupena.

Získání geodat ve formátu XML

Volba vhodného vzorku geodat pro uložení do již vybraných databázových systémů také nebyla triviální záležitostí. Pro objektivní posouzení možnosti využití XML nativních databází v oblasti geografických věd bylo vhodné vybrat reálná data.

Západočeská univerzita spolupracovala při tvorbě Atlasu mezinárodních vztahů, který vznikal především s využitím XML technologií. XML technologie byly uplatněny i pro zdrojová data, pro která byl zvolen plnohodnotný XML formát JML[1] (JUMP GML).

Protože zdrojová data Atlasu mezinárodních vztahů byla přímo v XML formátu a jednalo se o reálná data, bylo několik z těchto zdrojových souborů vybráno pro využití v nativních databázích. Šest získaných souborů se nemusí jevit jako dostatečný počet, ale jejich odlišné, a ve třech případech velmi velké, velikosti poskytnou podle mého názoru dobrý nástin využití nativních XML databází v oblasti geografických věd.

Název souboru	Velikost souboru [B]
body_miny1.jml	5 974
body_miny2.jml	27 182
magda.jml	14 720 878
miny.jml	14 612 497
zbrojeni.jml	14 625 758
zbrojeni_popisky.jml	6 673

Tabulka 2: Velikosti použitých souborů s geodaty

Testování vybraných databází při práci s geodaty

Obecné poznatky a závěry o jednotlivých databázích

1. **4Suite**

Databáze 4Suite byla první z aplikovaných databázových systémů. Její základní nedostatek byl hned z počátku zřejmý – databáze 4Suite nepodporuje dotazovací jazyk XQuery ani žádný jiný obdobně vyvinutý dotazovací jazyk, proto nemůže při dotazování poskytovat ani trochu podobné výsledky jako při dotazování v klasických GIS aplikacích.

Podporovaným dotazovacím jazykem je jazyk XPath, který by ale mohl být při jednoduchých úkonech dostačující a ve spojení s 4Suite Serverem by mohl získávat pohodlné výsledky.

[1] Tento formát má základ ve formátu GML (Geography Markup Language) a je využíván především v rámci Open Source programů JUMP a openJUMP. Soubory formátu JML mají v sobě již zabudovanou strukturu dokumentu, tedy jakési schéma v úvodu vlastního dokumentu, narozdíl od klasických souborů ve formátu GML.

4Suite Server sice umožňoval vytvoření kolekce dokumentů a z části i správu uložených dokumentů, ale bohužel nebylo možné této nadstavby využít při vlastní aplikaci dotazu, a tak mohly být dotazy aplikovány pouze na dokumenty uložené v běžných souborových úložištích.

Pro práci umožňoval 4Suite společně s 4Suite Serverem používat tři pracovní prostředí: klasickou příkazovou řádku, pracovní prostředí programovacího jazyka Python a GUI serveru. Protože podle mého názoru nemusí být uživatel ani správce databáze zdárným programátorem, testovala jsem 4Suite pouze v prostředí příkazové řádky a GUI. Příkazová řádka byla pro práci, i přes chudou nápovědu, příjemným pracovním prostředím bez zjevných nedostatků. V GUI se vyskytovaly nesčetné problémy, jako jeden z nejnápadnějších bych uvedla nemožnost mazání uložených dokumentů (tuto operaci bylo možné provést pouze z příkazové řádky), proto nebylo možné toto prostředí příliš využívat.

Obrovskou předností tohoto systému je podpora XSLT a XUpdate. Právě díky této velké výhodě bych 4Suite pro ukládání geodat nezavrhovala, a pokud by bylo možné načítat soubory i z kolekcí dokumentů na serveru (tuto skutečnost znemožňovala serverová chyba), bych spíše věřila, že by měl tento systém velkou budoucnost alespoň v oblastech transformací XML souborů s geodaty.

2. Berkeley DB XML

Dalším vybraným zástupcem z nativních XML databázových systémů byl Berkeley DB XML. Tato databáze je velice dobře uzpůsobená práci s různými XML i ne XML soubory uloženými v kolekcí dokumentů, což by mohlo být velmi vhodné například ve spojení geodat (ve formátu XML) s rastry.

Protože z dotazovacích jazyků podporuje Berkeley jak XPath, tak XQuery, může teoreticky v dotazování, právě díky jazyku XQuery, podat také velmi dobré výsledky.

Z ostatních XML technologií, které se týkají nějakým způsobem přímo dokumentů, nepodporuje Berkeley DB XML bohužel žádný, tedy ani XSLT ani XUpdate. Aktualizace je zajištěna vestavěnými funkcemi a pro transformaci dat do jiné podoby musí každý uživatel sáhnout do jiného systému. Pro geografická data je nemožnost využití XSLT dost nevýhodná, ale pro celkové využití Berkeley DB XML pro ukládání těchto dat zase není prioritou, tudíž i přes tento nedostatek by mohla být databáze Berkeley DB XML pro ukládání geodat snadno, a myslím si, že vcelku úspěšně, využita.

Pracovní prostředí by se mohlo mnohým uživatelům jevit jako velice strohé, však se také jedná pouze o prostředí příkazové řádky, ale z vlastní zkušenosti mohu s klidným svědomím tvrdit, že GUI je zde zcela zbytečné, protože díky kvalitní nápovědě a jednoduchým věcným příkazům si každý uživatel brzy osvojí základní pracovní postupy a zdánlivá počáteční nepřehlednost zcela zmizí.

3. eXist

Poslední využitou nativní XML databází byla databáze eXist. Tento systém je jistě pro uživatele na první pohled z hlediska pracovního prostředí nejpříjemnější. Dokumenty uložené v kolekcích jsou dobře viditelné, editovatelné, a způsob jejich načítání či odstraňování se ničím neliší od způsobu v běžných programových vybaveních. Nedostatečná nápověda systému je nahrazena obrovským zázemím na webových stránkách, kde se dá zjistit téměř cokoli.

I pro uživatele pracujícího s geodaty by mohl být tento systém velice atraktivní, a to nejen díky příjemnému pracovnímu prostředí. Jsou zde opět podporovány dva dotazovací jazyky, XQuery a XPath, přičemž XQuery má v rámci eXistu další podstatná rozšíření umožňující například aktualizace či transformace dokumentů. Jazyk XSLT je však tímto způsobem zcela nahrazen a tedy bohužel dále již není nijak podporován.

Dotazování

Dotazování nad geodaty v prostředí XML nativních databází je zřejmě jediná oblast, ve které se geodata od ostatních dat podstatně odlišují. Vzhledem k existenci prostorové složky roste i náročnost dotazů, obzvláště prostorových, které se samozřejmě nad běžnými daty nevytvářejí. Proto bylo nutné vytvořit několik jednoduchých i poněkud složitějších dotazů, jak atributových, tak prostorových, aby bylo možné zhodnotit vhodnost využití dané databáze pro ukládání geodat.

Protože většina XML nativních databází podporuje dva základní dotazovací jazyky, XPath a XQuery, vznikaly dotazy právě v těchto formátech, a to pomocí softwaru Altova XML Spy® 2007 Enterprise Edition.

Prvotně byly zformulovány dotazy v jazyce XQuery. Konečný počet těchto dotazů je dvanáct – osm atributových dotazů [(XQ_1) ... (XQ_8)] a čtyři dotazy prostorové [(XQ_9) ... (XQ_12)]. Z některých jednodušších dotazů pak vycházely obdobné dotazy v jazyce XPath, konečný počet XPath dotazů je pět – tři atributové [(XP_1) ... (XP_3)] a dva prostorové [(XP_4) ... (XP_5)].

Při vytváření dotazů jsem získala přibližnou představu o tom, jak dlouho by mohl průběh každého dotazu trvat. Nemohla jsem tedy očekávat, že výsledky dotazů (XQ_11) a (XQ_12) budou známy v několika málo sekundách, když v nedatabázovém programu vyhodnocení těchto dotazů bylo měřeno na minuty či desítky minut. Přesto však zastávám názor, že vyhodnocení dotazu, které trvá více než 10 sekund, není optimální. Bohužel ale, jak je patrné z tabulky 3 uvádějící všechny časy XQuery dotazů odzkoušených v databázích Berkeley DB XML a eXist (4Suite dotazovací jazyk XQuery nepodporuje), dotazů překračujících 10vteřinou hranici není málo.

Protože každý dotaz má jinou strukturu a obě databáze pracovaly při jejich zpracovávání jiným způsobem, není možné najisto říci, která z databází si poradila s vyhodnocením XQuery dotazů lépe. Databáze Berkeley DB XML zpracovávala v konečném výsledku všechny dotazy v rychlejším čase (tabulka 3) a rovnoměrněji (obrázky 1, 2), ale nebyla schopna pro nedostatek paměti vyhodnotit složité prostorové dotazy. Naproti tomu databáze eXist ve svých výsledcích velice kolísala (obrázky 1, 2), ale dokázala vyhodnotit dotazy všechny – otázkou však zůstává zda výsledný čas dotazu (XQ_12) 6 h 52 min 42,647 s (= 24 762,647 s) je v praxi akceptovatelný.

Vytvořené XPath dotazy byly vyzkoušeny ve všech použitých databázových systémech. Protože struktura těchto dotazů si je mnohem bližší, než tomu bylo u dotazů XQuery, bylo jejich vyhodnocení v rámci jednotlivých databází velice podobné – z hlediska času však kvůli různým přístupům jednotlivých databází opět odlišné (tabulka 4). Například u databází 4Suite a eXist bylo po srovnání časových hodnot podle velikosti dosaženo totožných výsledků, tj. pořadí dotazů se shoduje. U těchto dotazů již nedocházelo k výkyvům, pouze databáze eXist vyhodnotila dotaz (XP_4) nad 10vteřinovou hranici, což byl celkově nejhorší dotaz celého

testování XPath dotazů (13,389 s), který této databázi velmi ovlivnil celkový čas potřebný k vyhodnocení všech XPath dotazů (tabulka 5).

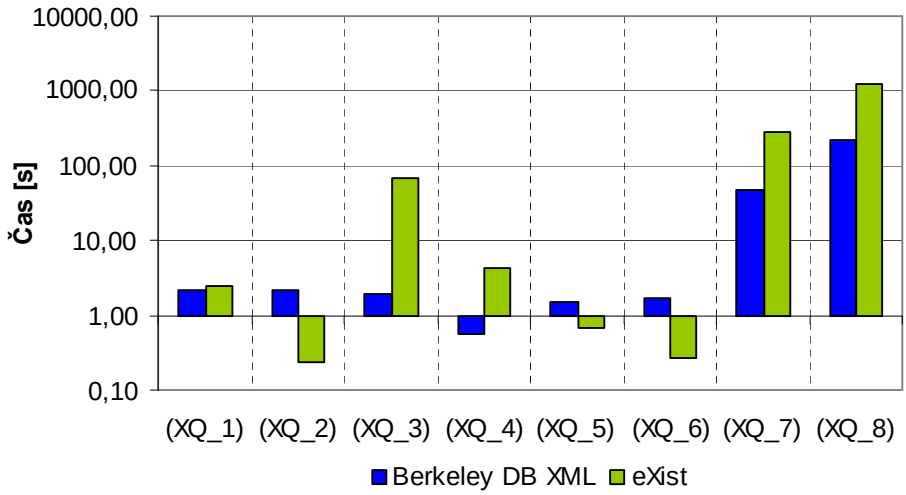
	Berkeley DB XML	eXist
(XQ_1) atrib_dotaz_CR_radka.xquery	2,169	2,454
(XQ_2) atrib_dotaz_UK_radka.xquery	2,130	0,240
(XQ_3) atrib_dotaz_UK.xquery	1,898	68,538
(XQ_4) atrib_dotaz_hustota_200_300.xquery	0,552	4,286
(XQ_5) atrib_dotaz_kontinenty.xquery	1,473	0,661
(XQ_6) atrib_dotaz_hustoty.xquery	1,716	0,271
(XQ_7) atrib_dotaz_hustota_Evropa.xquery	46,160	289,816
(XQ_8) atrib_dotaz_prum_hustota_svet.xquery	226,688	1 257,899
(XQ_9) prostor_dotaz_souradnice.xquery	2,415	9,543
(XQ_10) prostor_dotaz_multipolygony.xquery	1,054	0,531
(XQ_11) prostor_dotaz_sousedi.xquery	"out of memory"	2 028,918
(XQ_12) prostor_dotaz_okruh_1000km.xquery	"out of memory"	24 762,647

Tabulka 3: Výsledné časy vyhodnocení XQuery dotazů v sekundách

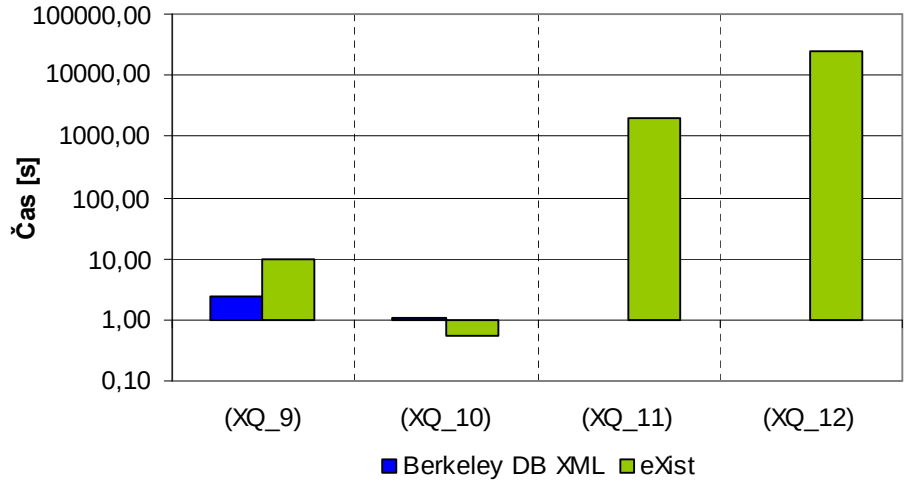
	4Suite	Berkeley DB XML	eXist
(XP_1) XPath_atrib_dotaz_CR_magda.txt	2,078	0,793	0,074
(XP_2) XPath_atrib_dotaz_UK_magda.txt	2,484	0,731	0,108
(XP_3) XPath_atrib_dotaz_prum_hustota_sveta_magda.txt	2,648	1,407	0,138
(XP_4) XPath_prostor_dotaz_souradnice_magda.txt	3,242	1,736	13,389
(XP_5) XPath_prostor_dotaz_multipolygony_magda.txt	1,886	0,593	0,058

Tabulka 4: Výsledné časy vyhodnocení XPath dotazů v sekundách

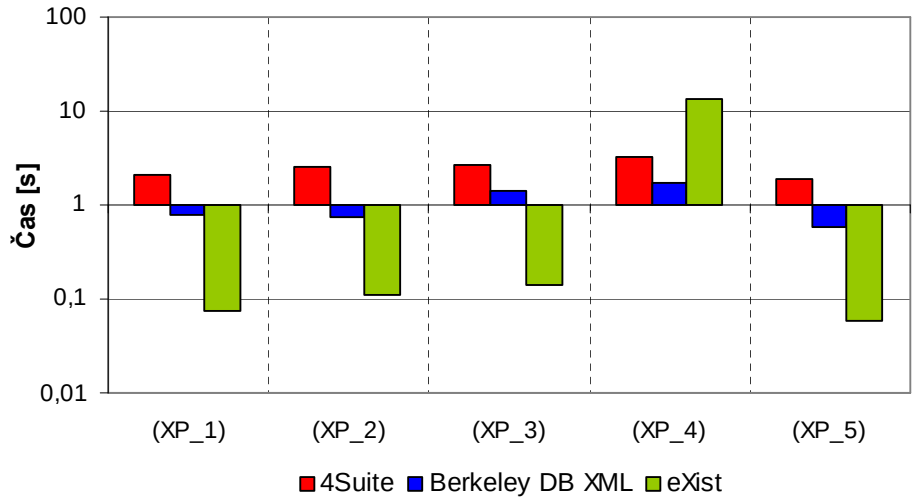
Pozn. Pro grafické znázornění časů vyhodnocení dotazů jsem z důvodu velkých rozdílů jednotlivých hodnot použila pro větší přehlednost na ose *y* logaritmické měřítko.



Obrázek 1: Grafické znázornění času vyhodnocení atributových XQuery dotazů jednotlivých databází



Obrázek 2: Grafické znázornění času vyhodnocení prostorových XQuery dotazů jednotlivých databází



Obrázek 3: Grafické znázornění času vyhodnocení XPath dotazů v jednotlivých databázových systémech

Celkové zhodnocení dotazování

Po dokončení testování XQuery a XPath dotazů jsem dospěla k závěru, že ze všech použitých databází pracovala nejrychleji a nejstabilněji databáze Berkeley DB XML. Pro ukládání geodat by však tato databáze zřejmě musela pracovat na výkonnějším počítači, aby dokázala vyhodnotit i složitější dotazy. Databáze eXist sice byla schopná dokončit všechny dotazy, ale celkový potřebný čas k jejich vyhodnocení je pro praktické využití (alespoň na počítači s podobnou konfigurací) příliš vysoký, což je patrné z tabulky 5. Databáze 4Suite dokázala vyhodnotit XPath dotazy velice spolehlivě a v podobných časech, její využití bych ale pro ukládání geodat pro nepodporu XQuery nevolila.

	4Suite	Berkeley DB XML	eXist
atributové XQuery dotazy	x	4 min 42,786 s	27 min 4,165 s
prostorové XQuery dotazy	x	"nelze vyhodnotit"	7 h 26 min 41,639 s
Xpath dotazy	12,338 s	5,259 s	13,768 s

Tabulka 5: Sumarizace časů jednotlivých databází po vyhodnocení všech dotazů

Závěr

Vzhledem k výsledkům, ke kterým jsem dospěla při aplikaci dotazů, považuji využití nativních XML databází pro ukládání geodat na počítači s podobnou konfigurací (1,8 GHz, 768 RAM) jako ne příliš vhodné. Pevně věřím, že na výkonnějších počítačích by například databáze Berkeley DB XML byla schopná vyhodnotit všechny dotazy a databáze eXist by složitější dotazy řešila poněkud rychleji.

Negativní výsledky však neovlivnila pouze konfigurace počítače, hlavní příčinou byla podle mého názoru špatná vnitřní struktura zdrojových dokumentů, kde obrovský počet hraničních souřadnic jednotlivých států byl uveden v jednom elementu s nadbytečnými mezerami (pro jednu dvojici souřadnic cca 25 mezer), proto k problémům docházelo především při prostorových dotazech, kde bylo nutné jednotlivé souřadnice porovnávat.

Vlastní velikost souboru samozřejmě také hrála významnou roli. S velkými velikostmi souborů se však v oblasti geověd musí počítat. Ke zmenšení velikostí by mohlo vést například nahrazení názvů elementů kratšími názvy nebo také využití binárního XML.

Po optimalizaci struktury i velikosti jednotlivých souborů věřím, že nativní XML databáze by na výkonnějších počítačích bylo možné v praxi využít. Jejich vlastnímu využití by však musel předcházet složitý výběr z existujících databází, protože pro geodata, jejich správu a zpracovávání by měla databáze podporovat co možná nejvíce XML technologií, a to především dotazovací jazyky (zejména XQuery), transformační jazyky a jazyky pro aktualizaci.

Seznam zdrojů

[1] BOURRET, Ronald. *XML and Databases* [online]. c2006 [cit. 2006-08-15]. < http://www.rpbourret.com/index.htm>

[2] The Apache Software Foundation. *4SUITE.org* [online].c2000 [cit. 2006-09-05]. < http://4suite.org/index.xhtml >

[3] Oracle. *Sleepycat Products: Berkeley DB XML* [nline]. c2006 [cit. 2006-08-15]. < http://sleepycat2.inetu.net/products/bdbxml.html >

[4] *eXist – Open Source Native XML Databáze*[online].[cit. 2006-09-05]. <http://exist.sourceforge.net/>

[5] Technologie XML / Irena Mlýnková ... [et al.]. -- 1. vyd.. -- Praha : Karolinum, 2006. -- 186 s. :. -- (Učební texty Univerzity Karlovy v Praze). -- 500 výt.. -- ISBN 80-246-1272-0 (brož.) :