

ULOŽENÍ, DISTRIBUCE A SDÍLENÍ DAT INTERNETOVÝMI PORTÁLY

DATA STORING, DISTRIBUTION AND SHARING IN INTERNET PORTALS

Vratislav Čermák, Pavel Šimek

Anotace:

V dnešním globálním světě není možné pro výměnu dat používat nějaké proprietární formáty, které jsou svázány s konkrétním softwarem nebo hardwarem. Proto by každý portál měl svá rozsáhlá data uchovávat v určitém, snadno rozšiřitelném, standardu, který lze velmi lehce distribuovat do různých koncových zobrazení. Jako nejlepší pro uchovávání strukturovaných a semistrukturovaných dat se jeví formát XML (eXtensible Markup Language - rozšiřitelný značkovací jazyk). Práci s XML usnadňuje i to, že celý formát je založen na obyčejném textu. I když pro většinu lidí zůstane XML kód skryt a budou ho používat pouze aplikace pro vzájemnou komunikaci, není problém kdykoliv otevřít XML dokument v jakémkoliv jednoduchém textovém editoru, ve kterém je možné provést veškeré potřebné úpravy.

Klíčová slova:

distribuce, sdílení, XML, XSL, XSLT, RSS, PHP, portál, Internet

Abstract:

In today's world it isn't possible to use incompatible formats that may be used only with specific software and hardware. Because of this fact portal should store its data in easily extensible format that can be distributed to many different devices. The best format for this purpose seems to be XML (eXtensible Markup Language). The work with XML makes easy, that it is just a plain text. In most cases is a XML code hidden and it is used only for communication between applications, but it is no problem to open the XML document in any text editor and change the document there.

Keywords:

distribution, sharing, XML, XSL, XSLT, RSS, PHP, portal, Internet

ÚVOD

Téměř na všech stránkách internetových e-zinů je portál označován jako nejvhodnější nástroj pro zpřístupnění informací určité skupině uživatelů. Z technologického hlediska může být takovým portálem webová aplikace komunikující s různými jinými aplikacemi skrze různá rozhraní. V současné době není možné pro výměnu dat používat nějaké proprietární formáty, které jsou svázány s konkrétním softwarem nebo hardwarem. Proto by každý portál měl svá rozsáhlá data uchovávat v určitém, strukturovaném, snadno rozšiřitelném, standardu, který lze velmi lehce distribuovat do různých koncových zobrazení. Jazyk XML vyhovuje všem náročným požadavkům na ukládání dat a jejich následné efektivní zpracovávání.

CÍL PRÁCE A METODIKA

V roce 2002 došlo díky neúprosnému pokroku v oblasti ICT k úplnému přepracování agrárního WWW portálu Agris. Jedním z hlavních požadavků na uchovávání dat byla jejich snadná distribuce do rozličných aplikací a různých koncových zařízení. Proto bylo zvoleno XSLT (Extensible Stylesheet Language Transformations) a XML (Extensible Markup Language). Původní data bylo nutné převést do XML formátu a zároveň vytvořit podporu pro snadné ukládání nových dat.

TRANSFORMACE XML DOKUMENTŮ

XSLT je v podstatě mechanismus pro transformaci XML dokumentů a manipulaci s nimi. Jazyk XML umožňuje strukturovat data uložená v jednotlivých dokumentech. XSLT ovšem umožňuje s obsahem těchto dokumentů pracovat – manipulovat s ním a na jeho základě vytvářet další dokumenty. Mezi základní manipulace s obsahem patří setřídění záznamů ve formátu XML nebo uložení dat v dokumentu XHTML, PDF, RTF, či jiném formátu. XSLT ovšem umožňuje nastavit i velmi pokročilé formátování.

S obsahem dokumentů XML lze pracovat několika způsoby. Je možné vytvořit vlastní program, který spolupracuje s aplikacemi, jejichž úkolem je analýza kódu XML. To však nevyhnutelně vyžaduje sestavování vlastního kódu. Pomocí XSLT však lze vykonávat stejné úlohy i bez programování. Místo psaní vlastního kódu v jazyce Visual Basic, v Javě nebo v C++ lze s obsahem dokumentů XML manipulovat jinak: pomocí XSLT je možno jednoduše určit, co se má s určenými částmi zdrojového dokumentu provést, a procesor XSLT toto provede.

XSLT je ve skutečnosti součástí širší specifikace, specifikace jazyka XSL (Extensible Stylesheet Language). Podstatou jazyka XSL je určení přesného formátu dokumentu. Formátovací část jazyka XSL, která podstatně překračuje rámec specifikace XSLT, je založena na speciálních formátovacích objektech. Zmíněná část XSL je často označována jako XSL-FO (nebo XSL:FO či XSLFO).

Technologie XSLT umožňuje uživatelům přímou manipulaci s obsahem dokumentů XML. Kód XSLT může být načten kterýmkoli procesorem XSLT, jenž pracuje s dokumenty XML. Na vývojáři zůstane pouze vytvoření dokumentu obsahujícího vhodný styl XSLT. Ten určí pravidla, která jsou aplikována při transformaci jednoho dokumentu na jiný. Kromě běžné transformace jednoho dokumentu XML na jiný téhož formátu lze dokumenty XML převádět i na další typy dokumentů, jako jsou třeba dokumenty XHTML, RTF, TXT, dokumenty ve formátu XSL-FO nebo jiné. Dokumenty XML lze transformovat na všechny formáty založené na XML: například MathML, MusicML, WML, HTML a mnoho dalších.

SPRÁVNĚ STRUKTUROVANÉ A VALIDNÍ DOKUMENTY XML

Aby mohl být dokument XML pokládán za správně strukturovaný, musí být dodržena syntaktická pravidla stanovená konsorciem W3C v doporučení XML 1.0, resp. XML 1.1 (www.w3c.org/TR/REC-xml). Neoficiálně znamená fráze „správně strukturovaný“ především to, že dokument obsahuje jeden nebo více elementů, přičemž jen jeden z nich, tj. kořenový element (root element), musí obsahovat všechny ostatní elementy. Kromě toho musí být všechny elementy správně ohraničeny příslušnými značkami (tags).

Většina XML parserů (nejčastěji např. internetový prohlížeč) ověřuje, zda je dokument správně strukturován. Některé z nich ovšem ověřují i skutečnost, zda je dokument validní (platný). Dokument XML je validní, je-li v něm přidružena deklarace typu dokumentu (DTD) nebo schéma XML a pokud dokument uvedenému DTD nebo schématu plně odpovídá. Znamená to, že DTD nebo schéma definují množinu pravidel, která si vynucují vnitřní konzistenci příslušného dokumentu. Je-li prohlížeč schopen určit, zda se dokument zmíněnými pravidly řídí, je dokument validní (platný).

XML je asi vůbec první formát, který hned od samého počátku myslel na potřeby i jiných jazyků než je angličtina. Jako znaková sada se používá ISO 10646. ISO 10646 je

32bitová znaková sada, která dokáže pojmut všechny dnes používané znaky všech jazyků.

V XML proto lze vytvářet dokumenty, které obsahují texty v mnoha jazycích najednou – je možno kombinovat např. češtinu, angličtinu, ruštinu, arabštinu a korejštinu zcela dle libosti. Pokud by dokumenty obsahovaly pouze český text, bylo by ukládání přímo v ISO 10646 zbytečné plýtvání místem. XML dokument proto může být v libovolném kódování (např. windows-1250, ISO-8859-2). Kódování je však v každém dokumentu přesně určeno, takže odpadají problémy s konverzí z jednoho kódování do druhého.

VÝSLEDKY

Naprostá většina zpravodajského materiálu, který je nabízen agrárním WWW portálem Agris je uložena ve formátu XML. K tomuto formátu bylo přistoupeno z důvodů, které byly popsány v předchozí části tohoto článku. Ovšem jako nejvýznamnější se jeví možnost snadné výměny dokumentů mezi aplikacemi a uchovávání dat ve strukturované a semistrukturované formě. Předchozí verze celého agrárního WWW portálu, jejíž provoz byl ukončen v první čtvrtině roku 2002, tuto formu uchovávání rozsáhlejších dat postrádal. Nicméně, veškeré texty z archívu se zdárně podařilo převést do současné verze Agrisu, tudíž celý WWW portál obsahuje dokumenty uložené pouze ve formátu XML.

VYUŽITÍ XML V AGRÁRNÍM WWW PORTÁLU

V databázi jsou uložena pouze data, která jsou pro daný dokument (text) klíčová, jako je titulěk, perex, datum pořízení, jméno autora, datum zveřejnění, identifikační číslo dokumentu, krátká charakteristika obsahu článku, data pro fulltextové vyhledávání, apod. Samotné XML dokumenty jsou uloženy na disku pod unikátním identifikátorem.

Pro názornou ilustraci je poskytnuta ukázka dvou článků z agrárního WWW portálu Agris, kde stojí za povšimnutí struktura samotných dokumentů.

```
<?xml version = "1.0" encoding = "windows-1250"?>
<nitf id = "128780" cesta = "51">
<body>
<body.content>
<p>Boneco Alima. Tak vyslovují od začátku roku obchodní název společnosti její zaměstnanci, když se představují do telefonu. Společnost Boneco, jejíž benešovský provoz zaměstnává několik desítek lidí, končí. Většina lidí ale podle všeho o práci nepřijde. Na přelomu roku se uskutečnila fúze tří potravinářských společností Jemča, Nanuk a Boneco do jediné společnosti Alima. Vedení společnosti si od transformace slibuje větší prosperitu a nedojde ani k většímu propouštění lidí. Podle obchodního ředitele firmy Zdeňka Hanuše se sloučením společností významně zefektivní výroba, sníží náklady a posílí pozice na trhu potravin. </p>
<em class="nitf_bold">
Nezaměstnaných nebude více
</em>
<p>Ani ředitel úřadu práce Petr Němeček se přívalu nezaměstnaných neobává. "Propouštění se dotkne jen několika administrativních pracovníků, kterých bude méně než deset," uvedl Petr Němeček s tím, že jde převážně o zaměstnance, kterým byla nabídnuta práce v jiné lokalitě, kvůli nutnosti dojíždět ji ale nepřijali. </p>
</body.content>
</body>
</nitf>

<?xml version = "1.0" encoding = "windows-1250"?>
```

```
<nitf id = "128947" cesta = "51">
<body>
<body.content>
<p>Společnost Kostecké uzeniny, jejíž manažery obvinili minulý týden policisté z poškození firmy o nejméně 200 miliónů korun, včera poprvé uvedla podrobnosti, jichž se vyšetřování týká.</p>
<p>"Policie podala trestní oznámení na pět lidí, kteří byli v představenstvu v letech 1996 až 2001. Čtyři z nich jsou v představenstvu i dnes," připustila Vladka Pivoňková, marketinková šéfka firmy, tuzemské jedničky v oboru.</p>
<p>Vyšetřování se podle ní týká možnosti, že před několika lety firma měla slevami zvýhodnit jednoho ze svých odběratelů. Šlo o družstvo Maso uzeniny, jež je zároveň vlastníkem Kosteckých uzenin. „Družstvo zvýhodňováno nebylo; slevy jsme dávali i dalším odběratelům. Všechny podklady jsme předali policii už dávno. Neoficiálně totiž věc šetří už téměř dva roky, a nám na vyjasnění věci záleží," prohlásila Pivoňková.
<media>
<media-reference source="Imagel.jpg"
alternate-text="Ilustrační
foto">Imagel.jpg
</media-reference>
</media></p>
</body.content>
</body>
</nitf>
```

Ukázka XML

Pomocí XSLT lze velmi snadno přenést tyto dokumenty do jakéhokoli formátu (XHTML, XML, TXT, RTF, PDF apod.). Tudiž všechny aplikace, ze kterých se samotný Agris skládá i přidružené aplikace mohou velmi snadno mezi sebou komunikovat a vyměňovat si data.

Veškeré šablony jsou v případě Agrisu uloženy v jednom samostatném souboru s příponou .xsl, dle kterého se formátuje výstup dat portálu ve validním (platném) kódu XHTML. Koncepce šablon XSLT umožňuje rozdělovat jednotlivé části XSLT dokumentu do samostatných částí (souborů), stejně tak jako vlastní XML data. Tímto se dosahuje větší přehlednosti. V ukázce transformace je zobrazena transformace elementu media, který se v NITF formátu používá především pro vkládání obrázků, do kódu XHTML:

```
<xsl:template match="media">
  <div class="media-obj">
    <p class="mediacaption"><xsl:value-of select="media-caption"/></p>
    <xsl:element name="div">
      <xsl:attribute name="style">text-align:left; width:<xsl:value-
of select="media-reference/@width"/>
      </xsl:attribute>

      <xsl:element name="img">
        <xsl:attribute name="src">/_images/<xsl:value-of
select="/nitf/@cesta"/></xsl:value-of
select="/nitf/@id"/></xsl:value-of select="media-
reference/@source"/>
        </xsl:attribute>
        <xsl:attribute name="alt"><xsl:value-of select="media-
reference/@alternate-text"/>
        </xsl:attribute>
      </xsl:element>

      <xsl:if test="string-length(//media-reference/@source-credit) != 0" >
        <div style="text-align:left">
          Zdroj:
          <xsl:value-of select="//media-reference/@source-credit"/>
        </div>
      </xsl:if>
    </xsl:element>
  </div>
</xsl:template>
```

Ukázka XSLT

Díky univerzálnímu uložení strukturovaných dat je možné je použít v různých dalších projektech. Navazující projekty pak nemusí být na jednotné platformě, nemusí využívat stejné prostředí, či prostředky pro transformaci. Data tedy mohou být využita např. jiným portálem, čtečkami XML, jako modulární součást informačních systémů, součást klasických windows aplikací, apod. XML formát může být načten a používán produkty různých výrobců. Vhodnost použití formátu XML dokládá i jeho úspěšné nasazení v projektech Asociace soukromých zemědělců, Šance pro jihozápad, Klub zemědělských novinářů a publicistů. Všechny tyto webové projekty přebírají data z portálu Agris a pomocí vhodných XSLT transformací se upravují do vzhledu konkrétního webu.

Díky standardizaci formátu XML a použitím standardizovaného DTD NITF je možné nejen transformovat strukturovaná data do jiného výstupního formátu, ale je možné je přetransformovat v budoucnosti do jiného formátu XML (do jiného DTD).

Největším problémem s využíváním XML formátu v reálném provozu internetového portálu Agris bylo pořizování textů administrátory – neodborníky na internetové technologie. Administrátoři portálu potřebují pro pořizování nejlépe prostředí tzv. WYSIWYG (What You See Is What You Get) editorů, kterých není na současném trhu příliš. V současné době je pro redakční systémy nejčastěji využíváno webového přístupu – tj. informace do portálu jsou zadávány prostřednictvím speciální internetové stránky, která například obsahuje vestavěný HTML editor. Tyto editory ovšem produkují nevalidní (neplatné) XML dokumenty a není proto možné je využívat pro pořizování strukturovaných XML textů. Pro tyto účely bylo třeba vyvinout klasickou aplikaci, pracující v prostředí operačního systému, založenou stejně jako

klasické webové aplikace (web-based application), na principu tenkého klienta. Výhodou této aplikace (napsané v Delphi) je integrace komponenty umožňující WYSIWYG editaci XML textů podle zadaného DTD. Aplikace komunikuje se serverovým modulem také na bázi XML formou XML požadavků a zpráv.

RSS

RSS slouží pro syndikaci - terminus technicus vyjadřující možnost přebírat obsah z různých zdrojů a dále ho používat pro vlastní stránky či služby - většina syndikací zpracovává pouze titulky, odkazy a perexy konkrétních článků. RSS je zkratka z **R**DF **S**ite **S**ummary, kde zkratka RDF znamená **R**esource **D**escription **F**ormat. V principu jde o XML vyjádření obsahu (povětšinou) zpravodajských zdrojů - může ale jít o vyjádření jakýchkoliv jiných libovolných "dat".

Při tvorbě RSS souboru je dbáno určitých pravidel, hlavně:

- Vždy je uváděna úplná a správná adresa zdroje.
- Nejlepší je používat kódování UTF-8, autor se tak vyhne možným problémům při čtení RSS.
- https:, file:, mailto:, news:, a javascript: odkazy nejsou povoleny.
- Vždy je uváděn smysluplný obsah

RSS je tedy dalším způsobem využití XML v agrárním www portálu. Z Agrisu je možné stáhnout hlavní zprávy aktuálního dne ve formátu RSS/XML a potom je pomocí vlastní šablony přidat do jiného webu, nebo různě transformovat do uživatelem požadovaných formátů. Agris RSS je vytvářeno ve verzi 2.0 a je možné si jej stáhnout z adresy: <http://www.agris.cz/etc/rss.php>. Některé další portály přebírají tuto formu zpravodajství a integrují tak různé, takto získané informace, do jednoho souhrnného zdroje informací, i když se ve skutečnosti jedná pouze o jakýsi zpravodajský rozcestník, jelikož formát RSS neobsahuje plné texty, ale pouze perexy (anotace) jednotlivých článků a odkaz na jejich plné znění. Příkladem portálu integrujícího zprávy z formátu RSS je www.kratce.cz.

Uvnitř portálu je RSS/XML využíváno pro postranní lištu využitelnou v prohlížečích Mozilla. Je to uživatelsky nastavitelný rám v levé části uživatelského prohlížeče, kde lze v různých panelech přistupovat k aktuálním informacím. Mezi panely je možné se přepínat klepnutím na název panelu. Výhody, které postranní lišta v prohlížeči poskytuje jsou:

- novinky agrárního WWW portálu Agris má uživatel neustále aktualizované kdykoliv k dispozici
- uživatel tedy nemusí stále otevírat kompletní stránky a čekat na jejich načtení
- uživatel může prohlížet různé rubriky portálu a aktuality má neustále na očích
- na postranní liště jsou všechny novinky z Agrisu a pouhým kliknutím se uživatel dostane na požadovanou rubriku, či konkrétní článek.
- každých 10 minut jsou k dispozici nové aktuální informace
- postranní lišta nezpomaluje ani nijak neomezuje uživatelské prohlížení internetu

Ve skutečnosti je použita pouze vhodná XSLT transformace pro vygenerování postranní lišty z RSS formátu. Tímto je možné využít vhodnou formu XML i pro jiné aplikace.



Postranní lišta v prohlížeči

ZÁVĚR

Standard XML je bezesporu nejúspěšnějším dokumentem, který kdy vzešel z konsorcia W3C. Jeho použití je skutečně univerzální, ať již je řeč třeba o jeho aplikacích používaných na internetu (XHTML, SVG,...) nebo jako o formátu pro ukládání dat.

Veškeré články agrárního WWW portálu Agris jsou uloženy právě v XML a jsou snadno transformovatelné do jiných formátů. Pro syndikaci – přebírání obsahu z různých zdrojů – je využito RSS verze 2.0. Využití XML bylo jedním z nejzásadnějších kroků při programování nové verze celého Agrisu.

Literatura:

- BRADLEY, Neil. *XML : Kompletní průvodce*. Přeložil BRÁZDA, Jiří. Praha : Grada, 2000. 537 s. ISBN 80-7169-949-7.
- HOLZNER, Steven. *XSLT : příručka internetového vývojáře*. 1. vyd. Praha : Computer Press, 2002. 515 s. ISBN 8072266004.
- CASTAGNETTO, Jesus – RAWAT, Harish – SCHUMANN, Sascha – SCOLLO, Chris – VELIATH, Deepak. *Programujeme PHP profesionálně*. 2. vyd. Praha : Computer Press, 2001. 656 s. ISBN 8072263102.
- KOČÍ, Michal. *Co je XML?*. [online]. c2000 [cit. 2004-03-15]. <<http://interval.cz/clanek.asp?article=145>>.
- BUREŠ, Jiří. *RSS? RSS!*. [online]. c2003 [cit. 2004-03-15]. <<http://interval.cz/clanek.asp?article=2050>>.