Sémantický web a jeho technologie (2)

Petr Matulík, Tomáš Pitner, FI MU

V minulé části tohoto seriálu jsme skončili stručným představením funkce klasifikačních schémat a dotkli jsme se i nejznámějších z nich, kterými jsou schémata Dublin Core a vCard. V dnešní části se zmíníme o některých dalších schématech, představíme si řízené slovníky a podrobněji se podíváme na metadatový fenomén posledních měsíců, standard RSS.

6 Další klasifikační schémata a jejich registry

Stručně uveďme další zajímavá klasifikační schémata:

- Konsorcium *PRISM*[1] (*Publishing Requirements for Industry Standard Metadata*, standardizuje metadata v oblasti publikace a výměny obsahu ve zpravodajství (publikace, licencování a znovupoužití obsahu, práva k digitálním dokumentům atd.).
- *W3C XPackage*[2] je specifikací, jak vytvářet metadatové popisy *kolekcí zdrojů*.
- *DAML+OIL*[3] je jedním z pilířů sémantického webu, umožňuje zachycovat ontologie a sémantiku webových zdrojů.

V prostředí Internetu se začínají objevovat i registry (seznamy) klasifikačních schémat[4].

7 Řízené slovníky

Řízeným slovníkem, tezaurem či terminologickou ontologií rozumíme soubor předmětových hesel s definovanou strukturou nadřazených a podřazených termínů a určením synonym či jiných pojmových relací. Přestože řízený slovník si může pro vlastní potřebu definovat úzká skupina uživatelů, pro využití v zájmu sémantického webu je třeba, aby byl slovník dostupný všeobecně. V kontextu RDF je význam řízených slovníků zřejmý. Pro určení hodnoty určité vlastnosti mohou klasifikační schémata vyžadovat použití hesla z konkrétního slovníku, což podstatným způsobem přispívá k interoperabilitě metadat. Toho také mohou využít tvorbu metadat usnadňující aplikace, které jsou kompatibilní s daným schématem a daným slovníkem a které tak mohou průběžně nabízet hesla ze slovníku jako možné hodnoty vlastností. Řízené slovníky jsou používány jako obor hodnot i pro některé vlastnosti klasifikačního schématu Dublin Core. V prostředí kvalifikovaného Dublin Core patří řízené slovníky mezi tzv. *kvalifikátory hodnoty*, které různými způsoby omezují obor hodnot dané vlastnosti. V podstatě lze při RDF popisu nejen specifikovat hodnoty vybraných vlastností schématu Dublin Core, ale i určit řízený slovník, ze kterého jsme hodnotu vlastnosti vybrali. Vidět to můžeme na následujícím příkladu, který znázorňuje použití řízeného slovníku v rámci kvalifikovaného DC v hlavičce HTML dokumentu.

```
<meta name=DC.Subject scheme=LCSH con-
tent=Dublin Core; DC; RDF; XML>
```

Použití řízeného slovníku *LCSH* (Library of Congress Subject Headings) je dáno jeho uvedením v atributu `scheme`. Atribut `content` pak obsahuje hesla z tohoto slovníku vybraná. K nejznámějším terminologickým ontologiím patří například *WordNet*[5] či jeho následníci *Sensus*[6] a vícejazyčný *EuroWordNet*, do nějž přispívá i tým *Laboratoře zpracování přirozeného jazyka* FI, který rovněž pořádal *Global Wordnet Conference (GWC) 2004*[7]. Další informace o řízených slovnících včetně jejich dostupnosti na webu lze najít např. v dokumentu na `http://info.sks.cz/users/ku/MTI/sjazyky.htm`.

8 RSS

V dnešní době je zřejmě nejrozšířenějším reálně používaným metadatovým formátem v pro-

[1] `http://www.prismstandard.org`
[2] `http://xpackage.org`
[3] `http://www.daml.org/2001/03/daml+oil`
[4] `http://metadata.net/`, `http://xmlns.com/`, `http://www.schemas-forum.org/`, `http://www-ksl.stanford.edu/knowledge-sharing/ontologies/README.html`, `http://desire.ukoln.ac.uk/registry/`, `http://athena.ics.forth.gr:9090/RDF/Examples.html`

[5] `http://www.cogsci.princeton.edu/~wn/`
[6] `http://www.isi.edu/natural-language/projects/ONTOLOGIES.html`
[7] `http://www.fi.muni.cz/gwc2004`

středí Internetu standard *RSS*, který sice s ideou sémantického webu souvisí jen volně, pro jeho praktickou významnost a takřka „každodenní použitelnost“ jej však nemůžeme pominout. Komunita vyvíjející standardy RSS je, bohužel, natolik rozštěpená, že vznikající standardy jsou navzájem nekompatibilní a shoda neexistuje ani ve významu akronymu RSS Pokusme se proto alespoň o obecné přiblížení podstaty RSS standardů.

RSS je univerzální široce použitelný metadatový formát pro agregaci a syndikaci internetového obsahu. *Syndikace* je pak ve specifikaci jedné z verzí RSS definována jako vytváření on-line přístupných dat, která slouží k dalšímu přenosu, agregaci a následnému znovupublikování. V době totální informační přesycenosti je pro zájemce o určitou specifickou oblast obtížné, ne-li nemožné pravidelně „brouzdat“ po desítkách webových míst, kde (občas) nalézá požadované informace. RSS umožňuje koncentrovat aktuální informace ze vybraných webových zdrojů na jedno místo, například na webový portál.

8.1 Exportní soubor RSS

K zprostředkování výše zmíněných informací využívá RSS takzvaný exportní soubor (kanál), jehož syntaxe odpovídá standardu RSS dané verze. Jednou z mála společných vlastností všech verzí RSS je fakt, že formáty souborů kanálů jsou vždy aplikací XML, z čehož vyplývá možnost procházet a zpracovat syntakticky správné exportní soubory pomocí běžných XML nástrojů. Soubor se nazývá exportním proto, že umožňuje export výtahu z nových informací, které se objeví na daném webu, a to ve formě stručné a srozumitelné lidem i počítačům. Většinou jde o *názvy* a stručné *popisy* obsahu aktuálních článků na zpravodajském serveru, další využití je však prakticky neomezené.

Exportní soubor může být buď dynamicky generován, nebo ručně vytvářen webmasterem daného zdroje. Doporučovanými příponami jsou `.xml`, `.rss` nebo `.rdf`. Soubor je zveřejněn (má vlastní URL) a na jeho přístupnost by mělo být adekvátně upozorněno, nejlépe na titulní straně daného webu, obvykle typickou oranžovou ikonkou „XML“. Vhodná je také registrace do významných webových agregačních portálů. Soubor se skládá z popisu daného webu a jednotlivých položek, které reprezentují popis nových informací včetně odkazu na jejich zdroj. Takto lze například (s využitím některých níže popsaných nástrojů) dosáhnout toho, že budeme schopni vidět seznam všech nových článků týkajících se oblasti našeho zájmu v jednom okně internetového prohlížeče, a snadno se prostřednictvím odkazu dostat k originálu článku, který si zvolíme.

8.2 Historie RSS

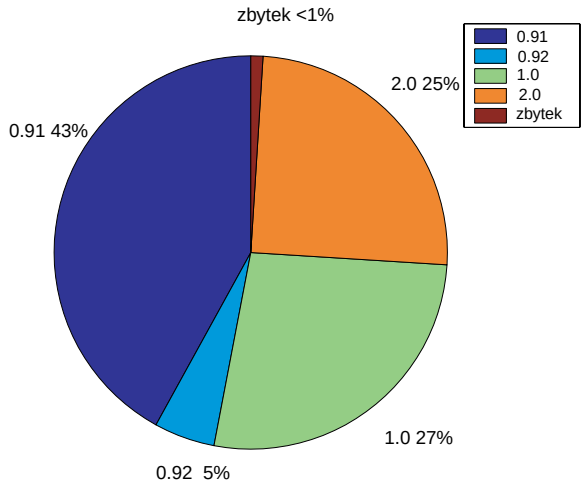
Naznačili jsme, že vývoj specifikace RSS byl spletitý a jednotlivé verze na sebe nenavazují. Formát RSS vznikl prapůvodně v dílně firmy Netscape, která jej chtěla využít jako mechanismus pro výtah obsahu na svém portálu my.netscape.com.

V březnu roku 1999 spatřila světlo světa specifikace *RSS 0.9*, jejíž jádro bylo založeno na RDF. Později došlo k zjednodušení standardu, odstranění RDF syntaxe a zařazení nových vlastností používaných v konkurenčním formátu *scriptingNews* firmy UserLand.

Tak vznikla verze *RSS 0.91*[8], která se dodnes hojně používá a jejíž jednoduchost pravděpodobně stimulovala pozdější rozšíření RSS. Netscape mezitím ztrácí o RSS zájem a hlavní slovo při vývoji standardu získává David Winer z firmy UserLand. Podle mnohých jde o nekonvenčního a impulsivního člověka, se kterým není snadné se domluvit a který je příčinou dnešního chaosu na poli RSS. V dalším období dospívají někteří uživatelé RSS k názoru, že struktura RSS 0.91 je nerozšiřitelná, její použití má příliš úzký obzor a použitá XML syntaxe je spíše intuitivní než přesně definovaná.

Proto vzniká samostatná mezinárodní skupina vývojářů, která v prosinci 2000 navrhuje *RSS 1.0*[9] založenou znovu na RDF a rozšiřitelnosti (modularizaci) pomocí jmenných prostorů. UserLand

[8] `http://backend.userland.com/rss091`
[9] `http://www.purl.org/rss/1.0/spec`



Obrázek 1: Poměr využití jednotlivých verzí RSS na portálu Syndic8

reaguje rozšířením své verze na RSS 0.92[10] později RSS 0.93 a 0.94. U těchto verzí jde v podstatě jen o přidávání nových prvků.

Na požadavek rozšiřitelnosti odpovídá až v srpnu 2002, kdy publikuje verzi *RSS 2.0*[11] používající jmenné prostory.

Teď už můžeme objasnit význam akronymu RSS pro jednotlivé verze standardu. RSS 0.9x představovalo zkratku pro *Rich Site Summary*. Verze 1.0 vykládá RSS jako *RDF Site Summary* a korunu těmto zmatkům nasazuje Dave Winer a jeho RSS 2.0 s významem *Really Simple Syndication*.

V dnešní době najdeme jen málo zpravodajských serverů a významnějších weblogů, které nepoužívají RSS. Využití tří nejoblíbenějších verzí, tedy 0.91, 1.0 a 2.0 je však poměrně vyrovnané, což klade zvýšené nároky na tvůrce nástrojů pro zpracování RSS. Graf na obr. 1 zachycuje poměr využití jednotlivých verzí RSS v exportních souborech sdružovaných na portálu Syndic8[12], který je zřejmě nejobsáhlejším zdrojem pro každého, kdo by se chtěl o RSS zajímat podrobněji.

8.3 Příklady RSS

Pro ilustraci uveďme příklady nejpoužívanějších verzí RSS, a to na již dříve použitém příkladu zpravodajského serveru *sport.cz*, který pomocí RSS upozorňuje na nové články. Z důvodu úspory místa použijeme v exportním souboru jen dvě položky. Jejich množství je obecně neomezené, běžnou praxí je však 10 až 15 položek na jeden exportní soubor. U všech tří souborů chybí definice použitého DTD (Document Type Definition), přestože se - jako nepovinná - objevit může. Po prostudování struktury RSS exportního souboru je jasné, že může sloužit k agregaci jakýchkoli diskrétních jednotek informace, tedy například autorů a jejich e-mailových adres na daném webu, informací o produktech dané firmy, informací o zboží nabízeném daným elektronickým obchodem, sportovních výsledků, atd.

RSS 0.91 má povinný kořenový element `rss`, který používá rovněž povinný atribut `version` pro udání verze RSS. Následuje element `channel`, obsahující svůj vlastní popis a jednotlivé položky.

```
<?xml version=‘‘1.0’’?>
<rss version=‘‘0.91’’>
  <channel>
    <title>sport.cz</title>
    <link>http://www.sport.cz/</link>
    <description>sport.cz poskytuje široké
       spektrum informací ze všech
       sportovních odvětví</description>
    <language>cs</language>
    <item>
       <title>Sparta  Chelsea 0:1</title>
       <link>http://www.sport.cz/fotbal/
         2003/12/04/spartachelsea.html
       </link>
       <description>V tomto článku se
          zaměříme na průběh zápasu Sparta
          - Chelsea, na rozbor obranné hry
          pražského týmu a jeho perspektivy
          v dalším průběhu Ligy mistrů
       </description>
    </item>
    <item>
       <title>Zlín Hood 2003  turnaj
          v lukostřelbě</title>
       <link>http://www.sport.cz/ostatni/
          2003/12/04/zlinhood.html</link>
       <description>Extrémně zajímavé klání
         našich předních lukostřelců se
         odehrálo ve městě obuvi.
         Vzrušující atmosféru jsme se
         snažili zprostředkovat v tomto
         článku. </description>
    </item>
```

[10] http://backend.userland.com/rss092)
[11] http://blogs.law.harvard.edu/tech/rss
[12] http://www.syndic8.com

```
  </channel>
</rss>
```

Exportní soubor RSS 1.0 je poněkud méně čitelný pro běžného uživatele a jeho kód je trochu rozsáhlejší než u RSS 0.91, zároveň však poskytuje snadnou rozšiřitelnost pomocí jmenných prostorů definovaných v kořenovém elementu `rdf:RDF`. Použití předpony (prefixu) `rdf` pro implicitní jmenný prostor RDF je povinné, ostatní předpony jsou volitelné.

```
<?xml version=''1.0''?>
<rss version=''2.0'' xmlns:dc=''
      http://purl.org/dc/elements/1.1/''>
  <channel>
    <title>sport.cz</title>
    <link>http://www.sport.cz/</link>
    <description>sport.cz poskytuje široké
      spektrum informací ze všech
      sportovních odvětví</description>
    <language>cs</language>
    <item>
      <title>Sparta  Chelsea 0:1</title>
      <link>http://www.sport.cz/fotbal/
        2003/12/04/spartachelsea.html
      </link>
      <description>V tomto článku se
        zaměříme na průběh zápasu Sparta
        - Chelsea, na rozbor obranné hry
        pražského týmu a jeho perspektivy
        v dalším průběhu ligy mistrů.
      </description>
      <dc:creator>Petr Matulík</dc:creator>
      <dc:date>2003-12-04</dc:date>
    </item>
    <item>
      <title>Zlín Hood 2003  turnaj
        v lukostřelbě</title>
      <link>http://www.sport.cz/ostatni/
        2003/12/04/zlinhood.html</link>
      <description>Extrémně zajímavé klání
        našich předních lukostřelců se
        odehrálo ve městě obuvi.
        Vzrušující atmosféru jsme se
        snažili zprostředkovat v tomto
        článku.  </description>
      <dc:creator>Petr Matulík</dc:creator>
      <dc:date>2003-12-04</dc:date>
    </item>
  </channel>
</rss>
```

RSS 2.0 rovněž umožňuje rozšiřitelnost pomocí jmenných prostorů, uvedených v kořenovém elementu. Je také zpětně kompatibilní s předchozími verzemi 0.9x, což je umožněno faktem, že nástroje pro zpracování RSS ignorují elementy, které neznají.

8.4 Nástroje pro zpracování RSS

Udělejme si nyní přehled o nástrojích, které jsou určeny ke zpracování a využití standardu RSS. Zmíníme se o aplikacích, které jsou na internetu volně dostupné.

Nejčastěji používaným nástrojem jsou tzv. *desktopové čtečky* (neboli *agregátory*) *RSS*. Jde o aplikace instalované přímo na počítač uživatele a lze do nich zaregistrovat jednotlivé exportní soubory námi vybraných internetových zdrojů (webů). Čtečka pak periodicky stahuje všechny zaregistrované exportní soubory a jejich obsah prezentuje vhodným způsobem uživateli. Vzniká tak dojem, že data nejsou stahována uživatelem, ale spíše tlačena webovými zdroji k uživateli („push“). Typickým použitím je upozorňování na nové informace (například právě publikovaný článek) na zaregistrovaných zdrojích, přičemž je možné rychle přejít na původní zdroj informace. Uživatel také nemusí procházet všechny zdroje, které poskytují informace v oblasti, která ho zajímá, ale stačí mu pouze přečíst si popisy jednotlivých nových informací (řekněme výtahy z článků) a vybrat si přesně to téma, které ho právě zaujme.

K nejrozšířenějším čtečkám patří *FeedDemon*[13], *FeedReader*[14] a *ActiveRefresh*[15] pro operační systémy Windows a například *Shrook*[16] pro MacOS X. Existují také čtečky ve formě modulu (plug-in) pro existující aplikace, například *RSS Miranda Plugin* [17] pro Instant Messenger Miranda, *Newsgator*[18] běžící pod Microsoft Outlook nebo panel nástrojů (takzvaný „sidebar“) do prohlížeče Mozilla[19]. Dnes už je

[13] `http://www.feeddemon.com/feeddemon/`
[14] `http://www.feedreader.com`
[15] `http://www.activerefresh.com`
[16] `http://www.fondantfancies.com/shrook/`
[17] `http://miranda-im.org/download/details.php?action=viewfile&id=409`
[18] `http://www.newsgator.com/`
[19] `http://www.theonering.net/staff/corvar/cgi-bin/sidebar-inst.pl`

jasné, že panel nástrojů RSS bude i přímo v základní výbavě nového operačního systému Windows Longhorn.

Ne každý ovšem má důvěru k freewarovým aplikacím stahovaným z internetu a raději dá přednost RSS agregátorům ve formě webové aplikace. Tyto aplikace samy sdružují informace z exportních souborů a on-line je publikují na svých webových stránkách. Uživatel se obvykle může registrovat k používání personalizovaného agregátoru a vytvářet vlastní seznam sledovaných exportních souborů.

V zahraničí je nejvýznamnějším hráčem na tomto poli portál *Moreover*[20]. Naopak v českém prostředí jsou nejznámějšími agregátory tohoto typu zejména všeobecně zaměřené portály *Právě dnes*[21] a *rss.pooh.cz*[22]. Přehled obsahu českých weblogů poskytuje portál *RSSky*[23] a na oblast hardware a software je zaměřen *Minasite*[24]. Tyto on-line agregátory často akceptují i jiné formáty pro agregaci obsahu. Na takových je pak třeba se s webmasterem daného webu dohodnout.

I průměrný webmaster by měl být schopen využít RSS nejen pro publikaci agregovaného obsahu vlastního webu, ale i pro začlenění obsahu cizího exportu na své stránky. K tomu musí na svém webu implementovat zpracování RSS exportních souborů, což není zrovna triviální záležitost. Pomoci mu mohou veřejně poskytované služby[25], kde stačí zadat URL exportního souboru, který chce zahrnout do svých stránek, a po odeslání je mu vygenerován krátký kód v javascriptu, který jednoduše vloží do svého kódu. Problémy ovšem vznikají při kódování českých znaků – bezchybně funguje jen pro RSS kódované v UTF-8. Služba *Feed Validator*[26] dovoluje ověřit syntaktickou správnost zvoleného exportního souboru. Existují i služby, které po zadání URL validního XHTML zdroje vygenerují jeho reprezentaci v RSS 1.0[27]. Další zajímavou službou může být RSS agregátor *Novobot*[28], který má nejen klasické vlastnosti, ale dokáže také procházet zdroje na zadaných URL a generovat jejich hlavičky (s využitím nadpisů, odkazů atd.) i bez existence exportních souborů.

9 Závěrem

Obrovská síla standardu RSS je v jeho jednoduchosti. Formát, který nelze pochopit prakticky okamžitě a který tak nemůže průměrný webmaster ihned začít používat, má jen malou šanci se v prostředí Internetu masově prosadit. RSS však tuto vlastnost má, a to mu spolu s širokou použitelností a faktem, že je založen na XML, zaručuje ještě dlouho udržení značného náskoku před ostatními – byť možná sofistikovanějšími – metadatovými standardy. □

[20] http://www.moreover.com)
[21] http://www.pravednes.cz/)
[22] http://www.pooh.cz/rss/
[23] http://www.websky.cz/rsssky/rsssky.aspx
[24] http://www.minasite.cz)
[25] http://jade.mcli.dist.maricopa.edu/feed/
[26] http://feedvalidator.org
[27] http://www.ilrt.bris.ac.uk/discovery/2000/08/hss/sw.html
[28] http://www.proggle.com/novobot/