Sémantický web a jeho technologie

Petr Matulík, Tomáš Pitner, FI MU

1 Co je sémantický web

Idea *sémantického webu* byla světu poprvé prezentována v květnu roku 2001. Tim Berners-Lee, tvůrce současného webu a ředitel Konsorcia W3C (`http://www.w3.org`) spolu s dalšími spoluautory čtivě, s nadhledem, ale důrazně upozornili v článku [1] v Scientific American na skutečnost, že současná síť WWW je v podstatě jen haldou webových stránek, která neustále roste a ve které je stále složitější nalézt relevantní informace. Východisko z tohoto chaosu spatřují v postupném přerodu stávajícího webu v tzv. *sémantický web*, jehož uživatelská představa je vyjádřena hned v úvodu jejich článku. Hovoří se tam o světě, kde jsou inteligentní (často mobilní) zařízení schopná navzájem automaticky komunikovat, jednat a řešit za člověka nejrůznější praktické úlohy, jejichž řešení se opírá o informace, znalosti a jejich důvěryhodné sdílení. Popusťme uzdu fantazii a podívejme se, jak by v představách vizionářů sémantického webu mohla vypadat nepříliš vzdálená budoucnost (překlad z [1]):

> Úvodní příběh začíná vyzváněním telefonu, které vpadlo do libozvučných tónů jedné písně Beatles. Telefon je však inteligentní a ví, že jeho pán nikdy nechce být během hovoru rušen. Přístroj proto vyšle všem blízkým zařízením s vlastností nastavení hlasitosti zprávu, aby se ztišila. Teď už nic nebrání přijetí hovoru. Volá sestra: „Matka potřebuje pravidelné schůzky s fyzioterapeutem, zařizuji to …“ Sestra, zdržující se dosud v ordinaci lékaře, instruuje svůj přenosný webový prohlížeč a dává mu pokyn, aby si od lékařova „agenta“ opatřil údaje o rehabilitační péči předepsané matce. Vše se automaticky konfrontuje s možnostmi jejího zdravotního pojištění a samozřejmě též s geografickými údaji – péče musí být dostupná v okruhu 20 mil. Kromě toho by fyzioterapeut měl být hodnocen alespoň jako velmi dobrý – pochopitelně důvěryhodnou ratingovou agenturou. Fyzioterapeuti připadající v úvahu mají své časové rozpisy, které je třeba dát dohromady s „našlapaným“ pracovním diářem sourozenců pečujících o matku. I o časové sladění se však může postarat technika, máli potřebné vstupní údaje a disponuje-li nezbytnými znalostmi … Předběžný návrh návštěv u fyzioterapeuta zpracovaný počítačem se ale bratrovi nelíbí: musel by jet s matkou ve špičce vozem přes celé město. Stahuje si proto od sestry všechny dosud získané údaje z vyhledávání a jeho vlastní agent (jistěže softwarový) se pokouší za zpřísněných podmínek ohledně místa a času o nalezení jiného terapeuta. Výměna dat (zejména takto citlivých) se sestrou probíhá na základě vzájemné důvěry a zabezpečené komunikace. Bratrův agent uspěl – má řešení, kvůli němuž stačí odsunout pár méně významných schůzek … A pohádka je téměř u konce – pro někoho možná děsivá, pro jiného lákavá vize světa, kde je běžná výměna kvalifikovaně reprezentovaných informací s jednotně chápaným významem a kde je možné nad těmito informace i strojově „uvažovat“ a řešit praktické problémy. Tato vize budoucnosti se jmenuje sémantický web.

Poněkud suchopárněji lze sémantický web [2] charakterizovat takto: Sémantický web je rozšířením současného webu, v němž informace mají přidělen dobře definovaný význam lépe umožňující počítačům a lidem spolupracovat. Sémantický web představuje reprezentaci dat na WWW. Je založen na technologii Resource Description Framework (RDF), která integruje širokou škálu aplikací využívajících syntaktický zápis v XML a identifikátory URI pro pojmenovávání.

Jde tedy o to, aby data prezentovaná na internetu měla přesně definovaný význam a dovolovala do značné míry automatizované (strojové) zpracování.

2 Hlavní prvky sémantického webu

Jedním ze základních kroků k vytvoření sémantického webu je *konceptualizace* dat dostupných na internetu. Jedním z klíčových nástrojů konceptualizace jsou *ontologie*. Ontologie lze charakterizovat jako formalizované reprezentace

znalostí určené k jejich sdílení a znovupoužití. Ontologie jsou často doménového (oborového) zaměření a bývají konstruovány jako *pojmové (konceptuální) hierarchie* nebo *sítě*. Přehledným úvodním zdrojem informací o ontologiích je [3].

(Polo)automatizované zpracování informací v sémantickém webu může být realizováno pomocí softwarových *agentů*, což jsou do určité míry autonomní inteligentní programové komponenty pohybující se obvykle v distribuovaném prostředí a schopné realizovat „na účet toho, kdo je pověřil" požadavky na vyhledávání informací, realizaci transakcí apod.

Důležitým předpokladem sémantického webu je rovněž *standardizovaný popis webových zdrojů*. Zdrojem se v této souvislosti rozumí cokoliv, co je dosažitelné prostřednictvím sítě WWW, tedy textové dokumenty, obrázky, videosekvence, zvukové soubory apod. Každý zdroj by byl vybaven stejnými charakteristikami (autor, typ zdroje, klíčová slova atd.), což by umožnilo uživatelům internetu pracovat se sítí WWW jako s relační databází a dotazovat se na její obsah prostřednictvím jazyků podobných SQL. Významným důsledkem by například byla velmi vysoká přesnost a relevance odpovědi na vyhledávací dotaz, což znamená, že by byl uživateli při vyhledávání určité informace vrácen seznam všech zdrojů, které se této informace týkají, a žádný zdroj navíc.

3 Metadata

Metadata mohou být stručně definována jako (strukturovaná) *data o datech*. Zachycují obsah, kontext a strukturu dat, která popisují. Síťová metadata, na která se zaměříme, jsou nejčastěji zapisována prostřednictvím XML [4], které svými vlastnostmi nejvíce odpovídá požadavkům na otevřenost, přenositelnost a interoperabilitu formátu pro výměnu a ukládání dat.

Pro vyjádření vztahů mezi jednotlivými metadatovými prvky a schémata byl navržen standard RDF a skutečné zachycení sémantiky popisovaných dat je zajištěno prostřednictvím *klasifikačních schémat* a *řízených slovníků*. Než se dostaneme k vysvětlení těchto pojmů, představíme si nejprve samotný rámec RDF.

4 RDF

Technologickým základem sémantického webu by se podle organizace W3C měl stát její standard RDF, Resource Description Framework. Podle oficiální definice jde o *obecný rámec pro popis, výměnu a znovupoužití metadat*.

Rámec RDF poskytuje jednoduchý model pro popis zdrojů, který není závislý na konkrétní implementaci. Datový model RDF zjednodušeně řečeno umožní specifikovat trojice {zdroj, vlastnost, hodnota vlastnosti} s významem: „Daný *zdroj* má danou *hodnotu* dané *vlastnosti*."

Trojice jsou v oficiální terminologii nazývány *tvrzení* a v rámci daného tvrzení je zdroj *subjektem*, vlastnost *predikátem* a hodnota vlastnosti *objektem*. Hodnotou vlastnosti může být buď řetězec znaků (literál) nebo jiný zdroj. Tento abstraktní datový model může být využit v mnoha oblastech různými způsoby. Příkladem je vytváření katalogů popisujících obsah, usnadnění sdílení a výměny znalostí, podpora tzv. *důvěryhodného webu* nebo přiřazení sémantiky webovým zdrojům. My se budeme věnovat pouze poslednímu zmíněnému případu, protože právě ten je pro sémantický web klíčový.

4.1 Reprezentace RDF v XML:

Datový model RDF lze reprezentovat například prostřednictvím grafů či trojic, pro vyjádření sémantiky webových zdrojů se však jako nejvhodnější jeví XML syntaxe RDF (dále jen RDF/XML). RDF/XML v podstatě umožňuje přiřazení vybraných vlastností určitému webovému zdroji, případně vyjádření vztahů mezi takovými zdroji. Webovým zdrojem pak rozumíme každý objekt, kterému je přiřazen jednoznačný identifikátor ve formátu URI (*Uniform Resource Identifier*, popsaný např. v RFC 1630) a který je dostupný prostřednictvím služby WWW.

Obrázek na str. 3 ilustruje použití RDF k popisu článku publikovaného na webu. Podrobnější informace k syntaxi RDF lze najít v posledních specifikacích dostupných na [5] nebo v četných tutoriálech, např. [6].

```
<rdf:RDF xmlns:rdf=‘‘http://www.w3.org/1999/02/22-rdf-syntax-ns\#’’
         xmlns:vcard=‘‘http://www.imc.org/vcard/3.0/’’
         xmlns:dc=‘‘http://purl.org/dc/elements/1.1/’’>
   <rdf:Description
        about=‘‘http://www.sport.cz/fotbal/2003/12/04/spartachelsea.html’’>
      <dc:Title>Sparta  Chelsea 0:1</dc:Title>
      <dc:creator rdf:resource=‘‘http://www.sport.cz/authors/PetrMatulik’’/>
      <dc:date>2003-12-04</dc:date>
   </rdf:Description>
   <rdf:Description about=‘‘http://www.sport.cz/authors/PetrMatulik’’>
      <vcard:fn>Petr Matulík</vcard:fn>
      <vcard:email>petrmatulik@email.cz</vcard:email>
</rdf:Description>
</rdf:RDF>
```

5 Klasifikační schémata

RDF/XML je vlastně jen *metajazykem* (rámcem) pro popis dalších jazyků (podobně jako XML). Umožňuje využít jednotným způsobem *klasifikačních schémat*, tedy souborů vlastností s definovanou sémantikou a omezeními kladenými na možné hodnoty těchto vlastností. Výše uvedená ukázka používala jako klasifikační schémata standardy *vCard* a *Dublin Core* (DC). Podrobnější údaje o Dublin Core lze najít na webu *Dublin Core Metadata Initiative* [7]; český překlad specifikace DC je k dispozici na `http://www.ics.muni.cz/dublin_core/`. Stručné přiblížení schématu vCard je možné najít například v internetové encyklopedii *Whatis* (`http://whatis.techtarget.com/definition/0,,sid9_gci213281,00.html`).

Začlenění klasifikačních schémat, které kromě vlastností často obsahují i definici hierarchie tříd a objektů nesoucích dané vlastnosti, do struktury RDF/XML je zajištěno prostřednictvím *jmenných prostorů XML*. Každá vlastnost použitá v RDF/XML dokumentu musí patřit do nějakého jmenného prostoru a každý jmenný prostor musí mít vlastní jednoznačný identifikátor ve formě URI. URI většinou ukazuje na místo, kde je uloženo tzv. RDF schéma, což je strojově čitelná definice daného klasifikačního schématu implementovaná opět v RDF/XML. Například naše ukázka metadatového popisu článku obsahovala elementy ve jmenných prostorech s URI `http://www.imc.org/vcard/3.0/` (prefix `vcard`) a `http://purl.org/dc/elements/1.1/` (prefix `dc`).

Alternativou k RDF schématům jsou novější jazyky pro popis ontologií DAML+OIL [8] a OWL [9], které jsou složitější, ale mají větší vyjadřovací sílu. K těmto standardům se podrobněji vrátíme v příštím pokračování seriálu o sémantickém webu.

Literatura

[1] Tim Berners-Lee, James Hendler, Ora Lassila. *The Semantic Web.* Scientific American, May 2001. `http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2`

[2] Semantic Web. `http://www.w3.org/2001/sw`

[3] Svátek, V. *Ontologie a WWW*. Tutoriál konference DATAKON 2002. Též dostupný na `http://www.datakon.cz/datakon02/d02_svatek.pdf`

[4] XML. `http://www.w3.org/XML`

[5] Resource Description Framework (RDF). `http://www.w3.org/RDF`

[6] Miloslav Nic. RDF Tutorial. `http://zvon.org/xxl/RDFTutorial/General/book.html`

[7] Dublin Core Metadata Initiative. `http://dublincore.org`

[8] DAML+OIL, `http://www.w3.org/TR/daml+oil-reference`

[9] OWL, `http://www.w3.org/TR/owl-ref` □