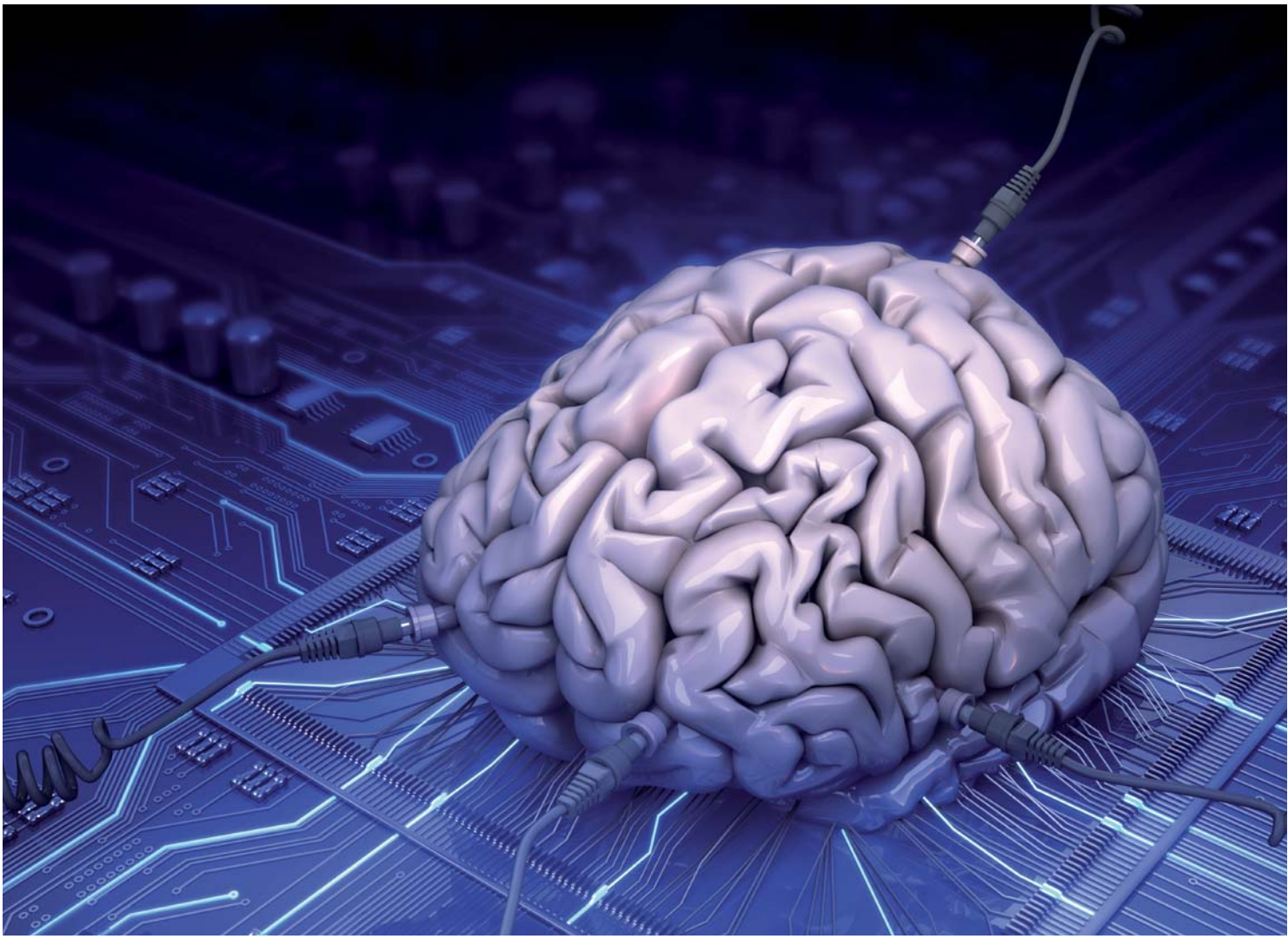


Sémantický web: Google se učí kombinovat

Při současném datovém chaosu na webu zažívají vyhledávací algoritmy kruté časy. Strukturované zálohování by však v budoucnu mělo vést k **PRAKTICKÝM VYHLEDÁVACÍM** výsledkům.

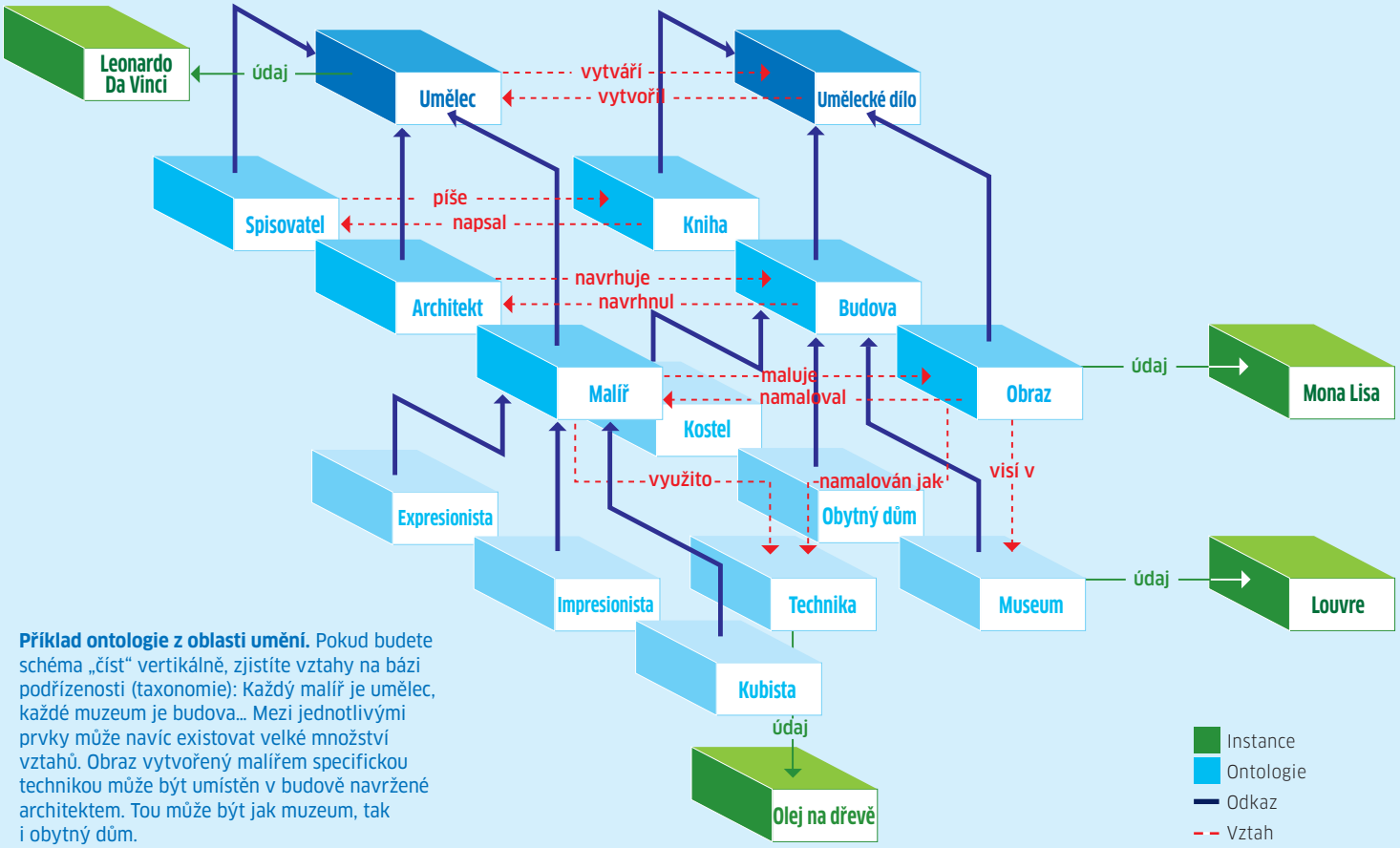
STEFAN MARTIN

V roce 2004 měl vyhledávač Google ve svých databázích „indexováno" osm miliard stránek, v loňském roce oznámil prolomení další „bariéry": v létě roku 2008 překročil počet indexovaných stránek 1 bilion (ano, jednička a dvanáct nul!). Když k tomuto obrovskému balíku informací přidáte ještě videa a obrázky, určitě vás nepřekvapí další fakt: internet je na pokraji zvratu. Miliardy informací jsou stále obtížněji přístupné, chybí jejich propojení a problematická je i možnost jejich počítačového zpracování. Tento problém se v současné době řeší na mnoha konferencích a blozích i v celé řadě článků v časopisech různého zaměření. Řešení však existuje a skrývá se pod názvem „sémantický web". Co si pod tímto pojmem představit? Sémantický web je charakterizován určitými znaky, které jsou na míle vzdáleny od „klasického" webu zaměřeného na dokumenty. Dříve než si je ale popíšeme, pojďme se podívat, jak sémantické weby vznikly. Problém „s informacemi" se řešil na konferencích typu ISWC (International Semantic Web Conference) nebo SemTech (Semantic Technology Conference) a právě tam se objevily první výsledky aktivního výzkumu. Poslední slovo měla organizace W3C (WWW Consorcium), která definovala a stanovila vhodné standardy.

Aktuální scénář: Stroje nemohou číst informace z webu

V současné době je internet obrovským zdrojem informací, které jsou dostupné ve

Ontologie: komplexní vztahy přehledněji



Příklad ontologie z oblasti umění. Pokud budete schéma „číst" vertikálně, zjistíte vztahy na bázi podřízenosti (taxonomie): Každý malíř je umělec, každé muzeum je budova... Mezi jednotlivými prvky může navíc existovat velké množství vztahů. Obraz vytvořený malířem specifickou technikou může být umístěn v budově navržené architektem. Tou může být jak muzeum, tak i obytný dům.

formě webů vytvářených lidmi tak, aby je mohli používat jiní lidé. Díky tomu mohou lidé tyto informace nejen snadno pochopit, ale také si je spojit s dalšími informacemi. Problém však nastává v případě „strojů": pro počítače je těžké vyfiltrovat z bludiště internetu určitou informaci. Současné řešení pomalu přestává stačit – vyhledávače (Google, Yahoo...) „pouze" nabízí pomocí specifických statistických metod (viz článek o algoritmech hledání Googlu v Chipu 10/2008) seznamy „zásahů" pro konkrétní zadané pojmy. Problém ale spočívá v tom, že tento typ hledání jen málokdy nabídne komplexnější informace. Vyhledávání vhodné odpovědi obvykle také „přináší" zbytečnou práci v podobě otvírání mnoha stránek s neužitečnými informacemi.

Existuje však i další problém: Nyní jsou jednotlivé informace na webu formulovány rozdílně, takže není možné jejich přímé porovnání. Výsledkem je, že informace na internetu není snadné shrnout a jednotně prezentovat. Hledáte-li pojem „semantic web", najdete množství stránek s dublujícími, doplňujícími se, či dokonce protikladnými informacemi. Důsledkem toho je ve finále nemožnost najít jednoznačné informace a nutnost získávat informace z jednotlivých výsledků hledání.

A to navíc nemluvíme o otázkách formulovaných „jak starý je václav havel". Zjistit odpověď na otázky tohoto typu vždy vyžaduje více práce při zpracování nalezených informací. Nejlepším řešením je prozatím upravit frázi pro vyhledávání do poněkud kostrbaté podoby „václav havel narozen".

Řešení: Organizace webového obsahu a jeho propojení

Termín „sémantický web" znamená uspořádání dat takovým způsobem, aby je počítače mohly logicky zpracovat. Aby to bylo možné, měly by sémantické webové stránky nabízet informace o vztazích mezi informacemi jako metatext.

Pro řešení tohoto problému definovala organizace W3C sérii otevřených standardů, přičemž v tomto případě hrají nejdůležitější

INFO

Co je W3C

W3C je organizace, která vytváří webové standardy a pravidla, dostupné poté jako Recommendation (doporučení). Ředitelem konsorcia W3C byl (již od jeho založení v roce 1994) známý Tim Berners-Lee, který během své práce pro výzkumné centrum CERN v roce 1989 „vynalezl" World Wide Web.

Na vývoji standardů spolupracují americké laboratoře univerzity MIT a Computer Science and Artificial Inteligence Laboratory (CSAIL) spolu s Evropským výzkumným centrem pro informatiku a matematiku (ERCIM) se sídlem ve Francii a s japonskou univerzitou Keio. W3C funguje na bázi příspěvků jednotlivých členů, výzkumných fondů a jiných veřejných a soukromých finančních zdrojů. Mimochodem, vizi „sémantického webu" formuloval Tim Berners-Lee již v roce 1999.

Zdroj: www.w3c.org

roli standardy XML, RDF (schéma), OWL a SPARQL. Ty umožňují uložení informace formulované sémanticky ve formě ontologií (vztahy mezi pojmy a významy) a taxonomií (rozdělení pojmů podle určitých pravidel). Více informací najdete kapitole Správné nástroje pro programátory – RDF". U jazyka SPARQL existuje dokonce plně vyvinutý dotazovací jazyk pro RDF ontologie...

Metody: Tři cesty k vytvoření sémantického webu

Jak tedy změníme web zaměřený na „dokumenty" na web zaměřený na obsah? První možností je změnit původní informaci takovým způsobem, aby byla strukturována sémanticky. Technologie umožňující takový převod je zatím ve fázi výzkumů, což není zrovna důvod k optimismu. Při hledání informací o zmiňovaném výzkumu můžete například narazit na pojem „Natural Language Processing" (přirozené jazykové zpracování), což je metoda, která umožňuje analyzovat text imitací lidského čtenáře. V tomto případě je nejprve text rozčleněn na věty a pomocí již známé struktury (podmět – sloveso – předmět) může být významový obsah věty jednodušeji využit. Vyhledávání pak může identifikovat osoby, místa a události a propojit je mezi sebou.

Další přístup je znám pod obchodním jménem „mikroformáty" (**http://microformats.org**). Ty v červnu 2008 oslavily tři roky své existence. Myšlenkou je rozšíření existující (X)HTML stránky manuálně, pomocí zvláštních prvků standardu (X)HTML, a tak je učinit „čitelnými" i pro počítače.

Jako příklad lze také uvést specifické formáty, které definují zadávání kontaktů, rozvrhů a záložek. Známé platformy typu Facebook, Flickr, Google Maps či Yahoo už tyto formáty používají. Pomocí RDFa už dokonce i W3C nabízí možnost vkládání informací, které budou čitelné pro „počítač", na klasické (X)HTML stránky. Zmiňované přístupy jsou však bohužel navzájem odlišné – vývojáři mikroformátů postupují podle pravidla „co nejvíce informací s co nejmenšími náklady".

Řešení by se opět mohlo objevit od organizace W3C, která se pokouší definovat obecnou „konstrukci", jež by mohla být využívána k vytváření metadat. V praxi je však tento přístup dražší, a je tedy otázkou, jestli se v praxi uchytí. Nicméně RDFa je již na dobré cestě k označení „Doporučení", a má tudíž v dlouhodobém horizontu i potenciál k nahrazení mikroformátů. Třetí cestou k sémantickému webu je používání speciální standardů.

Správné nástroje pro programátory sémantických webů

Které standardy se nyní používají pro vytváření „sémantických" webů? Představíme vám ty nejdůležitější programovací a vyhledávací „nástroje".

EXTENSIBLE MARKUP LANGUAGE (XML)

Základním standardem pro sémantický web je XML. Zjednodušuje strukturování informací a umožňuje snadnou konverzi do jiných formátů. Jeho syntaxe je poměrně přísná, určuje pravidla pro strukturu znaků a znakových řetězců, které musí dokument obsahovat.

Známým příkladem aplikace je jazyk (X)HTML, což je nástupce jazyka HTML na bázi XML. Technologickým základem pro vytváření strukturovaných dokumentů by měly být standardy RDF a OWL.

RESOURCE DESCRIPTION FRAMEWORK (RDF)

RDF je formální jazyk pro popis strukturované informace. V porovnání s HTML se nezajímá o „korektní reprezentaci obsahu", ale spíše o možnost dalšího zpracování informace či kombinace s jinými informacemi.

RDF dokument je popsán jako orientovaný graf, který je opět definován jako shluk uzlů, které jsou spojeny prostřednictvím hran. Každá hrana a každý uzel nesou určitou identifikaci, tzv. URI (Uniform Ressource Indentifier). Tato řada znaků indikuje fyzický či abstraktní zdroj a v podstatě se skládá ze syntaxe „schéma: sekvenční část".

URI mohou být webové adresy (**www.chip.cz**) nebo e-mailové adresy (mailto:info@chip.cz). Koncept URI může být však použit zcela odděleně od webu jako obecný mechanismus pro generování přesné identifikace. URI v RDF dokumentech tudíž obvykle neodkazují na existující webové stránky.

Google indexuje více než 1 bilion stránek.

RDF graf může být plně popsán pomocí specifikace svých hran. Každá hrana odkazuje na jednu z tzv. trojic (podmět, přísudek, předmět) a to umožňuje přenášet graf do zápisu typu XML. Mimochodem, existují další formy RDF, jako například „Triple Syntax Turtle", ty se ale v důsledku dominance XML používají jen zřídkakdy. RDF(S) – „S" zde znamená schéma – představuje rozšíření RDF a umožňuje specifikaci terminologické znalosti či znalosti schématu. Zde RDF nejdříve popíše předměty a jejich vzájemné vztahy. RDF(S) může tyto předměty dodatečně rozřadit do tříd. Tato třída potom dále může být částí jedné nebo více podřazených nebo nadřazených tříd. Výsledkem toho je jednoduchá reprezentace například „tříd", protože podobnou strukturu má i samotná taxonomie. Vtip je v tom, že předmět není pouze ukázkou své třídy, ale také ukázkou všech nadřazených tříd (viz tabulka).

Taxonomie: Síla hierarchie

Taxonomie Taxonomie je pokus o systematické a hierarchické uspořádání existujících informací. Známý příklad – taxonomie ve světě zvířat, která se v průběhu doby vyvíjela. Původní Brehmův seznam obsahoval pouze několik úrovní, v moderním systému třídění jich už ale najdete sedmnáct.

- **Říše** Fauna
- **Podříše** Mnohobuněční
- **Třída** Praví mnohobuněční
- **Skupina** Kruhoústí
- **Kmen** Strunatci
- **Podkmen** Obratlovci
- **Třída** Savci
- **Podtřída** Vyšší savci
- **Nadřád** Nadprimáti
- **Řád** Primáti
- **Podřád** Vyšší primáti
- **Nadčeleď** Opice starého světa
- **Čeleď** Opice
- **Rodina** Lidoopi
- **Druh** Šimpanzi
- **Třída** Běžní šimpanzi
- **Podtřída** Šimpanz západní

ZDROJ: WIKIPEDIA

Vlastnost (property) reprezentuje další rozšíření. Popisuje vlastnost předmětu. Například vlastnost „je šťastně ženat“ může být podřazena jedné či více vlastnostem, tedy například vlastnosti „je ženat“. V případě vlastností může být pole definice stejně jako pole hodnoty omezeno na určité třídy.

WEB ONTOLOGY LANGUAGE – (OWL)

Základy OWL jsou vlastnosti a třídy známé už z případu RDF(S), stejně jako deklarované RDF příklady tříd. Avšak OWL dokáže vytvořit komplexní vztahy i mezi těmito třídami a vlastnostmi. Třídy navíc mohou být deklarovány jako oddělené či stejné. Třídy dále můžete uzavírat a určovat tak, které prvky patří k určité třídě. Dalšími možnostmi pro zjemnění systému jsou některé operace, například kombinování, oddělování a skládání. Vlastnosti třídy mohou být omezeny či vyjmuty vymezením rolí. Další informace poskytují věty obsahující určitý počet vlastností, jako je např. „alespoň jeden“, „všechno“, „alespoň X“ či „maximálně Y“. OWL je k dispozici ve třech „subjazycích“, které jsou založeny na společném základě: v OWL Lite, v OWL DL a v OWL Full. Posledně jmenovaný subjazyk nabízí největší volnost ve smyslu popisování znalosti, to ale zvyšuje problémy, které právě vedly k vývoji OWL DL a OWL Lite. OWL DL je vytvořen tak, že se tyto problémy neobjevují. Je tedy téměř kompletně podporován současnými aplikacemi (viz sekce Applications Protége, Pellet, Kaon 2). OWL Lite nabízí pouze nejdůležitější elementy jazyka, a je tudíž v praxi méně důležitý.

SIMPLE PROTOCOL AND RDF QUERY LANGUAGE (SPARQL)

Jak z výše popsaných ontologií získat informace? RDF k tomu má SPARQL, nově vytvořený standard pro dotazování RDF a pro zobrazování výsledků. V tomto případě obsahuje dotaz tři základní části:

1. PREFIX – používá se k selekci jmenného prostoru (namespace),
2. SELECT – používá se k definování zobrazovacího formátu,
3. WHERE – používá se pro formulování aktuálního dotazu.

Filtrovací podmínky je možné formulovat v části WHERE, přičemž tyto podmínky ověří hodnoty vlastností (viz výpis 2). Je možné také hledat předměty, které mají určitou vlastnost. Výsledky mohou být dodatečně vytříděny pomocí SORT BY a použitím čísla OFFSET mohou být omezeny do specifického čísla předmětu.

W3C ještě nemá dotazovací jazyk pro OWL hotový, ale existují iniciativy, které si

VÝPIS 1

```
<?xml version='1.0' encoding='utf-8'?>
<rdf:RDF xmlns:rdf ='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:rdfs='http://www.w3.org/2000/01/rdf-schema#'
xmlns:ns ='http://example.org/'>

<rdf:Description rdf:about='http://example.org/ tistene medium'>
<rdf:type rdf:resource='http://www.w3.org/2000/01/rdfschema#
Class'/>
</rdf:Description>

<rdf:Description rdf:about='http://example.org/casopis'>
<rdfs:subClassOf rdfs:resource='http://example.org/tistene medium'/>
</rdf:Description>

<rdf:Description rdf:about='http://example.org/Chip'>
<rdf:type rdf:resource='http://example.org/casopis'/>
</rdf:Description>

<rdf:Description rdf:about='http://example.org/SemanticWeb'>
<ns:ArticleIn rdf:resource='http://example.org/Chip'/>
</rdf:Description>

</rdf:RDF>
Query:
WHERE
Result:
</div>
</div>
</p>
<p>
</p>
<p>
</p>
</div>
```

VÝPIS 2

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

_:a foaf:name „Johnny Lee Outlaw“ .
_:a foaf:mbox <mailto:jlow@example.com> .
_:b foaf:name „Peter Goodguy“ .
_:b foaf:mbox <mailto:peter@example.org> .
_:c foaf:mbox <mailto:carol@example.org> .

PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox
{ ?x foaf:name ?name .
?x foaf:mbox ?mbox }

name mbox
„Johnny Lee Outlaw“ <mailto:jlow@example.com>
„Peter Goodguy“ <mailto:peter@example.org>
```

VÝPIS 3

```
<div xmlns:dc="http://purl.org/dc/elements/1.1/">
<h2 property="dc:title">Semanticky Web</h2>
<h3 property="dc:creator">Chytry Martin</h3>

<div about="http://example.com/sw/dc.jpg">
<img src="http://example.com/sw/dc.jpg" />
<span property="dc:title">Dublin Core</span>
autor <span property="dc:creator">Chytry Martin</span>.
</div>

</div>
```

VÝPIS 4

```
<div typeof="foaf:Person" xmlns:foaf="http://xmlns.com/foaf/0.1/">
<p property="foaf:name">
Alice Vokounova
</p>

<p>
Email: <a rel="foaf:mbox" href="mailto:alice@example.com">alice@
example.com</a>
</p>

<p>
Phone: <a rel="foaf:phone" href="tel:+420-617-555-7332">+420
617.555.7332</a>
</p>

</div>
```

ZDROJ: HTTP://WWW.W3.ORG/TR/XHTML-RDFA-PRIMER/#ID84912

přejí učinit SPARQL dostupným pro OWL ontologie. Na rozdíl od RDF nabízí OWL DL výrazy jako součást jazyka. Ten může být používán k nalezení všech prvků, které odpovídají popisu třídy.

RESSOURCE DESCRIPTION FRAMEWORK IN ATTRIBUTES (RDF A)

RDFa se stejně jako mikroformáty používá k aktualizaci existujících XHTML stránek s metadaty. K tomu RDFa používá jmenné prostory a slovník pro indikaci informace.

Organizace „Dublin Core Matadata Initiative“ (**http://dublincore.org**) publikovala v roce 1994 pod názvem „Dublin Core“ jeden z nejpoužívanějších jmenných prostorů. Jde o často používaný zdroj, který obsahuje informace o dokumentech, například jméno autora, název nebo datum vytvoření.

Nyní se v RDFa používá následovně: „xmlns:dc“ odkazují přímo na jmenný prostor „Dublin Core“ a „DC:title“ je vlastně zkrácená verze zadání **http://purl.org/dc/elements/1.1/title**

a ukazuje název dokumentu. Položka „dccreator“ ukazuje autora. Tato informace může být prezentována také jako RDF trojice (document, dc:title, title) a (dokument,dc:creator, author). Položka „about“ poukazuje na část webové stránky.

Příklad výsledku vyhledávání obrázku s detaily o názvu a autorovi můžete vidět na výpisu 3.

Dalším příkladem může být jmenný prostor Friend-of-a-Friend (FOAF, **www.foaf-project.org**), který obsahuje slovník pro popsání kontaktních informací. Zde například položky „foaf:name“, „foaf:mbox“ a „foaf:phone“ určují kontaktní informace tak, aby byly „čitelné“ i pro počítačové zpracování (viz výpis 4).

Použití: sémantické editory a vyhledávače

Jak ale mohou být ontologie vytvářeny a které nástroje lze použít? Nejznámější editor ontologií je Protégé (**http://protege.stanford.edu**) – je k dispozici zdarma a je to open-source. Může být používán ke správě ontologií, jejich vizualizaci a jejich exportu, stejně tak jako k přiřazení jednotlivých položek do tříd. Implikační nástroje Pellet (**http://clarkparsia.com/pelle**t) a KAON2 (**http://kaon2.semanticweb.org**) umožňují derivovat nové „tvrzení“ z ontologií. Pellet je také zdarma a open-source, zatímco KAON2 je poskytován zdarma pouze pro nekomerční použití.

Editory a implikační nástroje vytvářejí základ pro sérii aplikací, které používají technologii sémantického webu ve formě „start-upů“. Otevřená databáze Freebase

VÝPIS 5

```
No

I used the following facts to provide this answer:

* thing that was created is the left class of ‚is older than‘
* thing that was created is the right class of ‚is older than‘
* the 26th of October 1947 is the birthdate of Hillary Clinton
(endorse)  (contradict)
* the 4th of August 1961 is the birthdate of Barack Obama
(endorse)  (contradict)
```

(ZDROJ: HTTP://BETA.TRUEKNOWLEDGE.COM/ANSWER.PHP?INPUT=IS+BARACK+OBAMA+OLDER+THAN+HILLARY+CLINTON%3F)

(**www.freebase.org**), vytvořená firmou Metaweb technologies, má za cíl vytvářet strukturované informace dostupné tak, aby počítače – stejně jako lidé – mohly tuto informaci optimálně využívat. Základem pro to jsou texty ze známých „otevřených“ databází typu Wikipedia nebo Musicbrainz (**http://musicbrainz.org**).

Stránky jsou analyzovány z hlediska svého obsahu a poté přidány k ontologii. Tímto způsobem má Freebase nyní více než tři miliony položek (750 tisíc osob, 450 tisíc míst, 50 tisíc firem a 40 tisíc filmů). Freebase ale využívá svoji vlastní ontologii a dotazovací jazyk, který se od jazyka SPARQL liší.

Další otevřená databáze založená na RDF je DBpedia (**http://dbpedia.org**). Ta také využívá textů z Wikipedie, která obsahuje téměř 2,5 milionu položek (108 tisíc osob, 392 tisíc míst, 57 tisíc hudebních alb, 36 tisíc filmů...). I na tato data se lze dotazovat jazykem SPARQL. Dodatečně je k dispozici řada aplikací, které se v kontextu výzkumu mohou ukázat jako užitečné.

Sémantický vyhledávač Powerset (**www.powerset.com**) prohledává dva zdroje informací. V první řadě používá sémanticky připravené texty z Wikipedie a v poslední době i z Freebase, které jsou již také „počítačově zpracovatelné“. Powerset umožňuje zadávat dotazy v přirozeném jazyce – například na dotaz „how old is George Bush?“ dostanete odpověď „June 12, 1924 (84 years ago)“, spolu se seznamem dokumentů obsahujících zdrojové informace.

Dalším sémantickým vyhledávačem je True Knowledge (**www.trueknowledge.com**). Vzhledově je podobný vyhledávači Powerset a také on dokáže nabídnout stejné výsledky. Důležitým rozdílem je ale podpora pro odlišný dotazovací jazyk – například: „Je Barrack Obama starší než Hillary Clinton?“ (viz výpis 5)

Dalším zajímavým webem je prohlížeč dokumentů Open Calais (**www.opencalais.com/DocViewer**), který hledá na klasických webech informace v podobě objektů, faktů a incidentů a exportuje je do formátu RDF. Jeho praktické využití si můžete vyzkoušet pomocí doplňku pro Firefox nazvaného Gnosis (**http://addons.mozilla.org/firefox/3999**).

Vyhlídky: sémantické weby budou přibývat

Sémantické weby nabízejí efektivní pomoc při vyhledávání relevantních informací z různých zdrojů. Standardy pro popsání informací již existují a jsou v praxi vyzkoušeny, k dispozici jsou i nástroje pro zpřístupnění těchto informací. Jediným problémem je prozatím příliš malý objem zdrojů dat, ale i zde se situace zlepšuje. Navíc se objevují nástroje schopné analyzovat klasické weby sémanticky (případně tyto informace v tzv. metadatech analyzovat pomocí RDF).

A právě dostupnost rozmanitých zdrojů dat je pro fungování sémantického webu důležitá. Zjednodušuje totiž vyhledávání určitého obsahu, a navíc napomáhá k lepší provázanosti mezi nalezenými informacemi. Sémantický web je tedy prvním a důležitým krokem ke zpřístupnění informací hledaných „klasickými otázkami“, a tedy i prvním akordem pro očekávanou hudbu budoucnosti.

AUTOR@CHIP.CZ

INFO ODKAZY

STANDARDY
W3C: www.w3.org
Mikroformáty: http://microformats.org
Dublin Core Metadata Initiative: http://dublincore.org
Friend of a Friend (FOAF) Project: www.foaf-project.org

OWL EDITORY
Protégé: http://protege.stanford.edu

OWL NÁSTROJE
Pellet: http://clarkparsia.com/pellet
KAON2: http://kaon2.semanticweb.org

DATABANKY
Freebase: www.freebase.com
DBpedia: http://dbpedia.org
MusicBrainz: http://musicbrainz.org

SÉMANTICKÉ VYHLEDÁVAČE
Powerset: www.powerset.com
True Knowledge: www.trueknowledge.com

TEXTOVÁ ANALÝZA
Open Calais: www.opencalais.com
Gnosis: https://addons.mozilla.org/firefox/3999