

Metadata, sémantika a sémantický web

Ing. Vilém Sklenák, CSc.

Vysoká škola ekonomická, fakulta informatiky a statistiky,
katedra informačního a znalostního inženýrství

sklenak@vse.cz

INFORUM 2004: 10. konference o profesionálních informačních zdrojích
Praha, 25.–27. 5. 2004

Abstrakt

Metadata a jejich využití pro zachycení a vyjádření sémantiky jsou nezbytným předpokladem pro fungování sémantického webu. Příspěvek se snaží rekapitulovat aktuální situaci, uvést problémy. Sémantický web byl a je nadějí pro všechny úlohy vyžadující sémantický pohled na webový obsah – jednou z takových oblastí je vyhledávání informací.

1 Úvodem

Metadata jsou dnes často skloňovaným pojmem, a to v řadě různých souvislostí. Jinak se na ně dívá správce webu (použití tagu <META> v záhlavích webových dokumentů, DC, RSS¹ apod.), jinak správce integrovaného informačního systému (slovníky metadat) anebo informační profesionál (DC, MARC21, ...). Proto existují různé definice, a také proto může docházet k různým nedorozuměním. Tak např. v [2] se uvádí pět pohledů na metadata:

1. *data o datech;*
2. *informace o informacích;*
3. *význam nebo sémantika dat;*
4. *zdroj poskytující informace o jiném zdroji;*
5. *popisná informace o webovém zdroji;*

Současně je zřejmé, že metadata jsou klíčovým problémem sémantického webu. Proto se nelze smířit s obvyklým konstatováním, že metadata jsou *data o datech* s mlhavým dovětkem o zvýšení efektivnosti vyhledávání. Kromě toho také platí, že některá data mohou být jak metadaty, tak daty samotnými. Role závisí na způsobu použití. Například bibliografický záznam o knize uložený v katalogu jsou metadata pro uživatele, který vyhledává podklady k nějakému tématu. Současně se však jedná také o data pro vlastní knihovní systém, který je používá např. k různému shlukování: knihy téhož autora, téhož názvu, téhož zaměření, téže lokace apod.

¹Rich Site Summary

2 Metadata obecně

Obecné vymezení „data o datech“ je pravdivé, ale je příliš zjednodušující a zastírající, že v mnoha případech obsahují metadata mnohem víc, např. informace o tom, jak jsou data uložena nebo spravována, nebo o jejich významu vzhledem ke konkrétnímu způsobu použití. Informace mohou být velmi rozmanité, může jít o popis nebo shrnutí obsahu, jindy může jít o informace nezávislé na obsahu, jako je umístění nebo datum vytvoření.

Použití metadat přináší dvě zásadní výhody:

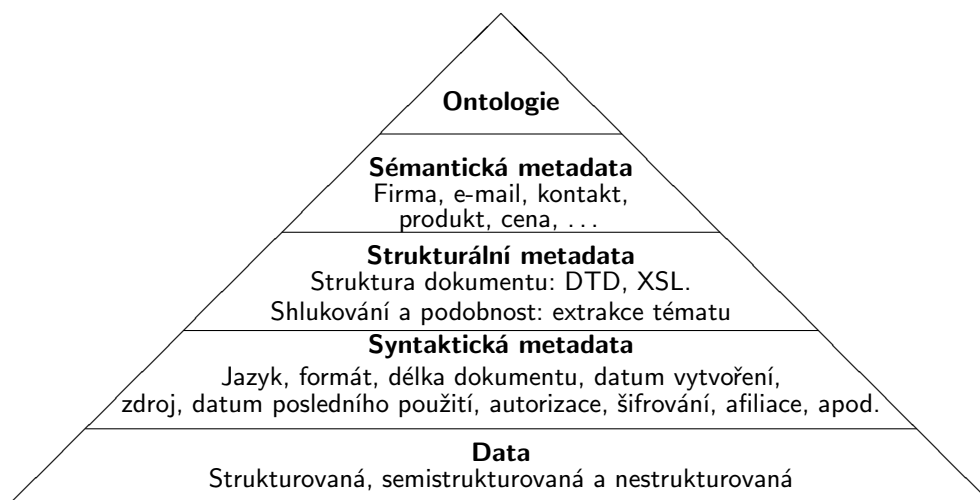
1. Umožňují abstrahovat od detailů reprezentace, jako jsou např. formát nebo organizace dat, a zachycují informační obsah nezávisle na původní formě dat. Toto představuje první krok směrem k redukci informačního zahlcení, protože cílené popisy ve formě metadat jsou obecně mnohem úspornější (ve smyslu počtu bajtů) než data samotná.
2. Umožňují reprezentovat doménové znalosti popisem informační oblasti, k níž přísluší výchozí data. Tato znalost pak může být použita k domněnkám (úsudkům, odvozování) o daných datech. Tím lze omezit informační přehlčení, protože odvozování může posloužit k určení relevance příslušných dat, aniž by bylo nutno k datům přistupovat.

Metadata lze klasifikovat podle různých kritérií. Podle úrovně abstrakce, s níž metadata popisují obsah, lze metadata odlišit následovně:

- **Syntaktická metadata** se soustřeďují na podrobnosti o zdroji dat (dokumentu), a proto poskytují lepší pohled na data. Tento druh metadat dobře poslouží při katalogizaci nebo kategorizaci.
- **Strukturální metadata** se zaměřují na strukturu dokumentu, což pak lze využít při ukládání, zpracování nebo prezentaci, jako je navigace, zjednodušuje vyhledávání informací a umožňuje přehlednější zobrazení. Typickým příkladem je XML schéma.
- **Sémantická metadata** popisují kontextově relevantní informace, přičemž se soustřeďují na doménově specifické elementy založené na ontologii, která je známá nebo srozumitelná uživatelům orientujícím se v dané doméně. Užitím sémantických metadat lze získat smyslupnou interpretaci dat a současně je podporována interoperabilita na vyšší úrovni.
- **Ontologie** představují nejvyšší formu metadat a současně jsou klíčovým principem sémantického webu. Termín *ontologie* pochází původně z filozofie, ve které se používá jako označení pro *nauku o bytí*, tedy o tom „co je“. V souvislosti s vývojem databází a zejména rozsáhlých bází znalostí uložených v nesčetných souborech na webu došlo k posunu významu a především k jeho rozšíření. V novém významu se ontologií myslí zhruba označení pro všechny možné metody získávání znalostí „o tom, co je v datech“ v nejširším slova smyslu. Protože tyto metody jsou vždy spojeny s použitím jistého

(dotazovacího) jazyka a s problémy spojenými s *významem* výrazů, pomocí nichž jsou data popisována a získávána, hovoří se proto o tzv. *konceptualizaci* jednotlivých oblastí, neboli o *dílčích ontologiích*.

V posledních deseti letech vznikla řada definic ontologie. Snad nejlépe vystihuje podstatu ontologií Gruberova definice: *Ontologie je formální, explicitní specifikace sdílené konceptualizace. Konceptualizace* říká, že se jedná o takový abstraktní model výseku reálného světa, který identifikuje relevantní koncepty daného výseku. *Explicitní* znamená, že je jednoznačně definován typ konceptu i podmínky jeho použití. *Formální* poukazuje na to, že ontologie by měla být strojově zpracovatelná. Je možná různá míra formality. Rozsáhlé ontologie, jako je třeba WordNet, obsahují více než 100 000 termínů, které jsou vyloženy pomocí přirozeného jazyka. Na druhé straně existují ontologie, které nabízejí formální axiomatické teorie. *Sdílený* odráží představu, že ontologie zachycuje konsenzuální znalosti, tj. že není omezena na jedince, ale je akceptována šířeji.



Obrázek 1: Typy metadat

3 Sémantika

Současný vývoj webu lze charakterizovat z různých perspektiv:

- **Lokalizace zdrojů** – způsob, jak lidé vyhledávají informace (věci) na webu, se mění od jednoduchého, založeného na volném textu a klíčových slovech, směrem k více sofistikovaným sémantickým technikám [9].
- **Uživatelé** – webové zdroje se vyvíjejí od primární podoby určené pro vnímání člověkem do podoby použitelné jak pro člověka, tak pro počítače.

- **Webové úlohy a služby** – samotný web prochází proměnou, kdy již nebude chápán jen jako místo pro nalezení věcí, ale bude schopen věci dělat.

Všechny tyto nové možnosti závisí podstatným způsobem na myšlence sémantiky – web se vyvíjí od informačních zdrojů, které explicitní sémantiku obsahují velmi málo nebo vůbec, ke zdrojům s bohatou sémantickou infrastrukturou.

Navzdory velmi častému používání pojmu *sémantický web*, rostoucímu počtu článků či monografií, je jeho fungování stále předmětem experimentů. Námětem diskusí a polemik se také někdy stává výklad pojmu samotného. Nesporný je důraz na informační obsah webu, který:

- je *strojově zpracovatelný* a
- je *asociován s významem*.

Označením „strojově“ se zkráceně myslí počítače, resp. software, jež vykonávají určité webové úlohy. Podobné programy se obvykle označují jako *softwarové agenty*, nebo *softwarové roboty*.

Strojově zpracovatelný obsah předpokládá, že stroji je známo, co má dělat s informacemi získanými z webu. Jedna cesta, jak lze dosáhnout tohoto cíle, spočívá v čtení a zpracování strojově srozumitelné specifikace sémantiky informace. Toto je zatím ideál, který překračuje současný stav poznání. Jednodušší alternativu pro vývojáře webových aplikací představuje přímé zapracování znalostí do softwaru, který pak dokáže s informacemi pracovat rozumným způsobem. V tomto druhém případě již lze uvést řadu příkladů užití informací získávaných z webu. Existují obchodní agenty, které dokáží rutinně pracovat s webovým obsahem spojeným se slovy „price“, „weight“, „airport“ apod. Díky vestavěnému „porozumění“ dokáží takové agenty najít obchody nabízející požadované zboží s nejnižší cenou, nejlevnější letenky pro let mezi dvěma městy apod.

Jazyk RDF² byl konsorciem W3C prohlášen za jazyk sémantického webu, proto dochází někdy k neadekvátním závěrům typu „aplikace používá RDF, a proto je to aplikace sémantického webu“. Toto připomíná nedávnou minulost, kdy také mnohokrát byly používány argumenty typu „je to naprogramováno v Lispu nebo Prologu, a proto se jedná o umělou inteligenci“.

3.1 Sémantika?

Co je vlastně sémantika a jak je chápána v konceptu sémantického webu? Tyto dvě otázky jsou určitě zajímavé, nicméně podrobný rozbor překračuje rámec tohoto článku (podrobněji viz třeba [12]), a proto je následující text trochu zkratkovitý.

²Resource Description Framework

3.1.1 Implicitní sémantika

V nejjednodušším případě lze chápat sémantiku jen jako implicitní. Význam pak vychází ze sdíleného porozumění založeného na konsensu lidí. Obecně známým příkladem je používání XML značek, jako jsou *cena*, *adresa* nebo *datum dodání*. Nikde v XML dokumentu, DTD³ nebo XML schématu, není řečeno, co tyto značky znamenají. Nicméně, existuje-li implicitní sdílený konsensus o významu těchto značek, pak mohou lidé zapracovat tuto implicitní sémantiku do webových aplikací. Tímto způsobem obvykle dnes fungují nákupní agenty (shopping agents), když hledají na webu nejlepší obchodní podmínky. Z hlediska perspektivy komerčních aplikací, které automaticky používají webový obsah, jak je obsaženo ve vizích sémantického webu, se jedná asi o maximum možného při současném stavu poznání.

Nevýhoda implicitní sémantiky spočívá v množství mnohoznačností. Lidé se totiž často neshodnou na významu pojmu. Například *cena* může být vyjádřena v různých měnách, může nebo nemusí zahrnovat určité daně nebo přepravní náklady apod. Odstranění nejednoznačnosti je hlavní motivace používání zvláštního jazyka při sestavování právních dokumentů. Cena za identifikaci a vymýcení nejednoznačnosti je příliš vysoká.

3.1.2 Neformální sémantika

Jiným přístupem je explicitní sémantika, která je vyjádřena neformálním způsobem. Například se může jednat o slovník nebo textový dokument. Ze složitosti přirozeného jazyka pak plyne omezená možnost přímého strojového zpracování sémantiky vyjádřené neformálně. Ta je určena především člověku. Na webu lze nalézt hojnost příkladů neformální sémantiky:

- Význam HTML značky `<h2>`, která se používá pro nadpis druhé úrovně.
- Význam výrazů modelovacích jazyků, jako je třeba UML⁴ a původní specifikace RDF schémat.
- Význam údajů Dublin Core.

Typicky je sémantika vyjádřená v neformálních dokumentech napevno zapracována lidmi do softwarových produktů.

Hlavní nevýhodou implicitní sémantiky je prostor pro nejednoznačnosti. Tím se však snižuje důvěra, že dvě rozdílné implementace (třeba RDF schémat) jsou konzistentní a kompatibilní. Implementace se mohou lišit jen v drobných detailech. To však v důsledku může vést k problémům interoperability nebo při změně implementace. Z těchto důvodů

³Document Type Definition

⁴Unified Modelling Language

někdy neformální specifikace nevyhovuje. O to větší motivace pak je spojena s úsilím o vznik formální sémantiky, např. pro UML, RDF nebo DAML+OIL⁵.

3.1.3 Formální sémantika pro zpracování člověkem

O krok dále již směřuje explicitní sémantika vyjádřená formálním jazykem, nicméně je určena pouze pro zpracování člověkem. Lze ji uvažovat jako formální dokumentaci, nebo jako formální specifikaci významu. Některé příklady:

- Modální logika se používá často k definování sémantiky ontologických kategorií, jako jsou rigidita nebo identita.
- Modální logika se také používá k definování sémantiky výkonných akcí, jako jsou třeba *inform* a *request* v jazyce ACL⁶. Formální definice poslouží člověku k pochopení, vyhodnocení a porovnání alternativ ACL.

3.1.4 Formální sémantika pro strojové zpracování

Explicitní, formálně vyjádřená sémantika určená pro stroje umožňuje přímé zpracování včetně automatické inference. Idea spočívá v tom, že pokud se narazí na nový pojem, lze o tomto pojmu něco automaticky odvodit a pak adekvátně zpracovat. Inferenční stroje mohou sloužit různým účelům.

Určující vlastnost sémantického webu je strojově zpracovatelný *obsah*. Což jinými slovy znamená, že stroj ví, co má udělat s webovým obsahem, na který narazí. To ale vůbec neznamená, že zde musí být vůbec nějak vyjádřena explicitní sémantika. Místo toho sémantika (ať implicitní, neformální, nebo formální) může být přímo zapracována do webových aplikací. Robustnější přístup směřuje pochopitelně k formální reprezentaci sémantiky, tím umožnit její strojové *zpracování* a dynamicky odhalovat význam obsahu a jak jej použít – toto se označuje jako *strojově zpracovatelná sémantika*. Otázkou pak zůstává, jak se může stroj (softwarový agent) dozvědět něco o významu pojmu, na který dosud nikdy nenarazil.

K uvedenému problému lze přistoupit dvěma způsoby – procedurálně nebo deklarativně. Procedurální přístup spojuje sémantiku s vyvoláním určité procedury v okamžiku, kdy se narazí na určitý příznak. K sémantickému webu má však blíže přístup deklarativní, který sémantiku zachycuje formální deklarací. Toto však nemůže fungovat zcela obecně. Je zapotřebí rozumět symbolům, znát syntaktická pravidla použitého jazyka. Jinak se jedná o úlohu z oblasti pokročilé kryptografie, která je extrémně obtížná pro člověka, natož pak pro stroje. Proto je nutno vycházet z určitých předpokladů:

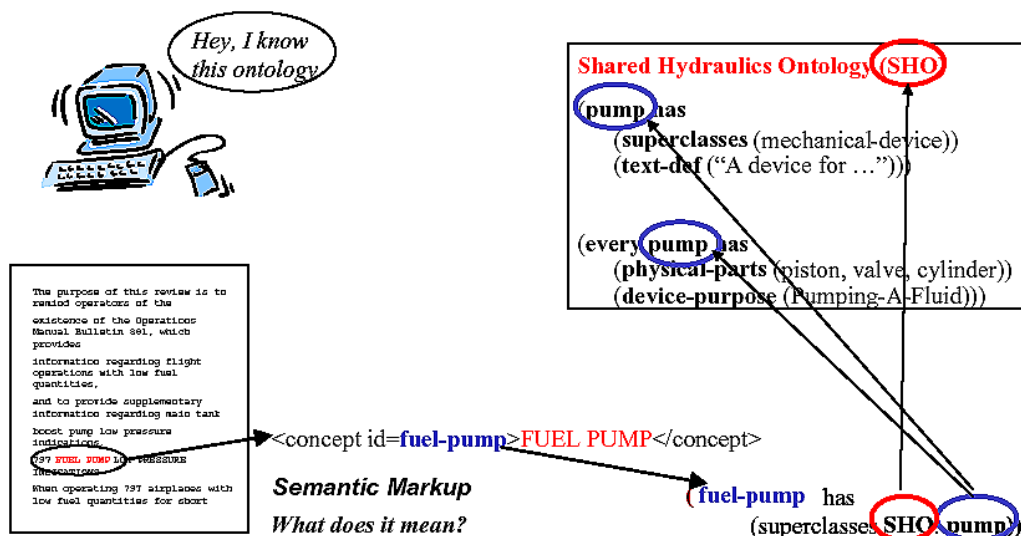
1. Stejný jazyk reprezentace – různé ontologické jazyky vycházejí z různých paradigmat (např. deskripční logika, logika prvního řádu, reprezentace založená na rámcích,

⁵DARPA Agent Markup Language+Ontology Inference Layer

⁶Agent Communication Language

taxonomie, sémantická síť apod.), mají různou vyjadřovací sílu, různá míra formální podpory pro vyjádření sémantiky, různá míra schopnosti inference apod. Proto je nutnou podmínkou, aby byl obsah zapsán pomocí jazyka, který je stroji (agentu) znám – např. DAML+OIL, RDF Schema.

2. Logicky kompatibilní konceptualizace – stejný jazyk nezaručuje „dorozumění“ dvou stran. Například v [12] je uveden příklad dvou reprezentací času, jedna je založena na časových intervalech a druhá pak na časových bodech. Tento podstatný rozdíl je příčinou toho, že agent vycházející z ontologie založené jedním způsobem nemůže přijmout axiomy druhého agenta používajícího ontologii principiálně jinou. Z hlediska logiky lze dvě reprezentace připodobnit oleji a vodě.
3. Veřejně deklarované koncepty – ani při sdíleném jazyku a kompatibilní konceptualizaci nelze vyloučit, že dva různí lidé nebudou pro stejnou doménu používat různé ontologie. Dva různé pojmy mohou mít stejný význam a stejný pojem může mít dva různé významy. Tentýž koncept může být modelován s různou mírou podrobnosti. Určitá myšlenka může být vyjádřena s využitím rozdílných jazykových primitivů. Například jde o to modelovat „být červený“ – použít atribut barvy s hodnotou „červená“, nebo použít třídu typu „ČervenáVěc“? Nebo použít obojí, ale jak to je potom s nezávislostí obou modelů?



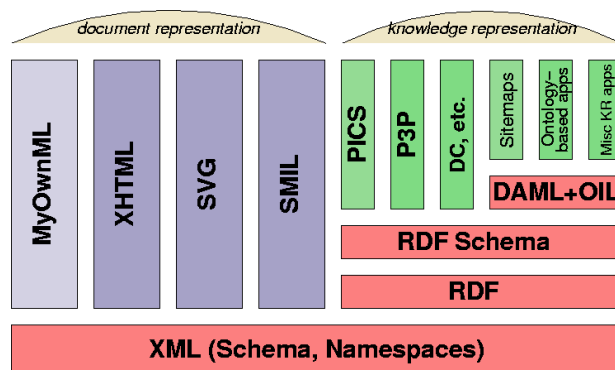
Obrázek 2: Formální sémantika pro strojové zpracování

Předpokládejme agenta, který má za úkol shromažďovat informace o různých mechanických zařízeních (viz obr. 2). Narazí na webovou stránku s textem „fuel pump“ (palivové čerpadlo). Bez schopnosti porozumění přirozenému jazyku je termín nejednoznačný. Víceznačnost lze redukovat asociací textu „fuel pump“ s formálně definovaným pojmem

„*fuel-pump*“ (sémantické značkování). Agent nemusel dosud na tento pojem narazit. V takovém případě bude definice nového pojmu bude získána z definice pojmu *pump*, která je součástí externí ontologie. Agent může zjistit, že *fuel-pump* je podtřídou *pump*, která je zase podtřídou *mechanical-device*. Agent proto nyní ví, že *fuel-pump* není ani psací stroj, ani kosmická loď, protože ty nejsou zahrnuty mezi typy čerpadel. Agent nemá znalost o jaký druh čerpadla se jedná, pouze ví, že jde o nějaké čerpadlo. Nicméně tato úroveň porozumění postačuje, aby agent daný dokument vyhodnotil jako relevantní při dotazu na mechanická zařízení. Podobný přístup je již dnes implementován i v některých komerčních produktech⁷. Uvedený příklad ukazuje důležitost sémantického značkování a sdílení ontologií, rovněž tak zvýrazňuje význam formálních ontologií a automatické inference.

4 Metadata a sémantický web

Sémantický web je podmíněn existencí vhodných metadat, právě díky nim bude možné strojově zpracování obsahu. Jedním z problémů strojového zpracování je jejich vhodná reprezentace. Zdálo by se, že zde zde žádný problém není, protože od počátku propagace sémantického webu byl jako základní vyjadřovací prostředek uváděn jazyk RDF (viz obr. 3). Na druhou stranu se však sluší uvést, že to není jediná možnost. Vedle jazyka RDF, který vznikl pod záštitou konsorcia W3C, totiž existuje také standard *Topic Maps*. Ten vznikl jako standard ISO 13250 nezávisle na W3C a nabízí trochu více vlastností než RDF, a proto je zajímavou alternativou.



Obrázek 3: Jazyky sémantického webu

Metadata sama o sobě pro potřeby sémantického webu nestačí. Jak již bylo řečeno jsou zapotřebí standardy pro syntaxi jejich zápisu a reprezentaci sémantických znalostí tak, aby je agenty mohly efektivně a mnohostranně používat. V posledních letech byla navržena řada vhodných technologií, některé vznikly ještě před úvahami o sémantickém webu, jiné zase

⁷<http://www.ontoprise.com>

byly vyvinuty a zpřesněny právě s ohledem na specifické potřeby sémantického webu. Mezi všemi klíčové postavení mají tyto čtyři:

- XML, XML Schema
- RDF, RDF Schema
- DAML+OIL
- OWL⁸

Úvahy o významu metadat a jejich použití při strojovém zpracování obsahu vycházejí z mlčky přijímaného předpokladu, že metadata existují. Takový předpoklad je však velmi odvážný, protože vytvořit metadata, která jsou přesná a konzistentní, je úloha obtížná. O tom ostatně mohou říci své informační profesionálové. Jde o velkou výzvu, jak decentralizovat činnost, která je obvykle v informačních institucích centralizována a je zajišťována speciálně školenými pracovníky. Rozvoj sémantického webu je tak podmíněn také vývojem nástrojů, které umožní relativně snadno pořizovat sémantická metadata o webových dokumentech.

Jednou z vizí Tima Berners-Lee bylo anotování založeného na sdíleném autorství. K tomuto účelu vznikla řada anotačních pomůcek – od jednoduchých, umožňujících prostý popis, až po sofistikované systémy, které nabízejí víceúrovňové pohledy v různých formátech.

Bezproblémová tvorba RDF metadat závisí na dvou věcech: na předem dané ontologii, která zahrnuje potřebné pojmy a vztahy mezi nimi, a uživatelské rozhraní pro snadnou, až intuitivní tvorbu instancí metadat. Obvykle se v jednom okně zobrazuje ontologie, ve druhém popisovaný dokument. Autor zvýrazňuje anotované prvky dokumentu, pak používá ontologii pro jejich vymezení a určení případných vzájemných vztahů. Metadata ve tvaru RDF se pak generují automaticky.

Přehled prostředků pro tvorbu sémantických metadat lze najít např. na adrese <http://annotation.semanticweb.org/tools>. Mezi nimi lze uvést:

- OntoMat
- Annotea
- Annozilla
- COHSE
- SMORE

⁸Web Ontology Language

5 Alternativní cesty získávání metadat

V poslední době se rozvíjejí metody, jejichž cílem je ze semistrukturovaných webových dokumentů získat sémantická metadata. Ta potom pochopitelně umožňují další manipulace s dokumenty včetně strukturovaných dotazů. Nejde o obvyklou úlohu extrakce informací (sumarizace, anotování), ale jde o využití různých příznaků a z nich potom usuzovat na výskyty entit a vztahů mezi nimi. Navrhovaných metod je celá řada, např. lze uvést:

- vizuální vzory (design patterns)
- lexikografická analýza
- využití ontologií [11]

6 Závěr

Role metadat a sémantiky pro potřeby sémantického webu je zřejmá. Lze uvést řadu podpůrných příkladů, ať již učebnicového charakteru, nebo vycházejících z experimentálních systémů. Vyhledávání informací je jednou z oblastí, která „volá“ po sémantické podpoře. Vyhledávací stroje založené na práci s plnými texty jsou stále dokonalejší a pořád dokáží své uživatele něčím překvapovat⁹. Nicméně se ukazuje, že efektivnost jejich vyhledávání je nižší než u vyhledávání založeného na sémantické bázi. Například [7] uvádí výsledky rozsáhlého experimentu, který porovnává klasické vyhledávání s použitím klíčových slov a vyhledávání založeného na použití ontologie. Pro porovnání byly použity různé typy dotazů – jednak široké s použitím obecných slov, jednak specifické s použitím úzce vymezených slov nebo určitých sousloví. Hodnota F score¹⁰ se u obecných dotazů s využitím klíčových slov pohybovala v rozmezí 20–44 %, zatímco s využitím ontologií byla efektivnost výrazně vyšší – 92–94 %. U dotazů specifických byl výsledek trochu těsnější, ale i zde byly dotazy s ontologií přesvědčivější: 71–83 % vs. 83–91 %.

Sémantický web je vizí, jejíž naplňování je během na dlouhou trať, krůček po krůčku k cíli, který je zdánlivě v nedohlednu. Je to trochu rozpor oproti očekáváním a zvyklostem, kdy uživatelé i firmy mají rádi okamžité efekty. Sémantický web má za sebou již několik let výzkumu, ale přesto ještě nedospěl. Zdá se, že je stabilizovaný stav v otázce *jak* metadata vyjadřovat. Nyní se výzkum mj. ubírá různými směry, které se snaží nalézt metody efektivního získávání metadat. Závěrem lze jen souhlasit se slovy analytiků: „*Sémantický web je do značné míry vizí, vypadající, jako by patřila do velmi vzdálené budoucnosti. Snaha o to, aby spolu počítače nejenom komunikovaly, ale aby rovněž chápaly obsah své komunikace,*

⁹Kolik uživatelů třeba zná a používá schopnost služby Google vyhledávat synonyma? Stačí jen málo – uvést před vyhledávané slovo znak ~. Funguje to však jen pro angličtinu.

¹⁰F score se vypočítává jako harmonický průměr z hodnot úplnosti a přesnosti, obvyklých charakteristik používaných pro měření efektivnosti vyhledávání.

představuje tradičně obtížně řešitelný problém.“¹¹ A pak se třeba naplní obsah fikce *August 2009: How Google beat Amazon and Ebay to the Semantic Web*.¹²

Reference

- [1] Abramowicz, Witold (ed.). *Knowledge-Based Information Retrieval and Filtering from the Web*. The Kluwer International Series in Engineering and Computing Science. Boston: Kluwer Academic Publishers, 2003. 303 s. ISBN 1-4020-7523-5.
- [2] Geroimenko, Vladimir. *Dictionary of XML Technologies and the Semantic Web*. Springer Professional Computing. London: Springer, 2004. ISBN 1-85233-768-0.
- [3] Greenberg, Jane; Sutton, Stuart; Cambell, D. Grant. Metadata: A fundamental component of the semantic web. *Bulletin of the American Society for Information Science and Technology*, 2003, roč. 29, č. 4, s. 16–18.
- [4] Hess, Andreas; Kushmerick, Nicholas. Learning to attach semantic metadata to web services. In Fensel, Dieter (ed.) *ISWC*, Lecture Notes in Computer Science 2870. Springer, 2003, s. 258–273.
- [5] Hunter, Jane L. A survey of metadata research for organizing web. *Library Trends*, 2003, roč. 52, č. 2, s. 318–344.
- [6] Jacob, Elin K. Ontologies and the semantic web. *Bulletin of the American Society for Information Science and Technology Bulletin of the American Society for Information Science and Technology*, 2003, roč. 29, č. 4, s. 19–22.
- [7] Khan, Latifur; McLeod, Dennis; Hovy, Eduard. Retrieval effectiveness of an ontology-based model for information selection. *The VLDB Journal*, 2004, roč. 13, s. 71–85.
- [8] Morris, Jeff. Putting it together: Taxonomy, classification & search. *Transform Magazine*, 2003, roč. 12, č. 9, s. 29–35.
- [9] Özsoyoglu, Gultekin; Al-Hamdani, Abdullah. Web information resource discovery: Past, present, and future. In Yazici, Adnan; Sener, Cevat (eds.) *ISCIS*, Lecture Notes in Computer Science 2869. Springer, 2003, s. 9–18. URL <http://art.cwru.edu/T0papers/ISCIS03Short.pdf>.
- [10] Sivashanmugam, Kaarthik; et al. Kap. Metadata and Semantics for Web Services and Processes. In Benn, W.; et al. (eds.) *Datenbanken und Informationssysteme*. Praktische Informatik I. Hagen, 2003, s. 245–271. URL <http://lsdis.cs.uga.edu/lib/download/Schlageter-book-chapter-final.pdf>.
- [11] Stuckenschmidt, Heiner; van Harmelen, Frank. Ontology-based metadata generation from semi-structured information. In *International Conference On Knowledge Capture: Proceedings of the international conference on Knowledge capture*. New York: ACM Press, 2001, s. 163–170. URL <http://www.cs.vu.nl/~frankh/postscript/K-CAP01.pdf>.
- [12] Uschold, Michael. Where are the semantics in the semantics web? *AI Magazine*, 2003, roč. 24, č. 3, s. 25–36. URL http://lsdis.cs.uga.edu/SemWebCourse_files/WhereAreSemantics-AI-Mag-FinalSubmittedVersion2.pdf.
- [13] Vassiliadis, Panos; Stavrakas, Yannis. Different perspectives of metadata for web-based information systems. In Aberer, Karl; Read, Brian J (eds.) *11th ERCIM Database Research Group Workshop on Metadata for Web Databases*. European Research Consortium for Informatics and Mathematics, 1998. URL <http://www.ercim.org/publication/ws-proceedings/11th-EDRG/pvassil1.pdf>.

¹¹Ronald Schmelzer, firma ZapThink, <http://www.zapthink.com>

¹²http://www.ftrain.com/google_takes_all.html

- [14] Wittenburg, P.; Broeder, Daan. Metadata overview and the semantic web. In *IMDI 2002 Workshop*. Nijmegen: Max-Planck-Institute for Psycholinguistics, 2002. URL http://www.ling.lu.se/projects/echo/contributors/pdf/metadata_overview.pdf.