AKTUÁLNÍ PROBLÉMY A PERSPEKTIVY SÉMANTICKÉHO WEBU

Doc. Ing. Vojtěch Svátek, Dr.
Vysoká škola ekonomická v Praze
Katedra informačního a znalostního inženýrství

Studijní materiál ke kurzu 4IZ440 Propojená data na webu
(upraveno z textu pro předmět Znalostní technologie II na Univerzitě Hradec Králové)

1. Úvod

Sémantický web je v mnoha směrech kontroverzním a víceznačným pojmem. Pro někoho jde o ad hoc „nálepku“ pro množinu specifických standardů (s ústřední rolí datového formátu RDF a dotazovacího jazyka SPARQL) a nad nimi postavených technologických řešení. Někdo pod něj naopak zahrnuje jakékoli aktivity nebo artefakty obsahující sémantické prvky (přičemž pojem „sémantický“ je sám o sobě mnohovýznamový), které se odehrávají ve webovém prostředí nebo v určité vazbě na ně. Množství různých odborných komunit, které v oblasti sémantického webu vycítily příležitost uplatnit své dřívější výsledky, je enormní: během posledních přibližně 10 let do této oblasti fakticky „migrovala“, ať už dlouhodobě nebo krátkodobě, řada odborníků, jejichž původními oblastmi byly např. databázové inženýrství, výpočtová logika, zpracování přirozeného jazyka, dolování z dat nebo knihovnicko-informační věda. Pro nové zájemce, kteří s tématikou sémantického webu přijdou v současnosti do styku, je proto velmi nesnadné rozpoznat, co je jádrem disciplíny a co okrajové, co je ustálené a co novinka, ale dokonce i to, zda se obor jako celek posouvá směrem k masivnímu uplatnění v praxi nebo se naopak stahuje do zákoutí akademických laboratoří. Autor tohoto textu měl příležitost se na výzkumu sémantického webu skromným dílem podílet téměř od začátku,[1] a měl proto příležitost dynamiku oboru a jeho tematické vymezení zblízka sledovat. Hlavním cílem předkládaného textu je podělit se o tyto zkušenosti, a také nové zájemce z řad studentů (případně i mladých akademických pracovníků) povzbudit k aktivnímu zapojení do těch směrů sémantického webu, které se v blízké době zřejmě budou výrazněji rozvíjet.

Kapitola 2 tohoto studijního textu se ohlíží do minulosti a stručně mapuje historii sémantického webu, ale i toho, co jeho vzniku bezprostředně předcházelo. Kapitola 3 je zaměřená na současnost: pokouší se charakterizovat, v jaké fázi svého životního cyklu se technologie sémantického webu nacházejí, jaké jsou hybné síly pro jejich široké uplatnění a co naopak tomuto uplatnění brání (včetně možných způsobů odstraňování těchto překážek). Velmi krátká Kapitola 4 mapuje situaci v České republice, a závěrečná Kapitola 5 shrnuje hlavní myšlenky textu.

2. Stručná historie sémantického webu

2.1. Prehistorie: sémantické modelování dat a znalostí

O „sémantice“ v oblasti zpracování strukturovaných dat[2] můžeme hovořit tehdy, když jsou data doprovázena metadaty specifikujícími jejich význam. Metadata mají zpravidla charakter odkazu do určitého *slovníku*, kde je příslušný věcný typ dat vymezen a vysvětlen, ať už formou textu, nebo i s pomocí určitého logického aparátu. V tomto smyslu můžeme jako sémantická data chápat např. data uložená v relační databázi, za předpokladu, že jednotlivé hodnoty nebo názvy sloupců

[1] Konkrétně, od zahájení projektu 5. rámcového programu EU nazvaného OntoWeb (v r. 2001), v jehož rámci se utvářely základy současných standardů, a také od prvního ročníku International Semantic Web Conference (2002) jako nejvýznamnějšího mezinárodního setkávání této komunity.

[2] V tomto textu se nebudeme věnovat sémantice jako odvětví jazykovědy, která se zabývá významem řečových aktů. Sémantický web totiž na toto pojetí sémantiky navazuje jen zcela minimálně.

obsahují např. odkaz na určité standardizované seznamy hodnot (označované nejčastěji jako číselníky nebo kódovníky).

Míra „sémantičnosti“ dat se ovšem odvíjí od bohatosti a přesnosti informačního obsahu příslušných slovníků. V tomto směru „ploché“ seznamy hodnot nebo i jednoduché hierarchické číselníky výrazněji zaostávají za formálními konceptuálními modely reality – *ontologiemi* – které se staly ve světě informatického výzkumu (zejména té jeho části, která se označuje jako „umělá inteligence“, a ještě přesněji lze mluvit o komunitě reprezentace znalostí – „knowledge representation“) populárním tématem v polovině 90. let. Již v té době se principiálně nelišily od mnohých ontologií navrhovaných a zkoumaných dnes: jednalo se o množiny logických formulí, vztahujících se ke třídám objektů z určité věcné domény a ke vztahům mezi nimi. Nepoužívaly však globální webové identifikátory. Mezi hlavní zástupce ontologických jazyků z té doby patří dnes již zcela odložená Ontolingua [12], dale CyCL (který je stále do jisté míry používán jako původní jazyk znalostní báze CyC [23]), nebo z Ontolinguy odvozený OCML (Operational Conceptual Modelling Language) [26], založený na LISPu a reálně využitelný pro praktické úlohy spojené s logickým odvozováním i s procedurálním výpočtem.

2.2. Sémantická data vstupují na web

Pokud jde o World-Wide Web, jeho tvůrce, Tim Berners-Lee, ho od samého začátku plánoval jako „sémantický“, tj. se zachycením věcného významu dat obsažených v elementech stránek. Prvotní jednoduchá verze webového jazyka HTML, zaměřená především na prezentační aspekty, se ale rychle rozšířila rychleji, než mohly být sémantické prvky do základního standardu zapracovány. Teprve později byl standard HTML obohacen o kaskádové styly (CSS), umožňující zavést uživatelské třídy elementů a tím i zachytit vedle instrukcí pro prohlížeč také věcné kategorie informací (alespoň na úrovni názvů tříd). Někteří weboví vývojáři proto v té době interpretovali pojem „sémantický web“ jako synonymum pro technologii CSS. Určitý prostor pro „sémantiku na webu“ poskytovaly také, i dnes využívané, elementy META v hlavičce dokumentů HTML.

Prvním reálně uskutečněným pokusem přenést reprezentaci znalostí z uzavřených systémů do webového prostředí se stal kolem roku 1997 projekt SHOE (Simple HTML Ontology Extension) [18]. Jak název naznačuje, jednalo se o rozšíření HTML o sémantické elementy, umožňující vztahovat webové stránky k reálným objektům, jejichž typy jsou definovány v ontologii (rovněž vystavené na webu). Příkladem je element

<INSTANCE KEY="http://novak.cz">

který pomocí identifikátoru URL definuje novou entitu – člověka, o kterém se na stránce vystavují informace, dále

<CATEGORY NAME="cs.GraduateStudent">

vyjadřující příslušnost dotyčné (implicitně odkazované) instance ke třídě postgraduálních studentů, a nebo

<RELATION NAME="cs.name">

<ARG POS=1 VALUE=" http://novak.cz">

<ARG POS=2 VALUE="Jan Novák">

</RELATION>

definující vztah o dvou argumentech, platící mezi člověkem (identifikovaným daným URL) a jeho jménem.

Stránka využívající SHOE ovšem pak již nebyla plně validní stránkou HTML. Sémantické elementy se v ní ne zcela organicky směšovaly s původními elementy HTML zaměřenými na prezentaci. Poněkud problematicky bylo řešeno i odlišení reálných objektů od webových stránek, které o nich pojednávají.[3] SHOE byl proto zakrátko opuštěn, a nejvýznamnější z jeho spolutvůrců, J. Hendler, se stal zastáncem nových jazyků RDF [25] a OWL [20].

Jazyk HTML je založen na starším značkovacím jazyku SGML, ze kterého přibližně ve stejné době vznikl zjednodušený jazyk XML. Když se na konci 90. let začaly zintenzivňovat kontakty mezi odbornými komunitami reprezentace znalostí a webového inženýrství, naskytla se otázka, zda je pro reprezentaci významu (sémantiky) na webu vhodný právě XML, jako postupně stále populárnější, hierarchicky orientovaný jazyk, nebo jeho méně známá alternativa, síťově orientovaný jazyk RDF (navržený v roce 1999) [25], vyjadřující fakta pomocí trojic „subjekt – predikát – objekt“. Pro XML hovořila úzká spojitost s HTML, která se ještě prohloubila se vznikem jazyka XHTML – dialektu HTML zcela odpovídajícího standardu XML. Hlavní zbraní RDF v tomto „souboji“ byl jeho důraz na využívání URI jako globálních identifikátorů entit, ale také vyšší flexibilita síťové reprezentace a její lepší shoda s požadavky na logické odvozování. Zástupci komunity reprezentace znalostí (kteří v té době chápali logické odvozování jako ústřední operaci, která se na sémantickém webu bude provádět), těsně po roce 2000 uspěli při získávání finanční podpory z tzv. 5. rámcového programu EU, což se stalo rozhodujícím momentem pro dominanci RDF ve výzkumu sémantického webu. Určitým ústupkem vůči rozsáhlé komunitě XML bylo využití XML jako prostředku *serializace* RDF pro ukládání v dokumentech. Syntaxe RDF/XML se stala vůbec první standardizovanou syntaxí RDF, a stále proto patří, navzdory své často kritizované neúspornosti a velmi špatné čitelnosti pro člověka, mezi nejpoužívanější (až v posledních letech ji v roli „první volby“ stále častěji nahrazuje syntaxe Turtle).

Rozsáhlá komunita webových vývojářů ovšem v počáteční fázi RDF převážně ignorovala, a vytvořila si vlastní, pragmatický přístup k zachycení strukturovaných dat na webu: *mikroformáty*. Jedná se o soubory metadatových hodnot navržené pro každý typ obsahu (např. vizitkové informace, recenze, kalendářové informace)

[3] Tento problém není plně uspokojivě vyřešen ani v současných standardech a „best practices“. Jedná se totiž o problém principiální: reálné objekty nemohou být „vystaveny“ na webu; zároveň je ale žádoucí, aby na webu měly svůj jednoznačný identifikátor, a aby tento identifikátor byl *dereferencovatelný*, tj. aby bylo pomocí HTTP požadavku možné získat *dokument* shrnující informace o daném objektu. Takový dokument ovšem může být rovněž potřebné opatřit jeho vlastním identifikátorem, což vyžaduje zavedení specifických konvencí v rámci adresování URL.

samostatně, které lze vložit do kódu stránky – typicky jako hodnoty určitého atributu. Jednoduchým příkladem je deklarování určitého řetězce jako jména osoby, analogicky k příkladu uvedeném pro SHOE:

```
<div class="vcard">
 <a class="url fn" href="http://novak.cz"> Jan Novák</a>
</div>
```

Souběžně ale určitá část „ambicióznějších“ vývojářů výhody plynoucí z flexibility modelu RDF (a definování sémantiky pomocí ontologií) rozpoznala a podílela se na vzniku syntaxe RDF spočívající v zanoření trojic RDF do kódu HTML – na rozdíl od SHOE však již ne prostřednictvím nových elementů, ale pouze *atributů*. Nová syntaxe, standardizovaná W3C v r. 2008, se proto nazývá *RDFa* („RDF in attributes“) [1]. Jako příklad trojic zanořených do HTML můžeme uvést následující fragment, přibližně odpovídající předchozímu použití mikroformátu:

```
<html xmlns:foaf="http://xmlns.com/foaf/0.1/">
…
  <div about="#me" typeof="foaf:Person">
      <a property=”foaf:homepage” href="http://novak.cz">
      <span property="foaf:name">Jan Novák</span></a>
      </div>
```

V hlavičce dokumentu je deklarován prefix a URI ontologie FOAF (Friend of a Friend) [6], ze které je následně uvnitř stránky použita třída *Person* a vlastnosti *homepage* a *name*. Tento kód zachycuje tři trojice; všechny mají jako subjekt URI složené z URL stránky a fragmentu *#me* (tímto fragmentem je osoba, o které stránka pojednává, jasně odlišena od stránky samotné).

Souběžná existence mikroformátů a RDFa trvá až do dnešní doby, s tím, že jako třetí „do hry“ vstoupila tzv. *mikrodata*.[4] Jejich předností je těsné spojení vývoje jejich specifikace se standardem HTML 5 (v počáteční fázi byly přímo jeho součástí).

2.3. Sémantický web jako distribuovaná databáze

Počáteční koncepce sémantického webu, zpopularizovaná vizionářským článkem [5], se výrazně opírala o techniky *umělé inteligence* – stála na webovém zpřístupnění znalostních bází, nad kterými budou softwaroví agenti automaticky odvozovat nové informace. Přibližně v letech 2004-5 se ale jako významný začal ukazovat alternativní pohled na sémantický web, totiž, jako na rozsáhlou, distribuovanou databázi. Odlišnost tohoto pohledu spočívá zejména ve dvou aspektech:

- Znalostní báze jsou zpravidla získávány rozhovorem s doménovým expertem, případně sofistikovanou analýzou základních dat. Tyto procesy jsou velmi časově, případně výpočetně náročné (hovoří se o tzv. „knowledge

[4] Viz např. http://html5doctor.com/microdata/

acquisition bottleneck“, tj. získávání znalostí jako úzkém místu procesu tvorby znalostních systémů), a vzniklé báze proto nemohou mít příliš velký rozsah. Naopak databázové zdroje shromažďují „běžná“ data vznikající při rutinních činnostech (výrobních a obchodních podniků, veřejné správy, médií, univerzit atd.), mohou být proto i v případě převedení do datových formátů sémantického webu velmi *rozsáhlé*.

- Zpracování takových dat nemá obvykle charakter odvozování zcela nových znalostí, ale spíše jen *vyhledání* stávajících dat a jejich *propojení*. Pro vyhledávání dat je přirozené využít *dotazovací jazyky*, podobné těm, které byly dávno předtím navrženy pro oblast relačních (nebo jiných) databázových systémů.

První jednoduché jazyky pro dotazování do RDF se objevily prakticky současně s datovým formátem RDF samotným. Za první dotazovací jazyk plně kompatibilní se specifikací RDF a současně poskytující dostatečnou vyjadřovací sílu (včetně možnosti vyjádřit nepovinnost části hledaného grafu, která je s ohledem na reprezentační flexibilitu RDF často nezbytná) je obvykle považován SeRQL. Z něj se následně (se zahrnutím inspirace dalšími zdroji) vyvinul jazyk SPARQL, který je dnes již široce používaným standardem podporovaným W3C [14].

Soustředění na zpracování rozsáhlých dat pomocí dotazovacích jazyků v první fázi trochu oslabilo „webovost“ sémantického webu: databázová komunita začala spíše zkoumat obecnou efektivitu indexování a dotazování v síťové reprezentaci[5] založené na trojicích, ve srovnání s reprezentací relační. Vznikla celá řada propracovaných implementací *úložišť RDF trojic* (tzv. „triple stores“), a to jak pro nativní ukládání trojic přímo v RDF, tak i jako „sémantická“ vrstva nad běžnou relační databází (často MySQL).

V další fázi, počínaje přibližně rokem 2007, se však v komunitě sémantického webu začal prosazovat nový slogan „Less semantics, more web“. Jeho první polovina odrážela rostoucí nedůvěru v praktickou uplatnitelnost sofistikovaného strojového odvozování (a „inteligentních agentů“) pro praktické úlohy nad rozsáhlými webovými daty. Lze říci, že od té doby již prototypy „sémantického“ software nasazované v praxi využívají formálně-logické odvozování jen v minimální míře, a nahrazují ho spíše transformačními dotazy v jazyce SPARQL (s klauzulí CONSTRUCT umožňující zkonstruování nových RDF trojic na výstupu) nebo procedurálním programovým kódem.[6] Druhá polovina sloganu pak upozorňovala na otevřenost a heterogenitu webového prostoru a s ní spojenou potřebu vnést do samého centra vývoje sémantických technologií předpoklad decentralizace dat a jejich zpracování. Zásadní roli v tomto posunu těžiště sehrál opět „otec“ webu Tim Berners-Lee, který již v r. 2006 přišel s konceptem propojených dat neboli *linked data*. Ten se v krátké době stal

[5] Pro zajímavost si můžeme připomenout, že „konkurence“ mezi relačním a síťovým modelem se v historii databázového inženýrství v minulosti již objevila: v 70. letech hrál dokonce síťový model (ač od RDF výrazně odlišný), prominentní roli, a tehdy vyvinutý systém IDMS se místy používá dodnes.

[6] Jakkoli zároveň stále ještě „pod vlajkou“ sémantického webu probíhá řada akademických projektů orientovaných na logické odvozování, případně i agentové technologie.

nejvýraznějším hybným momentem pro uvádění výsledků (v té době již poněkud stagnujícího) výzkumu sémantického webu do reálné praxe.

2.4. Linked data jako ucelený soubor principů vystavování dat

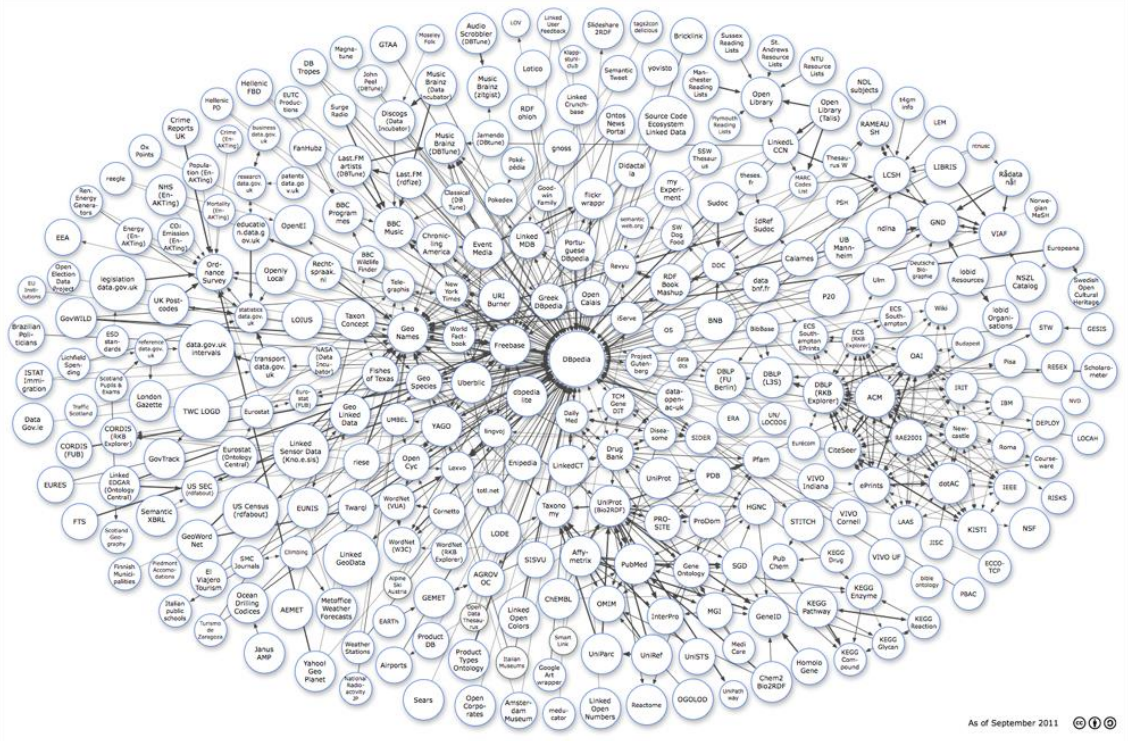
Deklarovanými hlavními principy Linked data, poprvé zveřejněnými v [4], jsou tyto:

- *Identifikace entit* pomocí globálních identifikátorů ve tvaru *URI*.
- Využívání *HTTP URI*, na které je možné se dotazovat pomocí generických prostředků webového adresování (servery doménových jmen – DNS).
- Na dotaz položený formou URI je potřeba vrátit *užitečnou informaci* o entitě, která je tímto URI identifikována. Tato informace by měla být reprezentována ve standardním jazyce, kterým je *RDF*.
- Informace o dané entitě by měly zahrnovat *propojení*, opět formou URI, na další, související, entity (případně může jít o jiná URI používaná pro stejnou entitu v jiných datových zdrojích).

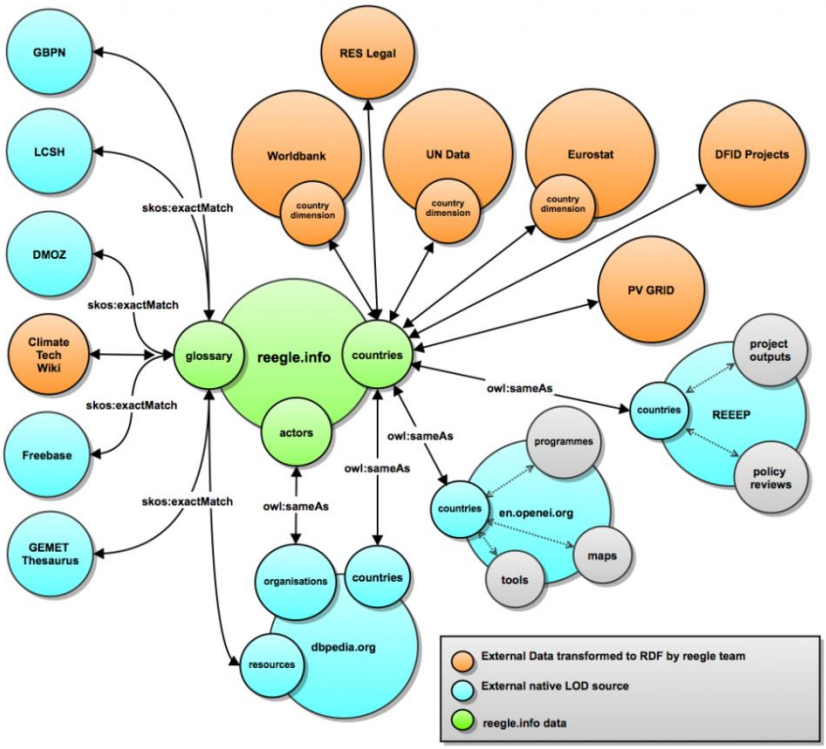
Vidíme, že na rozdíl od původního konceptu sémantického webu se zde nepředjímá existence automaticky usuzujících agentů. Cílem je pouze to, aby byla data vystavena na webu a vzájemně propojena, a to s využitím jednotných formátů (URI, RDF).

V letech následujících bezprostředně po publikování těchto principů objem dat vystavených v souladu s nimi (měřeno počtem tématicky vymezených datových sad – datasetů) rapidně rostl. Ty z nich, které splňovaly i detailnější kritéria, týkající se mj. rozsahu a počtu linků, byly zařazovány do tzv. *Linked Data Cloud*. Poslední publikovaná verze tohoto „oblaku datasetů“ je na Obr. 1. V jeho centru je umístěna *DBpedia*, strukturovaná verze Wikipedie, obsahující informace automaticky vyextrahované z jejích informačních boxů, úvodních textových pasáží, rozcestníků atd. DBpedia je nejčastějším zdrojem, na které se jiné datasety odvolávají, což je dáno jejím encyklopedickým charakterem: téměř v jakékoliv oblasti lidské činnosti lze nalézt témata (ať už charakteru individuí – konkrétních osob, organizací, lokalit, historických událostí… nebo charakteru typů – označení lidských profesí, biologických taxonů, předmětů denní potřeby…) natolik populární, že pro ně vznikl článek ve Wikipedii a tudíž následně i „záznam“ (resp. URI entity, propojené na jiná URI i datové hodnoty) v DBpedii.

V posledních dvou letech již ucelené zobrazení Linked Data Cloudu jako „celého sémantického webu“ (alespoň, v pojetí Linked Data) přestalo být udržitelné a pomalu postrádá smysl. Místo něj se objevují schémata specifičtější pro konkrétní doménu: příkladem je *Linked Linguistic Data Cloud* pro oblast lingvistických zdrojů, nebo tzv. *Clean Energy Linked Open Data cloud* zachycený na Obr. 2.



Obr. 1: Linked Open Data cloud
(Zdroj: Richard Cyganiak a Anja Jentzsch, http://lod-cloud.net/)



Obr. 2: Clean Energy Linked Open Data cloud
(Zdroj: Andreas Blumauer,
http://blog.semantic-web.at/wp-content/uploads/2013/06/reegle-lod-cloud.png)

2.5. Alternativní způsoby vystavení dat v RDF

V současnosti existují čtyři hlavní způsoby, jak data ve formátu RDF na webu vystavit. Každý z nich má své výhody i nevýhody, které můžeme stručně shrnout.

1. *Linked data* vystavená podle výše zmíněných principů.
    - *Výhody*: Umožňují rychlý *přímý přístup* k informacím o konkrétní entitě, i postupný sběr rozsáhlejších dat formou *navigace* po datovém grafu, analogicky k indexování webových dokumentů. Pomocí tzv. *negociace obsahu* lze na stejném URI poskytovat informace určené pro strojové zpracování (RDF) i pro lidského uživatele (HTML).
    - *Nevýhody*: Nelze získat velký objem dat naráz, a celý tento způsob vystavování klade nemalé nároky na vybavení serveru a znalosti jeho administrátora.
2. *Koncové body* (tzv. endpointy) využívající jazyk SPARQL. Na webu lze data v RDF vystavit prostřednictvím služby umožňující dotazování v jazyce SPARQL.
    - *Výhody*: Je možné získávat data ve *velkém rozsahu*, a přímo při dotazování provádět jejich *transformace* do požadované struktury (pomocí dotazů typu CONSTRUCT).
    - *Nevýhody*: Obecnost jazyka SPARQL (ve srovnání s běžnými webovými API, přes která se dnes data z databázových systémů poskytují) vede k riziku zahlcení serveru sofistikovanějšími dotazy. Určitým omezením je i samotná nutnost zvládnout syntaxe SPARQLu, zejména v případě, kdy chce uživatel pouze získat rámcovou představu o jejich obsahu; pro tento případ však existuje možnost zprovoznit nad koncovým bodem facetový prohlížeč, který převádí akce uživatele (odpovídající navigaci po linked data, jako je tomu u předchozího modelu vystavení) na dotazy ve SPARQL, viz např. [24].
3. Jednoduché vystavení *datových souborů* (dumpů). V případě velkých dat se jedná nejčastěji o datový formát N-Triples, který lze dobře komprimovat.
    - *Výhody*: Data jsou k dispozici ke stažení celá na jeden požadavek.
    - *Nevýhody*: Uživatel dat si musí data sám následně zpracovávat, nevyužívají se tedy v plné míře možnosti datové infrastruktury sémantického webu.
4. Vnoření RDF do HTML: *RDFa*.
    - *Výhody*: Pro data, která jsou z databází standardně produkována v HTML, není nutné zajišťovat další výstupní „kanál“.
    - *Nevýhody*: Pro další zpracování je nutno data ze stránek extrahovat.

První dva přístupy jsou obecně technologicky pokročilejší a poskytují spotřebiteli dat komfortnější možnosti. Vzhledem k vyšším nárokům na takové publikování a

nižší míře obeznámenosti webových vývojářů s potřebnými technologiemi však takovými rozhraními není, a v dohledné době zřejmě nebude, pokrytý dostatečně vysoký počet datových zdrojů. Zatímco počet běžných webových API (postavených na modelu REST a poskytující data ve formátech JSON nebo XML), roste v posledních letech exponenciálně, a v r. 2012 jejich počet, podle přehledu na serveru *ProgrammableWeb.com*[7] překročil 6 000, počet datasetů evidovaných na serveru *Datahub* s dostupností ve formátu RDF je přibližně desetinový.[8] Počet datasetů zařazených do Linked Data Cloudu je ještě menší.

Zejména v případě koncových bodů je také problémem nedostatečná udržovanost v chodu: jak lze ověřit pomocí služeb, které dostupnost koncových bodů monitorují,[9] dokonce i koncové body, které jsou v rámci komunity sémantického webu velmi populární, jako je DBpedia, mívají období nedostupnosti. Z celkového počtu necelých 500 koncových bodů jich je obvykle v provozu jen přibližně polovina. Proto je prozatím obtížně představitelné, že by na takové externí službě bylo dlouhodobě postaveno podnikání komerčního subjektu nebo fungování orgánu veřejné správy. Z toho důvodu se v praxi zatím prosazují spíše „okleštěné" varianty sémantického webu, ve kterém pokročilé modely vystavení slouží jen pro úvodní prozkoumání obsahu dat, zatímco provozní využití se opírá o data stažená (ve formě dumpu) spotřebitelskou organizací na vlastní softwarovou platformu.

2.6. Od otevřených dat k linked open data

Vedle míry pokročilosti technologického způsobu vystavení linked data v RDF, jako základním jazyce sémantického webu, má smysl se zabývat i přechodem tzv. otevřených dat („open data") vystavovaných na webu od „ne-sémantických" datových formátů k „sémantickým".

Pojem „otevřená data" zpočátku vznikl relativně nezávisle na technologiích sémantického webu. Tlak veřejnosti na dostupnost informací o fungování veřejné správy vedl v řadě zemí ke vzniku zákonů, které měly tuto dostupnost zajišťovat.[10] Deklarace politických špiček (např. prezident Obama v USA, premiér Cameron ve Velké Británii) a jimi podpořené iniciativy postupně napomáhaly zvýšení ochoty úředníků na všech úrovních podílet se na veřejném vystavování dat, která spravují. Kromě technologického vystavení je samozřejmě podstatný i právní rámec – *licenční podmínky*, za kterých lze vystavená data dále používat a šířit. Vzhledem k technologickému zaměření předmětu se mu zde ale nebudeme dále věnovat, budeme pouze předpokládat, že data jsou vystavena pod otevřenou licencí, umožňující jejich bezúplatné používání a šíření.

V r. 2010 přišel (opět) Tim Berners-Lee s elegantním znázorněním postupného přechodu od „obyčejných" otevřených dat k „linked open data" pomocí přidělování „hvězdiček", podle následujícího schématu (opět zveřejněného v „živém" webovém dokumentu [4]):

---

[7] http://blog.programmableweb.com/2012/08/23/7000-apis-twice-as-many-as-this-time-last-year/
[8] Ke dni 4.7.2013 se jednalo o 680 datasetů, viz http://datahub.io/dataset?tags=format-rdf.
[9] Viz zejména http://labs.mondeca.com/sparqlEndpointsStatus/index.html.
[10] V ČR se jedná o Zákon č.106/1999 Sb., o svobodném přístupu k informacím.

- Jedna hvězdička: zveřejnění dat s otevřenou licencí v *libovolném formátu*. Může se tedy jednat o pouhé bitové mapy naskenovaných textových dokumentů, jejichž obsah není prakticky nijak strojově zpracovatelný. Je však ručně zpracovatelný lidským uživatelem.
- Dvě hvězdičky: strukturovaná data ve *strojově čitelné* podobě. Tento požadavek splňují data uložená v proprietárních formátech tabulkových kalkulátorů nebo databázových systémů. Jejich zpracování je již nesrovnatelně efektivnější, závisí však na přístupu k příslušnému proprietárnímu software.
- Tři hvězdičky: strukturovaná data v *neproprietárním formátu*, nejčastěji CSV. Export do takového formátu zpravidla pro majitele dat nepředstavuje velký problém, a okruh možných konzumentů dat se tím stane neomezeným.
- Čtyři hvězdičky: strukturovaná data ve *formátu RDF*. Datový formát RDF poskytuje možnost globální identifikace entit pomocí URI, možnost data bez problémů slučovat (díky modulárnímu charakteru „trojic"), a flexibilitu datového schématu (není třeba řešit reprezentaci „chybějících" hodnot – informace, které nejsou známy, se prostě neuvedou). Využívání RDF ovšem nese vyšší režii na straně vystavovatele (data zpravidla v RDF nativně nejsou, a je tedy nutné je do tohoto formátu transformovat) a předpokládá znalost RDF na straně spotřebitele dat.
- Pět hvězdiček: strukturovaná data *propojená na externí data*. Teprve zde se jedná o plnohodnotná „linked open data" – propojená otevřená data.

Lineární schéma „hvězdiček" samozřejmě nepostihuje plně všechny eventuality. Existují např. návrhy na „přemostění" mezi třetí a pátou úrovní formou „propojených dat CSV".[11] To umožní vystavovateli i spotřebiteli oprostit se od nutnosti znát syntaxi a sémantiku RDF, a přitom využívat (pro data, která jsou tabulkově strukturovaná spíše než „přirozeně grafová") výhody propojitelnosti.

Schéma také (jak jeho autor zmiňuje) nezohledňuje kritérium, které je vůči samotným datům externí, ale má zásadní význam: existenci metadat popisujících samotná data a zveřejněných v relevantním *datovém katalogu* otevřených dat. Datové katalogy jsou provozovány s pomocí speciálního software, systémů pro správu dat (DMS), z nichž nejrozšířenější je v současnosti *CKAN*, vyvíjený Open Knowledge Foundation (OKFN).

2.7. Ontologie a datové slovníky

Prozatím jsme se jen velmi stručně dotkli oblasti tvorby ontologií jako modelů definujících *typy* objektů a vztahů, o kterých vypovídají data na sémantickém webu. V tomto odstavci se pokusíme velmi stručně zachytit další vývoj tzv. *ontologického inženýrství* pro potřeby sémantického webu.

Metodiky tvorby ontologií si i po roce 2000 do značné míry zachovávaly své pojetí z období před sémantickým webem, v tom smyslu, že ontologie převážně vznikaly

[11] Viz např. http://jenit.github.io/linked-csv/.

nezávisle jedna na druhé jako „monolitické“ modely. Scénáře jejich využívání nepočítaly s rozsáhlými daty, a soustřeďovaly se na efektivní logické usuzování nad ontologií jako takovou, případně nad nevelkými bázemi faktů cíleně shromažďovanými již s ohledem na danou ontologii.

V oblasti linked data se však ukázal jako nezbytný obrácený postup. Data ve formátu RDF zde vznikají z dat dříve existujících v jiných formátech, a pro nově vytvořený dataset proto málokdy existuje ucelená ontologie. Přesto je žádoucí, aby se konkrétní výskyty objektů a vztahů v datech odvolávaly na typy, které budou „někde“ definované – v ideálním případě v dokumentu, který bude mít sám charakter linked data: URI těchto typů (tříd a vlastností) budou dereferencovatelná právě ve formě tohoto dokumentu. Existují dvě základní možnosti, jak toto udělat:

1. Typy odpovídající objektům a vztahům v dotyčném datasetu budou shromážděny do nově navržené „ontologie“ – protože ovšem takový model vzniká „odspodu“, podle konkrétního datasetu a ne důkladnou analýzou pojmového systému určité věcné oblasti, je vhodnější hovořit o *slovníku* daného datasetu. Tento způsob byl použit mj. v případě DBpedie, pro kterou vznikla tzv. *DBpedia ontology*,[12] ale i v případě mnoha specializovanějších datasetů.
2. Dohledat vhodné typy v existujících (pokud možno, široce rozšířených) ontologiích nebo datových slovnících, a pro daný dataset je přepoužít. V případě datasetů s různorodým obsahem může být často počet zapojených ontologií/slovníků poměrně vysoký, přičemž z každého z nich se použije třeba jen několik málo typů.

Častým případem je ovšem hybridní řešení, kdy se použijí běžné typy z rozšířených ontologií a slovníků, a doplní méně obvyklými typy v rámci slovníku nového.

V zájmu toho, aby byly často používané typy běžně dostupné a nemusely se pro nové datasety navrhovat znova, experimentuje komunita linked data s koncepcí tzv. *vocampů*[13] (z „vocabulary camp“ – setkání určené k vývoji slovníku). Na rozdíl od klasického postupu ontologického inženýrství [16], kdy vývoj ontologie probíhá v dlouhém časovém horizontu, počínaje sběrem relevantních textových termínů ze zvolené domény, přes jejich definování ve slovním glosáři, formalizaci v jazyce s vysokou vyjadřovací silou (přinejmenším predikátová logika 1. řádu), až po kódování v cílovém, omezenějším jazyce umožňujícím efektivní odvozování, je vznik slovníku v rámci vocampu takřka bleskovou záležitostí. Zúčastnění odborníci předloží svůj názor na to, které vlastnosti objektů předem zvoleného typu považují za zásadní a v datech se často vyskytující, navrhne se pro ně konsensuální název, a příslušná URI zavedou, s textovým názvem (hodnota vlastnosti *rdfs:label*) a stručným komentářem (hodnota vlastnosti *rdfs:comment*) do nově vzniklého slovníku.

Charakter linked data se ovšem projevuje ve snaze mezi sebou jednotlivé slovníky provázat, nejčastěji pomocí predikátů *rdfs:subClassOf* a *rdfs:subPropertyOf*. Systém

[12] Viz http://dbpedia.org/Ontology.
[13] Přehled pořádaných vocampů s odkazy na jejich výstupy lze nalézt na http://vocamp.org/.

takto propojených slovníků[14] byl nedávno označen jako *linked open vocabularies*, a zpřístupněn v rámci portálu[15] vytvořeného francouzskou firmou Mondeca (v r. 2012 byl pak zařazen pod správu organizace Open Knowledge Foundation). Aktuálně (červenec 2013) je zde zahrnuto 360 slovníků, pro které lze zobrazit metadata, vzájemné vazby, i časový vývoj příslušné specifikace.

2.8. Shrnutí

Předložené historické „ohlédnutí" je samozřejmě do jisté míry subjektivní. Lze snadno argumentovat, že pojetí sémantického webu založené na agentech, logickém odvozování a dalších prostředcích umělé inteligence je na „vyšší vývojové úrovni" než pojetí založené na postupné transformaci „obyčejných" otevřených dat na linked data, která budou často zpracovávána poměrně jednoduchými aplikacemi, často označovanými jako „mesh-upy" (spojení běžnějšího termínu „mash-up" a termínu „mesh" označujícího síťovou strukturu). Z hlediska časového vývoje i z hlediska velikosti odborné (i laické) komunity, která se v příslušném pojetí sémantického webu angažuje, je však předložený sled dobře zdůvodnitelný. V následující kapitole se pokusíme na vývoj sémantického webu podívat z trochu jiného, analytického spíše než jen popisného, pohledu.

3. Předpoklady dalšího rozvoje sémantického webu

3.1. Sémantický web a model životního cyklu technologie

Jak už jsme naznačili v úvodu, na sémantický web se můžeme dívat buďto jako na obecnou myšlenku, nebo jako na soubor konkrétních technologií, které mají tuto myšlenku realizovat. Tyto pohledy někdy splývají, což může vést k nedorozumění při hodnocení pozice sémantického webu v rámci informatického výzkumu a praxe.

Často používaným stereotypem pro hodnocení pozice určité technologie je tzv. Gartnerovský model životního cyklu (označovaný i jako „hype cycle").[16] Jednotlivé technologie se podle něj pohybují pro trajektorii složené z pěti fází, jejichž anglické termíny vesměs odrážejí charakteristické tvary křivky v dané fázi:

1. Vstup nové technologie na scénu („technology trigger"), dosud bez reálného komerčního využití a často ještě i bez funkčního produktu, avšak s rostoucím zájmem odborných médií.
2. Fáze přepjatých očekávání („peak of inflated expectations"), ve které se již projevuje silný, médii podněcovaný a z velké části neinformovaný zájem ze strany komerčních subjektů. Občasné úspěchy jsou zdůrazňovány a neúspěchy spíše skrývány nebo bagatelizovány.
3. Fáze deziluze („trough of disillusionment"), ve které nastává odliv odběratelů nové technologie, a postupně i jejích poskytovatelů. Pokud

[14] Včetně některých sofistikovanějších modelů, které svým charakterem zaslouží (a často i v názvu mají) označení „ontologie".
[15] http://lov.okfn.org/dataset/lov/
[16] Viz http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

technologie v této fázi zcela nevymizí, je to jen díky tomu, že její zbylí poskytovatelé usilovně vylepšují nabízená řešení podle požadavků zbylých odběratelů.

4. Fáze růstu informovanosti („slope of enlightment“), kdy jsou přínosy pro odběratele jasněji specifikovány a opírají se o existenci úspěšných případů nasazení. Zároveň již existují „nové generace“ produktů, které jsou zacíleny na dosahování těchto přínosů. Okruh odběratelů technologie je však stále ještě omezený a většina projektů má charakter prototypů.
5. Fáze produktivity („plateau of productivity“), ve které je již technologie hromadně přijímána a nasazována, o jejích přínosech už nejsou zásadní pochybnosti.

Když se na sémantický web díváme touto optikou, nemusí být zdaleka jasné, zda jde o technologii, která se nachází ve fázi 2, 3 nebo 4 (a v případě fází 2 nebo 3, zda v „prohlubni“ zanikne nebo nastoupí růst podle fází 4 a 5). Důvodem je zřejmě právě propletenost obecné myšlenky sémantiky na webu s konkrétními technologiemi. Oproti „modelovému“ pohybu jasněji vymezených technologií po klasické „gartnerovské“ křivce má popularita a míra adopce sémantického webu spíše charakter stálého, nepravidelného „vlnění“, vyvolaného souběžným působením řady pozitivních i negativních faktorů.

3.2. Pozitivní faktory pro rozvoj sémantického webu

Dlouhodobě působícími pozitivními faktory jsou zejména tyto:

1. Obě jeho ústřední ingredience jsou trvale přítomné:
    - *Webová* struktura je a v představitelné budoucnosti bude k dispozici, protože se o ni opírá obrovské množství „běžných“, masově využívaných služeb.
    - Totéž platí o podchycení *sémantiky* dat: v naprosté většině případů se vyplatí mít obsahové aspekty strojově zpracovávaných dat zachyceny strukturovanými metadaty. Dokonce i kdyby byla tato metadata (v nejomezenější variantě) zamýšlena jen pro lepší přehled lidských vývojářů aplikací, které s těmito daty mají pracovat, byla uchovávána v různých proprietárních formátech, a nebyla veřejná, stále budou i v období případné dočasné „zimy sémantického webu“[17] představovat skrytý potenciál; z něj pak může kdykoli vyrůst „nová generace“ sémantického webu, byť třeba s využitím jiných standardů, než jsou ty dnešní.
2. Proces převodu „běžných“ dat do podoby dat sémantického webu, jde v případě *veřejnosprávních* dat velkou část cesty společně s procesem „pouhého“ otevírání těchto dat (viz schéma pěti hvězdiček představené v předchozí části). Otevírání datových zdrojů veřejné správy je pak

[17] Termín „zima“ byl pro období útlumu popularity a finanční podpory určitého odvětví informatiky v minulosti použit zejména v souvislosti s umělou inteligencí (období označovaná jako „AI winter“ v 70. a 80. letech minulého století).

neodmyslitelnou součástí „otevřeného vládnutí“, které je v demokratických zemích dlouhodobě podporováno občanskými iniciativami i médii, včetně takových, které technologický koncept sémantického webu sám o sobě příliš nezajímá.

3. Jak už částečně naznačil historický přehled uvedený v předchozí kapitole, sémantický web propojuje velký počet oblastí, které jsou předmětem výzkumu na informatických katedrách *univerzit* po celém světě:
    - umělou inteligence a agentové technologie
    - webové inženýrství
    - databázové technologie
    - formální logiku, včetně práce s neurčitou informací (která je nutně spojena s otevřeným a dynamickým prostředím webu)
    - zpracování přirozeného jazyka (jak kvůli souvislostem mezi abstraktní sémantikou „pojmů“ a jazykovými výrazy, tak s ohledem na možnost automaticky extrahovat strukturovaná data z textových, která převládají na současném „dokumentovém“ webu).

    Absolventi doktorského nebo i nižšího stupně studia na těchto katedrách, kteří se sémantickým webem přišli do styku, pak po nástupu do praxe sehrávají roli „kontaktních bodů“, přes které může akademická sféra získávat firmy a orgány veřejné správy jako partnery do nových projektů v oblasti sémantického webu.

3.3. Faktory brzdící rozvoj sémantického webu

To, že se sémantický web navzdory těmto „motorům“ dosud nestal součástí hlavního proudu informatické praxe, je zřejmě způsobeno brzdicími faktory, za které lze považovat zejména tyto:

1. Sémantický web stojí a padá s ochotou dostatečného počtu subjektů otvírat své datové zdroje druhým. Otevření dat ale u komerčních subjektů může často znamenat vzdání se konkurenční výhody, kterou na trhu mají. To může vést k postoji, kdy firma deklarativně otevírání dat podporuje, ale fakticky svá vlastní data zadržuje a chce spíše jen využívat data vystavená druhými. Reálné otevírání dat je pak podmíněno identifikací *byznys modelu*, který firmě ukáže dostatečně vysoké přínosy, které budou kompenzovat možnou ztrátu ze zřeknutí se exkluzivního přístupu k vlastním datům.
2. V počátcích sémantického webu, kdy se formovaly základy jeho současných standardů pro reprezentaci dat, měly hlavní slovo odborné skupiny vycházející z představy „umělé inteligence na webu“. *Standardy* (zejména jazyk OWL) proto vznikaly s ohledem na úlohy strojového usuzování, a obsahují mnoho sofistikovaných prvků, které nejsou zásadní pro efektivní indexování, vyhledávání a propojování dat. Existence obsáhlých standardů, ze kterých se významněji využívá jen malá část, je pro zájemce z praxe odrazující.

3. Používání otevřených datových standardů jde také proti zájmům velkých korporací, které chtějí v okruhu svých obchodních partnerů nadále prosazovat své vlastní, *proprietární formáty*, aby si partnery udržovaly v závislosti.
4. Skutečnost, že je většina projektů v oblasti sémantického webu buďto závislá na externím grantovém financování (toto platí zejména v Evropě), nebo je realizována firmami typu „start-up“, předem nasměrovanými na akvizici „velkými hráči“ (toto platí zejména v USA), vede k *nedostatečné perzistenci* předkládaných „vzorů k následování“. Vystavená data se brzy stávají neaktuálními, a aplikace pro jejich zpracování nejsou dlouhodobě udržovány. Navzdory tomu, že v každém časovém okamžiku lze na webu nalézt stovky „živých“ výsledků výzkumných i (drobných) komerčních projektů sémantického webu, je velké riziko, že korporátní nebo veřejnosprávní zájemce o vyzkoušení těchto technologií „padne“ při náhodné navigaci či vyhledávání dříve na data či aplikace, které jsou již „mrtvé“, a jeho ochota se v této oblasti angažovat proto ochladne.
5. Rychlý růst popularity klasického, dokumentového webu stál do značné míry na „nízkoprahovosti“ využívání těchto technologií pro *běžné uživatele*, nejen v profesionální sféře, ale i ve sféře volného času. Současný stav technologií sémantického webu naopak vyžaduje jejich spravování odborníky se speciálními znalostmi. Běžní, ani internetově relativně gramotní uživatelé si tudíž k sémantickému webu nemohou získat bezprostřední vztah.
6. Datové slovníky a ontologie pro linked data vznikají spontánně a nekoordinovaně. V důsledku toho jsou často stejné typy objektů reálného světa modelovány odlišně, a to nejen na úrovni názvů (identifikátorů), ale i na úrovni struktury. Příkladem strukturní heterogenity slovníků je situace, kdy je v jednom slovníku pro hudební tématiku navrženo přiřazení prodávaného „hudebního objektu“ XYZ ke třídě hudebních alb pomocí instanciace třídy jako vestavěného konstruktu jazyka RDF (predikát *rdf:type*), např. *„zboziXYZ rdf:type slovnik1:Album“*, zatímco v jiném pomocí predikátu specificky definovaného v daném slovníku, např. *„zboziXYZ slovnik2:jeTypu slovnik2:Album“*. V obou případech je přitom reálná situace, představující *ontologické pozadí* obou slovníků, shodná. Strukturní heterogenita způsobuje, že jsou takové slovníky a na nich založená data např. odlišně zobrazována ve vizualizačních prostředích (takže je vztah mezi nimi špatně patrný pro lidského uživatele), ale je také obtížnější mezi nimi hledat mapování pomocí sofistikovaných nástrojů pro automatické *mapování ontologií* [11], které si zpravidla umějí poradit jen s heterogenitou lexikální.[18]

[18] Tento poslední bod zřejmě není svým významem srovnatelný s předcházejícími, vzhledem k relativně malému rozsahu a nízkému tempu změn slovníků (obsahujícím typy objektů a vztahů) oproti samotným datům (instancím objektů a vztahů). Autor tohoto textu ho zařadil spíše proto, že jde o oblast, ve které je sám vědecky aktivní.

Ve zbývajících šesti sekcích této kapitoly se podíváme na to, jakým způsobem postupně jsou, nebo by mohly být, odbourávány tyto překážky širšího uplatnění sémantického webu.

3.4. Ekonomické modely pro sémantický web

Způsoby, jakými může komerční subjekt profitovat na otevřeném poskytování vlastních dat (případně dat třetích stran), jsou v posledních letech předmětem zájmu po třech „liniích“:

- Samotné firmy tyto možnosti, prozatím s opatrností, prozkoumávají ve svém vlastním podnikání, tj. specifickou praxí.
- Akademické skupiny v oblasti aplikované informatiky, řízení a marketingu do této oblasti rozšiřují předchozí výzkum zaměřený na modely podnikání obecně, a usilují o nalezení obecných modelů jako nových vědeckých poznatků.[19]
- Podobně, avšak praktičtěji, jsou pojaté rozbory prováděné marketingovými analytiky [7] nebo aktivisty otevřené demokracie [3].

Následující přehled možností „monetizace“ otevřených dat (tj. dosažení efektu, kdy akt otevření dat přímo nebo nepřímo vede k finančnímu prospěchu pro vystavovatele) rozebral analytik S. Brinker [7] a doplnil L. Dodds [9]. Z jejich analýzy vybíráme pouze ty varianty, kdy se skutečně jedná o otevřená data a nikoliv jen data poskytovaná uzavřenému okruhu partnerů prostřednictvím webové infrastruktury:

1. *Přímá finanční podpora vystavení dat ve veřejném zájmu.* Vystavení otevřených dat může být chápáno jako prospěšné pro veřejnost, a tudíž podpořeno rozpočtovými a grantovými prostředky z veřejných zdrojů. Tento model financování je v současnosti dominantní. Netýká se ovšem jen dat veřejné správy jako takové, ale také dat organizací, které jsou dlouhodobě podporovány z rozpočtových prostředků, jako je BBC, a v menší míře i ziskových organizací.
2. *Datové SEO.* Indexační algoritmy vyhledávacích služeb jako je Google v současnosti již zohledňují existenci strukturovaných metadat na webu jako pozitivní faktor, indikující důvěryhodnost a obsahovou kvalitu daného zdroje (doporučení pro tvorbu takových metadat jsou uvedena i na jejich stránkách určených pro webmastery[20]). Stránky obsahující RDF se proto dostávají na vyšší pozice ve výsledcích vyhledávání, a v důsledku toho dosahují vyššího počtu návštěv. Tento model je v současnosti hlavním motorem vystavování otevřených dat čistě komerčními subjekty.
3. *Přilákání tvůrců aplikací*. Existence otevřených dat o určité dlouhodobě poskytované službě, zejména určené pro širší veřejnost (např. dopravní

[19] Zřejmě nejvýznamnější výzkumnou organizací specificky orientovanou na tuto problematiku je v současnosti Open Data Institute ve Velké Británii, viz http://www.theodi.org/. V ČR je v r. 2013 na toto téma řešen interní grant VŠE Praha „Ekonomické modely otevřených dat“ (IG406023).

[20] V případě Google nejnověji už i v české verzi:
https://support.google.com/webmasters/answer/99170?hl=cs&ref_topic=1088472

služby), může motivovat různé dobrovolníky k tvorbě nápaditých softwarových aplikací nad těmito daty. Existence webových aplikací, jejichž vývoj firmu nic nestojí, následně na trhu znamená konkurenční výhodu.

4. *Model „freemium"*. Data jsou v omezené formě poskytována zdarma („free"), a v plné či vylepšené („premium") formě za úplatu. Omezenost volných dat může být dána jejich rozsahem, ale ještě spíše propojením na jiná data nebo časovou aktualitou (zájemce, který chce data před určeným časem vystavení, za ně musí zaplatit). Také je možné vydělávat jen na dodatečných službách – konzultacích a úpravách dat na zakázku.
5. *Reklama na úrovni dat*. Součástí dat (která musí být, podle podmínek příslušné otevřené licence, šířena spolu s nimi) je reklama produktů nebo služeb poskytovatele. Výhodou tohoto typu reklamy je, v kontextu sémantického webu, možnost dynamicky generovat reklamní informace vztažené k poskytovaným datům samotným. Zatímco v případě reklamy ve výsledcích webového vyhledávání je vztah mezi reklamou a výsledky zprostředkován pouze společnými klíčovými slovy, v prostředí sémantického webu je možné využívat přesnější ukotvení významu pomocí konceptů z taxonomií a ontologií.
6. *Affiliate marketing*. V tomto případě jde opět o reklamu šířenou prostřednictvím dat, reklama se však týká produktů/služeb třetích stran. Zdrojem příjmů jsou proto partnerské společnosti, pro které je reklama zprostředkována.
7. *Data zvyšující hodnotu.* Data mohou být nabízena ne jen sama o sobě, ale jako doplněk k jiným bezplatným službám, který navyšuje jejich hodnotu pro spotřebitele, tím zvyšuje pravděpodobnost jejich využívání spotřebitelem, které může postupně vést i k přechodu na služby placené.
8. *Budování značky prostřednictvím dat.* Firma má možnost prostřednictvím dat (a zejména slovníků) propagovat svůj pohled na dané odvětví, např. upozorňovat na existenci parametrů, ve kterých její výrobky/služby vynikají. Vedle toho samozřejmě nemůžeme opomenout možnou roli otevřeného poskytování dat jako součásti „filantropického image" (zejména) velkých korporací.

Kromě samotného otevírání dat je pro sémantický web, zejména v pojetí linked data, zásadní také *propojování* dat, zejména na úrovni vazeb mezi identifikátory URI označujícími stejné objekty reálného světa. Růst počtu takových vazeb, ve formě tzv. *sad propojení* („link sets"), výrazně zaostává za růstem objemu individuálně vystavovaných dat. Zatímco pro samotné vystavení dat postačuje rozumět vlastním datům a vhodným způsobem převést interní identifikátory datových objektů na URI, pro propojení vlastních dat na data externí je nutné alespoň do jisté míry porozumět datům cizím. Propojovací nástroje, jako je Silk,[21] sice dosahují přijatelné spolehlivosti, avšak pouze za podmínky předchozí normalizace dat a vyladění parametrů porovnávacích metrik (nehledě na ruční

[21] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/.

dohledání vzájemně odpovídajících prvků datového schématu v obou propojovaných databázích). Tvorba sad propojení dnes stále ještě převážně stojí buďto na „nadstandardním“ entuziasmu kurátora dat nebo podpoře ze strany některého akademického partnera. Zvolna se ale rozbíhá i propojování dat na zakázku jako forma podnikání.[22]

Vedle finančního zisku z obratu může být na druhé straně motivací k vystavování dat *úspora nákladů*. V případě veřejné správy je systematické vystavování dat alternativou k zodpovídání zákonných požadavků mnoha subjektů, které využívají právo na otevřené informace; tímto způsobem lze výrazně ušetřit čas příslušných úřadníků. Pro komerční subjekty může být obdobně zajímavá možnost uplatnění otevřených dat pro úsporu nákladů ve fázi předběžné identifikace možných obchodních partnerů: především s tímto zamýšleným účelem vznikla před několika lety rámcová ontologie pro elektronický obchod *GoodRelations*,[23] která je dnes jedním z nejvýznamnějších datových slovníků na webu.

3.5. Přizpůsobování standardů potřebám praxe

Zřejmě největším břemenem pro praktické uplatnění sémantického webu byl v jeho první fázi ontologický jazyk OWL. Jeho struktura totiž víceméně přesně kopírovala konkrétní variantu tzv. *deskripční logiky*, bez ohledu na to, jak jsou příslušné konstrukce tohoto kalkulu srozumitelné pro neodborníka, i na to, jakou frekvenci jejich využívání lze předpokládat.

Druhá verze jazyka, OWL 2 [20], standardizovaná v r. 2009, komplexitu jazyka sice ještě o něco zvýšila, avšak již rozlišila tři specifické dialekty. Dialekt OWL-EL obsahuje konstrukty používané v rámci rozsáhlých *taxonomických* (zejména biomedicínských) systémů, kde jsou pojmy formálně definované na základě jiných pojmů, zpravidla bez zachycení konkrétních instancí dat. Dialekt OWL-RL obsahuje konstrukty umožňující zachytit pomocí vztahu třídy a podtřídy obdobné jevy, jaké se obvykle vyjadřují pomocí *pravidel*, ontologie je tudíž potenciálně zpracovatelná pravidlovými znalostními systémy. Dialekt OWL-QL obsahuje konstrukty podporující efektivní *dotazování do rozsáhlých bází dat*.

Ještě dále ve směru práce s rozsáhlými daty jde prozatím předběžný návrh dialektu OWL-LD [15], specificky podporujícího požadavky *linked data* na základě rozsáhlé empirické analýzy.

3.6. Synergie s proprietárními modely firem

Rok 2011 byl pro sémantický web průlomový z hlediska zapojení velkých korporací webového byznysu. Tři velké vyhledávače, Google, Yahoo! a (Microsoft) Bing, zveřejnily své společné schéma pro strukturovaná metadata pod názvem *Schema.org*.[24] V témže roce se naplno rozjela propagace protokolu *OpenGraph*[25]

[22] Jedním z takto podnikajících subjektů je německá firma Uberblic, viz http://uberblic.com.
[23] Specifikace ontologie samotné je přístupná z http://purl.org/goodrelations/; informace o jejím poslání jsou shromážděny na http://www.heppnetz.de/projects/goodrelations/.
[24] Viz http://schema.org.
[25] Viz http://ogp.me/. Podobný charakter má i protokol Twitter Cards zavedený mikroblogovací službou Twitter, viz https://dev.twitter.com/docs/cards.

firmy Facebook jako nástroje pro zapojení externích stránek do sítě Facebooku. V obou případech dotyčné firmy zavedly nové ontologické entity (třídy a vlastnosti) pro pojmy, které už byly pokryty existujícími a poměrně populárními slovníky, jako je FOAF nebo GoodRelations. V případě Schema.org navíc došlo k odklonu od RDF, protože jako jazyk „první volby“ pro metadata doporučuje mikrodata a nikoliv RDFa. OpenGraph sice na RDF staví, ale z jeho možností využívá jen základní syntaxi. Proto se okamžitě rozproudily diskuse, zda tento vývoj znamená pro sémantický web úspěch nebo neúspěch (či dokonce krach).

Současný pohled nicméně vyznívá, že podpora webového publikování strukturovaných dat velkými firmami je pro sémantický web z principu pozitivní záležitost. Jakmile jsou data vystavena, byť méně pokročilým způsobem, otevírá se prostor pro jejich další zpracování. Nová verze slovníku GoodRelations např. již explicitně zahrnuje mapování na třídy ze Schema.org.[26] Byla navržena také efektivní transformace mikrodat na RDF [19].

3.7. Perzistentní demonstrace přínosů sémantického webu

Perspektivním a dosud nedostatečně využívaným prostředkem, jak demonstrovat smysluplnost technologií sémantického webu, je jejich nasazení přímo v *akademické sféře*, která je jejich hlavním propagátorem. Univerzity a vědecké instituty mají k dispozici velké objemy dat, která jsou často na webu nesystematicky zveřejňovaná, ať už v podobě statických stránek nebo výstupu z různorodých databází – jde o informace o organizační struktuře, projektech, publikacích (včetně kvalifikačních prací), ale často také o výukové prezentace a jiné učební objekty.

Tato data by mohla sloužit nejen pro prvotní vyzkoušení nových technologií (což se pravidelně děje), ale také pro dosažení dlouhodobého hmatatelného přínosu. Nad otevřenými daty by mohly následně vznikat zejména aplikace pro vyhledávání partnerů pro vědecké projekty, oponentů pro kvalifikační práce, recenzentů pro vědecké konference a časopisy, ale také pro agregování zpráv o pořádaných odborných seminářích a jiných setkáních. Skutečnost, že nad daty vystavenými v souladu s principy Linked Data a s využitím populárních slovníků mohou takové aplikace vytvářet i vývojářské týmy malých firem (nebo vědeckých pracovišť samotných) bez nutnosti zkoumat detaily proprietární reprezentace různých univerzitních informačních systémů, pro vedení univerzit znamenají potenciální úsporu rozpočtových nákladů na nákup externího software i na lidskou práci, oproti situaci, kdy by se podobné funkčnosti mělo dosahovat rozšiřováním oficiálních univerzitních systémů samotných. Univerzity a vědecké instituty mají, oproti grantovým projektům a „startupům“, nesrovnatelně delší dobu trvání, což by právě mělo zajistit relativně vysokou perzistenci vystavovaných dat.

Prvním rozsáhlejším pokusem aplikovat sémantický web na vlastní produkci akademických institucí byl zřejmě Semantic Web Dog Food.[27] V současnosti je

[26] Tento krok se naopak rychle dočkal pozitivní odezvy ze strany konsorcia Schema.org, viz http://blog.schema.org/2012/11/good-relations-and-schemaorg.html.
[27] http://data.semanticweb.org/

iniciativa za akademická linked data představována zejména portálem LinkedUniversities[28] a projektem LinkedUp.[29] V tuzemsku se tento směr pokouší propagovat neformální iniciativa Semanti-CS.[30]

3.8. Sémantický web a koncoví uživatelé

O koncových uživatelích nejčastěji hovoříme jako o těch, kdo pouze „konzumují" data produkovaná softwarovými aplikacemi prostřednictvím uživatelského rozhraní. V tomto směru jejich role v sémantickém webu není příliš velká – uživatelé by v podstatě jako konzumenti dat neměli poznat rozdíl mezi aplikací běžnou a „sémantickou", založenou na otevřené a propojené datové infrastruktuře (s výjimkou toho, že takovou aplikaci dostanou k dispozici v podstatně kratší době, a je také možné ji nechat v podstatně kratší době změnit).

Nelze však zapomínat, že i „koncoví uživatelé", neprofesionálové, mohou mít zájem *vystavovat data*, která vytvoří nebo zpracují. Tento zájem se významně projevil v rámci aplikací „webu 2.0" (zahrnujících zejména vystavování textů, fotografií a videí). Pro sémantický web však mají význam především data strukturovaná. Asi nejzajímavějším nástrojem pro jejich snadné publikování v rámci stránek HTML je *Exhibit* [21], v současnosti podporovaný již ve verzi 3.[31] Na rozdíl od hlavního proudu sémantického webu, určeného spíš pro vystavování dat větších organizací, nevyžaduje znalost jazyků, jako je RDF nebo OWL, ani existujících ontologií; data jsou ve formátu JSON, a schéma je jejich součástí. Prostřednictvím volání specializovaných skriptů je možné přizpůsobit zobrazení dat na webové stránce.

Podobně jako v případě proprietárních schémat korporací je pozitivem samotný fakt, že se strukturovaná data na webu objevují a je možné je kromě vizualizace zpřístupnit i pomocí API. Tvorba mapování mezi etablovanými slovníky a ad hoc vzniklými schématy velkého množství malých vystavovatelů je ovšem méně efektivní. Na druhou stranu může pro rozvoj sémantického webu sehrát pozitivní roli efekt „první zkušenosti" s vystavováním strukturovaných dat u lidí, kteří později ve své profesionální praxi (v korporacích nebo veřejné správě) přijdou do styku s pokročilejšími možnostmi vystavování.

3.9. Zachycení „ontologického pozadí" datových slovníků

Vzhledem k tomu, že významné slovníky jsou typicky odkazovány z mnoha rozsáhlých datasetů, není myslitelné, přinejmenším v krátkodobém horizontu realizovat jejich strukturní sjednocování na syntaktické úrovni. Ontologické inženýrství na sémantickém webu dominovaném instancemi dat musí mít z principu „reaktivní" charakter: zachovávat stávající modely navržené „odspodu" a často ad hoc, a budovat nad nimi nadstavby, které na stávající struktury nabídnou nový pohled.

[28] Viz http://linkeduniversities.org/lu/.
[29] Viz http://linkedup-project.eu/.
[30] Viz http://semanti-cs.org/.
[31] Viz http://www.simile-widgets.org/exhibit3/.

Dosud byly nejpoužívanějším nástrojem pro zachycením hlubší podstaty ontologických konceptů tzv. *vrcholové* („upper-level“) ontologie, obsahující typy a vztahy s nejvyšší úrovní obecnosti a relevantní pro široký okruh oblastí zájmu (fyzické vs. abstraktní entity, individuální vs. kolektivní entity, události probíhající v čase vs. objekty, které se událostí účastní, atd.). Části vrcholových ontologií, jako je SUMO [27] nebo DOLCE [13], bývají do doménových ontologií importovány jako jejich nejvyšší (nejobecnější) úroveň. Požadavek „reaktivního“ ontologického inženýrství ovšem lépe splňují tzv. *modely ontologického pozadí* („ontological background models“), které „značkují“ jednotlivé entity z ontologií na samostatné úrovni, bez přímého zásahu do jejich struktury. Rozšířený je formální jazyk ontologického pozadí OntoClean [17], který je primárně určený pro ontologie s výrazně taxonomickou strukturou, a zachycuje pojmy jako je rigidita (zda je příslušnost objektů k dané třídě v čase proměnná nebo daná nastálo) nebo existence jednoznačného kritéria identity (pro objekty dané třídy). Pro linked data s jejich převážně síťovou strukturou faktů byl nově navržen jazyk PURO [28], který zkoumá odlišení jednotlivých objektů („particulars“ - P) od typů objektů („universals“ - U), a vztahů („relationships“ - R) od skutečných objektů („objects“ - O). V obou případech lze následně testovat ontologickou koherenci „označkovaných“ modelů, případně i získat hlubší vhled do jejich podstaty.

4. Výzkum a praxe sémantického webu v ČR

Historicky se z hlavních tematických oblastí sémantického webu nejprve rozvíjelo ontologické inženýrství. Na katedře kybernetiky FEL ČVUT v Praze byla vyvinuta softwarová řešení pro přístup k bázím znalostí s pomocí ontologií [22]. Katedra informačního a znalostního inženýrství VŠE v Praze se podílela na vzniku COMM - Core Ontology of Multimedia [2], později na témže pracovišti vznikla sada nástrojů pro transformaci ontologií s využitím transformačních vzorů[32] (PatOMat) [29] a již zmíněný jazyk modelů ontologického pozadí [28]. Tým Masarykovy univerzity se dlouhodobě věnuje vývoji tzv. wordnetů jako specifického typu lexikální ontologie.

V oblasti linked data stojí za zmínku vytvoření první verze české DBpedie[33] týmem FIT ČVUT v Praze; tým z VŠE v Praze (v rámci projektu EU LinkedTV) zase navrhl obohacení DBpedie o přiřazení dodatečných typů,[34] s využitím analýzy přirozeného jazyka na stránkách Wikipedie. Publikování linked data transformovaných z českých datových zdrojů, ale i některých zahraničních zdrojů (zejména pro oblast veřejných zakázek) se věnuje český tým zapojený do projektu EU LOD2,[35] složený z pracovníků MFF UK a VŠE v Praze; výsledky jeho činnosti jsou prezentovány na serveru http://opendata.cz.

Pokud jde o další specifické podoblasti výzkumu sémantického webu, lze kupříkladu zmínit, že tým FIT ČVUT vyvinul sémantickou nadstavbu pro výběr webových API [10], a tým FIT VUT Brno se dlouhodobě věnuje zejména propojení technik zpracování přirozeného jazyka se sémantickým webem. Výměnu informací

[32] http://patomat.vse.cz
[33] http://cs.dbpedia.org/.
[34] http://ner.vse.cz/datasets/linkedhypernyms/
[35] http://lod2.eu.

v rámci české (a slovenské) „sémantické“ komunity, případně i vystavování (zejména akademických) dat v podobě linked data se snaží podporovat neformální iniciativa *Semanti-CS*.[36]

5. Závěry

Sémantický web jako obor výzkumu prošel za téměř 15 let své existence pestrým vývojem. Z okrajového tématu se stal oborem, o kterém se často hovoří, a který propojuje řadu jinak velmi odlišných vědeckých komunit. Vývoj nových technologických řešení se opírá o etablované standardy (navržené zejména po linii W3C), zároveň však stále vznikají návrhy na standardy nové. Jak bylo v textu ukázáno, reálné přínosy sémantického webu se nejvíce ukazují v oblasti linked data. Technologická podpora jejich životního cyklu, od jejich vystavení až po využití v aplikacích, je již poměrně rozsáhlá. Na druhou stranu u velké části software nelze mluvit o zralosti a stabilitě. Pro významné rozšíření těchto technologií do praxe však nejvíce chybí ustálené procesy a ekonomické modely.

V textu jsme záměrně (téměř) opomenuli souvislost mezi strukturovanými daty na webu a technologiemi zpracování přirozeného jazyka. Přestože jsou mezi oběma oblastmi mnohostranné vazby, uplatnění lingvistických technik v rámci sémantického webu a linked data má zatím spíše ad hoc charakter, a doporučené postupy teprve vznikají.

Literatura

[1] ADIDA, Ben, HERMAN, Ivan, SPORNY, Manu, BIRBECK, Mark. *RDFa 1.1 Primer. Rich Structured Data Markup for Web Documents. W3C Working Group Note*. [online]. 2012-06-07 [cit. 2013-07-04]. Dostupné z: http://www.w3.org/TR/xhtml-rdfa-primer/

[2] ARNDT, Richard, TRONCY, Raphaël, STAAB, Steffen, HARDMAN, Lynda, VACURA, Miroslav. In: Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web konference (ISWC'07/ASWC'07), 30-43, Springer-Verlag, 2007.

[3] BERG, Michal. Otevřená data a jejich byznys modely: kde v nich hledat peníze? [online]. 2013-04-08 [cit. 2013-07-04]. Dostupné z: http://www.datablog.cz/clanky/byznys-modely

[4] BERNERS-LEE, Tim. Design Issues: Linked Data. [online]. [cit. 2013-07-04]. Dostupné z: http://www.w3.org/DesignIssues/LinkedData.html

[5] BERNERS-LEE, Tim, HENDLER, Jim, LASSILA, Ora. The Semantic Web. *Scientific American*, Květen 2001, 29-37.

[6] BRICKLEY, Dan, MILLER, Libby. FOAF *Vocabulary Specification 0.98. Namespace Document*. [online]. 2010-08-09 [cit. 2013-07-04]. Dostupné z: http://xmlns.com/foaf/spec/

[36] http://semanti-cs.org/

[7] BRINKER, Scott. 7 business models for linked data. [online] 2010-01-08 [cit. 2013-07-04]. Dostupné z: http://chiefmartec.com/2010/01/7-business-models-for-linked-data/

[8] COBDEN, Marcus, BLACK, Jennifer, GIBBINS, Nicholas, CARR, Les, SHADBOLT, Nigel. A Research Agenda for Linked Closed Data. In: Proceedings of the Second International Workshop on Consuming Linked Data (COLD2011), held in conjunction with the 10th International Semantic Web Conference 2011 (ISWC 2011), Bonn. Editoři Olaf Hartig, Andreas Harth, Juan Sequeda. Aachen : CEUR-WS, 2011. Dostupné z: http://ceur-ws.org/Vol-782/CobdenEtAl_COLD2011.pdf

[9] DODDS, Leigh. Thoughts on linked data business models. [online] 2013-01-10 [cit. 2013-07-04]. Dostupné z: http://blog.ldodds.com/2010/01/10/thoughts-on-linked-data-business-models/

[10] DOJCHINOVSKI, Milan, KUCHAŘ, Jaroslav, VITVAR, Tomáš, ZAREMBA, Maciej. Personalised Graph-Based Selection of Web APIs. In: The Semantic Web – ISWC 2012. Springer-Verlag, LNCS 7649, 2012, 34-48.

[11] EUZENAT, Jerôme, SHVAIKO, Pavel. *Ontology matching*. Berlin: Springer-Verlag, 2007, ix, 333 s. ISBN 978-3-540-49611-3.

[12] FARQUHAR, Adam, FIKES, Richard RICE, James. The Ontolingua Server: a tool for collaborative ontology construction. *International Journal of Human-Computer Studies*. 1997, vol. 46, issue 6, s. 707-727. DOI: 10.1006/ijhc.1996.0121. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S1071581996901214

[13] GANGEMI, Aldo, GUARINO, Nicola, MASOLO, Claudio, OLTRAMARI, Alessandro, SCHNEIDER, Luc. Sweetening Ontologies with DOLCE. In: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW '02). Springer-Verlag, 2002, 166-181.

[14] GEARON, Paul, PASSANT, Alexandre, POLLERES, Axel. *SPARQL 1.1 Update. W3C Recommendation*. [online]. 2013-03-21 [cit. 2013-07-04]. Dostupné z: http://www.w3.org/TR/sparql11-update/

[15] GLIMM, Birte, HOGAN, Aidan, KRÖTZSCH, Markus, POLLERES, Axel. OWL: Yet to arrive on the Web of Data? In: Proceedings of the WWW12 Workshop on Linked Data on the Web (LDOW2012), Lyon. Aachen : CEUR-WS, 2012. Dostupné z: http://ceur-ws.org/Vol-937/ldow2012-paper-16.pdf.

[16] GÓMEZ-PÉREZ, Asunción, FERNÁNDEZ-LOPEZ, Mariano, CORCHO, Oscar. *Ontological engineering: with examples from the areas of knowledge management, e-commerce and the semantic web.* London: Springer-Verlag, 2004, xii, 403 s. ISBN 18-523-3551-3.

[17] GUARINO, Nicola, WELTY, Christopher A. An Overview of OntoClean. In: *Handbook on Ontologies*. International Handbooks on Information Systems Springer-Verlag, 2009, 201-220.

[18] HEFLIN, Jeff, HENDLER, James, LUKE, Sean. SHOE: *A Knowledge Representation Language for Internet Applications.* Technical Report CS-TR-4078 (UMIACS TR-99-71), Dept. of Computer Science, University of Maryland at College Park. 1999. Dostupné z: http://www.cs.umd.edu/projects/plus/SHOE/pubs/techrpt99.pdf

[19] HICKSON, Ian, KELLOGG, Gregg, TENNISON, Jeni, HERMAN, Ivan. *Microdata to RDF. Transformation from HTML+Microdata to RDF. W3C Interest Group Note.* 2012-10-09 [cit. 2013-07-04]. Dostupné z: http://www.w3.org/TR/microdata-rdf/.

[20] HITZLER, Pascal, KRÖTZSCH, Markus, PARSIA, Bijan, PATEL-SCHNEIDER, Peter F., RUDOLPH, Sebastian. OWL 2 Web Ontology Language Primer (Second Edition). W3C Recommendation. 2012-12-11 [cit. 2013-07-04]. Dostupné z: http://www.w3.org/TR/owl2-primer/.

[21] HUYNH, David F., KARGER, David R., MILLER, Robert C. Exhibit: lightweight structured data publishing. In: Proceedings of the 16th international conference on World Wide Web (WWW'07). New York : ACM, 2007.

[22] KŘEMEN, Petr, KOUBA, Zdeněk. Ontology-Driven Information System Design. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 42(3): 334-344 (2012).

[23] LENAT, Douglas B, GUHA, Ramanathan V. *Building large knowledge-based systems: representation and inference in the Cyc project*. Reading, Mass.: Addison-Wesley Pub. Co., 1989c1990, xix, 372 p. ISBN 02-015-1752-3.

[24] MAALI, Fadi, LOUTAS, Nikolaos. SPARQL 1.1 and RDF Faceted Browsing. [online]. [cit. 2013-07-04]. Dostupné z: http://140.203.154.100/content/sparql-11-and-rdf-faceted-browsing

[25] MANOLA, Frank, MILLER, Eric. *RDF Primer. W3C Recommendation*. [online]. 2004-04-10 [cit. 2013-07-04]. Dostupné z: http://www.w3.org/TR/rdf-primer/

[26] MOTTA, Enrico. *Reusable Components for Knowledge Models: Principles and Case Studies in Parametric Design*. Amsterdam: IOS Press, 1999.

[27] NILES, Ian, PEASE, Adam. *Towards a standard upper ontology*. In: Proceedings of the international conference on Formal Ontology in Information Systems, FOIS 2001, ACM, 2-9.

[28] SVÁTEK, Vojtěch, HOMOLA, Martin, KLUKA, Ján, VACURA, Miroslav. Metamodeling-Based Coherence Checking of OWL Vocabulary Background Models. In: 10th OWL: Experiences and Directions Workshop (OWLED 2013), Montpellier, France, 26th-27th May, 2013.

[29] ŠVÁB-ZAMAZAL, Ondřej, SVÁTEK, Vojtěch, IANNONE, Luigi. Pattern-based ontology transformation service exploiting OPPL and OWL-API. In: Proceedings of the 17th International conference on Knowledge engineering and Knowledge management (EKAW'10), 105-119, Springer-Verlag, 2010.