Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Ing. Michal Šmiraus

Výzkum metod optimalizace Sémantického webu s využitím ontologií

Research on optimization methods of Semantic Web using ontologies

Disertační práce

Studijní program:	Inženýrská informatika
Studijní obor:	Inženýrská informatika
Školitel:	doc. Ing. Zdenka Prokopová, CSc.

Zlín, duben 2015

RESUMÉ

Informace poskytované na celosvětové síti WWW (World Wide Web) v současné době zahrnují velké množství informací a dokumentů, jež jsou sice z velké části srozumitelné lidem, avšak již méně srozumitelné pro automatizované vyhledávací stroje, které zatím nedokáží přesněji uspokojivě identifikovat, co obsah dané stránky ve smyslu svého skutečného významu vyjadřuje. Spolu se vzrůstajícím množstvím dostupných informací na současném Webu tak vzniká sílící potřeba efektivně označit, rozeznat a zpracovat relevantní informace nikoli jen na základě prostého full-textového vyhledávání pomocí klíčových slov, ale také na základě znalostních bází s využitím technologií Sémantického webu a ontologií, jejichž předmětem je na jedné straně vývoj obecných jazyků, metodik či softwarových nástrojů a na druhé straně pak také konstrukce samotných ontologií, popisující nejrůznější věcné oblasti a aplikace, které je využívají.

Klíčová slova: RDF, sémantika, ontologie, optimalizace

ABSTRACT

The information provided on the global network WWW (World Wide Web) currently include a lot of information and documents which are although largely comprehensible to people, but has been less clear for automated search engines which are currently unable to satisfactorily identify precisely what the content of this page in terms of its importance expresses. Along with the increasing amount of information available on the global Web network so there is a need to effectively identify, recognize, and process the relevant information not only on the basis of a simple full-text search using key words, but also on the basis of knowledge bases using Semantic Web technologies and ontologies, designed on the one hand, the general development languages, methodologies and software tools, but on the other hand, also design their own ontologies describing various substantive areas even applications that will use them.

Keywords: RDF, semantic, ontology, optimization

Na tomto místě bych rád, ještě před uvedením vlastní práce, poděkoval své školitelce **doc. Ing. Zdence Prokopové, Csc.** za kvalitní odborné vedení, věcné připomínky a poskytnuté konzultace při zpracování této disertační práce.

Poděkování patří také všem kolegům a přátelům z akademického prostředí mé „alma mater“, se kterými jsem v průběhu svého doktorandského studia mohl vzájemně sdílet či diskutovat výsledky své práce a v neposlední řadě také přítelkyni a rodině za morální podporu při studiu.

„The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.“

Tim Berners-Lee

Prohlašuji, že jsem na předkládané disertační práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků, je-li to uvolněno na základě licenční smlouvy, budu veden jako spoluautor.

Ve Zlíně, dne 1.5. 2015

OBSAH

RESUMÉ **2**
ABSTRACT **2**
ÚVOD DO PROBLEMATIKY **6**
1. ZHODNOCENÍ SOUČASNÉHO STAVU **8**
1.1 Stávající web 8
1.2 Sémantický web 9
2. CÍLE DISERTAČNÍ PRÁCE **11**
3. TEORETICKÝ ZÁKLAD **12**
3.1 Technologie Sémantického webu 12
3.1.1 URI indentifikátory 15
3.1.2 XML syntaxe 16
3.1.3 RDF a vyjádření metadat 17
3.1.4 RDFS pro popis slovníků 19
3.1.5 Mikroformáty a Mikrodata 22
3.1.6 SPARQL dotazovací jazyk 25
3.1.7 GRDDL transformace 27
3.2 Ontologie a znalostní inženýrství 28
3.2.1 OWL jazyk pro zápis ontologií 30
3.2.2 RIF pravidla pro odvozování 32
3.2.3 Extrakční ontologie 34
3.2.4 Učení ontologií 36
3.2.5 Řízení znalostí 38
3.3 Interoperabilita Sémantického webu 40
3.4 Sémantická analýza a interpretace 42
3.4.1 Formální definice 43
3.4.2 Statistický model 44
3.4.3 Rozpoznání jmenných entit 45
3.5 Dostupné softwarové nástroje 47
3.5.1 Protégé 47
3.5.1 GoodRelations 49
3.5.1 Ontopia 51
3.5.2 Anzo 52
3.5.3 mSpace 53
3.6 Sémantické vyhledávače 54
3.6.1 Swoogle 54
3.6.2 Open Calais 56
3.6.3 Kngine 57
3.6.1 Sindice 58
3.6.2 DuckDuckGo 59

4. DOSAŽENÉ VÝSLEDKY 60
4.1 Provedení dotazníkového šetření 60
4.1.1 Cíle a metodika výzkumu 61
4.1.2 Předpoklady 61
4.1.3 Vyhodnocení výsledků 62
4.2 Vytvoření informačního portálu pro podporu výzkumu 69
4.2.1 Uživatelsky rozšiřitelná encyklopedie 69
4.2.2 Edukační tutoriál 70
4.2.3 Znalostní testy 70
4.3 Návrh a implementace softwarového anotačního nástroje 71
4.3.1 Specifikace požadavků 71
4.3.2 Funkční programové atributy 72
4.3.3 Analýza a návrh anotačního procesu 73
4.4 Praktické možnosti využití zpracování přirozeného jazyka 79
4.4.1 Znalostní báze a doménová ontologie 80
4.4.2 Systémová architektura a zpracování dotazu 81
4.4.3 Sémantická interpretace a vyhledávání 82
4.5 Komparace sémantických inferenčních mechanismů 85
4.5.1 Sémantické frameworky pro práci s ontologiemi 85
4.5.2 Funkční parametry odvozovacích modulů 86
4.5.3 Výsledky experimentálního měření 89
5. ZVOLENÉ METODY ZPRACOVÁNÍ 93
6. VYUŽITELNOST VÝSLEDKŮ 94
6.1 Přínos pro vědu 94
6.2 Přínos pro praxi 94
ZÁVĚR 95
POUŽITÁ LITERATURA A ZDROJE 99
SEZNAM SYMBOLŮ A ZKRATEK 103
SEZNAM OBRÁZKŮ 104
SEZNAM TABULEK A GRAFŮ 105
PUBLIKAČNÍ ČINNOST AUTORA 105
PROFESNÍ ŽIVOTOPIS 107

ÚVOD DO PROBLEMATIKY

Web a jeho služby se vlivem masivního rozvoje Internetu staly nedílnou součástí našich životů, využívanou prakticky denně, v mnoha případech i nevědomě. S touto již plně zažitou a rozšířenou technologií nám procházení Webu přijde jednoduché a přirozené, i když tomu tak ještě v nedávné historii zdaleka nebylo. Nyní navíc, díky mohutnému rozvoji technologie rychlých datových přenosů v mobilních sítích, se stává Web přístupným prakticky odkudkoliv, což současně implikuje možnosti jeho dalšího využití.

Primárními konzumenty webového obsahu na síti měli být z historického pohledu vždy především lidé. Při neustále rostoucí velikostí Webu je však s neustále nabývajícím množstvím dostupných dokumentů stále obtížnější efektivně vytěžit požadované informace, a to zejména právě kvůli velkému objemu dat a jejich nedostatečnému provázání, což ve své podstatě umožnilo vznik dnes prakticky nepostradatelných webových vyhledávačů, které pracují se systémem klíčových slov a vytvářením indexů jsou následně schopny pro dané klíčové slovo zjistit odpovídající množinu URL adres dokumentů korespondujících se zadaným klíčovým slovem (resp. množinou klíčových slov), jenž se v daném umístění vyskytuje. Najde-li však vyhledávač vhodný dokument pouze na základě analýzy klíčových slov, nemusí to ještě znamenat, že v něm uživatel automaticky nalezne to, co v daný moment hledá a očekává.

Způsob, jakým vyhledávače pracují, je totiž determinován architekturou stávajícího Webu, která je ve své podstatě přizpůsobena výše uvedenému faktu, tedy že data na Webu jsou primárně určena lidskému uživateli. Vyhledávač tak s obsahem (viditelná data či skrytá metadata) pracuje pouze jako s prostým textem, resp. binárními daty, ale o významu nebo kontextu jednotlivých spojení netuší nic. Člověk přitom většinou hledá informaci, nikoli jen klíčové slovo, které je jen jakýmsi prostředkem k tomu, jak se ke kýžené informaci dostat.

Základní myšlenka Sémantického webu, prezentovaného jako rozšíření Webu klasického, je tedy v principu založená na datech, která jsou čitelná jak pro člověka, tak i pro stroje. Uživatelům Internetu je v současné době k dispozici celá řada webových aplikací a služeb (elektronické noviny, sociální sítě atd.), jež jsou však často pouhými uzavřenými systémy a nesdílí mezi sebou žádná data, přestože by mohly. Typickým příkladem mohou být uživatelské účty pro různé aplikace, e-shopy apod. Uživatel musí opakovaně vkládat své údaje během registrace na každém jednotlivém serveru. Je to způsobeno tím, že dané aplikace nemají interoperabilní vazby na ostatní aplikace a nemohou proto využít již existující údaje. Sémantický web je naopak na interoperabilitě založený, avšak k tomu je zapotřebí sdílet společně s daty také jejich význam [32].

Tim Berners-Lee – autor a průkopník původní myšlenky Sémantického webu, vidí budoucnost celé platformy spojenou s nutností učinit Web více spolupracujícím, srozumitelným a strojově zpracovatelným médiem, což by však znamenalo posun dále od prostého získávání HTML stránek z webového serveru k automatizovanému inteligentnímu vyhledávání nejen informací, ale také vlastních faktických znalostí vyvozených pomocí usuzovacího principu podobnému lidskému způsobu uvažování [16].

Přístupem zvoleným pro zajištění sémantických funkcí přitom není vytvoření jediného univerzálního přístupu k vyjadřování významu informací, ale zajištění podpory procesu práce s významovými informacemi na základě integrovaného využívání sémantických jazyků a přístupů ke konceptuálnímu modelování.

Berners-Lee [6] charakterizuje stávající Web, označovaný někdy též Web 2.0, jako množinu sídel, majících svá vlastní data, která však mezi sebou navzájem nesdílejí. Koncept (dosud plně nespecifikovaného) Webu 3.0, nebo také tzv. *Webu Dat*, je stejně jako původní verze založen na indexovaném množství různých dokumentů na Webu, ale současně i na množství nezávislých webových aplikací, kdy každá aplikace je schopna používat různá data a může běžet lokálně či v prohlížeči. Vše by také mělo být „bezešvé“ díky integraci.

Myšlenku Sémantického webu vyslovil Berners-Lee již v roce 2001, avšak tento koncept se sám o sobě dosud nesetkal s širším uplatněním, i přes rozsáhlé možnosti využití, které nabízí. Berners-Lee a mnozí další propagátoři tento fakt připisují tomu, že je ve své současné podobě příliš složitý. Lidé musí mít nejprve jednoduchou možnost vložit či upravit sémantická data na stávajícím Webu a sami tak dopomoci k masivnějšímu rozšíření Webu sémantického.

Širší podpora uživatelsky přívětivých softwarových nástrojů pro sémantickou anotaci nového či již stávajícího obsahu bez předchozích teoretických znalostí je v současnosti zásadní překážkou pro postupné přetváření Sémantického webu z dosud převážně akademické sféry do oblasti praktického využití.

Pro následný další rozvoj celého odvětví může dopomoci zavedení uživatelsky snadno ovladatelného anotačního nástroje dostupného online prostřednictvím některého z již rozšířených a populárních systémů pro správu obsahu (CMS) společně se zvýšenou měrou informovanosti o tom, co přesně Sémantický web je a jaké výhody běžnému uživateli poskytuje.

Obě tyto vysoce aktuální potřeby se snaží reflektovat i předkládaná disertační práce, která zejména ve své praktické části dokládá aplikovatelné výsledky spojené s širším využitím možností technologie Sémantického webu.

1. ZHODNOCENÍ SOUČASNÉHO STAVU

Pro stanovení teoretických východisek disertační práce bylo nutno nejprve zmapovat a zhodnotit aktuální stav zkoumané problematiky. V následujících podkapitolách proto bude nejprve prezentován nezávislý pohled na současnou podobu již zavedeného Webu následovaný úvodním teoretickým porovnáním s jeho stále doposud plně neprosazeným sémantickým rozšířením.

1.1 Stávající web

World Wide Web ve své stávající podobě je globální služba poskytovaná v rámci vzájemně propojené počítačové sítě Internet. Je to však také v současnosti již značně obsáhlý informační prostor, ve kterém jsou jeho jednotlivé prvky a popisované koncepty adresovány pomocí systému globálních identifikátorů *URI (Uniform Resource Identifier)*, které umožňují identifikovat libovolnou reprezentaci informačního zdroje. Specifikace URI identifikátorů je založena na závazné syntaxi, která se skládá z několika komponent [33]:

```
schema : [//autorita][/cesta][?dotaz][#fragid]
```

kde *schema* označuje nejčastěji komunikační protokol, prostřednictvím něhož je daný informační zdroj dostupný, *autorita* označuje vlastníka zdroje, kterým bývá zpravidla doménové jméno serveru, *dotaz* určuje hodnoty odesílaných argumentů a *fragid* představuje identifikátor sekundárního zdroje „fragmentID“, ke kterému nevede přímá cesta (např. kotva v hypertextovém dokumentu). Vlastnictví URI je delegováno společností IANA URI scheme registry [60], která v případě některých URI schémat deleguje tuto pravomoc dále. Proces alokace jmenných prostorů se v různých URI schématech liší. Podle toho, zda je URI dereferencovatelné se dle funkční specifikace RFC 1630 dále rozlišuje *URL (Uniform Resource Locator)* často reprezentován adresou a *URN (Uniform Resource Name)*, který však nutně nemusí být dereferencovatelný.

Dalšími základními prvky současného Webu zajišťující manipulaci s reprezentací jednotlivě identifikovaných zdrojů jsou *HTTP (Hypertext Transfer Protocol)* – dle RFC 2616 v aktuální ve verzi 1.1, který zprostředkuje interakci mezi klientem a serverem při bezestavové komunikaci typu dotaz-odpověď a *HTML (Hypertext Markup Language)*, což je značkovací jazyk pro strukturované formátování textu určeného pro Web. Provázání HTML jednotlivých dokumentů pomocí odkazů je charakteristickým rysem Webu, kterého využívají jak *weboví agenti* (servery, proxy, prohlížeče, multimediální přehrávače nebo vyhledávací roboti) tak i *webové vyhledávače* při vytváření svých indexů během procházení webu (tzv. *crawling*, viz kapitola 3.6).

1.2 Sémantický web

Sémantický web[1] poskytuje obecný rámec umožňující sdílení dat a jejich znovupoužitelnost mezi aplikacemi i lidmi navzájem. Od klasického konceptu WWW se Sémantický web významně neliší a je prezentován jako rozšíření či „vylepšení“, které dává původnímu Webu mnohem větší upotřebení a stává se skutečným v okamžiku, kdy se lidé různého oboru nebo povolání dohodnou na společném schématu reprezentace informací, o které se zajímají [16].

V rámci mezinárodního konsorcia W3C (w3.org) tak vznikla iniciativa, která zformulovala Sémantický web jako množinu technologických vrstev, které mají informacím na internetu zajistit strojovou čitelnost a interpretovatelnost.

Při vytváření Sémantického webu je prvotním předpokladem zajištění konceptualizace dostupných dat, přičemž klíčový nástroj pro další rozšíření zde představuje *ontologie*, což je svým způsobem serializovaná reprezentace znalostí, určená k následnému sdílení a opakovanému použití v nejrůznějších aplikacích. Standardizovaným popisem Sémantického webu mohou být tedy bez omezení prakticky všechny známé webové zdroje, ať již se jedná o klasické webové stránky, textové dokumenty, obrázky, videa, zvuky apod.

Všechny tyto zdroje pak mohou obsahovat stejné charakteristické údaje (vypovídající o autorovi, klíčových slovech, typu zdroje a dalších metadatech), což teoreticky dává konečnému uživateli možnost vyhledávání a práce v síti WWW obdobně jako při filtrování výsledků z relační databáze, a to pomocí dotazovacího jazyka velmi podobného klasickému SQL, avšak s důrazem na vysokou přesnost a relevanci v odpovědích na hledaný dotaz [19].

URI v prostředí Sémantického webu je definováno modelem *RDF (Resource Description Framework)*, který jednotlivé informace formuluje prostřednictvím tzv. *trojic (subjekt-predikát-objekt)*, kde predikát vymezuje relaci mezi objektem a subjektem. Jelikož RDF samotný je pouze model, pro jeho textovou reprezentaci bylo vyvinulo několik druhů serializací – od nejstarší RDF/XML přes dobře čitelnou Turtle nebo primitivní N-Triples až po v praktickém nasazení zatím nejvhodnější RDFa (viz softwarový anotační nástroj, v části 4.3).

Abychom však RDF mohli smysluplně využít, je třeba jednotlivým slovům přiřadit jejich vlastní význam, k čemuž v praxi slouží technologie již výše uvedených ontologií a popisných slovníků, neboli RDF schémat (např. FOAF, SIOC, GoodRelations apod.), která jsou definována pomocí standardů *RDFS (Resource Description Framework Schema)* případně složitějším *OWL (Ontology Web Language)*, které jsou ontologiemi samy o sobě.

[1] Z hlediska užívaného názvosloví, které v této oblasti není doposud pevně ustanoveno, bude v rámci textu této disertační práce namísto obecného výrazu „sémantický web“ použit termín „Sémantický web“, který označuje a lépe zde vystihuje konkrétní využití technologie RDF jako hlavního modelu pro reprezentaci informací.

Důležitým krokem ve vývoji Sémantického webu bylo také uvedení *SPARQL (Simple Protocol And RDF Query Language)*, dotazovacího jazyka, určeného k manipulaci s RDF databázemi a ontologiemi, který tak umožňuje naplnění koncepce strojově čitelných dokumentů a „indikuje schopnost strojů řešit dobře definované problémy prováděním dobře definovaných operací na existujících dobře definovaných datech“ [7].

Z této vize vychází také publikační model Linked Data (označovaný často jako podmnožina Sémantického webu) pro zveřejňování strukturovaných dat na Webu, jenž je vymezen několika principy a doporučeními, která stanovují, jakým způsobem mají být tato data v sémantické podobě na Web publikována, přičemž je zde patrná určitá snaha o zajištění úrovně užitečnosti informací, kdy například komunita kolem projektu Linking Open Data takto zveřejňuje veřejné databáze jako Linked Data. Příkladem praktické realizace takového modelu pro reprezentaci informací je DBpedia využívající data extrahovaná z Wikipedie.
Praktická část této práce (v podkapitole 4.2) za pomoci rozšíření Semantic MediaWiki realizuje obdobný encyklopedický projekt, avšak v omezené podobě se striktním zaměřením pouze na informační oblast Sémantického webu.

W3C uvolnilo specifikace pro jazyky a technologie, které Sémantický web potřebuje, aby se mohl rozvíjet dále za hranice konkrétních či úzce specializovaných systémů. Prozatím jde o převážně vývojovou aktivitu, na níž se podílí řada univerzitních pracovišť, nicméně existují již také první aplikace používané v průmyslových odvětvích, s potřebou sdílení faktických informací a automatickou dedukcí následných relací. Například distributoři elektrické energie v USA užívají technologie Sémantického webu pro výměnu modelů energetických systémů mezi systémovými operátory. MITRE Corporation zase uplatňuje nástroje Sémantického webu ve snaze lépe interpretovat pravidla vzniku ozbrojených střetů. U.K.’s National Mapping Agency používá sémantický web za účelem přesnějšího a méně nákladného generování geografických map [16]. Harper’s Magazine nasadil sémantické ontologie na svém Webu, aby představovaly anotované časové osy aktuálních událostí, které jsou automaticky propojeny se souvisejícími články o těchto událostech.

To všechno jsou jen některé z dosud prakticky uskutečněných projektů, které dokazují, že Sémantický web je široká množina technologií mající možný potenciál v mnoha oborech lidské činnosti. Berners-Lee [6] však současně vyjadřuje názor, že doposud není k dispozici valné množství dostupné literatury, která by pomohla dalším projektům, podnikům a organizacím uvést Sémantický web do reality v širším pojetí původně zamýšleného konceptu. Tento faktický nedostatek dostupných informačních zdrojů o dané problematice (konkrétně také v českém jazyce) konec konců odhaluje také provedené dotazníkové šetření v rámci praktické části této práce, v podkapitole 4.1.

2. CÍLE DISERTAČNÍ PRÁCE

Za motivací k sepsání této disertační práce stála aktuální absence souhrnné koncepce a technologií, které by spojovaly možnosti současného Webu s technologiemi určenými pro Web sémantický a z toho vyplývající konkrétní požadavky na široce dostupný a v praxi snadno použitelný prostředek pro sémantickou anotaci webových dokumentů, která je nezbytným předpokladem pro jakoukoli formu následného zpracování takto označených dat.

Na základě motivace k vlastnímu výzkumu, vycházející též z potřeby rozvoje povědomí o optimalizaci a použitelnosti vize Sémantického webu, bude jedním z cílů práce navrhnout a realizovat univerzální a snadno rozšiřitelný softwarový nástroj pro podporu vkládání sémantiky do obsahu webových dokumentů. Součástí řešení bude i analýza problematiky Sémantického webu a návrh prakticky využitelných možností zpracování dotazů v přirozeném jazyce.

Teoretická část práce bude věnována analýze a zmapování stavu současného Webu, kde bude poukázáno na některé jeho aktuální nedostatky. Následně bude představena myšlenka Sémantického webu, přičemž budou prezentovány výhody a možnosti souvisejících technologií, které budou dále rozvedeny v praktické ukázce návrhu pro využití technologie dotazování prostřednictvím přirozeného jazyka nad znalostní bází specificky orientované ontologie.

Výzkumná část dále počítá s vyhodnocením dotazníkového šetření ohledně zpracovávané problematiky a s vytvořením webového portálu na bázi otevřené informační encyklopedie, za účelem sdružování zájemců o danou problematiku Sémantického webu z řad odborné i laické veřejnosti, k následné diskusi či předávání vlastních znalostí a zkušeností.

Praktická část práce bude věnována návrhu a implementaci softwarového nástroje pro sémantickou anotaci k zajištění širší podpory rozvoje Sémantického webu prostřednictvím realizace editačního rozšíření pro uživatelsky oblíbený způsob práce s informacemi prostřednictvím WYSIWYG editoru. Účelem navržené komponenty je zprostředkovat uživateli možnost pokročilé sémantické anotace ve formátu RDFa s cílem podpory rozvoje Sémantického webu, který pak může být dále využíván jak aplikacemi v rámci téhož systému (např. znalostní báze v prostředí podnikového intranetu), tak i externími nástroji jakými jsou sémantické vyhledávače, které zahrnují podporu inferenčních systémů, jež mohou produkovat nové informace a znalosti.

Závěrem praktické části bude zhodnoceno a statisticky porovnáno užití několika vybraných inferenčních systémů v prostředí doménové ontologie.

3. TEORETICKÝ ZÁKLAD

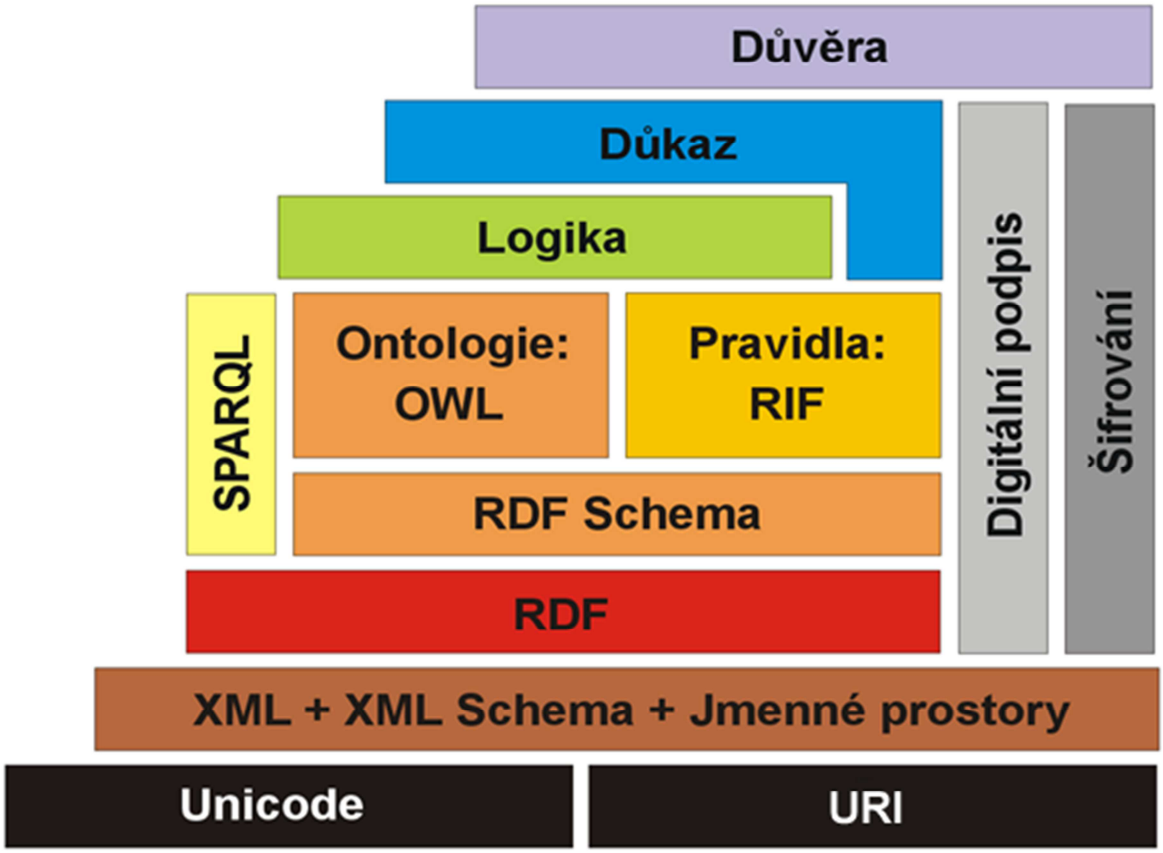
Jak plyne z již v úvodu citovaného článku [6], termín Sémantický web nepředstavuje novou verzi Webu, ale pouhé rozšíření konceptu a doplnění popisných meta dat do toho stávajícího. Tato data by přitom měla být zapsána pomocí strojově čitelných jazyků a jejich součástí by měla být také použitá slovní zásoba a soubor vztahů mezi jednotlivými pojmy.

Na Webu je však téměř nemožné prosazení jednotného jazyka a vymezení jednotné slovní zásoby, což je fakt plynoucí jednak z principu decentralizace samotného Webu, ale také z povahy zpřístupňovaných informací – z globálního hlediska se v podstatě jedná o všechny myslitelné oblasti znalostí, v nichž je obecně problém najít byť jen částečnou jednotu či shodu.

O to se však Sémantický web nesnaží. Jeho myšlenka je založena především na flexibilním a otevřeném datovém modelu s odpovídajícími datovými jazyky, tak aby vyhovoval této nekonečné varietě Webu.

3.1 Technologie Sémantického webu

Sémantický web je ve své podstatě faktická kompozice několika technologií, mezi kterými je zaveden určitý hierarchický vztah jednotlivých vrstev:

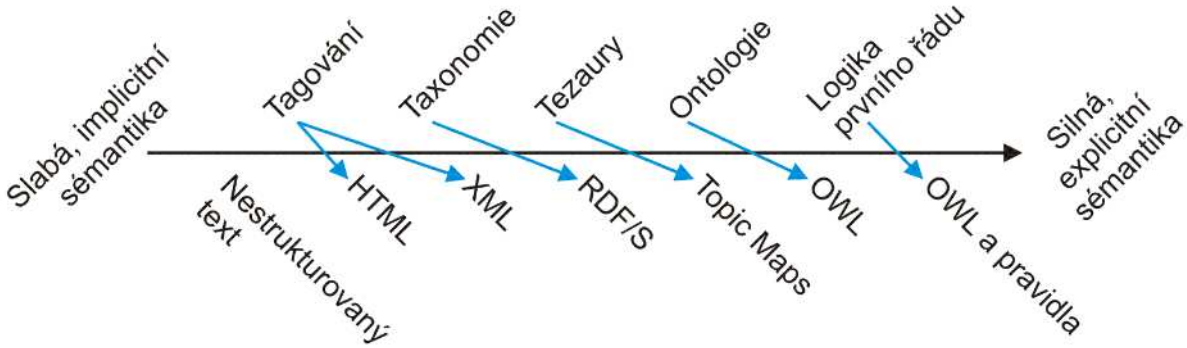


Obr. 3.1: Schématická struktura vrstev Sémantického webu

Celý koncept Sémantického webu je (podobně jako u Webu stávajícího) postaven na veřejném identifikátoru URI, který pomoci řetězce znaků dokáže identifikovat zdroj informace. Technologickým základem pomyslné pyramidy jsou také značkovací postupy a technologie XML [16], které pomáhají vytvoření strukturovaného dokumentu s vlastními značkami (tagy). Následuje technologická vrstva RDF [11], která umožňuje definovat vztahy mezi objekty (zdroji). Následující vrstva, která umožňuje zachycování složitějších ontologických struktur, je realizována prostřednictvím jazyka OWL [59]. Logická vrstva nám dovoluje popsat vztahy mezi jednotlivými objekty komplexněji a díky aplikování použitelné deskripční logiky provádí odvozování implicitních informací za pomocí dotazování SPARQL [58] na základě určitých pravidel RIF [10]. Poslední vrstva Důvěry má pak zajistit spolehlivost a pravdivost jednotlivých informací s využitím digitálního podpisu či šifrování.

Sémantické informace vpletené do běžného webu umožňují počítači manipulovat s daty inteligentněji. Například slovo „škola“ vyskytující se na běžném Webu je pro počítač pouze řetězec pěti znaků. Na Sémantickém webu je však možné označit slovo „škola“ identifikátorem pojmu škola v mnohem širším popisu pojmů a jejich vzájemných vztahů, běžně označovaných jako ontologie. Počítač pak v ontologii například zjistí, že škola je vzdělávací zařízení, které přijímá studenty a že student je člověk, který má studentský průkaz.

Vyskytuje-li se pak ve vyhodnocovaném textu třeba informace *„doktorand studuje na této škole“*, je následně pro počítač snazší odvodit, že doktorand je student, který má také studentský průkaz, což však je *znalost* z úvodní formulace jasně nevyplývající. Nejdříve ale musí existovat ontologie, která takové vztahy popisuje, a text musí být anotovaný (doplněný o významové značky). Samotná tvorba ontologií povětšinou spočívá v manuálním vývoji. Automatické odvozování ontologií je stále předmětem aktivního výzkumu a obdobně je tomu i se sémantickým značkováním, které ve většině případů také často probíhá ručně, ale existuje již také několik automatických a poloautomatických variant, přičemž jedna z nich bude blíže představena v části 4.3 této disertační práce.



Obr. 3.2: Sémantické spektrum a jeho vyjádření v závislosti na užité technologii zápisu

Na obrázku 3.2 je vyobrazeno sémantické spektrum pokrývající rozsah výsledné sémantiky od slabé, implicitní až po silnou, explicitní vyjadřovací schopnost. Je zcela zřejmé, že prostý nijak nestrukturovaný text bude mít pro následné zpracování webovým vyhledávacím strojem mnohem menší vypovídající hodnotu než dokument, v kterém byl uplatněn alespoň jeden ze základních principů sémantické úpravy informací:

- **Tagování** – označování obsahu prostřednictvím přiřazování klíčových slov bez ohledu na formát. Tagy zajišťují *klasifikaci obsahu*, popisují webový dokument a umožňují následné zpětné dohledání informace.

- **Taxonomie** – vyjadřuje hierarchické uspořádání ve sjednoceném tvaru klasifikačního systému zařazení zpracovávané informace do určité skupiny, třídy, které jsou sdružovány nejčastěji ve formě *stromových struktur*, kdy pokud uživatel hledá vhodnou třídu, pak se postupuje nejprve od vyšších obecnějších skupin směrem dolů k detailnějším.

- **Tezaurus** – seznam slov, který ke každému z nich navíc může přiřadit jeho alternativní vysvětlení a význam pro synonyma, antonyma, případně další doplňující či naopak zobecňující termíny významu daného slova na základě definovaného *řízeného slovníku* deskriptorů.

- **Ontologie** – soubor tříd, jejich vzájemných vazeb a atributů z určité vyčleněné oblasti zájmu, poskytující možnost uchování a předávání znalostí ve formě otevřeného datového modelu grafové struktury.

- **Logika prvního řádu** – je rozšířením jednoduché logiky výrokové o kvantifikační výrazy jako každý, všichni, někteří či žádný, díky čemuž je možno rozlišit v každé větě individuum, o němž se něco predikuje (odtud také analogické pojmenování *predikátová logika*) a následně také porovnává ekvivalentnost dvou výrazů.

Sémantický web sám jeho zakladatel Tim Berners-Lee propaguje již léta, přesto se však dosud nedočkal významnějšího rozmachu, pravděpodobně protože je oproti stávajícímu Webu pro běžného uživatele příliš komplikovaný. Postupem času proto vznikají jednodušší způsoby, jak do stávajícího či nově vytvářeného obsahu sémantickou informaci přidat. Jedná se zejména o již v současnosti využívané *RDF*, *Mikrodata a Mikroformáty*, které představují snadnější podporu pro následné strojové zpracování takto sémanticky označených dokumentů nebo dnes již běžnou sumarizaci obsahu pomocí *RSS* nebo ukládání a klasifikace pomocí popisných ontologií (např. FOAF) [8].

Slibné oblasti dalšího rozvoje Sémantického webu včetně možných překážek vývoje celého odvětví jsou blíže diskutovány v závěrečném shrnutí této práce.

3.1.1 URI indentifikátory

„Jednotným identifikátorem zdroje“ (URI) rozumíme kompaktní řetězec znaků, používaný k jednoznačné identifikaci nebo pojmenování zdroje za účelem umožnění interpretace takového zdroje, typicky přes síť World Wide Web v prostředí Internetu, za použití specifických protokolů [60].

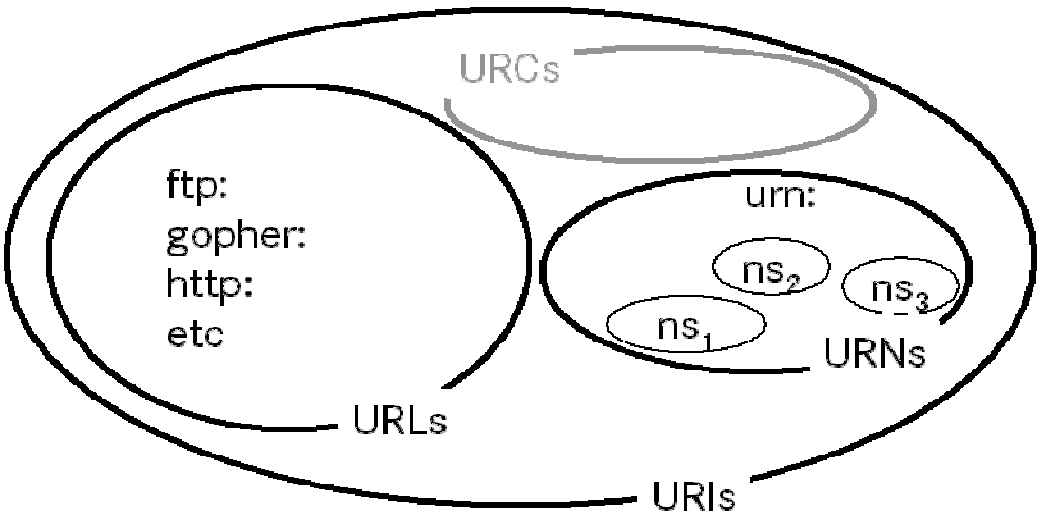
Při obecném popisu zdrojových dat se využívá formátu URI, který je však ve svém základu velmi volný a jde v podstatě pouze o syntaxi definující povolená schémata, hierarchickou část, dotaz a fragment zdroje, tak jak bylo teoreticky popsáno v úvodním přehledu zhodnocení stávajícího Webu (podkapitola 1.1).

URI – umožňuje popis zdroje informací jednak z pohledu jeho vlastní identity, (přičemž nemusí nutně určovat, kde je možné daný zdroj nalézt) nebo také z hlediska toho, jak je možné zdroj nalézt (ale nepopisovat jeho identitu) případně kombinací obou výše uvedených postupů současně, tedy zajistit přesnou identitu zdroje i způsob (protokol), jakým je možné jej dosáhnout.

URL – *„jednotný identifikátor adresy“* je synonymem pro internetové adresy a obecně nejznámějším indentifikátorem, který definuje způsob, jak se ke zdroji dostat, tedy jeho přesné umístění. V HTML specifikaci se URL používá jak pro vlastní zacílení odkazu, tak i pro načítání podpůrného obsahu (např. skriptů).

URN – *„jednotný identifikátor jména“* slouží jako lokačně-nezávislá možnost trvalého určení zdroje bez ohledu na doménu či server, na kterém je uložen.

URC – *„jednotná charakteristika zdroje“*, která slouží pro zařazení meta infomací, které jsou základem funkční koncepce Sémantického webu.



Obr. 3.3: Eulerův diagram vztahů mezi URI identifikátory

3.1.2 XML syntaxe

XML *(eXtensible Markup Language)* ve své podstatě není analogií programovacího jazyka, ale jedná se spíše fakticky o sadu pravidel pro tvorbu syntakticky bohatých značkovacích jazyků. Následně pak každý jazyk, získaný na základě takovýchto pravidel se nazývá XML aplikace [16].

Role XML v oblasti Sémantického webu je zásadní zejména z těchto důvodů:

- XML vytváří dokumenty a data, která jsou nezávislá na dané aplikaci
- XML má vlastní syntaktický standard pro meta-data
- XML má standardizovanou strukturu jak pro dokumenty, tak i data a
- je v současnosti poměrně široce známou a využívanou technologií

Díky tomu, že XML data nejen popisuje, ale rovnou jim prostřednictvím definovaných elementů přiřazuje vlastní význam, lze docílit také přehlednějšího zápisu zdrojového kódu ve srovnání s HTML, např. zápis produktu z ukázkového internetového obchodu s barvami by vypadal v HTML takto:

```
<b>Barva:</b> modrá <i>balení 5L, Odstín 1111</i> - cena 999 Kč
```

zatímco v XML „tagovém“ zápisu jednotlivých elementů takto:

```
<produkt>
    <barva>modrá</barva>
    <baleni>5L</baleni>
    <odstin>1111</odstin>
    <cena_mena="CZK">999</cena>
</produkt>
```

Aby však XML syntaxe zdárně fungovala a takto označené informační zdroje mohly být dále nezávisle na sobě sdíleny a zpracovány mnoha různými aplikacemi, za předpokladu, že daný dokument je strukturovaný (povinně) anebo validní (volitelně), je třeba při jejím zápisu dodržet alespoň několik základních pravidel, vycházejících ze specifikace (X)HTML [49]:

- každý element je párový (možný zápis: <barva></barva> nebo </barva>)
- elementy se nesmí křížit (toto je chybný zápis: barva)
- obsah dokumentu musí být uzavřen v jednom kořenovém elementu
- hodnoty atributů musí být v uvozovkách, a to i v případě pokud jde o číslo
- jména elementů mohou obsahovat písmena, čísla, tečky, podtržítka a pomlčky (ale nesmí začínat tečkou, pomlčkou ani podtržítkem)

3.1.3 RDF a vyjádření metadat

V oficiální specifikaci konsorcia W3C [37] je RDF zaveden jako standardizovaný datový model/rámec umožňující vyjádření popisných informací o webových zdrojích, přičemž tento zdroj může být i reprezentací informace, která nemusí být v prostředí Webu přímo dostupná, ale lze ji identifikovat na základě obecně-funkčního principu URI identifikátoru popsat jak z hlediska prostých vlastností, tak i jejích hodnot. Jinými slovy aplikací RDF jsme schopni odvozovat obsah sdělení z různých větných konstrukcí jednoduchých tvrzení a tyto následně znázornit jako grafické uzly a orientované hrany, kde právě jedna hrana společně s koncovým a počátečním uzlem tvoří jedno dané tvrzení.

Tvrzením zde rozumíme takovou základní jednotku informace, která je ztvárněna ve specifickém a snadno zpracovatelném formátu, který se dále může sdružovat do formy strukturovaných informací, tzv. *kolekcí*.

Informace jako taková může být také definována pomocí přirozeného jazyka, jehož zpracování není mnohdy strojově zcela jednoduše proveditelné. Jestliže však informace postupně převádíme do množin méně komplexně definovaných tvrzení, pak lze na základě tohoto přístupu zachytit i poměrně složité kolekce, které mohou být navíc také dobře strojově zpracovatelné. Tuto vlastnost strojové zpracovatelnosti RDF zajišťuje pomocí definovaných struktur pro jednotlivá tvrzení s využitím tzv. *tripletů* (trojic), sestávajících ze tří komponent:

- **Subjekt** – v mluvnickém pojetí *podmět* věty, značící její podstatu, v logice značí takový termín, o kterém někdo něco tvrdí a v RDF je subjekt reprezentací zdroje, který je popisován predikátem a objektem a nabývá hodnot z oborů *RDF URI reference* nebo *blank node* [41].
- **Predikát** – *přísudek* ve větě ukazuje na její část modifikující podmět a obsahující slovesnou frázi. V logice plní predikát funkci přiřazení reálné hodnoty konkrétnímu typu subjektu a její arita je určena počtem argumentů. V případě RDF jde pak o vazbu mezi subjektem a objektem, která nabývá hodnoty z oboru *RDF URI reference* [16]
- **Objekt** – v mluvnici *podstatné jméno*, vyvozené podle přísudku, v logice je objekt závislý na predikátu a v RDF může být buď zdrojem, na který odkazuje predikát nebo jeho doslovná hodnota [16] a nabývat může hodnoty z oborů *RDF URI reference*, *literal* nebo *blank node* [41].

RDF sám o sobě je jen abstraktním formátem systému popisu zdrojů, jenž pouze udává způsob zápisu informace o zdroji, ovšem nemá přitom žádnou předem definovanou syntaxi. Z tohoto důvodu není obsah RDF v dokumentu na první pohled v klasickém zobrazení (bez úprav pomocí XML) rozpoznatelný.

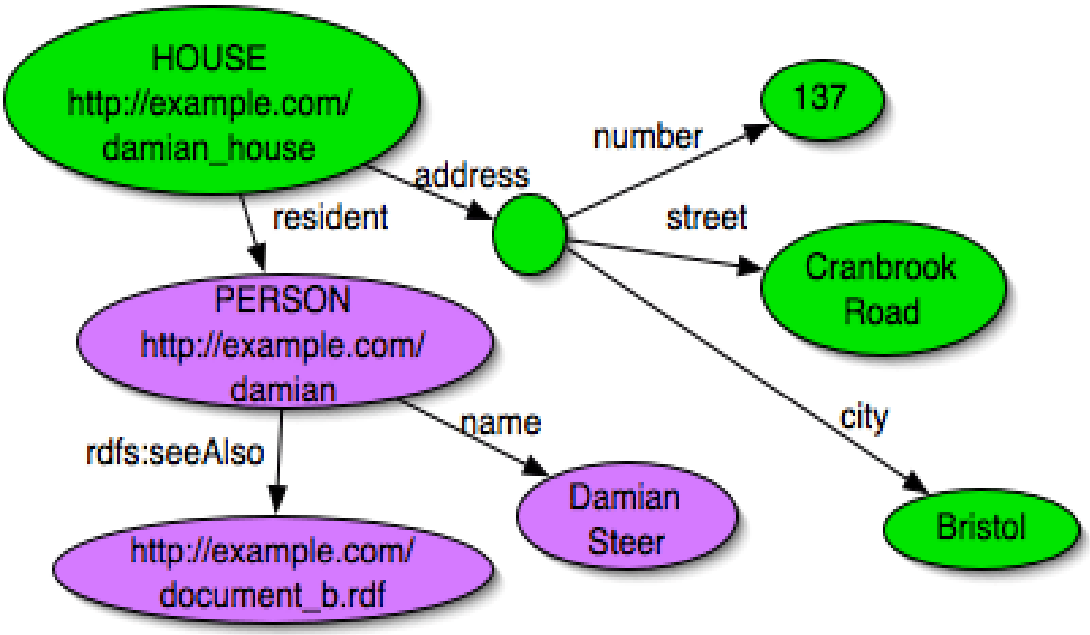
Následující příklad může pomoci lépe pochopit řetězec, vznikající při vytváření RDF trojice – mějme dvě prohlášení:

1) Autorem disertace je doktorand
2) Doktorand je autorem disertace

Zde je možno vysledovat zásadní rozdíl mezi vnímáním člověka (který má schopnost vyvodit z obsahu těchto dvou tvrzení tu skutečnost, že pojednávají o jednom a tomtéž faktu) a stroje, jenž však tato sdělení vnímá jako dva zcela rozdílné řetězce a pokud oba tyto výroky rozložíme, tak jak to dělá stroj (tedy do trojice subjekt-predikát-objekt), dostáváme následující zápis:

1) "disertace" = podmět, "má Autora" = vlastnost, "doktoranda" = předmět
2) "Doktorand" = podmět, "je autorem" = vlastnost, "disertace" = předmět

Seskupení více množin takových RDF trojic pak nazýváme RDF graf. Množinou uzlů pak rozumíme množinu subjektů a objektů v grafu. Uzel může být reprezentován URI s volitelným identifikátorem fragmentu dokumentu (URI reference, literál nebo prázdný) viz obrázek 3.4. URI reference a literál použitý coby uzel značí specifický pojem, který daný uzel představuje. URI reference použitá jako predikát identifikuje vztah mezi pojmy, který je představován uzly, které spojuje. URI reference predikátu pak také může být uzlem v RDF grafu. Prázdným uzlem je pak označován takový uzel, který neodpovídá ani URI referenci ani literálu, ale je pouze unikátním uzlem, který nemá vlastní jméno, ale je možné jej použít současně v jednom nebo více RDF grafech.



Obr. 3.4 RDF Graf s reprezentací množiny uzlů

RDF také současně uvozuje specifikaci datových typů pro určení hodnoty literálu sestávající z tzv. lexikálního prostoru, prostoru hodnot a mapování tohoto lexikálního prostoru do prostoru hodnot. Lexikální prostor představuje množina unicode řetězců, mapováním lexikálního prostoru do prostoru hodnot pak rozumíme takovou množinu dvojic, kde prvním elementem dvojice je ten z lexikálního prostoru a druhým pak ten z prostoru hodnot [44].

Je možno párovat každý prvek z prostoru hodnot s libovolným počet prvků prostoru lexikálního. Jedna nebo více referencí přitom identifikuje datový typ, který v případě RDF je předdefinován pouze jedinný a slouží pro vkládání dat do RDF prostřednictvím formátu XML se zápisem ve tvaru *rdf:XmlLiteral*. Pro ostatní běžné datové typy jako jsou celá čísla, časové razítko či datumy je využito předefinovaných datových typů převzatých z XML Schema, kde je navíc možnost nadefinovat si i datové typy vlastní a specifické, neboť samotné RDF touto možností vybaveno není [41].

V případě vyjádření dat pomocí datového typu *Literal* v RDF platí, že jakákoliv takto reprezentovaná informace (klasicky řetězce, čísla, data apod.) mohou být reprezentovány také ve formě URI. Avšak častější a intuitivnější je právě použití literálu, neboť tento může být pouze objektem RDF tvrzení, ale nikoli již subjektem či predikátem. Ve specifikaci RDFS se rozlišují *prosté (plain)* a *typované (typed)* literály. Prostý literál může být jako řetězec volitelně kombinován s jazykovým tagen *(xml:lang)* a používá se pro vyjádření prostého textu přirozeného jazyka. Kombinací řetězce s datovým typem a URI je pak typovaný literál, spadající do prostoru hodnot, jenž tento URI identifikuje [37].

3.1.4 RDFS pro popis slovníků

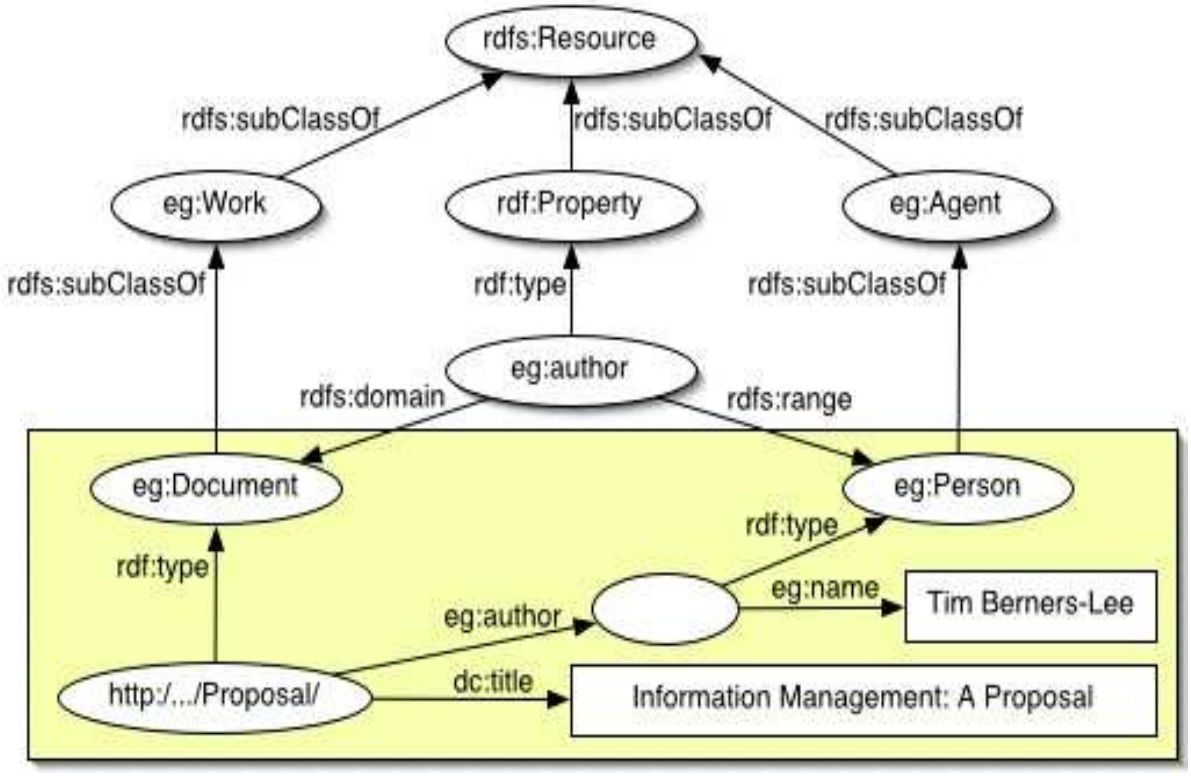
RDF-Schema, zkráceně RDFS, je sémantickým rozšířením stávajícího modelu RDF blíže popisovaného v předchozí podkapitole. Ve svém funkčním principu vychází z toho, že vlastnosti RDF sice lze chápat jako vyjádření atributů jednotlivých zdrojů, kdy je možno tyto odpovídajícím způsobem zápsat jako dvojice „atribut-hodnota“ s přidanou další vazbou mezi těmito dvěma zdroji, které jsou zapsány ve formě uspořádaných trojic do RDF grafů.

Problém nastává ve chvíli, kdy se snažíme popsat vlastnosti takového zdroje a jeho možné vztahy mezi dalšími zdroji. Pro tento účel jednoduchý model RDF již nestačí a je nutné zavedení jednotného jazyka pro popis schematického slovníku, kdy pomocí RDFS nejprve definujeme třídy a vlastnosti, jež se dají následně použít také pro popis jiných tříd a vlastností u jiných zdrojů. Systém popisování tříd a vlastností je obdobný jako u objektově orientovaných programovacích jazyků, ale na rozdíl od nich jsou zde vlastnosti popisovány pomocí výčtu tříd, jejimiž členy mohou samy být.

Vlastnost v praktickém použití RDFS je definována jako doména, tedy obor tříd *(rdfs:Domain)* s rozsahem hodnot určitého datového typu *(rdfs:Range).* Tento koncepční přístup poskytuje možnost dodatečného definování nové vlastnosti bez nutnosti přidání celé nové třídy.

V rámci pojetí RDFS lze každou třídu považovat současně za zdroj informací, který může být popsán vlastnostmi a má i svůj vlastní URI indentifikátor. Pokud je daný zdroj instancí určité třídy zapisuje se toto tvrzení jako RDF vlastnost *(rdf:type)* indentifikaci a jednotlivé členy pak lze označit jako instance třídy. RDFS však dále rozlišuje mezi třídou a množinou jejich instancí, kdy dvě třídy mohou mít současně sice stejné množiny instancí, ale mohou to být dvě naprosto rozdílné třídy – tedy třída může být členem množiny svých instancí a současně také může být instancí sebe sama.

Taková skupina zdrojů se v RDFS definuje jako *(rdf:Class).* Platí přitom, že jestliže třída C je podtřídou třídy C', pak také všechny instance třídy C budou rovněž instancemi třídy C'. Další vlastností *(rdfs:subClassOf)* pak lze konstatovat, že jedna třída je podtřídou druhé.



Obr. 3.5: Vzorové RDF Schema s definicí vlastností tříd a podtříd

Celou problematiku blíže osvětluje ukázka schématického znázornění tříd a potříd (obrázek 3.5) včetně následného uvedení výčtu dalších doplňkových v praxi nejčastěji používaných tříd, včetně stručného popisu v přehledové Tabulce 1, která dle standardu RDF Schema specifikuje tyto třídy [11]:

Tabulka 1 – Definice tříd v rámci standardu RDF Schema

Třída	**Definice**
rdfs:Resource	všechny elementy v RDF představují zdroje, které jsou instancemi třídy *rdfs:Resource* a současně pak také platí, že všechny ostatní třídy jsou podtřídou této třídy
rdfs:Class	třída zdrojů, které jsou součástí *rdfs:Class*
rdfs:Literal	třída literálových hodnot (prostých nebo typovaných), kde typovaný literál je také současně instancí uvozující třídy *rdfs:Datatype*
rdfs:Datatype	třída datových typů, kdy každá instance této třídy je současně podřídou třídy *rdfs:Literal*
rdf:XMLLiteral	třída hodnot, které jsou XML Literál *rdf:XMLLiteral* je instancí třídy *rdfs:Datatype* a současně implikující podtřídou třídy *rdfs:Literal*
rdf:HTML	je třídou HTML literálových hodnot, kde *rdf:HTML* je instancí *rdfs:Datatype* a podtřídou *rdfs:Literal*
rdf:Property	představuje třídu RDF, zachycující vlastnosti zdroje
rdfs:seeAlso	je instancí třídy *rdf:Property*, která označuje, že zdroj může obsahovat další doplňující informace
rdfs:isDefinedBy	je instancí třídy *rdf:Property*, která je definuje subjekt zdroje RDF slovníku, který tento zdroj popisuje
rdfs:member	je instancí třídy *rdf:Property* současně zahrnující všechny třídy spadající do dané oblasti či domény
rdfs:subPropertyOf	určuje fakt, že jedna vlastnost je podřazenou vlastností jiné vlastnosti, což následně umožnuje postupné vytváření hierarchie vlastností
rdfs:range	prohlašuje, že hodnoty vlastnosti jsou instance jedné nebo více tříd, jejiž rozsah je tímto definován
rdfs:domain	jejímž úcelem je indikovat, že jakýkoliv zdroj, který má danou vlastnost, je instance jedné ci více tríd (která je specifikovaná jako hodnota této vlastnosti)
rdf:type	používaná pro konstatování, že daný zdroj je instancí trídy uvededné jako atribut této vlastnosti
rdfs:subClassOf	uvádí, že všechny instance aktuální trídy jsou zároven instancemi trídy, která je uvedena hodnota vlastnosti
rdfs:subPropertyOf	používá se pro označení všech zdrojů, které jsou svázány vlastností uvedenou jako hodnota vlastnosti a které jsou zároveň svázány aktuální vlastností
rdfs:label	poskytuje název zdroje dobře čitelný pro člověka
rdfs:comment	poskytuje popis zdroje dobře čitelný pro člověka

3.1.5 Mikroformáty a Mikrodata

Mikroformáty, jakožto i otevřené datově formalizované zápisy ve formě Mikrodat, představují jednoduchou možnost přidání sémantické informace do stávajícího obsahu existujících (X)HTML značek, což umožuje učinit původně významově nepopsaný dokument strojově čitelným a zpracovatelným informačním zdrojem, který je schopen plnit ideu Sémantického webu, a to bez nutnosti složitějších zásahů do funkčnosti celé stránky i možného rizika, že by přidání těchto popisných metadat ohrozilo validitu celého dokumentu.

Mikroformáty (ve zkratce uváděné někdy jako: µF)

Samy o sobě nejsou označením žádného konkrétního pojmu či názvosloví. Jde spíše o označení přístupu, jakým se jednotlivé atributy mikroformátů používají. Ono „mikro“ v názvu i zkráceném označení totiž označuje skutečnost, že vlastní formátování v takto označeném dokumentu nijak nezasáhne do celé jeho struktury, ale vždy se bude týkat pouze konkrétní takto označené informace. Technicky je taková sémantická informace uložena ve formě hodnoty určitého atributu a vyjádřena svým textovým významem.

Například adresa naší fakulty může být ve zdrojovém kódu popsána takto:

```
  <h3><strong>Fakulta aplikované informatiky</strong></h3>
  <p class="itAddrLine">Nad Stráněmi 4511, Zlín</p>
```

což sice může být z pohledu uživatele dostatečně čitelný zápis, ne tak však již pro stroje, které takto nerozeznají, jaká část zápisu odpovídá názvu či adrese. Konkrétně pro sémantický zápis adresy existuje samostatný mikroformát *hCard*, z jehož pomocí by se celý zápis mohl přepsat následovně:

```
  <div class="vcard">
  <div class="fn org">Fakulta aplikované informatiky</div>
  <div class="adr">
  <div class="street-address">Nad Stráněmi 4511</div>
  <span class="locality">Zlín</span>
  <span class="country-name">Česká republika</span></div>
</div>
```

Nespornou výhodou mikroformátů je tak poskytnutí rychlého a efektivního způsobu, jakým lze jednoduše do stávajících či nově vytvářených webových stránek zanést sémantickou informaci. Na druhé straně, vývoj aplikace, která na jedné stránce rozpozná vícero mikroformátů, může narazit na problém jednoznačného určení toho, o který mikroformát se vlastně jedná.

Počet možných mikroformátů totiž neustále narůstá a zevrubný přehled o tomto trendu je patrný i v Tabulce 2 a 3, kde je uveden jak výčet již v praxi zavedených, tak i na své širší prosazení stále čekající mikroformáty.

Tabulka 2 – Přehled standardizovaných mikroformátů

Mikroformát	Definice
hCalendar	slouží pro popis událostí a schůzek
hCard	obdoba vizitky popisující lidi, organizace, kontakt
rel-license	užívá se při vymezení popisu licencovaného obsahu
rel-nofollow	zamezení indexace stránky vyhledávacími roboty
rel-tag	označení příspěvků a stránek subjektem
XFN	spojení mezi profily osob na sociálních sítích
XMDP	určuje slovníkový profil Mikroformátů

Vzhledem k tomu, že specifikace mikroformátů nemá prozatím žádnou ustálenou podobu, nejedná se o řešení elegantní, za to však překvapivě funkční a s postupným rozšířením Mikroformátů za hranice akademické sféry vznikají postupně nové tagy pro specifický obsah, od vizitek až po kuchařské recepty.

Tabulka 3 – Nově zaváděné koncepty mikroformátů

Mikroformát	Definice
Adr	informace o místní adrese či lokalitě
Geo	zeměpisná šířka a délka dle geo souřadnic WGS84
hAtom	označení obsahu v příspěvcích s časovým razítkem
hListing	výpis produktů nebo služeb
hMedia	informace o multimediálních souborech
hNews	novinové články, rozšíření pro hAtom
hProduct	sémantické označení zboží, značky, výrobce
hRecipe	označení kuchařských receptů na vaření a pečení
hResume	individuální resumé a životopisy
hReview	individuální recenze a hodnocení produktů
hReview-aggregate	agregace dostupných recenzí a hodnocení
rel-author	propojení s home page stránkou autora
rel-home	odkaz na úvodní stranu webové prezentace
rel-payment	odkaz na platební bránu pro zaplacení zboží

Sémantiku zakódovanou do webového dokumentu pomocí mikroformátů je navíc možno následně extrahovat a získat tak validní RDF data, což je možné například s využitím technologie GRDDL, která bude popisována v části 3.1.7.

Mikrodata

Principiálně vycházejí z mikroformátů, avšak podstatně zjednodušují jejich syntaxi a odkazující primárně na již existující schéma a v současnosti se těší velké oblibě nejen v souvislosti se začleněním jejich koncepce do nově zavedeného standardu HTML 5.1. Velkou výhodou při podpoře zavádění Mikrodat do praxe je také fakt, že jejich podporu v rámci zavedeného již do své funkcionality zahrnuli také všichni zástupci předních světových vyhledávačů (Google, Bing, Yahoo, Yandex), kteří participují na společném projektu *Schema.org*, odkud je také následně globálně přebírán kompletní přehled dostupných schémat pro Mikrodata, vč. užitých tagů a jejich vlastností.

Pro značkování kódu mikrodaty jsou přitom dostačující tyto atributy:

- **Itemscope** – je označením nového objektu, kterému je poté následně přiřazen jeho popis a rozširující popisné schéma pomocí absolutní adresy

- **Itemtype** – specifikuje schéma, které se použije pro daný objekt. Možná je i tvorba vlastního schématu, ale obecně se doporučuje přebírat a odkazovat na ty, které jsou již hotové a platné

- **Itemprop** – označení vlastnosti pro detailní popis objektu, kterému náleží již konkrétní hodnota, která má být tímto atributem popsána

Pro srovnání s předchozím příkladem označení adresy instituce mikroformáty, je zde uveden obdobný příklad s vyžitím mikrodat a zavedených schémat.

```
<div class="info" itemscope
itemtype="http://schema.org/Organization">
     <span itemprop="name">Tomas Bata University in Zlin</span>
     <img itemprop="logo" src="/logo.jpg" />

     <span itemprop="streetAddress">Nad Stráněmi 4511</span>
     <span itemprop="location">Zlín</span>
     <span itemprop="postalCode">76001</span>
     <span itemprop="addressCountry">CZE</span>

     <span itemprop="description">Fakulta aplikované informatiky je
     dynamicky se rozvíjející fakulta se zaměřením na informatiku,
     bezpečnostní technologie a automatizaci. </span>

     <a href="http://fai.utb.cz/" itemprop="url">Odkaz na Web</a>
</div>
```

3.1.6 SPARQL dotazovací jazyk

SPARQL *(Protocol and RDF Query Language)* je schváleným standardem konsorcia W3C. Jak je již ze zkráceného akronymu zřejmé, jedná se o dotazovací jazyk pro RDF, respektive jeho reprezentací RDF grafu. Plní však zároveň i funkci protokolu umožňující dotazování skrze metody GET a POST.

V prostředí Sémantického webu plní roli obdobnou jako SQL v oblasti relačních databází, tedy pokud si představíme pomylsnou globální kolekci navzájem provázaných sémanticky označených dat, je možné se nad ní dotazovat obdobně jako právě u relačních databází, a to právě díky SPARQL, kterým definujeme syntaxi a sémantické prvky dotazu pro různé datové zdroje, které se tak mohou stát nativním či zprostředkovaným RDF úložištěm.

Umožněno je dotazovat povinné i volitelné vzory v grafech, stejně tak i jejich spojení a průniky. Obsahem dotazu je zpravidla množina trojic ve formě základní šablony grafu (basic graph pattern), přičemž v této šabloně RDF trojic může být proměnná zastoupena na pozici všech elementů (subjekt, predikát, objekt). Nejčastějšími typy SPARQL dotazů jsou:

SELECT – proměnným přiřadí hodnoty a vrátí je v tabulce
ASK – vrací hodnotu boolean, testuje zda basic graph pattern má řešení
DESCRIBE – vrací podgraf vyhovující basic graph patternu
CONSTRUCT – obdoba SELECTu, jen místo hodnot v tabulce vrací graf

Shoda v základní šabloně grafu a podgrafu RDF nastává v momentě, kdy termy podgrafu mohou být substituovány za proměnné a výsledkem pak mohou být množiny (při použití klauzule *SELECT*) nebo RDF grafy (s využitím klauzule *CONSTRUCT*).

Pro RDF existuje řada serializací (RDF/XML, N3, N-Triples, Turtle, Trix aj.), z nichž některé slouží čistě k přenosu a uchovávání RDF dat, zatímco jiné byly zavedeny spíše s ohledem na lepší čitelnost následného zápisu.

Následující jednoduchý příklad, s využitím jedné z popularizovaných serializací (RDF – Turtle), znázorňuje vzorová RDF data (tvořená URI a literály) s identifikátorem knihy, jejíž název (a tedy současně vlastnost s vlastním identifikátorem) je *„SPARQL Tutorial“* (literálový objekt). V dotazu, který následuje je poté deklarováno, které termy mají být vybrány *(?title)* a šablona RDF grafu. Výsledkem je pak množina dat (proměnná, hodnota) [52].

Data:

```
<http://example.org/book/book1>
<http://purl.org/dc/elements/1.1/title>
"SPARQL Tutorial" .
```

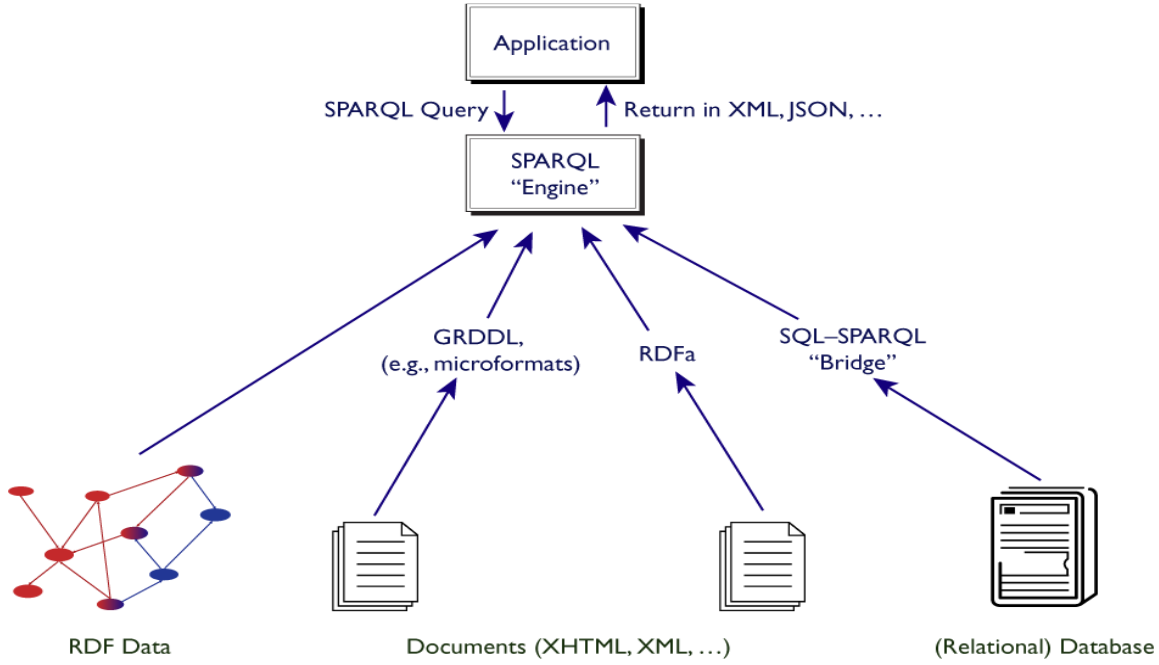
Dotaz:

```
SELECT ?title
WHERE
{
<http://example.org/book/book1>
<http://purl.org/dc/elements/1.1/title>
?title .
}
```

Výsledek:

title
"SPARQL tutorial"

Známé pozice v trojici jsou zadefinovány, ostatní jsou ponechány volně. Výsledkem pak bude ta z uspořádaných trojic, která určené podmínce vyhovuje. Jak ukazuje schéma na obrázku 3.6, v současné době jádro SPARQL enginu podporuje několik stěžejních formátů pro zpracování nejen sémanticky orientovaných dat. Nově je např. s využitím syntaktického přemostění možné dotazování i nad klasickými SQL relačními databázemi, což v praxi otvírá dveře mnohem širšímu využití jazyka SPARQL i v jeho formě coby RDF protokolu.



Obr. 3.1 Schéma aplikace využívající dotazovací engine SPARQL

SPARQL jako dotazovací protokol může být specifikován buďto jako abstraktní rozhraní nezávislé na konkrétní implementaci, případně jako tzv. *endpoint*, představující vazbu na rozhraní HTTP a SOAP webové služby, které jsou nezbytnou součástí pro zajištění celkové interoperability Sémantického webu a budou podrobněji teoreticky popisovány v části 3.3 této práce.

3.1.7 GRDDL transformace

GRDDL *(Gleaning Resource Description from Dialects of Language)* by se volně dalo interpretovat jako sběr popisu zdrojů z jazykových dialektů u doménově specifických jazyků, kdy jde o techniku pro získávání RDF dat z XML a XHTML dokumentů (včetně mikroformátů) na základě jejích dialektu, k čemuž se běžně používá podpora v podobě XSLT 1.0 transformací, ačkoli povolené jsou i jiné metody (jazyk C nebo Xquery) [28].

Transformace se k dokumentu připojují buďto přímou referencí nebo nepřímo přes profilové a namespace dokumenty. Autor obsahu si tak sám může určit transformaci, jejímž použitím lze takto vytvořená data získat ve formě RDF. U korektně strukturovaných XML dokumentů se GRDDL vkládá jako deklarace jmenného prostoru v kořenovém uzlu a současně definicí hodnoty atributu *grddl:transformation*, jehož hodnotou může být seznam URI identifikátorů, které lze použít k lokalizaci jednak transformace samotné nebo rovněž také aplikace, která transformaci provádí [29].

```
<html xmlns="http://www.w3.org/1999/xhtml"
xmlns:grddl="http://www.w3.org/2003/g/data-view#"
grddl:transformation="glean_title.xsl
http://www.w3.org/2001/sw/grddl-wg/td/getAuthor.xsl">
<head>
[...]
</html>
```

Je-li dokumentem validní XHTML, pak se GRDDL transformace připojuje nejprve uvedeným profile atributem v elementu *head* a posléze odkazem přes URI na transformační skript, jenž je definován jako hodnota atributu *href* v *link* elementu s nastaveným atributem *rel* na hodnotu *transformation*.

```
<html xmlns="http://www.w3.org/1999/xhtml">
<head profile="http://www.w3.org/2003/g/data-view">
<title>Popsaný zdroj</title>
<link rel="transformation"
href="http://www.w3.org/2000/06/dc-extract/dc-extract.xsl" />
<meta name="DC.Subject"
content="ADAM; Simple Search; Index+; prototype" />
...
</head>
...
</html>
```

3.2 Ontologie a znalostní inženýrství

Ontologií se podle dostupných zdrojů [1][32] rozumí ujednané pochopení dané domény, axiomatizované a formálně reprezentovatelné jako logická teorie ve formě zdrojů. Na základě sdílení ontologií pak mohou aplikace mezi sebou vzájemně komunikovat a vyměňovat data obohacená o jejich významovou složku a tím tak následně zajistit interoperabilní spolupráci několika různých systémů nezávisle na jejich interních ontologiích.

Ontologie jsou aktuálním předmětem výzkumu hned několika vědních oblastí současně: filozofie, lingvistiky, logiky a informatiky. V oblasti informatiky lze pak výzkum ontologií rozdělit do dvou hlavních subdomén: umělá inteligence (budování sdílených bází znalostí) a databáze (budování konceptuálních datových schémat) [32]. Hlavní a fundamentální výhodou ontologií oproti klasickým znalostním bázím a konceptuálním datovým schématům, je vysoká nezávislost na dané aplikaci.

Samotný pojem *ontologie*, vycházející z původně latinského slova *ontologia,* může podle slovníku [43] představovat dva možné významy:

1. Odvětví metafyziky, zaměřené na povahu a vztahy bytí.
2. Konkrétní teorie o podstatě bytí nebo formách existence.

Z těchto definic je patrné, že původní význam slova je zakotven ve filozofii. Nejčastěji uváděné definice vhodně adaptované na vědní oblast informačních technologií pak jsou:

- Ontologie definující běžná slova a pojmy, které jsou následně použity k popsání a znázornění obecné nebo konkrétní znalostní oblasti
- Ontologie jako produkt inženýrství, skládající se z konkrétní slovní zásoby použité pro popis reality (nebo její části) a tvrzení ohledně významu této slovní zásoby – jinými slovy specifikace konceptualizace.

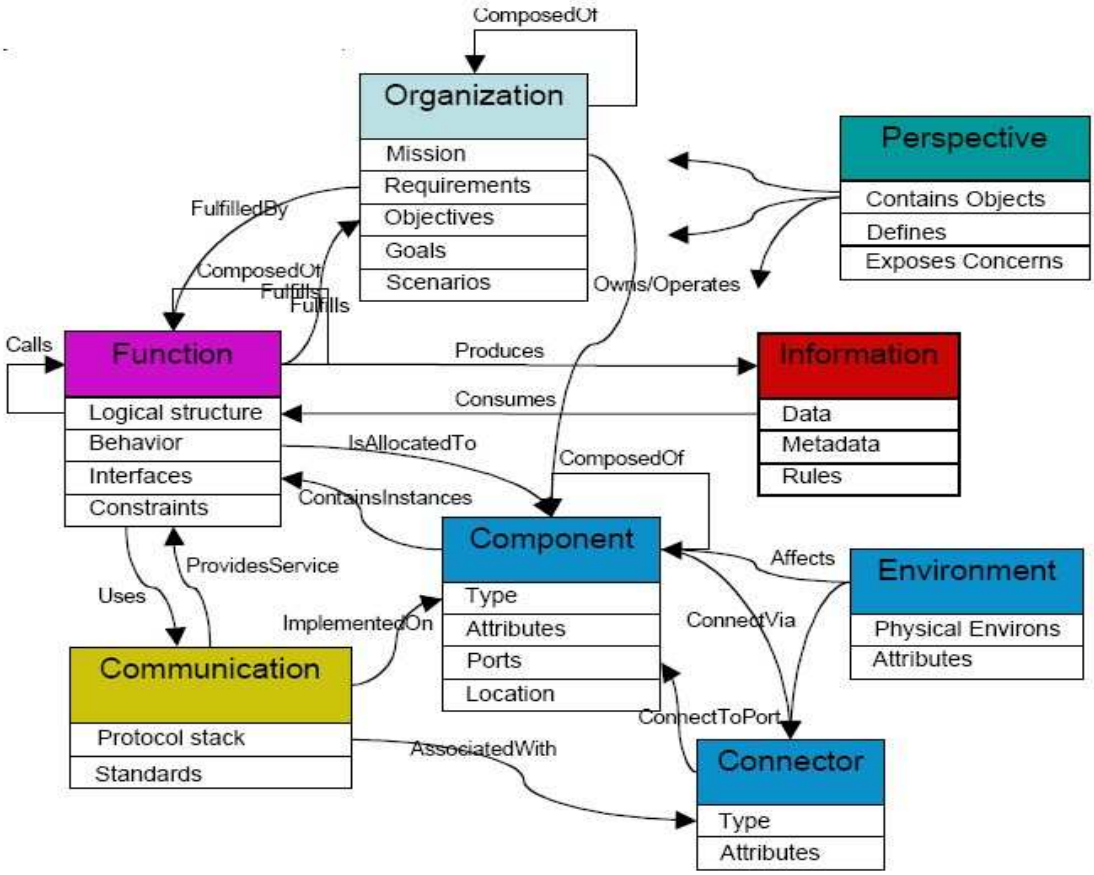
Konceptualizací rozumíme *způsob smýšlení o světě* [16]. Aby bylo možné strojové zpracování dat včetně jejich sémantiky, bylo nutné definovat formální jazyk pro popis ontologií – OWL, který má větší vyjadřovací schopnost než RDF a používá se v případě potřeby explicitního vyjádření významu termu nebo vztahu mezi dvěma termy se zajištěním popisu vlastností disjunkce, kardinality, rovnosti, bohatšího typování vlastností (např. symetrie) a výčtových tříd.

Ontologické inženýrství by se tak na rozdíl od databázového / softwarového inženýrství mělo snažit o zachycení reality „tak, jak je“, relativně nezávisle na požadavcích databází a konkrétních softwarových systémů.

To zahrnuje také soubor aktivit, týkající se procesu vývoje životního cyklu a metodické tvorby ontologií, který lze zobecnit následující posloupností kroků:

- ujasnění účelu a rozsahu ontologie
- specifikace terminologické části ontologie
- odlišení ontologických typů
- specifikace struktury taxonomie
- vytvoření netaxonomických relací, atributů a instancí
- nasazení a údržba ontologie

V praxi některé důkladnější metodiky, jako METHONTOLOGY [20], předpokládají vznik samotné ontologie nejprve v neomezené podobě cílového logického formalismu, kdy se jednotlivé konstrukty a vztahy mezi nimi zapisují do tabulek a logické formule formou pseudojazyka, nad kterým se ještě nepředpokládá odvozování. První verze ontologie tak může existovat jen ve formě schematického textového dokumentu (viz obrázek 3.2), ale existují i některé nástroje WebODE [34], které umožňují následnou konverzi takového schématu do strojově zpracovatelných formátů. Zcela automatická konverze by však vyžadovala zohlednění sofistikovaných logických vzorců pro cílový jazyk, což v současnosti ještě v žádném známém systému implementováno není.



Obr. 3.7 Příklad konstrukce doménové ontologie v organizaci

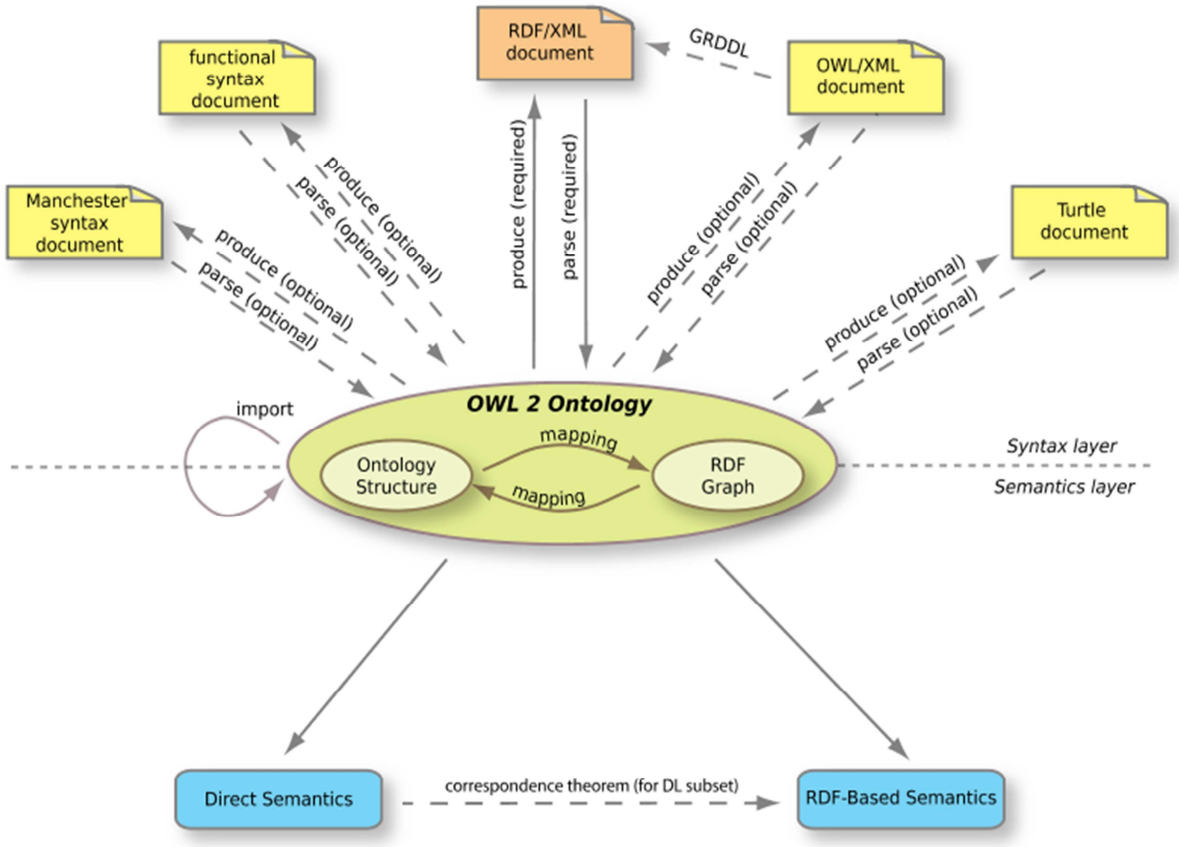
3.2.1 OWL jazyk pro zápis ontologií

Ontology Web Language je jazyk formálního popisu ontologií, založený na despkripční, někdy též terminologické logice, která vychází z predikátové logiky 1. řádu, ale je syntakticky omezenější. Role deskripční logiky je formálním základem při odvozování nových pojmů nad ontologiemi a OWL tak zde plní funkci nádstavbového rozšíření RDFS, ze kterého vychází a jehož konstrukce (*subClassOf, subPropertyOf, range, domain aj.*) také využívá [42].

Nejznatelnějším rozšířením je zde přitom možnost použití nejen standardně pojmenované třídy, ale i anonymní třídy, jež může být definována logickým výrazem. Lze tak snadno provádět operace konjunkce i disjunkce tříd, třídy do sebe navzájem vnořovat a sestavovat tak postupně z jednodušších pojmů složitější. Vlastnosti tříd zde rozdělujeme explicitně na *objektové*, kde hodnota představuje instanci nějaké třídy a *datové*, jejíž hodnota je vyjádřena literálem.

Takto definovaným vlastnostem je pak následně možno přiřadit ještě charakteristiky funkčnosti, symetrie, vzájemné inverznosti apod.

Schéma na obrázku 3.8 zachycuje možnou strukturu sestavení ontologie na základě akuálně schváleného standardu OWL 2. Znázorněno je zde přitom několik druhů syntaxí, které OWL ke svému zápisu může využívat.



Obr. 3.3: Ontologická struktura možného syntaktického vyjádření ve standardu OWL 2

Samotný programový zápis v podobě RDF trojic (jenž jsou vyjádřeny např. v XML či Turtle syntaxi) je sice v praxi málo přehledný, ale umožňuje následné bezproblémové zpracování stejnými nástroji jako RDF. Pro lidsky čitelnější zápis můžeme užít tzv. *Manchesterskou syntaxi*, která však nepokrývá celý jazyk, ale jen jeho základní části, což však pro běžné případy povětšinou dostačuje. Další možností zapisu ontologií v OWL je využití normativní funkční syntaxe, případně lze syntaxi OWL vyjádřit i přímo prostřednictvím XML [4].

Vytvářením tříd, definováním vztahů mezi nimi a popisem vlastností prvků těchto tříd je možné modelovat určitý systém znalostí o vybrané zájmové oblasti. Samotnému zápisu ontologie v jazyce OWL předchází vymezení použitých slovníků (jmenných prostorů), které jednoznačně popisují značky používané v celém takto popisovaném dokumentu, přičemž deklarace jmenných prostorů je uzavřená v rámci tagu <*rdf:RDF*>, např.:

```
<rdf:RDF  xmlns:nc = "http://www.neco.cz/#"
 xmlns:osoby = "http://www.neco.cz/osoby#"
 xmlns:owl ="http://www.w3.org/2002/07/owl#"
 xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
 xmlns:xsd ="http://www.w3.org/2001/XMLSchema#">
```

Hlavička definované ontologie se pak nachází za deklarací jmenného prostoru:

```
<owl:Ontology rdf:about="">
 <rdfs:comment>Příklad OWL ontologie</rdfs:comment>
 <owl:priorVersion rdf:resource="http://www.neco.cz/ovoce"/>
 <owl:imports rdf:resource="http://www.neco.cz/jidlo"/>
 <rdfs:label>Ovocná ontologie</rdfs:label>
</owl:Ontology>
```

Atribut *rdf:about* označuje zdroj, který je v dokumentu popisovaný *rdf:label*. V případě, že je jeho hodnota prázdný řětězec ("") je zdroj představovaný aktuálním dokumentem (tj. dokument samotný popisuje danou ontologii). Atribut *owl:comment* umožňuje k dané ontologii zadefinovat vlastní poznámku. Prostřednictvím *owl:priorVersion* se předává informace o verzi ontologie, která předchází té aktuálně definované a pomocí atributu *owl:imports* můžeme do stávající struktury vkládat také jiné, již existující ontologie.

Avšak ani v jinak zásadně pokročilejším formátu OWL nedokážeme vždy dostatečně explicitně vyjádřit některá tvrzení, tudíž tento způsob zápisu nemusí být vždy zcela vhodným pro všechny aplikace, a to zejména vlivem omezenosti logiky prvního řádu, na které stojí. Řešením může být následné zavedení určité sady pravidel, která budou blíže popsána v následující části a podle kterých lze takto zachycené znalosti dále podrobněji modelovat.

3.2.2 RIF pravidla pro odvozování

V návaznosti na úvodní představu o modulárním uspořádání Sémantického webu (viz obrázek 3.1 v technologickém úvodu teoretické části této práce) lze vysledovat provázanost jednotlivých dosud zde uváděných technologií, kdy v zásadě celá koncepce Sémantického webu stojí na základech RDF, popisující data ve formě grafů, SPARQL pro dotazování a OWL, pro zachycení složitějších struktur a vazeb mezi nimi ve formě ontologií. Problém však nastane při pokusu o sdílení takto zachycených faktů mezi různorodými aplikacemi.

V takovém případě však pro jejich spojování a možnost vzájemného odvozování nových faktů musí existovat jednotný způsob sdílení předem definovaných pravidel mezi systémy – RIF *(Rules Interchange format)* [10].

Vize RIF je založena na kolekci *dialektů*, což jsou rozšiřitelné jazykové sady se strikně definovanou syntaxí a sémantikou. Za předpokladu splnění myšlenky vzájemné provázanosti Sémantického webu, by pak aplikace napsaná pro jednu sadu pravidel měla být kompatibilní s aplikací řídící se odlišnou sadou pravidel pro odvozování nových faktů, přičemž mezi nejrozšířenější z nich v současnosti patří SWRL *(Semantic Web Rule Language)*, RuleML *(Rule Markup Language)* a R2ML *(Reverse Rule Markup Language)*.

Právě i s ohledem na fakt, že těchto ustálených sad pravidel existuje početně markantnější množství, bude zcela jistě výhledově zapotřebí, aby se tyto časem buďto sjednotily pod unifikovaným formátem nebo byly navzájem kompatibilní v rámci specifické mezivrstvy, která komunikaci mezi nimi navzájem zajistí.

I proto se RIF spíše než na pokus o sestavení jednoho všeobecně platného standardu zaměřuje na zprostředkování univerzálně použitelného prostředku, pro výměnu informací, neboť je zřejmé, že v oblasti Sémantického webu skutečně neexistuje formát, který by současně pokrýval složitost explicitního vyjádření logicky obtížně odvoditelných paradigmat a současně byl natolik výkonově dostačující, aby vracel odpovídající výsledky vyhledávání v co nejkratším možném čase. Problematice výkonnostního srovnání inferenčních mechanismů se ostatně věnuje i praktická část 4.5 této disertační práce.

V rámci konsorcia W3C stále probíhají práce na vývoji několika základních dialektů pro použití v rámci RIF formátu. Mezi dva, které jsou v současné době nejblíže vlastní standardizaci se řadí BLD *(Basic Logic Dialect)*, založený na dobře známých a stabilních principech sémantiky, jež vychází z jazyka predikátové logiky prvního řádu a PRD *(Production Rule Dialect)* navržený pro produkční verze již zavedených či vyvíjených systémů jako Jess či Drools [34].

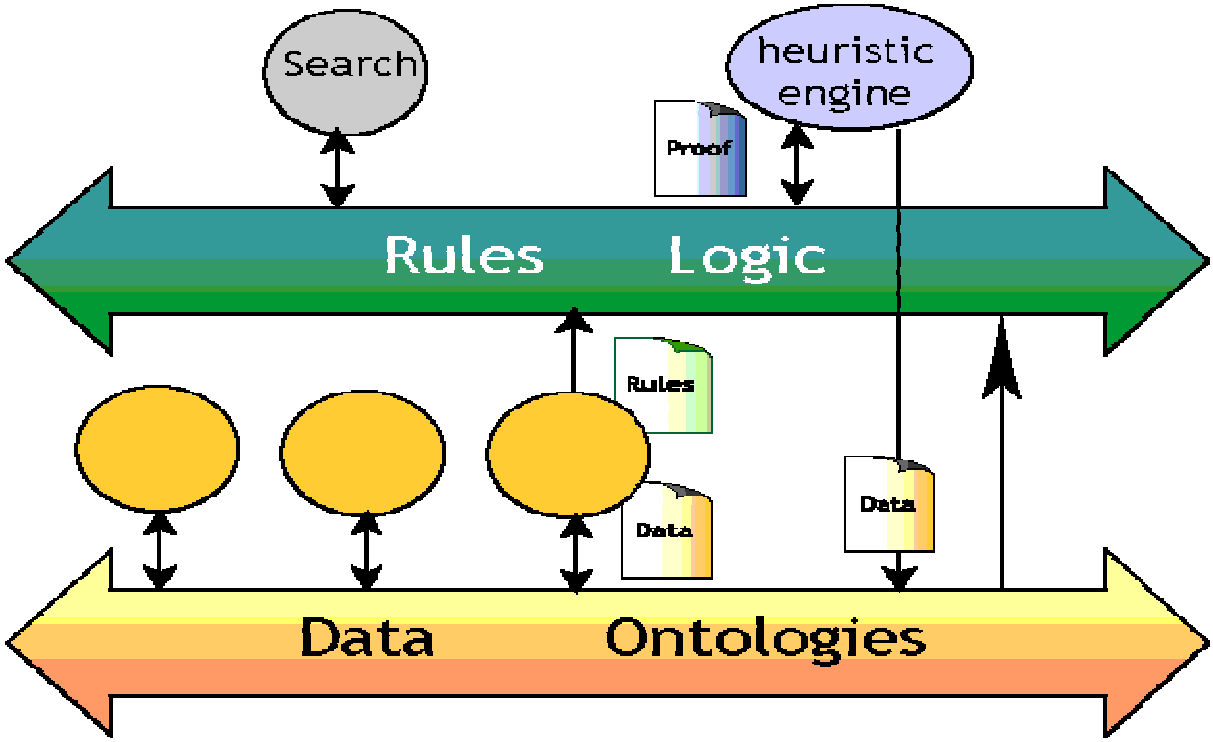
Odlišné sady pravidel vyžadují často také odlišné požadavky při vlastní implementaci, a to zejména dle toho, zda se jedná v základu o:

- **Dedukční pravidla** – představující výroky schopné vyvození nových poznatků z jiných znalostí pomocí logické inference. Popisují statické závislosti mezi entitami, které mohou být použity k dalšímu rozšíření znalosti vycházející z toho, co je předem definováno v znalostní databázi.

- **Normativní pravidla** – popisují omezení a zamezují nesrovnalostem v datech či logice aplikace spíše než aby samy byly zdrojem odvození nové znalosti, kdy například jeden uživatel v databázi smí mít v rámci svého účtu přiděleno pouze jedno identifikační číslo.

- **Reaktivní pravidla** – představují reakční chování a implementaci reaktivního systému, schopného automaticky vykonávat předem specifikované akce při určitých událostech či splněných podmínkách.

Všechna tato pravidla se mohou navzájem kombinovat a prolínat, stejně tak mohou být mnohdy v praxi součástí sofistikovaného heuristického enginu (obrázek 3.4), který se v rámci aplikace Sémantického webu přímo spolupodílí na procesu zpracování hledaného dotazu a vyvození nových faktů ze stávajících dat (RDF) a ontologií (OWL) s využitím logických konsekvencí na základě zpracování datových kolekcí z předem definovaných dialektů [36].



Obr. 3.9 Role RIF pravidel v logickém procesu vyhledávání výsledků

3.2.3 Extrakční ontologie

Extrakce informací z nestrukturovaných dat je disciplínou, jež se v rámci strojové lingvistiky řeší již několik desítek let, často se označuje také jako rozpoznání jmenných entit – NER *(Named Entity Recognition)*.

V rámci NER se kromě vlastního rozpoznání místa, kde se daná pojmenovaná entita nachází (formálně blíže probíráno v části 3.4.3), řeší také problematika *disambiguace*, tedy rozpoznání typu pojmenované entity (např. *je Paris město nebo osoba?*) a následné mapování typu na tuto konkrétní entitu (např. *pokud jde o město, tak o kterou z mnoha měst, pojmenovaných Paříž jde?* Při rychlém dotazu na Wikipedii lze dohledat, že jich je 26 jen v USA).

Abychom zamezili této nejednoznačnosti pojmů, rozlišitelných pouze v širším kontextu vznikaly postupně přístupy k NER založené na slovnících (gazetters), ručně vytvářených vzorech a pravidlech až po strojové učení, založené např. na neuronových sítích či skrytých Markovových modelech (HMM).

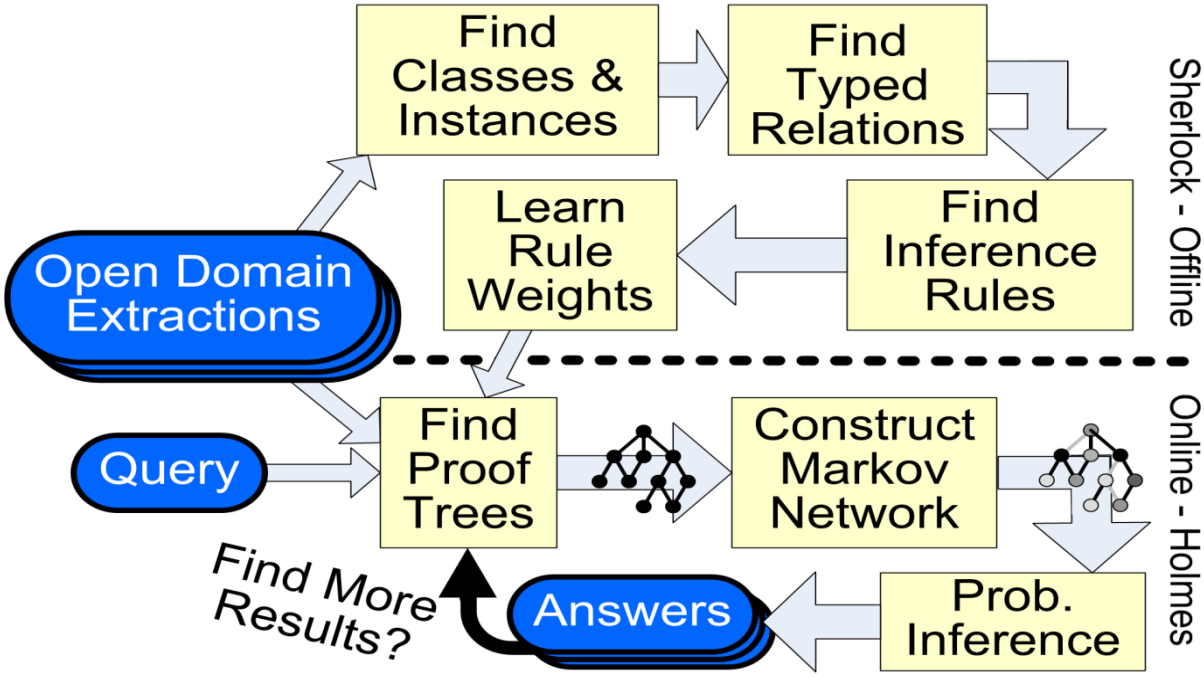
I v případě učení vzorů je ovšem vyžadována ručně anotovaná trénovací sada dat s příklady, které mohou být reprezentovány tzv. *wrappery* (kdy postačuje několik málo příkladů, v závislosti na struktuře stránky jsou však jejich možnosti omezené) nebo *pravidly* (viz předchozí část 3.3.3) či *statistickými daty*, které za cenu prvotně požadovaného většího objemu korpusu trénovacích dat, nakonec dosahují podstatně lepších výsledků.

Čistě ruční příprava takových dat je však nákladná a časově náročná, proto se v praxi při jejich tvorbě často používají iterativní procesy definované jako:

- *Statistický bootstrapping* – testuje vzory, na malém ručně anotovaném vzorku dat, kdy pokud se osvědčí, jsou následně použity pro sémantickou anotaci dalších dat, což však s sebou vždy nese riziko vzniku regresních chyb, zapříčiněných vlivem chybně interpretovaného vzorku.

- *Bootstrapping založený na redundanci informací* – nalezené vzory na různých zdrojích či v různé stuktuře jsou nejprve převedeny do formálního tvaru daného zdroje, podle nejž jsou poté následně extrahovány informace o dosud neznámých objektech.

Na principu bootstrappingu pracuje také softwarový anotační nástroj, navržený v praktické části této práce (kapitola 4.3), kde je taktéž provedena analýza anotačního procesu, který může být s využitím datových korpusů několika globálně dostupných ontologií také částečně automatizován.



Obr. 3.5: Extrakce informací v prostředí otevřených domén

Na výše uvedeném obrázku 3.10 je schématicky znázorněn proces extrakce informací, kdy na základě dotazu mezi otevřenými doménami hledáme odpovídající výsledky, vyvozené na základě tříd, jejich instancí a mapovacích pravidel, která jsou schopna na základě konstruktů a naučených testovacích dat vyvodit patřičné logické vztahy a vracet v odpovědi dané výsledky vyhledávání korespondující se skutečným významem původně pokládaného dotazu.

Perspektivní jsou dnes v tomto ohledu zejména přístupy založené na extrakčních ontologiích, jejichž charakteristickým zástupcem může být nástroj:

DBpedia Spotlight – s využitím znalostní báze Wikipedia řeší rozpoznání a disambiguaci jmenných entit, kdy v případě více možných významů daných pojmů, je zde možné odvození nejpravděpodobnější kombinace konkrétních entit, a to podle toho, jak jsou tyto entity vzájemně prolinkovány na Wikipedii a také na základě míry shody jejich popisků vůči zpracovávanému textu.

Např. najdi všechny filmy, které natočil režisér filmu Sherlock Holmes:

```
PREFIX dbprop: <http://dbpedia.org/property/>
PREFIX db: <http://dbpedia.org/resource/>
SELECT ?who ?work ?genre WHERE{ db:Sherlock_Holmes
dbprop:director  ?who  .
?work dbprop:director ?who  .
OPTIONAL{  ?work  dbprop:genre ?genre }
}
```

3.2.4 Učení ontologií

Učením ontologií rozumíme proces, který je ve své podstatě obdobou manuální (intelektuální) tvorby extrakčních ontologií (viz předchozí část 3.2.3), ale je zde snaha o větší zapojení automatizace do fáze komplexního zpracování informací, kdy se s využitím nástrojů umělé inteligence plně automaticky či poloautomaticky nejprve extrahují klíčové termíny, které se následně rozdělují na třídy a instance pojmů. Následuje tvorba taxonomických a netaxonomických relaci a složitějších axiomů s jejich následnou charakterizací. V praxi se při tomto postupu uplatňují (a často vzájemně také prolínají) dva hlavní směry:

- *směr založený na četnostech termínů*, kde např. pokud se ve většině dokumentů obsahující termín t_2 někde v jeho blízkosti vyskytuje také termín t_1, lze následně usoudit, že t_2 by mohl být označením podtřídy vzhledem k t_1 a současně díky jejich obsahové blízkosti ve většině prohledávaných dokumetů by se mohlo jednat o netaxonomickou relaci.
- *směr založený na strukturních vzorech, tzv. Hearst patterns*, což má opět souvislost s klasickou informační extrakcí, kdy je možné odvození složitějších vlastností a vazeb mezi zachytávanými termíny s možností vyjádření např.: *„X a jiné Y“*, *„X je Y, který…“*, *„…tyto Y: X, …“*

Z hlediska životního cyklu, uplatnitelného při návrhu modelů pro učení ontologí nepanuje zjevná shoda na postupu při jejich tvorbě a použité terminologií ani mezi odborníky v oblasti znalostního inženýrství.

Avšak vycházejíce z všeobecně rozšířeného přístupu dle Fowlera [22], lze v zásadě rozlišit dvojí typové metodiky, uplatňující *waterfall* – vodopádový model v různých variacích (fontánový či V-model) nebo *iterativní metodiky* (a jejich varianty v podobě spirálového, inkrementálního, evolučního či prototypového modelu). Podstatný vliv na výslednou podobu metodiky přitom mohou mít potřeby, požadavky a omezení formulované jejími autory, a to obvykle i s ohledem na budoucí uživatele.

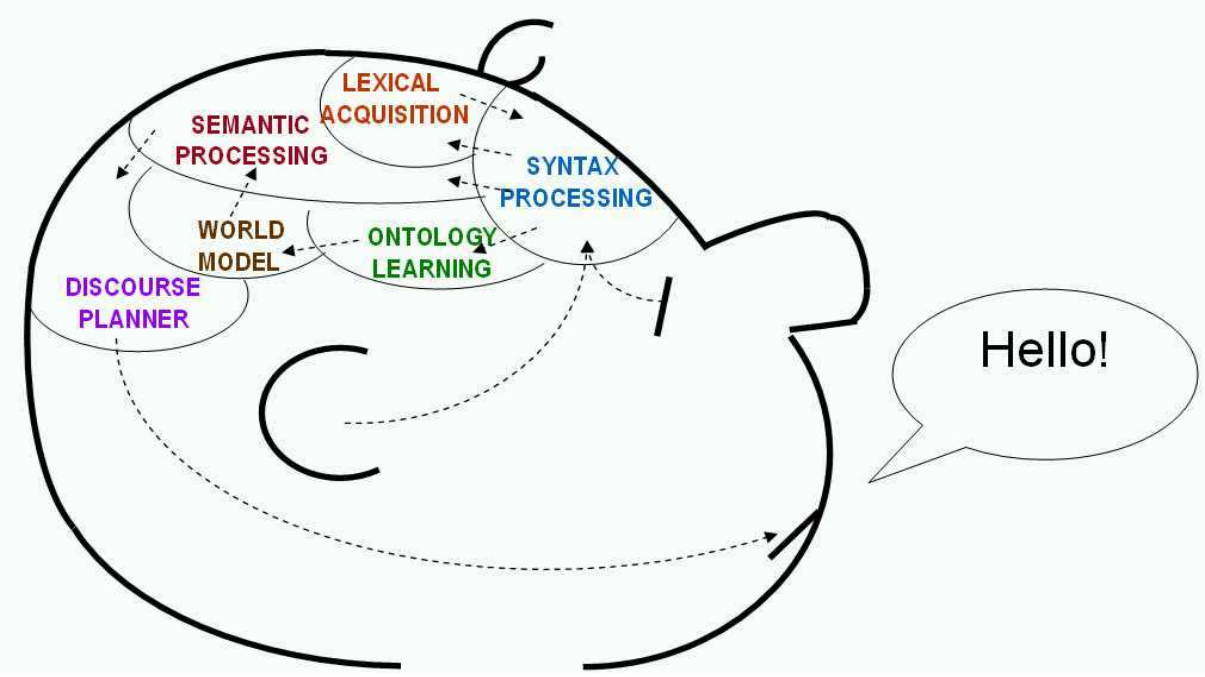
Vodopádový model – separuje projekt na jednotlivé činnosti, v závislosti na specifikaci požadavků z hlediska analýzy, návrhu, testování a implementace. Tyto procedury následují jedna za druhou v pevně stanoveném pořadí, kdy je sice možné vrátit se k předchozí fázi, avšak v praxi je snaha tyto zpětné kroky, spojené s následným opakováním těchto jednotlivých fází, minimalizovat.

Iterativní model – je naopak na plánovaném opakování fází jednotlivých procesů založen. Vychází přitom s myšlenky, že opakovaným řešením, spojeným s verzováním a prototypováním lze dosáhnout kvalitnějšího výsledku. Před zahájením každé iterace je zpravidla navržena základní architektura systému a až následně se řeší a průběžně testují jednotlivé části systému.

Výsledkem každé iterace by měl být funkční vzorek v rozsahu předem stanovené specifikace. V průběhu dalších iterací se pak ověřují nejen nově přidané funkcionality, ale také již hotové a naimplementované části, což podstatně eliminuje riziko zanesení chyb.

Nejobvyklejším způsobem inženýrského návrhu v praxi často bývá deduktivní postup označovaný jako *top-down*, jenž směruje od obecného modelu k implementaci konkrétní části – tzv. **dopředné inženýrství** *(forward engineering)*, kdy se celá doména nejprve rozčlení na základní kategorie, které se poté postupně naplňují konkrétními termíny a pojmy. Pokud se z již hotové funkční aplikace ve směru *bottom-up* dodatečně odvozuje (indukuje) abstraktní konceptuální model, kdy se naopak nejprve shromáždí všechny termíny, které jsou pak následně seskupeny do kategorií, od konkrétní části k obecnému celku, hovoří se o tzv. **zpětném inženýrství** *(reverse engineering)*. Oba tyto směry vývoje mohou být v současné době podporovány softwarovými prostředky pro tvorbu ontologií a do jisté míry automatizovány.

Míra automatizované schopnosti učení ontologií je však vždy přímo úměrná dostatečnému objemu sémanticky anotovaných dat. Pro účely jejich získání je v praxi často žádoucí využití technik pro dolování informací z nestrukturovaných textů s následným uplatněním možností NLP *(Natural Language Processing)*, blíže diskutováno v části 4.4, kde učení ontologií plní svou roli při komplexním procesu zpracování přirozeného jazyka (viz obrázek 3.11), který je dále rozvíjen v kombinaci se sémantickým a syntaktickým zpracováním lexikální akvizice, s využitím modelování situací reálného světa a interpretací mluveného projevu.



Obr. 3.11: Proces učení ontologií při zpracování dotazů v přirozeném jazyce

3.2.5 Řízení znalostí

Pokročilé informační technologie již delší dobu sehrávají důležitou roli v tzv. znalostním managementu (knowledge management), kdy např. v rámci organizace státní správy či větších soukromých podniků, je oprávněnému uživateli k dispozici značné množství informací, které přicházejí z nejrůznějších zdrojů, přičemž každý takový zdroj navíc může poskytovat informace ve svém vlastním specifickém formátu, což dále může znesnadňovat jejich sdílení.

Bez možnosti kvalitativního řízení toku těchto informací pak lze snadno dojít do bodu, kdy sice bude existovat fakticky velké množství informací, ale jen velmi málo prakticky upotřebitelných znalostí, které z těchto informací vycházejí a které by konečnému uživateli mohly být prospěšné, díky čemuž se pak systém stává celkově nepřehledným, a to právě vlivem chaotického zpracování informací, způsobující jejich obtížnou znovupoužitelnost.

Idea knowledge managementu je založena na tom, že informační systémy mohou mezi sebou komunikovat a současně sdílet repozitáře nejen prostých informací, nýbrž také především do znalostních struktur předdefinovaných bází (knowledge base), což by mohlo představovat jednu z možností řešení nevýhod chaotického zpracování informací při současném zavedení centralizované znalostní struktury s využitím Sémantického webu.

Vychází se přitom z předpokladu, že jakákoli informace před vlastní integrací do stávajícího systému musí být nejprve namapována na témata v taxonomii a na entity ve stávající struktuře doménové ontologie, což nám však při následném procesu vyhledávání umožní získání výsledků, které by jinak za normálních okolností zůstaly skryty.

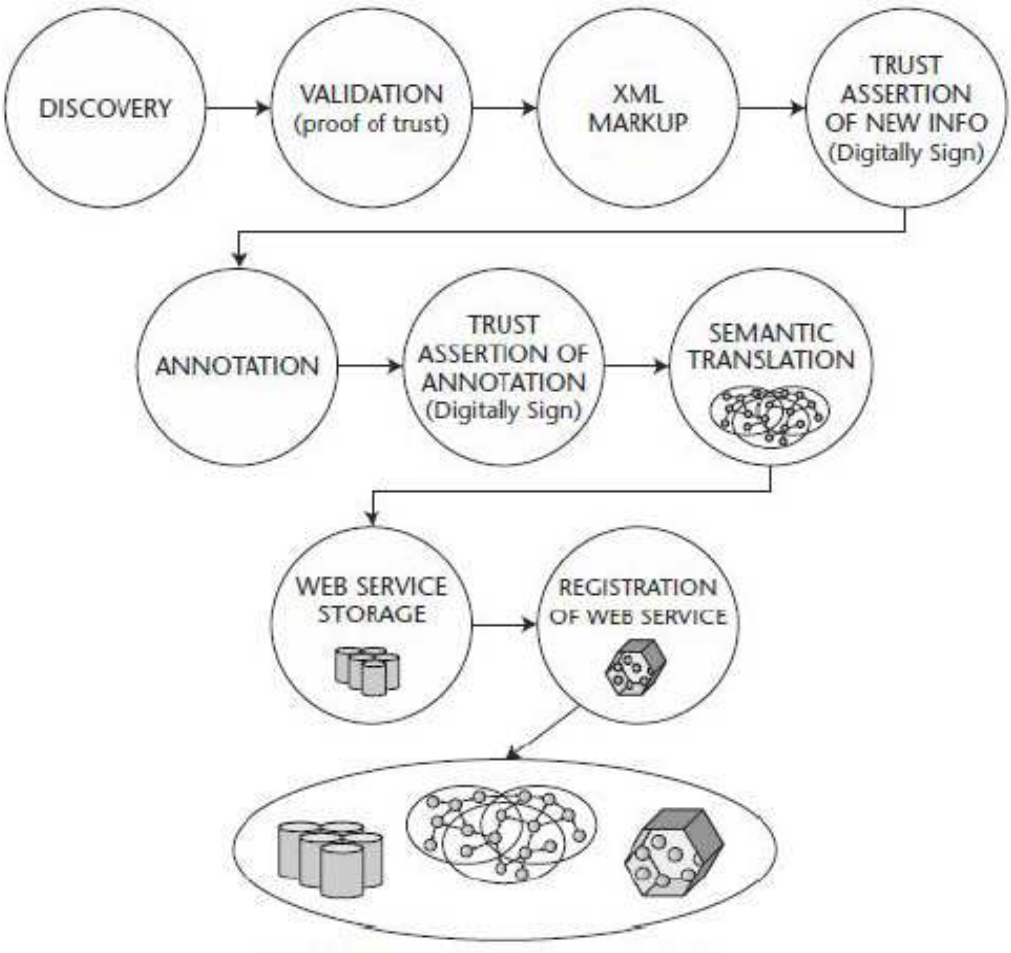
Efektivní znovupoužití informací je přitom přímoúměrné kvalitě i kvantitě takto získávaných informací, což se však aktuálně na pozadí knowledge managementu jeví jako podstatný problém, a to zejména v rámci stavu současného Internetu, kdy je globálně patrná:

- celková informační zahlcenost a přetíženost
- neefektivnost vyhledávání pouze na principu klíčových slov
- nedostatečná důvěryhodnost informací
- nedostatek systémů schopných strojově zpracovat informace v přirozeném jazyce a automaticky nabízet související fakta
- nové typy dokumentů, které nejsou dostatečně indexované
- neustálý nárůst hostitelských počítačů a serverů
- obsah a lokalizace dokumentů podléhající častým změnám

Centralizované zpracování znalostí v typickém procesu knowledge managementu (dle obrázku 3.6) lze rozdělit na několik fází, kdy v první fázi nejprve objevujeme (získáváme) informace z různých zdrojů, které je však následně třeba ověřit, aby nedocházelo k poruše či ztrátě informace, což by mohlo mít za následek pozdější neúplnost a nedokonalost dané znalostní báze.

Plynulým přechodem do produkční fáze se pak postupně (nejčastěji pomocí XML) budoucí znalost značkuje a anotuje (pro potvrzení může být také digitálně podepsána, šifrována a pomocí protokolu XKMS dále bezpečně přenášena).

Následuje fáze sémantické anotace, kdy se blíže vymezí jednotlivé pojmy, vlastnosti těchto pojmů a vztahy mezi nimi (použitím RDF bez porušení původního digitálního podpisu) a posledním krokem je pak jejich samotná aplikace nebo uložení sdílené znalostní báze (pro pozdější korektury nejlépe verzovatelné, umožňující následné zpřesňování a doplňování znalostí), která může být v případě Sémantického webu dále distribuována pomocí webových služeb, které budou blíže teoreticky představeny v následující části 3.3.



Obr. 3.12: Centralizované zpracování informací v rámci znalostního managementu

3.3 Interoperabilita Sémantického webu

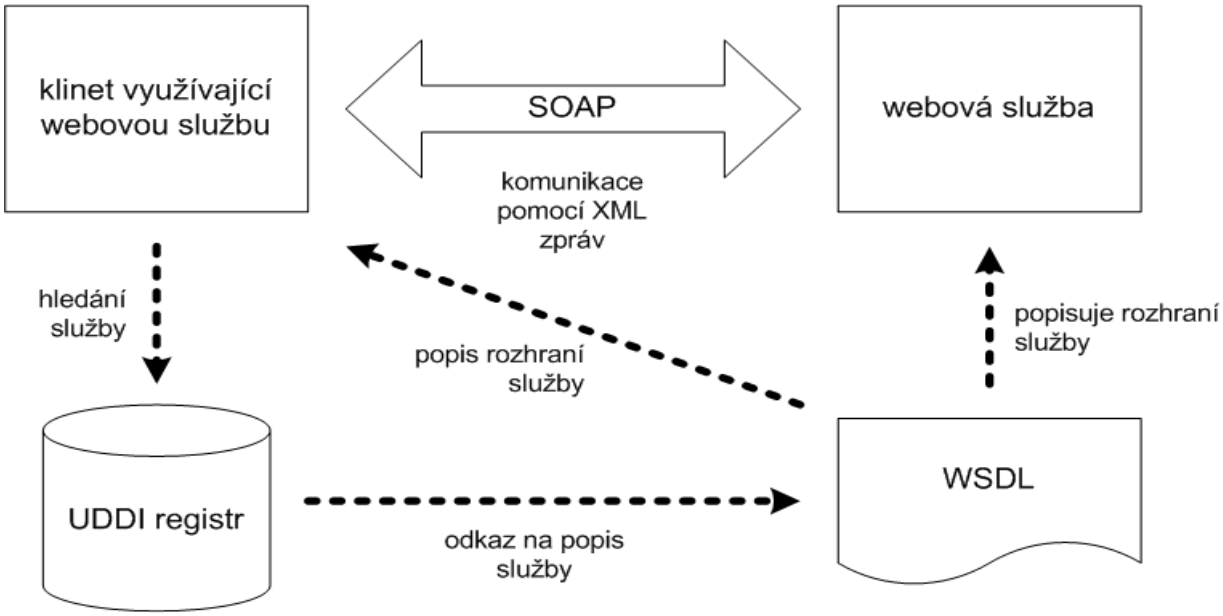
Prostředek pro dosažení interoperability Webu představují *webové služby*, které jsou charakteristické tím, že přijímají i produkují strojově zpracovatelné XML, což umožnuje sloučit koncepce webových služeb a Sémantického webu.

Daconta [16] ve své monografii uvádí tuto definici webových služeb:
Webové služby jsou softwarové aplikace, které mohou být objeveny, popsány a přistoupeny na základě XML a standardních webových protokolů napříč intranety, extranety a Internetem.

Systém webových služeb je založen na souboru tří základních technologií [56]:

- *SOAP (Simple Object Access Protocol)* – protokol používaný pro obousměrnou komunikaci prostřednictvím XML zpráv
- *WSDL (Web Services Description Language)* – standardní formát pro popis veřejného rozhraní webové služby
- *UDDI (Universal Description, Discovery and Integration)* – standardní mechanismus umožňující registraci a vyhledávání webových služeb

Vzájemné vztahy mezi těmito technologiemi zachycuje obrázek 3.7. Ke každé webové službě by měl existovat její formální popis v jazyce WSDL, na základě něhož pak lze automaticky vygenerovat SOAP požadavek. Vlastní formální popis se ve větších systémech nebo přímo v otevřeném Internetu může zaregistrovat do UDDI registru, který následně slouží jako určitá obdoba telefonního seznamu a umožňuje vyhledávání služeb s určitými parametry.



Obr. 3.13: Vztah základních technologií webových služeb

Klient, který chce využít webovou službu, nejprve získá její popis buď přímým dotazem nebo prostřednictvím UDDI. Z něj je jasné, jakou strukturu má mít SOAP zpráva a kam se má webové službě poslat, aby ji rozpoznala [56].

Každá webová služba je identifikována svým URI, na který je možno přistupovat pomocí univerzálně použitelných webových protokolů. Cílem je umožnit zpřístupnění a snadnou spolupráci programů běžících na libovolných platformách. V duchu této vize se pak mohou uživatelé dostat k požadovaným informacím kdykoli a odkudkoli – navíc z libovolného zařízení, tedy nejen z klasického počítače či notebooku, ale také z tabletu či mobilního telefonu. To však předpokládá, že mezi datovými zdroji a uživatelem musí existovat dostatečně spolehlivé a kompatibilní řešení, které je navzájem propojí.

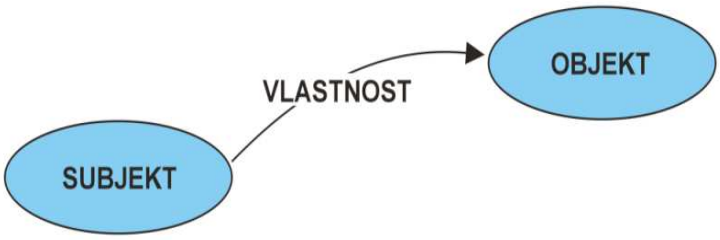
Na Webu lze prohlásit cokoliv o čemkoliv. To znamená, že vztah mezi dvěma objekty může být uložen odděleně mimo informace o těchto objektech. Tímto se Sémantický web liší od objektově orientovaných systémů, které často implementují entitně-relační modely, jež obecně předpokládají, že informace o objektu je uložena v objektu samotném; definice třídy objektu tedy definuje úložiště odvozené od jeho vlastností [7].

To je současně jedním z východisek pro zajištění interoperability aplikací Sémantického webu formou strojově čitelných taxonomií. *Taxonomie* je obecně způsob klasifikace či kategorizace množiny věcí do *hierarchie*, tvořené stromovou strukturou. V kontextu informačních technologií chápe Daconta [16] význam taxonomie jako klasifikaci informačních entit ve formě hierarchie podle předpokládaných vztahů entit reálného světa, které mohou zastupovat. Jinými slovy se člověk či spíše stroj pomocí taxonomie dokáže vyznat v množině pojmů dané domény, mezi nimiž mohou existovat *generalizační* nebo *specializační* vazby, přičemž generalizací rozumíme procházení taxonomií směrem nahoru (zobecňování), procházení směrem dolů je pak specializací dané ontologie.

Taxonomie může vytvářet jednoduchou sémantiku, jestliže entity jsou tříděny podle jejich významu. Entity jsou rozlišovány pomocí rozlišovacích vlastností, které tak určují hranici mezi podřazenou a nadřazenou entitou (třídou).

Jak již bylo v úvodu této podkapitoly uvedeno, syntaktická interoperabilita je v prostředí Sémantického webu zajištěna použitím jazyka XML. Základ sémantické interoperability je poté postaven na pěti stěžejních pilířích [16]:

- logická tvrzení,
- klasifikace,
- formální modely dat,
- pravidla, a
- důvěryhodnost.



Logická tvrzení lze konstruovat z klíčových částí věty propojením podmětu a předmětu přísudkem. Na tomto principu je založeno RDF, které zachycuje vazby mezi subjekty a objekty, které tímto definují logické výroky o zdroji. Cílem *klasifikace* je pak vytvořit seskupení jednotlivých *tříd* a *podtříd*, na kterých je následně možné provedení generalizace. Strojové odvozování navíc vyžaduje precizní *formalismus* při modelování tříd i podtříd a vazeb mezi nimi.

Např. vlastnost lze blíže rozlišit jako symetrickou či tranzitivní a lze ji definovat jako binární relaci ρ, která o každé uspořádané dvojici (a, b) ∈ ρ prohlašuje, že x *má vlastnost* y. Taková symetrická vlastnost pak může být formálně popsána takto:

$$x\rho y \Rightarrow y\rho x$$

a tranzitivní takto:

$$x\rho y \wedge y\rho z \Rightarrow x\rho z$$

Pravidla umožňují odvodit závěr na základě množiny premis. Jedním ze základních pravidel je pravidlo *modus ponens* (pravidlo odloučení):

$$A, A \Rightarrow B \vdash B$$

Na základě pravidel obsažených v ontologiích tak lze odvozovat nové informace, přičemž důležitým pilířem pro zajištění pravdivosti informací v Sémantickém webu je taktéž *důvěryhodnost*, zajišťovaná kryptografickými metodami elektronického podpisu a šifrování [16].

3.4 Sémantická analýza a interpretace

Proces sémantické analýzy je založen na statistických modelech a algoritmech pro trénování ontologií. Pro tato modelová učení je však zapotřebí existence tzv. anotovaného datového korpusu, který může využívat již hotové nebo nově zaváděné nástroje pro předzpracování (preprocessing tools) jako: tokenizéry, morfologické taggery nebo rozpoznání jmenných entit (NER).

To je je obzvláště důležité zejména v případě, kdy se na základě modelových parametrů z anotovaného korpusu vyhodnocují dotazy interpretované prostřednictvím přirozeného jazyka. Vlastní proces zpracování pak probíhá v chronologickém pořadí: předzpracování (lemmatizace a morfologická kategorizace), následné rozpoznání jmenných entit a nakonec vlastní sémantická analýza s využitím trénovacího modelu sestaveného na základě anotace pojmů a větné skladby v analyzovaném textu. Formální definice sémantického modelu pro dotazové zpracování bude předmětem následující části.

3.4.1 Formální definice

Formální vlastnosti, s ohledem na úvodem uvedený příklad zavedené definice modelu sémantické analýzy, mohou formálně být navrženy následovně:

Nechť $S = w_1 \ldots w_W$, je větou, kde S (otázka) obsahuje W slov $w_1 \ldots_W$ a budiž $T(Subj, Pred, Obj)$ jako uspořádaná trojice (triplet) obsahující subjekt, predikát a objekt. Sémantickou anotaci *Sem* takové věty je možno zapsat jako neuspořádanou množinu M takových tripletů:

$$Sem(S) = \{T_1, T_2, \ldots T_M\}.$$

Uvnitř $Sem(S)$ lze triplety také spojovat. Formálně, pak tedy dva triplety

$$T_x(Subj_1, Pred_1, Obj_1),\ T_y(Subj_2, Pred_2, Obj_2)$$

mohou sdílet své subjekty a objekty, tedy $Obj_1 = Subj_2$ or $Subj_1 = Subj_2$, formují defakto orientovaný graf.

Jmenné entity *(NE)* – je-li N^{τ} definováno jako instance pojmenované entitou *(Named Entity)* typu τ, poté bude platit, že:

$$N^{\tau}{}_{j,k} = N^{\tau}{}_{span}(w_j \ldots w_k)$$

je jmennou entitou typu τ , která leží v oblasti rozmezí *(span)* výskytu slov od w_j do w_k, kde $1 \leq j < k \leq W$. To znamená, že jmenné entity jsou asociovány se slovy zastoupenými ve větě. Nechť také C je definováno jako sémantický koncept a současně C není asociováno s žádným slovem.

Obě N^{τ} a C současně mohou být částí tripletu, a to tak, že tento může nabývat jedné z následujících forem:

$$T(Subj, Pred, Obj) = \begin{cases} T(C, Pred, N^{\tau}_{j,k}) & \text{je-li } subj\ C \text{ a } obj \text{ je } NE \\ T(N^{\tau 1}_{j,k}, Pred, N^{\tau 2}_{o,p}) & \text{je-li } subj \text{ a } obj\ NE \\ T(N^{\tau}_{j,k}, Pred, C) & \text{je-li } subj\ NE \text{ a } obj \text{ je } C \\ T(C_a, Pred, C_b) & \text{je-li } subj \text{ a } obj\ C \end{cases}$$

kde j, k, o, p jsou možná rozmezí výskytu jmenných entit, a, b jsou zde použity k determinaci možných rozdílných konceptů C a $\tau 1$ a $\tau 2$ mohou být rozdílné typy použitých jmenných entit uvažovaných v rámci užitého statistického modelu.

3.4.2 Statistický model

Na základě výše zmíněného formálního zápisu lze vyvodit, že úkolem sémantické analýzy je nalezení odpovídající sémantické anotace Sem(S), jež je obsažena ve větě S. S využitím pravděpodobnosti tak lze poté celý problém formulovat následovně:

$$P(Sem(S)|S) = P(T1, T2, \ldots TM|S).$$

Současně s předpokladem, že triplety jsou nezávislé, vyplývá

$$P(T_1, T_2 \ldots T_M|S) \approx \prod^{M} P(T_m|S)$$

kde

$$P(T_m|S) = P(T_m(Subj, Pred, Obj)|w_1 \ldots w_w)$$

Za předpokladu, že uvažovaný model bude omezen pouze na v praxi nejčastěji zastoupené první dvě z možných forem tripletů (viz předchozí část 3.4.1), tedy formálně $T = (C,Pred,N^{\tau})$ a $T = (N^{\tau 1},Před,N^{\tau 2})$, pak pravděpodobnost výskytu pro jednotlivé triplety lze aproximovat jako:

$$P(T(C, Pred, N^{\tau})|S) \approx P\left(T(C, Pred, N^{\tau})\middle|N_{j,k}^{\tau}\right) . P(N_{j,k}^{\tau}|S)$$

a

$$P(T(N^{\tau 1}, Pred, N^{\tau 2})|S) \approx P\left(T(N^{\tau 1}, Pred, N^{\tau 2})\middle|N_{j,k}^{\tau 1}, N_{o,p}^{\tau 2}\right) . P(N_{j,k}^{\tau 1}|S). P(N_{o,p}^{\tau 2}|S)$$

Pro odhad modelových parametrů budiž uvažován korpus trénovacích dat Ŝ, obsahující soubor vět s jejich sémantickou anotací Ŝ = {Sn, Sem(Sn)}. Pravděpodobnost výskytu jednotlivých tripletů pak lze zapsat jako:

$$P(T(C, Pred, N^{\tau})|N^{\tau}) = \frac{Cnt(T(C, Pred, N^{\tau}))}{Cnt(N^{\tau})}$$

a

$$P\left(T\left(N_{j,k}^{\tau 1}, Pred, N_{o,p}^{\tau 2}\right)\middle|N_{j,k}^{\tau 1}, N_{o,p}^{\tau 2}\right) = \frac{Cnt(T\left(N_{j,k}^{\tau 1}, Pred, N_{o,p}^{\tau 2}\right), N_{j,k}^{\tau 1}, N_{o,p}^{\tau 2})}{Cnt(N_{j,k}^{\tau 1}, N_{o,p}^{\tau 2})}$$

Takto představený model určující pravděpodobnost výskytu tripletů však dále nebere v úvahu všechny souvislosti mezi rozpoznanými jmennými entitami. Ke zlepšení odhadu pravděpodobnosti výskytu tripletů v návaznosti na kontext jmenných entit je třeba modelový zápis pozměnit do tvaru:

$$P(T(C, Pred, N^{\tau})|S) \approx P\left(T\left(C, Pred, N^{\tau}\right)\middle|N_{j,k}^{\tau}, w_{j-1}\right).P(N_{j,k}^{\tau}|S)$$

a

$$P(T(N^{\tau 1}, Pred, N^{\tau 2})|S) \approx$$
$$P\left(T\left(N_{j,k}^{\tau 1}, Pred, N_{o,p}^{\tau 2}\right)\middle|N_{j,k}^{\tau 1}, N_{o,p}^{\tau 2}, w_{j-1}, w_{o-1}\right)$$
$$.P\left(N_{j,k}^{\tau 1}\middle|S\right).P(N_{o,p}^{\tau 2}|S)$$

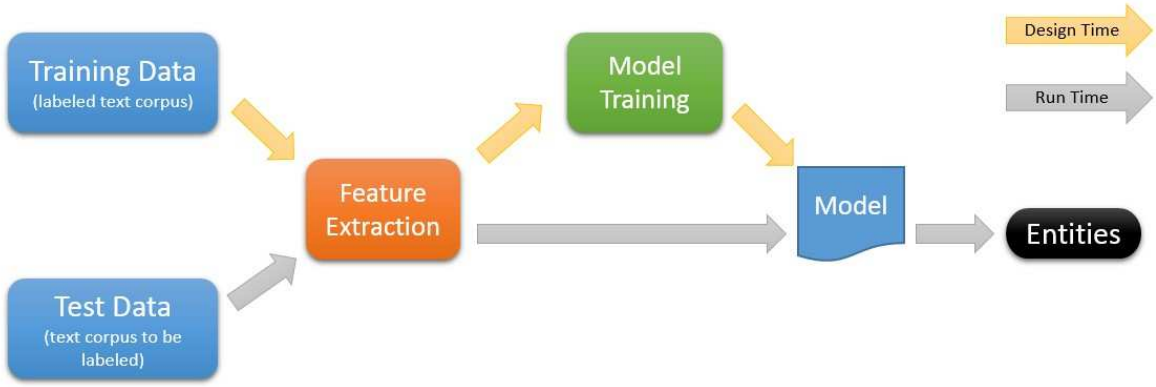
což představuje finální verze zápisu modelu využívající princip *lemmatizace* v kontextu předchozího slova. Úkolem lemmatizace je převod slova na jeho základní tvar *(lemma)*, který následně plní úlohu selekčního znaku, vytvořeného ze zdrojového slova. Z pohledu zaměření procesu indexace a vyhledávání, jehož cílem je prvotně poskytnutí relevantních dokumentů. Není však přitom apriori nutné, aby použité algoritmy stavěly na poznatcích jazykovědeckých – tedy například lemmatem nemusí, ale může být základní slovníkový tvar či kořen slova. Tyto selekční znaky vstupující do procesu indexování je pak následně možné ze samotného dokumentu a jeho obsahu odvodit pomocí metodiky rozpoznávání jmenných entit, které bude blíže popsáno v následující části.

3.4.3 Rozpoznání jmenných entit

NER *(Named Entity Recognition)* je speciální podobor extrakce informací, kterým se v teoretické aplikaci extrakčních ontologií zabývá část 3.2.3 této práce. NER je současně také stěžejní částí *preprocessing* fáze zpracování dotazů v přirozeném jazyce, kde se jmenné entity vyskytují jako gramatické a lexikální jevy v běžném textu a lze jich využít pro následné zpřesnění detekce. Některé z jejich vlastností mohou být univerzální prakticky pro všechny lingvistické jazyky, jiné jsou však specifické pouze pro jeden konkrétní jazyk, což proces rozpoznávání jmenných entit dále komplikuje [51].

Například čeština ve srovnání s angličtinou je jazykem, kde se vyskytuje poměrně komplikované skloňování a časování, přičemž současně také každé slovo může mít v závislosti na kontextu několik možných podob. Použitím morfologického analyzátoru a dříve uváděné lemmatizace se s mnohotvárností českého jazyka lze vyrovnat. Tyto přístupy ovšem vyžadují existenci dostatečně obsáhlého sémanticky označkovaného datového korpusu, na základě kterého se pak samotný systém může postupně učit a dále zpřesňovat zachytávané vazby, vztahy a závislosti mezi jednotlivě označenými entitami.

Schématické znázonění procesu zpracování jmenných entit s využitím trénovacích dat je zobrazeno na obrázku 3.14. Vyznačené šipky představují vazby při procesu učení (Design Time), kdy systém využívá existujícího korpusu trénovacích dat (Training Data) k vytváření modelu (Model Training), který představuje soubor s připravenými příklady zobecněných dotazů, ve kterých jsou jednotlivé entity již předem sémanticky anotovány, což při následném porovnání s dosud neanotovaným textem v samotném procesu rozpoznávání na testovacích datech (Run Time) může podstatně zvýšit šance na korektní zachycení vlastností a vztahů mezi jednotlivými entitami (Feature Extraction), na čemž ve finále závisí i samotná učinnost navrženého postupu.



Obr. 3.14: Schéma procesu zpracování jmenných entit s využitím trénovacího modulu

Testovací data (Test Data), tvoří dosud nijak předem anotovaná data, která se někdy označují jako slovní rovina. Prvním krokem v procesu generování slovní roviny, je princip tokenizace, kdy dochází nejprve k základnímu rozdělení textu na odstavce, dělení podle mezer, interpunkce, konců řádků a vymezení samotných slov. Následně pak provedením *morfologické analýzy* lze takto odděleným slovům přiřadit množinu možných kombinací lemmat a značek (tagů), čímž dále dochází ke zjednoznačnění kontextu, ve kterém se slovo ve větě nachází – např. věta: „*Hlavní výhodou Webu je bezesporu jeho flexibilita*“, může při procesu morfolologické analýzy se zaměřením na slovo *hlavní* nabývat hned několika možných výsledků. Může jít například o podstatné jméno *hlaveň* v druhém pádu jednotného nebo sedmém pádu množného čísla, ale může také jít o přídavné jméno *hlavní*. To ovšem se samotného tvaru nelze jednoznačně určit a je třeba zohlednit také kontext, v jakém se slovo ve větě nachází.

S morfologickou analýzou se pojí také další z procedur zpracování přirozeného jazyka – *morfologické generování*, kdy je cílem pro dané lemma a značku vygenerovat odpovídající tvar slova, tak aby každé dvojici lemma-značka, odpovídalo vždy nanejvýš jedno konkrétní slovo. Relativní nepřesnost výsledků při automatickém rozpoznání entit však stále zůstává ovlivněná faktem, že text není zpracováván lidským mozkem, ale softwarem a závisí tak vždy na kvalitním předzpracování a nutné znalosti analyzovaného vzorku.

3.5 Dostupné softwarové nástroje

Pro vývoj a správu Sémantického webu byla od jeho zrodu představena již řada řešení, rozvíjená zejména na poli otevřených open-source řešení, ale jak postupem času docházelo k prolínání původně čistě vědecké oblasti s praxí, začali se objevovat také zástupci komerční sféry. Významným produktem v této oblasti je např. Oracle Database 11g Semantic Technologies [48], který je typickým zástupcem tzv. RDFStore, představující kategorii permanentního úložiště RDF dat s možností všech běžných databázových operací.

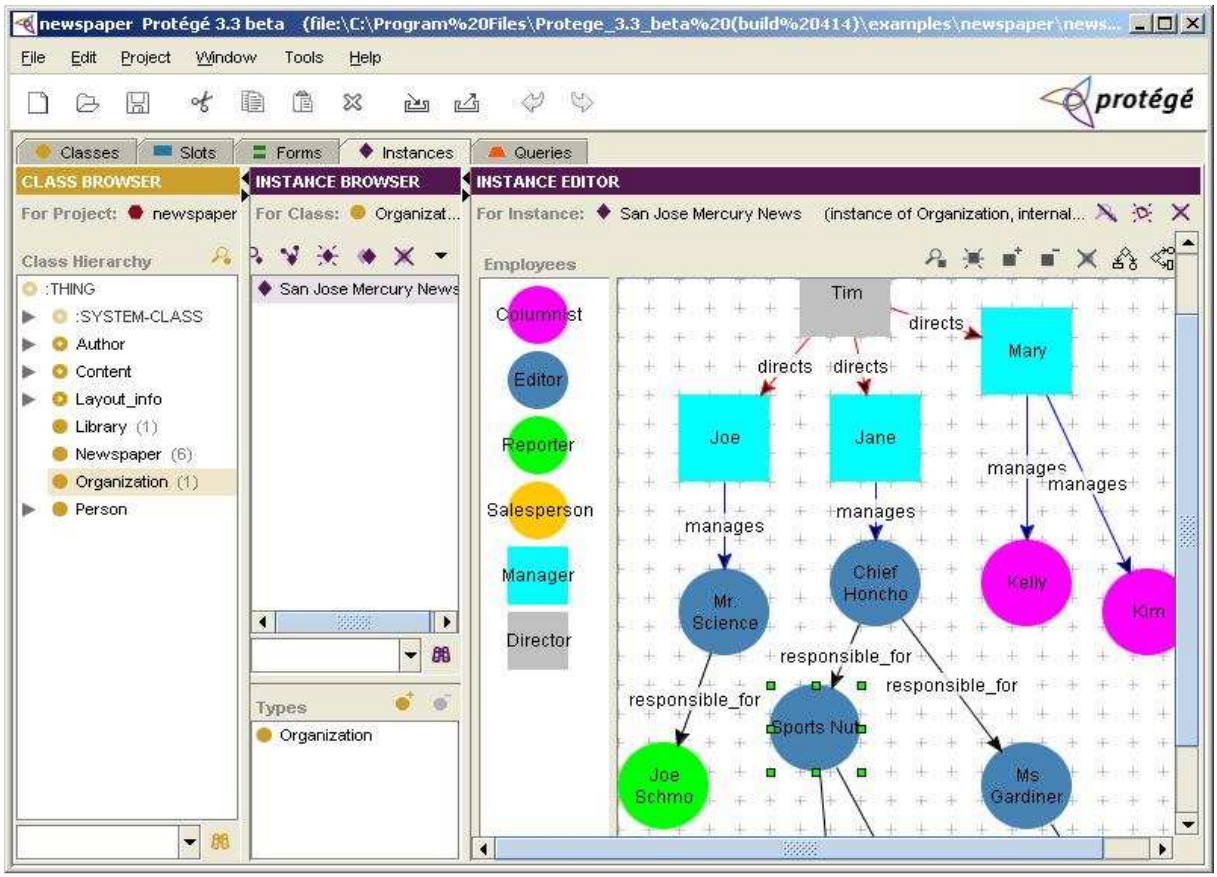
Obecně krom datových úložišť sémantických dat (ve formě uspořádaných tripletů) rozlišujeme také software pro implementaci jednotlivých komponent. Od textových editorů až ke komplexním frameworkům a odvozovacím systémům (statisticky porovnávaných na konkrétním případě v rámci praktické části 4.5 této práce). Z dostupných softwarových nástrojů bude obsahem následujících částí stručně představeno a specifikováno několik v současné praxi poměrně rozšířených nástrojů, určených zejména pro práci s RDF daty a ontologiemi, na jejichž funkčním principu Sémantický web staví.

3.5.1 Protégé

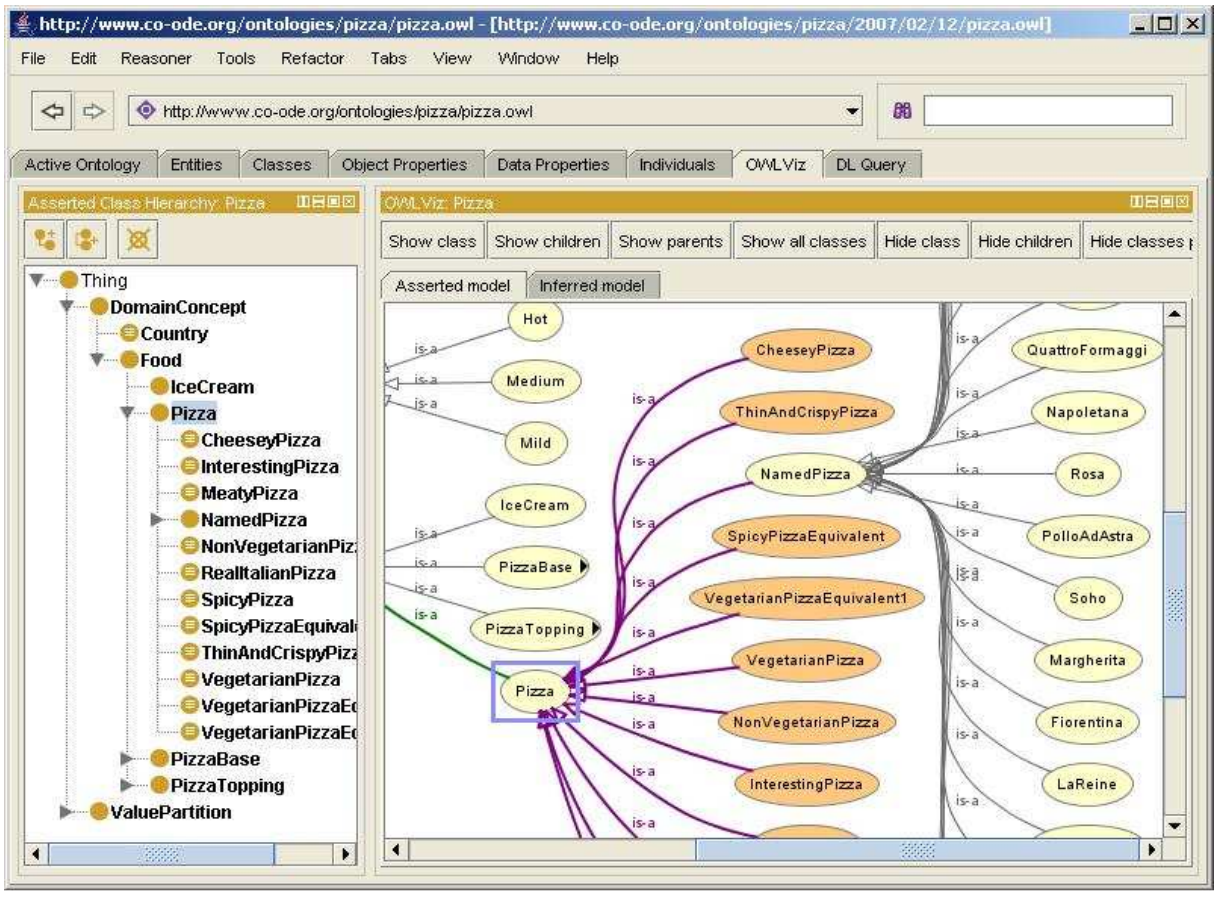
Protégé (http://protege.stanford.edu/) je volně šiřitelný platformově nezávislý prostředek pro vývoj a správu ontologií a znalostních bází. Aplikace je vyvíjena v Javě a je dostupná prakticky na všech platformách: Win, MacOS, Linux, Unix. S výhodou lze využít také rozšiřujících přídavných modulů, tzv. pluginů např. pro vizualizaci ontologií, zpracování přirozeného jazyka apod.

Celosvětová komunita čítá v současnosti již přes 250 tisíc členů z řad vývojářů akademické i soukromé sféry, což z této aplikace činí základ pro rychlé vytváření prototypů a vývoj nových ontologií v rámci inteligentních systémů. Výsledný export ontologií je možno provést pro různé formáty, zejména pak RDF(S), OWL a XML Schema. Platforma Protégé podporuje dva typy modelování ontologií, prostřednictvím základního editoru Protégé-Frames (obrázek 3.15) nebo nádstavbového řešení Protégé-OWL (obrázek 3.16) [38].

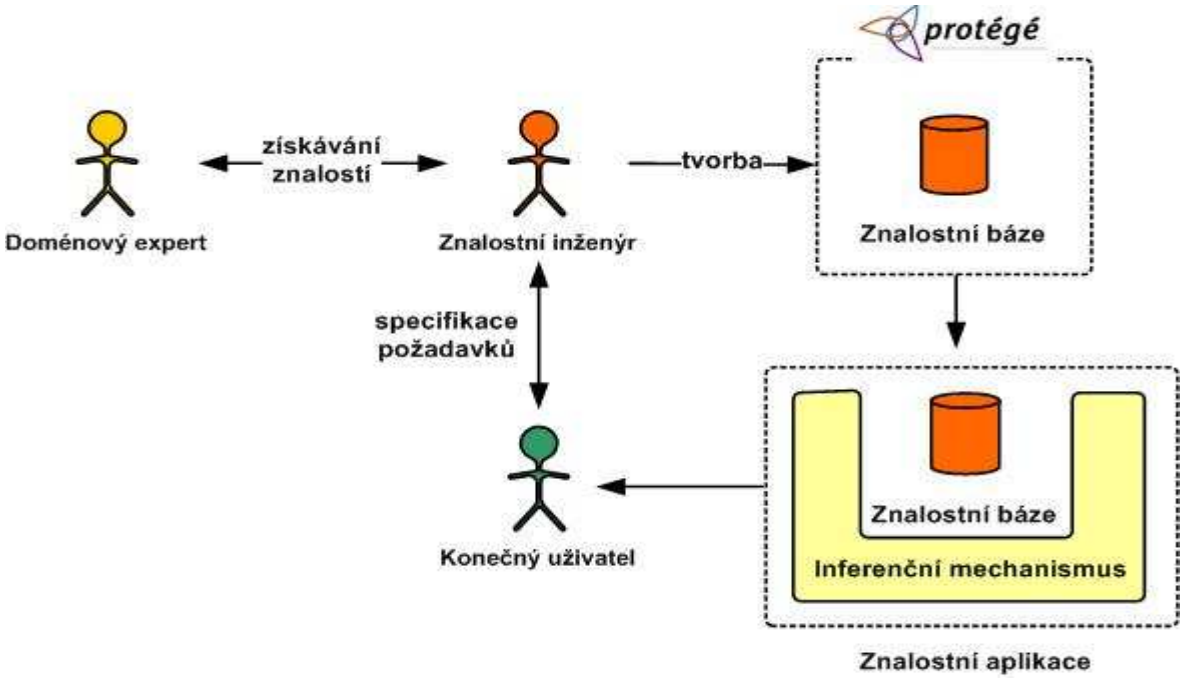
Protégé-Frames umožňuje základní klasifikaci tříd a instancí pomocí plnohodnotně uživatelsky uzpůsobeného rozhraní pro tvorbu základny budoucí doménové ontologie, kterou pak následně s využitím Protégé-OWL je možno dále analyzovat a všechny takto zadefinované vztahy a relace mezi jednotlivými klasifikovanými třídami též také posléze vizualizovat v interaktivním editoru. Z uživatelského pohledu se tak jedná o jedinečnou možnost přehledného zobrazení i relativně složitých struktur, které se tak stanou mnohem čitelnější.



Obr. 3.15: Protégé-Frames pro definici a správu jednotlivých tříd či jejich instancí



Obr. 3.16: Protégé-OWL pro vizualizaci ontologií a inferenčních modelů



Obr. 3.17: Role softwarového nástroje Protégé při tvorbě znalostní báze

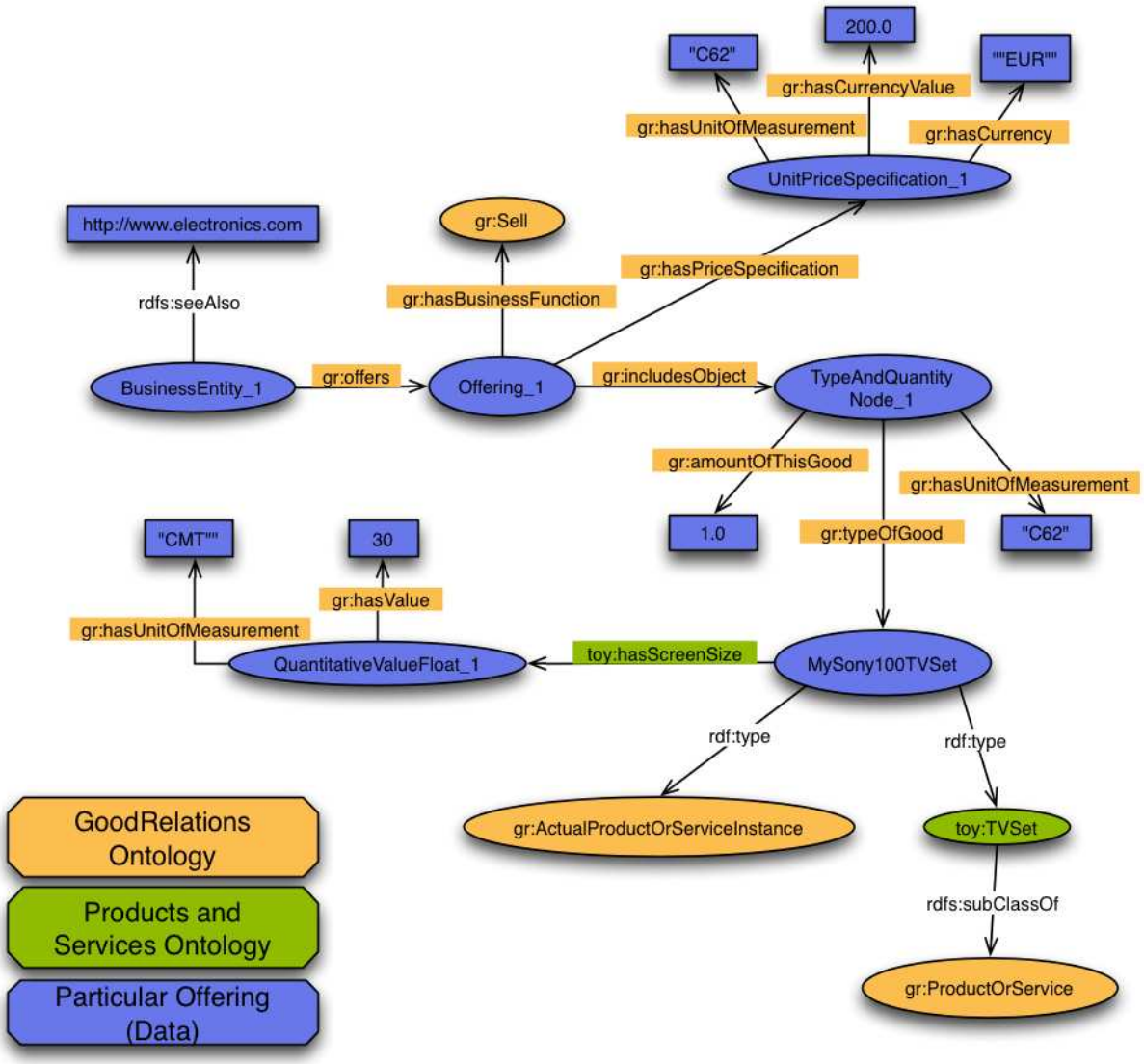
Původním cílem při vývoji aplikace bylo usnadnění práce ontologickým inženýrům při tvorbě a implementaci tzv. *znalostních bází*. Obrázek 3.17 ukazuje souvislosti mezi Protégé a znalostní aplikací obsahující bázi znalostí.

Je však nutno dodat, že (byť se tak na první pohled tváří) Protégé není expertní systém ani program, který by sloužil přímo k jejich tvorbě a poskytoval tak expertní rady a doporučení vhodné pro danou situaci. Může však významně dopomoci při vytváření stěžejní části takovýchto expertních systémů, kterou je právě báze znalostí. Tím, že bude báze znalostí vytvářena odděleně od tvorby znalostní aplikace, bude ji pak možno lépe udržovat a spravovat.

3.5.1 GoodRelations

GoodRelations je standardizovaným slovníkem pro popis produktů v oblasti e-commerce, který může být snadno integrován do již stávajícího obsahu statických i dynamických stráněk webové prezentace či e-shopu.

Nově jsou GoodRelations plně kompatibilní se specifikací mikroformátů, zavedenou odsouhlaseným standardem HTML 5 a mohou být také použita jako rozšíření pro slovník mikrodat ze *Schema.org*, což dále může pomoci ve zvýšení viditelnosti takto označených produktů a služeb ve vyhledávačích, které tuto syntaxi zohledňují. Je proto víc než zřejmé, že vlivem tohoto faktu tak již v současnosti sami obchodníci mají a budou mít zájem na tom tímto způsobem upravená a o další informace obohacená data generovat, což dále může dopomoci k rozšíření technologií Sémantického webu mezi širší masu uživatelů.



Obr. 3.18: Příklad GoodRelations ontologie s využitím zápisu pomocí RDF Grafu

Obrázek 3.18 popisuje strukturu RDF grafu jednoduchého příkladu pro určení úhlopříčky a ceny televizoru na základě ontologické struktury, která je již od počátku vyvíjená jako doménově nezávislá a je tudíž univerzálně použitelná na jakýkoli typ zboží. Cílem GoodRelations je definovat strukturu, která je validní napříč poskytovanými platformami, syntakticky nezávislá a schopná práce s mikrodaty na základě zápisu v populárních formátech (RDFa, RDF/XML, Turtle, JSON, OData, GData apod.) [54].

Základní třídy GoodRelations pak tvoří:

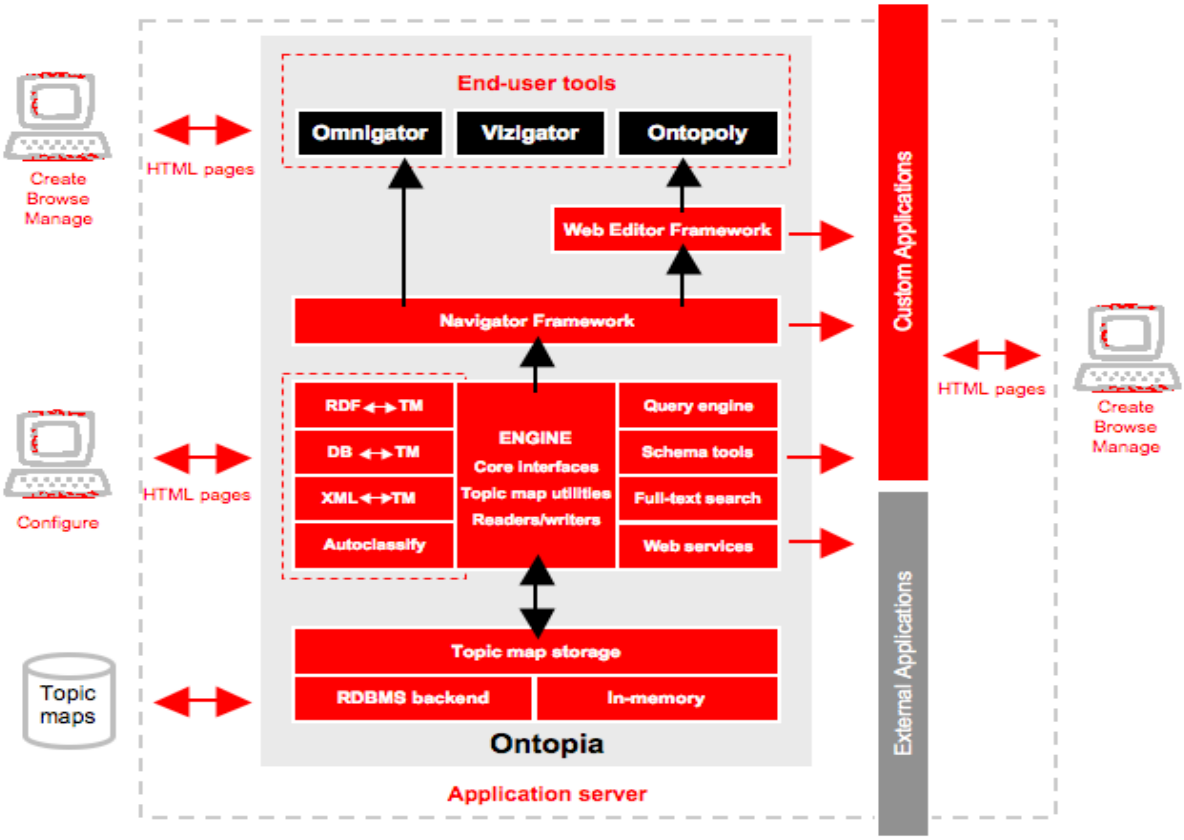
- **BusinessEntity** – hlavní doména, která se pojí s jednotlivými nabídkami
- **Offering** – vlastní nabídka produktu, který prodejce v dané entitě prodává
- **ProductOrService** – bližší specifikace instance produktu (např. model)
- **Location** – určení oblasti, pro kterou nabídka platí, popř. adresa obchodu

3.5.1 Ontopia

Ontopia je sada nástrojů pro tvorbu ontologií na základě TM (*Topic Maps)*, což je mezinárodní standard pro reprezentaci a výměnu znalostí s podporou navigace a vyhledávání v rozsáhlých kolekcích. Aplikace samotná, jejíž celá platforma je založená na rozšiřitelném Java API, pak plní funkci jakéhosi RDF kovertoru, schopného pokročilých operací s uvedenými Topic mapami [50]:

- číst a modifikovat jakoukoli část topic mapy
- importovat topic mapu ze souboru (v CTM, TM/XML či LTM formátu)
- exportovat topic mapu do souboru (v XTM, TM/XML či LTM formátu)
- konverze z RDF do topic map (a naopak)
- provádět zpracování dotazů prostřednictvím vlastního jazyka Tolog
- možnost full-textového vyhledávání a automatické klasifikace

Engine aplikace udržuje topic mapy v paměti nebo je ukládá do relační databáze (Oracle, MySQL, PostgreSQL). Díky editoru *Ontopoly* je možné navržení vlastní mapy prostřednictvím uživatelsky přívětivé aplikace přímo ve webovém prohlížeči a možná je také jejich grafická vizualizace s využitím rozšíření *Vizigator*. Pomocí browseru *Omnigator* zase vývojáři mohou lépe unifikovat a ladit své nově vytvořené Topic mapy s již aktuálně vytvořenými.



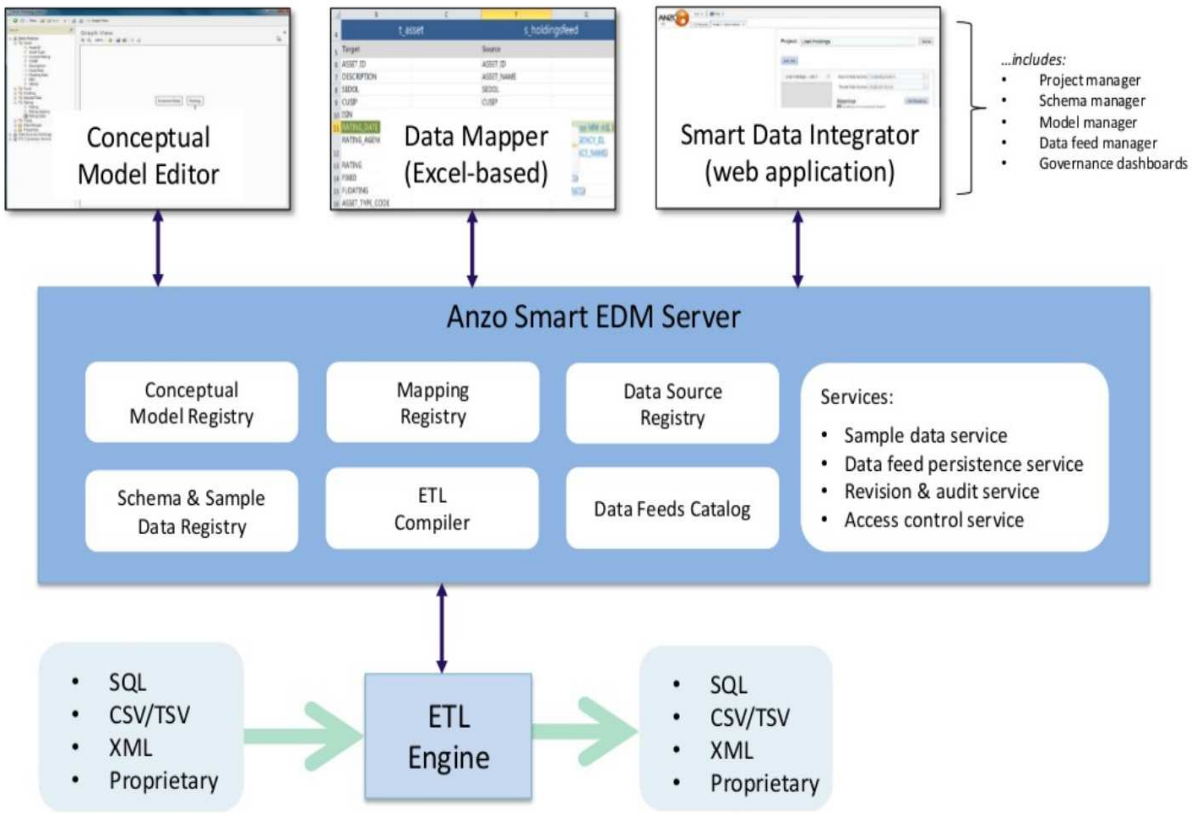
Obr. 3.19: Funkční schéma softwarového nástroje Ontopia

3.5.2 Anzo

Anzo je zástupcem komerční multiplatformní aplikace vyvíjené pod iniciativou Cambridge Semantics, jež umožňuje i netechnicky založeným uživatelům snadno vytvářet individuální řešení zejména v oblasti bussiness analýzy EDM *(Electronic Document Management)* dle standardů W3C.

Obsahuje Smart Data Integrator, který využívá klasické konceptuální modely k mapování a integraci datových zdrojů ve formě schémat, datových katalogů a registů, které je možné pomocí ETL *(Extract-Tranform-Load)* kompilátoru dále zpracovávat. Mapování ontologií je možné provádět jak přes konceptuální modelovací editor, tak i online přes webovou aplikaci Smart Data Integratoru.

Co je však hlavní výhodou Anzo řešení oproti obdobným produktům je možnost napojení dat ze sešitů hojně používaného kancelářského produktu MS Office, respektive známého tabulkového procesoru MS Excel. Tento rozšiřující modul umožňuje, aby jednotlivé sešity (spreadsheety) mohly být mapovány do ontologií, a to i prostřednictvím vestavěného vizuálního editoru, což dále ještě může usnadnit práci s tímto nástrojem i pro méně technicky zdatné uživatele. Data v těchto sešitech jsou následně transformována do RDF formátu uspořádaných trojic a připravena k uložení na Anzo EDM serveru, který slouží zároveň jako cloudové řešení pro sdílení dokumentů mezi uživateli [46].



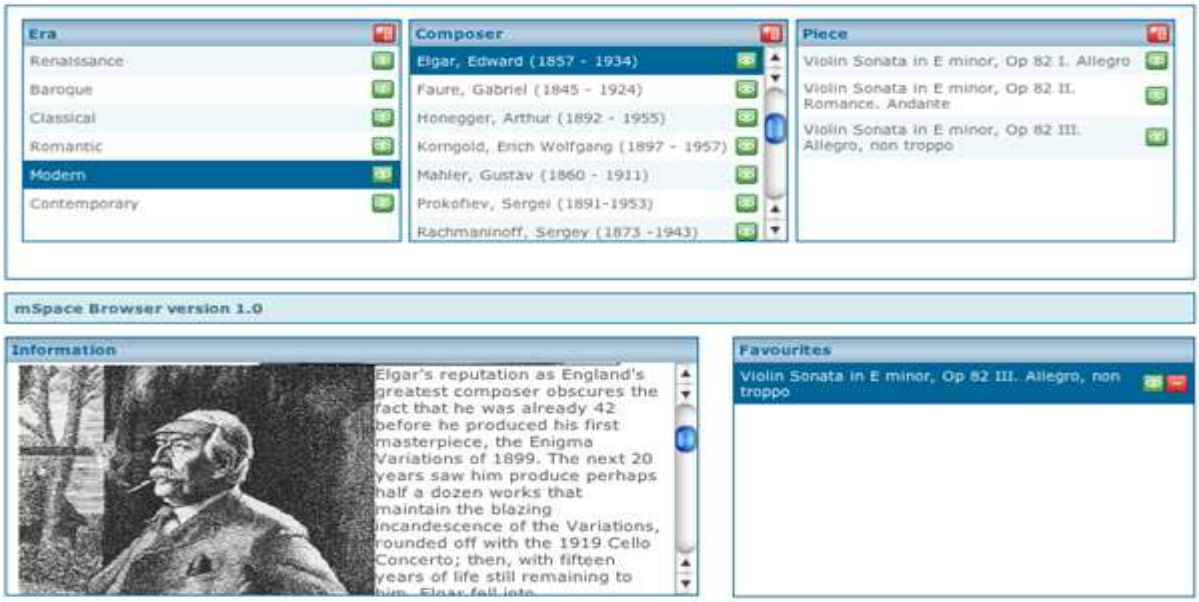
Obr. 3.8: Základní architektura Anzo EDM serveru

3.5.3 mSpace

Služba *mSpace* je ukázkou praktického využití interakčního modelu softwarového frameworku v multimediálním prostředí se snahou o interaktivní zmapování a učení nových faktů prostřednictvím individuálních přístupů k vyhledávání při zkoumání vztahu mezi informacemi samotnými uživateli. Jako doménová ontologie představuje několik sdružených kategorií (původně úzce zaměřená na doménovou oblast klasické hudby v jednom velkém informačním prostoru – odtud tedy zřejmě název mSpace), kdy umožňuje uživateli manipulovat s předkládanými výsledky, tak aby lépe vyhovovaly hledaným zájmovým oblastem, ve kterých však daný uživatel nemá detailnější znalosti.

Příkladem může být volně představená myšlenka: *„Nevím mnoho o klasické hudbě, ale poznám, co se mi líbí, když to uslyším. Jak ale najít klasickou hudbu k poslechu když nevím, jak zní melodie od Bethowena či Bacha? Jaký je rozdíl mezi romantickou a barokní houslovou hudbou?“*

Tradiční postup při řešení tohoto problému ve formě obecného dotazování se pomocí klíčových slov („klasická hudba“) v některém ze známých vyhledávačů vrací obsáhlý seznam terminologických pojmů z dané oblasti a trvá zpravidla velmi dlouho než je požadovaná znalost mezi tímto kvantem informací objevena. mSpace zejména v oblasti multimediálního obsahu (hudba, filmy) nabídne uživateli nejprve možnost přehrání krátkých ukázek a podle nich následně uživateli představí seznam tématicky podobného obsahu (např. soupis těch autorů, kteří napsali houslové sonáty v E-moll). Uživateli je opět následně k poslechu nabídnuta ukázka skladby konkrétního autora a sám si pak může navíc ještě přidávat kategorie pro detailnější řazení dle nejrůznějších žánrů [61].



Obr. 3.21: Proces vyhledání dotazu od úvodní ukázky až ke konkrétnímu autorovi

3.6 Sémantické vyhledávače

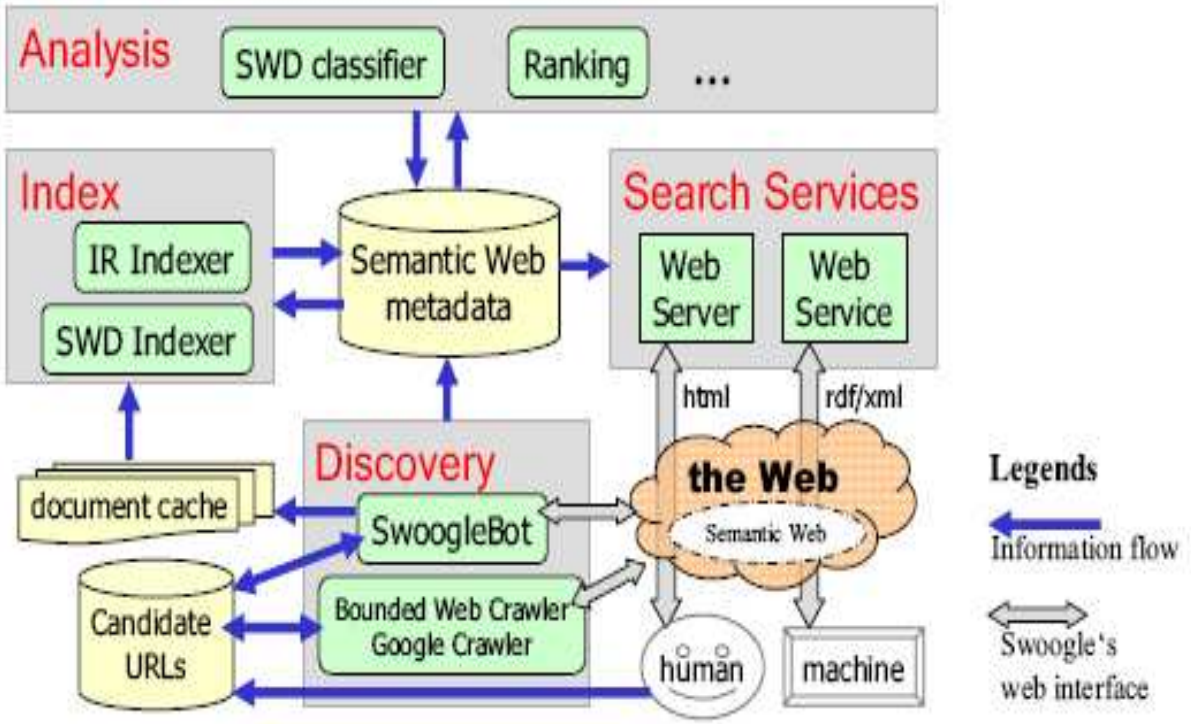
Použití sémantické informace v procesu vyhledávání umožňuje uživatelům získávat vysoce relevantní informace, přičemž při vlastním hledání není vyloučeno dotazování v přirozeném jazyce formou klasických otázek nikoli pouze slovních spojení. Sémantické neboli znalostní vyhledávání tedy není založeno jen na výskytech slova, ale také na jeho významových vztazích.

Rozšířené sémantické funkce postupně do svých algoritmů zapracovávají nejznámější vyhledávače jako *Ask*, *Bing*, *Google* či *Yahoo!* Např. Google zaznamenává hledané výrazy a využívá je následně k lepšímu zacílení zobrazované reklamy. Čistě sémantickému vyhledávači se však ve své veřejné podobě ale v současnosti blíží pouze Bing, který integrací původně samostatného vyhledávacího systému *Powerset* pomohl ke zpřesnění výsledků zejména v oblasti zdravotnictví, díky napojení na obsáhlou ontologii *Medstory*. *DBpedia* je zase kompletní export informací ze známé Wikipedia.org do RDF. Oblast sémantického vyhledávání však byla prozatím spíše pouze předmětem akademického výzkumu než masivnějšího praktického nasazení [57].

Za sémantický vyhledávač bývá někdy mylně považován *Wolfram Alpha*, ale je třeba ho zde zmínit jako zástupce tzv. expertního systému, jehož základem je výpočetní engine, který na základě matematických a statistických modelů vypočítává výsledky. Je založen na sofistikovaném systému Mathematica a pracuje s jazykovou (linguistickou) sémantikou, která však nemá se Sémantickým webem co do činění. Podkladovými daty jsou zde odborné studie a vědecké materiály, které jsou podobně jako u jiných vyhledávačů procházeny crawlery. Spolehlivost informací však může být vyšší než u neodborných textů.

3.6.1 Swoogle

Je jedním z příkladů původně čistě vědeckého projektu sémantického vyhledávání (http://swoogle.umbc.edu/), vyvíjený výzkumníky na Univerzitě v Marylandu (USA), který indexuje RDF či OWL dokumenty a webové stránky s vloženými RDF metadaty (eRDF, RDFa) [2]. Sémantický popis je charakterizován pojmy a vazbami ve specifické oblasti, kterou se dokument zabývá a popisuje tedy vnitřní ontologii indexovaných dokumentů. Výsledky vyhledávání pak řadí podle ranku užitečnosti a oblíbenosti (obdobně jako PageRank od Googlu), ale s tím rozdílem, že v tomto případě je důraz kladen na faktickou hodnotu a relevantnost informací uzpůsobenou pro Sémantický web. V praktické části této práce (viz podkapitola 4.3) byl tento vyhledávač použit jako prostředník pro vyhledání URI v rámci návrhu anotačního procesu.



Obr. 3.22: Schéma funkčních principů webového vyhledávacího stroje Swoogle

Dle představeného funkčního schématu na obrázku 3.22, systém sestává z databáze uložených metadat dokumentů Sémantického webu (SWD), která je vytvářena na základě analýzy (klasifikace a hodnocení), indexace (s využitím technik tvorby extrakčních ontologií) umožňující současné objevování nových dokumentů (crawlery) a jejich přistupování skrze webové vyhledávací služby pomocí REST rozhraní zpracovávajícího HTTP GET dotazy a vrací výsledky ve formě dynamických stránek enkodovaných v RDF/XML syntaxi [45].

- **Objevování** (discovery) – vytváří kolekci URL vhodných pro následné přidání do indexu buďto ručním vložením odkazu na sémantický dokument nebo automaticky pomocí robota, který stránky prochází.

- **Indexování** (indexing) – analyzuje objevené dokumenty a extrahuje z nich metadata, která charakterizují nejen vlastnosti zachycených dokumentů a termů, ale také individuální vazby v dané doméně mezi nimi.

- **Analýza** (analysis) – zapojuje hodnotící mechanismus relevantnosti gererovaných a extrahovaných metadat na základě klasifikace sémantických dokumentů obsažených ve vytvářeném indexu.

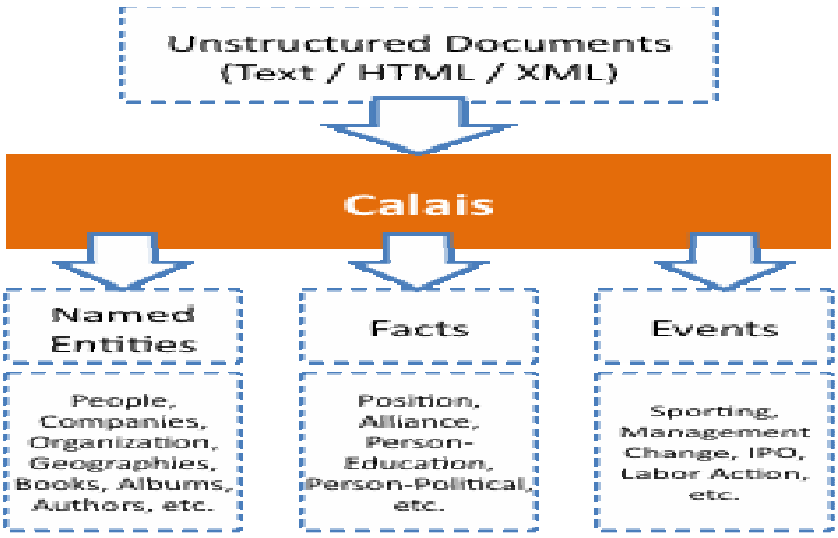
- **Vyhledávání** (search services) – poskytuje přístup k metadatům prostřednictvím webové vyhledávací služby, vlastní zpracování i zvýraznění výsledků či podporu ontologických slovníků na úrovni obsažených termů pro další rozšíření a zpřesnění navigace.

3.6.2 Open Calais

Open Calais (http://opencalais.com/) je dalším nástrojem k prohledávání dokumentů, který je specifický v tom, že hledá na webových stránkách informace v podobě objektů, faktů a událostí, které pak následně exportuje do formátu RDF. Jeho praktické použití je možné vyzkoušet pomocí doplňku pro Firefox – *ClearForest Gnosis*, který v reálném čase při prohlížení stránky podtrhuje pojmy v textu různými barvami a zvýrazňuje tak lidi, organizace, společnosti, produkty a geografické informace. O jeho rostoucí popularitě svědčí i praktické nasazení v rámci mezinárodně uznávané databáze Thomson Reuters, kde plní funkci kontextového syndikátoru.

Zachycená metadata postupně ve strukturách map, grafů a sítí propojují jednotlivé dokumenty napříč doménovými oblastmi. S využitím analýzy v procesu zpracování přirozeného jazyka, strojového učení a dalších metod Open Calais v dokumentech následně dokáže rozpoznat jmenné entity (teoreticky rozebíráno v 3.4.3) a stejně tak i fakta skrytá v jinak nestrukturovaném dokumentu, ze kterého se jeho následným zpracováním a provedením sémantické anotace stává strojově zpracovatelná RDF varianta [23].

Velkým potenciálem celého projektu je jeho otevřenost, kdy použití a dotazování se nad službou OpenCalais v rámci vlastní API je k dispozici zcela zdarma, a to jak pro komerční tak i nekomerční použití. Tento fakt tak může v budoucnu jednak výrazně napomoci nárůstu sémanticky anotovaných dokumentů, jejichž dostatečné množství je základním předpokladem pro funkční myšlenku Sémantického webu v globálním měřítku, ale stejně tak je existence takto široce dostupného nástroje důležitá i pro následné učení a zdokonalování ontologií a doménových oblastí, nad kterými vlastní vyhledávání probíhá.



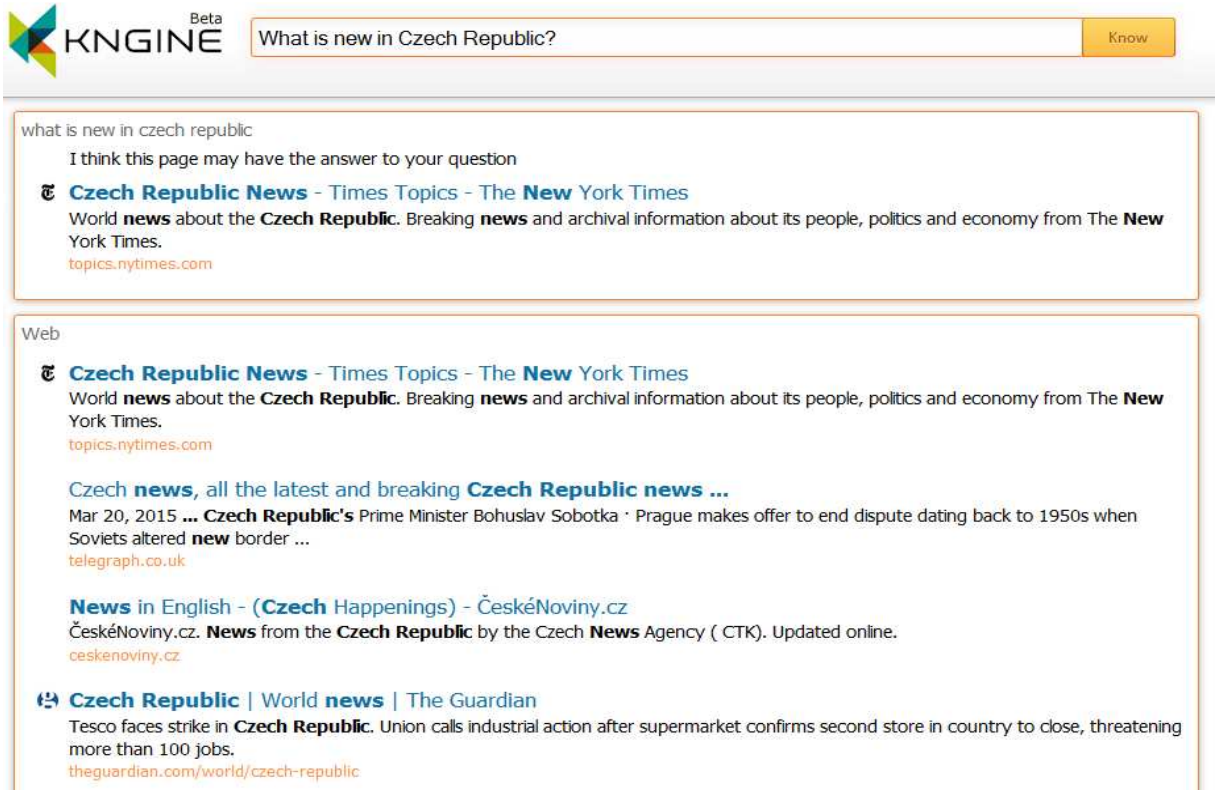
Obr. 3.23: Open Calais prostředníkem procesu zpracování nestrukturovaného obsahu

3.6.3 Kngine

Kngine (http://kngine.com/) je inteligentním asistentem, jež generuje odpovědi na otázky pokládané prostřednictvím přirozeného jazyka. Jeho hlavním cílem je co nejvíce přiblížit formu strojového vyhledávání procesu lidského uvažování. Typický vyhledávací engine se totiž chová jako kniha, která v rejstříku nabízí seznam relevantních pojmů z pokládaného dotazu – samotné pojmy však zpravidla vždy nestačí a je třeba následně požadované informace dále dohledávat, což může někdy představovat značně zdlouhavý proces.

Aplikace na bázi znalostních systémů typu Kngine však umí extrahovat odpovědi na pokládané dotazy přímo z indexovaných stránek a je tak možné získat relevantní výsledky i na tak obecně formulovaný dotaz jako: *„Co je nového v České republice?“*, viz obrázek 3.24. Odpovědí pak může být aktuální přehled zpráv soudobého dění zprostředkovaný renomovanými zpravodajskými servery, navíc doplněný o grafy, mapy, tabulky či multimediální obsah.

Kngine kontinuálně prochází nestrukturované stránky v prostředí stávajícího Webu a automaticky z nich vytěžuje statistické údaje, hypotézy a možné odpovědi, které si pak již ve strukturované formě ukládá do vlastního cache indexu, což v konečném důsledku znamená vyšší efektivitu a rychlost při zpracovávání dotazů oproti jiným sémantickým vyhledávačům, kdy statisticky až 99% všech požadavků je zde zpracováno pod hranicí 3 sekund [35].



Obr. 3.24: Ukázka obecně formulovaného dotazu zpracovaného nástrojem Kngine

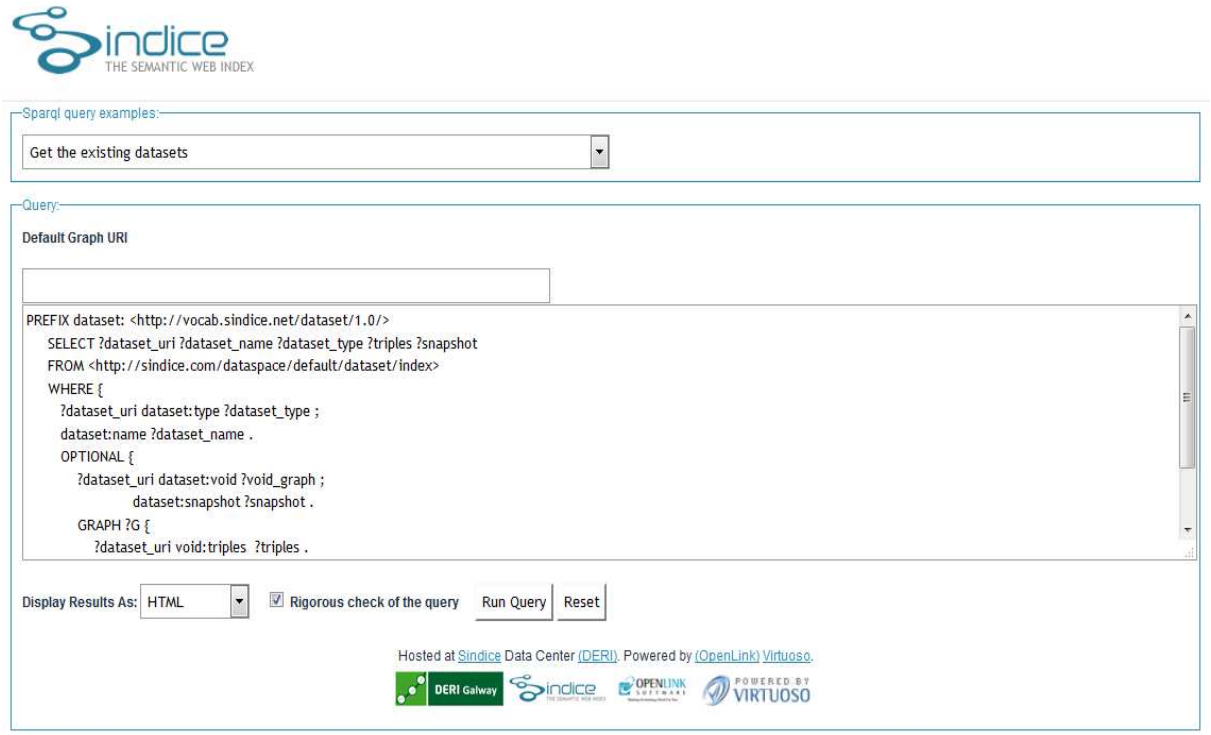
3.6.1 Sindice

Sindice (http://sindice.com/) není front-endovým sémantickým vyhledávačem v pravém slova smyslu, ale umožňuje uživateli rychlé nalezení použitelného RDF zdroje na základě principu reverzního indexu vyhledávání v Sémantickém webu. Jeho myšlenka je založena na představě, že výsledkem hledaného dotazu není klasická množina odkazů s relevantními informacemi, ale dynamické URL adresy obsahující všechna možná místa výskytu zdroje hledaného URI.

Úkolem reverzního indexu je zprostředkování rychlého vyhledávání za cenu dodatečných operací při přidávání nového dokumentu do databáze. Typicky je takový index představován seřazeným souborem významných slov, lemmatizátorem upravených na základní tvar, kde následně ke každému slovu je přiřazen seznam dokumentů se zachyceným výskytem takového slova.

Ačkoli Sindice používá a propaguje model propojených dat (linked data) ve formě URI identifikátorů, které na rozdíl od URL identifikátorů nemohou být dereferencovatelné, zůstává jeho hlavním cílem a současně výhodou oproti jiným nástrojům možnost zachycení těch vazeb na stávající zdroj, které byly vymezeny i mimo autoritativní oblast tohoto zdroje [13].

Nad takto uspořádaným indexem slov ve formě datasetů je pak možné se dotazovat prostřednictvím SPARQL jazyka s možným zobrazením výsledků v nejrůznějších formátech (HTML, XML, JSON, Javascript, NTriples a další).



Obr. 3.25: Příklad SPARQL dotazu v prostředí dotazovacího okna indexu Sindice

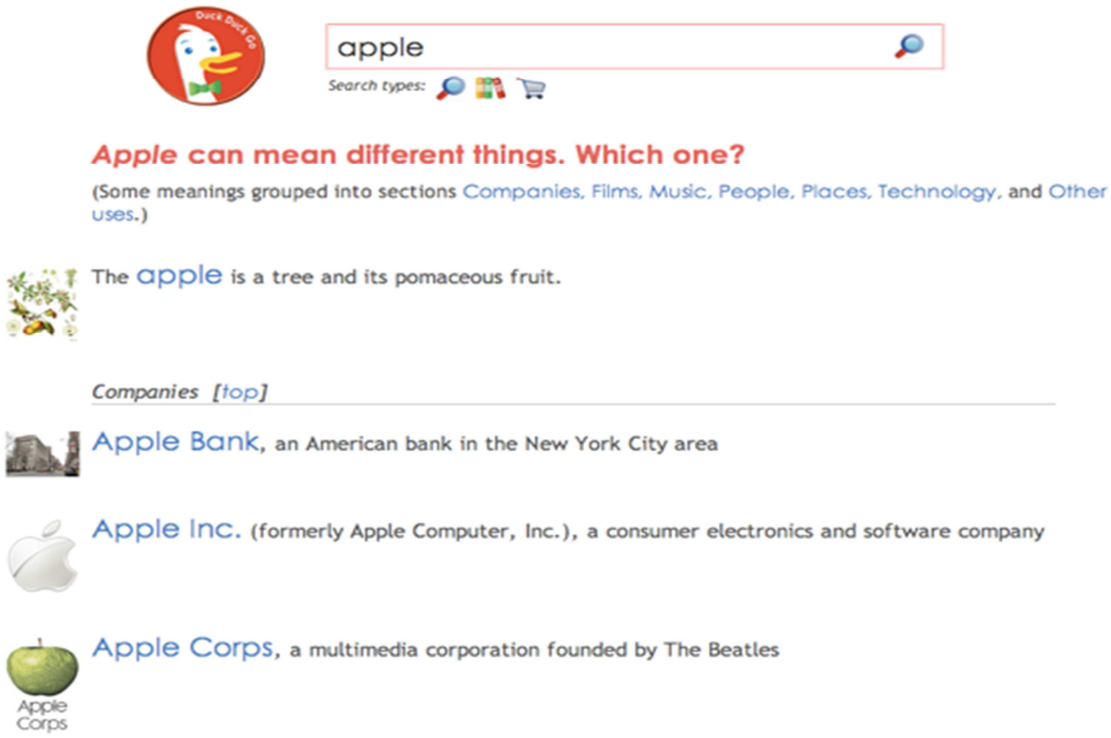
3.6.2 DuckDuckGo

DuckDuckGo (http://duckduckgo.com/) je jedním z vývojově nejmladších a v současné době také jedním z nejprogresivnějších nástrojů, které pro proces vyhledávání znalostí a dat využívají technologií Sémantického webu.

Mezi neustále rostoucí skupinou uživatelů je oblíben zejména z důvodu ochrany soukromí, kdy na rozdíl od Google nesleduje uživatelskou aktivitu, nezobrazuje žádné reklamní bloky ani neprofituje prodejem osobních informací a nedostává tak uživatele do tzv. filtrovací bubliny personalizovaných výsledků vyhledávání. Jako jeden z mála dostupných sémantických vyhledávačů nabízí také rozšíření pro mobilní platformy a do seznamu svých výchozích vyhledávačů jej postupně zařadily také internetové prohlížeče Firefox a Safari.

Výsledky vyhledávání jsou generovány z crowdsourgingových databází, např. Wikipedie, ale také z dostupných indexů partnerských vyhledávačů (v současnosti Yandex, Yahoo!, Bing a Yummly).

Částečně je zde umožněno dotazování formou jednoduchých otázek a obstojně je řešena také problematika mnohoznačnosti slov – *disambiguace* (viz obrázek 3.26), kdy již přímo ve výsledcích je názorně označeno a zdůvodněno, jaké mohou být další možné významy hledaného výrazu.



Obr. 3.26: Řešení slovní disambiguace v prostředí vyhledávače DuckDuckGo

4. DOSAŽENÉ VÝSLEDKY

S využitím aplikovaných znalostí a technologií, popsaných v předchozích kapitolách teoretické části této práce, bude postup při řešení praktické části, s cílem podpory a využití možností Sémantického webu v praxi, následující:

1. Provedení dotazníkového šetření s cílem zmapování informovanosti o existenci koncepce a možnostech praktického využití Sémantického webu s následnou formulací hypotéz dalšího možného vývoje.
2. Vytvoření, propagace a postupné rozšiřování encyklopedické znalostní báze o technologiích Sémantického webu prostřednictvím portálu (http://www.semanticweb.cz/) za účelem poskytnutí uceleného souboru relevantních informací o dané problematice široké veřejnosti.
3. Návrh rozšiřujícího modulu pro vybraný WYSIWYG editor, s možností univerzální integrace do CMS a následnou podporou sémantického anotačního procesu dle standardu RDFa, umožňující uživatelsky snadnou implementaci sémantiky do nových nebo již existujících dokumentů při současném zachování validity výsledného kódu.
4. Zmapování praktických možností pro využití vyhledávání za pomoci počítačově zpracovávaných dotazů v přirozeném jazyce, namísto dosud upřednostňovaného vyhledávání pomocí klíčových slovních spojení, s uvedením výčtu možných záporů i předností takového způsobu práce s dotazováním nad ontologiemi a možnou prognózou budoucího vývoje.
5. Statistické výkonnostní srovnání v současné praxi nejrozšířenějších sémantických reasonerů, schopných inference logických konsekvencí z předložených tvrzení, faktických informací či axiomů.

Jednotlivé body budou postupně podrobněji diskutovány v následujících podkapitolách, vždy s uvedením ukázky praktického výstupu z daného úkolu. V závěrečné kapitole 6 pak bude zhodnocen přínos jednotlivých výsledků pro vědu i praktickou oblast s možným dalším využitím v rámci zkoumané tématiky.

4.1 Provedení dotazníkového šetření

Navzdory skutečnosti, že uplynula již více než dekáda od prvotního vyslovení myšlenky Sémantického webu, později označovaného souhrnně jako Web 3.0, je tato technologie mezi širokou veřejností vnímána stále jako relativně nová a neznámá. Monografické zdroje o dané problematice jsou v současné době dostupné především v cizojazyčné literatuře. V češtině zde existuje sice řada zejména internetových a informačně spíše zevrubně popisných článků, avšak celkově vzato je v rámci dané oblasti stále ještě postrádáno ucelené obšírnější zpracování dostupných informací společně s rozvojem dalšího výzkumu.

4.1.1 Cíle a metodika výzkumu

Hlavní motivací prováděného dotazníkového šetření bylo zjištění aktuálního povědomí o problematice Sémantického webu mezi širokou veřejností v České republice a zkoumání potřeby rozšíření komplexně zpracovaných informačních zdrojů o těchto technologiích a používaných standardech v českém jazyce.

Pro vlastní výzkum byla zvolena forma online dotazníkového formuláře s uzavřenými položkami, přičemž výzkumná část byla zahájena v únoru 2012 a ukončena v dubnu 2014 s následným vyhodnocením výsledků. V teoretické části výzkumu bylo nejprve nutno provést zmapování oblasti, tedy zjištění aktuálního stavu existujících technologií a následné zaměření na některé technologie doposud méně známé či aktuálně rozvíjené.

Po nastudování a kritické rešerši dostupných literární zdrojů, byl na základě teoretického průzkumu sestaven a naprogramován elektronický dotazník, který byl ve své online podobě (http://www.semanticweb.cz/dotaznik) propagován prostřednictvím sociálních sítí, zároveň rozeslán také do několika firem a středních i vysokých škol s IT zaměřením.

Současně byly kontaktovány odborné servery zabývající se informačními technologiemi – minimum z nich však na výzkum odkazovalo, přičemž přístupy na dotazník z těchto odkazujících serverů byly především prvních několik dní po jeho uveřejnění – poté byly odkazující články postupně odsunuty novějšími a další respondenti přicházeli již převážně pouze z akademické a firemní sféry. Osloveni byli také autoři odborných článků o Sémantickém webu. Všichni respondenti mohli zároveň využít nepovinného bloku přímo v dotazníku a rozeslat doporučení k vyplnění na jimi uvedené emailové adresy známých či kolegů z oboru a pomoci tak dalšímu rozšíření celého průzkumu.

4.1.2 Předpoklady

V souvislosti se záměrem na provedení výzkumného šetření byly sestaveny a formulovány následující hypotézy:

1. Sémantický web je technologií veřejnosti spíše neznámou.
2. Nejčastějším zdrojem informací pro zájemce o danou problematiku jsou cizojazyčné literární monografie a odborné vědecké články.
3. Respondenti, kteří problematiku dosud neznají, ji možná začnou využívat.
4. Mezi nejpřínosnější odvětví Sémantického webu bude patřit možnost webového vyhledávání pomocí dotazování v přirozeném jazyce.
5. Sémantický web není doposud hojně prakticky nasazován z důvodu nedostatečného rozšíření souvisejících technologií, resp. chybějících standardů a celkově malému povědomí o jeho praktickém užití.
6. Komplexní zpracování problematiky v českém jazyce bude vítáno.

4.1.3 Vyhodnocení výsledků

V rámci daného časového rámce se probíhajícího výzkumu zúčastnilo celkem 973 respondentů, z toho 134 žen (14%) a 839 mužů (86%). Z demografického hlediska nejvíce respondentů dosahovalo středoškolského (34%) a následně vysokoškolského vzdělání (42%). Převážnou většinu dotazovaných tvořili studenti s délkou praxe v IT do 10 let, přičemž nejčastější věk respondentů byl do 30 let a dále bylo zjištěno, že převažují obyvatelé z obcí nad 10 000 obyvatel.

Výzkum ukázal, že poměrná část respondentů problematiku Sémantického webu zná (74%), přičemž většina z nich se o ní dozvěděla z českého internetového článku (67%), naopak minimum prostřednictvím klasické monografické literatury, ať již české (4%) nebo cizojazyčné (12%).

O tom, zda by nové technologie využili, je přesvědčena nadpoloviční většina respondentů (56%). Důvodem k prozatímnímu nevyužívání možností Sémantického webu je ve většině případů to, že respondenti sami netvoří webové aplikace (47%), nemají nedostatek zdrojů v češtině (34%) a technologie jako taková zatím není příliš rozšířena (19%).

Jako vysoce přínosnou se naopak mezi respondenty jeví myšlenka využití webového vyhledávání prostřednictvím dotazů přirozeného jazyka (43%). Zkušenost s tímto druhem vyhledávání má však pouze zlomek dotazovaných (13%). Drtivá většina všech zúčastněných respondentů (82%) by přitom uvítala komplexní zpracování problematiky Sémantického webu v českém jazyce pro zvýšení povědomí o této stále ještě rozvíjející se technologii.

Z výsledků výzkumu tak vyplývá, že hypotéza č. 1 byla vyvrácena, neboť většina zúčastněných technologii Sémantického webu zná. Zároveň je vyvrácena i hypotéza č. 2, protože nejvíce respondentů se poprvé dozvědělo o existenci Sémantického webu z českého internetového článku.

Naopak hypotéza č. 3 byla potvrzena, protože uživatelé, kteří Sémantický web neznají, zatím skutečně neví, zda jej budou využívat. Splněn byl též předpoklad č. 4, že co by možnou nejpřínosnější technologií Sémantického webu, je lidmi vnímána možnost vyhledávání na webu prostřednictvím přirozeného jazyka a získání relevantní odpovědi na takto pokládané dotazy.

Hypotéza č. 5 však byla výzkumem vyvrácena, neboť technologii nevyužijí především respondenti, kteří netvoří webové aplikace. Závěrečná hypotéza č. 6 pak byla potvrzena, protože zpracování dané problematiky v českém jazyce (odpovídající prvotnímu jazykovému zdroji) je respondenty obecně vítáno.

Následující přehled je souborným shrnutím dat zaznamenaných výzkumným šetřením ve formě doplňkových grafů a tabulek se statistickým vyhodnocením počtu odpovědí jednotlivých respondentů na zvolené možnosti, jež mapují celkové povědomí širší veřenosti o diskutované problematice.

Otázka: *„Setkali jste se někdy s pojmem Sémantický web?“*

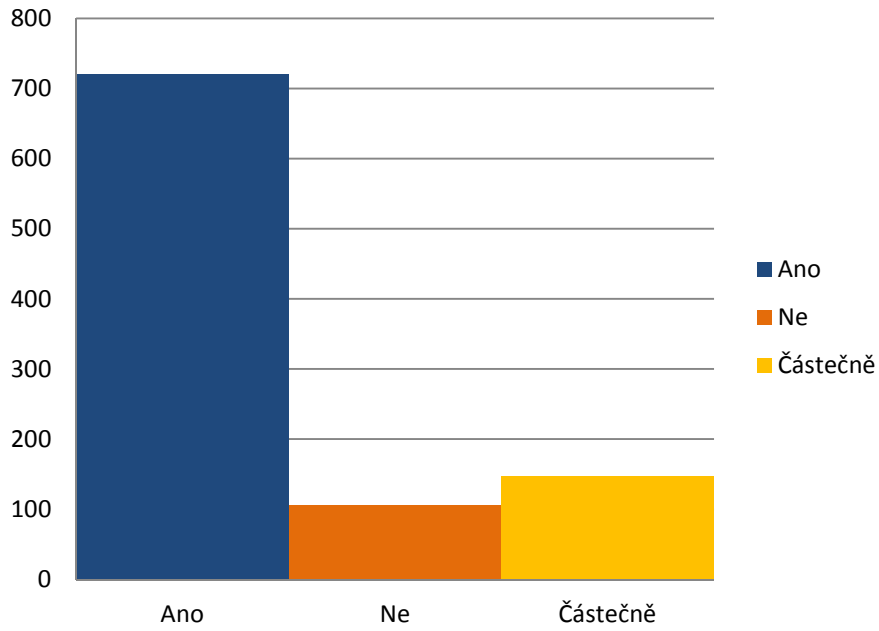
Výsledky jsou odpovídajícím předpokladem reflexe skutečného stavu. Při převažující většině vysokoškolsky vzdělaných respondentů s delší praxi v oblasti IT se všeobecně vyšší míra informovanosti o nových trendech očekává.

Demograficky menší zastoupení žen v této oblasti se však odráží i na celkově menším počtu uváděné znalosti problematiky Sémantického webu.

Uspokojující je počet respondentů s alespoň částečným povědomím o zkoumané oblasti, což svědčí o neustále sílícím vlivu Sémantického webu, který coby současný, avšak současně stále ještě stále více futurologický trend, může sehrát důležitou roli v oblasti nového přístupu při vyhledávání informací a znalostí na Webu, tak jak jej známe dnes, a to celosvětově.

Tabulka 4 – Úvodní rozdělení respondentů

Odpověď	Počet	Z toho muži	Z toho ženy
Ano	720	697	23
Ne	106	39	67
Částečně	147	118	29



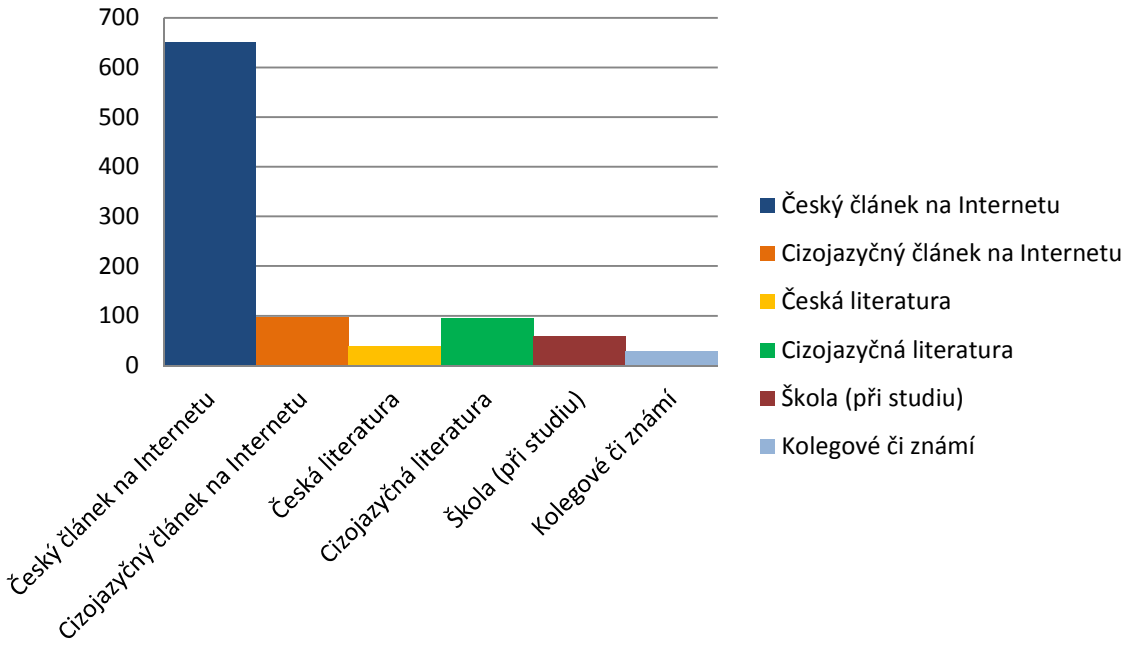
Graf 4.1 – Počty respondentů dle obecného povědomí o zkoumané problematice

Otázka: *„Jak jste se o Sémantickém Webu dozvěděli?“*

Zachycená data zcela jasně ukazují na fakt, že nejčastějším zdrojem pro získávání prvotních informací mezi dotazované respondenty byly česky psané internetové články a články na Internetu obecně. Po prostudování a kritické rešerši těchto česky psaných zdrojů však bylo zjištěno, že se jedná zpravidla vždy pouze o krátké obecně informativní, jejichž primárním cílem je informovat širší veřejnost o tom, že nějaká oblast Sémantického webu existuje, nepopisují však již následně detailněji jeho jednotlivé technologie a vazby na již existující standardy napříč vrstvami sémantického spektra.

Tabulka 5 – Počty respondentů v závislosti na informačních zdrojích

Odpověď	Počet
Český článek na Internetu	652
Cizojazyčný článek na Internetu	97
Česká literatura	39
Cizojazyčná literatura	96
Škola (při studiu)	60
Od kolegů či známých	29



Graf 4.2 - Prvotní informační zdroje o problematice Sémantického webu

Je zde tedy sice fakticky patrná existence řady dostupných textů popisujících zevrubně danou problematiku, je však jen velmi málo z nich, které by na celou věc nahlížely detajlněji a byly současně k dispozici širší veřejnosti v českém jazyce, s čímž, jak se později ukázalo, souvisí i následující dotaz:

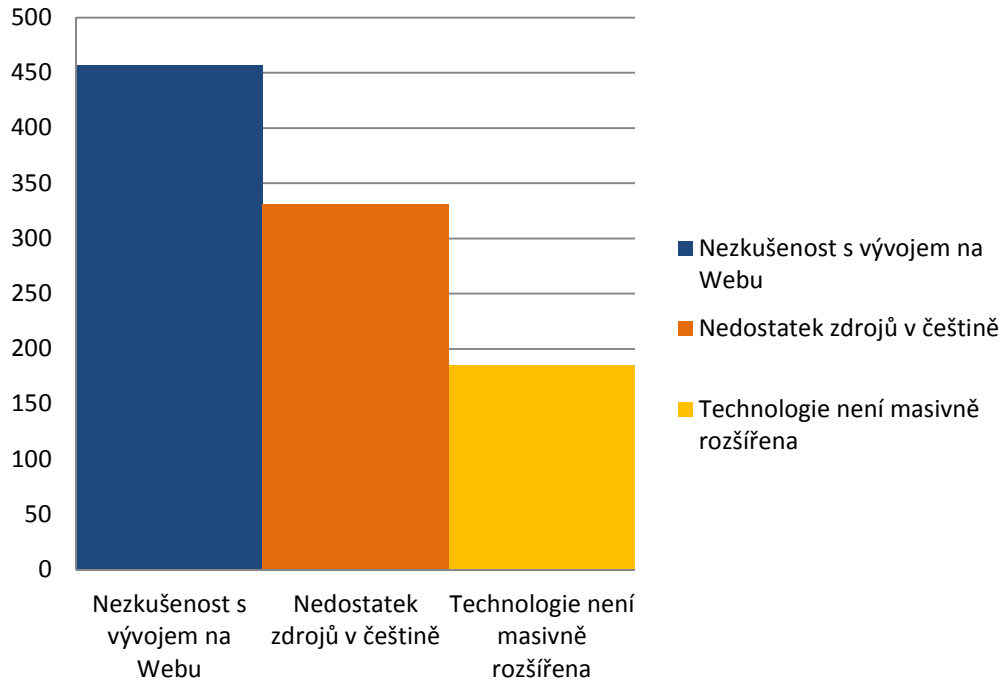
Otázka: *„Z jakého důvodu byste možností Sémantického webu nevyužili?“*

Nedostatek dostupných zdrojů v mateřském jazyce by byl pro ty respondenty, kteří se aktivně začínají zabývat vývojem webových aplikací tou nejproblematičtější překážkou, což může být pochopitelné, neboť sami pokud nemají přístup k cizojazyčné literatuře a neporozumí jí, nedokáží pak ani bez detailnějšího vhledu do dané oblasti pochopit poměrně komplexní myšlenku Sémantického webu jako takovou a do následného vývoje nových aplikací pro tuto platformu se tak v praxi mnohdy ani nepustí.

Svou roli v tom může sehrát i fakt, že menší část dotazovaných respondentů si myslí, že Sémantický web, prozatím není natolik hojně rozšířenou technologií, aby stálo za to s ní počítat při otrém vývoji. Neustále rostoucí počet v praxi nasazovaných aplikací v posledních letech však svědčí o opaku a je zřejmé, že i do budoucna se se Sémantickým webem počítá, jako s rozšířením Webu stávajícího, tak jak to ostatně predikoval i sám jeho průkopník Tim Berners-Lee.

Tabulka 6 – Překážky rozvoje Sémantického webu v praxi

Odpověď	Počet
Nezkušenost s vývojem webových aplikací	457
Nedostatek zdrojů v češtině	331
Technologie není rozšířena	185



Graf 4.3 – Možná omezení aktivního vývoje Sémantického webu

Otázka: *„Co řadíte mezi hlavní přínos technologií Sémantického webu?“*

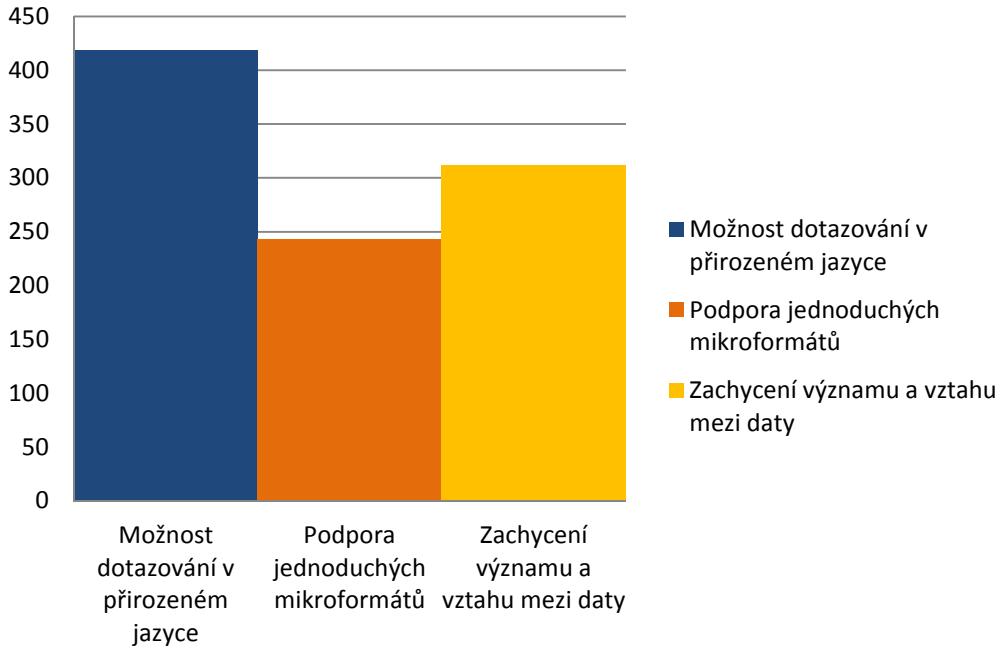
Dle dotazovaných respondentů se jako bezpochyby nejzajímavější možností využití Sémantického webu jeví zachycení souvislostí s možným odvozením nových znalostí mezi strukturovanými daty na webu a technologiemi pro zpracování přirozeného jazyka, což lze z větší míry zdůvodnit jistou atraktivitou tohoto přístupu k vyhledávání známého dosud spíše ze sci-fi filmů a je tedy na pohled zřejmé, že v tomto ohledu je veřejnost plna očekávání věcí budoucích.

I přesto, že jsou mezi těmito oblastmi mnohostranné vazby, uplatnění lingvistických technik v rámci sémantického webu a linked data mají zatím spíše ad hoc charakter, a doporučené postupy teprve vznikají. Na rozdíl od Mikroformátů, které již několik let plní úlohu jednoduché a spolehlivé sémantické anotace informací a uživatelé si na jejich použití pomalu zvykají.

Ostatně zájem „koncových“, neprofesionálních uživatelů vystavovat svá data, která sami vytvoří a zpracují, se výrazněji projevil již před časem s nástupem aplikací „webu 2.0“ (sociální sítě, sdílená videa, fotografie atd.).

Tabulka 7 – Praktický přínos technologií Sémantického webu

Odpověď	Počet
Možnost dotazování v přirozeném jazyce	418
Podpora jednoduchých mikroformátů	243
Zachycení významu a vztahu mezi daty	312

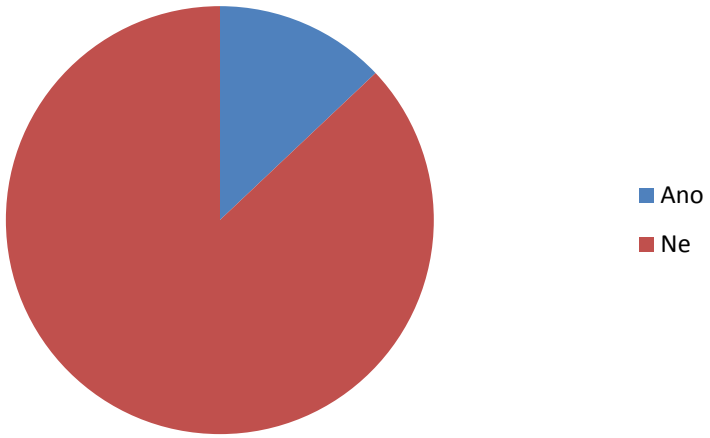


Graf 4.4 - Hlavní výhody Sémantického webu dle dotazovaných respondentů

Otázka: *„Máte osobní zkušenost s vyhledáváním na Sémantickém webu?“*

Tabulka 8 – Praktické využití sémantických vyhledávačů

Odpověď	Počet	Z toho muži	Z toho ženy
Ano	126	102	24
Ne	847	745	102



Graf 4.5 – Zkušenost se sémantickým vyhledáváním informací

Jelikož sémantické vyhledávání na Webu je disciplínou stále poměrně novou a taktéž samotné sémantické vyhledávače dosud nedosahují takových kvalit, které by je předurčily k masivnějšímu rozšíření, zcela oprávněně pak lze potvrdit výsledky odpovědi na tento dotaz. Většina dotazovaných sémantické vyhledávání v praxi nikdy nepoužila a nepoužívá. Menšina z nich tak činí pouze ze zvědavosti nebo v rámci vlastního pokusného testování.

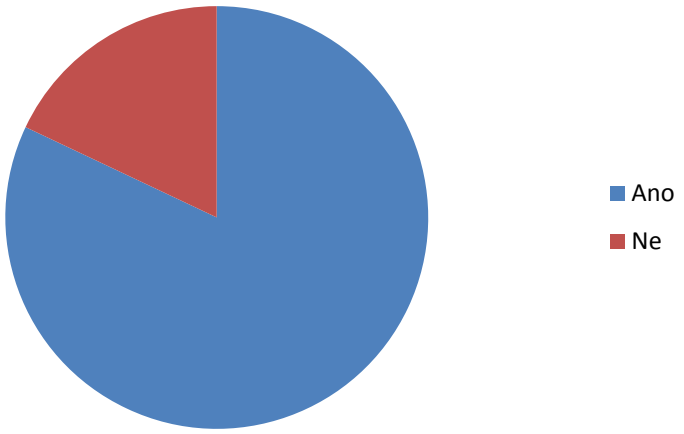
Zažité zvyklosti současných uživatelů Internetu pak dozajista mohou také hrát v celé věci svou roli. Hledání informací prostřednictvím webového vyhledávacího stroje s pomocí klíčových slov se využívá odnepaměti. Sémantický web však nabízí nejen vyhledání informací (identifikaci relevantních dokumentů a jejich relevantní řazení), ale také jednoduché i komplexní odpovědi na obecně formulované otázky v přirozeném jazyce (např. *„Co je nového ve větě?“*, *„Kdo byl českým prezidentem před 10 lety?“* apod.).

Je však zřejmé, že ne všem uživatelům by mohl v budoucnu tento představený způsob pokládání dotazů vyhovovat. Každá zásadní změna v koncepci hledání na Internetu totiž s sebou vždy nesla také větší či menší vlnu nevole a nespokojených reakcí. Je tedy otázkou, zda by časem měl být Sémantický web prosazován cestou evoluční či revoluční. Jako celkově přijatelnější se může jevit kompromis inklinující spíše k první variantě, ovšem za předpokladu dostatečné informovanosti široké veřejnosti současně s dostupnými informačními zdroji.

Otázka: *„Uvítali byste širší zpracování této problematiky v českém jazyce?“*

Tabulka 9 – Potřeba komplexního zpracování informací

Odpověď	Počet	Z toho muži	Z toho ženy
Ano	798	674	124
Ne	175	167	8



Graf 4.6 – Zpracování souhrnného přehledu v českém jazyce

S ohledem na již dříve zjištěné skutečnosti o nedostatečném množství komplexně zpracovaných informací v oblasti Sémantického webu v českém jazyce, se myšlenka na souborné sepsání základního přehledu nejpoužívanějších technologií a postupů při jeho tvorbě jeví obecně jako velmi přínosná.

Pro nové zájemce, jenž s problematikou Sémantického webu přijdou v současnosti do kontaktu, je totiž zpravidla jen velmi obtížně rozpoznatelné, co je skutečným ustáleným jádrem této oblasti a co již pouze okrajové či naopak zcela nové téma. Pro nezainteresovaného člověka může být taktéž jen velmi nesnadné usoudit, zda celý obor směřuje spíše směrem k masivnímu praktickému uplatnění či se naopak stahuje do zákoutí akademických laboratoří.

Budou-li však dále v budoucnu vznikat názorně a čtivě sepsané texty v českém jazyce, obsahující vysvětlení základní podstaty praktického uplatnění stěžejních pilířů celé oblasti, dle zavedené struktury zmíněné v úvodní části technologií Sémantického webu (obrázek 3.1), může to výrazně dopomoci lokálnímu rozšíření celého konceptu nejen mezi naší vědeckou komunitou, ale postupně také i širokou veřejností, což je ostatně také hlavním cílem teoretické části předkládané disertační práce, a sice podělit se o tyto zkušenosti a povzbudit tak nové zájemce (ať už z řad studentů či akademických pracovníků) k tomu, aby se aktivně zapojili zejména do těch směrů Sémantického webu, u kterých je předpoklad masivnějšího rozvoje v následujících letech.

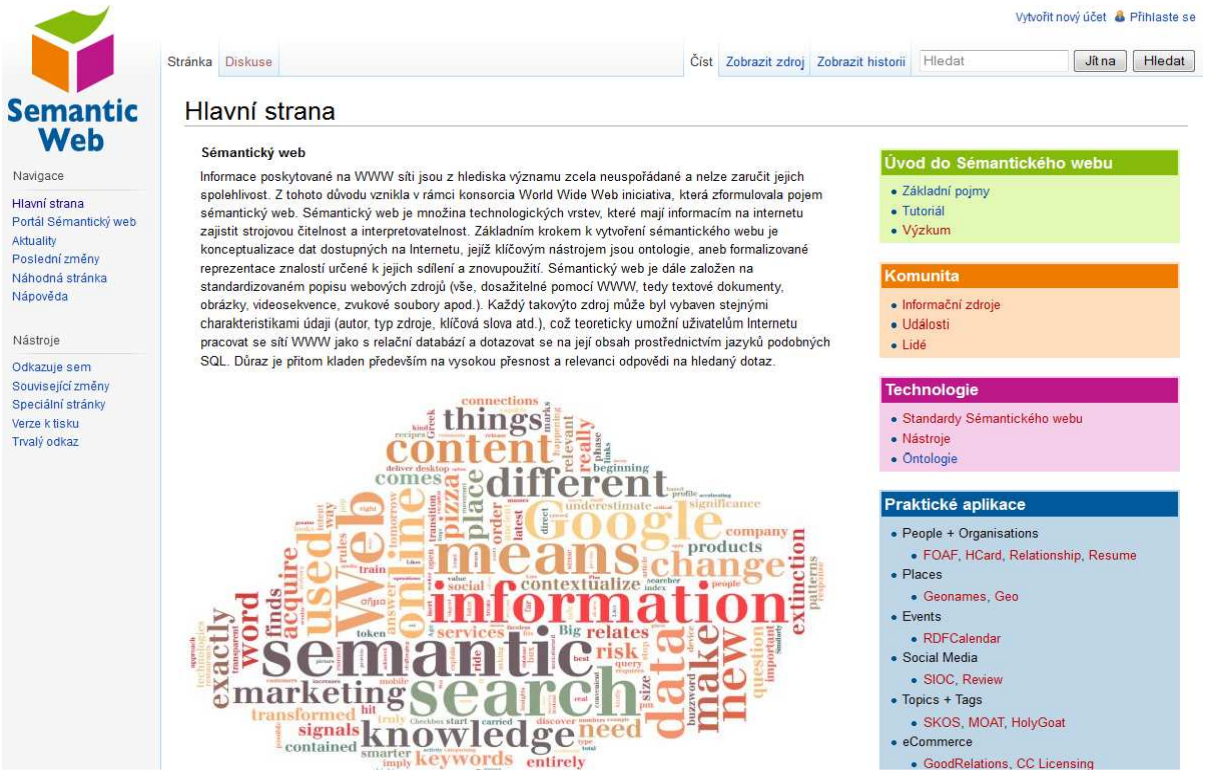
4.2 Vytvoření informačního portálu pro podporu výzkumu

V návaznosti na vyhodnocené výsledky provedeného dotazníkového šetření je zřejmé, že potřeba rozšíření povědomí o Sémantickém webu je aktuálním požadavkem současné doby, a to jak mezi odbornou, tak i laickou veřejností.

V této souvislosti byl v rámci praktické části disertační práce vytvořen informační web obsahující přehled odborných termínů, pojmů a zkratek s českým výkladem. K dispozici je také výukový tutoriál a sada testovacích otázek pro závěrečné ověření znalostí dané problematiky.

4.2.1 Uživatelsky rozšiřitelná encyklopedie

Při volbě formy prezentace informací o Sémantickém webu, se právě možnosti formátu otevřené encyklopedie ukázaly jako nejvhodnější. Vzhledem k neustálému rozvoji této stále ještě poměrně mladé vědní oblasti, je třeba nové informace mezi komunitou sdílet rychle a efektivně, nezávisle na jediném autorovi, avšak současně se zachováním určitého řádu pro následná ověření vkládaných dat, kdy každý takto nově přidaný záznam podléhá schválení ze strany redaktora či administrátora daného serveru. Tento ověřený koncept uživatelsky rozšiřitelné informační báze (inspirovaný celosvětově rozšířenou encyklopedií Wikipedia.org) se stal základem také pro vlastní informační portál: **SemanticWeb** – (http://www.semanticweb.cz/)



Obr. 4.1: Hlavní strana informačního portálu SemanticWeb.cz

4.2.2 Edukační tutoriál

V rámci komplexního zpracování dané problematiky, je součástí otevřené encyklopedie také přehledný tutoriál (http://www.semanticweb.cz/tutorial), poskytující přehled základní terminologie a vývojové linie Sémantického webu od jeho historických počátků po současnost s uvedením výčtu nejčastěji používaných standardizovaných formátů, typických pro danou oblast.

Jedná se tak ve své podstatě o rozšíření teoretické části této disertační práce, přičemž celý koncept edukačního tutoriálu je postaven na snaze o širší přiblížení této zájmově vyhledávané problematiky populární formou.

Teoretický výklad je doplněn o sekci dostupných softwarových nástrojů pro tvorbu a správu Sémantického webu s uvedenými odkazy na stažení, v případě volné licence pro další šíření. V návaznosti na dostupné programové prostředky je zde dále uvedeno i několik vizualizačních schémat ontologií a ukázkové příklady s částí kódu vytvořené prostřednictvím nástroje Protégé 4.3 [24].

To, že zájem veřejnosti o problematiku Sémantického webu v současné době vykazuje vzestupnou tendenci, dokazuje také provedené výzkumné šetření (diskutované v podkapitole 4.1.3). Lze tedy předpokládat, že význam podobných souhrnně-teoretických přehledů bude následně v širších kruzích nezanedbatelným praktickým přínosem pro další rozvoj v této oblasti.

4.2.3 Znalostní testy

Navazují na obsah teoretického tutoriálu a jsou rozděleny do samostatných bloků testovacích otázek, které následují vždy na konci rámcové části výkladu. Realizovaná online verze (http://www.semanticweb.cz/testy) umožňuje okamžité vyhodnocení po ukončení každého testu a současně také porovnání procentuální úspěšnosti daného pokusu s výsledky předchozích účastníků.

S využitím dostupných softwarových nástrojů pro export dat na bázi maker (např. Word Quiz Template) lze navíc z původní offline verze v textovém editoru snadno připravit import všech testových otázek pro použití v některém ze známých výukových systémů typu Moodle, což implikuje další možné využití v edukační oblasti při průběžném ověřování znalostí.

Společně s teoretickým tutoriálem lze pak všechny tyto materiály souhrnně využít jako podklady k vytvoření či možného rozšíření kurzu pro výuku specializovaného předmětu v rámci studijního programu Inženýrská informatika, zaměřeného na technologie vývoje Sémantického webu a práci s ontologiemi.

4.3 Návrh a implementace softwarového anotačního nástroje

Za účelem zajištění podpory masivnějšího rozvoje stávajícího Sémantického webu formou implementace sémantiky do nových či již existujících dokumentů, je zapotřebí běžnému uživateli poskytnout jednoduchý, přitom však dané věci plně dostačující prostředek pro správu webového obsahu.

Výchozím bodem při realizaci softwarového nástroje pro pokročilou sémantickou anotaci bude implementace zásuvného modulu pro WYSIWYG softwarovou komponentu CKEditor, jejíž kompletní dokumentaci je možno nalézt na webových stránkách (http://docs.ckresource.com/).

Aplikační platformou komponenty bude na straně serveru Microsoft .NET Framework 3.5 a programovací jazyk C#, na straně klienta pak programovací jazyk Javascript, ve kterém je implementována i vlastní komponenta CKEditor. Vyžadováno bude připojení k Internetu na straně klienta a komponenta využije externí knihovnu Spring.Rest jako REST klienta.

4.3.1 Specifikace požadavků

Základní požadavky na funkčnost komponenty jsou stanoveny následovně:

1. Uživateli je umožněno vytvoření nové a editace i odstranění již stávající RDFa anotace pomocí kontextové nabídky nad označeným textem.
2. Uživateli je umožněno dohledání webových zdrojů, jejichž URI nezná (jako prevence duplicitních tvrzení).
3. Komponenta bude reprezentována zásuvným modulem pro volně dostupný WYSIWYG editační nástroj CKEditor
4. Mělo by se jednat o univerzální řešení, integrovatelné do jakéhokoliv CMS na bázi .NET Framework s CKEditorem a vlastní integrace komponenty by měla být pro vývojáře snadná.
5. Výstupem komponenty musí být validní XHTML.
6. Koncepce anotace by měla odpovídat standardům specifikace RDFa
7. Pro dohledávání výsledků pro zdrojová data URI bude použito předávaných výsledků sémantického vyhledávače Swoogle.
8. Měla by být zajištěna kompatibilita posledních verzí nejpoužívanějších webových prohlížečů (Internet Explorer, Mozilla Firefox, Chrome).
9. Uživatelsky použitelné a přívětivé ovládací rozhraní.

Provozní podmínky se omezují pouze na běhové prostředí webového prohlížeče, framework Microsoft .NET verze 3.5 a funkční připojení k Internetu. Cílovým uživatelem výsledné aplikace může být každý, kdo má v rámci hostitelského CMS oprávnění k vytváření obsahu. Protože se však jedná o pokročilou metodu tvorby a editace obsahu, měl by zainteresovaný uživatel mít alespoň základní povědomí o problematice RDFa.

4.3.2 Funkční programové atributy

Po načtení stránky s CKEditorem v prohlížeči se provede inicializace. Následně je možná uživatelská úprava obsahu, na níž zásuvný modul nijak nereaguje, naopak interakce je vyvolána pomocí kontextové nabídky k anotování či editování vybraného textu, kdy se registrují následující příkazy a jejich obslužné rutiny ve formě anonymních funkcí:

1. *addSubject*
2. *addRelation*
3. *addProperty*
4. *editSubject*
5. *editRelation*
6. *editProperty*
7. *removeAnotation*

Příkaz *addSubject* odpovídající anotaci samostatného subjektu vychází z konstruktu komponenty pro zjednodušení práce se zdroji v daném dokumentu a slouží také jako prevence duplicit.

Pomocí API *editor.getSelection().getSelectedText()* se získá uživatelem označený výběr ve formě textového uzlu typu *CKEDITOR.dom.element*.
Výsledná selekce se poté uzamkne, aby nedošlo k její změně v průběhu anotace.

Nakonec dojde k otevření dialogového okna s anotačním formulářem, kde je možno vyplnit jak relaci, tak vlastnost daného subjektu, aby následně došlo k vytvoření kompletní sémantické trojice {*subject*, *relation*, *property*}.

Příkazy *addRelation* a *addProperty* pracují na podobném principu jako výše zmíněný *addSubject*, rozdíl je pouze v query-string parametrech, které zohledňuje anotační formulář a se kterými se následně otevírá stránka v dialogovém okně. Přípustné parametry jsou: *mode*, k indikaci typu anotace s hodnotou v množině {*subject*, *relation*, *property*} a *action* z množiny {*fnew*, *editg*} identifikující, zda se jedná o novou anotaci či úpravu již stávající.

Příkazem *editRelation*, jež vychází z kódu funkce *addRelation*, se před otevřením dialogového okna provede navíc parsování XHTML z označeného textu či elementu, aby došlo k naplnění anotačního objektu, který pak následně zohledňuje dialog při svém načtení.

A konečně příkazem *removeAnotation* se namísto otevření dialogového okna rovnou odstraní uzly vložení při anotaci a nahradí je původním textovým uzlem, při čemž je zohledněn typ anotace, logika se pro jednotlivé typy mírně liší.

4.3.3 Analýza a návrh anotačního procesu

Načítaná stránka prochází v průběhu svého zpracování několika dílčími stavy *(Page Request, Start, Initialization, Load, Postback handling, Rendering, Unload)*. Při každé z těchto fází pak vyvolává relevantní události *(PreInit, Init, InitComplete, PreLoad, Load, LoadComplete, PreRender, PreRenderComplete, SaveStateComplete, Render, Unload)*, jimž mohou naslouchat handlery dalších komponent ve stránce a připojovat vlastní logiku.

Stromová struktura je dána kolekcí controls následníků, vycházející z terminologie ASP.NET:

- *Control* ovládací prvek gererující HTML kód do stránky, přičemž každý takový ovládací prvek dědí z třídy System.Web.UI.Control.
- *User-control* je uživatelsky ovládaný prvek, složený z několika dalších ovládacích prvků, které zapouzdřuje – typické je, že v rámci user-controlu se neřeší samotné generování HTML kódu.
- *Server-control* musí naopak přetěžovat metodu Render, v níž se logika generování HTML kódu nachází.

Dialogové okno, jež je součástí ASPX stránky obsahuje user-control, reprezentovaný anotačním formulářem, který může být vložen na jakoukoliv ASPX stránku, kterou lze ovlivnit hostitelským systémem a vhodným nastavením a definováním CSS stylů také snadno integrovat do stávajícího prostředí vybraného WYSIWYG editoru.

Životní cyklus stránky

Při načtení je vykonán skript, který provede asynchronní odeslání formuláře (postback), čemuž předchází vyzískání seznamu všech subjektů, které se uvnitř editoru vyskytují. K objektu CKEditoru se přistoupí prostřednictvím DOM, tzn.

window.opener.CKEDITOR.instances[instanceName]

kde instance Name představuje jméno instance CKEditoru ukládané v anotačním objektu. Za pomocí jQuery kódu se současně získají všechny uzly, které mají neprázdný atribut about. Výsledný seznam hodnot daných atributů se pak následně uloží do skrytého formulářového prvku ve stránce a teprve poté je proveden asynchronní požadavek. Smyslem celého procesu je naplnění drop-down listu hodnotami, ze kterých pak uživatel vybírá subjekt, umístěný v konkrétním dokumentu. Důvodem použití tohoto řešení byl tzv. viewstate validation mechanismus, který zabezpečuje aplikaci proti podvržení neplatných hodnot do formulářových prvků, které například nemají povolenou editaci, nebo drop-down listů k zamezení možného podvržení položek.

V případě, že by se drop-down list plnil javascriptem při načtení stránky bez postbacku, by byla při ukládání změn vyvolána výjimka ze strany .NETu. Viewstate validaci je možné vypnout, ale pouze na úrovni stránky nebo aplikace, nikoliv user controlu.

Při editaci záznamu se pak ještě před asynchronním odesláním navíc mimo výše zmíněné, provádí také akce, při níž jsou hodnoty získané z atributu anotačního objektu (window.annotationInstance.result) serializovány (funkce SerializeResult()) do formátu:

$$n_i; h_i \,|n_i+1; h_i+1|\ldots|n_n; h_n$$

kde n_i je název *i*-tého atributu a h_i je hodnota *i*-tého atributu, a následně nastaveny jako hodnota skrytého formulářového prvku. Ve chvíli zpracování na straně serveru, je hodnota tohoto prvku deserializována a relevantní formulářové prvky jsou odpovídajícím způsobem přednastaveny.

Uložením formuláře, dojde k odeslání hodnot prvků na server, kde se následně provede validace oproti přípustným hodnotám. Stiskem potvrzovacího tlačítka se vyvolá událost FormSaved, na kterou je navázán handler zajišťující uložení dat. Tato koncepce byla zvolena kvůli možnosti navázání dodatečné funkcionality z jiných komponent na akci při stisknutí tlačítka. Např.:

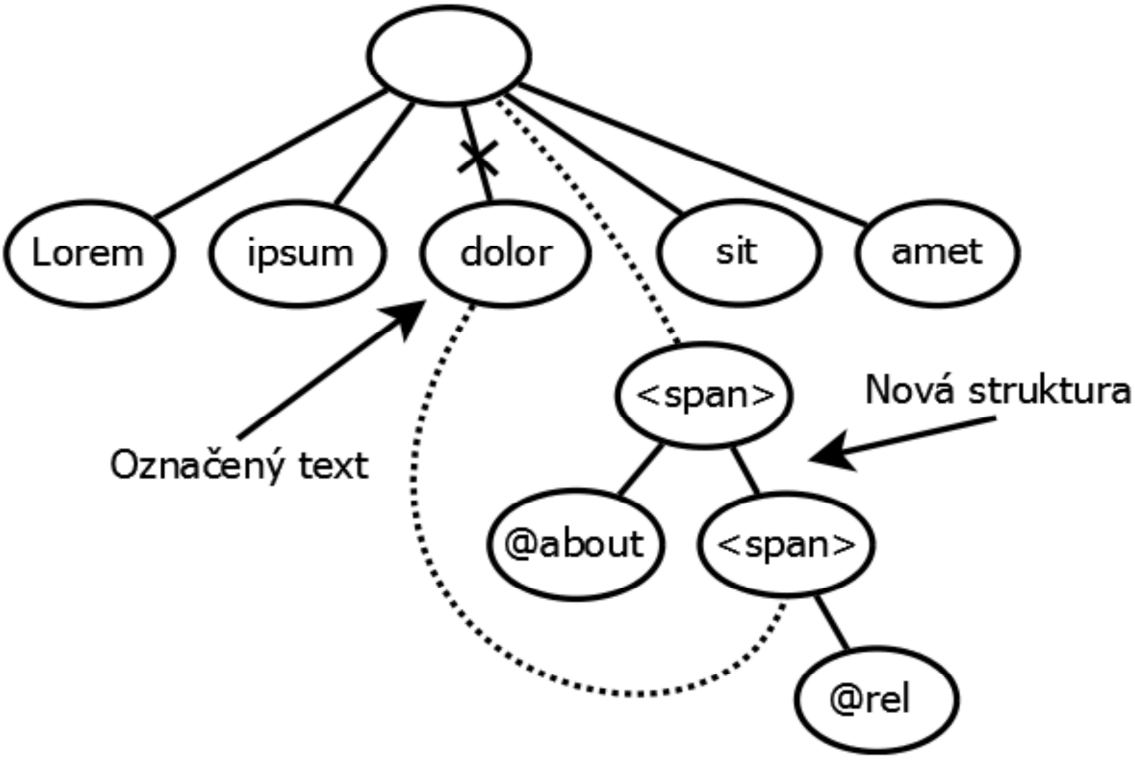
```
ucAnnotationForm.OnFormSaved
                    +=
                        System.EventHandler(customCallbackFunction);
```

Formulování výchozího handleru je definováno javascriptovým výrazem ve formátu objektového literálu, složených z hodnot formulářových prvků, jenž je pak nastaven jako vlastnost *result* anotačního objektu a zajištuje tak, aby došlo ke spuštění na straně klienta a následnému zavolání funkce *SaveAnnotation()* v případě nové anotace, resp. *EditAnnotation()* pro editaci již existující.

Obecně lze říci, že funkce *SaveAnnotation()* zohledňuje tyto typy anotace:

Subjekt – fyzická extrakce textového uzlu z DOMu, vytvoření uzlu nového s nastavením id atributu, přičemž původní textový uzel se stává potomkem nově vytvořeného a celá struktura je následně vložena zpět do místa vyjmutí.

Vazba – na místo vybraného textového uzlu je vložena struktura nová, tvořená elementem *span* a atributem *about*, jehož potomkem je rovněž *span* element s nastaveným atributem *rel* podle orientace vazby a původním textovým uzlem jakožto potomkem s uvedením jeho zdroje (src, href, resource). Modifikace DOM stromu je zobrazena na schématickém obrázku 4.2.



Obr. 4.2: Modifikace struktury DOM stromu

Vlastnost – podobně jako v předchozím případě je původní označený text extrahován a na jeho místo vložena struktura sestávající z nového *span* elementu a atributu *about*, identifikující subjekt prostřednictvím následnického span uzlu s atributem property a k němu připojeným literálovým textovým uzlem, obsahující původní text.

Logika funkce *EditAnnotation()* zohledňuje navíc také typ anotace. Ve všech případech platí, že nejsou vytvářeny nové uzly, nýbrž pouze modifikovány již stávající. Z uzamknuté selekce se získá právě vybraný uzel, tj. text včetně obalujícího *span* uzlu, čímž lze modifikovat uzel udržující informaci o vazbě či vlastnosti. Pro editaci subjektu je potřeba získat uzel s *about* atributem, který je předkem aktuálně zvoleného uzlu.

Po provedení jedné z uvedených anotačních funkcí (pro vytvoření či editaci) se dialogové okno zavře. Autor při anotaci může navíc také využít komponenty pro dohledání identifikátoru zdroje URI. Pole v anotačním formuláři jsou vybaveny dohledávacím tlačítkem, jehož stisknutí otevírá nové dialogové okno. Původní obsah vybraného textu je předán do vyhledávacího pole v dialogu a následně je odeslán dotaz na sémantický vyhledávač.

Celá funkcionalita je zahrnuta do uživatelského ovládacího prvku (user-control), stejně jako v případě anotačního formuláře, z důvodu zachování možnosti přizpůsobit vzhled i obsah vyhledávací stránky.

Vyhledání URI pomocí Swoogle

Sémantický vyhledávač Swoogle podporuje na dvě desítky variabilních typů dotazu, z nichž pro účely této práce postačuje zmínit tyto tři: sémantické webové dokumenty, které jsou klasifikované jako ontologie; všechny sémantické webové dokumenty; a sémantické webové termy. Výchozí je nastavení pro hledání sémantických webových termů.

Vlastní funkcionalitu pro zprostředkování vyhledávání zapouzdřuje metoda *SearchWithSwoogle*, která vrací výsledky hledání sémantického vyhledávače, jež se poté předávají zpět na klientskou stranu ve formě popisného RDF, ve kterém jsou výsledky a jejich popis anotovány pomocí speciálních ontologií. Generovaná XML data jsou parsována prostřednictvím XPath výrazů, které se mohou lišit v závislosti na upřednostněné variantě hledání (např. hledání v sémantických termech představuje výraz *//wob:SemanticWebTerm*).

Tímto se získají uzly obsahující konkrétní URI zdroje. Metoda vrací datovou strukturu *System.Data.DataSet* a takto získaná data jsou poté dále obecně použitelná, čímž tento celkový koncept oddělení vyhledávání od zobrazování výsledku poskytuje CMS vývojáři prostor pro tzv. bezešvou integraci.

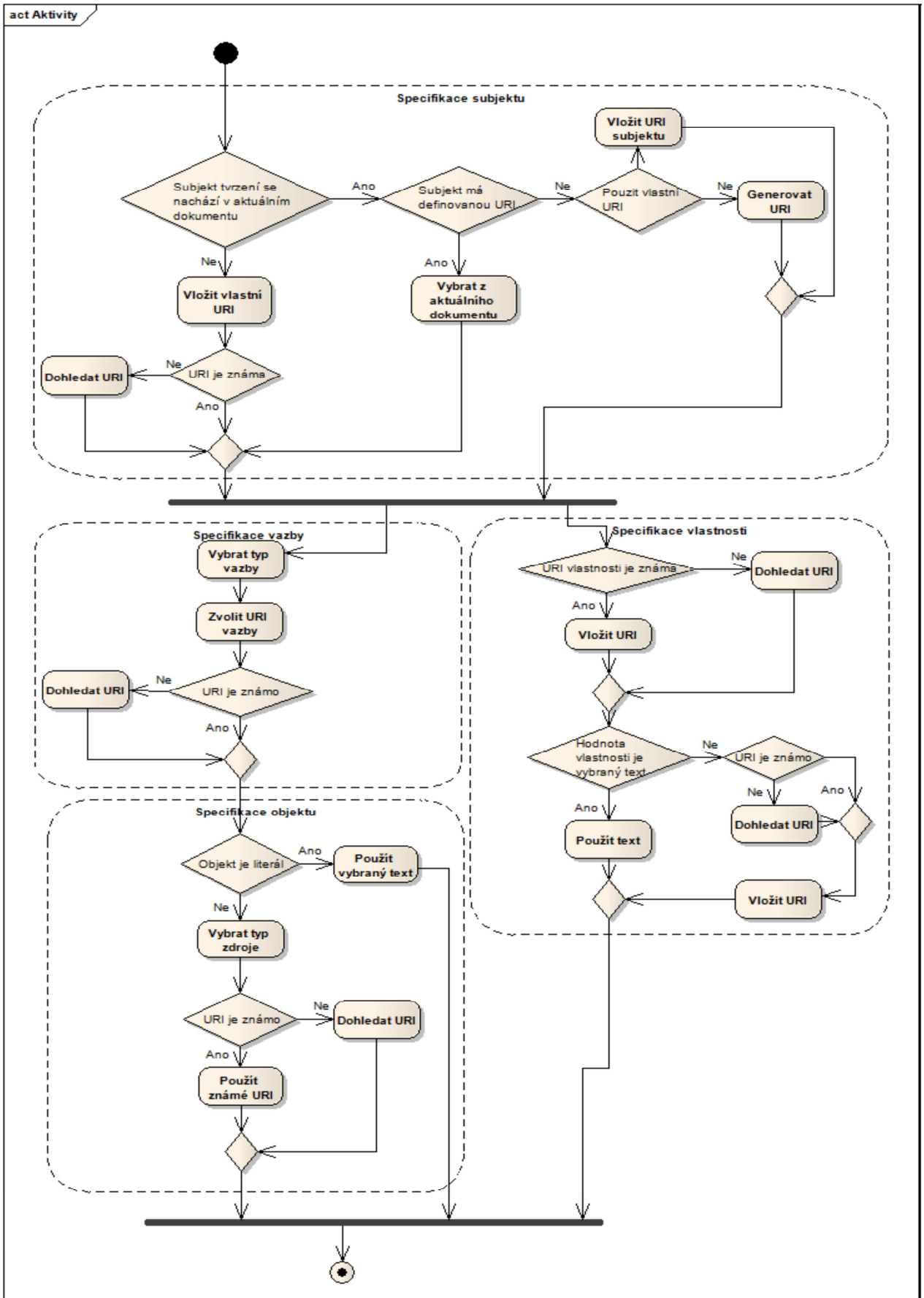
Samotný proces vyhledávání spočívá v odeslání požadavku na Swoogle prostřednictvím REST klienta. Podle typu hledání je pak následně formulován dotaz (v podmínkách RESTu jde prakticky o předávané URL), který je metodou HTTP GET odeslán na Swoogle ke zpracování. Podoba takového dotazu pak může vypadat například takto:

```
http://sparql.cs.umbc.edu:80/swoogle31/q?queryType=
search_swt&searchString=place&key=demo
```

ResourceLookup user-control reší zobrazování výsledku buď pomocí vlastního zobrazovacího prvku, nebo s pomocí externího prvku, jehož ID je předáváno user-controlu a na němž proběhne tzv. data-binding.

V případě vlastního prvku probíhá selekce daného identifikátoru zdroje URI tak, že na něj uživatel klikne. Pokud by bylo užito externího prvku, musí se zajistit vyvolání události *UriSelected* objektu *ResourceLookup*, které zajistí předání vybraného URI zpět do anotačního formuláře. Po výběru URI se pak dialogové okno uzavře a celý anotační proces je u konce.

Schématický návrh celé výše popsané koncepce sémantického anotačního procesu včetně dohledávání zdrojových URI je přehledně popsán prostřednictvím UML diagramového modelu na obrázku 4.3.



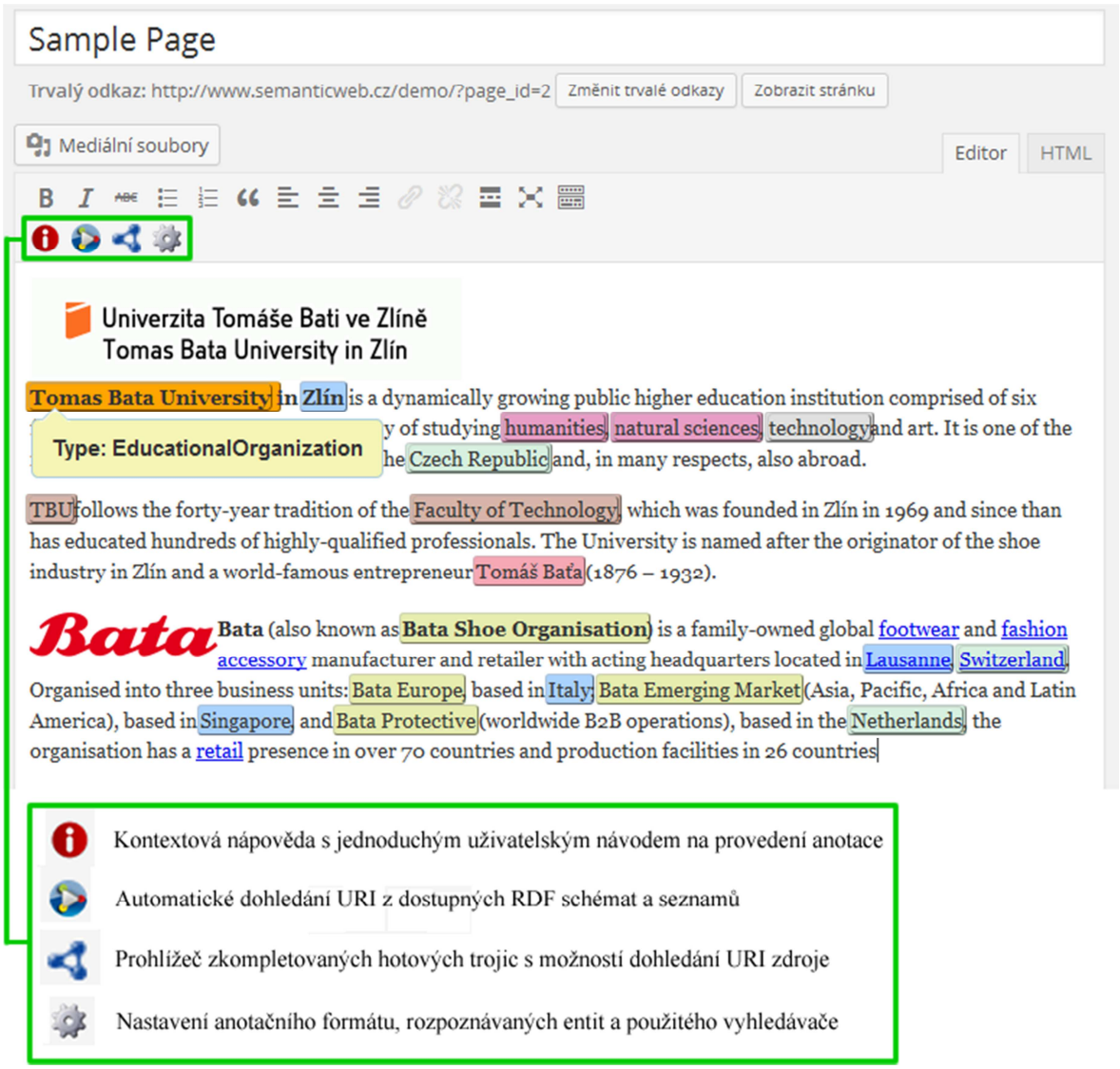
Obr. 4.3: Schematický návrh sémantického anotačního procesu

Metodická omezení daného řešení

S ohledem na algoritmizaci sémantické anotace bylo třeba přistoupit k omezení některých možností, jak s pomocí RDFa anotaci provést. Jedná se především o tato omezení:

- Nelze provádět řetězení tvrzení, tzn. objekt jednoho tvrzení nemůže být současně subjektem druhého.
- Subjekt musí být vždy vyjádřen atributem *about* – RDFa povoluje uvést subjekt pomocí sekundárního atributu pro subjekt *src*.
- Nelze používat URI ve zkrácené formě, tzn. formáty identifikátorů nejsou povoleny jako prefixované.

Výsledná podoba implementace do WYSIWYG editoru



Obr. 4.4: Ukázka sémantické anotace v prostředí CMS WP.NET 3.8.1

4.4 Praktické možnosti využití zpracování přirozeného jazyka

Myšlenka počítačového zpracování přirozeného jazyka v souvislosti s pokročilými technikami vyhledávání na Webu, je v posledních několika letech podrobena intenzivnímu zkoumání a tento trend je znatelný i v počtu stále častějších experimentech mezi moderními vyhledávači, spočívajících ve snaze o začlenění dotazování v přirozeném jazyce do vlastní funkcionality.

V návaznosti na tuto vědecky neustále aktuální oblast výzkumu, bude v následujících kapitolách představen koncept návrhu systémové architektury pro dotazování v přirozeném jazyce s rozšířením pro Sémantický web, což představuje zcela nové možnosti optimalizace dané technologie pro masovější užití a poukazuje na dosud nepříliš zmapovanou oblast ve zkoumání této problematiky, která bude v následujících částech diskutována při demonstraci řešení praktického příkladu z oblasti doménově orientované ontologie.

Jako zdroj pro dolování webových dat k dalšímu zpracování byl vybrán internetový obchod obsahující množství produktů, odpovídajících zájmové doméně barvy-laky. Zastoupeny jsou zde produkty různého druhu, rozdělených podle specifického účelu použití, velikosti balení, barev a odstínů.

Z pohledu zákazníka neznalého problematiky, může být i v sebelépe organizovaném katalogu zboží problém najít přesně ten produkt, který pro svůj konkrétní účel potřebuje. Nejčastějším problémem v této oblasti je zejména neurčité a o dalších vlastnostech nevypovídající kódové označení barevných odstínů, vhodnost či naopak nevhodnost některých barev pro různé povrchy, možnost aplikace v interiéru či exteriéru, odolnost proti UV záření apod.

Bez předchozích zkušeností nebo dalších doplňujících informací (poskytnutých např. kvalifikovaným prodejcem) se pak výběr a nákup tohoto druhu zboží stává leckdy značně problematickým.

Jaké by to však bylo, kdyby se zákazník při nákupu v internetovém obchodě nemusel ptát žádného „živého“ prodejce, suplujícího ve své podstatě fakty a vědomostmi naplněnou ontologii, nýbrž svůj dotaz v přirozeném jazyce vepsal přímo do vyhledávacího formuláře na stránce? Vzorový příklad dotazu formulovaného přirozeným jazykem by pak mohl vypadat např. takto:

„Chtěl bych levnou bledě modrou omyvatelnou a paropropustnou barvu použitelnou pro nátěr vnitřních prostor, tak aby balení vystačilo na 20m2 plochy.“

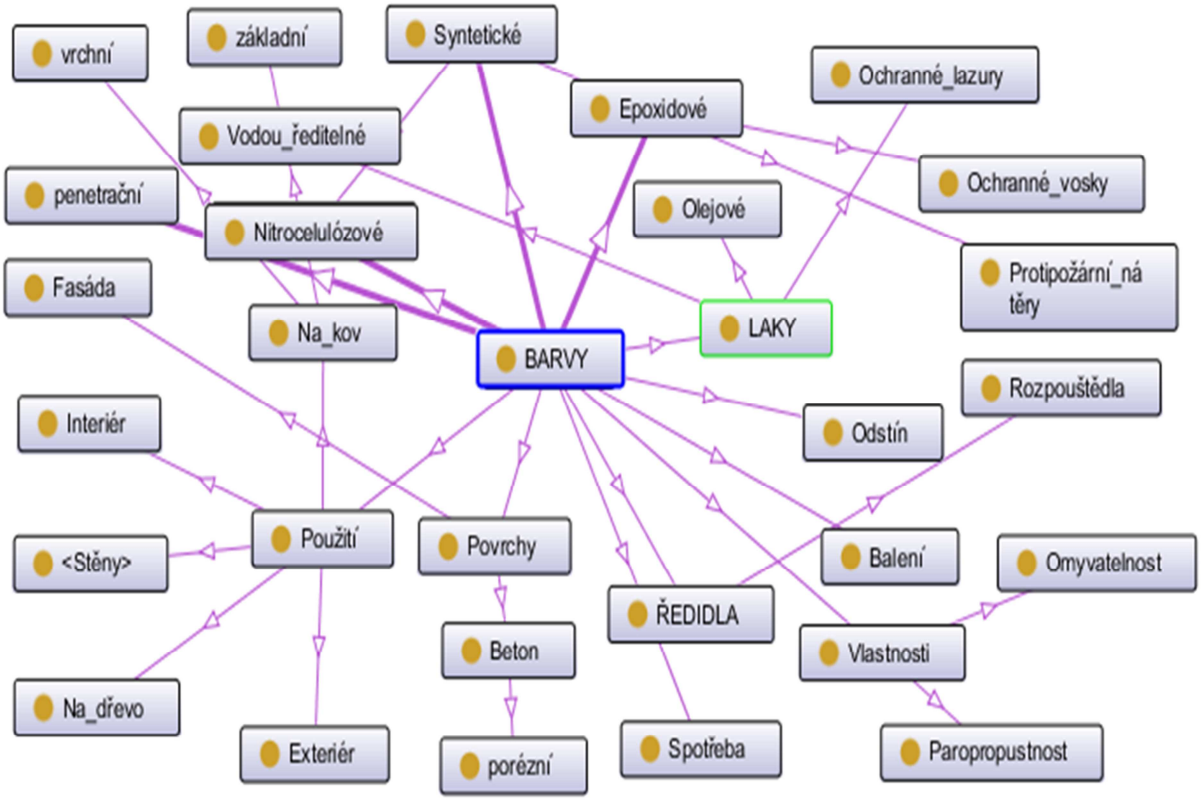
Níže navrhované rozhraní pro následné zpracování takového dotazu pak na svém výstupu zobrazí seznam relevantních výsledků, odpovídajících alespoň z větší části takto položenému dotazu (viz obrázek 4.6).

4.4.1 Znalostní báze a doménová ontologie

V pojetí navrhovaného rozšíření systému pro zpracování dotazů prostřednictvím přirozeného jazyka, představuje znalostní báze samostatný modul, který zprostředkuje co možná nejkomplexnější popisné informace o specifické zájmové doméně ve formě strukturovaných dat, přičemž v prostředí Sémantického webu jsou hlavním nástrojem pro modelování takovýchto znalostních bází právě instancemi naplněné ontologie.

Východiskem při vývoji ontologií je zpravidla seznam relevantních termínů, s následným odlišením ontologických typů, specifikace taxonomie a také vytvoření netaxonomických relací, atributů a instancí, přičemž již v této fázi je možno s výhodou využít podporu softwarového anotačního nástroje navrženého v kapitole 4.3, který proces při vytváření nové ontologie značně usnadňuje.

Každá ontologie má svoji vlastní jedinečnou strukturu v závislosti na doméně, kterou popisuje. Pro tento konkrétní příklad byl zvolen návrh struktury vycházející z datových podkladů produktového portfolia firmy Chemolak, a.s. popisovaném v rámci uvažované doménové ontologie barvy-laky. Taxonomická struktura byla vytvořena ve specializovaném software pro tvorbu ontologií Protegé 4.3, jejíž výstup včetně vztahů mezi jednotlivými třídami je pomocí vykreslovací technologie OntoGraph schematicky znázorněn na obrázku 4.5.



Obr. 4.5: *Ukázka části struktury zkoumané doménové ontologie*

Výsledná struktura ontologie je více či méně plochá, vyznačuje se jednou hlavní třídou s mnoha objekty a datovými vlastnostmi, kde zároveň platí, že jedna produktová karta je ekvivalentní jedné instanci hlavní třídy. V souvislosti s vlastnostmi daných instancí se mohou v ontologiích objevovat data strukturovaná (obsahující sémanticky definované objektové vlastnosti) a data textová (v nestrukturovaném a tedy sémanticky nepopsaném obsahu).

Z pohledu počítačového zpracování přirozeného jazyka, jsou přitom pro dosažení relevantních výsledků klíčová právě data strukturovaná. Na druhou stranu je však drtivá většina informací dnes dostupná pouze v čistě textové podobě a bez dalších doplňujících vlastností, což vlastní sémantické vyhledávání značně komplikuje. Bez předchozí automatické či manuální sémantické anotace je tedy také nasazení technologie pro zpracování dotazů v přirozeném jazyce prakticky bezpředmětné, neboť dosažené výsledky jsou zcela irelevantní v porovnání s použitím uvedených sémanticky strukturovaných dat.

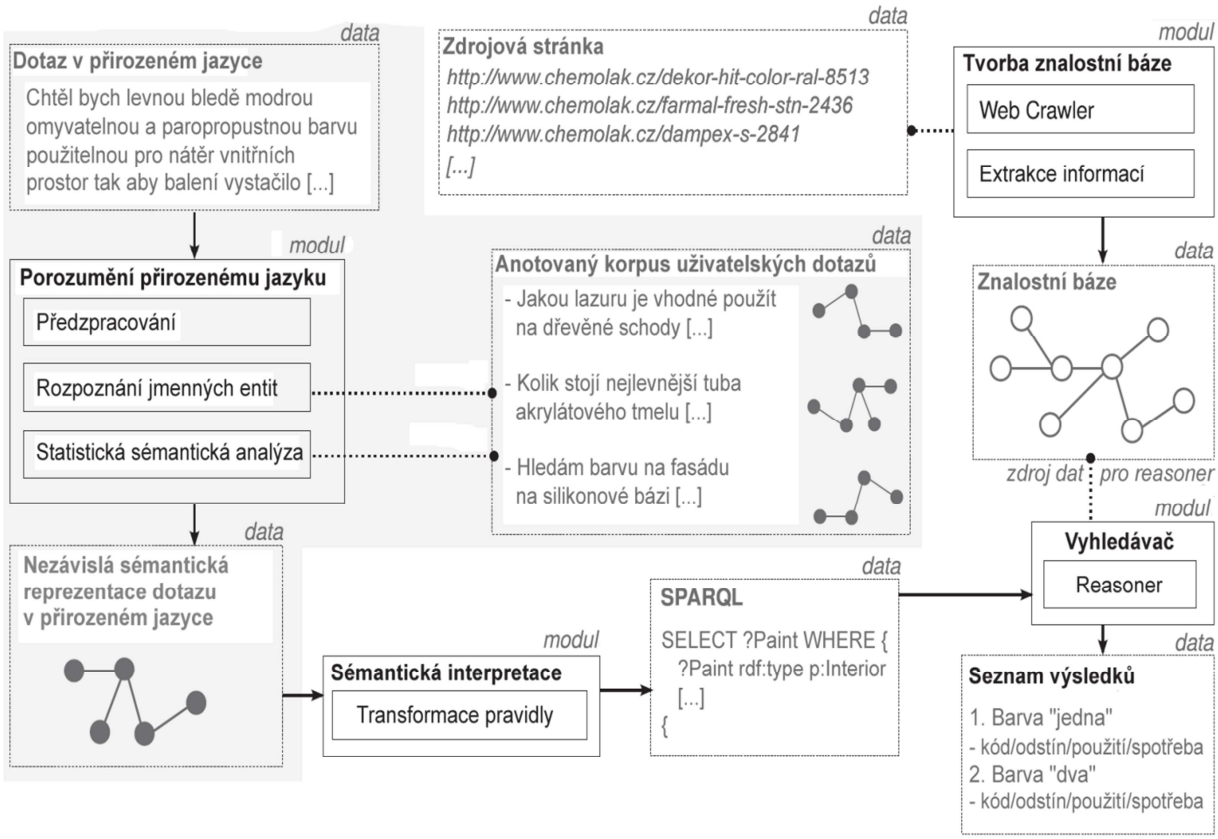
Právě precizní sémantika je hlavní výhodou každé dobře navržené ontologie, protože umožňuje za pomoci komplexních sémantických konstrukcí vyvozovat nová fakta a relace mezi jednotlivými subjekty.

4.4.2 Systémová architektura a zpracování dotazu

Porozumění přirozenému jazyku – NLU *(Natural Language Understanding)*, představuje jádro celého procesu počítačového zpracování dotazů v přirozeném jazyce, které jde ruku v ruce s aktuálním vývojem Sémantického webu – s tím rozdílem, že NLU z historického hlediska (např. v aplikacích pro převod lidské řeči na text) byl vnímán jako prostředník mezi člověkem a počítačem. Zde však v duchu myšlenky Sémantického webu půjde především o formalizovanou výměnu webových dat, přístupných současně jak pro lidi, tak i počítače.

Funkční schéma na obrázku 4.6 popisuje procesní systémovou strukturu na příkladu testovacího dotazu zmíněného v úvodní části této kapitoly. Uživatel na vstupu položí dotaz v přirozeném jazyce, přičemž zde nejsou žádná omezení, co se délky dotazu týče a lze tedy následně analyzovat jak jednoduché věty či klíčová slovní spojení, tak i rozvitá souvětí nebo dokonce celé odstavce textu. Dotaz je analyzován prostřednictvím NLU komponenty a následně porovnán s výstupem sémantické reprezentace stejného dotazu z nezávislé znalostní báze.

NLU komponenta sestává ze tří základních bloků pro předzpracování dotazu, rozpoznání jmenné entity a sémantického analyzátoru, který navíc využívá statistického modelu a k němu přidruženého korpusu sémanticky anotovaných dat, tvořených souborem hojně frekventovaných uživatelských dotazů.



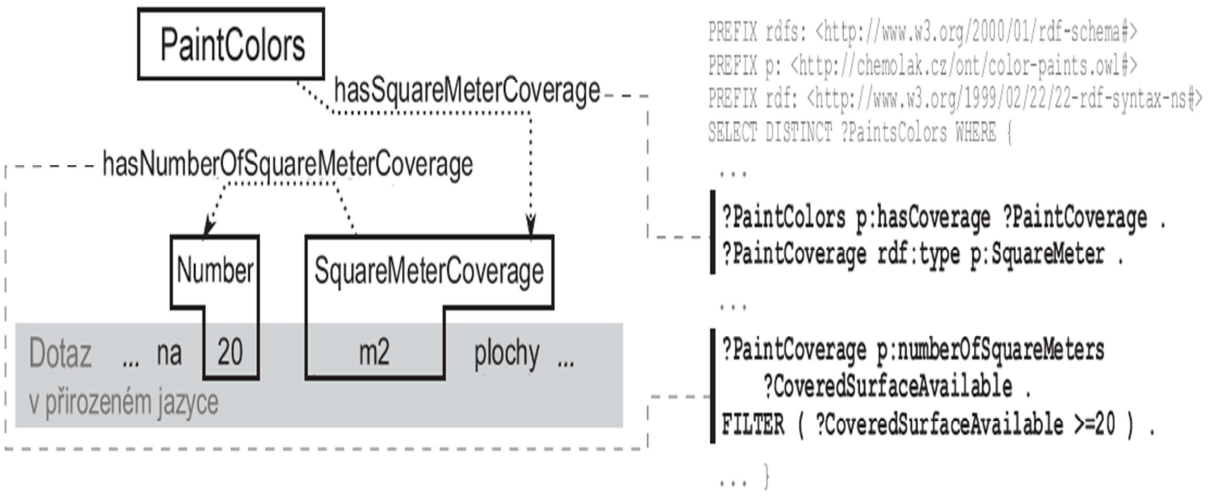
Obr. 4.6: Schéma procesu zpracování dotazu v přirozeném jazyce

Modul sémantické interpretace pak na základě transformačních pravidel takto anotovaná data z konkrétního dotazu v jazyce přirozeném převádí do podoby dotazovacího jazyka SPARQL a následně se již ve formě strojově zpracovatelného výrazu posílají dále proti znalostní bázi, tvořené již předem extrahovanými informacemi a tedy sémanticky obohacenými znalostmi, kde za využití sémantického odvozovacího modulu tzv. reasoneru (dále podrobněji popisováno v kapitole 4.5), dochází k inferenci relevantních dat, která jsou nakonec ve formě seznamu odpovídajících výsledků zobrazena uživateli.

4.4.3 Sémantická interpretace a vyhledávání

Jak již bylo v předchozí části nastíněno, k provedení vlastního vyhledání je nutno zpracovávaný dotaz na vstupu nejprve anotovat sémantickým popisem a následně upravit do tvaru výstupního SPARQL dotazu.

Příklad na obrázku 4.7 ukazuje jakým způsobem je vytvořená sémantická anotace z dotazu v přirozeném jazyce transformována do formy ontologického dotazovacího jazyka, kterým se pak následně v rámci back-end procesu provádí prohledání znalostní báze, kdy je prakticky možno (na základě předchozí sémantické anotace) procesovat konkrétní triple-trojici z původně na vstupu předkládaného dotazu v přirozeném jazyce na výstupní formát SPARQL jazyka.



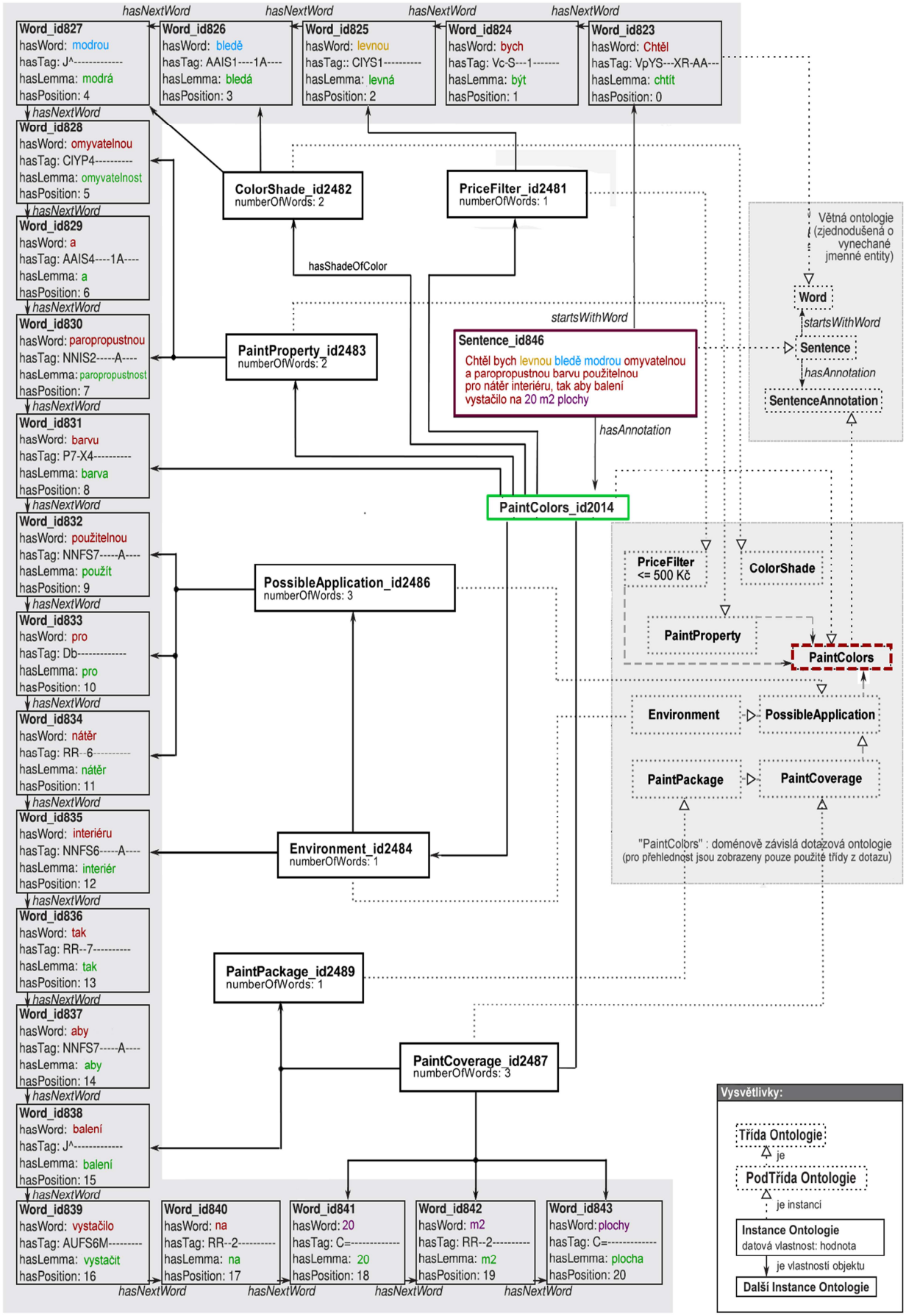
Obr. 4.7: *Rozdělení a úprava dotazu na strojově čitelný zápis*

Formulovaný dotaz ve SPARQL syntaxi dále postupuje ke zpracování do modulu inferenčního mechanismu, tzv. *reasoneru* (podrobněji samostatně popisováno v podkapitole 4.5) na jehož expresivitě a dalších stěžejních parametrech při transformaci ze sady doménově-specifických pravidel závisí také celková přesnost výsledků vyhledávání vzhledem k požadovanému dotazu.

Aby systém pro zpracování přirozeného jazyka byl schopný automatické sémantické inference z předkládaného dotazu, neobejde se bez znalostí příslušné jazykové struktuře a musí v něm být zabudovány znalosti o tom:

- Co jsou to slova (slovní tvary a jejich složky – morfémy),
- Jak se slova (větné složky) kombinují do vět,
- Co slova označují a jaké jsou jejich významy,
- Jak se význam věty skládá z významů slov a slovních spojení.

Samotný proces zpracování vyhledávaného dotazu (schématicky znázorněný na obrázku 4.8) tedy začíná nejprve postupným rozdělením vstupního dotazu na samostatné větné celky a dále jednotlivá slova (větné složky), která jsou poté anotována přiřazením unikátního identifikačního tagu. Současně je provedena lemmatizace, která vytvoří (vyhledá v databázi) k určitému tvaru slova jeho základní tvar, tzv. *lemma* (např. *„chtěl bych“* odvozeno na *„chtít být“)* a proces poté pokračuje vyvozením logických vztahů mezi zachycenými trojicemi (subjekt – predikát – objekt) a následným porovnáním se znalostní bází, kdy se dále vyhodnocují vlastnosti mezi jednotlivými třídami, podtřídami a instancemi (např. instance *PaintCoverage* je jednou z vlastností objektu instance ontologie *PaintColors*, která současně vyvozuje jakému balení daná spotřeba odpovídá a zohledňuje i další vlastnosti barvy (omyvatelnost, paropropustnost) až ke konečnému nalezení a filtraci odpovídajícího výsledku na základě požadovaných kritérií (např. levná – odpovídá ceně do 500 Kč apod.).



Obr. 4.8: Příklad vyhodnocení vstupního dotazu v přirozeném jazyce

4.5 Komparace sémantických inferenčních mechanismů

Data v ontologiích jsou modelována jako množiny vzájemně pojmenovaných vazeb mezi zdroji. Zásadním předpokladem pro další rozšiřování každé ontologie je schopnost inference nových vazeb na základě dodatečných informací ve formě slovníku nebo množiny pravidel.

Konsorcium W3C pro tento přístup poskytuje specifikaci RIF *(Rule Interchange Format)*, tedy formát pro výměnu pravidel mezi systémy, zaměřený na definici obecných inferenčních mechanismů, které jsou v oblasti Sémantického webu zastoupeny tzv. *reasonery*, což mohou být ve své podstatě softwarové moduly schopné vyvození logických konsekvencí ze sady uspořádaných faktů či axiómů, reprezentovaných znalostní bází, která obsahuje nashromážděné znalosti a zkušenosti expertů v dané oblasti se záběrem od nejobecnějších, učebnicových, až k úzce specializovaným informacím.

Mezi množstvím již dostupných či aktuálně vyvíjených sémantických reasonerů, jejichž podrobnější teoretický přehled lze nalézt v dostupné literatuře [17], bude pro jejich následné funkční a parametrové porovnání uvažováno pouze několik nejpopulárnějších, pro které již v současnosti existuje dostupné rozšíření v softwarovém konstrukčním nástroji Protégé či NeOn toolkit.

4.5.1 Sémantické frameworky pro práci s ontologiemi

V praxi při vývoji a poskytování abstraktních funkcionalit pro správu ontologií prostřednictvím API (Application Programming Interface) slouží sada softwarových knihoven, obecně označovaných jako Framework. Hlavní výhodou sémantických frameworků z pohledu vývojáře je především jednotné rozhraní pro snadný přístup k ontologiím a odvozovacím mechanismům.

Mezi trojici v současnosti nejvyužívanějších frameworků se řadí:

- **Jena –** Java Framework pro vývoj sémantických aplikací [14]. Ve své aktuální verzi (2.11.1) podporuje Resource Description Framework (RDF) [11], RDF Schema (RDFS) a jazyk OWL. Obsahuje také vlastní reasoner založený na RIF pravidlech, který umožňuje dotazování nad znalostními bázemi pomocí SPARQL jazyka. Vstupně výstupními formáty tohoto frameworku mohou být RDF ve formě RDF/XML a N-Triples trojice. Jedním z hlavních účelů architektury Jena frameworku je také poskytnutí jednoduché správy dat v RDF grafové vrstvě se současnou možností zobrazovat data jako N-Triple trojice.

- **OWLApi** – vzniklo jako pokus o poskytnutí snadno využitelné komponenty pro konstrukci aplikací na bázi jazyka OWL. Poskytuje rozhraní pro vytváření, manipulaci a seřazování OWL ontologií [3]. Aktuální OWLApi (v3) podporuje parsování formátů RDF/XML, OWL či Turtle. K provádění odvozování OWL ontologií na rozdíl od Jena nevyužívá vlastní reasoner – možnost využití externích reasonerů, které je však třeba implementovat zvlášť s umožněním přístupu přes OWLApi.

- **OWLlink** – představuje inferenční protokol pro výměnu informací mezi OWL aplikacemi a sémantickými reasonery, přičemž procesy inferenčních mechanismů mohou být delegovány také na jiné počítače s využitím distribuovaného výpočetního schématu. OWLlink (1.2.2) sestává z vlastního protokolového jádra a sady pro bindování dat, které může být realizováno jako XML přes HTTP (primární bindování), funkcionální syntaxe přes HTTP a S-Expression přes HTTP.

Na základě stručného výčtu sémantických frameworků pro práci s ontologiemi se nabízí otázka, která z představených variant bude svými vlastnostmi vhodná pro nasazení na některé z rozsáhlejších ontologií, obsahující velké množství vzájemných relačních vazeb.

Možností může být použití frameworku a integrovaného či externího reasoneru. Porovnání funkčních atributů jednotlivých inferenčních mechanismů a následné experimentální měření na široce profilované doménové ontologii bude obsahem následujících dvou podkapitol.

4.5.2 Funkční parametry odvozovacích modulů

Sémantický inferenční mechanismus každého reasoneru sestává z navzájem kooperujících součástí, zabezpečujících procedurální složku činnosti systému, přičemž takový odvozovací modul umožňuje v určitém rozsahu napodobovat expertní schopnost uvažování, kdy simuluje především ty činnosti, které souvisí s efektivním využíváním poznatků a zkušeností, získaných na základě asociací, hierarchií, příčinně-důsledkových vazeb, kontextů a spojování poznatků do vhodně souvisejících celků a posloupností.

Takto zavedený odvozovací modul tedy odpovídá mechanismům všeobecného uvažování, opírajícího se o bázi znalostí, na jejímž základě je možno konkrétní problémy řešit. Inferenční pravidla sémantických reasonerů jsou odvozována prostřednictvím deskripčního jazyka, vycházejícího z predikátové logiky prvního řádu, přičemž samotné odvození obvykle probíhá formou zpětného či dopředného řetězení [55].

Role deskripční logiky pro Sémantický web je formálním základem pro odvozování nad ontologiemi a její využití dodává ontologiím přidanou hodnotu oproti „ad-hoc“ konceptuálním modelům.

Ucelený přehled deskripčních logik využívaných v ontologickém jazyce *OWL* lze nalézt v literatuře [3], čistě pro účely této práce však postačí teoretické uvedení nejrozšířenější varianty deskripční logiky *SROIQ*, která představuje syntaktickou variantu poslední verze standardu *OWL 2* [59].

Dalšími stěžejními parametry sémantických reasonerů jsou:

- **Metodologie** – procedura nebo výpočetní algoritmus, který je reasonerem využíván k řešení základních problémů v deskripční logice.

- **Spolehlivost a kompletnost** – mohou významně urychlit celý proces; hodnotí se zda, všechny možné inference byly vyvozeny či nikoliv.

- **Inkrementální klasifikace (+/-)** – pokud již jednou dojde ke klasifikaci ontologie a tato je pak ještě později měněna (přidáváním/odebíráním), může reasoner využít předchozí informaci o klasifikaci s upravenými axiomy pro vyvození nového hierarchického konceptu.

- **Podpora pravidel** – určuje, zda je reasoner schopen vyvozování nových skutečností na základě podporovaných pravidel, nejčastěji SWRL *(Semantic Web Rule Language)*.

- **Zdůvodnitelnost** – tato funkce zajišťuje poskytnutí vysvětlení v případě nekonzistencí, které se mohou v ontologii objevit.

- **ABOX Reasoning** – vyvozování individuálních vztahů, zprostředkuje kontrolu konzistence instancí a zodpovídání konjunktivních dotazů.

Z několika desítek v současné době existujících sémantických reasonerů, bude nyní vyselektována pětice nejpoužívanějších zástupců dané kategorie samostatných odvozovacích modulů. Dvojice nejvhodnějších z nich pak bude vybrána také pro závěrečné praktické experimentální měření a porovnání se zástupcem vybraného sémantického frameworku s integrovaným reasonerem.

V následném přehledu funkčních parametrů sémantických reasonerů (Tabulka 10) bude hodnocena zejména použitá varianta deskripční logiky a metodologie, schopnost inferenční spolehlivosti a kompletnosti. Dále pak bude zohledněna podpora odvozovacích pravidel a dostupných programových prostředků [40].

Tabulka 10 – Přehled funkčních parametrů sémantických reasonerů

Reasoner	Pellet	RACER	FACT++	HermiT	CEL
Deskripční logika	SROIQ	SHIQ	SROIQ	SROIQ	EL+
Nativní profil	DL, EL	DL	DL	DL	EL
Metodologie	Tableaux	Tableaux	Tableaux	HyperTab	Completition
Spolehlivost	✓	✓	✓	✓	✓
Kompletnost	✓	✓	✓	✓	✓
Klasifikace (+)	✓	✗	✗	✗	✓
Klasifikace (–)	✓	✗	✗	✗	✗
Podpora pravidel	✓ (SWRL)	✓ (SWRL)	✗	✓ (SWRL)	✗
Zdůvodnitelnost	✓	✓	✗	✗	✓
ABOX Reasoning	✓	✓	✓	✓	✓
OWL API	✓	✓	✓	✓	✓
OWL Link API	✓	✓	✓	✓	✓
Jena framework	✓	✗	✗	✗	✗
Podpora Protégé	✓	✓	✓	✓	✓
Podpora NeOn	✓	✗	✗	✓	✗
Progr. jazyk	Java	LISP	C++	Java	LISP
Licence	Open source	Commercial	Open source	Open source	Commercial

Z uvedené tabulky vyplývá, že všechny reasonery využívají hodnotících parametrů pro spolehlivost a kompletnost. Pouze Pellet je schopen vyvození plné inkrementální klasifikace, ostatní nikoliv. Podporu vyvozovacích pravidel SWRL vykazují Pellet, RACER a HermiT, avšak pouze Pellet, RACER a CEL poskytují zdůvodnění pro nekonzistenci dat v ontologii.

Na základě uvedeného přehledového srovnání parametrů jednotlivých reasonerů, bude pro další rozbor a zkoumání uvažována dvojice samostatných modulů Pellet a HermiT, a to zejména z důvodu plné podpory obou v současnosti nejvyužívanějších softwarových nástrojů pro tvorbu a správu ontologií – Protégé i NeOn.

Tato volba se pro další srovnání jeví jako výhodná i z důvodu shodné platformy implementačního jazyka (Java) v obou případech i vzhledem ke stejně založenému Jena frameworku, který byl vybrán jako zástupce integrovaných reasonerů pro následné experimentální měření výkonnosti a paměťové náročnosti, které bude předmětem následující části 4.5.3.

Funkční parametry vybraných reasonerů budou nyní popsány podrobněji:

- **Jena –** ač ve své podstatě framework, jehož základní distribuce obsahuje několik inferenčních mechanismů pro odvozování prostřednictvím RDFS či OWL, obsahuje navíc Jena také dva interní rule enginy: RETE [21] a tabled datalog engine , které lze použít buď samostatně, nebo jako jeden hybridní rule engine provozovaný přes implementační třídu: *GenericRuleReasoner*. V hybridním módu jsou data zprocesována nejprve prostřednictvím dopředného vyvozování prostřednictvím RETE a poté jsou výsledky předány ke zpracování zpětně řetězícího tablet datalog enginu, který nakonec zobrazí výsledky na základě položeného dotazu.

- **HermiT** – je popisný odvozovací mechanismus pro *SROIQ* logiku, distribuovaný pod licencí LGPL, na který může být přistupováno prostřednictvím rozhraní OWLlink nebo OWLApi. Hlavním rozdílem oproti ostatním reasonerům je to, že pro svůj algoritmus hyper-tabulkového kalkulu využívá první pravidlo větné (nikoli predikátové) logiky [3], což umožňuje redukci počtu možných uvažovaných modelů (zaměřit se pouze na jednu možnost významné podskupiny ontologií).

- **Pellet** – představuje samostatný reasoner pro formát OWL 2 s distribucí AGPLv3 pro open source projekty a speciální licencí pro komerční aplikace, který v rámci *SROIQ* logiky poskytuje techniky pro plně inkrementální klasifikaci a také několik optimalizačních technik jako zpětný skok, zjednodušení či absorbce [55]. Je založený na tabulkovém algoritmu vyvinutém pro expresivní deskripční logiky a umožňuje přímé propojení s frameworkem Jena. Dostupná je také podpora pro přímé dotazování prostřednictvím SPARQL a SWRL pravidel a samotný reasoner může být přistupován pomocí OWLApi a OWLlink rozhraní.

4.5.3 Výsledky experimentálního měření

Jako výchozího souboru zdrojových dat, pro následná praktická experimentální měření výkonu a paměťové náročnosti, bylo použito volně dostupné verze doménové ontologie GALEN [53] v anglickém jazyce, která v sobě zahrnuje komplexní přehled problematiky zdravotnické terminologie v oblasti anatomie, chorob a dostupných medikamentů. Pro účely měření bylo vytvořeno rozšíření zdrojové ontologie o množinu uživatelů-pacientů, tak aby výsledné rozvržení korespondovalo se zamýšleným experimentem na vytvoření pravidel, která budou kontrolovat, zda pacienti berou správné léky, ve správný čas a na základě indikace jejich skutečné nemoci, předem diagnostikované na základě zjevných příznaků typických pro danou chorobu.

K celkovému zjednodušení zdrojové ontologie, byly pro měření uvažovány a vyselektovány pouze tři oblasti: (i) *activity* (ii) *routine* (iii) *time*, přičemž takto modifikovaný dataset byl sestaven jako týdenní plán aktivit uživatele-pacienta a v souvislosti s tím byly také zavedeny tři nové stěžejní OWL třídy:

- DisabledPerson – osoba, která potřebuje zdravotní péči
- MedicalTreatment – léčba, druh léků a jejich předepisování
- Activity – proces aplikace léčby / braní léků

Na základě předchozího teoretického uvedení sémantických frameworků (v části 4.5.1) a po následném srovnání parametrů sémantických reasonerů (část 4.5.2), byly pro další praktická testování zvoleny tyto kombinace:

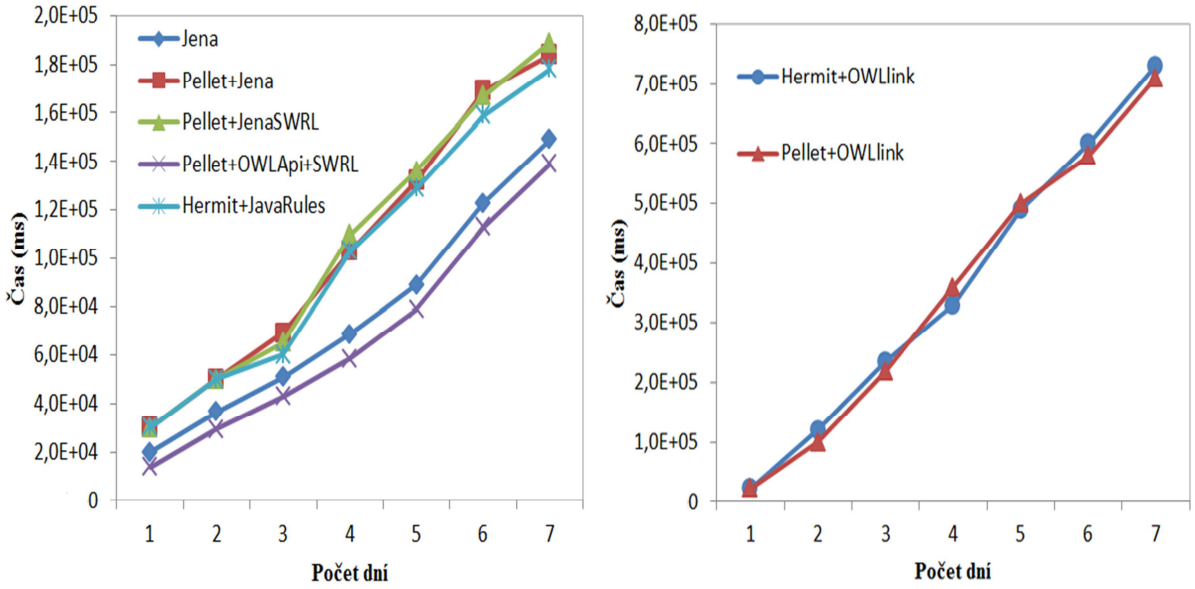
1. **Jena** (současně jako framework a integrovaný reasoner)
2. **Pellet + Jena** (reasoner + framework – samostatně)
3. **Pellet + Jena + SWRL** (reasoner + framework s využitím SWRL)
4. **Pellet + OWLApi + SWRL** (reasoner + framework s využitím SWRL)
5. **HermiT + Java rules** (reasoner + vlastní Java pravidla)
6. **HermiT + OWLlink** (reasoner + framework – samostatně)
7. **Pellet + OWLlink** (reasoner + framework – samostatně)

Každá z kombinací byla se stejným datasetem testována celkem desetkrát, přičemž data uvedená v grafech 4.7 a 4.8 zachycují průměrné hodnoty se standardními odchylkami. Protože spouštění sémantických pravidel (SWRL) tzv. *out-of-the-box* [31] je standardně dostupné pouze pro Jena Framework a Pellet reasoner, byla pro zbylé kombinace reasoneru HermiT a frameworku OWLlink a OWLApi vytvořena vlastní Java pravidla k pokrytí dané ontologie.

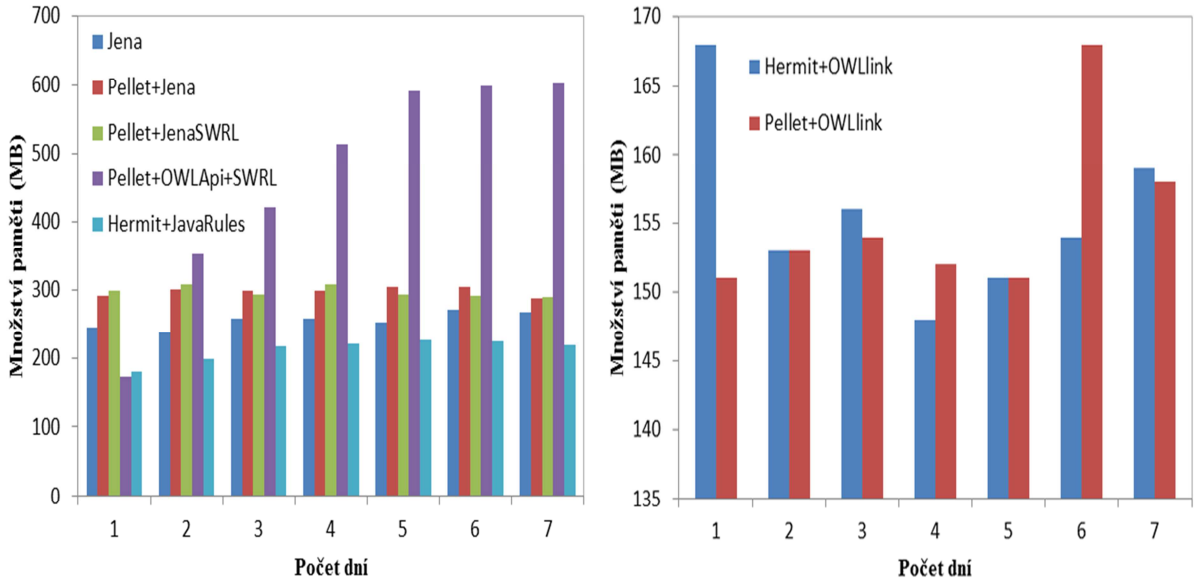
Konfigurace testovacího stroje:
Notebook: Acer Aspire 4830GT s 4 GB RAM
Procesor: Intel Core i5-2410M
OS: Ubuntu 11.10 64bit.
Java: 1.6.0_35

Vlastní experiment pro měření času potřebného k vyhodnocení simulované aktivity z předem připraveného datasetu v závislosti na paměťovém zatížení počítače pak probíhal následovně:

Po spuštění programu dojde nejprve k načtení celé ontologie reasonerem (i několik desítek minut, v závislosti na hardwarové konfiguraci počítače). Následně se začnou načítat simulované aktivity z použitého datasetu a postupně je tak neustále upravován a rozšiřován i kontext uživatele – pacienta, a to s každým dalším dnem, přičemž se sleduje čas potřebný k vyhodnocení a množství obsluhované operační paměti v rámci daného procesu.



Graf 4.7: Odezva na vyhodnocení požadavku v závislosti na čase



Graf 4.8: Množství využité operační paměti v závislosti na čase

Grafy 4.7 ukazují čas potřebný na vyvození logických konsekvencí v závislosti na simulované aktivitě (např. pacient, který bere léky na jednu chorobu, nemůže zároveň ve stejnou dobu brát léky na chorobu jinou apod.). Jak je patrné z grafu pro kombinace (vlevo i vpravo), tento čas, potřebný k vyhodnocení s počtem dnů v rámci rozplánovaného týdne neustále narůstá, a to vlivem stále se zvětšujícího objemu možných odvozovaných relačních spojení (např. když si pacient jeden den vzal jiný lék, než měl, je třeba vyvodit možná rizika a ovlivnitelnost léků navzájem apod.). Testovaný dataset neobsahuje ani zdaleka všechny možné situace, které by v dané oblasti mohly nastat, ale snaží se pokrýt jen ty nejzásadnější případy i s ohledem na množství operační paměti, jejíž využití v závislosti na testované kombinaci je zobrazeno v grafech 4.8.

Z výsledků měření a praktického srovnání jednotlivých kombinací sémantických inferenčních mechanismů vyplývá, že z hlediska množství času potřebného pro vyhodnocení se jako nejrychlejší ukázalo použití kombinace OWLApi+SWRL s přístupem přes Pellet, přičemž tato kombinace byla o 14% rychlejší než druhá v pořadí Jena. Na druhou stranu však nutno dodat, že současně tato nejrychlejší kombinace měla také celkově největší spotřebu operační paměti, která v přímé úměře s časem dále narůstala až o třetinu.

Obecně lze říci, že samostatný reasoner, užitý v kombinaci s Jena framework (Pellet+Jena, Pellet+JenaSWRL) je pomalejší než ostatní. Nadruhou stranu, při použití vlastního integrovaného reasoneru ve frameworku Jena, je znatelné zlepšení výkonu, tedy i dosahovaného času a lze tedy říci, že framework Jena funguje podstatně lépe samostatně nežli v kombinaci s externími reasonery.

Co se týče vyhodnocení z hlediska paměťové náročnosti celého procesu, tak zde se jako nejvhodnější jeví kombinace HermiT+Java rules přes OWLApi. V konkrétním případě Pellet+OWLApi+SWRL lze vypozorovat inverzní relaci mezi množstvím využité paměti a časem běhu, protože tato kombinace má nejnižší prováděcí čas a současně také nejvyšší paměťové nároky. V případě použití OWLlink protokolu je také znatelný vyšší nárůst času potřebného k provedení simulace, což je dáno nutností roztřídění a následné zpětné serializace dat přidáním abstraktní vrstvy při přístupu klient/server.

Na základě uvedeného výčtu a výsledků vycházejících z praktického porovnání vlastností jednotlivých frameworků i reasonerů lze potvrdit, že inferenční schopnosti každého z nich závisí vždy současně také na možnostech toho druhého, se kterým figuruje v kombinaci, přičemž použití kombinace (samostatný reasoner a Framework) nemusí být vždy za každé situace výhodné.

I proto je závěrem nutno konstatovat, že v praxi obecně se parametry jednotlivých reasonerů mohou značně lišit, a to zejména jsou-li pro finální nasazení zahrnuty všechny jeho atributy bez dalšího omezení pravidly.

Neexistuje tedy univerzální reasoner vhodný pro všechny druhy aplikací. Stejně tak rovněž rozdílné výsledky patrné v časové náročnosti zpracování automatického odvozování poukazují na fakt, že samotné praktické nasazení v budoucnu musí počítat s opravdu výkonným hardwarovým vybavením a před samotným praktickým nasazením bude třeba vždy provést nejprve kvalifikovaný kritický odhad a hodnocení požadavků na aplikaci reálného světa, tak aby následně extrakce a odvozování nových faktů nad rozsáhlými ontologiemi (např. v lékařství či Sémantickém webu obecně) bylo nejen technologicky funkční, ale též hlavně uživatelsky snadno použitelnou oblastí.

5. ZVOLENÉ METODY ZPRACOVÁNÍ

Ke komplexnímu výzkumu dané problematiky v oblasti Sémantického webu bylo přistupováno na základě normativního přístupu s využitím teoretických vědeckých metod analýzy, syntézy a dedukce.

- K získání základních východisek pro další výzkum oblasti využití sémantických technologií byla použita metoda analýzy.
- K analýze v teoretické části práce byly citovány informační zdroje, uvedené v přehledu použité literatury a zdrojů.
- Na základě syntézy zjištěných problémů byly definovány cíle práce, obecný rámec zamýšleného řešení a jeho požadované charakteristiky.
- Při zpracování praktických možností využití dotazování v přirozeném jazyce, byla použita zejména metoda deduktivní.
- K návrhu softwarového řešení pro podporu sémantického anotačního procesu, byly vytvářeny schématické UML modely, na kterých byla ověřována vhodnost řešení dílčích problémů, zejména principů pro správu požadavků, diagramů tříd pro definici struktury systému, diagramů aktivit pro popis vnitřních procesů systému a sekvenčních diagramů pro popis interakce systému a jeho komponent.
- Aplikační platformou pro navrženou komponentu sémantické anotace open source WYSIWYG nástroje CKEditor je na straně serveru Microsoft .NET Framework 3.5 a programovací jazyk C#, na straně klienta pak programovací jazyk JavaScript.
- Při zkoumání rozdílů mezi vlastnostmi v praxi nejčastěji užívaných ontologických reasonerů byla použita metoda komparační.
- V období od února 2012 do dubna 2014 byly dotazníkovou metodou ověřovány vyslovené hypotézy v souvislosti s informovaností o existenci koncepce a možného využití Sémantického webu mezi širokou veřejností v České republice.
- V návaznosti na výsledky dotazníkového šetření, ze kterého vyplynula vítaná potřeba komplexního zmapování této problematiky v českém jazyce, byl následně vytvořen veřejně dostupný informační portál na bázi uživatelsky rozšiřitelné encyklopedie, realizované s využitím opensource frameworku MediaWiki ve verzi 1.21.7, s modifikovaným rozšířením pro generování obsahu dle standardů Sémantického webu.

Pro větší přehlednost je k disertační práci připojen seznam použitých pojmů a zkratek, shrnující nejčastěji používané odborné termíny v rámci dané oblasti. Samotná práce je ve své formální úpravě psána dle normy *ČSN ISO 690 Informace a dokumentace – Pravidla pro bibliografické odkazy a citace informačních zdrojů* a doporučení v literatuře [18].

6. VYUŽITELNOST VÝSLEDKŮ

V návaznosti na provedení revize cílů disertační práce (vyslovených v kapitole 2), budou nyní ve stručnosti zhodnoceny výsledné výstupy, společně s uvedením možného dalšího využití a přínosu pro vědecké i praktické účely.

6.1 Přínos pro vědu

Výsledky provedeného dotazníkového šetření napomohly k lepšímu vymezení budoucího vývoje tematické oblasti, v rámci které veřejnost očekává postupnou širší informovanost o technologiích a standardech Sémantického webu. Současně bylo poukázáno na neexistenci dostatečného množství vzájemně propojených a sémanticky anotovaných dat, což by v praxi pomohla vyřešit masivnější podpora některého z uživatelsky přívětivých anotačních nástrojů.

Statistická porovnání kombinací dostupných sémantických frameworků a reasonerů mohou být užitečná pro odborné pracovníky, kteří se zkoumáním v oblasti vývoje a testování ontologií začínají a uvítají tak případná doporučení i návrhy vhodných programových prostředků a postupů pro odvozování nových znalostí v rámci aplikovaného výzkumu Sémantického webu, což jistě dopomůže k rychlejší a snazší orientaci v dané problematice.

Realizace encyklopedického informačního portálu dle očekávání pomohla k navázání bližší spolupráce se stejně zaměřenými badateli při shromažďování dostupných informací. Součástí řešení byla také realizace výukového tutoriálu spojená s modulem hodnocených znalostních testů. Tento praktický výstup lze pak následně jako kompaktní celek využít např. v rámci podpory akademické výuky předmětu, specializovaného na Sémantický web a jeho technologie.

6.2 Přínos pro praxi

Realizace a rozšíření univerzálně použitelného anotačního nástroje s podporou stěžejních formátů Sémantického webu je možno chápat jako snahu bližší o zpřístupnění techniky sémantické anotace webových dokumentů, která byla dosud na úrovni běžného uživatele jen obtížně aplikovatelná.

Představené částečně univerzální rozšíření WYSIWIG editoru umožní bližší seznámení s uživatelsky přívětivou technikou sémantické anotace i těm uživatelům, kterým tato problematika byla doposud neznámou.

Praktické využití komponenta nalezne v komerčním prostředí firem (např. znalostní báze v podnikovém intranetu) i pro osobní účely při publikaci na stávajícím Webu (např. blogy, články apod.), což představuje jednu ze slibných možností budoucího rozvoje Sémantického webu právě prostřednictvím v současné době populárních CMS – systémů pro správu obsahu.

ZÁVĚR

Sémantický web se má stát novým evolučním stupněm stávajícího webu. Jedná se o Web, kde jsou informace strukturovány a uloženy podle standardizovaných pravidel, což usnadňuje jejich vyhledání a zpracování. Skutečný Web 3.0, za který bývá mnohdy Sémantický web nepřesně považován, je však zatím poněkud vzdálen realitě, a to zejména proto, že dosud neexistuje dostatečné množství sémantických dat, na kterých by Sémantický web mohl stavět. Jedním z aktuálních trendů při vytváření Sémantického webu, jsou Linked Data. Vztah mezi těmito termíny je vhodně vyjádřen v [6]:

„Vize Sémantického webu je založena na budování globálního Webu strojově zpracovatelných dat. Zatímco Sémantický web, nebo Web dat je cílem či výsledkem, Linked Data představují prostředek nebo způsob jeho dosažení.“

Myšlenky i celková koncepce Sémantického webu jsou neustálým předmětem diskusí, především mezi softwarovými a vědomostními inženýry, kterých se vývoj i budoucí realizace dotýká nejvíce. Většina z nich přitom očekává další vývoj v této oblasti se zájmem, i když postoje nejsou vždy jednoznačně kladné.

Původní představa Sémantického webu je sice na první pohled velmi líbivá a částečně dnes již i funkční, nicméně její technická realizace poněkud předběhla mentální schopnosti většiny běžných uživatelů. Sémantický web zahrnuje technologie, které tvoří pomyslný most mezi prvotní myšlenkou a jejím následným naplněním, avšak pro jeho smysluplné využití je třeba nejprve tyto technologie znát a umět používat, což může představovat komplikace při následném širším praktickém nasazení. Tuto technologickou a dovednostní bariéru může fakticky z části překonat postupné zavádění softwarových nástrojů pro vytváření sémanticky anotovaných webových dokumentů prostřednictvím uživatelsky přívětivého prostředí WYSIWYG editorů, které umožňují vlastní tvorbu či editaci takového obsahu bez nutné znalosti značkovacího jazyka.

Jako další v současné praxi uplatnitelné řešení se také ukazují mikroformáty [1], které již prokázaly svoji životaschopnost (běžné je například používání tagů vCard, hCalendar apod.). Z mikroformátů ostatně vychází i oficiální HTML 5 specifikace skupiny WHATWG, která zavádí sémantiku v rámci stávajícího obsahu webových stránek ve formě mikrodat, která používají podpůrné slovníky pro popis položek dvojice-hodnota k přiřazení hodnoty a její vlastnosti.

V budoucnu se dá očekávat také větší využití popisného GRDDL a komplexnějšího RDFa formátu, tak aby se postupně informace obsažené v mikrodatech a mikroformátech hladce začlenily do prostoru stávajícího Webu.

Mnoho sociálních sítí exportuje svá data jako ontologii FOAF, čímž dosud nejvýznamněji přispívají k rozvoji tzv. Linked Data [9], přičemž úspěch takových přístupů závisí do značné míry na tom, zda vývojáři webových aplikací mají zájem na zveřejnění dat a zda jsou ochotni používat standardy.

Protože ne všechny informace na Webu mohou být vždy veřejně přístupné, vznikl v rámci zajištění ochrany dat systém WebAccessControl [30], využívající autentizační protokol FOAF+SSL [39], který kombinuje již zavedenou ontologii s kryptografickým zabezpečením. Tento protokol je rychlejší a ve většině případů i uživatelsky přívětivější než rovněž aktuální OpenID [47] (uživatel si nemusí pamatovat jméno/id ani heslo, stačí pouze vlastnit osobní certifikát).

Právě využití ontologií při vývoji a následné optimalizaci Sémantického webu je jedním z vysoce aktuálních témat výzkumu. Je zřejmé, že jednotnosti (ať již formátové či názorové) se obecně nejsnáze dosáhne v určité uzavřené oblasti. Proto již dnes právě v souvislosti s aplikacemi Sémantického webu existuje poměrně velké množství samostatných doménových ontologií.

Otázkou však zůstává, zda pro úspěšnou realizaci komplexnějších informačních systémů v budoucnosti postačí rozšiřování počtu těchto příslovečných kamínků mozaiky Sémantického Webu, či zda bude nutné zapojit i svým charakterem obecnou, široce pojatou „všeobjímající“ ontologii.

Nejvýznamnější překážkou v tomto ohledu je patrně časová náročnost tvorby takového zdroje. K překonání tohoto slabého místa však může vést tradiční cesta a sice – pokusit se použít to, co je již hotovo. Cílem takového přístupu k budování ontologií může být integrace již existujících dokumentů, databází a dalších zdrojů, jejich „vyčištění“, zpřesnění uložených informací a nakonec integrace do jednotné ontologické struktury – obdobně jako Linked Data.

Jako jedno z výhledově masověji aplikovatelných řešení se jeví využití zpracování přirozeného jazyka a automatické určení kontextu textu, což by mohlo umožnit automatické generování RDF [27][49].V krátkodobém horizontu se však zatím širší podpora zejména ze strany majoritních vyhledávačů očekávat nedá, a to zejména vlivem chybějící motivace k implementaci technologií pro strojové zpracování informací. Stávající vyhledávače Sémantického Webu jsou prozatím zcela minoritními projekty a investice do tvorby jakkoli sémanticky uzpůsobených stránek je tak i z pohledu uživatele vcelku bezpředmětná.

Nedostatečná je také kvalita samotných vyhledávačů [57], které by dokázaly RDF dostatečně efektivně využít a nedostatek je také i samotných RDF dat a tedy v konečném důsledku i málo dostupných výsledků při vyhledávání.

V souvislosti s kritikou totoho přístupu při zpracování dokumentů je často diskutovaná ta skutečnost, že s využitím RDF dat se faktický problém reálného strojového porozumění textu dále jen odsouvá s odůvodněním, že vyhledávací stroje sice vědí, co o zdrojích tvrdí jejich autor, ale zatím nedokáží nerozlišit, co je opravdu reálným obsahem dokumentu, čímž se tak současně otevírá prostor pro otázky etického charakteru a o možném zneužití takového přístupu.

Vyhledávací stroj totiž na základě svého algoritmu poskytne vždy právě jen ten výsledek, který poskytnout má – je však na zvážení, jaké budou v případě Sémantického webu možnosti při ověřování jednotlivých zdrojů a zda nelze tímto způsobem manipulovat s realitou, někdo tak např. na poli vědeckých prací může více propagovat vlastní objevy, přičemž zde existuje jen velice slabé konkurenční prostředí, což dále snižuje možnosti kontroly. Nebo zda naopak v silně konkurenčním prostředí podniků a firem budou mít tyto subjekty vůbec zájem otvírat svá data (a tím i své „know-how“) veřejně celému světu.

Problém nastává také v případě, kdy data v prostředí Sémantického webu mají být reprezentována v několika jazykových mutacích. Google Translator [26] by však výhledově mohl být nástrojem, umožnující předklad v reálném čase nejen pro části dokumentů zobrazovaných uživateli, ale i metadat určených pro strojové zpracování. V úvahu také přichází a v praxi se již ojediněle vyskytují mnohojazyčné implementace informací v RDF [25], které však mnohdy neumožňují univerzálnější použití. Takové úpravy navíc častokrát neúměrně zvyšují velikost celého dokumentu, s čímž se ruku v ruce pojí i zvýšené nároky na administraci a problémy mohou nastat i vlivem zvýšeného datového přenosu.

Možnost praktického využití Sémantického webu je však i přesto značná. Za příklad možno uvést aplikace knihovních systémů, které na základě extrakčních technik dolování informací, spojených s následnou sémantickou anotací dat, pomohou ve specifických doménách vytvářet robustní a online dostupné specializované databáze znalostí. Navíc pokud se fyzicky změní lokační adresa zdroje (URL), je možno ji zcela automaticky najít a doplnit [7].

Význam knihoven se v tomto prostředí přímo nabízí. Kromě výše uvedeného vytváření exaktních databází znalostí a pojmů zde může jít i o generování rešerší na základě spolupráce takovýchto databází, což však předpokládá, že tyto služby by nebylo možné poskytovat pouze odděleně, ale muselo by se jednat o celkové jednotné rozhraní, k němuž se následně budou moci připojovat i další služby.

Názorným příkladem tohoto spojení může být kooperace knihovnického a e-learningového systému, případně nástroje pro univerzitní a vědecké účely jako např. automatický monitoring citací v publikacích, analyzování oblastí s největším potenciálem na základě zájmu odborné vědecké veřejnosti apod.

Výzkum v oblasti Sémantického webu zaznamenal za bezmála 15 let své existence pestrý vývoj. Původně spíše okrajové téma se postupně stávalo hojně diskutovaným oborem, který dokázal spojit i řadu odlišně zaměřených vědeckých komunit napříč nejrůznějšími zájmovými skupinami.

Nově vyvíjené technologie mají již dnes oporu v zavedených standardech (definovaných v rámci konsorcia W3C), stejně tak ale neustále vznikají návrhy na standardy nové. Jak ostatně i výzkum prezentovaný v této práci ukázal, reálné možnosti využití Sémantického webu se dají očekávat zejména v oblasti Linked Data, to ovšem pouze za předpokladu existence dostatečného množství sémanticky anotovaných a veřejně sdílených dokumentů, přičemž technologická podpora jejich životního cyklu je již dnes (od vystavení až po využití v aplikacích) poměrně rozsáhlá. Na druhou stranu ani u velké části dostupných softwarových nástrojů zatím nelze mluvit o zralosti a stabilitě odpovídající ostrému nasazení, přičemž dalšímu významnějšímu rozvoji těchto technologií v praxi brání také doposud neustálené procesy a ekonomické modely.

Perspektivním avšak stále ještě nedostatečně využívaným prostředkem pro demonstraci smysluplnosti Sémantického webu, jsou praktické aplikace v oblasti akademické sféry, která je dosud také jejich hlavním propagátorem. Univerzity a vědecké instituce mají k dispozici velké objemy dat, která jsou často na Webu nesystematicky zveřejňovaná, ať už v podobě statických stránek nebo výstupu z různorodých databází – jedná se především o dokumenty organizační, projektové či publikační (včetně kvalifikačních prací), ale často také o výukové materiály, prezentace či jiné učební pomůcky [15].

Tato data by tak následně mohla sloužit nejen pro pilotní testování nových technologií (což se již v praxi pravidelně děje), ale také pro dosažení a prokázání dlouhodobě udržitelného přínosu. Nad otevřenými daty by tak postupně mohly vzniknout 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. Samotné univerzity a vědecké instituce mají navíc, oproti omezeným grantovým projektům a „startupům“, nesrovnatelně delší dobu trvání, což by právě mělo zajistit také relativně vysokou perzistenci vystavovaných dat.

Pro celkový předpoklad širšího využití bude však bezpodmínečně nutná další praktická implementace, a to nejen ve zmíněné oblasti webových projektů z univerzitního či výzkumného prostředí, nýbrž také jako zamýšlené rozšíření v podmínkách široce využívaného Webu současného. Bez toho by Sémantický web zůstal jinak sice teoreticky velmi zajímavým, ale prakticky neuchopitelným konceptem s minimálními možnostmi reálného využití v celosvětovém měřítku.

POUŽITÁ LITERATURA A ZDROJE

[1] ADIDA, Ben, Mark BIRBECK a Ivan HERMAN. Semantic Annotation and Retrieval: Web of Hypertext – RDFa and Microformats. *Handbook of Semantic Web Technologies*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, s. 157. DOI: 10.1007/978-3-540-92913-0_5. Dostupné z: http://link.springer.com/10.1007/978-3-540-92913-0_5 .

[2] ADIDA, Ben, Mark BIRBECK, M., McCarron, S., Pemberton, S. *RDFa in XHTML: Syntax and Processing A collection of attributes and processing rules for extending XHTML to support RDF*. W3C Recommendation 14 October 2008. Dostupné z: http://www.w3.org/TR/rdfa-syntax .

[3] BAUMGARTNER, Peter; FURBACH, Ulrich; NIEMELÄ, Ilkka. Hyper tableaux. In: *Logics in Artificial Intelligence*. Springer, 1996. p. 1-17.

[4] BECHHOFER, Sean; VOLZ, Raphael; LORD, Phillip. Cooking the Semantic Web with the OWL API. In: *The Semantic Web-ISWC 2003*. Springer Berlin Heidelberg, 2003. p. 659-675.

[5] BECHHOFER, Sean. OWL: Web ontology language. In: *Encyclopedia of Database Systems*. Springer US, 2009. p. 2008-2009.

[6] BERNERS-LEE, Tim, James HENDLER, Ora LASSILLA. The Semantic Web. *Scientic American*, 2001, 284, str. 35-43. Dostupné z: http://www. sciam.com/2001/0501issue/0501berners-lee.html .

[7] BERNERS-LEE, Tim. W3C – Design Issues : *What the Semantic Web can represent, Semantic Web is not Artificial Intelligence.* 1998. Dostupné z: http://www.w3.org/DesignIssues/RDFnot.html .

[8] BERNERS-LEE, Tim. *Sir Tim Berners-Lee Talks with Talis about the Semantic Web Transcript of an interview* recorded on 7 February 2008. Dostupné z: http://talispodcasts.s3.amazonaws.com/twt20080207BL.html .

[9] BIZER, Chris, Tom, HEARTH, Tim, BERNERS-LEE. 2009. Linked Data The Story So Far. *International Journal on Semantic Web and Information Systems.* 2009, vol. 5, no.3, s. 1-22. Dostupné z: http://eprints.ecs. soton.ac.uk/21285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf .

[10] BOLEY, Harold, et al. Rule interchange on the web. In: *Reasoning web*. Springer Berlin Heidelberg, 2007. p. 269-309.

[11] BRICKLEY, Dan, et al. *RDF Vocabulary Description Language 1.0: RDF Schema W3C Recommendation 10 February 2004*. Dostupné z: http://www.w3.org/TR/2004/REC-rdf-schema-20040210/ .

[12] BURKE, Mary. The semantic web and the digital library. In: CORNELIUS, Ian (ed.). *Aslib Proceedings*. Emerald Group Publishing Limited, 2009. p. 316-322.

[13] CAMPINAS, Stéphane, et al. The Sindice-2011 dataset for entity-oriented search in the web of data. In: *Proceedings of the 1st International Workshop on Entity-Oriented Search (EOS)*. 2011. p. 26-32.

[14] CARROLL, Jeremy J., et al. Jena: implementing the semantic web recommendations. In: *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. ACM, 2004. p 74-83.
[15] ČERNÝ, Michal, et al. Sémantický web–jak dál?. *Ikaros* [online]. 2009, roč. 13, č. 5 [cit. 24.04.2014], urn:nbn:cz:ik-005437, ISSN 1212-5075. Dostupné z: http://www.ikaros.cz/node/5437 .
[16] DACONTA, Michael C.; OBRST, Leo J.; SMITH, Kevin T. *The semantic web: a guide to the future of XML, web services, and knowledge management*. John Wiley & Sons, 2003, ISBN 0-471-43257-1.
[17] DENTLER, K., R. CORNET. Comparison of Reasoners for large Ontologies in the OWL 2 EL Profile. *Semantic Web Journal*. 1-17. 2011. [online]. [cit. 2014-04-21]. Dostupné z: http://www.semantic-web-journal.net/sites/default/files/swj120_2.pdf .
[18] FABIÁNOVÁ Kamila. *Informační výchova na UTB ve Zlíně – Jak správně citovat*. [online]. [cit. 2014-04-25]. Dostupné z: http://iva.k.utb.cz/?page_id=34 .
[19] FEIGENBAUM, Lee; HERMAN, Thomas; HONGSERMEIER, E., NEUMANN, S., STEPHENS. *Scientific American*. The Semantic Web in Action. Vol. 297, December 2007, pp. 90-97. Dostupné z: http://thefigtrees.net/lee/sw/sciam/semantic-web-in-action .
[20] FERNANDEZ-LOPEZ, Mariano; CORCHO, Oscar. *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer, Incorporated, 2010.
[21] FORGY C., Rete: A fast algorithm for the many pattern/many object pattern match problem, *Artificial intelligence*, vol. 19, no. 1, pp. 17– 37.
[22] FOWLER, Martin. *Destilované UML*. Grada Publishing as, 2009, 1. vyd. Praha, Kapitola 2, Proces vývoje, s. 37-49, ISBN 978-80-247-2062-3.
[23] GANGEMI, Aldo. A comparison of knowledge extraction tools for the semantic web. In: *The Semantic Web: Semantics and Big Data*. Springer Berlin Heidelberg, 2013. p. 351-366.
[24] GENNARI, John H., et al. The evolution of Protégé: an environment for knowledge-based systems development. *International Journal of Human-computer studies*, 2003, 58.1: 89-123.
[25] GÓMEZ-PÉREZ, Asunción; EUZENAT, Jérôme (ed.). *The Semantic Web: Research and Applications: Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29--June 1, 2005, Proceedings*. Springer Science & Business Media, 2005.
[26] GOSRAVIZADEH, Parvaneh. Google Translation Semantic Structure. Dostupné z: http://www.google.com/intl/cs/help/faq_translation.html .
[27] GREGAR, Tomáš, et al. Využití nástrojů pro zpracování přirozeného jazyka v e-learningu. 2005, *sborník 2. ročníku konference o elektronické podpoře výuky*. 1. vyd. Brno, s. 45-50, ISBN 80-210-3699-0.

[28] GRUBER, T.R. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition 5*: 199-200. 2003.
[29] HAZAËL-MASSIEUX, Dominique; CONNOLLY, Dan. Gleaning Resource Descriptions from Dialects of Languages (GRDDL). *World Wide Web Consortium, W3C Coordination Group Note NOTE-grddl,* 2004.
[30] HARBULOT, Bruno. *FOAF & SSL: creating a global decentralised authentication protocol.* [online]. [cit. 2014-04-25]. Dostupné z: http://www.w3.org/2008/09/msnws/papers/foaf+ssl.html
[31] HORROCKS, Ian, et al. SWRL: A semantic web rule language combining OWL and RuleML. *W3C Member submission*, 2004, 21: 79.
[32] HUTCHINSON, David, et al. Advances in Web Semantics I: Ontologies, Web Services and Applied Semantic Web. Berlin, *Heidelberg*. 2009. Dostupné z: http://dx.doi.org/10.1007/978-3-540-89784-2 .
[33] JACOBS, Ian; WALSH, Norman. Architecture of the World Wide Web, Volume One. W3C Recommendation, December 2004. 2008. Dostupné z: http://www.w3.org/TR/webarch/ .
[34] JULIO, Arpírez C., et al. WebODE: a scalable workbench for ontological engineering. In: *Proceedings of the 1st international conference on Knowledge capture*. ACM, 2001. p. 6-13.
[35] KHAN, Javed Ahmad, et al. A performance evaluation of semantic based search engines and keyword based search engines. In: *Medical Imaging, m-Health and Emerging Communication Systems (MedCom), 2014 International Conference on*. IEEE, 2014. p. 168-173.
[36] KIFER, Michael. Rule interchange format: The framework. In: *Web reasoning and rule systems*. Springer Berlin Heidelberg, 2008. p. 1-11.
[37] KLYNE, Graham; CARROLL, Jeremy J. Resource description framework (RDF): Concepts and abstract syntax. W3C Recommendation 2006. Dostupné z: http://www.w3.org/TR/rdf-concepts/ .
[38] KNUBLAUCH, Holger, et al. The Protégé OWL plugin: An open development environment for semantic web applications. In: *The Semantic Web–ISWC 2004*. Springer Berlin Heidelberg, 2004. p. 229-243.
[39] KOLOVSKI, Vladimir; HENDLER, James; PARSIA, Bijan. Analyzing web access control policies. In: *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007. p. 677-686.
[40] LUTZ, Carsten; WOLTER Frank. Temporal Description Logics: A Survey. *Computational Logic Center*: 1-13. 2008
[41] MANOLA, Frank, et al. RDF primer. *W3C recommendation*, 2004, 10.1-107: 6. Dostupné z: http://www.w3.org/TR/rdf-primer/ .
[42] MCGUINNESS, Deborah L., et al. OWL web ontology language overview. *W3C recommendation*, 2004, 10.10: 2004.
[43] MERRIAM-WEBSTER [online]. 2012. [cit. 24.04.2014], Dostupné z: http://www.merriam-webster.com/dictionary/ontology .

[44] MORRISON, Michael. Hour 5: Putting Namespaces to Use. *Sams Teach Yourself XML*. Sams Publishing. p. 91. 2002.
[45] MOTTA, Yolanda Gil Enrico; MUSEN, V. Richard Benjamins Mark A. The Semantic Web–ISWC 2005.
[46] O'CONNOR, Martin J.; HALASCHEK-WIENER, Christian. M2: A Language for Mapping Spreadsheets to OWL. In: *OWLED*. 2010.
[47] OPENID FOUNDATION. *OpenID Specification* [cit. 2011-01-31]. Dostupné z: http://openid.net/specs/openid-authentication-2_0.html
[48] ORACLE Database 11g Semantic Technologies Overview. WU, Zhe. ORACLE. [online]. [cit. 2014-04-25]. Dostupné z: http://download. oracle.com/otndocs/tech/semantic_web/pdf/oow10_semtech_oraover.pdf
[49] PALMER, Sean B. *The Early History of HTML*. [cit. 2014-04-26], 2009.
[50] PEPPER, Steve. The TAO of topic maps. In: *Proceedings of XML Europe*. 2000. Dostupné z: http://ontopia.net/ .
[51] POPELÍNSKÝ, Lubomír. *Počítače a porozumění textu* [online]. Brno : Fakulta informatiky MU, 2007 [cit. 2009-04-30]. Prezentace. Dostupný z: http://www.fi.muni.cz/~popel/nll/czv06-final.ppt .
[52] PRUD, Eric, et al. Sparql query language for rdf. 2006 Dostupné z: http://www.w3.org/TR/rdf-sparql-query/ .
[53] RECTOR, A. L., et al. OpenGALEN: open source medical terminology and tools. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2003. p. 982.
[54] REPP, Martin. Goodrelations: An ontology for describing products and services offers on the web. In: *Knowledge Engineering: Practice and Patterns*. Springer Berlin Heidelberg, 2008. p. 329-346.
[55] SIRIN, Evren, et al. Pellet: A practical owl-dl reasoner. *Web Semantics: science, services and agents on the World Wide Web*, 2007, 5.2: 51-53.
[56] SYSTINET. *Vývoj Web Services*. Praktické ukázky technologií SOAP, WSDL a UDDI. 2001.
[57] ŠTENCEK, Jiří. *Průzkum portálů využívajících technologie sémantického webu* [online]. 2008 [cit. 2014-04-25]. Dostupné z: http://vse.stencek.com/semanticky-web/ch05s01.html
[58] W3C, SPARQL query language for RDF, W3C, W3C Recommendation, 2008, http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115.
[59] W3C, OWL 2 web ontology language document overview, W3C, Tech. Rep., 2009, http://www.w3.org/TR/2009/REC-owl2-overview- 20091027/.
[60] W3C, Uniform Resource Identifier (URI) Schemes. [online]. [cit. 2014-04-25]. Dostupné z: http://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml
[61] WILSON, Max, et al. mSpace: improving information access to multimedia domains with multimodal exploratory search. *Communications of the ACM*, 2006, 49.4: 47-49.

SEZNAM SYMBOLŮ A ZKRATEK

API – *Application Programming Interface*
ASPX – *Active Server Page Extended File*
CMS – *Content Management System*
CTM – *Compact Topic Maps*
DAML+OIL – *DARPA Agent Markup Language + Ontology Interface Layer*
EDM – *Electronic Document Management*
ETL – *Extract-Tranform-Load Compilator*
FOAF – *Friend Of A Friend*
GRDDL – *Gleaning Resource Descriptions from Dialects of Languages*
HMM – *Hidden Markov Model*
HTML – *Hypertext Markup Language*
HTTP – *HyperText Transfer Protocol*
IANA – *Internet Assigned Numbers Authority*
LGPL – *Lesser General Public License*
LTM – *Linear Topic Map Notation*
NER – *Named Entity Recognition*
NLU – *Natural Language Understanding*
OWL – *Ontology Web Language*
RDF – *Resource Description Framework*
RDFa – *Resource Description Framework in attributes*
RDFS – *Resource Description Framework*
REST – *Representational State Transfer*
RSS – *Rich Site Summary*
RIF – *Rule Interchange Format*
SIOC – *Semantically-Interlinked Online Communities*
SOAP – *Simple Object Access Protocol*
SPARQL – *SPARQL Protocol and RDF Query Language*
SQL – *Structured Query Language*
UDDI – *Universal Description, Discovery and Integration*
URI – *Uniform Resource Locator*
URL – *Uniform Resource Locator*
URN – *Uniform Resource Name*
WSDL – *Web Services Description Language*
WWW – *World Wide Web*
WHATWG – *Web Hypertext Application Technology Working Group*
WYSIWYG – *What You See Is What You Get - editor*
XKMS – *XML Key Management Specification*
XML – *eXtensible Markup Language*
XTM – *International Better Translation Technology*

SEZNAM OBRÁZKŮ

Obr. 3.1: Schématická struktura vrstev Sémantického webu 12
Obr. 3.2: Sémantické spektrum v závislosti na užité technologii zápisu 40
Obr. 3.3: Eulerův diagram vztahů mezi URI identifikátory 49
Obr. 3.4: RDF Graf s reprezentací množiny uzlů 18
Obr. 3.5: Vzorové RDF Schema s definicí vlastností tříd a podtříd..... 20
Obr. 3.6: Schéma aplikace využívající dotazovací engine SPARQL..... 26
Obr. 3.7: Příklad konstrukce doménové ontologie v organizaci 29
Obr. 3.8: Ontologická struktura ve standardu OWL 2..... 30
Obr. 3.9: Role RIF pravidel v logickém procesu vyhledávání výsledků..... 78
Obr. 3.10: Extrakce informací v prostředí otevřených domén 84
Obr. 3.11: Proces učení ontologií při zpracování dotazů v přirozeném jazyce .. 37
Obr. 3.12: Centralizované zpracování informací znalostního managementu..... 39
Obr. 3.13: Vztah základních technologií webových služeb 40
Obr. 3.14: Schéma procesu zpracování jmenných entit..... 46
Obr. 3.15: Protégé-Frames pro správu tříd a jejich instancí..... 48
Obr. 3.16: Protégé-OWL pro vizualizaci ontologií a inferenčních modelů 48
Obr. 3.17: Role softwarového nástroje Protégé při tvorbě znalostní báze 49
Obr. 3.18: Příklad GoodRelations ontologie pomocí RDF Grafu 50
Obr. 3.19: Funkční schéma softwarového nástroje Ontopia..... 51
Obr. 3.20: Základní architektura Anzo EDM serveru 52
Obr. 3.21: Proces vyhledání dotazu v prostředí mSpace..... 53
Obr. 3.22: Schéma funkčních principů webového vyhledávače Swoogle..... 55
Obr. 3.23: Open Calais v procesu zpracování nestrukturovaného obsahu 56
Obr. 3.24: Ukázka obecně dotazu zpracovaného nástrojem Kngine..... 57
Obr. 3.25: Příklad SPARQL dotazu v prostředí indexu Sindice 58
Obr. 3.26: Řešení slovní disambiguace ve vyhledávači DuckDuckGo..... 59
Obr. 4.1: Hlavní strana informačního portálu SemanticWeb.cz..... 69
Obr. 4.2: Modifikace struktury DOM stromu..... 75
Obr. 4.3: Schematický návrh sémantického anotačního procesu 77
Obr. 4.4: Ukázka sémantické anotace v prostředí CMS WP.NET 3.8.1 78
Obr. 4.5: Ukázka části struktury zkoumané doménové ontologie..... 80
Obr. 4.6: Schéma procesu zpracování dotazu v přirozeném jazyce 82
Obr. 4.7: Rozdělení a úprava dotazu na strojově čitelný zápis..... 83
Obr. 4.8: Příklad vyhodnocení vstupního dotazu v přirozeném jazyce..... 84

SEZNAM TABULEK A GRAFŮ

Tabulka 1: Schématická struktura vrstev Sémantického webu 21
Tabulka 2: Přehled standardizovaných mikroformátů..... 23
Tabulka 3: Nově zaváděné koncepty mikroformátů..... 23
Tabulka 4: Úvodní rozdělení respondentů..... 63
Tabulka 5: Počty respondentů v závislosti na informačních zdrojích..... 64
Tabulka 6: Překážky rozvoje Sémantického webu v praxi..... 65
Tabulka 7: Praktický přínos technologií Sémantického webu 66
Tabulka 8: Praktické využití sémantických vyhledávačů..... 67
Tabulka 9: Potřeba komplexního zpracování informací..... 68
Tabulka 10: Přehled funkčních parametrů sémantických reasonerů 88

Graf 4.1: Počet respondentů dle obecného povědomí o zkoumané oblasti..... 63
Graf 4.2: Prvotní informační zdroje o problematice Sémantického webu 64
Graf 4.3: Možná omezení aktivního vývoje Sémantického webu..... 65
Graf 4.4: Hlavní výhody Sémantického webu dle dotazovaných respondentů.. 66
Graf 4.5: Zkušenost se sémantickým vyhledáváním informací..... 67
Graf 4.6: Zpracování souhrnného přehledu v českém jazyce..... 68
Graf 4.7: Odezva na vyhodnocení požadavku v závislosti na čase 91
Graf 4.8: Množství využité operační paměti v závislosti na čase..... 91

PUBLIKAČNÍ ČINNOST AUTORA

Články ve sbornících mezinárodních konferencí:

ŠMIRAUS, Michal, JAŠEK Roman. Query answering under evolving ontologies with mapping redefinition. In Annals of *CER International Scientific conference 2015*. Sciemcee Publishing, London, ISBN 978-0-9928772-2-4.

ŠMIRAUS, Michal, BUDÍKOVÁ Věra. Data Integration with Evolving Ontologies. In Annals of *SCIECONF international conference 2013*. Scientific Conference, 2013, 1, ISBN 978-80-554-0726-5.

ŠMIRAUS, Michal, JAŠEK Roman. Risks of Advanced persistent threats and defense against them. In Annals of *DAAAM for 2011 & Proceedings of the 22nd International DAAAM Symposium* "Intelligent Manufacturing & Automation: Power of Knowledge and Creativity". Vienna : DAAAM International Vienna, 2011, 22, ISBN 978-3-901509-83-4.

ŠMIRAUS Michal, DRGA Rudolf. Software pro dohledová řídící centra a jejich další vývoj. Mezinárodní konference Bezpečnostní technologie Systémy a Management. 2011, 1, ISBN 978-80-87500-05-7.

Články v recenzovaných časopisech:

ŠMIRAUS, Michal. Information processing across ontologies based on semantic mapping. *Trilobit : odborný vědecký časopis*. 2013, 2, s. 1-8. ISSN 1804-1795.

ŠMIRAUS, Michal. Současné trendy a směr vývoje modelové architektury webových vyhledávacích strojů. *Trilobit : odborný vědecký časopis*. 2011, 3, 1, s. 1-5. ISSN 1804-1795.

ŠMIRAUS, Michal. Evaluation model for ontology-based information extraction [online]. [cit. 2013-08-25]. *Posterus*. ISSN 1338-0087. Dostupné z: http://www.posterus.sk/?p=16157

ŠMIRAUS, Michal. Natural Language Processing with Semantic Web Search ontologies. *IOS Press Semantic Web journal*. 2014, 4, s. 1-10. 1570-0844.

ŠMIRAUS, Michal. Přístup k sémantickému webu založený na ontologiích [online]. [cit. 2013-05-20]. *Posterus*. ISSN 1338-0087. Dostupné z: http://www.posterus.sk/?p=15724

PROFESNÍ ŽIVOTOPIS

Osobní informace:

Jméno a příjmení: Ing. Michal Šmiraus
Datum narození: 9. prosinec 1985
Email: smiraus@fai.utb.cz
Bydliště: Zlín
Státní příslušnost: ČR

Vzdělání:

2001 – 2005
Střední průmyslová škola elektrotechnická Zlín
Maturitní zkouška: Elektronika, Mikroprocesorová technika
Anglický jazyk, Český jazyk

2005 – 2008
Univerzita Tomáše Bati ve Zlíně, Fakulta aplikované informatiky
Bakalářské studium
Studijní obor: Bezpečnostní technologie, systémy a management
Získaný titul: Bc.

2008 – 2010
Univerzita Tomáše Bati ve Zlíně, Fakulta aplikované informatiky
Magisterské studium
Studijní obor: Bezpečnostní technologie, systémy a management
Získaný titul: Ing.

2010 – 2014
Univerzita Tomáše Bati ve Zlíně, Fakulta aplikované informatiky
Doktorské studium
Studijní obor: Inženýrská informatika

Projekty:

2012 – 2013
FRVŠ / MŠMT č. 504: Inovace předmětu Matematika II o funkcích více proměnných webovou prezentací s grafikou i animacemi a využitím pro univerzity s inženýrskými obory, *spoluřešitelé:* RNDr. Miloslav Fialka, doc. Zdenka Prokopová, Ing. Hana Charvátová, Ph.D.

Pedagogická činnost:

A1ZPT – Základy počítačové techniky

A9EPW – Elektronická příprava dokumentů a www stránek

A5EBS – Elektronické zabezpečovací systémy

A8NPB – Nadstandartní prvky objektové bezpečnosti

Odborná praxe:

2010 – 2011
CREALOGIX AG, Bern – Švýcarsko
automatizované testování bezpečnosti a optimalizace uživatelské použitelnosti internetového bankovnictví

2011 – 2012
AUDI AG, Ingolstadt – Německo
aplikační a konfirmační testování navigačního software pro palubní počítače koncernových automobilů

2013 – 2014
EXPERCASH GmbH, Mannheim – Německo
automatizované regresní testování nově zaváděných komponent virtuální platební brány

Odborné zájmy:

Redakční systémy CMS (Wordpress, Joomla, Drupal)
E-commerce systémy (Opencart, Prestashop, Magento)
SEO & SEM optimalizace pro webové vyhledávače

Školení & certifikáty:

ISTQB: International Software Testing Qualifications Board – foundation level
CEH: Certified Ethical Hacking course – EC Council
CQS: PHP Applications Security Testing – ICT PRO
CISCO: Certified Entry Networking Technician
StrictProfessionals: Test driven development (TDD) a Agilní metody vývoje