Doktorandské dny ’07

Ústav informatiky
Akademie věd České republiky
v.v.i.

Malá Úpa

17.– 19. září 2007

vydavatelství Matematicko-fyzikální fakulty
University Karlovy v Praze

Ústav Informatiky AV ČR v.v.i., Pod Vodárenskou věží 2, 182 07 Praha 8

ISBN *– not yet –*

Obsah

Pavel Tyl: **Problematika integrace ontologií** **1**

Problematika integrace ontologií

doktorand:
ING. PAVEL TYL

Ústav informatiky AV ČR v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8

Fakulta mechatroniky
Technická univerzita Liberec
Hálkova 6
461 17 Liberec 1

pavel.tyl@tul.cz

školitel:
ING. JÚLIUS ŠTULLER, CSC.

Ústav informatiky AV ČR v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8

stuller@cs.cas.cz

obor studia:
Technická kybernetika

Práce byla částečně podpořena výzkumným centrem 1M0554 Ministerstva školství, mládeže a tělovýchovy České republiky: Pokročilé sanační technologie a procesy, projektem 1ET100300419 programu Informační společnost (Tématického programu II – Národního programu výzkumu v ČR: Inteligentní modely, algoritmy, metody a nástroje pro vytváření sémantického webu) a výzkumným záměrem AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications".

Abstrakt

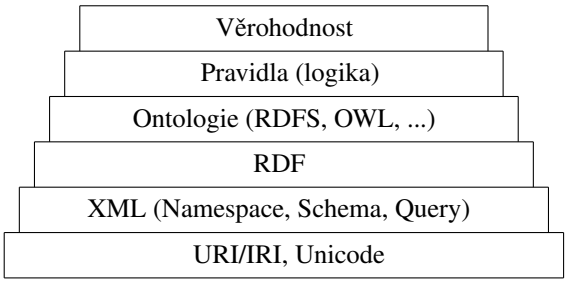
Internet je ohromným zdrojem provázaných, ale většinou neuspořádaných dat. Sémantický web, jako rozšíření webu současného, se snaží tuto neuspořádanost řešit a to nejen bezprostředně pro lidského uživatele, ale zejména z hlediska možnosti strojového zpracování informací. Cílem je doplnit data o metadata, která mají být srozumitelná jak pro člověka, tak pro počítač. Tato metadata jsou nejčastěji vyjádřena pomocí ontologií, které jsou jedním ze základních stavebních prvků sémantického webu. V příspěvku se snažím nastínit některé z možností integrace (slučování) ontologií za účelem sdílení informací.

1. Úvod

Internet je pozoruhodným informačním zdrojem. Svoboda, rozšířenost a téměř všudypřítomnost Internetu je ale zaplacena neuspořádaností většiny z něho dostupných informací, které jsou navíc velmi často bez logických návazností a vztahů. Hledané konkrétní informace nám tak mnohdy zůstanou skryty. Bez potřebné provázanosti zůstanou informace skrze svá podpůrná data obtížně vyhledatelné i pro stroje, které by jinak byly schopny nalezené výsledky dále zpracovávat pro různé skupiny uživatelů. Pokud by data na webu byla rozšířena o jejich význam, otevřelo by to široké možnosti při jejich sdílení, vyhledávání a znovupoužití. Tuto myšlenku se snaží rozvíjet vize sémantického webu [I4].

2. Sémantický web

Tvůrce webu Tim Berners-Lee říká, že sémantický web není separátním webem, nýbrž je rozšířením webu současného. Sémantický web přiřazuje datům na webu přesný význam umožňující spolupráci lidí a softwaru [2]. V tomto ohledu se chová jako informační systém, který má usnadnit cestu informace od jedné osoby ke druhé. Dnes se web dynamicky vyvíjí zejména jako zprostředkovatel dokumentů pro lidského uživatele. Sémantický web se snaží naopak vyzdvihnout automatické zpracování dat a informací pomocí počítačů a umožnit tak provoz inteligentních služeb. Aby mohl sémantický web vůbec fungovat, je třeba, aby počítače měly přístup ke strukturovaným souborům dat a zároveň srozumitelná pravidla k provádění automatických operací s těmito daty [1]. Na obrázku 1 je naznačena struktura sémantického webu.



Obrázek 1: Vrstvy sémantického webu

Jak je vidět, tak pro vývoj sémantického webu jsou důležité mnohé technologie. Pod zkratkou RDF [I2] si můžeme představit model pro reprezentaci dat uložených v jednotlivých zdrojích na webu. Zatímco XML [I5] umožňuje uživatelům vytvářet vlastní struktury dokumentů, ale neříká nic o jejich významu, RDF umožňuje zachycení významu, a to v podobě trojic objekt–atribut–hodnota (podmět–přísudek–předmět). Konkrétní věci (lidé, webové stránky, tabulky nebo cokoliv jiného) mají určité vlastnosti (atributy, predikáty – například být synem), které pak nabývají jistých hodnot (jiná osoba, jiná webová stránka atd.). Objekt, atribut i hodnota mohou být identifikovány pomocí URI či IRI (Internationalized Resource Identifier – URI s možností použití libovolného kódování, např. českého). RDF trojice vytvářejí pavučiny informací o souvisejících věcech. URI umožňují, že koncepty nemusejí být pouhými slovy v dokumentu, ale mohou být provázány na unikátní definici, kterou si každý může na webu najít. Na webu nejčastěji používaná forma zápisu RDF je pomocí XML [4].
Za těchto předpokladů je ovšem stále možné, že například dvě rozdílné webové databáze budou používat různé identifikátory příslušející stejnému konceptu. Proto je nutný další ze základních kamenů sémantického webu, konkrétně ontologie.

3. Ontologie

Podle jedné z definic je ontologie *formální specifikace sdílené konceptualizace*. *Konceptualizací* je myšlen abstraktní model výseku reálného světa, který popisuje relevantní koncepty daného výseku. Slova *formální* a *sdílené* mají důležitý význam ke (znovu-)použitelnosti ontologií, protože základním předpokladem jejich opakované (počítačové) použitelnosti je jejich formální vyjádřitelnost a možnost jejich sdílení; pokud by kteroukoli z těchto dvou vlastností postrádaly, byly by zřejmě k ničemu. Ontologie je tedy určitým systémem zachycení reality, který je znovupoužitelný a je možné ho sdílet.

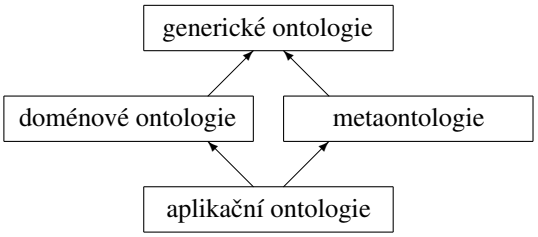
3.1. Meta model ontologie

Pod tímto pojmem si můžeme představit popisné a odvozovací schopnosti modelu. Jde o formální definice toho, co ontologie může obsahovat, co jsou uzly, co vazby, jaké typy vztahů připouští, jak je možné specifikovat pravidla a funkce apod. Každá ontologie má svůj model, více ontologií ale může být vystavěno podle stejného meta modelu.

3.2. Klasifikace ontologií

Ontologie lze obecně rozdělovat podle různých kritérií. První z možností je rozdělit veškeré ontologie do dvou skupin, *ontologie popsané formálním jazykem* a *ontologie v přirozeném jazyce* [5]. A protože samotný přirozený jazyk poskytuje mnoho různých prostředků konceptualizace, je zpracování ontologií popsaných přirozeným jazykem složité a provádí se většinou na lingvistické úrovni nebo se převádí na jiný (jednodušší) typ.
Druhou a nejčastěji používanou možností je dělit ontologie podle zdroje konceptualizace (viz obrázek 2):

- **generické ontologie** (*ontologie vyššího řádu*) -– zachycování obecných zákonitostí, (mohou sloužit také jako prostředek pro spojení jednotlivých doménově specifických ontologií a tak pomoci k jejich širší integraci),
- **doménové ontologie** (*doménově specifické ontologie*) — určeny pro specifickou věcnou oblast (nejčastější; např. pro oblast sportu, hudby atd.),
- **úlohové ontologie** (*reprezentační ontologie* či *metaontologie*) — zaměřeny na procesy odvozování,
- **aplikační ontologie** — adaptovány na konkrétní aplikaci, (nejspecifičtější; zpravidla zahrnují doménovou i úlohovou část).



Obrázek 2: Druhy ontologií a jejich vztahy z pohledu konceptualizace

V dalším textu se pod pojmem ontologie uvažují zejména aplikační a doménové ontologie.

3.3. Jazyk ontologií

Jedná se o jazyk, který se používá pro reprezentaci ontologií. Nejčastěji používanými jazyky jsou RDFS [I3] a OWL [I1]. Jazyk ontologií pro sémantický web se skládá ze dvou částí, logické a mimologické. Logická část se obvykle skládá z *axiomů* pro *definici tříd*, *vlastností*, *instancí* atd. Prvky mimologické části jsou většinou vlastnosti, které se netýkají funkčnosti

(*jméno autora*, *datum vytvoření*, *komentáře*, ale i *deklaraci jmenných prostorů* či *import dalších ontologií*). Mimologická část je určena především pro lidi, přestože množství výše uvedených vlastností je strojově zpracovatelné (příkladem jsou jmenné prostory nebo import ontologií: ten může být proveden buď pomocí přidání logické části importované ontologie do logické části ontologie, do níž importujeme, vytvoříme tak jeden logický popis, nebo použitím jakéhosi prostředníka, který řeší nestejnorodost dvou ontologií).

3.4. Využití ontologií

Agregace, integrace, unifikace
Jak již bylo zmíněno, Internet je prostoupen informacemi ve všemožné podobě, struktuře a kvalitě. Ontologie by mohly být prostředkem propojení a následné agregace takových heterogenních zdrojů. Databáze, které obsahují cenná data, by mohly sloužit ještě mnohem lépe v integrovaném celku. Ontologie by se mohly stát jádrem systému, prostředkem pro kompozici nezávislých webových služeb.

Snížení redundance
Přestože již mnohokrát vytvořené, nashromážděné, zpracované, ověřené a porovnané informace jsou znovu a znovu vytvářeny, shromažďovány, zpracovávány, ..., zvyšuje se jejich redundance, která může vést až k nekonzistenci, když si duplikovaná data vzájemně protiřečí. S použitím ontologií by mohla být data místo duplikace sdílena, a tak by redundance i nekonzistence mohla klesnout, mohla by být lépe kontrolována či úplně eliminována.

Znovupoužití
Konceptualizovaná data je mnohem snažší použít, a to i vícekrát a různými způsoby.

4. Integrace ontologií

Existující ontologie se hodí jako zdroje znalostí pro vytváření ontologií nových: ontologie mohou být převáděny a slučovány tak, aby k nim bylo možné přistupovat jako k jednomu většímu celku. Výsledkem je nová ontologie.
Se systémy a daty integrovanými pomocí ontologií se zvýší možnosti interoperability. Současným aplikacím schází především možnost budovat z nich kompaktní celky a poskytovat společně realizované služby pro uživatele. Slovo kompaktní v tomto případě neznamená monolitické, ale spíše poskládané z mnoha nezávislých komponent, které jsou překryty jednotící vrstvou.

Je potřeba rozlišovat několik operací či činností, které je možné s ontologiemi provádět:

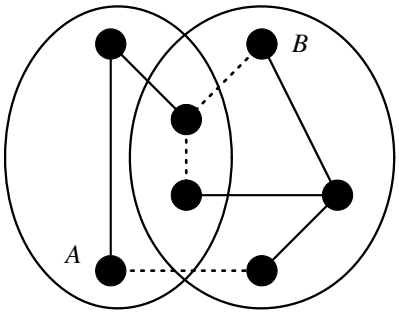
Transformace ontologií
Může být dvojího druhu:

- *meziformátová* – mezi jazyky pro zachycení ontologií (*RDF* → *OIL*),
- *sémantická* – změna vnitřní struktury podle jiného metamodelu nebo pro jiné použití.

Vývoj ontologií
Vývojem ontologií myslíme jejich *údržbu*, *doplňování nových konceptů*, *slaďování se současnými poznatky* o doméně nebo o ontologiích.

Spojování ontologií (*Ontology Merging*)
Výsledkem této operace je jedna nová ontologie, která zahrnuje informace ze dvou či více ontologií. Integrovaná ontologie je již nezávislá na ontologiích původních, které v podstatě nahradí.

Integrace ontologií (*Ontology Alignment*)
Integrace ontologií má význam především tam, kde se očekává budoucí rozvoj a údržba spojovaných ontologií.
Více se zde zajímáme o dvojice ontologií, které se určitým způsobem překrývají, a kdy spolu některé jejich elementy více či méně souvisejí. Výsledkem integrace dvou ontologií *A* a *B* jsou stále dvě ontologie (nové), ale s definovanými společnými místy a přesahujícími vztahy, jak ukazuje obrázek 3. Snahou je, aby nesouvisející elementy byly ponechány stranou tak, aby nedošlo k porušení struktury ontologií.



Obrázek 3: Graficky znázorněná integrace ontologií

Zde je možné rozlišit dva typy situací, jak odlišit dvojice ontologií:

- každá z ontologií popisuje *odlišnou doménu* – tyto ontologie mohou být spojeny do jedné "superontologie" přes společné části, jsou-li takové, nebo přes nějakou obecnější ontologii [6].
- obě ontologie popisují *stejnou doménu*, ale z *různých úhlů* pohledu nebo *různými prostředky* – v tomto případě musí být provedeno *srovnání ontologií* za účelem vytvoření překrytí odlišností ontologií.

Srovnáním ontologií (*Ontology Matching*) se nazývá proces nacházení podobností mezi dvěma zdrojovými ontologiemi. Výsledkem je specifikace těchto podobností, která slouží jako vstup tzv. *mapování* (viz níže). Srovnání ontologií je věnována samostatná část 5.

Mapování ontologií (*Ontology Mapping*) je deklarativní specifikací sémantického překrytí mezi dvěma ontologiemi O a O'. Shody mezi odlišnými entitami jsou typicky vyjádřeny použitím axiomů formulovaných v "mapovacím" jazyce (jazyk k reprezentaci mapování ontologií). Mapování může být *jednosměrné* (specifikuje, jak termy z jedné ontologie mohou být vyjádřeny použitím termů ontologie druhé) či *obousměrné* (funguje oběma směry).

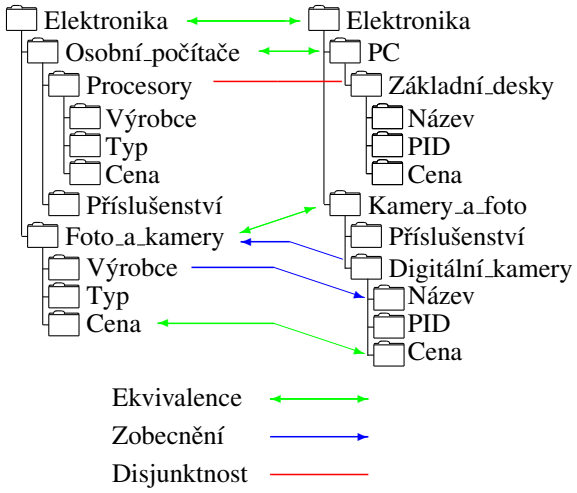
Integrace ontologií specifikuje, jak spolu ontologie souvisí v logickém smyslu. To znamená, že se původní ontologie nezmění, ale vzniknou další axiomy popisující vztahy mezi jejich koncepty. Ponechání původních ontologií v nezměněné podobě většinou znamená, že lze zintegrovat ontologie pouze částečně, neboť hlavní odlišnosti by vedly k nutné úpravě vstupních ontologií. Integrace ontologií je tedy určitým zobecněním mapování – *dvě ontologie mohou být zintegrovány pomocí mapování* [6].

4.1. Problémy při integrací ontologií

Problémy mohou nastat v mnoha případech. Třeba v tom, že tvůrci ontologií neuvažují stejně a vzájemně si leckdy neporozumí. Jedna ontologie může například reprezentovat červenou barvu jako *vztah*, druhá jako *hodnotu*. Přitom zvolená reprezentace je v rámci ontologie vždy *správná* a *pravdivá* -– správná je z definice, neboť jde o definici. Další potíže jsou na *jazykové úrovni*. To může komplikovat proces automatické integrace, protože je složité zjistit, zda jsou dva uzly (podíváme-li se na ontologii jako na graf) *stejné*, *podobné* nebo *zcela odlišné* [3].

5. Metody řešení srovnání ontologií

Předpokládejme, že máme dvě ontologie, z nichž každá se skládá z množiny entit (elementů, relací, tříd, vlastností atd.). Ty jsou v tomto případě vstupem pro srovnání. Výstupem pak budou vztahy (*ekvivalence*, *subsumpce*, neboli podřazení, *disjunktnost* atd.). Pro zjednodušení můžeme srovnání ontologií přirovnat ke srovnání XML schémat, jak ukazuje obrázek 4.



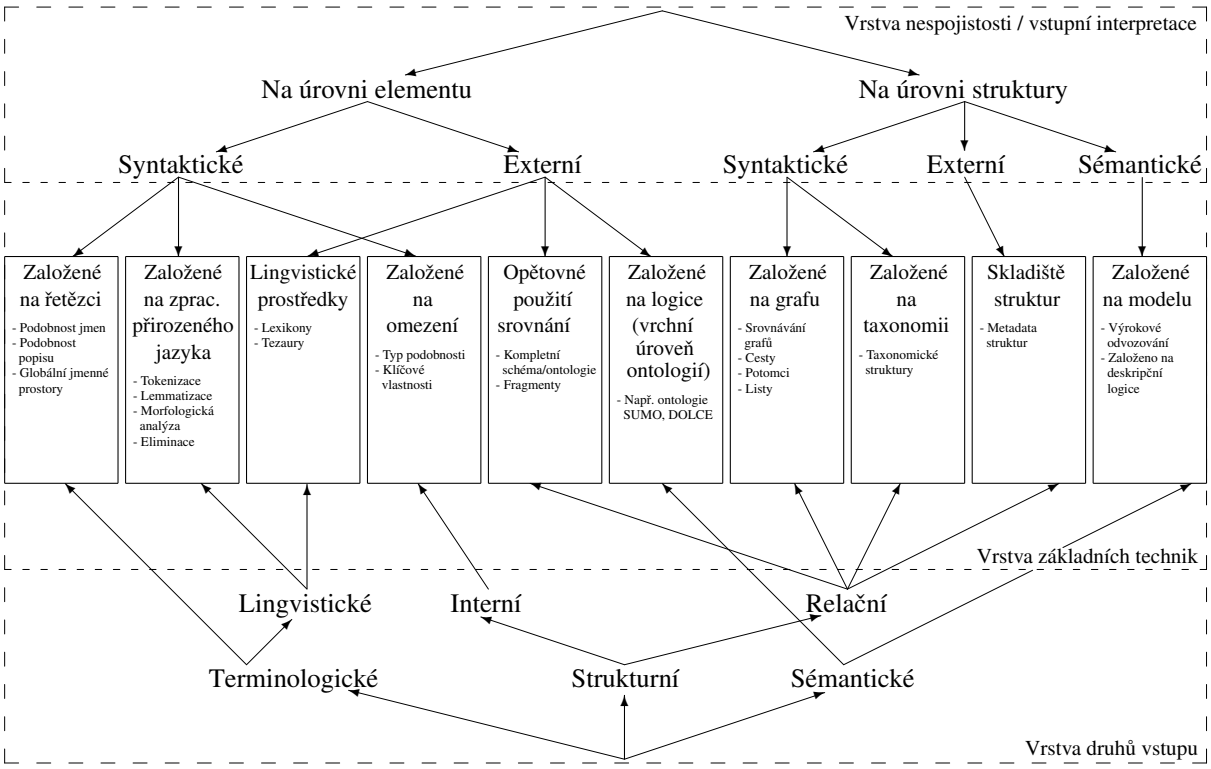
Obrázek 4: Ukázka možného srovnání dvou XML schémat

Obrázek 4 ukazuje možné vztahy srovnání dvou XML schémat. Od srovnání ontologií se však v některých aspektech odlišuje. Schémata často neposkytují explicitní sémantiku pro svá data. Ontologie, jako logické systémy, se omezují na význam. Ontologické definice jsou množiny (logických) axiomů. Ontologie a schémata mají ale i společné rysy. Oboje mají své slovníky pojmů, které popisují oblast zájmu (doménu) a oboje zároveň vymezují význam těchto pojmů.
Nestejnorodost schémat či ontologií se redukuje ve dvou základních krocích:

1. *vymezení* (viz obrázek 4),
2. *zpracování* (transformace, spojení, ...).

Máme-li dvě ontologie (schémata) O a O', je srovnáním mezi O a O' množina odpovídajících si prvků, trojic $\langle e, e', r \rangle$, kde $e \in O$ a $e' \in O'$ a r je vztah mezi e a e' (*ekvivalence*, *zobecnění*, *disjunktnost*). Na obrázku 5 je zobrazeno rozdělení metod pro srovnání na základě schématu. Jednotlivé metody vyžadují alespoň stručný popis:

- **Metody založené na řetězci** – Pracují s předponami (resp. příponami) slov, kdy jsou vstupem dva řetězce a kontroluje se, zda první řetězec začíná (resp. končí) druhým řetězcem.
  Např.: *hotel* $\rightarrow$ *hot*



Obrázek 5: Klasifikace metod srovnání ontologií na základě schémat [7]

Dále je možné určovat počet stejných N-gramů (počet N-tic písmen, které mají dva řetězce společné) či vzdálenost dvou řetězců.
Např.: *Nokia* $versus$ *Nka*.

- **Metody založené na zpracování přirozeného jazyka** — Využívají analýzy přirozeného jazyka. *Tokenizace* je rozdělení textu na jednotlivé slovní tvary (tokeny).
Např.: *foto-aparát*.
Lemmatizace je analýza tokenů pro zjištění všech základních forem slov.
Eliminací odstraníme "bezvýznamná" slova.

- **Lingvistické prostředky** — Zabývají se významem slov, na tomto principu funguje třeba WordNet.
Např.: $A \sqsupseteq B$, neboli A *je hypernymem nebo holonymem* B, konkrétně *Evropa* $\sqsupseteq$ *Řecko*.

- **Metody založené na omezeních** — Metoda srovnání datových typů.
Např.: *integer* $\subset$ *real* nebo *datum* $\in$ *[1.4.2007, 30.6.2007]* $\subset$ *datum[year = 2007]*.

- **Znovupoužití srovnání** — Potřebujeme-li provést srovnání schématu/ontologie O' a O'' a již máme srovnání mezi O a O' a zároveň O a O'', využijeme ho.

- **Metody založené na taxonomii** — Na schémata/ontologie se díváme jako na grafy obsahující termy a vztahy mezi nimi. Například pokud se *shodují koncepty vyšší úrovně*, *aktuální koncepty se podobají*.

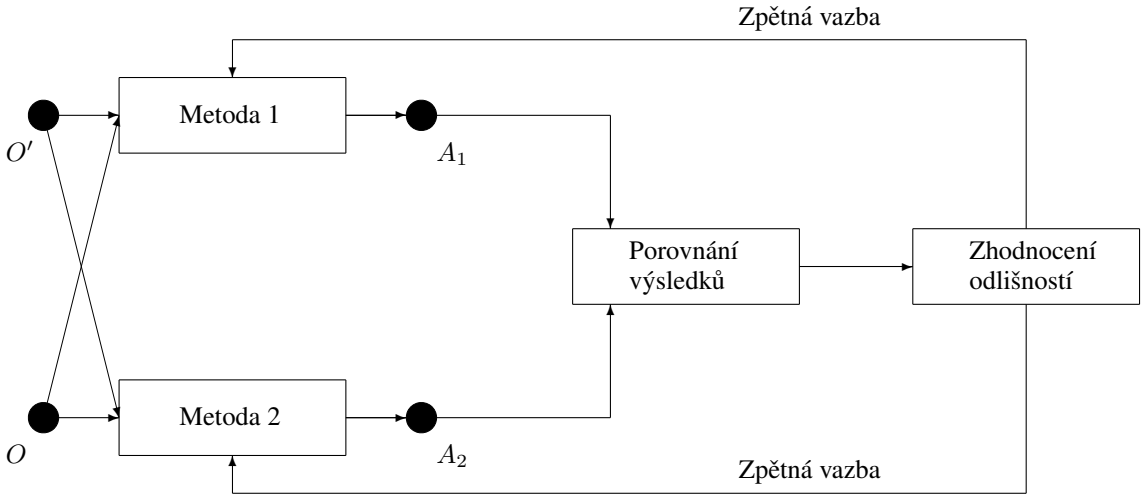
- **Metody založené na grafu** — Elementy dvou nelistových schémat jsou *strukturou podobné*, pokud jsou *množiny přímých potomků podobné* nebo pokud jsou *podobné jejich listové množiny*, i když množiny jejich přímých potomků nejsou. Jestliže *dva uzly dvou schémat/ontologií* jsou *podobné*, jejich *sourozenci mohou být rovněž podobní*.

- **Metody založené na modelu** — Převedeme srovnání grafu (stromu) na *srovnání množiny jeho uzlů*. Vytvoříme páry uzlů, které *spolu mohou souviset* a vztahy mezi nimi zapíšeme *výrokovými formulemi*. Poté kontrolujeme *platnost* jednotlivých formulí.
Např.: $(Elektronika_1 \Leftrightarrow Elektronika_2) \wedge (Osobní\_počítač_1 \Leftrightarrow PC_2) \Rightarrow (Elektronika_1 \wedge Osobní\_počítač_1 \Leftrightarrow Elektronika_2 \wedge PC_2)$.

5.1. Návrh spojení metod srovnání ontologií

Každá z metod uvedených v předchozí části má svoje omezení a svoji chybovost. Skutečně ideálním řešením by mohlo být vyvinout nástroj, který by podle typu ontologie využil více metod srovnání najednou. Jednotlivým metodám by šlo dávat váhy a metody by zároveň spolu mohli spolupracovat tak, aby jedna eliminovala nedostatky druhé. Schéma takového nástroje je naznačeno na Obrázku 6.

Srovnáním ontologií O a O' pomocí *metody 1* vznikne A_1. Dále srovnáním ontologií O a O' pomocí *metody 2* vznikne A_2. V ideálním případě, pokud by byly metody dokonalé, by $A_1 = A_2$. Tato situace je však málo pravděpodobná. V tom případě se provede porovnání A_1 a A_2. Tím zjistíme odlišnosti a rozdíl obou výsledků potom slouží jako zpětná vazba pro případnou úpravu *metody 1* a *2*. Části výsledku, které se naprosto liší by šlo potom ze srovnání úplně vynechat. Této problematice bych se chtěl v průběhu svého dalšího studia věnovat.



Obrázek 6: Návrh schématu spojení dvou metod srovnání ontologií

6. Závěr

Ontologie mohou v mnohém vylepšit fungování webu. V nejjednodušším případě se může jednat např. o přesnost vyhledávání, kdy se vyhledávač může zaměřit jen na stránky odpovídající danému konceptu (a nikoli dvojznačným nebo dokonce víceznačným klíčovým slovům). Jejich integrací navíc docílíme toho, že informační zdroje budou pro uživatele působit kompaktnějším dojmem. Úspěšné integraci napomáhají do určité míry v textu uvedené metody srovnání ontologií. Tyto metody by ji mohli dále vylepšovat, ale jejich skutečná síla by se mohla projevit, když jejich jednotlivé přednosti spojíme, popřípadě využijeme pro jejich úpravu poznatků, v čem se výsledky po srovnání ontologií liší.

Literatura

[1] ANTONIOU, Grigoris -– VAN HARMELEN, Frank. "A Semantic Web Primer". London: The Mit Press, 2004. ISBN 0-262-01210-3.

[2] BERNERS-LEE, Tim – HENDLER, James – LASSILA, Ora. "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities". *Scientific American*, vol. 284, 5, pp. 35–43. May 17, 2001.

[3] HITZLER, Pascal – KRÖTZSCH, Markus – EHRIG, Marc – SURE, York. "What Is Ontology Merging? – A Category-Theoretical Perspective Using Pushouts". *Proceedings of 1st International Workshop on Contexts and Ontologies (IWCO '05)*. 2005.

[4] HJELM, Johan. "Creating the Semantic Web with RDF". New York: Wiley, 2001. ISBN 0-471-40259-1.

[5] MAEDCHE, Alexander. "Ontology Learning for the Semantic Web". Norwell: Kluwer Academic Publisher, 2002. ISBN 0-7923-7656-0.

[6] PREDOIU, Livia – FEIER, Cristina – SCHARFFE, Francois – BRUIJN, Jos de – MARTÍN-RECUERDA, Francisco – MANOV, Dimitar – EHRIG, Marc. "State-of-the-art Survey on Ontology Merging and Aligning V2". *Digital Enterprise Research Institute*, University of Innsbruck. 2005–2006.

[7] SHVAIKO, Pavel – EUZENAT, Jérôme. "Tutorial on Schema and Ontology Matching". *Proceedings of 2nd European Semantic Web Conference (ESWC '05)*. 2005.

[I1] Web Ontology Language (OWL) / W3C Semantic Web Activity. `http://www.w3.org/2004/OWL`.

[I2] Resource Description Framework (RDF) / W3C Semantic Web Activity. `http://www.w3.org/RDF`.

[I3] RDF Vocabulary Description Language 1.0: RDF Schema. `http://www.w3.org/TR/rdf-schema`.

[I4] W3C Semantic Web Activity. `http://www.w3.org/2001/sw`.

[I5] Extensible Markup Language (XML). `http://www.w3.org/XML`.

Ústav Informatiky AV ČR v.v.i.
DOKTORANDSKÉ DNY '07

Vydal
MATFYZPRESS
vydavatelství
Matematicko-fyzikální fakulty
University Karlovy
Sokolovská 83, 186 75 Praha 8
jako svou *– not yet –* . publikaci

Obálku navrhl František Hakl

Z předloh připravených v systému LaTeX
vytisklo Reprostředisko MFF UK
Sokolovská 83, 186 75 Praha 8

Vydání první

Praha 2007

ISBN *– not yet –*