Integrace dat v prostředí Sémantického Webu*

Zdeňka Linková

2nd ročník PGS, email: linkova@cs.cas.cz
Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, ČVUT
školitel: Július Štuller, Ústav informatiky Akademie Věd ČR

Abstract. Data integration has been an acknowledged problem in data processing field. Its goal is usually to provide an unified view over several data sources. In case of nonmaterialized solution, it is crucial to establish relationships between provided virtual view and data in the sources. The paper deals with relationships establishment. Its approach is based on ontologies.

Abstrakt. Datová integrace je uznávaný problém v oblasti zpracování dat. Jejím cílem je obvykle poskytnout ucelený pohled na několik datových zdrojů. V případě nematerializovaného řešení je klíčové stanovení vazeb mezi poskytovaným virtuálním pohledem a daty uloženými ve zdrojích. Článek se zabývá řešením stanovení těchto vazeb. Svůj přístup zakládá na ontologiích.

1 Úvod

Dnešní svět je světem informací. Téměř vše je založeno na informacích, od pokroku v oblasti výzkumu po podnikatelský úspěch. Větší přístup k informacím umožnila také expanze World Wide Webu (WWW) - dnešní WWW obsahuje obrovské množství informací. Se stále vzrůstajícím objemem dat se však objevují nové problémy, které je třeba řešit. Nastávají potíže se správou dat a jejich zpracováním. Příčinou jsou různé datové formáty, nesourodost dat a prezentace dat způsobem, který je sice příjemný pro člověka, ale není počítačově "čitelný" a tím ztěžuje automatické zpracování. Manuální zpracování je však vzhledem k množství dat téměř nemožné. A tak, i když je požadovaná informace na WWW umístěna, může být obtížné ji najít či zpracovat a využít.

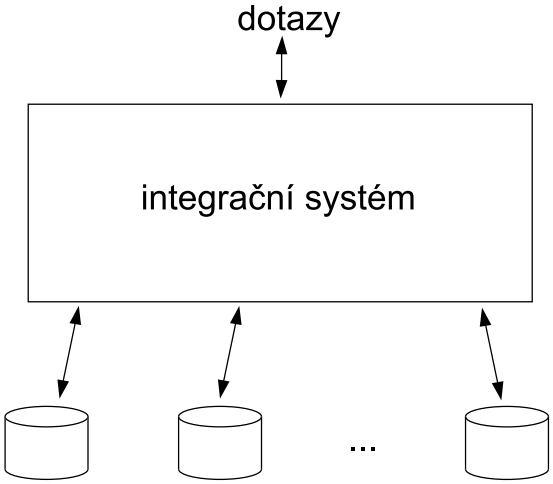
Jedním z problémů, kterými je třeba se zabývat, je tzv. *integrace dat*. Integrace dat se zabývá sloučením dat. Obvykle je jejím cílem prezentovat data pocházející z různých datových zdrojů jako jediný celek a umožnit je zpracovávat, jako by pocházela z jediného datového zdroje. Potřebu datové integrace nacházíme v mnoha oblastech práce s daty: při jejich vkládání do databáze, dotazování se nad několika tabulkami relační databáze, až po slučování více databází i jiných datových zdrojů. Například při vyhledávání informace na webu je často k odpovědi na dotaz potřeba práce s více zdroji - požadovaná informace se totiž nemusí nikde nacházet kompletní. Často je však možné na různých místech nalézt alespoň její části, ze kterých lze pak, jsou-li vhodně zkombinované, požadovanou informaci sestavit.

Datová integrace je dlouho uznávaným problémem zpracování dat. I když je předmětem mnoha výzkumů a projektů a některá data z určitých oblastí již byla integrována, stále není žádný univerzální nástroj, který by integraci dat řešil na obecné úrovni.

*Práce byla podpořena 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 "Informatika pro informační společnost: Modely, algoritmy, aplikace".

Různé přístupy řeší integraci na různých úrovních abstrakce, některé integrují pouze data samotná, jiné se zabývají také datovými schématy. Přístupy lze rozdělit také podle toho, v jaké formě onen ucelený přístup k datům poskytují. Jeden způsob spočívá ve vytvoření nového datového zdroje, který obsahuje všechna data z původních zdrojů. Data jsou z původních zdrojů do nového přesunuta, nebo jsou vytvořeny jejich kopie, případně jsou některá data dále nakombinována. Protože nový datový zdroj data opravdu fyzicky obsahuje, označuje se tento přístup jako *materializovaný*. Výhodou tohoto řešení je, že data jsou stále dostupná, není nutný další přístup k původním zdrojům. S tím také souvisí výhoda relativně rychlé odpovědi, nebo alespoň možnosti rychlost přístupu k datům ovlivnit. Nevýhodou mohou být však paměťové nároky a především aktuálnost dat. Při změně původních dat se integrovaná data stávají neaktuálními a je nutná změna. Tento přístup proto není vhodný pro data, která se rychle mění.

Při integraci dat na webu, a nejen tam, se stále častěji využívá *nematerializovaný* přístup. Jeho podstatou je vytvoření tzv. *virtuálního pohledu* [20] na data. Ten data fyzicky neobsahuje, data zůstávají v původních zdrojích, ale je možné k němu přistupovat a s ním pracovat, jako by data skutečně obsahoval. Data jsou pak stále aktuální. Je však nutné zajistit dostupnost zdrojů a rychlost odpovědi na dotaz je limitována rychlostí nejpomalejšího zdroje.



Obrázek 1: Nematerializovaný integrační systém

Aby bylo možné k integraci využít virtuální pohled, je třeba definovat jeho vazbu na fyzická data. Proto je třeba se v tomto přístupu zabývat také schématy dat. Vazba mezi pohledem a daty se pak zajistí definováním vztahů mezi jednotlivými částmi schématu pohledu a částmi schémat původních zdrojů. Stanovení takovýchto vztahů pak bývá označeno jako *mapování*. Mapování je pak využito při zpracování dotazů. Dotaz, který je kladen v prostředí - tj. jazyk a schéma - (globálního) integrovaného virtuálního pohledu, bývá pomocí mapování rozložen na části odpovídající jednotlivým (lokálním) zdrojům a přepsán do jejich prostředí. Tyto části jsou vyhodnoceny nad původními zdroji. Jejich lokální odpovědi jsou pak s pomocí mapování opět sestaveny do globální odpovědi, která je vrácena jako odpověď na kladený dotaz.

Hlavní komponenty integračního systému, založeném na nematerializované integraci, jsou zdroje se svými (lokálními) schématy, virtuální pohled s (globálním) schématem a systém mapování. Integrační systém lze tedy formalizovat jako trojici

$$I = (G, L, M),$$

kde G je globální schéma, L je množina lokálních schémat a M je systém mapování.

2 Sémantický Web

Idea *Sémantického Webu* [3], [11] vychází ze snahy umožnit zautomatizované zpracování dat. Snaží se data prezentovat v takové podobě, aby byla nejen srozumitelná pro člověka, ale také srozumitelná, čitelná a smysluplně zpracovatelná pro počítačový program. Toho chce dosáhnout doplněním sémantiky - je založen na myšlence, že spolu s daty bude definován také jejich význam.

Sémantický Web není nějaký nový samostatný web, ale je zamýšlen jako rozšíření toho současného. Je založen na několika webových technologiích a standardech, o jejichž definici usiluje W3C (WWW konsorcium) [21].

Důležitým požadavkem počítačem zpracovatelné informace je *strukturování dat*. Na webu je hlavní strukturovací technikou používání značek (tagů). Základním pilířem tvorby Sémantického Webu je v současné době značkovací jazyk XML (eXtensible Markup Language) [9]. Ovšem samotné XML a strukturování dat k počítačově čitelné informaci nestačí. Technikou, jak specifikovat význam informace je RDF (Resource Description Framework) [18]. RDF je základním nástrojem k *připojení metadat* (dat o datech). Poskytuje abstraktní model pro definici metadat a jejich použití. K vyjádření metadat modelu RDF je využívána syntaxe jazyka XML (tzv. RDF/XML). K dispozici je i rozšíření RDF zvané RDF Schema [17], s jehož pomocí lze definovat třídy a hierarchickou strukturu.

Nástrojem pro definici termínů použitých v popisu dat nebo metadat jsou *ontologie*. V kontextu webových ontologiích je ontologie soubor nebo dokument, který obsahuje formální definici termínů a vztahů mezi termíny. Technikou Sémantického Webu pro definici ontologií je jazyk OWL (Ontology Web Language) [22]. Díky používání ontologií budou moci aplikace sdílet své slovníky, což umožní jejich kooperaci.

Idea Sémantického Webu navíc zahrnuje i přidání *logiky* na web, především ve smyslu používání inferenčních pravidel. To přináší možnost odvozovat další vztahy a činit různé závěry.

Svého plného potenciálu může Sémantický Web dosáhnout jen tehdy, jestliže lidé vytvoří také programy, které budou jeho služeb využívat. Takovéto programy by měly zpracovávat obsah webových zdrojů a vzájemně kooperovat. Jejich práce bude tím efektivnější, čím více dat na webu bude počítačově zpracovatelná a čím budou přístupnější ostatní automatizované služby. Sémantický Web by tak měl poskytnout základnu pro ostatní technologie.

3 Integrace dat na Sémantickém Webu

S ohledem na množství datových zdrojů na WWW a na fakt, že některé zdroje jsou zde velmi rychle aktualizovány, je k datové integraci vhodný nematerializovaný přístup. Klíčové je v tomto případě určení vazeb mezi jednotlivými datovými schématy. I na Sémantickém Webu je možné použít klasické přístupy [14], [4], známé jako GAV a LAV.

GAV (*Global As View*) přístup je založen na tom, že globální virtuální pohled je definován jako pohled nad lokálními zdroji. Každý element globálního schématu je tedy charakterizován jako pohled nad lokálními schématy. Tento způsob mapování vlastně systému říká jak získat data. V jednoduchém případě je pak zpracování globálních dotazů relativně snadné, každý element globálního schématu je v dotazu nahrazen pohledem, kterým je definován. Ovšem idea GAV je vhodná v případě, že je množina integrovaných zdrojů stabilní. Změna ve zdrojích, například přidání nového zdroje, může být složitá. Nový či měněný zdroj může mít vliv na definici různých elementů globálního schématu, takže je systémový návrhář nucen přepracovat schéma a uvažovat všechny zdroje znovu.

LAV (*Local As View*) spočívá v definici lokálních schémat jako pohledů definovaných nad globálním schématem. V tomto přístupu je globální schéma voleno (relativně) nezávisle na schématech zdrojů. Každý zdroj je potom charakterizován v termínech globálního schématu. Mapování tak vlastně specifikuje roli každého zdroje v globálním schématu. V případě změny nebo přidání nového zdroje nejsou vyžadovány žádné jiné změny, jen je přepracováno nebo přidáno mapování dotčeného zdroje. Na druhou stranu, je v LAV obtížné zpracování dotazů. Jediná znalost o datech globálního pohledu je pomocí pohledů reprezentující lokální zdroje a ty obsahují pouze částečnou informaci o datech. Nemusí být proto zřejmé, jak datové zdroje pro zodpovězení dotazu využít.

Protože každý z přístupů GAV a LAV má své výhody i nevýhody v různých částech procesu integrace, objevují se i projekty, které oba přístupy různě kombinují. Například GLAV (*Global Local As View*) [10] specifikuje mapování LAV i GAV.

K popisu mapování, ať už získaném přístupem GAV, nebo LAV je možné využit různých struktur. Tyto struktury se často liší projekt od projektu, některé zachycují pouze dvojice ekvivalentních konceptů, jiné pro zachycení složitějších vztahů využívají struktur složitějších. Na Sémantickém Webu je také možné takovéto přístupy ke stanovení a popisu mapování využít. Ovšem na Sémantickém Webu je k dispozici jeden mocný prostředek, který může přispět v úloze stanovení vztahů mezi schématy a nabídnout více v případě jejich zachycení. Tím prostředkem jsou ontologie.

Pojem *ontologie* [8] bývá užíván v různých souvislostech. Velmi populární definicí ontologie v informatice je: ontologie je formální, explicitní specifikace konceptualizace. Konceptualizace se vztahuje k abstraktnímu modelu světa. Ovšem konceptualizace není platná universálně, není jednoznačný přístup k tomu, jak abstraktní model světa okolo nás vytvářet. Ontologie by měly řešit problém implicitního přístupu k modelování znalostí zavedením explicitní konceptualizace.

Je množství přístupů k datové integraci, které využívají ontologií. Ty mohou být využity v různých částech integračního procesu. Na počátku mohou být ontologie využity v datových zdrojích k popisu dat. Tyto ontologie pak mohou být použity při identifikaci sémanticky korespondujících konceptů. To je zásadní při stanovení mapování.

Některé projekty, které využívají ontologií dostupných se zdroji dat, řeší integraci

klasickým přístupem GAV nebo LAV, jako například [1]. V některých projektech mají ontologie i další úlohu. Například je vytvořena také tzv. globální ontologie, tj. ontologie globálního pohledu. Ta může být definována dvěma způsoby. Jednak může obsahovat slovník, který lokální zdroje sdílejí. V některých projektech obsahuje taková ontologie základní pojmy z konkrétní domény a je obvykle mnohem obecnější než lokální ontologie [15]. Druhou možností pak je definovat globální ontologii jako výsledek sloučení ontologií lokálních.

Mnohé projekty datové integrace zůstávají u definice mapování pro stanovení vztahů mezi globálním pohledem a lokálními zdroji, např. [7]. K mapování může být využita široká škála struktur, od jednoduchých mapovacích pravidel mezi dvěma koncepty (synonyma, homonyma, disjunktní pojmy atd.), přes mapování konceptu na jedné straně k dotazu či pohledu na straně druhé [5] (jako je tomu u GAV a LAV), až po přídavné mapovací struktury (např. model referencí v [19]). Některé projekty využívají v této fázi ontologie k určení svého pojetí mapování a vlastní mapování je vlastně instancí této ontologie mapování.

Přístup, který je prezentovaný v tomto příspěvku je podobný přístupu v [6] - je také založený na slučování lokálních ontologií. Rozdíl je v tom, že ačkoli je globální ontologie výsledkem těch lokálních, oni k popisu mezi lokálním a globálním prostředím zůstávají u "tradičního" pojetí mapování, konkrétně k mapování využívají mapovací tabulku. Přístup v tomto příspěvku navrhuje v mapování využít samotnou ontologii, která vznikne jako výsledek sloučení ontologií lokálních, ontologie globální a rovněž všech známých vztahů mezi nimi. Proto je tento přístup nejvíce podobný projektům, které byly primárně určeny ke slučování ontologií, jako např. [16].

3.1 Integrace dat založená na ontologiích

Integrační systém v tomto příspěvku je založen na nematerializovaném přístupu. Globální zdroj je reprezentován virtuálním pohledem. K zajištění přístupu k datovým zdrojům (a tím vlastně k fyzickým datům) je využit jistý druh mapování. Avšak na rozdíl od mapovacích pravidel jako tvrzení zachycujících vztahy mezi elementy jednotlivých schémat je využita složitější struktura. Vychází z předpokladu, že data, která integruje jsou data pro Sémantický Web, neboli předpokladem je dostupnost ontologií, které integrovaná data popisují.

Samotná úloha integrace je transformována na vytváření ontologie integračního systému. Tato ontologie ze své podstaty by měla pokrývat ontologie všech dat, která jsou v systému využita, a mapování, na které můžeme obecně nahlížet jako na stanovení vztahů mezi daty. Tak může být tato ontologie využita k datové integraci na úrovni schémat.

Ontologie a datová schémata obecně k sobě mají blízko. Hlavním rozdílem je především jejich účel. Ontologie bývá tvořena s úmyslem definovat pojmy používané v nějaké oblasti, zatímco schéma bývá využito k modelování nějakých určitých dat. Ačkoli není nutně platné tvrzení, že lze nalézt korespondenci mezi datovým modelem a k jeho vyjádření využitými pojmy, často tomu tak bývá. Obzvláště u schémat reprezentujících sémantický datový model, nemusí být patrný žádný rozdíl a tím ani způsob jak identifikovat co je schéma a co ontologie. V ostatních případech může být ontologie využitá pro ta konkrétní data obohacena o další koncepty a vztahy, aby zachytila i datové schéma.

Předpokládejme, že na počátku úlohy jsou k dispozici dva datové zdroje S_1 a S_2, které mají být integrovány. Každý zdroj je popsán ontologií: ontologii zdroje S_1 označme O_{S_1}, ontologii zdroje S_2 označme O_{S_2}. Globální integrovaný pohled poskytovaný integračním systémem bude popsán ontologií O_G. Integrační systém, který je v kapitole 1 formalizován jako trojice $I = (G, L, M)$, má v tomto případě reprezentaci $I = (O_G, \{O_{S_1}, O_{S_2}\}, O_I)$, kde O_I je ontologie integračního systému.

Ontologie O_I je určena k popsání mapování mezi elementy globálního pohledu a lokálních zdrojů. Mapování je stěžejní částí integračního systému a jeho stanovení, popis a vyjadřovací síla ovlivňují množství informace, které jsme schopni integračním systémem získat. O_I je zároveň ontologií všech dat v integračním systém. Z toho plyne, že pro ontologie lokálních zdrojů platí:

$$O_{S_1} \subseteq O_I$$
$$O_{S_2} \subseteq O_I$$

Zatímco ontologie O_{S_1} a O_{S_2} jsou dány spolu se zdroji, O_G a O_I je třeba stanovit. Popis O_G je relativně nezávislý na zdrojích. O_G obsahuje definici všech konceptů přístupných přímo pomocí globálního pohledu. Je proto úlohou designéra, který rozhodne, co bude pomocí integračního systému přístupné a v jaké formě.

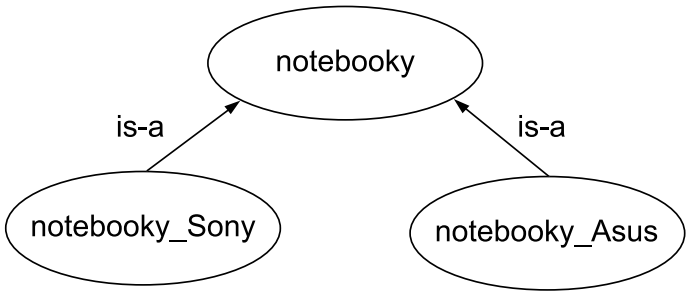
Určení O_I je zásadním krokem, nejde ovšem o nijak snadnou úlohu. Protože by O_I měla pokrývat O_{S_1}, O_{S_2}, O_G i jejich vzájemné vztahy, je O_I výsledkem úlohy nazývané slučování ontologií (*ontology merging*). Při procesu slučování je několik lokálních ontologií na vstupu, na výstupu je pak jako výsledek vrácena sloučená ontologie. Proces slučování ontologií je tématem mnoha výzkumných projektů,např. [12], a je i předmětem mnoha specializovaných konferencí.

Hlavní obtíž při sémantické integraci je stanovení korespondence. I když formální definice v ontologiích jsou nejlepší specifikace termínů, které jsou v současné době k dispozici, nemohou zachytit plný význam. Často je proto při stanovení korespondence nutná jistá lidská intervence. Ovšem i v případě, že počítačové programy nejsou schopny korespondenci odvodit, mohou být využity k návrhům možných vztahů či k ověření vztahů zadaných člověkem.

Stejně jako při integraci schémat v jiných přístupech, i zde mohou vyvstat některé konflikty [2], které je třeba řešit. Obecně mohou být takové konflikty několika typů [13], například to mohou být konflikty mezi jednotlivými termíny (synonyma, homonyma atd.), schématické nesourodosti, konflikty mezi vlastními daty a metadaty atd. Ve světě ontologií není obtížné uvažovat rozdílné koncepty, protože jsou zde prostředky, jak vyjádřit vztahy mezi nimi. V ontologii má každý pojem jednoznačnou referenci. I když mohou být dva termíny ve dvou ontologiích shodně nazvány, jsou odlišitelné díky kontextu - ontologii, ve které jsou definovány. Toto je například v syntaxi XML řešeno pomocí jmenných prostorů (namespaces). V ontologiích je také možné stanovit, že dva termíny jsou ekvivalentní a umožnit tak, aby byly podle toho zpracovány.

Příklad 1:
Předpokládejme obchod, který prodává notebooky od různých výrobců. Pro jednoduchost předpokládejme pouze dva výrobce: Sony a Asus. Každý z nich poskytuje data o svých produktech. Obchod by však přivítal přístup k těmto datům jako k jedinému celku, proto je nutné oba zdroje s daty integrovat.

K integraci jsou dva zdroje. Zdroj 1 obsahuje notebooky produkované firmou Sony. Jeho ontologie O_1 obsahuje pouze jedinou třídu nazvanou `notebooky_Sony`, které mají vlastnosti `Rychl_procesoru`, `Paměť` atd. Zdroj 2 obsahuje notebooky od producenta Asus. Jeho ontologie O_2 obsahuje třídu `notebooky_Asus` s vlastnostmi `RychlostProc`, `RAM` atd. Protože integrační systém by měl poskytovat popis notebooků z různých zdrojů, globální ontologie O_G obsahuje třídu `notebooky` s vlastnostmi `Rychlost_procesoru`, `RAM` atd. Pro získání ontologie celého systému, jsou O_1, O_2, O_G a znalosti o vztazích mezi koncepty sloučeny. Výsledná O_I je následující:



Obrázek 2: Ontologie O_I

Ontologie O_I obsahuje tři třídy: `notebooky`, `notebooky_Sony` a `notebooky_Asus`. Notebooky Sony i Asus jsou notebooky, proto je hierarchický vztah třída - podtřída mezi třídami `notebooky` a `notebooky_Sony` a mezi `notebooky` a `notebooky_Asus`. Notebooky Sony a Asus nemohou být sloučeny do jediné třídy, protože tyto třídy obsahují rozdílné notebooky. Se znalostí sémantiky jednotlivých vlastností tříd, lze podobně nalézt hierarchické vztahy vlastnost - podvlastnost, například mezi `Rychlost_procesoru` a `Rychl_procesoru`. Navíc, jsou-li na tyto vlastnosti kladeny stejná integritní omezení, je možné je sloučit a označit jako ekvivalentní. □

Uživatel integračního systému klade své dotazy v prostředí globálního pohledu (tj. schéma, jazyk atd.). Aby bylo možné tento dotaz nad lokálními zdroji vyhodnotit, je nutné jej zpracovat. V zásadě jsou dva přístupy ke zpracování globálního dotazu. V prvním je dotaz přepsán - je dekomponován na části odpovídající jednotlivým zdrojům a přepsán do příslušného lokálního prostředí. Druhý přístup se zabývá hledáním odpovědi na dotaz, kdy neklade žádné nároky na to, jak je dotaz proveden, ale jediným cílem je vyšetřit všechny možné informace s cílem najít množinu dat, která (s využitím znalostí) logicky implikuje, že jde o odpověď na dotaz.

Máme-li mapování vyjádřeno v ontologii a uvažujeme-li pouze hierarchický is-a vztah, je možné pro zpracování dotazu využít pravidlo dobře známé z objektově-orientovaného programování: potomek může zastoupit svého předka. Jestliže například hledáme všechny instance třídy T, které mají vlastnost $V = x$, je dotaz následující:

$$q := T(V = x).$$

S využitím ontologie O_I poskytuje is-a vztah prostředek, jak dotaz přepsat s ohledem na určitý lokální zdroj. Jestliže T není konceptem lokálního schématu, je třída T v dotazu nahrazena nejbližším svým potomkem T'. Toto pravidlo je rekurzivně aplikováno, dokud není malezen kontext lokálního schématu, nebo není-li již k dispozici další podtřída - v tom případě je odpověď na (lokální) dotaz prázdná. Podobně je toto pravidlo aplikovatelné pro přepisování vlastností stanovených v dotazu.

Při druhém přístupu k zodpovídání dotazů je is-a hierarchie také klíčová. Vyjadřuje totiž, že instance libovolného uzlu je zároveň instancí uzlu v hierarchii nad ním. Na základě tohoto poznatku je možné určit, zda informace z lokálního zdroje může být odpovědí na globální dotaz.

Příklad 2:
V pokračování jednoduchého příkladu integrace notebooků je ukázáno zpracování dotazů. Globální pohled poskytuje notebooky. Dotaz: vyber všechny notebooky s rychlostí procesoru 1.6 GHz, t.j.

$$q := notebooky(Rychlost\_procesoru =' 1.6'),$$

je zpracován následovně: `notebooky` není konceptem žádného z lokálních zdrojů, dotaz je přepsán. Třída `notebooky` má dva potomky `notebooky_Sony` ze zdroje 1 a `notebooky_Asus` ze zdroje 2. Dotaz je tedy přepsán do dvou podob:

$$q'_1 := notebooky\_Sony(Rychlost\_procesoru =' 1.6'),$$
$$q'_2 := notebooky\_Asus(Rychlost\_procesoru =' 1.6').$$

Protože vlastnost `Rychlost_procesoru` není koncept zdroje 1, je dotaz q'_1 dále přepsán s využitím vztahu vlastnost - podvlastnost na:

$$q''_1 := notebooky\_Sony(Rychl\_procesoru =' 1.6').$$

Dotaz q''_1 je proveden nad zdrojem 1. Analogicky je přepsán dotaz q'_2 a proveden nad zdrojem 2. □

Ontologie je tedy schopná zachytit jednoduchý is-a vztah. Ontologie je však mnohem silnější. Poskytuje dost silné prostředky, aby zachycovala množství různých vztahů. Aby bylo umožněno získat ze zdrojů co největší množství informace, je vhodné využít nějaký inferenční mechanismus, který umožní pracovat i se vztahy, které ontologie pokrývá, nicméně nejsou v ní přímo vyjádřeny - je však možné je z vyjádřených vztahů dále odvodit.

4 Diskuze a závěr

Řešení integrace dat navrhované v tomto příspěvku vychází z technik Sémantického Webu. Je založena na ontologiích - vychází z předpokladu, že integrovaná data jsou popsána v ontologiích, které jsou dostupné, a ontologii dále využívá jako prostředek ke stanovení vazeb mezi původními daty a poskytovaným integrovaným pohledem. Vlastní úloha integrace je převedena na úlohu slučování ontologií.

Použití ontologií (a jejich definice pomocí OWL jako standardního prostředku Sémantického Webu) v integračním systému může přinést řadu výhod. Ontologie jsou velmi silný prostředek pro zachycení vztahů mezi schématy jednotlivých zdrojů. Ontologie jako obecný prostředek, tedy na rozdíl od jiných struktur mapovacích pravidel definovaných primárně pro potřeby integrace, může být dále použita i v jiných souvislostech. Velkou výhodu použití ontologie v integraci lze nalézt při změně sytému, například při přidání nového zdroje. V takovém případě není nutná změna stávající ontologie - není tedy nutné jako ontologii nebo některou její část přepracovat, či odstranit. Původní ontologie je obohacena o nový stav změněného zdroje, či o nový zdroj, jiné změny nejsou nutné.

Literatura

[1] B. Amann, C. Beeri, I. Fundulaki, and M. Scholl. *Ontology-Based Integration of XML Web Resources*. Proceedings of the *1st Int. Semantic Web Conference (ISWC 2002)*, (2002), 117–131.

[2] S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano. *Semantic integration of heterogeneous information sources. Data & Knowledge Engineering* 36 (2001), 189–210.

[3] T. Berners-Lee, J. Hendler and O. Lassila. *The Semantic Web. Scientific American*, vol. 284, 5, (2001), 35–43.

[4] A. Cali, D. Calvanese, G. De Giacomo, and M. Lenzerini. *On the Expressive Power of Data Integration Systems*. In Proceedings of the *21st Int. Conf. On Conceptual Modeling (ER 2002)*, LNCS 2503, Springer, (2002), 338–350.

[5] D. Calvanese, G. De Giacomo, and M. Lenzerini. *Ontology of integration and integration of ontologies*. In Proceedings of the *2001 Description Logic Workshop (DL 2001)*, (2001).

[6] I. F. Cruz, H. Xiao, and F. Hsu. *An Ontology-based Framework for XML Semantic Integration*. In Proceedings of the *8th Int. Database Engineering and Application Symposium (IDEAS'04)*, Coimbra, Portugal, (2004), 217–226.

[7] Z. Cui, D. Jones, and P. O'Brien. *Issues in Ontology-based Information Integration*. In Proceedings of the *IJCAI Workshop: Ontologies and information sharing*, Seattle, USA, (2001).

[8] Y. Ding, D. Fensel, M. Klein, and B. Omelayenko. *The semantic web: yet another hip?. Data & Knowledge Engineering* 41 (2002), 205–227.

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

[10] M. Friedman, A. Levy, and T. Millstein. *Navigational plans for data integration.* In Proceedings of the *16th Nat. Conf. On Artificial Intelligence (AAAI'99)*, AAAI Press, (1999), 67–73.

[11] M.-R. Koivunen and E. Miller. *W3C Semantic Web Activity.* In the proceedings of the *Semantic Web Kick/off Seminar*, Finland, (2001).

[12] K. Kotis, G. A. Vouros, and K. Stergiou. *Towards automatic merging of domain ontologies: The HCONE-merge approach. Web Semantics: Science, Services and Agents on the World Wide Web* 4 (2006), 60–79.

[13] C.-Y. Lee and V.-W. Soo. *The conflict detection and resolution in knowledge merging for image annotation. Information Processing and Management* 42 (2006), 1030–1055.

[14] M. Lenzerini. *Data Integration: A Theoretical Perspective.* In Proceedings of the *21st ACM SIGMOD - SIGACT - SIGART symposium on Principles of database systems*, ACM Press, (2002), 233–246.

[15] N. F. Noy. *Semantic Integration: A Survey Of Ontology-Based Approaches.* In *ACM SIGMOD Record, Special Section on Semantic Integration*, vol.33, 4, (2004), 65–70.

[16] N. F. Noy and M. A. Musen. *The PROMPT suite: Interactive tools for ontologzy merging and mapping. International Journal of Human-Computer Studies* vol. 56, 6,(2003), 983–1024.

[17] RDF Vocabulary Description Language 1.0: RDF Schema. *W3C Recommendation.* http://www.w3.org/TR/2004/REC-rdf-schema-20040210, February, 2004.

[18] Resource Description Framework (RDF). http://www.w3.org/RDF/.

[19] H. T. Uitermark, P. J. M. van Oosterom, N. J. I. Mars, and M. Molenaar. *Ontology-based integration of topographic data sets. International Journal of Applied Earth Observation and Geoinformation* 7 (2005), 97–106.

[20] J. D. Ullman. *Information integration using logical views. Theoretical Computer Science* 239 (2000), 189–210.

[21] W3C (WWW Consortium). http://www.w3.org.

[22] Web Ontology Language (OWL). http://www.w3.org/2004/OWL.