České vysoké učení technické v Praze

Czech Technical University in Prague

Ing. Tomáš Vitvar, Ph.D.

Sémantický web

Semantic Web

Summary
The Semantic Web is an area of Web engineering focused on knowledge usage in the context of open and distributed Web environment. A vision of the Semantic Web is to create a universal platform for data, information and knowledge exchange. The Semantic Web technology should enable better utilization of information and services that the Web currently provides. Semantic Web uses knowledge representation techniques as a basis for knowledge models (ontologies), it defines languages supporting such representations, it defines and maintains various types of ontologies covering various activities of our daily lives, and it defines mechanisms allowing to incorporate ontologies in the existing Web environment.

In the core, the Semantic Web exploits knowledge representation theories based on logic. A fundamental requirement is to express rich taxonomic hierarchies enabling effective knowledge organization and reasoning. The Semantic Web supports various levels of semantic expresivity so that engineers can solve a tradeoff between requirements for semantic expressivity and requirements for system performance. The Semantic Web thus supports semantics at the basic level in a form of taxonomic hierarchies, objects, classes of objects (categories), objects' memberships in categories. At the higher level of semantic expressivity, the Semantic Web supports semantics of description logic that allows for advanced relations between objects, for example, disjunction, intersection of categories, cardinality constraints, etc. In addition, the Semantic Web adopts rule languages for definition of implicit knowledge that have a basis in Logic Programming.

The Semantic Web uses these representations for definition of languages and ontologies. The Semantic Web strongly adopts existing Web architecture principles, in particular, syntactical representation of data (XML language), identification of Web resources (URI), and communication (HTTP protocol). On those basis, the Semantic Web builds a family of languages reflecting incremental requirements for semantic expressivity, namely RDF (objects and categories of objects), RDFS (taxonomic hierarchies), OWL (advanced semantics of description logic), rule languages (arbitrary rules for definition of implicit knowledge). The Semantic Web also provides query languages for semantic data as well as reasoning mechanisms and tools.

One of the most important activities in the Semantic Web is the definition of ontologies for various application domains. Ontologies serve the basic interoperability in a domain and a common understanding among users and systems on the Web. A very active area of the Semantic Web is open link data (through Open Link Data project). Open link data currently involves around 4.7 billion RDF triples in ontologies that describe various areas of our daily lives. For example, open link data includes data from Wikipedia, bibliographic resources, user profiles, etc. In order to solve the interoperability between ontologies (in situations when domain ontologies overlap), the Semantic Web defines methods for ontology alignment and data transformations. The Semantic Web also works on mechanisms for annotation of existing resources on the Web so that semantic data can be used on top of existing information that the Web currently provides.

Research and development in the Semantic Web has recently produced many valuable technologies and applications. They slightly become the mainstream technologies on the Web while they will significantly influence the move towards the universal platform for data, information and knowledge exchange on the Web.

Souhrn
Sémantický web je oblast webového inženýrství, která se zabývá využitím znalostí v prostředí otevřeného a distribuovaného prostředí Webu. Vize Sémantického webu spočívá ve vytvoření universální platformy pro sdílení dat, informací a znalostí a vytoření inteligentních mechanismů pro vylepšenou práci uživatelů s informacemi a službami, které web dnes nabízí. Sémantický web využívá způsoby reprezentace znalostí jako základ znalostních modelů (ontologií), definuje jazyky pro popis těchto ontologií, vytváří různé typy ontologií pro využití v různých oblastech aktivit člověka a definuje mechanismy pro začlenění a využití ontologií v existujícím webovém prostředí.

Sémantický web využívá reprezentace znalostí založené na logice. Základním požadavkem na tyto reprezentace je vyjádření bohatých taxonomických hierarchií, které umožní efektivní organizaci znalostí a odvozování znalostí nových. Sémantický web podporuje různé úrovně vyjadřovacích schopností jazyků pro reprezentaci znalostí tak, aby bylo možné sladit požadavky na expresivitu modelů s požadavky na výkon systémů. Sémantický web tak podporuje základní sémantiku na úrovni taxonomické hierarchie, objektů, tříd (kategorií) objektů a začlenění objektů do kategoriií. Na vyšší úrovni sémantické expresivity je potom možno definovat modely s expresivitou deskripční logiky umožňující popisovat složitejší vztahy mezi objekty (např. vztah disjunkce, průniku kategorií, omezení kardinalit, atp.) a také modely s expresivitou pravidel pro definici implicitních (skrytých) znalostí založené na bázi logického programování a produkčních systémů.

Sémantický web využívá tyto reprezentace pro definici jazyků pro popis ontologií a zároveň zaručuje, aby tyto jazyky byly založeny na existující infrastruktuře Webu ve smyslu základních technologií pro: popis struktury a syntaxe dat (jazyk XML), jednoznačnou identifikaci zdrojů (schéma URI) a komunikaci (protokol HTTP). Na tomto základě Sémantický web definuje rodinu jazyků pro popis ontologií jejiž členění odpovídá požadavkům inkrementální expresivity modelu, tzn. jazyky RDF (základní definice objektů a kategorií), RDFS (taxonomické hierarchie), OWL (pokročilá sémantika na úrovni deskripční logiky) a jazyky pro popis pravidel (volně definovatelná pravidla pro popis implicitních znalostí). Sémantických web taky definuje jazyky pro dotazování nad znalostmi a mechanismy pro usuzování v ontologiích.

Důležitou oblastí aktivit v Sémantickém webu je tvorba ontologií jako prostředků pro modelování znalostí v různých aplikačních doménách, ale také jako prostředků pro zajištění základní interoperability a porozumění uživatelů a systémů na Webu. Velice aktivní oblastí Sémantického webu je oblast tzv. otevřených a spojených dat (projekt Open Link Data). Open Link Data v současnosti obsahuje kolem 4.7 biliónů sémantických dat (tzv. trojic–triples) v rámci ontologií pro popis různých oblastí (např. Wikipedie, bibliografikých zdrojů, uživatelských profilů, atp.). Pro zajištění interoperability ontologií potom existují mechanismy pro mapování ontologií a transformaci dat. Aby bylo možné využít technologie Sémantického webu v existujícím prostředí a podmínkách současného webu je nutné umožnit propojení existujích dat se sémantickými daty. K tomuto účelu jsou vytvářeny anotační mechanismy pro využití sémantických dat v existujících aplikacích.

Vývoj, věda a výzkum v Sémantickém webu za poslední roky vyprodukoval mnoho realizací a technologií, které se postupně stávají součástí hlavního proudu technologií na webu. Jejich využití v aplikacích postupně povede k naplnění vize universální a globální platformy pro sdílení dat, informací a znalostí.

Klíčová slova
Sémantický web, ontologie, znalosti, reprezentace znalostí, webové inženýrství, logické systémy, deskripční logika, pravidlové systémy, sémantické sítě, RDF, RDFS, OWL, XML, SPARQL, architektura Webu, anotace obsahu, interoperabilita.

Keywords
Semantic Web, ontology, knnowledge, knowledge representation, Web engineering, logic systems, description logic, rule systems, semantic networks, RDF, RDFS, OWL, XML, SPARQL, Web architecture, content annotation, interoperability.

Obsah

1 Úvod

Sémantický web patří mezi jednu z nejrychleji se rozvíjejících oblastí Webu. Základní myšlenkou iniciativy Sémantického webu[1] je umožnit aby informace, které jsou dnes na Webu k dispozici, byly srozumitelné nejenom koncovým uživatelům, ale také strojům—počítačům. Vize Sémantického webu je vytvoření universální platformy pro výměnu a sdílení dat, informací a znalostí. Tato platforma je extenzí existujícího Webu, která vyžaduje nové přístupy a technologie pro popis informací a nové nástroje umožňující práci s těmito informacemi. Z tohoto hlediska je Sémantický web v zásadě rozvíjen po třech základních liniích. První a nejdůležitější linie si klade za cíl využít, resp. rozšířit systémy pro reprezentaci a práci se znalostmi tak, aby je bylo možné využít v otevřeném a distribuovaném prostředí Webu. Tato iniciativa vytváří jazyky pro popis znalostních modelů (ontologií) a systémy pro dotazování a usuzování nad těmito modely. Druhá linie vytváří prostředky pro práci s ontologiemi zahrnující jejich interoperabilitu (ontology alignment), anotaci a extrakci informací pomocí ontologií, správu a udržování aktuálnosti ontologií z hlediska jejich vývoje, atd. Třetí linie si potom klade za cíl vytvořit a udržovat konkrétní ontologie pokrývající ruzné oblasti (domény) aktivit člověka.

Iniciativa Sémantického webu je v současnosti z větší části řízena a rozvíjena akademickou komunitou s podstatnou podporou standardizační organizace W3C. Sémantický web svými aktivitami, přístupy a nabízenými technologiemi dnes pokrývá mnoho oblastí informatiky. Patří sem například znalostní inženýrství, softwarové inženýrství, služby a middleware, interoperabilita dat a systémů, logické jazyky, interakce člověka a počítače (human-computer interaction), sociální sítě a různé aplikace v oblastech obchodu, zdravotnictví, státní správy (e-government), telekomunikací a dopravy. Komunita Sémantického webu je soustředěna kolem třech významných akademických konferencí: International Semantic Web Conference (ISWC), European Semantic Web Conference (ESWC) a Asian Semantic Web Conference (ASWC). Všechny tyto konference sdružují poslední poznatky vývoje Sémantického webu ve všech zmíněných oblastech.

2 Znalosti

Podstatné rozšíření, které Sémantický web prináší, je využití formálních prostředků pro reprezentaci znalostí pro popis modelů umožňujících realizaci inteligentního chování systémů na Webu. Sémantický web definuje jazyky pro *sémantickou vrstvu*, která umožňuje explicitní a jednoznačnou reprezentaci znalostí v podobě *ontologií* a zároveň zachovává existující způsoby pro reprezentaci dat na syntaktické úrovni.

2.1 Kvalita modelu

Na kvalitu modelu zobrazující objekty, jejich vazby a vlastnosti lze pohlížet z hlediska třech úrovní jako na *data*, *informace* a *znalosti*. Podle teorie systémového inženýrství[35] jsou data vyjádřena jako relace poznatku k rozlišené větné formě jazyka, informace je vyjádřena jako relace dat k poznatkové úrovni uživatele (jejím použitím se rozšiřuje dosud existující poznání uživatele) a znalost je vyjádřena jako relace informace k jiným údajům nebo informacím.

Znalosti a způsoby jejich reprezentací jsou klíčovým prostředkem pro popis modelů Sémantického webu. Iniciativa Sémantického webu klade velký důraz nejen na prostředky

[1] Jádro iniciativy sémantického webu je dnes soustředěno ve W3C (`http://www.w3.org/2001/sw/`)

pro popis znalostí ale také na způsoby jejich získávání, zpracování a využití. Iniciativa Sémantického webu přijímá základní členění znalostí (literatura např. [25]):

- *implicitní znalosti* jsou „skryté“ ve významu informace a zachycené v nestrukturované podobě (např. v podobě přirozeného jazyka) nebo je možno tyto znalosti odvodit (pomocí inference/usuzování) z jiných znalostí obsažených ve znalostním modelu,
- *explicitní znalosti* jsou oddělené od vlastní úlohy, která znalosti zpracovává a zároveň jsou vyjádřeny formálním způsobem,
- *deklarativní znalosti* svou reprezentací vyjadřují poznatky o množině objektů reálného světa, tzn. co je poznáno nebo dokázáno (např. pes je zvíře),
- *procedurální znalosti* svou reprezentací vypovídají o způsobu poznávání a odvozování; znalosti reprezentované procedurálně mají tvar pravidel (např. je-li X pes, potom je X zvíře);

Iniciativa Sémantického webu zohledňuje všechny typy znalostí. Významnou roli v současném rozvoji Sémantického webu představuje explicitní reprezentace znalostí v podobě pravidel nebo poznatků, umožňující jejich sdílení mezi systémy, k hledání výsledků, vysvětlení postupů jejich hledání nebo odvozování nových znalostí. Explicitní reprezentace znalostí je rovněž významná pro řešení problémů vzájemného porozumnění heterogenních systémů (interoperability).

2.2 Reprezentace znalostí

Pro reprezentaci znalostí o reálném světě v otevřeném prostředí jako je Web je důležité vytvářet a pracovat s *otevřenými znalostními modely*. Tyto modely, resp. reprezentace pomocí kterých je možno tyto modely definovat, musí umožňovat popisovat objekty reálného světa na různých úrovních abstrakce a zároveň musí být možno tyto modely měnit a rozšiřovat během jejich existence. Tento základní požadavek ovlivňuje volbu reprezentace znalostí ve znalostních modelech, která musí umožňovat organizovat objekty do kategorií a vyjadřovat základní vztahy mezi kategoriemi a objekty. Jazyk predikátové logiky prvního řádu je základním nástrojem pro reprezentaci takových znalostí. V literatuře jsou prostředky reprezentace znalostí uvedeny např. v [32, 25].

Kategorie je možné pomocí predikátové logiky prvního řádu reprezentovat pomocí *predikátů* a *objektů*. Například znalost *Fík je pes* můžeme zapsat jako $Pes(Fik)$. Predikát Pes můžeme chápat jako množinu všech psů, kterou můžeme „zhmotnit“ jako objekt Psi. Potom můžeme zapsat, že $Member(Fik, Psi)$ (zkráceně zapsáno vztahem přináležitosti prvku do množiny jako $Fik \in Psi$), tzn. Fík je prvkem kategorie psů. Dále můžeme zapsat $Subset(Psi, Zvirata)$ (zkráceně zapsáno vztahem podmnožiny jako $Psi \subset Zvirata$), tzn. psi jsou podmnožinou zvířat. Kategorii je tedy možné chápat jako množinu jejich prvků. Z pohledu predikátové logiky prvního řádu jde o „komplexní“ objekt, který má definovány relace $Member$ a $Subset$. Všem prvkům dané kategorie můžeme dále přiřadit vlastnosti. Například znalost, že všichni psi mají čtyři nohy zapíšeme jako $x \in Psi \Rightarrow Nohy(x) = 4$.

Vztah podmnožiny umožňuje definovat *taxonomické hierarchie* mezi kategoriemi a využít *dědičnosti* (*inheritance*) pro efektivní modelování znalostí o reálném světě. Například pokud všechny instance kategorie Psi mají 4 nohy a $Baseti$ jsou podmnožinou kategorie Psi, potom víme, že každý baset má 4 nohy (jinými slovy, individuální baseti dědí vlastnost „má 4 nohy“).

Podobně jako vztahy mezi kategoriemi, je někdy též vhodné definovat vztahy mezi objekty kdy jeden objekt je částí jiného objektu (relace *PartOf*). Například znalost, že Česká republika patří do Evropy a Morava patří do České republiky můžeme zapsat jako $PartOf(CR, Evropa)$ a $PartOf(Morava, CR)$.

Základní vztahy mezi kategoriemi a objekty umožňují definovat znalostní modely se základní sémantikou. Podle požadavků na tvorbu těchto modelů je ovšem někdy nutné vyjádřit složitější vztahy mezi kategoriemi a objekty. Prostředky pro reprezentaci znalostí tak umožňují definovat například vztahy průniku kategoriií, disjunkci kategorií, ekvivalenci, restrikci na minimální a maximální kardinalitu, nebo libovolně definovat procedurální znalosti.

Sémantické sítě. Cílem sémantických sítí bylo vytvoření notace pro reprezentaci deklarativních znalostí, která je „uživatelsky přívětivá“. V praxi existuje mnoho variant sémantických sítí pro popis znalostí, všechny varianty ovšem vycházejí z grafické notace uzlů a hran tzv. *existenčních grafů*[28].

Sémantické sítě definují grafickou notaci pro reprezentaci znalostí o objektech, kategoriích objektů, jejich vlastnostech a relacích a umožňují vyjadřovat taxonomické hierarchie. Jednotlivé složky sémantické sítě jsou:

- *Kategorie* vyjadřuje abstraktní popis třídy objekt,
- *Objekt* vyjadřuje konkrétní výskyt objektu,
- *Relace* vyjadřují vztahy
  - *kategorie-kategorie* ve smyslu „je druhem“ (is-kind-of), která odpovídá relaci $Subset$,
  - *objekt-kategorie* ve smyslu „je prvkem“ (is-member-of), která odpovídá relaci $Member$,
  - *objekt-objekt* ve smyslu „patří do“ (is-part-of), odpovídá relaci *PartOf*.

Sémantická síť umožňuje asociovat znalosti z několika pohledů a sdružovat objekty do obecnějších kategorií. Takové kategorie je možné na nižších úrovních konkretizovat pomocí speciálnějších druhů. Tímto způsobem je možné vytvářet členěné komplexy znalostí s možností odvozování znalostí nových. V sémantické síti lze díky taxonomickým vztahům provádět odvozování *specializací* nebo *generalizací*. Při specializaci se informace v taxonomii přenášejí od obecnějších typů ke speciálnějším, při generalizaci naopak. Sémantické sítě však neřeší problémy s konflitky, které mohou vzniknout při dědění již existujících vlastností.

Deskripční logika. Sémantické sítě se staly základem dalších formálních prostředků pro reprezentaci znalostí. Jedním z nich je deskripční logika (DL), která vznikla jako reakce na požadavek vytvoření formální specifikace pro sémantické sítě se zachováním principu taxonomické struktury znalostních modelů. DL používá označení koncept pro kategorii/třídu (chápáno jako množina objektů), role pro relaci (chápány jako binární relace na objektech) a individuum pro objekt. DL definuje množinu kalkulů, které se liší svou vyjadřovací schopností (např. průnik konceptů, sjednocení konceptů, negace konceptů, omezení pomocí kardinalit, atd.).

Syntaxe	Popis
A, B	Koncept
$R.C$	Relace konceptu C
$\top$	universální třída (všechny třídy)
$\bot$	prázdná třída
$A \equiv B$	ekvivalence třídy/relace
$A \sqsubseteq B$	subsumce třídy/relace
$\neg A$	doplněk (negace)
$A \sqcap B$	průnik (konjunkce)
$A \sqcup B$	sjednocení (disjunkce)
$\exists R.C$	existenční restrikce
$\forall R.C$	universální restrikce

Tabulka 1: Elementy DL

Základní stavební bloky DL (viz tabulka 1) jsou atomické koncepty, atomické role a individua. Atomické koncepty modelují abstraktní množinu objektů jako například kniha, osoba. Atomické role specifikují atributy jako věk, cena, pohlaví. Individua jsou konkrétními instancemi konceptů, například kniha Harry Potter, nebo osoba Pavel. Koncepty v DL mohou být tvořeny definicemi průniku (konjunkce) konceptů, sjednocení (disjunkce) konceptů, doplněk (negace) konceptu, universální restrikce (restrikce hodnot), existenční restrikce, subsumce a ekvivalence. V DL je znalostní báze rozdělena na dvě komponenty, komponenta terminologická (TBox) a komponenta tvrzení (ABox). TBox zahrnuje elementy jako jsou koncepty, taxonomické hierarchie, ekvivalence. V principu TBox obsahuje pouze abstraktní informace a neuvažuje konkrétní instance. ABox definuje tvrzení o individuích a jejich zařazení do hierarchie konceptů.

Například znalosti obsažené v TBox mohou definovat koncepty rodič, žena, matka, nebo matka jejíž děti jsou pouze ženy, následovně:

$$Rodic \equiv Osoba \sqcap \exists ma\_dite.Osoba$$
$$Matka \equiv Rodic \sqcap Zena$$
$$Matka\_deti\_zeny \equiv Matka \sqcap \forall ma\_dite.Zena$$

Základními úkoly při usuzování nad znalostním modelem jsou subsumce (*subsumption*), tj. ověření, že jedna kategorie je podmnožinou jiné kategorie a klasifikace (*classification*), tj. ověření, zda objekt patří do určité kategorie. Více informací o deskripční logice a usuzování v deskripční logice je uvedeno např. v [32].

Pravidla. Kromě deklarativních znalostí je někdy vyžadováno (podle konkrétních požadavků aplikace), aby znalosti o prostředí byly vyjádřeny pomocí libovolných pravidel o reálném světě. Systémy, které pracují s těmito pravidly, mohou být systémy využívající konceptů logického programování (LP), produkčních systémů nebo expertních systémů pro podporu rozhodování. Cílem pravidlových systémů je využití deklarativních a procedurálních znalostí ve tvaru

`if` A_1 `and` A_2 `and ... and` A_n `then B`,

kde A_1 `and` A_2 `and ... and` A_n je označován jako *antencedant* (*head*) pravidla a `B` jako *konsekvent* (*body*) pravidla. Úlohy pravidlových systémů zahrnují hledání řešení na dotazy, kdy dotaz představuje cíl (cíl je možno chápat jako pravidlo bez antecedentu), který

má systém na základě pravidel ve znalostní bázi odvodit. K tomuto účelu se využívají inferenční algoritmy zpětného řetězení (*backward chaining*). Další úlohou je nalezení všech řešení, které je možné odvodit z nějaké výchozí konfigurace (počáteční nastavení paměti systému). K tomuto účelu se využívají inferenční algoritmy dopředného řetězení (*forward chaining*). Podrobné informace o pravidlech, pravidlových systémech a inferenčních algoritmech je možné nalézt v [21].

Předpoklad otevřenosti světa. Usuzování v logických systémech se řídí *předpokladem otevřenosti světa* (open-world assumption–OWA). Pokud například znalostní báze obsahuje informace o dopravních prostředcích $A12$, $A23$, $T15$, potom tato báze obsahuje tři tvrzení

$$\mathit{Prostredek}(A, \mathit{12}), \mathit{Prostredek}(A, \mathit{23}), \mathit{Prostredek}(T, \mathit{15}),$$

kde $\mathit{Prostredek}(p, c)$ je predikát přiřazující prostředku p (A pro autobus a T pro tramvaj) jeho číslo c. Logický systém nemůže usoudit například nic o celkovém počtu dopravních prostředků v této bázi protože nic o tomto počtu neví[2] (logický systém pouze ví, že tři uvedené prostředky v bázi jsou). Aby mohl logický systém učinit takové závěry, tuto bázi je nutné rozšířit o následující znalost:

$$\begin{aligned} \mathit{Prostredek}(p, c) \quad \Leftrightarrow \quad & [p, c] = [A, 12] \vee [p, c] = [A, 23] \vee \\ & [p, c] = [T, 15]. \end{aligned}$$

Logické systémy a předpoklad otevřenosti světa odráží podmínky otevřeného prostředí Webu, kdy informace mohou nekontrolovaně vznikat, zanikat a měnit se. Na druhé straně, pokud implementujeme uvedenou bázi pomocí databázového (např. relačního) systému, při dotazu na počet prostředků dostaneme jasnou odpověď „tři“. Databázové systémy se řídí *předpokladem uzavřenosti světa*, které odpovídá podmínkám, ve kterých jsou databáze implementovány a používány, tzn. v organizacích s přesně stanovenými informačními pravidly.

3 Jazyky

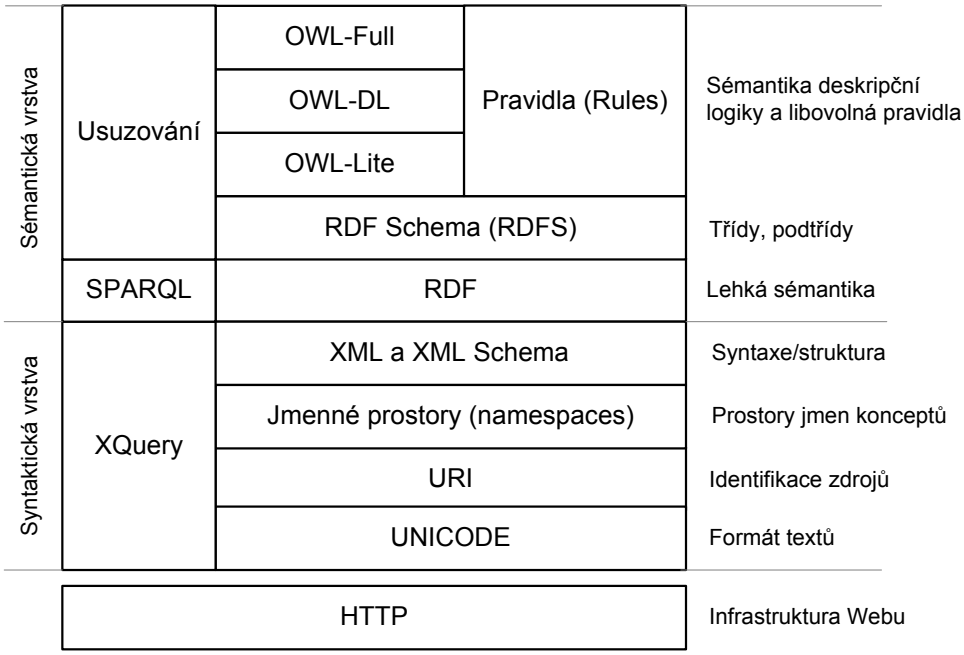
Sémantický web rozšiřuje existující prostředky Webu o sémantickou vrstvu a staví na základních principech infrastruktury WWW[4]:

- Universální propojování zdrojů[3] na Webu pomocí linků,
- Otevřenou architekturu Webu, která definuje otevřené, standardrní a volně dostupné technologie,
- Oddělení vrstev (seperation of layers), které využívá standardní rozhraní umožňující nezávislou inovaci.

[2]Logické systémy využívají princip NaF (*Negation as Failure*), který umožňuje „dokázat“ *not* P pouze v případě, že P nelze dokázat. Více informací je uvedeno v [32].

[3]Pojem zdroj je chápan ve smyslu *resource*, tzn. základní element Webu (volně lze přiřadit zdroj k dokumentu).

Klíčové technologie, které realizují tyto principy, a které jsou základem pro jazyky Sémantického webu zahrnují adresaci URI[14] (Uniform Resource Identification), jazyk XML[11] a protokol HTTP[18]. Jazyky Sémantického webu využívají URI pro jednoznačnou identifikaci zdrojů (konceptů, rolí, objektů), které jsou předmětem popisu znalostí v podobě ontologií. Jazyk XML definuje syntaktický formát pro výměnu zpráv na Webu, jazyky Sémantického webu jej využívají jako jeden z formátů pro serializaci ontologií (jejich syntaktickou reprezentaci). Protokol HTTP je základním aplikačním protokolem pro komunikaci na Webu, který tak umožňuje fyzický přístup ke zdrojům.



Obrázek 1: Jazyky Sémantického webu

Obrázek 1 zobrazuje jazyky Sémantického webu, tak jak jsou definovány ve W3C[4]. Jazyky Sémantického webu jsou postaveny na základní infrastruktuře Webu, kterou tvoří adresace URI a protokol HTTP. Jazyky Sémantického webu také respektují různorodost požadavků na reprezentaci znalostí. Proto jsou tyto jazyky definovány ve vrstvách, kdy jednotlivé vrstvy nabízí různou sílu vyjadřovací schopnosti jazyka ve smyslu reprezentace znalostí. Nejnižší vrstva je vrstvou syntaktické reprezentace, která je zastoupena jazykem XML, popř. dalšími serializačními formáty jako jsou N3[5] nebo Turtle[12]. Druhá vrstva (vrstva RDF[8]) nabízí reprezentaci umožňující definovat libovolné vztahy mezi objekty, popř. jejich kategoriemi bez explicitní specifikace významu vazeb či objektů. Třetí vrstva (vrstva RDF Schema—RDFS[5]) nabízí minimální sémantickou vyjadřovací schopnost, která definuje význam některých elementů, konkrétně třídy (Class) a taxonomické relace podtřídy (subClassOf). Sémantiku, kterou je možné vyjádřit pomocí jazyka RDFS je někdy též označována jako lehká sémantika (lightweight semantics). Čtvrtá vrstva (vrstva OWL[26]) nabízí pokročilou reprezentaci znalostí na úrovni deskripční logiky a poslední vrstva (vrstva pravidel) disponuje vyjadřovací schopností na úrovni procedurálních znalostí.

[4]W3C prezentuje skupinu jazyků Sémantického jako Semantic Web Layer Cake nebo Semantic Web Languages Stack, `http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html`

[5]`http://www.w3.org/DesignIssues/Notation3`

Vrstvení jazyků Sémantického webu má velký význam. Vyjadřovací schopnost vyšší vrstvy v zásadě zahrnuje vyjadřovací schopnost vrstvy nižší (platí pro vrstvu RDF, RDFS a částečně pro OWL). Jazyky je tak možné škálovat a využít podle konkrétních požadavků na expresivitu znalostního modelu. Inženýr má před vlastním modelováním za úkol zvolit úroveň této expresivity (podle požadavků aplikace) a zároveň zvážit možnosti nástrojů pro usuzování a dotazování. S rostoucí expresivitou jazyka roste složitost algoritmů pro usuzování. Další vlastností je extensibilita modelu. Jednotlivé konstrukty gramatiky jazyků jsou definovány ve jmenných prostorech příslušného jazyka[6]. Vlastní model, který je v daném jazyku popsán, definuje svoje vlastní jmenné prostory což umožňuje model rozšiřovat a přepoužít jeho znalosti v rámci distribuované sítě: model tak mohou využívat a nezávisle rozvíjet různé skupiny uživatelů, je možné oddělit různé verze modelu, atp. aniž by mezi těmito modely docházelo k interferenci. Základ infrastruktury Webu a jazyka XML v neposlední řadě umožňuje zacházet s modely standardním způsobem např. při komunikaci nebo manipulaci s daty na syntaktické úrovni.

3.1 XML

eXtensible Markup Language (XML) patří do množiny značkovacích jazyků, které jsou využívány pro popis hierarchických struktur textových dokumentů (řetězce UNICODE[7] znaků) s využitím tzv. tagů. Tag je značka, která vždy obsahuje počáteční a koncový tag společně definující element. Např. element `<jmeno>Tomas</jmeno>` je složen z tagu `<jmeno>` a textového konstruktu `Tomas`. Tagy mohou být chápány jako meta-data popisující jiné konstrukty, na úrovni XML ovšem nemají tyto meta-data žádný formální základ. Jazyk XML definuje specifikaci pro popis syntaxe (struktury) XML documentu pomocí XML Schema[9]. Ačkoliv XML Schema poskytuje možnosti zapsat pravidla, kterými se daný XML dokument musí řídit, tyto pravidla nemají logický základ. XML Schema není proto z pohledu Sémantického webu považováno za sémantiku, ale strutkuru či syntaxi. XML Schema také definuje základní datové typy pro řetězec, integer, atd., které jsou využívány v sémantických jazycích.

3.2 RDF

Resource Description Framework (RDF) je základní vrstvou v rodině sémantických jazyků. RDF reprezentuje informace v grafovém modelu pomocí trojic (triples), tzn. výrazů ve tvaru `<subjekt, predikát, objekt>`. Subjekty a objekty spojené predikáty vytvářejí grafovou strukturu. RDF nedefinuje sémantiku pro subjekty, objekty ani predikáty, ale definuje prostředky pro rozlišení objektů a kategorií (pomocí predikátu `rdf:type`). Když je objekt popsán vlastností `rdf:type`, hodnota této vlastnosti je chápána jako zdroj, který reprezentuje kategorii nebo třídu objektů a subjekt této vlastnosti je chápán jako instance této kategorie nebo třídy. Například trojice

```
<ex:Petr, rdf:type, ex:Osoba>
```

[6]Pro definici jmenných prostorů využívají jazyky na Webu identifikátory URI. Jmenné prostory jsou v modelech označovány zkráceně pomocí prefixů, v tomto dokumentu jsou použity standardní prefixy `rdf`, `rdfs`, `owl` a `ex` pro příklady. Např. prefix `rdf` odkazuje na jmenný prostor jazyka RDF, který je definován pomocí URI `http://www.w3.org/1999/02/22-rdf-syntax-ns#`. Obvykle platí, že jmenné prostory URI jsou odkazy URL, kde je možné nalézt specifikaci daného jmenného prostoru.

[7]UNICODE je standard pro reprezentaci textových dokumentů, `http://en.wikipedia.org/wiki/UNICODE`

popisuje instanci `Petr` třídy `Osoba`. RDF tedy umožňuje vyjádřit základní vztahy přináležitosti prvku do kategorie (ve smyslu relace *Member* v sekci 2). RDF dále definuje tzv. kontejnery (např. `rdf:bag`, `rdf:seq`) pro objekty definující skupinu objektů (v podobném smyslu relace *PartOf* v sekci 2).

RDF definuje způsob jak převést modely zachycené v RDF do reprezentace nutné pro přenos a další zpracování zpráv na nižších úrovních Webu a jeho infrastruktury. Jednou z takových reprezentací je RDF/XML[6], která definuje pravidla jak zapisovat model v RDF v jazyce XML. Mezi další serializační formáty pro jazyk RDF patří Notation 3 (N3) nebo Turtle. Tyto formáty mají jednodušší syntax (RDF dokumenty zapsané v N3 jsou obvykle čitelnější a fyzicky menší než dokumenty zapsané v RDF/XML). Výhoda RDF/XML ovšem spočívá hlavně v možnosti využití standardních nástrojů pro XML (např. parser, dotazování a transformace pomocí XQuery[10], XSLT[2] nebo XPath[1].

3.3 RDFS

Resource Description Framework Schema (RDF Schema, zkráceně RDFS) je základním jazykem pro reprezentaci ontologií s jednoduchou sémantikou. RDFS je rozšířením jazyka RDF a definuje konstrukty pro vyjádření tříd objektů, vlastností objektů a hierarchie tříd a vlastností.

`rdfs:Class` definuje ktegorii (třídu objektů). Třídou může být například osoba jejíž instance je potom spojená s třídou osoba pomocí predikátu `rdf:type`. Tento výraz můžeme zapsat jako trojici

```
<ex:Petr, rdf:type, ex:Osoba>
<ex:Osoba, rdf:type, rdfs:Class>
```

Třídy lze spojovat do hierarchií pomocí vlastnosti `rdfs:subClassOf` (ve smyslu relace *Subclass* v sekci 2). Například trojice

```
<ex:Osoba, rdfs:subClassOf, ex:Savec>
<ex:Savec, rdf:type, rdfs:Class>
```

definuje třídu `Osoba` jako podtřídu třídy `Savec`. RDFS dále definuje vlastnosti (property) jako třídu RDF vlastností. Každý prvek této třídy je predikátem v RDF. Např. vlastnost `ex:Studuje` v trojici

```
<ex:Petr, ex:Studuje, ex:CVUT>
```

je možno definovat pomocí trojic

```
<ex:Studuje, rdfs:domain, ex:Osoba>
<ex:Studuje, rdfs:range, ex:Skola>
<ex:Skola, rdf:type, rdfs:Class>
```

kde `rdfs:domain` je predikát definující doménu vlastnosti, tj. třídu pro kterou platí, že levá strana vlastnosti (subjekt) musí být instancí této třídy; `rdfs:range` je potom rozsah této vlastnosti, tj. třída pro kterou platí, že pravá strana vlastnosti (objekt) musí být instancí této třídy. Pro náš příklad tedy platí, že

```
<ex:Petr, rdf:type, ex:Osoba>
<ex:CVUT, rdf:type, ex:Skola>
```

RDFS také umožňuje definovat hierarchii vlastností pomocí predikátu `rdfs:subPropertyOf`. RDFS dále definuje další specializované vlastnosti umožňující popsat znalostní model, jako např. přidat textové poznámky nebo komentáře k definovaným konceptům, odkazy, atp.

3.4 OWL

Web Ontology Language (OWL) představuje vrstvu jazyků pro Sémantický web, které implementují sémantiku deskripční logiky. Jak již bylo řečeno, protože volbu úrovně expresivity jazyka je nutné zvažovat s ohledem na výpočetní složitost nástrojů pro usuzování, vrstva OWL specifikuje další 3 varianty jazyka: OWL-Lite, OWL-DL a OWL-Full.

OWL-Lite nabízí expresivitu na úrovni taxonomické hierarchie (stejně jako RDF a RDFS), ekvivalence tříd a vlastností, dále transitivní, symetrické, inverzní vlastnosti, existenční a universální restrikci, jednoduché omezení kardinality (omezenou na hodnoty 0 a 1), průnik tříd.

OWL-DL nabízí maximální expresivitu při současném zachování rozhodnutelnosti (decidability) a možnosti využít dostupné nástroje pro usuzování. OWL-DL obsahuje všechny jazykové konstrukty OWL, které mohou být ale použity s určitým omezením (např. nelze použít omezení kardinalit na vlastnosti, které jsou definovány jako transitivní). OWL-DL umožňuje (navíc k OWL-Lite) definovat disjunkci tříd, sjednocení tříd, doplněk třídy, libovolné omezení kardinality.

OWL-Full nabízí maximální expresivitu s použitím všech jazykových OWL konstruktů, nicméně nezaručuje rozhodnutelnost při usuzování. OWL Full navíc dovoluje měnit význam předdefinovaných konstruktů jazyka RDF a OWL. V současnoti neexistují efektivní algoritmy, které by umožňovaly usuzování v OWL-Full využívající všechny její vlastnosti.

Příklad jednoduché znalosti zapsané v OWL, definující kategorie mužů a žen jako disjunktní, je:

```
<ex:Osoba, rdf:type, owl:Class>
<ex:Zena, rdfs:subClassOf, ex:Osoba>
<ex:Muz, rdfs:subClassOf, ex:Osoba>
<ex:Zena, owl:disjointWith, ex:Muz>
```

3.5 Pravidla

V některých aplikacích Sémantického webu je vyžadováno vyjádření libovolných pravidel. Příkladem je aplikace technologií Sémantického webu pro popis rozhraní webových služeb. Webové služby reprezentují distribuované objekty na Webu, které je možné popsat pomocí podmínek, které musí platit ve stavu před jejich spuštěním (tzv. preconditions) a podmínek, které musí platit ve stavu po jejich spuštění (tzv. postconditions). Takové podmínky je možné zapsat jako logické výrazy a použití technologií Sémantického webu vyžaduje rozšíření jazyků o možnosti pracovat s libovolnými pravidly. Touto problematikou se zabývá iniciativa Sémantických webových služeb, která během posledních let významě přispěla do vývoje jazyků Sémantického webu pro pravidla. Důležitým jazykem této iniciativy je jazyk WSML[15] (Web Service Modeling Language), který je vyvíjen s cílem kombinovat možnosti jazyků pracující s deklarativními znalostmi založené na deskripční logice (varianta WSML-DL) s jazyky pracující s procedurálními znalostmi založené na logickém programování (varianta WSML-Rule). Zatímco jazyk WSML-DL je syntaktická varianta jazyka OWL-DL, jazyk WSML-Rule se stal vstupní specifikací pro W3C pracovní skupinu RIF[8] (Rule Interchange Format). Pracovní skupina RIF má za úkol vytvořit specifikaci pro jazyk Sémantického webu pracující s pravidly.

Následující příklad definuje pravidlo v jazyce WSML[9] „manažeři také pracují“, které

[8] `http://www.w3.org/2005/rules/wiki/RIF_Working_Group`

[9] WSML používá uživatelsky přívětivou (human-readable) syntaxi a také definuje pravidla reprezentace WSML v XML.

říká, že pokud manažer spravuje nějakou pobočku, potom musí mít tento manažer danou pobočku jako pracovní místo. V tomto příkladě je `Manager` třída, `spravuje` a `pracuje` vlastnosti (property), `?x` a `?y` jsou proměnné.

```
axiom manazeri_take_pracuji
  definedBy
    ?x[spravuje hasValue ?y] memberOf Manager
  implies
    ?x[pracuje hasValue ?y].
```

3.6 Dotazování a usuzování

Základním prostředkem pro dotazování nad daty Sémantického webu je jazyk SPARQL[33] (SPARQL Protocol and RDF Query Language[10]). SPARQL umožňuje formulovat dotazy pomocí vzorů RDF trojic a logických operací konjunkce a disjunkce. Současná verze jazyka SPARQL[33] umožňuje provádět jednoduché odvozování na úrovni RDF, ale neumožňuje pokročilé usuzování nad RDFS nebo OWL ontologiemi. Takové rozšíření jazyka SPARQL je v současnosti předmětem výzkumu[29].

Dotaz v jazyce SPARQL má 3 základní části: část prolog definuje jmenné prostory a prefixy (pomocí konstruktu `PREFIX`), které dotaz používá, část hlavička (head) definuje jaká data mají být výsledkem dotazu (pomocí konstruktů `CREATE`, `SELECT` nebo `ASK`) a část tělo (body) definuje množinu zdrojových RDF grafů (pomocí konstruktu `FROM`– tato část může být vynechána, protože graf může být implicitně zadán), podmínky v podobě grafových vzorů včetně proměných a výrazy pro řazení výsledků (order). Příklad jednoduchého dotazu v jazyce SPARQL, který vrací jméno a adresu osob je následující (klíčové slovo `a` reprezentuje predikát `rdf:type`):

```
PREFIX ex: <http://www.vitvar.com/ns/ex>
SELECT ?jmeno ?adresa
WHERE {
  ?osoba a ex:Osoba.
  ?osoba ex:jmeno ?jmeno.
  ?osoba ex:adresa ?adresa.
}
```

Kromě SPARQL existují ještě další mechanismy pro dotazování nad RDF daty, jako například RDFQ[11] nebo RDQL[7]. Tyto specifikace v současnosti nejsou W3C standardy.

Mezi nejdůležitější nástroje pro usuzování nad ontologiemi zapsané v jazycích s expresivitou na úrovni deskripční logiky patří Pellet[12], RacerPro[13], Fact++[14] a KAON2[15]. Nástroje pro usuzování s pravidly jsou Flora-2[16], XSB[17], MINS[18], IRIS[19] a KAON2.

[10] SPARQL akronym je tzv. rekurzivní akronym.
[11] `http://sw.nokia.com/rdfq/RDFQ.html`
[12] `http://clarkparsia.com/pellet`
[13] `http://www.racer-systems.com`
[14] `http://owl.man.ac.uk/factplusplus`
[15] `http://kaon2.semanticweb.org`
[16] `http://flora.sourceforge.net`
[17] `http://xsb.sourceforge.net`
[18] `http://dev1.sti2.at/mins`
[19] `http://www.iris-reasoner.org`

4 Ontologie

Jazyky Sémantického webu slouží pro popis znalostních modelů, které se nazývají ontologie. Ontologie není pouze znalostní model, který je popsán jazykem s vyšší vyjadřovací silou, ontologie navíc slouží jako primární prostředek pro dosažení interoperability na Webu. Pojem ontologie původně pochází z filozofie, Thomas Gruber definoval její použití pro účely znalostního inženýrství jako formální a explicitní specifikaci sdílené konceptualizace určité oblasti (domény)[20]. Význam této definice je následující:

- „Formální“ a „explicitní“ znamená, že ontologie vyjadřuje znalosti pomocí určitého ontologického jazyka s určitou vyjadřovací schopností, a která má formální logický základ.
- „Sdílená“ znamená, že ontologie je využívaná všemi členy komunity. Každý člen této komunity se zavazuje, že bude ontologie používat pro popis konceptů dané domény. Ontologie se tak stává sociálním závazkem pro danou komunitu.
- „Konceptualizace“ znamená, že ontologie definuje koncepty domény na určité úrovni abstrakce, která odpovídá požadavkům na modelování domény.

Ontologie je tedy popisem určité doménové znalosti, která může být využita v informačních nebo procesních modelech systémů. Ontologie je navíc produktem kolaborativního návrhu. Členové komunity se nejenom zavazují ontologii používat, ale také se podílejí na jejím návrhu, vývoji a rozvoji. Ontologie je tedy spravidla výsledkem sociálního konsensu. Z tohoto důvodu je ontologie modelem, který představuje spoječný jazyk komunikace mezi členy komunity a potažno systémy, které ontologii využívají. Slouží tedy jako jeden z prostředků pro dosažení interoperability.

Vzhledem k tomu, že prakticky není možné aby ontologie popisovala všechny možné koncepty, které se mohou v doméně vyskytnout, inženýr by měl po analýze požadavků na model nejříve zjistit existující ontologie, které by mohl pro modelování využít. Takových ontologií může být mnoho a mohou sem patřit tzv. vyšší ontologie (upper ontologies) i jiné doménové ontologie, které svým rozsahem zasahují do modelované domény. Inženýr tedy přijme takové ontologie jako výchozí modely pro svou doménu a rozšíří ji o koncepty podle svých požadavků. Tyto koncepty spravidla vytváří v odděleném jmenném prostoru. Vývoj ontologií je ale typicky otevřený proces, který umožňuje komukoliv se na vývoji ontologie podílet. Inženýr může proto svými poznatky o rozšířené verzi ontologie přispět i do základní definice původní ontologie (spravidla v rámci standardizačního procesu přislušné organizace/skupiny, která zodpovídá za vývoj dané ontologie).

4.1 Typy ontologií

Komunita Sémantického webu rozlišuje následující základní typy ontologií.

Vyšší ontologie. Vyšší ontologie (Upper Ontology) popisují velmi obecné koncepty které se typicky vyskytují v každé doméně, a které mohou být různými doménami sdíleny. Tímto způsobem je možné zaručit velmi základní interoperabilitu na nejvyšší úrovni mezi velkým množstvím ontologií. Důležité vyšší ontologie zahrnují Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)[3] a WordNet[20].

[20] `http://wordnet.princeton.edu`

Doménové ontologie. Doménové ontologie popisují koncepty konkrétní domény. Doménové ontologie mohou využívat vyšší ontologie a definovat specializovanější koncepty. Skupina W3C Semantic Web Education and Outreach group[21] zodpovídá za komunitní projekt s názvem Linking Open Data community[22], jejímž cílem je rozšířit Web o data různých datových sad. Datové sady Open Link projektu aktuálně zahrnují přes 4.7 biliónů RDF trojic. Tyto datové sady jsou velmi dobrým příkladem doménových ontologií, které vznikají kolaborativním způsobem. Mezi nejvyznamější patří: DBpedia[23] (datová sada obsahující RDF data z Wikipedie), FOAF[24] (ontologie pro popis osobních profilů), SIOC[25] (ontologie pro popis provázaných sociálních komunit a sítí), DBLP[26] (ontologie popisující bibliografické zdroje) a mnoho dalších.

Ontologie pro popis služeb a procesů. Ontologie služeb (service ontology) popisují meta-koncepty webových služeb jako například definici funkcionality služby, její kategorizaci, rozhraní, atd. Uznávané ontologie z této oblasti jsou WSMO[31], WSMO-Lite[34], OWL-S[24]. Procesní ontologie popisují koncepty procesů. Výzkumem této oblasti za zabývá EU projekt SUPER[27], který definuje několik typů takových ontologií.

4.2 Interoperabilita ontologií

Jak již bylo řečeno, důležité poslání ontologií je zajištění interoperabilty prostřednictvím sdílených znalostí. Realisticky ovšem nelze předpokládat, že bude vždy existovat jedna ontologie, která popisuje kompletní prostředí. Jednotlivé aplikační domény nemají jednoznačně vymezené hranice a proto mezi nimi existují překryvy. Někdy je rovněž složité docílení konsensu při vytváření ontologií (např. z organizačních důvodů nebo z důvodů vzájemného neporozumění), proto mohou při vývoji ontologií vznikat podobné vývojové větve. Z tohoto důvodu je důležité zajistit interoperabilitu ontologií pomocí mapování konceptů jedné ontologie na koncepty druhé ontologie (takové mapování se v Sémantickém webu označuje jako alignment–volně přeloženo jako slaďování).

Mapování ontologií definuje dvě fáze: (1) inženýr pomocí příslušných nástrojů namapuje koncepty zdrojové ontologie na koncepty cílové ontologie a (2) software, který se nazývá *mediator*, fyzicky provede transformaci instancí, které odpovídají těmto konceptům. Současný výzkum v oblasti mapování ontologií využívá jazyky pravidel, pomocí kterých je možno popsat mapování ve formě sémantických vztahů, které existují mezi dvěma ontologiemi. Konkrétně můžeme například pomocí takových pravidel vyjádřit, že třídy jedné ontologie jsou ekvivalentní s třídami druhé ontologie. Pomocí logických výrazů můžeme popsat pravidla, které jednoznačně definují jak data obsažena v instanci zdrojové třídy mohou být obsažena v instanci cílové třídy.

```
axiom mapovaci_pravidlo definedBy
  mediated(?x, Osoba)[jmeno hasValue ?a] memberOf o1#Podnik
    :- ?x[nazev hasValue ?a] memberOf o2#Firma.
```

[21] http://www.w3.org/2001/sw/sweo/
[22] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[23] http://dbpedia.org/About
[24] http://www.foaf-project.org/
[25] http://sioc-project.org/
[26] http://www4.wiwiss.fu-berlin.de/dblp/
[27] http://www.ip-super.org/

Příklad výše ukazuje pravidlo v jazyce WSML mezi konceptem `Podnik` definovaný ve zdrojové ontologii identifikované prefixem `o1` a konceptem `Firma` definovaný v cílové ontologii `o2`. Konstrukt `mediated(X,C)` reprezentuje identifikátor nově vytvořené cílové instance, kde `X` je zdrojová instance, ktará je transformována, a `C` je cílový koncept, na který se provádí mapování. Toto pravidlo provádí mapování vlastnosti `jmeno` konceptu `Podnik` na vlastnost `nazev` konceptu `Firma` (symbol `:-` vyjadřuje implikaci). Mapovací pravidla pomocí jazyka WSML jsou předmětem práce v [27, 22].

Problémy interoperability mezi ontologiemi mohou být velmi komplexní a proto není možné vždy zaručit plně automatizovanou podporu pro definici takových pravidel. Důkazem je studie iniciativy Ontology Alignment Evaluation[17], která ukazuje, že výsledky prvních 5 nejlepších systémů se pohybují mezi 60% a 80% pro koeficient přesnosti (precision) a mezi 65% a 70% pro koeficient úplnosti (recall)[28]. Z tohoto důvodu fáze mapování ontologií zatím vždy zavisí na manuální podpoře inženýra. Více informací o mapování ontologií je možno nalézt v [16].

4.3 Anotace obsahu na Webu

Sémantický web definuje sémantickou vrstvu z hlediska jazyků popisující zdroje na Webu. Cílem této sémantiky je obohatit existující popisy zdrojů tak aby mohly být využity inteligentní nástroje pro vyhledávání, usuzování nebo interoperabilitu systémů. Aby tyto technologie mohly být využity ve stávajícím prostředí, musí existovat prostředky, které umožní propojit (pomocí tzv. anotací) nesémantickou vrstvu se sémantickou vrstvou.

Anotace umožňuje propojení dat, které jsou k dispozici ve formátu XML nebo HTML, se sémantickými daty. Tato anotace je vyžadována pro aplikace, které pracují se sémantickými daty, ale informace jsou k dispozici v nesémantické formě (XML, HTML). Pro tento typ anotace je zapotřebí (1) nadefinovat reference mezi sémantickými konstrukty a nesémantickými koncepty na úrovni schémat a (2) nadefinovat transformace mezi nesémantickým modelem a sémantickým modelem pro potřeby transformace dat (instancí). Tomuto typu transformace se říká lifting (povýšení) a lowering (ponížení)[23][29]. V současné době jsou ve W3C zakotveny dva důležité standardy, které specifikují anotaci pro HTML/XML resp. XML Schema formou *začlenění sémantických dat* a *formou odkazů na sémantický model*.

Anotace formou začlenění sémantických dat. Tuto anotaci je možno využít na dokumenty v XML/HTML a je definována specifikací Resource Description Framework – in attributes (RDFa)[30] a Gleaning Resource Descriptions from Dialects of Languages (GRDDL)[19]. Výsledkem této anotace je jeden dokument HTML/XML, který obsahuje jak nesémantická tak sémantická data. RDFa definuje způsob začlenění RDF grafu do XML/HTML elementů pomocí speciálně definovaných atributů (např. `property`, `content`, `datatype`, `typeof`, apod.) podle definovaných pravidel. GRDDL potom definuje transformaci (např. v XSLT nebo v XQuery), která podle anotačních pravidel extrahuje sémantická data z dokumentu. V praxi může být tato specifikace využita například pro anotaci HTML dokumentů, kdy původní HTML je využito pro prezentaci v prohlížeči a RDF je využito k automatizovanému zpracování specializovanými systmémy, např. při in-

[28] Koeficienty přesnosti a úplnosti (precision, recall) se používají k hodnocení míry relevance u vyhledávaích systémů (information retrieval)

[29] Povýšení ve smyslu z nižší nesémantické (syntaktické) úrovně na vyšší (sémantickou) úroveň (pro ponížení platí opačně)

dexaci HTML dokumentů. Tuto anotaci aktuálně podporuje Google, který využívá takto získaná meta-data při zobrazování výsledků hledání (v tzv. *snippets*)[30].

Anotace formou odkazů na model. Tato anotace využívá odkazů (angl. *model references*) a je možno ji využít na XML Schema. Tuto anotaci definuje specifikace Semantic Annotations for Web Service Description Language and XML Schema (SAWSDL)[23], která byla vytvořena v rámci stejnojmenné pracovní skupiny ve W3C[31], a která je součástí specifikace pro anotaci webových služeb. Pomocí této specifikace je možné definovat (1) linky mezi XML elementy v XML Schema a ontologickými koncepty a (2) transformace mezi XML daty a ontologickými daty. Pro tyto transformace je možné využít standardních jazyků XQuery nebo XSLT, nicméně jejich použití je dosti omezené. Transformační dotazy definované v těchto jazycích jsou velice komplexní a složité protože musí mapovat stromovou strukturu XML do grafové struktury RDF. Z tohoto důvodu vzniká nový jazyk, který kombinuje možnosti XQuery pro formulaci dotazů a transformací na XML s možnostmi jazyka SPARQL pro formulaci dotazů a transformací na RDF. Tato specifikace má název XSPARQL[13] a v současnosti je předmětem standardizace ve W3C.

5 Závěr

Sémantický web rozšiřuje možnosti stávajícího Webu o znalostní vrstvu, která umožní aplikovat metody inteligentních systémů v tomto prostředí pro vylepšenou práci s informacemi a interoperabilitu systémů. Aktivity Sémantického webu jsou primárně zaměřeny na explicitní reprezentaci znalostí, která předpokládá, že uživatelé budou takové znalosti vytvářet, udržovat a používat. Reprezentace znalostí ovšem přináší vyšší míru složitosti jednak při tvorbě znalostních modelů, ale také při aplikaci metod, které tyto modely využívají. Sémantický web proto nabízí různé úrovně sémantiky, kterou je možno aplikovat na problémy webového inženýrství podle konkrétních požadavků. Je na zvážení inženýrů, kterou úroveň zvolit, aby výsledný systém odpovídal požadvkům inteligence a zároveň byl použitelný z hlediska jeho výkonu.

I když touha po světě plném inteligence a automatizace je v současnosti veliká, úplná realizace takového světa zatím není možná z důvodů omezených kapabilit nástrojů, které současné technologie nabízí. Sémantický web je proto horkým tématem v mnoha vědeckých programech a projektech, které se snaží o naplnění vize globální a inteligentní platformy pro využití všemi, všude a kdekoliv.

[30] `http://www.google.com/support/webmasters/bin/answer.py?answer=146898`
[31] `http://www.w3.org/2002/ws/sawsdl/`

Literatura

[1] XML Path Language (XPath). Recommendation, W3C, November 1999. http://www.w3.org/TR/xpath.

[2] XML Transformations. Recommendation, W3C, November 1999. http://www.w3.org/TR/xslt.

[3] Ontology Library. WonderWeb Deliverable, Laboratory for Applied Ontology, August 2003. http://www.loa-cnr.it/Papers/D18.pdf.

[4] Architecture of the World Wide Web, Volume One. Recommendation, W3C, December 2004. http://www.w3.org/TR/webarch/.

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

[6] RDF/XML Syntax Specification. Recommendation, W3C, February 2004. Http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/.

[7] RDQL - A Query Language for RDF. W3C Member Submission, W3C, January 2004. http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/.

[8] Resource Description Framework (RDF): Concepts and Abstract Syntax, February 2004. http://www.w3.org/TR/rdf-concepts/.

[9] XML Schema Part 1: Structures Second Edition. Recommendation, W3C, October 2004. http://www.w3.org/TR/xmlschema-1/.

[10] XQuery 1.0: An XML Query Language. Recommendation, W3C, January 2007. http://www.w3.org/TR/xquery/.

[11] Extensible Markup Language (XML) 1.0. Recommendation, W3C, November 2008. http://www.w3.org/TR/2008/REC-xml-20081126/.

[12] Turtle - Terse RDF Triple Language. W3C Team Submission, W3C, January 2008. http://www.w3.org/TeamSubmission/turtle/.

[13] W. Akhtar, J. Kopecký, T. Krennwallner, and A. Polleres. XSPARQL: Traveling between the XML and RDF Worlds - and Avoiding the XSLT Pilgrimage. In *ESWC*, pp. 432–447. 2008.

[14] T. Berners-Lee, *et al.* Uniform Resource Identifiers (URI): Generic Syntax. Tech. rep., The Internet Engineering Task Force (IETF), Aug 1998.

[15] J. de Bruijn *et al.* The Web Service Modeling Language WSML. Tech. rep., October 2005. Http://www.wsmo.org/TR/d16/d16.1/v0.3/20051005/.

[16] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag New York Inc, 2007.

[17] J. Euzenat, *et al.* Results of the Ontology Alignment Evaluation Initiative 2006. In *Proceeding of International Workshop on Ontology Matching (OM-2006)*, vol. 225, pp. 73–95. CEUR Workshop Proceedings, Athens, Georgia, USA, November 2006.

[18] R. Fielding, *et al.* Hypertext Transfer Protocol – HTTP/1.1 (RFC 2616). Tech. rep., The Internet Engineering Task Force (IETF), Jun 1999.

[19] Gleaning Resource Descriptions from Dialects of Languages (GRDDL). Recommendation, W3C, September 2007. `http://www.w3.org/TR/grddl/`.

[20] T. R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, 1993.

[21] M. Huth and M. Ryan. *Logic in computer science*. Cambridge University Press Cambridge, 2004.

[22] M. Kerrigan and A. Mocan. The Web Service Modeling Toolkit. In *ESWC*, pp. 812–816. 2008.

[23] J. Kopecký, T. Vitvar, C. Bournez, and J. Farrell. SAWSDL: Semantic Annotations for WSDL and XML Schema. *IEEE Internet Computing*, 11(6):60–67, 2007.

[24] D. Martin, *et al.* OWL-S: Semantic Markup for Web Services, W3C Member Submission. Tech. rep., W3C, 2004. `http://www.w3.org/Submission/OWL-S/`.

[25] V. Mařík, O. Štěpánková, and J. Lažanský, (eds.) *Umělá inteligence (1)*. Academia Praha, 1993.

[26] D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview. Recommendation 10 February 2004, W3C, 2004. `http://www.w3.org/TR/owl-features/`.

[27] A. Mocan and E. Cimpian. An Ontology-Based Data Mediation Framework for Semantic Environments. *Int. J. Semantic Web Inf. Syst.*, 3(2):69–98, 2007.

[28] C. Peirce, C. Hartshorne, and P. Weiss. *Collected papers of charles sanders peirce*. Harvard University Press, 1932.

[29] A. Polleres. From SPARQL to rules (and back). In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, (eds.) *WWW*, pp. 787–796. ACM, 2007.

[30] RDFa in XHTML: Syntax and Processing. Candidate Recommendation, W3C, June 2008. `http://www.w3.org/TR/rdfa-syntax/`.

[31] D. Roman, *et al.* Web Service Modeling Ontology. *Applied Ontology*, 1(1):77–106, 2005.

[32] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2003.

[33] SPARQL Query Language for RDF. W3c recommendation, W3C, October 2008. `http://www.w3.org/TR/rdf-sparql-query/`.

[34] T. Vitvar, J. Kopecký, J. Viskova, and D. Fensel. WSMO-Lite Annotations for Web Services. In *ESWC*, pp. 674–689. 2008.

[35] J. Vlček. *Systémové inženýrství*. Vydavatelství ČVUT, 1999.

Odborný životopis

Ing. Tomáš Vitvar, Ph.D. (*1974)

Vzdělání

- 2004 - titul Ph.D, ČVUT Praha, Fakulta dopravní, obor Inženýrská informatika v dopravě a spojích. Disertační práci obhájil na téma Metasystém jako základ pro datovou interoperabilitu.
- 1998 - titul Ing., ČVUT Praha, Fakulta stavební, obor Systémové inženýrství.

Odborná praxe

- 2010 - vědecký pracovník a pedagog na Fakultě dopravní ČVUT Praha, Ústav informatiky a telekomunikací.
- 2008 - vědecký pracovník v Institutu informatiky při Leopold-Franzens Universität Innsbruck, Rakousko.
- 2004 - vědecký pracovník v institutu Digital Enterprise Research Institute při National University of Ireland, Galway, Irsko.
- 2002 - asistent ředitele IT v NKT Cables, GmbH, Kolín nad Rýnem, Německo.
- 1998 - architekt softwarových projektů, ESV spol. s r.o., Česká republika

Pedagogická praxe

- ČVUT Praha, Fakulta dopravní, přednášky v předmětu Informační systémy a technologie (2010),
- Leopold-Franzens Universität Innsbruck, přednášky v předmetu Inteligentní systémy (2008).
- National Univeristy of Ireland, přednášky v předmětech Sémantický web a sémantické webové služby, Programování Java (2005, 2006, 2007)

Publikační činnost[32]

- Spolueditor a spoluautor knihy Semantic Technologies for E-Government (Springer-Verlag, 2010)
- Spoluautor 3 monografií (vydané Springer, Německo a USA).
- Autor či spoluautor 5 článků v impaktovaných nebo recenzovaných časopisech. Impaktované časopisy uvedené v ISI Web of Knowledge jsou IEEE Internet Computing, IEEE Software a Elsevier Advances in Computers.
- Autor či spoluautor více jak 30 článků ve sbornících mezinárodních konferencí (IEEE, Springer, ACM). Sborníky jsou uvedené v ISI Web of Knowledge.
- Celkový počet citací všech publikací autora přesahuje 400[33].

[32] Kompletní seznam publikací je k dispozici na adrese `http://www.vitvar.com/publications#t=all`

[33] Zdroj: Google Scholar (`http://www.vitvar.com/publications/#t=citations`). Počet citací podle ISI Web of Knowlege je kolem 47.