VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

VYUŽITÍ ONTOLOGIÍ K POPISU WEBOVÉ STRÁNKY

AUTOR PRÁCE
AUTHOR

Martin Milička

BRNO 2012

Obsah

1 Úvod

Symbolem dnešní doby je Internet. Uživatelé jej stále častěji využívají ke sdílení informací. Díky Internetu můžeme být během několika málo okamžiků informováni o událostech, které se stanou na druhém konci světa. V současné době nejsou velké vzdálenosti žádnou bariérou při sdílení informací.

K datům v Internetu lze přistupovat několika způsoby. Jedním z nich je World Wide Web (WWW), zkráceně web. Přestože jsou pojmy Internet a web dvě různé věci, v běžné komunikaci je slovo Internet často nahrazováno slovem web. Musíme si uvědomit, že Internet označuje pouze počítačovou síť, na které funguje služba WWW, jež nám umožňuje snazší prezentování informací.
Každým okamžikem se na webu objevuje víc a víc informací. Obrovským problémem takových dat je nemožnost úplného strojového zpracování. Od počátku webu byl kladen hlavní důraz pouze na zpracování informací člověkem. S narůstajícím potenciálem webu se začal klást důraz taky na strojové zpracování, které by mohlo usnadnit sdílení znalostí (informací) a to nejen těch vědeckých.

Problémem současného webu je existence dat, které postrádají explicitní vazby a sémantiku. Přestože je snaha takové data strojově zpracovávat, není úplně možné, abychom je strojově zpracovali tak, jak je vnímá člověk. K tomu je potřeba, aby byla data prezentována (anotována) tak, že ve výsledku budou prostředkem komunikace mezi lidmi a počítači.

V roce 2001 autoři Tim Berners-Lee, James Hendler a Ora Lassila představili ideu sémantického webu v [1]. Jedná se o rozšíření stávajícího webu, ve kterém jsou současná chaotická data na webu upravena tak, že je můžou autonomně používat inteligentní zařízení. Výhodou tohoto přístupu je stejná interpretace dat lidmi i stroji. K popisu takových dat se využívají *ontologie*.

Ontologie jsou hlavně používány k popisu pojmů určité domény. Tato práce navrhuje netradiční použití ontologií a to při zpracování vizuálních rysů webových dokumentů.

V kapitole 2 je představena idea sémantického webu. Její součástí je seznámení s pojmem ontologie a přidružených formátů k prezentaci dat v ontologiích. Kapitola 3 si bere za cíl seznámit čtenáře s vizuálními rysy webových dokumentů a návrhem možné ontologie pro popis takových rysů. Zhodnocení této práce je provedeno v kapitole 4.

2 Sémantický web

Sémantický web je charakterizován jako rozšíření stávajícího webu. V podstatě se jedná se o nový evoluční stupeň stávajícího webu. Jak již bylo zmíněno v úvodu, myšlenka sémantického webu byla poprvé představena v roce 2001 v časopise Scientific American kolektivem autorů vedených Timem Berners-Lee [1]. Autoři v tomto článku upozornili na množství informací vyskytujících se na webových stránkách. S narůstajícím množstvím takových informací je pak získávání relevantních znalostí stále víc a víc problematičtější.

Publikovaná idea sémantického webu pracuje s představou, kdy se ve světě vyskytují inteligentní zařízení, jež jsou schopná vzájemně komunikovat a řešit za člověka nejrůznější úkoly. Řešení takových úloh se pak opírá o informace, které jsou takovým zařízením poskytnuta. V ideálním případě by pak taková zařízení byla například schopná vhodně plánovat kalendář tak, aby v něm nenastaly kolize a co nejlépe vyhovoval požadavkům uživatele. Příkladem může být plán návštěvy lékaře, kdy by takové zařízení spolu se zařízením lékaře naplánovalo nejvhodnější termín návštěvy.
Jak již bylo předesláno, informace na webu je nutné v případě sémantického webu strukturovat tak, aby je mohl kromě člověka taky zpracovávat stroj a dále s nimi pracovat. Uložené informace musí mít pevně definovaný význam a zapsány podle přesně definovaných pravidel. Díky tomu se pak můžeme bavit o sémantickém webu jako o prostředku ke komunikaci mezi počítači a lidmi.

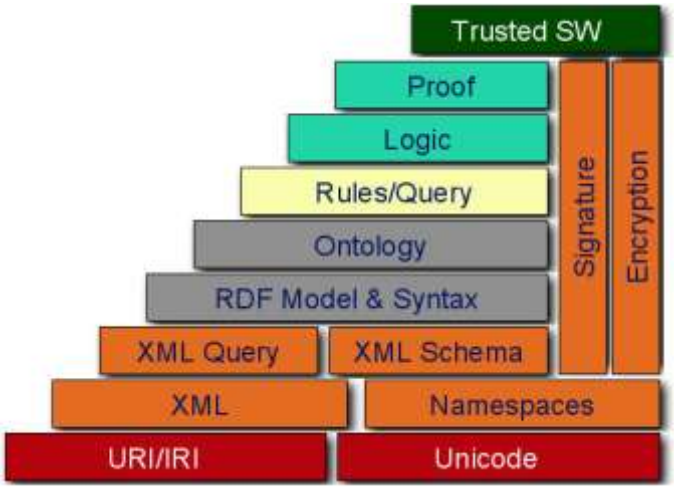
Sémantický web je inteligentní systém pro manipulaci a analýzu znalostní báze. Je schopný pracovat nad velkým množstvím dat. Díky definovaným vlastnostem uložených dat je možné pracovat s daty z různých zdrojů a ty pak dále zpracovávat. V sémantickém webu se odkazujeme na věci reálného světa jako na zdroje. Věc může být cokoliv, o čem chce někdo mluvit. Například „hodnota X je“ a nebo „Praha je …“. Obě tvrzení je možné modelovat v sémantické webu. Základní technologie sémantického webu je RDF[2].

Na specifikaci sémantického webu se podílí World Wide Web Consorcium W3C. Sémantický web je zejména postaven na Resource Description Framework RDF a OWL Ontology Web Language.

Ke sdílení dat na webu je třeba mít definován [3]:

- Datový model (RDF - Resource Description Framework)
- Datový model pro metadat (RDF Schema, OWL – rozšíření RDF Schema)
- Model pro dotazování (SPARQL)

2.1 Vrstvy sémantického webu



Obrázek 1: Vrstvy architektury sémantického webu

Při návrhu sémantického webu byl kladen důraz na využití stávajících technologií. Na obrázku 1 můžeme vidět vrstvy sémantického webu s využitím stávajících technologií. Při práci se sémantickým webem se můžeme setkat například s URI nebo XML, jež jsou v současné době hojně využívány v různých aplikacích. Ve vrstvovém modelu každá vrstva vychází z vrstvy bezprostředně pod ní a rozšiřuje její schopnosti.

Úplně nejnižší vrstva je složena z URI[1]/IRI[2], které umožňují jednoznačnou identifikace v prostředí webu. Unicode je kódování, které obsahuje všechny znaky dostupných abeced. Díky tomuto kódování je možné, aby v jednom textu byly uloženy znaky z různých abeced. Z toho vyplývá, že toto kódování usnadňuje přenositelnost dat, což má v případě sémantického webu obrovský význam [4].

Druhá a třetí vrstva je tvořena XML (Extensible Markup Language) [5] a jeho rozšířeními. XML je univerzální značkovací jazyk. Tento jazyk klade důraz na strukturu dat a neřeší tedy vizuální vzhled. Díky obecnosti tohoto značkovacího jazyka je možné vytvářet vlastní značkovacích jazyky, které budou pracovat pouze s definovanou abecedou. V dnešní době se XML hojně využívají například k výměně dat mezi heterogenními systémy. K definování struktury a omezení značek je možné použít DTD[6] nebo XML Schema [7].

Nad vrstvou XML a jejím rozšířením najdeme vrstvu RDF (Resource Description Framework) modelu a jeho syntaxe. RDF je umožňuje modelování informací ve tvaru podmět-vlastnost-předmět. Více informací o RDF je popsáno v kapitole 2.3.

[1] URI (Uniform Resource Identifier) – jedná se o řetězec znaků s definovanou strukturou k přesné identifikaci zdroje informací za účelem použití v Internetu

[2] IRI (Internationalized Resource Identifier) – jedná se o obecnější formu URI, která umožňuje použití různých znaků (není zde omezení jako v URI, která pracuje s ACSII znaky)

Ontologie nejčastěji popisují nějaké oblasti (domény). V ideálním případě by ontologie mohla popisovat celý svět, ale jelikož to není moc reálné, zavádějí se již zmíněné domény, jež se zaměří pouze na určitý obor, jako například ontologie vína. Detailnější popis ontologie je proveden v kapitole 2.2.

Nad ontologiemi jsou definovány specifické pravidla (Rules) a se nad uloženými znalostmi můžou volat dotazy (Queries), jež pak jsou základem sofistikovanějších aplikací.

Vrstva logiky (Logic) pracuje nad ontologiemi a jejími sémantickými daty tak, že umožňuje automatické odvozování informací.

Dokazování (Proof) kontroluje odvozené výrazy, zda jsou pravdivé. Vzhledem k tomu, že detaily dokazování nejsou předmětem této práce, nebudou dále hlouběji rozebírány.

Na obrázku 1 můžeme vidět, že aplikace podpis (Signature) a šifrování (Encryption) je možné použít ve více vrstvách. Podpis a šifrování mají smysl v případě zajištění důvěryhodnosti informací. Poslední vrstva sémantického webu je důvěryhodnost (Trusted SW). Ta se využívá samotnými aplikacemi.

2.2 Ontologie

Ontologie slouží k popisu pojmů vybrané domény (oblasti) lidského zájmu. V té jsou definovány třídy, které jsou propojeny vzájemnými vztahy (relacemi). Jedná se o jakýsi slovník, ve kterém jsou jednotlivé termíny jednoznačně definovány. Snahou je definovat společné a jednotné chápání pojmů. Díky ontologiím je přenos znalostí specifické oblasti mnohem jednodušší a hlavně jednoznačný.

Cílem ontologie je:

- Podpora porozumění mezi lidmi (i z různých oborů)
- Podpora komunikace mezi agenty (počítačovými systémy)
- Usnadnění návrhu znalostně orientovaných aplikací

Teoreticky je možné, abychom celý svět nadefinovali pomocí ontologií, resp. pojmů a jejich vzájemných vztahů. Avšak s ohledem na množství pojmů je takové modelování světa nereálné. Proto se vyváření ontologií vždy děje jenom pro specifickou doménovou oblast, kdy se jednotlivec nebo skupina snaží namodelovat určitou oblast (doménu). Při definování nové ontologie bychom měli vždy využívat pojmů, které byly již dříve definovány. Jednotlivé doménové ontologie se pak propojí v jeden celek, který v ideálním případě popíše celý svět.

Můžeme se setkat s následujícími typy ontologií:

- Terminologické (lexikální) – termíny dané oblasti a její vztahy (taxonomie)

- Informační – databázové systémy (pokročilejší schémata)
- Znalostní – aplikace umělé inteligence (formální definice pomocí logických formulí)
- Generické – zákonitosti a vztahy mezi obecnými pojmy
- Doménové – pro konkrétní oblast (astrofyzika, lékařství, atd.)
- Aplikační – pro konkrétní aplikaci

Mezi prvky ontologie řadíme[8]:

- Třídy (koncepty)
  - Jedná se o množiny konkrétních objektů, existuje zde dědičnost tříd (vícenásobná dědičnost).
- Individua (objekty, instance)
  - Konkrétní objekty reálného světa. Individuum nemusí být nutně instancí třídy.
- Vlastnosti (relace, atributy, sloty)
  - Pojetí vlastností je jiné než u OO modelování. Vlastnost je relace – samostatně definovaný prvek. Existuje zde dědičnost relací (např. maOtce). Nadřazená relace obsahuje všechny prvky podřazené relace.
- Meta-sloty (facety)
  - Jsou to vlastnosti vlastností. Existují globální (definiční obor a obor hodnot) a lokální (řeší kardinalitu) omezení.
- Primitivní datové typy
  - Argumentem relace může být primitivní hodnota (nemusí být objekt) – číslo, řetězec, výčtová hodnota.
- Axiomy (pravidla)
  - Logické formule, které vymezují vztahy tříd. Obvykle jsou součástí definice tříd.

Vytváření ontologií se může provádět následujícími způsoby:

- Shora dolů - velmi obecné doménové ontologie, přímo navázané na základní ontologie
- Zdola nahoru - např. ontologie orientované na používanou terminologii
- Ze středu „ven“ - od nejfrekventovanějších pojmů - asi nejefektivnější způsob

V souvislosti s ontologiemi se obvykle používá pojem *taxonomie* (nadřazenost pojmů), paranomie (celek-část), *struktury závislostí* apod. Při vytváření nové ontologie je vhodné, aby nejobecnější pojmy navázaly na nějakou existující ontologii resp. obsahový zdroj.

V současné době existuje spousta doménových ontologií. Můžeme se setkat s ontologiemi pro vinařství, zemědělství, geografii nebo například pro modelování vztahů mezi lidmi[3].

Seznamy již definovaných ontologií můžeme nalézt na několika webech[4]. Informace o ontologiích byly hlavně čerpány z [3, 8].

2.3 RDF

RDF (Resource Description Framework) tvoří technologický základ sémantického webu, který byl vypracovaný organizací World Wide Web Consorcium (W3C).

Model RDF je základní rámec pro reprezentaci, výměnu a znovupoužití dat a to nejen těch, která jsou přímo dostupná na webu. RDF provádí propojení webových zdrojů na základě významu dokumentů a to prostřednictvím speciálních informací o těchto datových zdrojích. Takové informace nazýváme metadata - slouží k popisu dat (strukturovaná data o datech).

Jednoduše řečeno, RDF tvoří základ pro zpracování metainformací, jež je možné bez problému zpracovávat strojově. Umožňuje konceptuální modelování znalostí bez ohledu na formát syntaxe.

RDF je možné s výhodou využít k vytváření znalostních databází, ve kterých jsou mezi daty vzájemné vztahy. Nad takovými daty je pak možné provádět dotazy a ty můžou být například používány softwarovými agenty, kteří získaná data můžou dále zpracovat (ohodnotit, upravit, atd.) a na základě nich se pak chovat.

Zdrojem v RDF může být entita, která může být popsána RDF výrazem. Kromě toho, zdrojem může být webová stránka (nebo její část), soubor webových stránek, element XML, objekt dostupný prostřednictvím webu (obrázek, kniha, atd.). RDF neumožňuje deklaraci konceptů (tříd, vlastností, vztahů). K tomuto účelu slouží RDF Schema (RDFS), o kterém je více psáno v kapitole 2.4.1. Informace pro tuto část textu pochází z [9, 10].

2.3.1 RDF model

Základem RDF je model reprezentující vlastnosti zdrojů a jejich hodnoty. Vlastnostmi můžou být atributy zdrojů, které odpovídají tradiční dvojici atribut-hodnota, nebo vztahy mezi zdroji. Tento případ pak připomíná diagramu vztahů

Základním prvkem tohoto modelu je RDF trojice, kde prvky můžou být *zdroje* identifikované pomocí URI. Pokud prvky nejsou zdroje, ale obsahují datové hodnoty, mluvíme o *literálech* [10].

RDF datový model umožňuje reprezentaci RDF výrazů. Dva RDF výrazy jsou stejné za předpokladu, že jsou stejné taky jejich datové modely.

[3] http://www.foaf-project.org/
[4] http://ontologydesignpatterns.org/

Základní datový model obsahuje tři typy objektů:

Zdroje (Resources): všechny prvky, které jsou popisovány RDF výrazy, se nazývají zdroje. Zdrojem může být webová stránka (webový dokument), specifická část stránky nebo kolekce stránek. Zdrojem může být taky objekt, který není dostupný přímo prostřednictvím webu (např. tištěná kniha). Zdroje jsou vždy identifikovány pomocí URI plus volitelných identifikátorů.

Vlastnosti (Properties): je specifický aspekt, charakteristika, atribut nebo vztah, který se používá k popisu zdrojů. Každá vlastnost má svůj specifický význam. Definuje své povolené hodnoty, typy zdrojů a taky může popsat vztahy s dalšími hodnotami.

Tvrzení (Statement): specifický zdroj spolu s konkrétní vlastností vytváří RDF tvrzení. Tyto tři individuální části se nazývají podmět, predikát a objekt. Objekt nějakého tvrzení může být jiný zdroj (identifikovaný URI) nebo to může být přímo hodnota (literál).

RDF tvrzení je možné taky zobrazit pomocí grafické notace (orientovaný graf - obrázek 2), kde uzly reprezentují zdroje a šipka reprezentuje predikát (vlastnost). Směr šipky je důležitý. Šipka vždy vychází ze subjektu směrem k objektu.

RDF trojice je ve tvaru:

<Subject> <Predikát> <Objekt>

Říkáme, že subject má vlastnost určenou objektem.

Příklad 1 (převzato z [11]):
Mějme větu:

Ora Lassila je autor zdroje http://www.w3.org/Home/Lassila

Věta má tedy následující části:

Subjekt (Zdroj)	http://www.w3.org/Home/Lassila
Predikát (Vlastnost)	Autor
Objekt (hodnota)	Ora Lassila

Tabulka 1: příklad RDF trojice

Obrázek 2: Grafická notace RDF trojice

RDF trojice se nejčastěji zapisují v XML syntaxi. RDF/XML umožňuje přiřazení určitých vlastností konkrétnímu webovému zdroji nebo vztahů mezi takovými zdroji. Jak již bylo zmíněno dříve, webovým zdrojem rozumíme objekt, kterému je přiřazen jednoznačný identifikátor URI. Ten je pak dostupný prostřednictvím služby WWW [12]. RDF/XML tvoří metajazyk, který umožňuje popis dalších jazyků. Chováním je tedy podobný univerzálnímu značkovacímu jazyku XML.

2.3.2 SPARQL

SPARQL (Simple Protocol and RDF query language) Jedná se o dotazovací jazyk nad RDF daty. Syntaxe dotazování je podobná SQL. První verze tohoto jazyka byla publikována v roce 2008.

Příklad použití SPARQL nad RDF daty (převzato z [13]):

Data:

```
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
_:a  foaf:name   "Alice" .
_:b  foaf:name   "Bob" .
```

Dotaz:

```
PREFIX foaf:   <http://xmlns.com/foaf/0.1/>
SELECT ?x ?name
WHERE  { ?x foaf:name ?name }
```

Odpověď:

x	Name
_:c	“Alice“
_:d	“Bob“

2.4 Jazyky pro zápis ontologií

2.4.1 RDF Schema

RDF Schema (označováno taky jako RDF's vocabulary description language) je sémantickým rozšířením RDF. Poskytuje mechanismus na popis skupin podobných zdrojů a jejich vzájemných vztahů. Definuje třídy, binární relace (definiční obor a obor hodnot) a hierarchie nad třídami a relacemi. Díky tomu RDF Schema umožňuje definovat ontologie.

Jedná se o slovník popisující vlastnosti a třídy RDF zdrojů se sémantikou pro zobecnění hierarchií takový vlastností a tříd. Informace pro tuto část byly čerpány z [14].

V RDFS můžeme použít:

rdfs:Class – třída

rdfs:subClassOf – podtřída třídy

rdf:Property – vlastnost

rdfs:range – rozsah hodnot

rdfs:domain – definovaná doména

atd.

Příklad RDFS kódu (převzato z [15]):

```
<rdfs:Class rdf:about="Osoba" rdfs:label="Osoba">
    <rdfs:subClassOf rdf:resource="Zivocich" />
</rdfs:Class>
<rdfs:Property rdf:about="maPritele">
    <rdfs:subPropertyOf rdf:resource="zna"/>
    <rdfs:domain rdf:resource="Osoba"/>
    <rdfs:range rdf:resource="Osoba"/>
</rdfs:Property>
```

2.4.2 OWL

Jedná se o složitější prostředek pro popis ontologií. Jeho vyjadřovací síla je větší než má RDF Schema. OWL byl primárně navržen pro použití v aplikacích, které zpracovávají data namísto jejich zobrazování uživatelům. Dokumenty OWL se nazývají OWL ontologie. Základním elementem je *rdf:RDF*, který zapouzdřuje jiné RDF a taky OWL obsah. Ontologie je reprezentována *owl:Ontology*, jež obsahuje popis ontologie.

S ohledem na výpočetní schopnosti existuje jazyk OWL ve třech variantách:

OWL Full

- Úplná varianta OWL, jež umožňuje používat všechny výrazy a konstrukty jazyka OWL a ty pak kombinovat s výrazy RDF a RDFS. Výhodou této varianty je možnost zpětné sémantické a syntaktické kompatibility s RDF. Tedy každý OWL Full dokument je taky dokumentem jazyka RDF. Složitost jazyka vede k nemožnosti úplné podpory pro odvozování a vysoké složitosti zpracování jazyka.

OWL DL

- Tato varianta je kompromisem mezi výpočetní výkonností a vyjadřovací silou. Nelze zde navzájem aplikovat výrazy jazyka – odpovídá standardu deskripční logiky (DL). Efektivnější zpracování jazyka a dobrá výpočetní podpora jsou vykoupeny možností plné kompatibility s RDF a RDFS. Platí, že platný OWL DL dokument je taky platným RDF dokumentem. Bohužel to neplatí naopak.

OWL Lite

- Jedná se o podmnožinu jazyka OWL DL. Tento jazyk obsahuje vzhledem k OWL Lite další omezení, která snižují vyjadřovací sílu jazyka. Díky tomu, že je pak jazyk zjednodušen, přináší snazší a efektivnější zpracování.

Detailnější informace k OWL je možné nalézt přímo na stránkách standardu [16].

Příklad OWL kódu (převzato z [17]):

```
<owl:Class rdf:ID="2+1">
    <rdfs:subClassOfrdf:resource="Byt" />
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="ma_soucast"/>
            <owl:someValuesFrom rdf:resource="Kuchyň”/>
        </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>
```

2.5 Mikroformáty, Mikrodata a RDFa

Webové stránky jsou v dnešní době obrovským zdrojem dat, které je nutné co nejjednodušeji upravit tak, aby je bylo možné zpracovávat strojově. Aby se nemusely všechny stránky kompletně předělávat do nové syntaxe jenom kvůli sémantice dat, objevily se přístupy, které pracují se stávajícím HTML kódem tak, že se do něj snaží sémantiku vložit. Příkladem můžou být *mikrodata, mikroformáty* a *RDFa*.

Mikroformáty využívají stávajících konvencí pro vkládání metainformací do HTML. Díky tomu taky nevadí prohlížečům a splňují standard HTML. Mikroformáty využívají atributy tříd v CSS. Některé vyhledávače tyto sémanticky pojmenované atributy začínají využívat k indexování. Přehled definovaných mikroformátů je dostupný na http://www.microformats.org. Mezi nejznámější mikroformáty patří *hCard* [5], který slouží k reprezentaci osob, organizací, atd. Jak již bylo zmíněno dříve, nevýhodou mikroformátů je, že zneužívají atribut class, jež byl definován za jiným účelem a tak může docházet ke kolizi identifikátorů. Dalším nevýhodou je, že prohlížeče mikroformáty zatím nepodporují a uživatel je nemůže nijak použít.

Mikrodata definují několik nových atributů v HTML. Snahou je vytvořit čistější vkládání metadat než je tomu v případě mikroformátů. Mikrodata byla představena s HTML5, avšak její začlenění není zcela stabilní.

RDFa (Resource Description Framework – in – attributes) je rozšíření XHTML o několik atributů, které umožní pohodlné vkládání libovolného RDF přímo do XHTML kódu. Výhodou tohoto přístupu je, že může využívat stávajících ontologií a nemusí vymýšlet novou syntaxi jako u mikroformátů. Díky tomu taky nemůže dojít ke kolizi identifikátorů. RDFa je možné bez problémů převést na RDF.

Uvedené informace jsou založeny na zdroji [18], kde je taky možné nalézt i konkrétní příklady použití.

[5] http://microformats.org/wiki/hcard

3 Ontologie pro popis webové stránky

V současné době slouží web jako hlavní médium ke sdílení informací. Kromě textových informací (obsah) je čtenář ovlivněn taky způsobem vizuální reprezentace takových informací. Jedná se o tzv. vizuální rysy. Tady může patřit například struktura dokumentu, barevná paleta dokumentu a taky vizuální vlastnosti jednotlivých částí (bloků) dokumentu. V kapitole 3.1 jsou tyto vizuální rysy blíže představeny. Vzhledem k tomu, že popis obsahu (samotných dat) dokumentu nemusí být vždy dostačující, má smysl, abychom se zabývali otázkou ukládání vizuální informace pomocí ontologie. Nejdříve je však nutné, abychom si stanovili atributy (konkrétní vizuální rysy), které budou dostačující k jednoznačnému popisu vizuální informace.

3.1 Vizuální rysy webové stránky

Získávání vizuálních rysů z webových dokumentů může být provedeno s různou rozlišovací úrovní detailů. V této souvislosti mluvíme o detailním získání vizuálních rysů (detailed visual features) nebo o globálním získání vizuálních rysů (overall visual features). Detailní získávání vizuální rysů má smysl v případě, kdy chceme provést detailní extrakci informací z dokumentu. Jedná se například o získávání barev textu, pozadí, vlastnosti orámování, atd.

V případě globálního získávání vizuálních rysů není kladen důraz na přesné hodnoty atributů. Mezi globální rysy můžeme zařadit například informaci o rozmístění jednotlivých bloků v dokumentu (document layout). Cílem je tedy získat informace o pozicích jednotlivých bloků, přičemž nás nezajímají žádné konkrétní rozměry nebo přesné pozice v rámci dokumentu. Globální vizuální rysy můžeme taky charakterizovat jako rysy, jež na uživatele vytváří první dojem.

Vizuální rysy webových stránek můžeme například rozdělit:

- Vizuální organizace dokumentu
- Barevná paleta
- Detailní rysy blokových elementů

3.1.1 Vizuální organizace dokumentu

Vizuální organizace dokumentu je jeden z vizuálních rysů, který člověka upoutá hned při prvním pohledu na dokument. V současné době existuje několik obecně známých rozmístění (layouts), která autoři webových stránek často používají. Proto jsou nejenom v odborné společnosti užívány pojmy jako jedno-sloupcový layout, dvou-sloupcový layout, atd.

Můžeme se taky setkat s tím, že na základě typu layoutu se dokumenty můžou kategorizovat. Například jedno-sloupcový layout je charakteristický pro knižní publikace.

Díky typu layoutu může být jeho čtenář schopen odhadnout, jak má s dokumentem pracovat. V případě jedno-sloupcového layoutu má čtenář všechna data pouze na jednom místě a to v šířce celého dokumentu. Odkazy k navigaci po webu, reklamní bannery a další ne zcela důležité prvky jsou pak umístěny buď v hlavičce anebo v patičce stránky.

Dvou sloupcový layout obvykle obsahuje jeden sloupec se specifickým obsahem a druhý sloupec s navigací, reklamními bloky, atd. V případě tří-sloupcového layoutu to může být podobné. Tam se taky vyskytuje jeden sloupec s obsahem a zbývající sloupce obsahují reklamní bannery nebo navigaci. U více sloupcových layoutů je snaha, aby byla navigace co nejblíže obsahu. Na druhou stranu to však v některých případech může přinášet neschopnost čtenáře k tomu, aby se na text plně soustředil.

Pokud má stránka více sloupců, jejich vzájemné rozměry mají taktéž vliv na vnímání dokumentu. Sloupce s větší šířkou obvykle označují klíčový obsah dokumentu. Datům v takových sloupcích jsou pak obvykle přiřazeny vyšší stupně důležitosti v případě nějakého zpracování. Sloupce, které obsahují navigaci nebo reklamními bannery jsou obvykle výrazně užší než sloupce s hlavním obsahem. Na data v takových sloupcích není obvykle kladen žádný důraz. Při strojovém zpracování jsou pak zahazována a zpracovávají se pouze data, která jsou detekována jako hlavní obsah dokumentu.

Kromě sloupců musí být brán zřetel taky na menší jednotky sloupců, které nazýváme *bloky*. Hranice bloků musí být vždy jednoznačně definovány. V mnoha případech jsou hranice bloků tvořeny jednoznačnou mezerou kolem bloku, rozdílnou barvou anebo například nějakým rámem. Jak již bylo zmíněno dříve, struktura dokumentu (vizuální organizace) je jedna z nejdůležitějších vizuálních rysů. Existuje několik způsobů jak ji ukládat. Informaci o rozložení můžeme ukládat například v mřížce. Návrh řešení byl představen Burgetem v [19]. Vzájemné závislosti je pak možné reprezentovat stromovou strukturou, která se však v některých případech ukázala jako omezující (vícenásobná závislosti). Pro tyto případy je pak nejlepší použití struktury obecného grafu, který využívají ontologie. Více informací o tomto použití bude prezentováno v kapitole 3.2.

3.1.2 Barevná paleta

Barevná paleta (schéma) je další vizuální rys, který má vliv na návštěvníka webového dokumentu. Každý web je specifický svou barevnou paletou. V mnoha případech mají barvy palety souvislost s oborem činnosti nebo zaměřením autora webu (jedná se však o nepsané pravidlo).
Pokud autor webu chce, aby na něm barvy dobře vypadaly, musí je zvolit tak, aby tzv. fungovaly. Za tímto účelem vznikl taky například online projekt na www.colorschemedesigner.com, který pomáhá takové barvy hledat.

Barevná paleta webového dokumentu definuje seznam barev a jejich množství s ohledem na velikost dokumentu. Jedná se o množství barvy přepočítané na zobrazenou část dokumentu - základ tvoří sto procent.

Barevné palety různých stránek můžeme pak porovnávat. Toto porovnávání má smysl pouze v případě, že se takové porovnávání použije jako doplněk nějaké sofistikovanější porovnávací metody.

Proces získání množství barev do barevné palety není úplně triviální, jak by se mohlo zdát. Webové dokumenty jsou tvořeny elementy, jež umožňují vzájemné překrývání. Vzhledem k tomu, že každý element může mít definovánu vlastní barvu na pozadí, je nutné, aby se podle toho taky množství viditelných barev spočítalo. Množství barvy nadřazeného elementu je tedy menší o množství barvy jeho překrývajících elementů. Příkladem může být obrázek 3, který obsahuje jeden blok se dvěma podřazenými (překrývajícími) bloky, jež v případě obrázku tvoří čtyřicet procent celkové plochy.

Obrázek 3: Blok obsahující dva podřazené bloky

Dalším problémem při získávání barev je barva textu a jeho vliv na barevnou paletu. V tomto případě má smysl, aby se množství barvy textu získávalo na základě nějakých heuristik, protože přesný výpočet by byl zdlouhavý a nepřinesl by větší přesnost.

3.1.3 Detailní rysy blokových elementů

Detailní rysy blokových elementů jsou úzce svázány se strukturovanými rysy a taky s barevnou paletou webového dokumentu. *Blokem* označujeme základní element. Bloky můžou být vzájemně překrývány (ve zdrojovém kódu vytváří strukturu). Blok je definován svou pozicí v dokumentu (je možné určit jeho rodiče a taky sousední bloky), rozměry, barvou, obsahem, písmem, atd. Pokud se v jednom bloku vyskytuje text s různými vlastnostmi, je nutné jej rozdělit na více částí. Obsahem blokových elementů můžou být taky obrázky nebo další objekty jako Flash, atd.
Bloky je možné detekovat pomocí vizuálních oddělovačů. Oddělovačem může být například mezera kolem bloku, rozdílné pozadí vzhledem k okolí anebo rámeček.

Bloky webových dokumentů jsou ve zdrojovém kódu reprezentovány stromovou strukturou. To znamená, že každý blok má svého rodiče a navíc může obsahovat dětské bloky.
Celý dokument obsahuje právě jeden kořenový element. Ten pak obsahuje dětské elementy, a tak se postupně vytváří stromová struktura pomocí zanořování elementů.

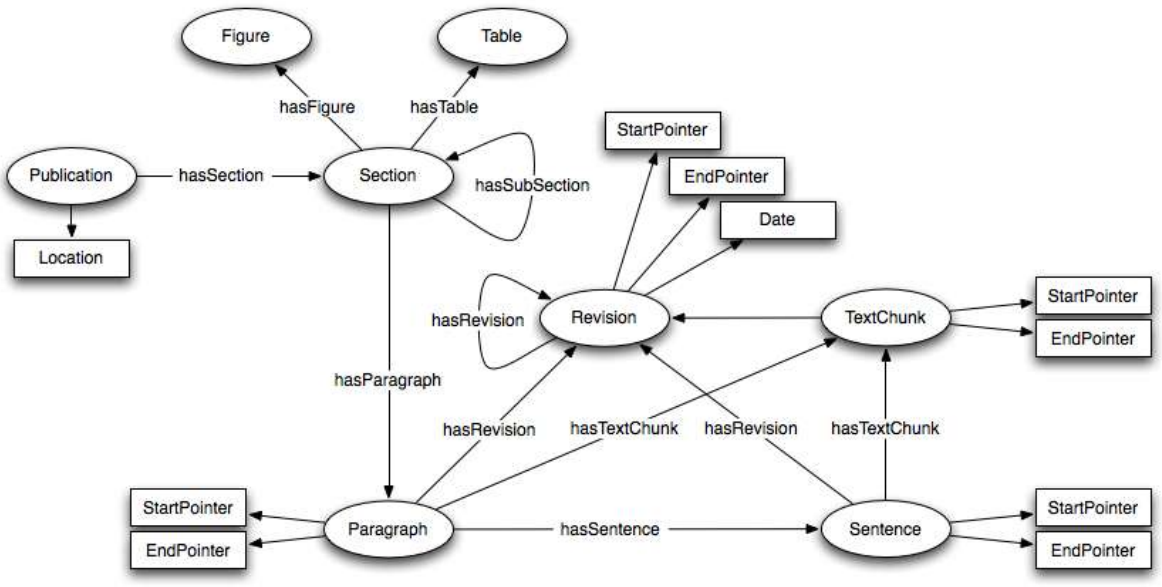
3.2 Návrh možné ontologie

Vzhledem k tomu, že v současné době neexistuje žádná ontologie, která by umožnila popsat vizuální vzhled dokumentu, v následujícím textu provedeme navržení možného řešení. Jak již bylo dříve zmíněno, při definici nové ontologie bychom měli vycházet z již definovaných ontologií, aby se nestalo to, že se v systému ontologií objeví duplicitní třídy pro stejnou doménu.

Návrh nové ontologie, s ohledem na použití existujících ontologií, rozdělíme na dvě části. První část je zaměřena na použití existujících ontologií a druhá část se zaměřuje na doplnění vlastností, které k uložení vizuálních rysů budou chybět. V Kapitole 3.2.1 je představena ontologie SALT Document Ontology, která se pro náš případ jeví nejvhodněji. Můžeme se taky setkat s ontologií *ALOCoM Content Structure Ontology* [20], která by se případně dala použít, avšak působí robustním dojmem. V kapitole 3.2.2 je provedeno seznámení s ontologií, jež umožní reprezentaci vzájemných hierarchických závislostí mezi prvky dokumentu. V kapitole 3.2.3 jsou definovány vizuální vlastnosti, jež je vhodné doplnit do ontologie, abychom byli schopni vyjádřit dostupné vizuální rysy.

3.2.1 SALT Document Ontology

SALT dokument ontology je ontologie pro popis lineární struktury dokumentů. Jedná se tedy o dokumenty, které mají obvykle jednoduchou strukturu. SALT je zkratkou Semantically Annotated LaTex. Tato ontologie vznikla hlavně za účelem modelování obsahu vědeckých článků. S výhodou je možné ji však použít pro libovolný dokument, který není nijak strukturován.



Obrázek 4: Přehledové schéma SALT Document Ontology (převzato z [21])

Na obrázku 4 můžeme vidět schéma navržené ontologie. Vzhledem k tomu, že se jedná o ontologii zaměřenou na vědecké články, můžeme ve schématu najít například extra informace o revizích textu, které bychom v běžných dokumentech na webu normálně asi nepoužili.

Tato ontologie k modelování obsahu dokumentu má definovány třídy: *kapitola, odstavec, věta, textový rámec, tabulka* a *obrázek*. K modelování běžného textu bychom možná ještě využili třídu popisující seznam, ale tu je případně možné nadefinovat pomocí existujících tříd.
Informace o SALT Document Ontology byly čerpány z [21].

3.2.2 Organizační struktura dokumentu

Jelikož je žádoucí, abychom byli schopni pomocí ontologií namodelovat taky hierarchickou strukturu, je nutné najít takovou ontologii, která nám v tom pomůže. S ohledem na HTML kód a jeho stromovou strukturu, potřebujeme modelovat vlastnost celku a jeho částí.

V [22] Valentina Pressutti vložil návrh, který je použitelný k modelování celku a jeho částí. S využitím tohoto návrhu je pak možné modelovat zanořené HTML elementy.

Navržená ontologie obsahuje tři základní elementy:

Entity (owl:Class)

- Označuje jakoukoliv třídu, kterou chceme při modelování používat.

hasPart (owl:ObjectProperty)

- Umožňuje vyjádřit tranzitivní relaci mezi celkem a jeho částí. Např. „Lidské tělo má část mozek.“ Při použití této vlastnosti musíme dbát na to, aby byla správně definovaná doména použití. Nemělo by se nám stát, že dojde k použití částí celku z jiných domén (např. zvířat a planet)

isPartOf (owl:ObjectProperty)

- Inverzní vlastnost k vlastnosti *hasPart*. Např. „Mozek je částí lidského těla.“ S rozsahem domény je zde stejný problém jako u vlastnosti *hasPart*.

Vstupem pro modelování organizační struktury dokumentu respektive její hierarchie bychom mohli s výhodou použít výstup renderovacího enginu *CSSBox*[6] od Burgeta [23] a to hlavně z toho důvodu, že je schopný existující HTML hierarchii výrazně zjednodušit.

3.2.3 Doplnění vizuálních atributů

Díky ontologiím zmíněným v kapitolách 3.2.1 a 3.2.2 jsme schopni namodelovat obsah webových dokumentů. Abychom však splnili dříve danou představu, kdy chceme modelovat

[6] http:// cssbox.sourceforge.net/

i vizuální rysy dokumentu, musíme tyto rysy specifikovat, protože zatím žádná ontologie s těmito informacemi nepočítá. Díky ukládání vizuálních rysů je možné provádět dotazy na daty tak, že bude možné zohlednit vizuální vnímání, které má člověk. Tím se uložená data můžou stát zajímavější v okamžiku, kdy budeme chtít řešit porovnávání dokumentů s ohledem na jejich vzhled a přesné umístění informací.

Burget v [24] navrhnul kategorie zajímavých vizuálních rysů, jež má smysl ukládat. Byly navrženy kategorie:

- Vizuální rysy písma (tabulka 2)
- Prostorové vizuální rysy (tabulka 3)
- Vizuální rysy textu (tabulka 4)
- Vizuální rysy barev (tabulka 5)

Kromě výše uvedených kategorií je nutné, aby byly některé ontologické třídy doplněny taky informací o šířce (width), výšce (height) nebo barvě pozadí (bgcolor).

font size	Průměrná velikost písma
font weight	Průměrná tloušťka písma v rozsahu 0-1 (0 označuje běžné písmo. 1 tučné.)
font style	Průměrný styl písma v rozsahu 0-1 (0 označuje běžné písmo, 1 kurzívu)

Tabulka 2: Vizuální rysy písma (převzato z [24])

aabove, abelow, aleft, aright	Počet oblastí, které jsou umístěny nad (above), pod (below),nalevo (aleft) a napravo (aright) vzhledem k dané oblasti
relx, rely	Relativní pozice oblasti v rámci celé stránky. (0 značí levý resp. horní okraj, 1 značí pravý resp. dolní okraj)
depth	Hloubka oblasti ve stromu hierarchií

Tabulka 3: Prostorové vizuální rysy (převzato z [24])

nlines	Počet textových řádku
ncols	Počet sloupců (počet podoblastí umístěných na rozdílné horizontální pozici)
tlenght	Celková délka textu
pdigits, plower, puper, pspaces, ppunct	Procento číslic, malých písmen, velkých písmen, bílých mezer a interpunkcí obsažených v textu

Tabulka 4: Vizuální rysy textu (převzato z [24])

Tlum	Průměrná hodnota vyzařování (luminiscence) textu
bglum	Vyzařování (luminiscence) pozadí. Pokud je barva pozadí průhledná, uvažujeme barvu pozadí rodičovského bloku.
contrast	Kontrast barev spočítaný z *tlum* a *bglum*
Cperc	Procento textu se stejnou barvou v dokumentu. Hodnota říká jak moc je tato barva unikátní vzhledem k celému dokumentu.

Tabulka 5: Vizuální rysy barev (převzato z [24])

Pro všechny definované atributy této kapitoly je nutné, aby se staly součástí nové ontologie, kde základ bude tvořen ontologiemi zmíněnými v kapitolách 3.2.1 a 3.2.2.

Vzhledem k tomu, že cílem této práce nebyl konkrétní návrh nové ontologie, nejsou zde prezentovány žádné konkrétní příklady.

4 Závěr

Cílem této práce bylo seznámit čtenáře s oblastí ontologií a jejím případném netradičním použití při zpracovávání vizuálních rysů dokumentů. Kapitola 2 se zabývala úvodem do problematiky sémantického webu a ontologií s prezentací základních jazyků pro zápis a dotazování nad ontologiemi.

V kapitole 3 byla představena myšlenka použití ontologií k popisu vizuálních rysů webových dokumentů. Součástí bylo taky přestavení problematiky získávání a zpracování dostupných vizuálních rysů.

Vzhledem k tomu, že si tato práce nekladla za cíl vytvoření kompletního návrhu nové ontologie, ale pouze prozkoumání možnosti použití s vizuálními rysy, nejsou její součástí žádné konkrétní příklady.

Výsledek této práce poslouží k návrhu zmíněné ontologie, jež by se následně měla stát vhodným základem pro článek na konferenci.

5 Literatura

[1] T. Berners-Lee, J. Hendler a O. Lassila, „Semantic web," *Scientific American,* Květen 2001.

[2] D. Allemang a J. Hendler, Semantic web for the working ontologies - effective modelling in RDFs and OWL, 2008.

[3] R. Burget, „Ontologie a sémantický web (přednáška do předmětu PIS)," [Online]. Dostupné na: není veřejně dostupné. [Navštíveno 8.srpna 2012].

[4] Mark Davis et al., „About the Unicode Standard," [Online]. Dostupné na: http://unicode.org/standard/standard.html. [Navštíveno 31.července 2012].

[5] „Extensible Markup Language (XML)," [Online]. Dostupné na: http://www.w3.org/XML/. [Navštíveno 31.července 2012].

[6] W3Schools, „DTD Tutorial," [Online]. Dostupné na: http://www.w3schools.com/dtd/default.asp. [Navštíveno 31.července 2012].

[7] A. Brown, M. Fuchs, J. Robie a P. Wadler, „XML Schema: Formal Description," 25 září 2001. [Online]. Dostupné na: http://www.w3.org/TR/2001/WD-xmlschema-formal-20010925/. [Navštíveno 31. července 2012].

[8] V. Svátek, „Ontologie a WWW," v *Datakon 2002,* Brno, 2002.

[9] B. Thuraisingham, XML Databases and The Semantic Web, 2002.

[10] „Resource Description Framework (RDF)," [Online]. Dostupné na: http://www.w3.org/RDF. [Navštíveno 2.srpna 2012].

[11] „Resource Description Framework (RDF) Model and Syntax Specification," [Online]. Dostupné na: http://www.w3.org/TR/PR-rdf-syntax/. [Navštíveno 2.srpna 2012].

[12] P. Matulík a T. Pitner, „Sémantický web a jeho technologie," *Zpravodaj ÚVT MU,* sv. XIV, pp. 15-17, 2004.

[13] „SPARQL Query Language for RDF," [Online]. Dostupné na: www.w3.org/TR/rdf-sparql-query/. [Navštíveno 8.srpna 2012].

[14] „RDF Vocabulary Description Language 1.0: RDF Schema," [Online]. Dostupné na: http://www.w3.org/TR/rdf-schema/. [Navštíveno 8.srpna 2012].

[15] M. Bureš, A. Morávek a I. Jelínek, Nová generace webových technologií, Praha: VOX, 2005.

[16] „OWL Web Ontology Language," [Online]. Dostupné na: http://www.w3.org/TR/owl-features/. [Navštíveno 8.srpna 2012].

[17] V. Svátek, „Sémantický web - úvodní seznámení," duben 2007. [Online]. Dostupné na:

http://nb.vse.cz/~svatek/rzzw/seweb-prehled.pdf. [Navštíveno 8.srpna 2012].

[18] J. Kosek, „Sémantika ve webových stránkách," [Online]. Dostupné na: http://www.kosek.cz/vyuka/4iz228/prednasky/semantika.pdf. [Navštíveno 2.srpna 2012].

[19] R. Burget, „Vizuálně orientované modelování dokumentů na WWW," v *Datakon*, 2006.

[20] J. Jovanovic, D. Gasevic, K. Verbert a E. Duval, „ALOCoM Content Structure Ontology," prosinec 2010. [Online]. Dostupné na: http://jelenajovanovic.net/ontologies/loco/alocom-content-structure/spec/. [Navštíveno 8.srpna 2012].

[21] T. Groza a S. Handschuh, „SALT Document Ontology," [Online]. Dostupné na: http://salt.semanticauthoring.org/ontologies/sdo. [Navštíveno 5.srpna 2012].

[22] V. Presutti, „PartOf," [Online]. Dostupné na: http://ontologydesignpatterns.org/wiki/Submissions:PartOf. [Navštíveno 8.srpna 2012].

[23] R. Burget, „CSSBox - Java HTML rendering engine," [Online]. Dostupné na: http://cssbox.sourceforge.net/. [Navštíveno 8.srpna 2012].

[24] R. Burget, „Visual area classification for article identification in web documents," v *21st International Workshop on Databases and Expert Systems Applications*, 2010.