

Sémantika ve webových stránkách

4IZ228 – tvorba webových stránek a aplikací

Jirka Kosek

Poslední modifikace: \$Date: 2011/09/29 10:47:25 \$

Copyright © 2010–2011 Jiří Kosek

Obsah

Proč potřebujeme sémantiku na webu	3
HTML a sémantika	4
HTML a sémantika	5
Nedostatky HTML	6
XML a sémantika	7
XML a sémantika	8
Problémy XML na webu	9
Využití XML ve specializovaných vyhledávačích	10
Sémantický web	12
Definice sémantického webu	13
Idea sémantického webu	14
Příklad výroku a reprezentace v RDF	15
Perspektiva sémantického webu	16
Problémy sémantického webu	17
Mikroformáty	18
Mikroformáty	19
Základní fakta	20
Nejpoužívanější mikroformáty	21
Problémy mikroformátů	22
RDFa	23
RDFa	24
Speciální atributy RDFa	25
CURIE	27
Problémy RDFa	28
Mikrodata	29
Mikrodata	30
Perspektiva mikrodat	31
GRDDL	32
GRDDL	33
Závěr	34
Shrnutí	35
Další zdroje informací	36
Dotazy	37

Proč potřebujeme sémantiku na webu

- množství informací na webu je obrovské
- současné vyhledávače nejsou zdaleka ideální
- „pokročilejší“ vyhledávání a automatické propojování informací je zatím v plenkách
- současné počítače nedokáží interpretovat text v přirozeném jazyce
- explicitní vyjádření sémantiky ve strojově zpracovatelné podobě jim může pomoci

HTML a sémantika

HTML a sémantika	5
Nedostatky HTML	6

HTML a sémantika

- HTML samo o sobě příliš sémantiku postihnout nedokáže
- informace o tom, co je seznam, co adresa a odkaz nelze příliš smysluplně využít
- do HTML lze vkládat základní metainformace jako autor, název a popis stránky

```
• <head>
  <title>Elektromix, a.s. </title>
  <meta name="description"
        content="Elektromix je firma zabývající se prodejem
                domácích spotřebičů na splátky">
  <meta name="keywords"
        content="Elektromix, prodej, elektrické
                spotřebiče, leasing">
</head>
```

Nedostatky HTML

- neexistuje rozšiřitelný mechanismus pro vkládání vlastní sémantiky
- prostředky pro vkládání základních metadat byly zneužívány a vyhledávači jsou proto používány v omezené míře

XML a sémantika

XML a sémantika	8
Problémy XML na webu	9
Využití XML ve specializovaných vyhledávačích	10

XML a sémantika

- jazyk XML umožňuje vytváření vlastních elementů/atributů a pomocí nich můžeme snadno označit význam informace
- pokud budou všichni pro jeden druh informace používat stejné elementy, půjde vše snadno indexovat a prohledávat

- <ceník>

...

```
<položka kategorie="CD" kód="04400148712">
```

```
  <název>Entropicture</název>
```

```
  <interpret>Dan Bárta</interpret>
```

```
  <cena měna="Kč">140<cena>
```

```
</položka>
```

...

```
</ceník>
```

Problémy XML na webu

- původní myšlenka, kdy mělo XML nahradit na webu HTML byla příliš revoluční
 - předběhla schopnosti autorů i prohlížečů
- schůdnější myšlenka kombinování XHTML a „sémantického“ XML v jednom dokumentu se také neprosadila
 - specifikace jazyka XHTML byla napsána tak nešťastně, že to formálně neumožňovala
 - nejrozšířenější prohlížeč nepodporoval XHTML

Využití XML ve specializovaných vyhledávačích

- některé vyhledávače si definují vlastní formát, ve kterém jim jde dodávat data k indexování
- tato data jsou poskytována paralelně k normálnímu webovému obsahu
- využívá např. Zbozi.cz¹ a Google Merchant Center (dříve Google Base)²

```
<feed xmlns="http://www.w3.org/2005/Atom"
  xmlns:sc="http://schemas.google.com/structuredcontent/2009"
  xmlns:gd="http://schemas.google.com/g/2005"
  xmlns:scp="http://schemas.google.com/structuredcontent/2009/products"
  xmlns:app="http://www.w3.org/2007/app">
  <entry>
    <app:control>
      <sc:required_destination dest="CommerceSearch"/>
      <sc:excluded_destination dest="ProductSearch"/>
    </app:control>
    <title>Android Shirt</title>
    <content type="text">Catch some air with this cool Android Cartwheel ▶
    Shirt. Since it's made of
      100% organic cotton and combed for extra softness. </content>
    <sc:id>1022316</sc:id>
    <link rel="alternate" type="text/html"
      ▶
      href="http://www.googlestore.com/Kid+s/Youth+Organic+Cotton+Android+T-Shirt.axd"/>
      ▶
    <sc:image_link>http://www.googlestore.com/content/images/standard/10%2081113%20blacka.jpg</sc:image_link>
    <sc:target_country>US</sc:target_country>
    <sc:content_language>en</sc:content_language>
    <sc:attribute name="myattribute" type="text" unit=""> Some Custom ▶
    Attribute </sc:attribute>
    <sc:attribute name="yourCustomAttribute" type="text" ▶
    access="private"> Another custom attribute
    for GCS </sc:attribute>
    <scp:brand>Acme</scp:brand>
    <scp:condition>new</scp:condition>
    <scp:gtin>AB23</scp:gtin>
    <scp:price unit="usd">15.20</scp:price>
    <scp:product_type>Clothing & Accessories &gt; Clothing &gt; ▶
    Outerwear &gt;
    Sweaters</scp:product_type>
    <scp:color>red</scp:color>
    <scp:color>blue</scp:color>
```

¹ <http://napoveda.seznam.cz/cz/specifikace-xml.html>

² <http://base.google.com/support/bin/answer.py?answer=58087>

Využití XML ve specializovaných vyhledávačích (Pokračování)

```
<scp:quantity>30</scp:quantity>  
</entry>  
...  
</feed>
```

Sémantický web

Definice sémantického webu	13
Idea sémantického webu	14
Příklad výroku a reprezentace v RDF	15
Perspektiva sémantického webu	16
Problémy sémantického webu	17

Definice sémantického webu

The Semantic Web is the representation of data on the World Wide Web. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming.

The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

—Tim Berners-Lee, James Hendler, Ora Lassila

Idea sémantického webu

- doplnit síť webových stránek sítí výroků
- výroky lze na rozdíl od webových stránek automatizovaně zpracovávat
- výroky se zapisují ve standardizovaném formátu RDF (Resource Description Format)
 - založen na XML
 - výrok = (subjekt, predikát, objekt)
 - jednotlivé části výroku jsou identifikovány URI adresou (případně hodnotou)
- z „webu dokumentů“ se stane „web znalostí“

Příklad výroku a reprezentace v RDF

- výroky:

Ema má maso

Webová stránka www.kosek.cz byla vytvořena 23. února 1999

- reprezentace v RDF:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:vztahy="http://example.org/vztahy/"
  xmlns:exterms="http://www.example.org/terms/">
  <rdf:Description rdf:about="http://example.org/lide/Ema">
    <vztahy:ma rdf:resource="http://example.org/jidlo/maso"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.kosek.cz/">
    <exterms:creation-date>1999-02-23</exterms:creation-date>
  </rdf:Description>
</rdf:RDF>
```

- reprezentace v N3:

```
@prefix exterms: <http://www.example.org/terms/> .
```

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```

```
@prefix vztahy: <http://example.org/vztahy/> .
```

```
<http://example.org/lide/Ema> vztahy:ma
```



```
<http://example.org/jidlo/maso> .
```

```
<http://www.kosek.cz/> exterms:creation-date "1999-02-23" .
```

Perspektiva sémantického webu

- na široké nasazení se zatím čeká
- ruční vytváření RDF anotací je vzhledem k syntaxi a pracnosti nemyslitelné
- RDF se proto dnes používá spíše opět jako alternativní formát pro prezentování informací, u kterých chceme ostatním usnadnit jejich automatické zpracování
- pro agregaci RDF metadat a jejich další zpracování je potřeba sjednotit „slovníky pojmů“ (tzv. ontologie)
- ontologie se zapisují pomocí jazyka OWL

Problémy sémantického webu

- neexistuje aplikace, která by oslovila větší množství uživatelů a způsobila tak větší zájem o technologii
- syntaxe RDF je zbytečně komplikovaná
- identifikace pomocí URI není vždy jednoznačná
- metadata musí být vytvářena odděleně od klasického webového obsahu

Mikroformáty

Mikroformáty	19
Základní fakta	20
Nejpoužívanější mikroformáty	21
Problémy mikroformátů	22

Mikroformáty

- konvence pro vkládání strukturovaných metainformací do HTML, která využívá pouze stávající vlastnosti jazyka
- „viditelná metadata“ – metadata jsou společná s viditelnou částí stránky a uživatel je nezapomene aktualizovat
- ```
<div class="vevent">
 Devátý ročník konference Znalosti 2010
 se bude se konat
 <abbr class="dtstart" title="2010-02-03">3.</abbr>-
 <abbr class="dtend" title="2010-02-06">5. února 2010</abbr>
 na

 fakultě managementu VŠE v Jindřichově Hradci
 (Pozice:
 <abbr class="latitude"
 title="49.14887111111111">49°8'55.936"N</abbr>,
 <abbr class="longitude"
 title="15.005985277777778">15°0'21.547"E</abbr>
).
 Podrobnosti
 o konferenci
</div>
```

# Základní fakta

- mikroformáty nevdí prohlížečům (mnoho prohlížečů nezvládne zpracovat XHTML dokument s vloženými fragmenty RDF nebo XML)
- mikroformáty lze strojově zpracovávat
- vyvinuly se a navazují na důsledné použití tříd v CSS
- postupně je začínají podporovat zejména vyhledávače
  - některé informace na stránkách s výsledky jsou označeny pomocí mikroformátů pro snazší další zpracování
  - mikroformáty ve stránce využívají vyhledávače pro zlepšení vyhledávání a zobrazení výsledků
- prohlížeče zatím nemají přímo integrovanou podporu, ale dají se používat pluginy<sup>3</sup>

<sup>3</sup> <http://microformats.org/wiki/user-interface>

# Nejpoužívanější mikroformáty

- přehled definovaných mikroformátů je dostupný na <http://microformats.org/>
- hCard<sup>4</sup>  
Vizitka – umožňuje reprezentovat osoby, organizace a jejich základní údaje jako jméno a a adresa
- XFN<sup>5</sup>  
Reprezentace vztahů mezi osobami
- hCalendar<sup>6</sup>  
Informace o údalostech jako je jejich místo a čas
- rel-license<sup>7</sup>  
Informace o licenci, pod kterou je vydán obsah na stránce
- rel-tag<sup>8</sup>  
„Tagování“ obsahu

<sup>4</sup> <http://microformats.org/wiki/hcard>

<sup>5</sup> <http://gmpg.org/xfn/>

<sup>6</sup> <http://microformats.org/wiki/hcalendar>

<sup>7</sup> <http://microformats.org/wiki/rel-license>

<sup>8</sup> <http://microformats.org/wiki/rel-tag>

# Problémy mikroformátů

- zatím chybí přímá podpora v prohlížečích, je potřeba používat pluginy nebo Web 2.0 aplikace
- zneužívá se atribut `class`
- není vyřešena kolize identifikátorů
  - částečně řeší profily

```
<head profile="http://microformats.org/profile/hcalendar">
```

```
<link rel="profile" href="http://microformats.org/profile/hcalendar">
```

- není definováno standardní API pro práci s mikroformáty v JavaScriptu

# RDFa

RDFa .....	24
Speciální atributy RDFa .....	25
CURIE .....	27
Problémy RDFa .....	28

# RDFa

- rozšíření XHTML o několik atributů, které umožní pohodlné vkládání libovolného RDF přímo do XHTML kódu
- RDFa = RDF in ... attributes
- lze využívat jakoukoliv ontologii, není potřeba pro každá data vymýšlet novou syntaxi jakou u mikroformátů
- princip je podobný jako u mikroformátů, ale jsou odstraněny nedostatky jako potencionální kolize identifikátorů

```
<html xmlns="http://www.w3.org/1999/xhtml"
 xmlns:cal="http://www.w3.org/2002/12/cal/ical#"
 xmlns:xs="http://www.w3.org/2001/XMLSchema#"
 xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#">
...
<p typeof="cal:Vevent">
 Devátý ročník konference
 Znalosti 2010
 se bude se konat
 <span property="cal:dtstart" content="2010-02-03"
 datatype="xs:date">3.-
 <span property="cal:dtend" content="2010-02-06"
 datatype="xs:date">5. února 2010
 na
 <a property="cal:location"
 href="http://www2.fm.vse.cz/znalosti/misto.html/">fakultě
 managementu VŠE v Jindřichově Hradci
 (Pozice:
 <span property="geo:lat"
 content="49.14887111111111">49°8'55.936"N,
 <span property="geo:long"
 content="15.005985277777778">15°0'21.547"E).
 Podrobnosti
 o konferenci
</p>
...
</html>
```

- je dán algoritmus, jak libovolné RDFa převést na RDF

# Speciální atributy RDFa

- RDFa definuje několik atributů, které lze používat pro obohacení v podstatě jakéhokoliv jazyka o možnost zápisu RDF tripletů
- dnes je definováno použití RDFa společně s XHTML<sup>9</sup>, podmnožinu RDFa bude podporovat ODF 1.2 a nyní se pracuje i na integraci do HTML5<sup>10</sup>

**Tabulka 1. Atributy používané pro zápis RDFa**

Atribut	Typ	Význam
rel	seznam CURIE	Zachycení vztah (predikát) mezi dvěma zdroji.
rev	seznam CURIE	Zachycení reverzního vztah (predikát) mezi dvěma zdroji.
content	řetězec	Strojově čitelný textový obsah hodnoty, pokud je jiný než obsah elementu.
href	URI	URI reprezentující objekt v závěru výroku externí objekty.
src	URI	URI reprezentující objekt v závěru výroku externí objekty vnořené stránky jsou například obrázky.

<sup>9</sup> <http://www.w3.org/TR/rdfa-syntax/>

<sup>10</sup> <http://www.w3.org/TR/rdfa-in-html/>

# Speciální atributy RDFa (Pokračování)

Atribut	Typ	Význam
about	URI nebo bezpečné CURIE	Určení předmětu výroku.
property	seznam CURIE	Určení vlastnosti (predikátu)
resource	URI nebo bezpečné CURIE	URI reprezentující objekt výroku (externí objekty, které nelze přejít pomocí odkazu)
datatype	CURIE	Určení datového typu hodnoty
typeof	seznam CURIE	Určení typu předmětu

# CURIE

- v RDF je vše identifikováno pomocí URI
- v jednom dokumentu se mohou opakovat URI se stejnou počáteční částí a ruční zápis je zbytečně zdlouhavý
- CURIE (compact URI) dovolují zkrátit zápis
- bezpečné CURIE = CURIE ve hranatých závorkách pro odlišení od URI
- zápis bez CURIE

```
<div about="http://dbpedia.org/resource/Albert_Einstein">...</div>
<div about="http://dbpedia.org/resource/Germany">...</div>
```

- různé alternativy s využitím CURIE

```
<html xmlns:db="http://dbpedia.org/">
 ...
 <div about="[db:resource/Albert_Einstein]">...</div>
 <div about="[db:resource/Germany]">...</div>
 ...
</html>
```

```
<html xmlns:dbr="http://dbpedia.org/resource/">
 ...
 <div about="[dbr:Albert_Einstein]">...</div>
 <div about="[dbr:Germany]">...</div>
 ...
</html>
```

# Problémy RDFa

- zatím je specifikována jen integrace s XHTML, na začlenění do HTML se teprve pracuje
- pomalu jej začínají používat největší poskytovatelé obsahu, ale mikroformáty zatím převažují
  - podpora RDFa ve vyhledávači Yahoo<sup>11</sup>
  - podpora RDFa ve vyhledávači Google<sup>12</sup>
- API pro čtení v JavaScriptu se bude teprve vytvářet v rámci nové pracovní skupiny W3C

<sup>11</sup> <http://developer.search.yahoo.com/start>

<sup>12</sup> <http://www.google.com/support/webmasters/bin/answer.py?answer=146898>

# Mikrodata

Mikrodata .....	30
Perspektiva mikrodat .....	31

# Mikrodata

- přidávají do HTML několik nových atributů, aby vkládání metadat bylo „čistší“ než v případě mikroformátů
- pro identifikaci typů objektů je možné používat URI a předejít tak problémům s kolizními identifikátory
- používají vlastní datový model (odlišný od RDF)

```
<div itemscope
 itemtype="http://schema.org/Event">
 Devátý ročník konference
 Znalosti 2010 se bude se konat
 <time itemprop="startDate" datetime="2010-02-03">3.</time>-
 <time itemprop="endDate" datetime="2010-02-06">5. února 2010</time>
 na

 <a itemprop="name"
 href="http://www2.fm.vse.cz/znalosti/misto.html/">fakultě
 managementu VŠE v Jindřichově Hradci
 (<span itemprop="geo" itemscope ►
itemtype="http://schema.org/GeoCoordinates">Pozice:
 49°8'55.936"N, 15°0'21.547"E
 <meta itemprop="latitude" content="49.148871"/>
 <meta itemprop="longitude" content="15.005985"/>
).

 Podrobnosti o konferenci
</div>
```

# Perspektiva mikrodat

- specifikace začlenění mikrodat do HTML5 není zcela stabilní
- je definováno API pro práci s mikrodaty v JavaScriptu
- nezapadá zcela do konceptu sémantického webu, ale podporují je silní hráči
- [schema.org](http://schema.org)<sup>13</sup> – slovníky běžných metadat podporované vyhledávači Google, Microsoft a Yahoo
- samotný vznik mikrodat byla poněkud partyzánská akce editora specifikace HTML5, který nemá rád RDFa (a sémantický web vůbec)

<sup>13</sup> <http://schema.org>

# GRDDL

GRDDL ..... 33

# GRDDL

- GRDDL (čti griddle) = Gleaning Resource Descriptions from Dialects of Languages
- umožňuje popsat, jak se mají metadata zachycená na stránce pomocí mikroformátů (nebo jakýmkoliv jiným způsobem) převést na RDF
- převod do RDF je popsán připojeným programem (nejčastěji transformací v XSLT)

```
<html xmlns="http://www.w3.org/1999/xhtml">
 <head profile="http://www.w3.org/2003/g/data-view">
 <title>Ukázkový dokument</title>

 <link rel="transformation"
 href="http://www.w3.org/2000/06/dc-extract/dc-extract.xsl" />
 <link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
 <meta name="DC.Subject"
 content="metadata, sémantický web, RDFa, mikroformáty" />
 ...
 </head>
 ...
</html>
```

# Závěr

Shrnutí .....	35
Další zdroje informací .....	36
Dotazy .....	37

# Shrnutí

- explicitně vyjádřená sémantika na stránkách může pomoci vyhledávačům a „mash-up“ aplikacím
- příklon k jednoduchosti a „viditelným“ metadatům
- několik soutěžících formátů – v nejbližší době jsou nejperspektivnější asi mikrodata a slovník schema.org
- důležité je sledovat především aktuální podporu ve vyhledávačích a v prohlížečích

# Další zdroje informací

- mikroformáty
  - <http://microformats.org> – hlavní stránka o mikroformátech
  - <https://addons.mozilla.org/cs/firefox/addon/4106> – rozšíření Operator pro práci s mikroformáty včleněnými do stránky
- RDFa
  - <http://www.w3.org/TR/xhtml-rdfa-primer/> – RDFa Primer (úvod do RDFa)
  - <http://www.w3.org/TR/rdfa-syntax/> – RDFa in XHTML: Syntax and Processing
- GRDDL
  - <http://esw.w3.org/topic/CustomRdfDialects> – transformace pro převod mnoha mikroformátů do RDF
- mikrodata
  - <http://dev.w3.org/html5/md/> – návrh specifikace
  - [schema.org](http://schema.org)<sup>14</sup> – slovník metadat podporovaný největšími vyhledávači

<sup>14</sup> <http://schema.org>

# Dotazy

???