Kvalitní dokument jako základ účinného vyhledávání informací

Daniela Tkačíková *
Daniela.Tkacikova@vsb.cz

INFORUM 2004: 10. konference o profesionálních informačních zdrojích
Praha, 25.-27. 5. 2004

Abstrakt: Kvalitně vytvořený dokument – bez ohledu na to, zda jde o dokument publikovaný tradiční cestou nebo o dokument elektronický – je jednoduché zpracovat pro účely vyhledání a využívání uživateli. V prostředí Webu má respektování zásad a standardů pro vytváření dokumentů snad ještě větší význam než v tradičním prostředí tištěných dokumentů, a to bez ohledu na možnosti a výhody automatizovaného fulltextového sběru dat, jež pozitivně ovlivňují vyhledávání informací uživateli. Respektování standardů usnadňuje a zkvalitňuje nejen zpracování a vyhledávání informací, ale také správu dokumentů, jejich aktualizaci, dlouhodobou dostupnost, použitelnost a nezávislost na technickém zařízení a cílovém formátu. Hodnotný obsah strukturovaného dokumentu je jednou z cest k dosažení kvalitních výsledků hledání. Příspěvek popisuje základní prvky XHTML dokumentů a na konkrétních příkladech ukazuje jejich význam pro práci vyhledávacích nástrojů.

1 Úvod

Výhodou a současně i nevýhodou elektronického publikování v prostředí Webu je jeho relativní jednoduchost a finanční nenáročnost. Díky tomu se mohou na komunikaci informací podílet i ti, pro něž by bylo šíření informací a poskytování informačních služeb tradičními postupy a cestami, především tiskem, prakticky nedostupné.

Psaní textů se řídí určitými pravidly a mělo by to platit i při publikování na Webu. Forma i způsoby šíření informací prošly dlouhým vývojem, během něhož se vyvíjela pravidla nejen pro psaní textů, ale také pro formální úpravu různých typů publikací šířených tradiční (tištěnou) formou. Jistě není nutné se dlouze rozepisovat o tom, jaké problémy při zpracování dokumentů a při jejich vyhledávání přináší, když vydavatelé nedodržují určité ustálené zvyklosti pro nakladatelskou úpravu knih a dalších publikací. Dávno již nejde jen o nepsané zásady, ale o mezinárodně zpracovaná a přijatá doporučení (viz například normy ISO vztahující se k problematice prezentace, identifikace a popisu dokumentů [4]). Přestože jejich respektování přináší řadu výhod, stále se například v knihovnické praxi nutností někdy i zbytečně komplikovaného řešení nedostatečných či dokonce chybných údajů musíme zabývat.

Analogie k nedodržování platných webových standardů je zřejmá. Doporučení týkající se vytváření webových dokumentů v mnohém vycházejí z tradičního způsobu publikování. Nerespektování těchto doporučení přináší obdobné problémy, jak při zpracování informací, tak při jejich vyhledávání. Elektronické prostředí má ovšem řadu výhod, které pomáhají ve snadnější orientaci v informacích na webu zveřejněných. Neznamená to však, že by si díky elektronickému prostředí s těmito nedostatky tvůrci vyhledávacích nástrojů jen tak jednoduše poradili. I pro ně to znamená, že je nezbytné s řadou chyb, občas i úmyslných, počítat a nějakým způsobem je ošetřit. Jestliže k tomu připočteme jinak pozitivní vlivy, které elektronické publikování v prostředí Internetu přineslo, například možnosti okamžitých aktualizací a změn

* Vysoká škola báňská-Technická univerzita Ostrava, Ústřední knihovna, 17. listopadu, 708 33 Ostrava-Poruba

v dokumentech, je zjevné, že také v elektronickém prostředí je výhodné akceptovat postupy a pravidla, která komunikaci informací usnadní.

Přitom je nutné si uvědomit, že například **webové standardy jsou jen nástrojem,** že tím, oč by autorům mělo jít především, je **obsah zveřejňovaných informací.** Webové dokumenty, ostatně stejně jako tradiční dokumenty, nejsou jen texty. Jde sice o prostředí hypermediální, nicméně textová složka je významná. I při psaní textů pro web se vyplatí dbát na dodržování určitých zásad, samozřejmě s ohledem na účel dokumentu či služby. Zajímavé informace, stručné a vyvážené, s jasným a jednotným vyjadřováním, uspořádané přehledně a logicky, by měly být vždy základem poskytovaných služeb.

2 Současnost webu

Současný přístup ke tvorbě webových stránek je charakterizován **důrazem na uživatelský prožitek a na přístupnost dokumentů.** Základem uživatelského prožitku jsou **informační architektura,** tj. organizace (uspořádání) informací, struktura a navigace, **a použitelnost** [5, 6]. Přidanou hodnotou takového přístupu je **ekonomický přínos** tvorby webových stránek. Vyjádřením tohoto přístupu jsou doporučení (normy) konsorcia W3C [7], jejichž dodržování je jednou ze záruk **všeobecné dostupnosti informací.** Kromě jiného tyto normy umožňují a usnadňují opětovné používání dat i jejich využívání pro různé účely, snižují náklady na modernizaci systémů a zajišťují nezávislost na konkrétních aplikacích, na určitém softwarovém či hardwarovém řešení. Jsou tak základem kompatibility, flexibility a dlouhodobé dostupnosti dat.

Každý, kdo vytváří webové dokumenty, víceméně vychází z doporučení, jež připravilo World Wide Web Consortium (W3C). Dodržuje tedy na určité úrovni tzv. *webové standardy.* Konsorcium W3C samo o sobě není standardizační institucí v běžném smyslu. Hlavní funkcí W3C je výzkum a vývoj a uveřejňování informací o technologiích a aktivitách týkajících se webu. Webové standardy – specifikace a doporučení vytvořená konsorciem W3C – nejsou normy, jimiž by se *museli* autoři bezpodmínečně řídit, zcela jistě však jde o zásady, které se vyplatí respektovat a jež jsou základním souhrnem důsledných, promyšlených a osvědčených postupů.

V současnosti se web zřejmě nachází v přechodném období mezi určitým překotným vývojem poznamenaným zmatky a nedostatky v technologiích 90. let minulého století a (doufejme) budoucností založenou na zvyklostech respektujících postupy vyjádřené ve webových standardech. Jejich cílem je usnadnění práce s webem pro všechny zúčastněné – počínaje autory a konče uživateli informací a služeb.

3 Potíže s HTML

Využívání HTML v průběhu rozvoje služby WWW přineslo řadu inovací, jež změnily původní jednoduchý jazyk pro popis struktury dokumentů (rozvržení obsahu) a definování vazeb mezi nimi (hypertextových odkazů) na složitý jazyk používaný k budování graficky a typograficky bohatých webových sídel. Jednou z nejproblematičtějších inovací je využívání tabulek pro formátování vzhledu webových dokumentů. Tabulky, zvlášť jsou-li složité a vnořené, přinášejí problémy s velikostí souborů i s rychlostí v zobrazování dokumentů. Je obtížné takto vytvořené dokumenty aktualizovat a udržovat, obzvlášť je-li správa webového sídla týmovou prací. Dokumenty formátované pomocí tabulek jsou navíc často nepřístupné skupinám uživatelů s různými omezeními, jedno zda zdravotními či technickými, nebo je obtížné s nimi

pracovat. **Zpracování takových dokumentů přináší problémy i robotům vyhledávacích služeb.**

Nevhodné používání jazyka HTML a jeho úpravy autory webových dokumentů není pochopitelně jediným problémem doprovázejícím rozvoj webu v uplynulých letech. Svoji roli sehrály i prohlížeče webu, jejichž podíl na používání nesprávných postupů je dodnes velmi významný, přestože právě výrobci prohlížečů patří mezi členy konsorcia W3C. Programové vybavení pro práci s Internetem je součástí výnosného obchodu, a tak je celkem pochopitelná snaha výrobců dosáhnout třeba i standardům odporujícími lákavými novinkami dominantního postavení. Dosavadní vývoj webu tak poznamenali kladně i záporně všichni: autoři dokumentů, výrobci prohlížečů i nástrojů pro budování webových stránek a svým způsobem i jeho uživatelé.

4 Webové standardy a jejich význam

Současný vývojový stav v oblasti webových standardů ovšem dosáhl úrovně, kdy se vyplatí z řady důvodů jejich respektováním využívat obrovského potenciálu, který je jednoznačně výhodný i z pohledu budoucnosti. A to i přesto, že v současnosti stále docela dobře „fungují" webové dokumenty založené na nestandardních postupech.

Existuje totiž řada důvodů, pro něž se vyplatí standardy pochopit a využívat:

- jejich respektování šetří čas a tím i peníze autorům a poskytovatelům webových informací a služeb,
- využívání postupů založených na standardních technologiích umožňuje, usnadňuje a urychluje práci uživatelům webu,
- porozumění standardům vede k chápání širších souvislostí a principů, na nichž je vznik i rozvoj webu vystavěn a mezi něž patří mj. i myšlenka všeobecné dostupnosti informací.

Z technologického pohledu je zřejmé, že respektování standardů vede k úspoře nákladů na budování a údržbu webových dokumentů a umožňuje jejich bezproblémové využívání různými koncovými zařízeními uživatelů. Ze společenského hlediska jsou standardy nástrojem, který odstraňuje bariéry v přístupu uživatelům, ať jde o překážky či omezení zdravotní, finanční či jazykové.

5 Technologie a aktivity W3C

Aktivity konsorcia W3C jsou velmi široké a je možné se o nich více dozvědět na webových stránkách W3C i z dalších zdrojů, například ze sborníků z konferencí pořádaných konsorciem (International World Wide Web Conferences [8]).

Hypertext Markup Language (HTML) [9]. Značkovací jazyk pro vytváření webových dokumentů, jehož konečnou verzí je HTML 4.01. Tato verze obsahuje jen menší změny oproti předchozí verzi HTML 4.0, její význam je však obrovský, protože DTD (Document Type Definition) HTML 4.01 jsou základem XHTML 1.0. *Jde dnes prakticky o uzavřenou kapitolu v činnosti W3C*.

Extensible Markup Language (XML) [10]. Značkovací jazyk pro univerzální formát strukturovaných dokumentů a dat.

Extensible Hypertext Markup Language (XHTML) [9]. Značkovací jazyk pro vytváření webových dokumentů a dokumentů pro alternativní zařízení. Navazuje na předchozí dvě aktivity.

Tabulky kaskádových stylů (Cascading Style Sheets, CSS) [11]. Stylový jazyk pro prezentaci, formátování vzhledu (X)HTML dokumentů.

Synchronized Multimedia Integration Language (SMIL) [12]. Jazyk založený na XML, jehož cílem je usnadnit synchronizaci multimedií (video, zvuk, text).

Scalable Vector Graphics (SVG) [13]. Jazyk založený na XML určený pro popis grafických objektů.

Mezinárodní prostředí (internationalization) [14]. Cílem této aktivity je usnadnit správné kódování a zobrazování dokumentů v mnohajazyčném mezinárodním prostředí webu.

Přístupnost (accessibility) [15]. Jejím cílem je zajištění přístupnosti dokumentů pro všechny uživatele.

Document Object Model (DOM) [16]. Aplikační programové rozhraní, jež definuje obecný standard pro přístup k jakémukoliv platnému HTML dokumentu nebo ke správně vytvořenému XML dokumentu; cílem je zajistit shodné objektové modely dokumentů v prohlížečích.

6 Oddělení struktury dokumentu od formátování jeho vzhledu

Jedním z prvořadých předpisů HTML 4 je oddělení způsobu prezentace dokumentu od jeho struktury. To je jedna z nejdůležitějších věcí, na které je kladen důraz i v dalším vývoji základních doporučení pro tvorbu webových dokumentů. Ve většině dokumentů, jež jsou dnes na webu zveřejněny, jsou však stále nesprávně používány prvky jazyka HTML pro formátování dokumentu. Tyto chyby vedou k problémům s přístupností, rychlostí zobrazování i s nejednotností vzhledu dokumentů v různých prohlížečích i jejich vývojových verzích.

Konečná verze jazyka HTML – specifikace HTML 4.01 – byla zveřejněna v prosinci 1999. Stala se základem pro specifikaci XHTML 1.0, publikovanou brzy poté v lednu 2000, (revidovaná verze pochází ze srpna 2002), jež je vlastně jen přeformulováním konečné verze HTML v XML. Zatímco první specifikace XHTML zachovává postupy zahrnuté ve standardu HTML 4.01 a jen přepisuje HTML jako aplikaci XML, v další vývojové verzi jazyka – v XHTML 1.1 z 31. května 2001, se z něj již stává pouze jazyk pro popis struktury dokumentu. Formátování vzhledu je přenecháno samostatnému stylovému předpisu.

Ve stadiu příprav je verze XHTML 2.0 [17], zveřejněný pracovní návrh doporučení je z 6. května 2003.

7 Modularizace XHTML

Jednou ze základních myšlenek rozvoje XHTML je jeho modularizace. Jednotlivé prvky jazyka jsou seskupeny do modulů odpovídajících jejich určení (funkci) společně s vlastnostmi, které se k nim mohou vztahovat, a s minimálním obsahovým modelem.

7.1 Moduly XHTML

Základními moduly XHTML jsou:

- **strukturální modul** zahrnující prvky, které tvoří základní strukturu XHTML dokumentu (body, head, html, title),
- **textový modul**, jenž definuje základní prvky sloužící k označení textu a obsahu dokumentů (h1 až h6, address, blockquote, div, p, pre, abbr, acronym, br, cite, code, dfn, em, kbd, q, samp, span, strong, var),

- **hypertextový modul** s prvkem a sloužícím pro hypertextové odkazy na jiné zdroje,
- **modul seznamů** obsahující prvky sloužící k vytváření seznamů (dl, dd, dt, ol, ul, li),
- **formulářový modul** s prvky pro tvorbu formulářů (form, input, label, select, textarea aj.).

Kromě toho obsahuje XHTML ještě řadu dalších modulů (modul tabulek, objektový modul, modul linků, metainformační modul aj.), jež pokrývají celou šíři povolených prvků dané verze jazyka. Standard XHTML 1.1 postavený na XHTML 1.0 Strict už neobsahuje ty prvky, jež byly v předchozích verzích sice povoleny, ale byly označeny jako překonané (*deprecated*).

Vezmeme-li v úvahu, co v relativně krátké historii webu představuje doba, jež uplynula od publikování první specifikace XHTML, a k tomu další fakt, že specifikace CSS2 pochází dokonce již z května 1998, je jistě podivné, že se využívání těchto webových standardů autory doposud nestalo běžnou zvyklostí. Zvlášť když jejich respektování přináší tolik výhod, jak autorům samotným, tak uživatelům webu. Svůj podíl na tom jistě mají výrobci prohlížečů, kteří se značným zpožděním začali akceptovat normy, na jejichž vývoji se jako členové konsorcia W3C sami určitým způsobem podíleli. Současné verze všech nejrozšířenějších prohlížečů však již s drobnými odchylkami víceméně podporují platné předpisy, a tak se tou překážkou možná mohou zdát uživatelé webu, kteří kupodivu stále z nějakého důvodu pracují se staršími verzemi prohlížečů... Každý autor webových dokumentů, který při své práci současné webové standardy respektuje, ovšem ví, že jednou z vlastností XHTML dokumentů je jejich zpětná kompatibilita, a že si s jejich obsahem poradí docela dobře i staré prohlížeče.

7.2 Co je to struktura dokumentu

Jde o základní kostru webového dokumentu.

Ta je v XHTML tvořena:

- deklarací verze jazyka XML a kódování (XML Prolog),
- deklarací typu dokumentu (`DOCTYPE`), jež definuje typ dokumentu a verzi DTD (Document Type Definition),
- kořenovým prvkem `<html>` se jmenným prostorem pro XHTML,
- záhlavím `<head>` se značkou `<title>`,
- tělem dokumentu (`<body>`),
- prvky popisujícími strukturu obsahu dokumentu a používanými logickým způsobem k řízení obsahu dokumentu, jako jsou především odstavce, případně zalomení řádků (`<p>`, `<br>`), nadpisy (`<h1>` až `<h6>`) a seznamy (`<ul>`, `<ol>`, `<dl>`) položek (`<li>`).

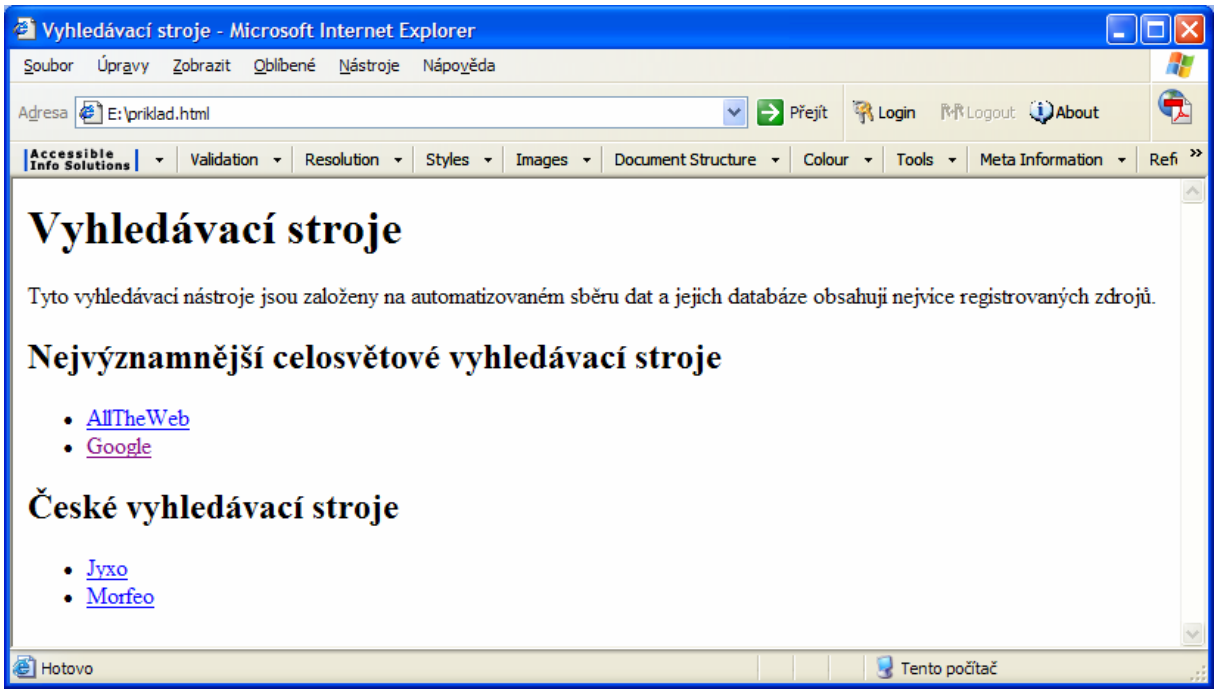
Všechny prvky v XHTML dokumentu musejí být uzavřeny mezi počáteční a koncovou značku (`<h1>Nadpis první úrovně...</h1>`), včetně ošetření prázdných značek (např. `<br />`), a musejí být zapsány malými písmeny.

Jednoduchý dokument XHTML s vyznačenou základní strukturou vypadá takto:

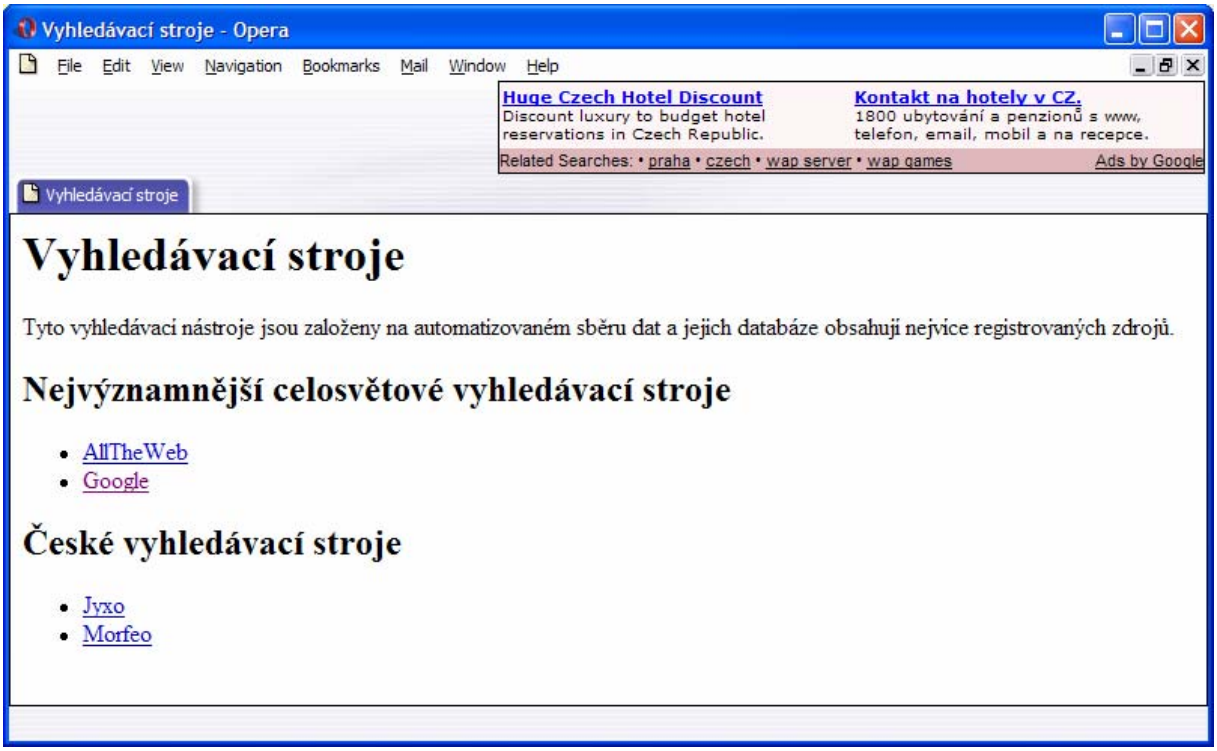
```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN"
    "http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="cs">
<head>
    <title>Vyhledávací stroje</title>
 </head>
```

```
 <body>
<h1>Vyhledávací stroje</h1>
<p>Tyto vyhledávací nástroje jsou založeny na automatizovaném
sběru dat a jejich databáze obsahují nejvíce registrovaných
zdrojů.</p>
<h2>Nejvýznamnější celosvětové vyhledávací stroje</h2>
<ul>
<li><a href="http://www.alltheweb.com/">AllTheWeb</a></li>
<li><a href="http://www.google.com/">Google</a></li>
</ul>
<h2>České vyhledávací stroje</h2>
<ul>
<li><a href="http://jyxo.cz/">Jyxo</a></li>
<li><a href="http://morfeo.centrum.cz/">Morfeo</a></li>
</ul>
</body>
</html>
```

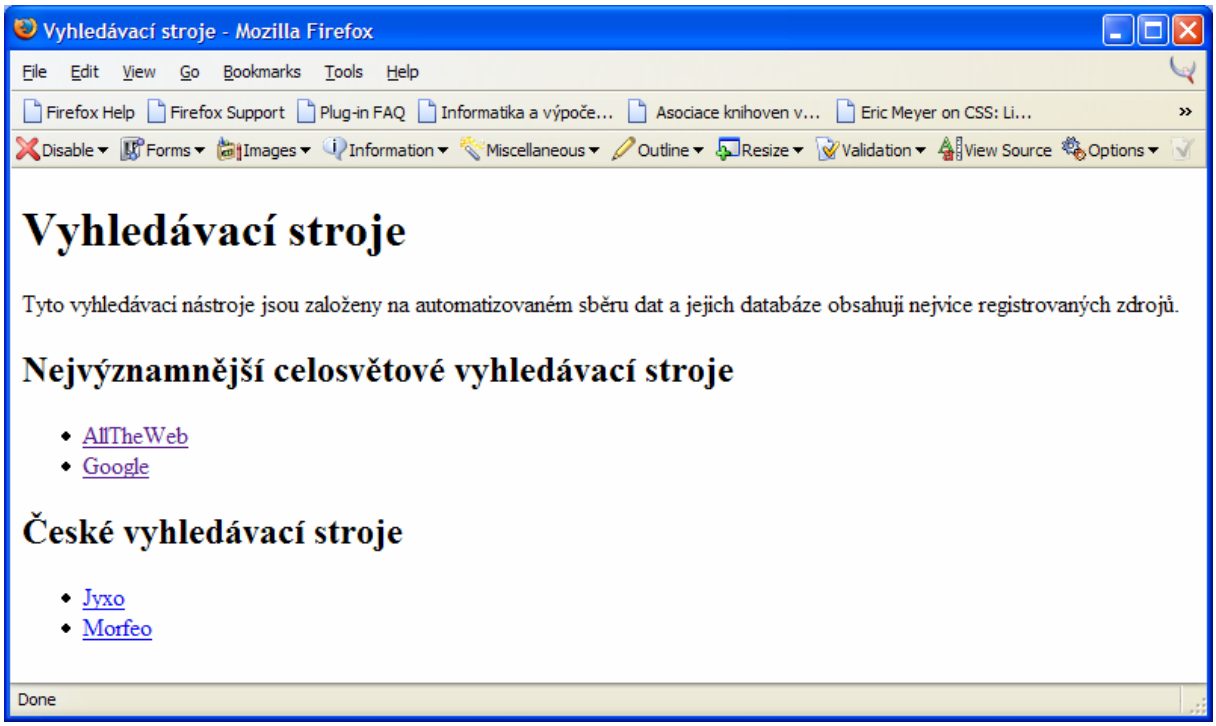
Na následujících třech obrázcích je vidět, jakým způsobem se tento jednoduchý XHTML dokument bez stylového předpisu zobrazí v prohlížečích Internet Explorer, Opera a Mozilla Firefox – naprosto shodně. Logická struktura je zřejmá.



Obr. 1 Jednoduchý XHTML dokument bez stylového předpisu v prohlížeči Microsoft Internet Explorer



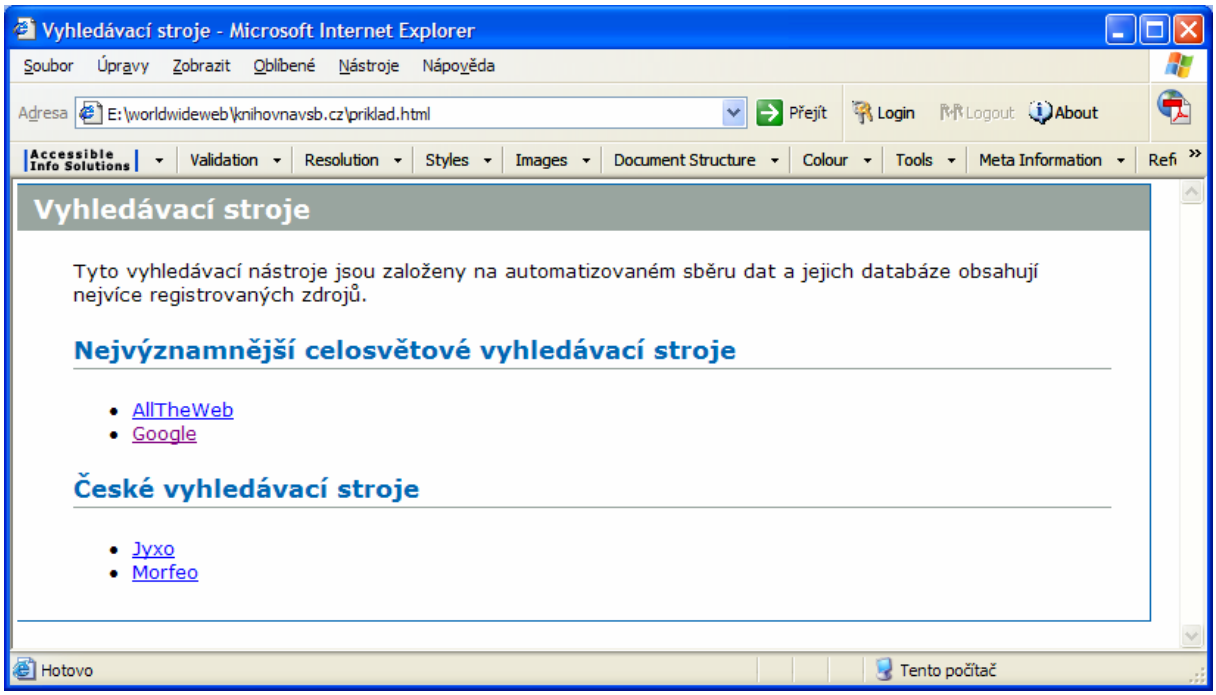
Obr. 2 Jednoduchý XHTML dokument bez stylového předpisu v prohlížeči Opera



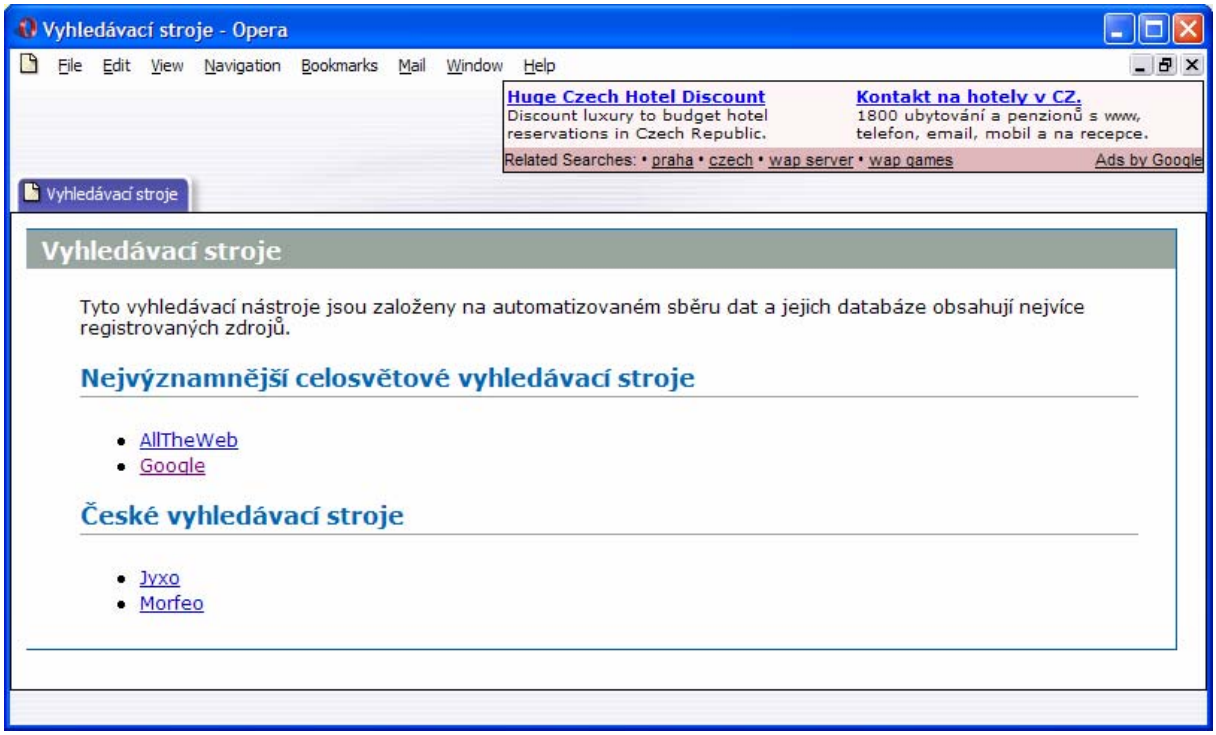
Obr. 3 Jednoduchý XHTML dokument bez stylového předpisu v prohlížeči Mozila Firefox

Připojíme-li k dokumentu stylový předpis, tj. do značky `<head>` přidáme odkaz na soubor se stylovým předpisem:

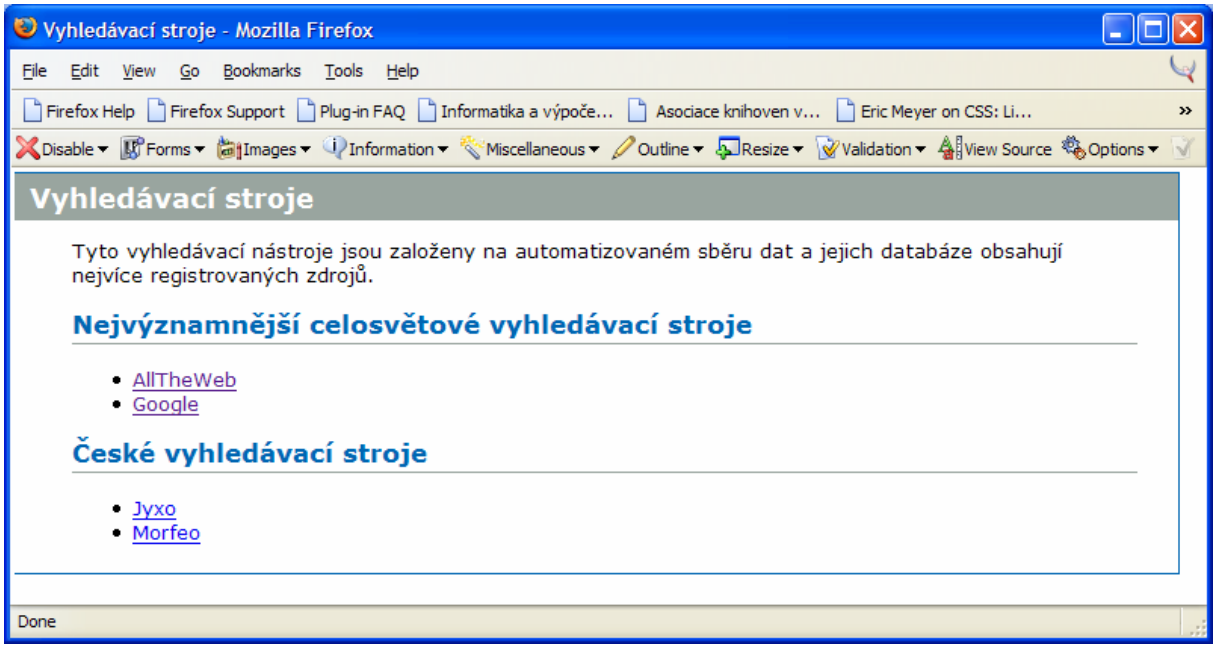
`<link rel="stylesheet" type="text/css" href="styly.css" />`,
a do zdrojového dokumentu vložíme značky `<div id="page">` a `<div id="obsah">`, které jimi označeným úseků nastaví společné vlastnosti pro formátování, zobrazí se dokument v prohlížečích takto, tedy opět takřka shodně:



Obr. 4 Jednoduchý XHTML dokument se stylovým předpisem v prohlížeči Microsoft Internet Explorer



Obr. 5 Jednoduchý XHTML dokument se stylovým předpisem v prohlížeči Opera



Obr. 6 Jednoduchý XHTML dokument se stylovým předpisem v prohlížeči Mozilla Firefox

7.3 Vlivy XML

K nejvýznamnějším vlivům XML na tvorbu webových dokumentů se řadí:

- důraz na logické značkování struktury dokumentů,
- rozlišování mezi malými a velkými písmeny (*case sensitivity*); XHTML používá malá písmena,
- nezbytnost dodržování syntaktických pravidel jazyka, jehož výsledkem je správně vytvořený (*well-formed*) dokument,
- nutnost používání koncových značek všech prvků, např. `</p>`, `</li>`, a tomu odpovídající ošetření prázdných značek doplněním lomítka: `<br />`, `<hr />`,
- používání uvozovek při zápisu hodnot vlastností prvků, např. u obrázkových prvků: `<img src="studdipl.jpg" alt="studovna diplomových a disertačních prací" />`.

Jednou z mnoha výhod respektování webových standardů při vytváření webových dokumentů je možnost ověření jejich platnosti prostřednictvím nástrojů, které jsou pro tento účel k dispozici (např. W3C Validator [18]). Validátor dokument zkontroluje, zda je správně strukturován a zda jeho zdrojový kód souhlasí s deklarovanou definicí typu dokumentu (DTD) v záhlaví.

7.4 Výhody oddělení struktury dokumentu od formátování jeho vzhledu

Oddělení struktury dokumentu od způsobu jeho zobrazení znamená, že v XHTML dokumentech **nejsou obsaženy** značky týkající se používání barev, pozadí dokumentu, způsobu formátování a typografické úpravy textu, rozvržení prvků na stránce apod. To vše je součástí stylového předpisu, který je obsažen v samostatném souboru. Tentýž dokument může být propojen s různými stylovými předpisy, které určují způsob jeho prezentace podle potřeb vý-

stupního zařízení, např. pro tiskárnu, kapesní počítač, organizér (PDA – *personal digital assistant*), mobilní telefon, televizor, zvukový syntezátor nebo pro čtecí zařízení pro Braillovo písmo. Díky oddělení vlastního obsahu dokumentu od informací o jeho výsledném vzhledu se webové dokumenty stávají přehlednějšími a pochopitelně vykazují i řadu dalších pozitivních vlastností nejen ve vztahu k uživatelům, jimž se tak stávají **přístupnými,** ale také s ohledem na možnosti jejich **zpracování pro účely vyhledávání.**

Na tomto místě je potřeba zdůraznit, že přístupnost dokumentu neznamená, že se dokument zobrazí stejným způsobem ve všech prohlížečích. Pojem přístupnost se vztahuje k obsahu dokumentu, k informacím.

Struktura dokumentu je velmi důležitá, bez ohledu na to, zda jde o dokument tradiční či elektronický. Obsah správně strukturovaného dokumentu se lépe čte a je také snadnější jej pochopit. Členění dokumentu na logické celky, zvýraznění určitých částí, vysvětlivky, citace, odkazy na další pasáže v textu či na jiné dokumenty související s tématem, to vše pomáhá uživateli při práci s jeho obsahem.

7.5 Sémantické značkování

Sémantické značkování založené na významu a smyslu součástí dokumentu znamená, že je obsah dokumentu, označen podle toho, o jaký druh informace (textu) jde. Pro tento účel jsou k dispozici logické značky založené na vyjádření obsahu, jež popisují význam textu, který je jimi označen.

Patří mezi ně např. značky `<abbr>`, `<acronym>`, `<address>`, `<blockquote>`, `<cite>`, `<code>`, `<kbd>`, `<q>`, `<samp>`, `<em>` nebo `<strong>`. Pokud autor nestanoví ve vlastním stylovém předpisu, jakým způsobem se má text uzavřený v takových značkách zobrazit, prohlížeče jej zobrazí podle vlastního stylu víceméně založeného na obecně používaných zvyklostech. Vytvoří zpravidla určitý vizuální efekt, aniž by ovšem narušily strukturu dokumentu či pozměnily autorem zamýšlený důvod pro jejich použití.

Například značky `<strong>` a `<em>` se používají pro označení těch pasáží textu, jimž přikládá autor zvláštní význam, klade na ně důraz, a chce je tedy z nějakého důvodu zvýraznit. Může jít například o důležité fráze či klíčové pojmy. Z hlediska obsahového má tedy význam to, že jde o důležitou část textu, výsledný vizuální efekt je vedlejší, neboť je závislý na koncovém zařízení uživatele. Text uzavřený do značky `<strong>` běžné prohlížeče zobrazí **tučným písmem,** text se značkou `<em>` *kurzívou* a text označený `<code> neproporcionálním písmem` s pevnou šířkou znaků. Význam takového logického značkování textu je zřetelný, uvědomíme-li si, že třeba značka `<b>` pro tučné písmo (*bold face*) nemá žádný význam pro uživatele používajícího čtecí zařízení pro Braillovo písmo nebo pro toho, kdo používá textový prohlížeč.

Při vytváření dokumentů je tedy nutné používat značky s vědomím toho, že nesou určitý význam, že byly zamýšleny pro vyjádření určitého smyslu ve vztahu k vlastnímu obsahu dokumentu. Je zřejmé, že tento způsob značkování určitých částí textu je velmi užitečný také pro zpracování dokumentů pro účely vyhledávání informací.

Je-li řeč o „webových standardech“, je důležité nezapomenout na to, že nejde jen o technologie, ale především o způsob, jakým tyto *nástroje* při své práci používají lidé. To, že někdo vytváří platné XHTML dokumenty a používá CSS pro řízení jejich vzhledu ani zdaleka neznamená, že se tím tyto dokumenty stávají automaticky přístupnými nebo přenosnými nebo že méně zatěžují přenosové linky. XHTML i CSS mohou být používány stejně špatně a nesmyslně, tak jako se to stává se staršími webovými technologiemi. Vedle toho se určitě dají na webu najít i dokumenty, jejichž autoři plně respektují nejnovější standardy a dokáží využít

možností CSS k vytvoření graficky i typograficky zajímavých webových stránek – schází jim však to hlavní, kvůli čemu web vznikl: **hodnotný a informačně bohatý obsah.**

8 Jak (X)HTML dokumenty ovlivňují webové vyhledávací služby?

Možnosti vyhledávání informací na webu jsou představovány dvěma základními typy vyhledávacích nástrojů, předmětovými katalogy založenými na ručním sběru dat a vyhledávacími stroji, jež využívají pro sběr a vytváření databází automatizovaných prostředků. Ačkoliv jsou předmětové katalogy stále oblíbeným informačním zdrojem pro řadu uživatelů webu, je nepochybné, že vyhledávací stroje poskytují nejen více informací o současném obsahu webových sídel, ale nabízejí i podstatně širší možnosti pro cílené vyhledávání.

Vyhledávací stroj je databáze zdrojů získaných z webu automatizovaným sběrem informací umístěných na webových serverech. Tuto databázi mohou uživatelé prohledávat prostřednictvím dotazů. Slova z dotazů jsou porovnávána s informacemi v databázi, nalezené informace odpovídající dotazu jsou setříděny podle toho, jak vyhovují položenému dotazu, a zobrazeny uživateli sestupně podle míry relevance.

8.1 Základní faktory ovlivňující hodnocení výsledků hledání vyhledávacími stroji

Způsob hodnocení výsledků vyhledávání vychází zpravidla ze dvou základních faktorů: z **vlastního obsahu webového dokumentu** (sídla) a z **vnějších vlivů,** jež rovněž významně přispívají k tomu, že se určitá webová sídla nebo dokumenty ocitají na předních místech při řazení výsledků vyhledávání.

Pokud jde o **obsah dokumentů,** hrají roli při vyhodnocování relevance především četnost výskytu slov v dokumentu, jejich pozice ve struktuře dokumentu, výskyt slov v dalších dokumentech v daném webovém sídle a způsob, jakým je webový dokument vytvořen (např. použití rámů může negativně ovlivnit indexaci webového sídla vyhledávacím strojem).

Mezi **vnější faktory,** které mají vliv na posouzení míry relevance vyhledaných odkazů, patří hlavně obliba zdroje, výsledky z partnerských vyhledávacích služeb a placené služby. Jestliže například na webová sídla nebo dokumenty vedou odkazy z jiných webových stránek (*link popularity*), viz například **PageRank** služby **Google,** nebo jde-li o hodně navštěvovaná webová sídla či dokumenty (*click popularity*), zvyšuje se tím míra relevance těchto zdrojů.

8.2 Hlavní funkční části vyhledávacích strojů

Vyhledávací stroje (*search engines*) jsou tvořeny čtyřmi základními funkčními částmi:

- roboty, jejichž hlavním úkolem je sběr informací na webových serverech,
- indexačním programem zpracovávajícím informace, jež získají z webu roboty,
- vyhledávacím programem, který na základě uživatelova dotazu vyhledává a zpracovává informace z databáze vytvořené indexačním programem tak, aby výsledky co nejlépe vyhovovaly položenému dotazu,
- rozhraním, jež dotazy uživatele předává vyhledávacímu stroji a zobrazuje výsledky hledání uživateli.

Z pohledu toho, jaké dokumenty nakonec služba na základě dotazu vyhledá, je velmi důležitý indexační program. Informace jsou z dokumentů získávány a zpracovávány do databází na základě rozhodnutí tvůrců těchto programů o tom, na kterých místech se v (X)HTML dokumentech vyskytují důležité informace. I z toho důvodu má význam kvalitní struktura a validní kód dokumentu. Rozsah načítaných informací bývá samozřejmě ovlivněn i

technologickým zázemím provozovatele vyhledávací služby. Sbírané údaje se proto u jednotlivých služeb liší, což je jednou z příčin rozdílného zpracování téhož dotazu několika vyhledávacími službami.

8.3 Jaké informace z webu roboty sbírají

Vyhledávací nástroje jsou jejich provozovateli zpravidla označovány jako fulltextové vyhledávače. Znamená to tedy, že by jejich databáze (indexy) měly být vytvářeny na základě zpracování úplných textů načtených z webových serverů.

Není to ovšem vždy zcela pravdivé tvrzení, i když v současnosti je možné říci, že většina nejznámějších vyhledávacích strojů – **Google, AllTheWeb, Yahoo! Search,** nebo **AltaVista** – opravdu načítá do svých databází plné texty webových dokumentů. Patří k nim i české vyhledávací stroje **Jyxo** a **Morfeo.** I zde však platí určitá omezení, jež se týkají dokumentů, jejichž velikost přesahuje jednotlivými vyhledávači stanovenou mez (například **Google** indexuje pouze úvodních 101 KB webového dokumentu, **Yahoo! Search** úvodních 500 KB).

Některé ze služeb neindexují tzv. stop-slova (členy, spojky, předložky, booleovské operátory, číslovky, velmi obecné a často se opakující slova apod.), nejsou-li součástí nějaké fráze, a ty **výrazy, jež jsou identifikovány jako spam.**

Mezi spam patří například tzv. „neviditelný text“ a velmi malé fonty. Pro neviditelný text je ve zdrojovém kódu použita stejná barva jako pro pozadí dokumentu, takže uživatel text v běžném prohlížeči nevidí. Malé fonty bývají v dokumentech používány například v zápatí pro informace, jež jsou sice důležitou součástí webových dokumentů, ale uživatelé jejich obsah zpravidla nevnímají (firemní informace, copyright apod.). Používání podobných metod autory je považováno za pokus o nežádoucí reklamu, jejímž cílem je zajistit výhodnější umístění daného zdroje ve výsledcích vyhledávání.

8.4 Nežádoucí reklama a metadata

Důsledkem podobných praktik je rovněž fakt, že prvku `<meta>`, jenž je součástí záhlaví (X)HTML dokumentu a jehož smyslem je umožnit autorům (kromě jiného) vložení informací o obsahu dokumentů (věcný popis) `<meta name="description" "content="... " />` a `<meta name="keywords" content="... " />`, žádný z významných celosvětových vyhledávacích strojů nepřikládá v současnosti význam, přestože tyto informace také jejich roboty sbírají. Tyto značky jsou obdobou **abstraktů** a **klíčových slov** v tradičních dokumentech, v článcích v odborných časopisech a v příspěvcích z konferencí. Vlivy tradičního publikování v oblasti komunikace odborných informací jsou tedy i zde zřejmé.

Z pohledu knihovníka, který zná problematiku věcného zpracování dokumentů ze své profese, je ovšem význam autorem volně tvořených klíčových slov použitých pro účely vyhledávání poněkud sporný. S ohledem na technologické možnosti současných vyhledávacích strojů rovněž. **Pro uživatele, stejně jako u tradičních publikací, jsou však tyto části dokumentu důležité při rozhodování o tom, který z vyhledaných informačních zdrojů zvolit.**

Informace umístěné ve značce `<meta>` jsou součástí „neviditelné“ části dokumentu v záhlaví `<head>` zdrojového kódu. Uživatelé obsah záhlaví v prohlížeči nevidí, s výjimkou textu značky `<title>`, jenž je „čitelný“ na titulní liště prohlížeče. Metadata jsou určena vyhledávacím nástrojům, ale ve skutečnosti jimi nebývají při vyhodnocování výsledků vyhledávání využívána z jednoho prostého důvodu: autoři webových dokumentů zneužívali

popis dokumentů a klíčová slova s cílem zajistit si lepší pozici ve výsledcích vyhledávání uváděním nepravdivých údajů nebo mnohonásobným opakováním stejných slov.

Vyhledávací nástroje takové postupy považují za spamming, neboť jednou z metod hodnocení relevance vyhledaných zdrojů je výskyt hledaných slov v dokumentech (jejich pozice v dokumentu a četnost výskytu). Čím vyšší je četnost výskytu hledaných termínů v daném dokumentu, tím výše se webové sídlo nebo jednotlivý dokument ocitne při zobrazení výsledku vyhledávání. Pokud někdo neoprávněně na svých webových stránkách použije některou z technik, jež používají vyhledávací stroje pro setřídění výsledků hledání, sníží tím vlastně úroveň jejich kvality. Některé ze služeb proto podobné metody trestají. Buď slova z textu dokumentu rozpoznaná jako spam nezahrnou do databáze nebo jim nepřikládají váhu, někdy dokonce do svých databází nezařadí takové dokumenty či webová sídla vůbec.

Vyhledávací služby metadata zpravidla zobrazují ve výsledcích vyhledávání jako doplněk k názvu vyhledaného zdroje (např. Google, AltaVista, AllTheWeb, Yahoo! Search, Jyxo, Morfeo – viz odkaz „detaily...“), i když je při hodnocení relevance neberou v úvahu. Zobrazují rovněž informace o datu vytvoření nebo aktualizace dokumentu, viz značka `<meta name="date" content="2004-04-17" />`. To je z pohledu uživatele důležitý údaj a měl by tedy být samozřejmou součástí zdrojového kódu. Informace vložené do značky `<meta>` totiž mohou být užitečným vodítkem pro uživatele při jeho vlastním vyhodnocování výsledků vyhledávání zobrazených vyhledávací službou. Metadatům se proto vyplatí při tvorbě dokumentů pro web věnovat pozornost. Pokud jde o věcný popis dokumentu, má význam zejména „meta description“. Tento prvek by proto měl obsahovat krátký text výstižně vyjadřující hlavní obsah dokumentu.

8.5 Význam součástí XHTML dokumentů pro práci vyhledávacích strojů

Mezi informace, jimž vyhledávací stroje věnují zvláštní pozornost a jimž přikládají význam, patří:

- **názvy dokumentů**, tedy text, který je označen v záhlaví dokumentu značkou `<title>` – je to první a nejspíš ta nejdůležitější část dokumentu vztahující se k jeho obsahu, kterou vyhledávací stroj získává; proto také je slovům z názvu dokumentu přikládán zásadní význam a mělo by tomu tak být i z pohledu autora: je to pole pro umístění „klíčových slov“ vyjadřujících co nejvýstižněji obsah dokumentu; názvy dokumentů jsou důležité z řady důvodů, jedním z nich je i způsob zobrazování výsledků vyhledávání – seznam nalezených odkazů na webové zdroje, v němž je text načtený z prvku `<title>` nejvýraznější položkou; právě na základě tohoto textu se uživatelé často rozhodují o tom, který z vyhledaných dokumentů odpovídá nejlépe jimi hledanému zdroji,
- **nadpisy** (titulky a mezititulky), tj. značky `<h1>` až `<h6>` členící vlastní text dokumentu do menších logických celků; vyhledávací stroje předpokládají, že autoři shrnují v těchto prvcích dokumentu obsah následujících úseků textu,
- **názvy hypertextových odkazů a URL,** názvy by měly výstižně vyjádřit obsah odkazovaného zdroje, měly by být informativní (*jaký význam asi má oblíbený pokyn „klikněte zde“?*),
- **vlastní text dokumentu,**
- **URL dokumentu,**
- **obsah značky ALT,** pokud obrázek nese významnou informaci; zvláště tehdy, nahrazuje-li text, který by jinak byl obsažen ve zdrojovém kódu (obrázkové nadpisy), je al-

ternativní text užitečný nejen pro uživatele, kteří z nějakého důvodu obrázek nevidí, ale také pro účely vyhledávání informací.

Velmi užitečným zdrojem informací je **obsah webového sídla** (*site map, site index*), tedy webová stránka, která obsahuje seznam buď všech nebo alespoň těch nejdůležitějších odkazů na dokumenty, které tvoří webové sídlo. Má význam nejen pro uživatele, jimž usnadňuje orientaci, ale také pro roboty vyhledávacích strojů, kterým pomáhá nalézt při načítání informací o daném webovém sídle všechny dokumenty.

9 Závěr

Vyhledávací stroje přikládají obsahu webových dokumentů různou váhu, jak při sběru a indexování informací, tak při hodnocení vyhledaných zdrojů ve vztahu k uživatelovu dotazu a při řazení výsledků vyhledávání. Musejí se vypořádat s řadou problémů, často vyplývajících z demokratické povahy webu, při vytváření a údržbě databází i při snahách o zvýšení úrovně kvality výsledků vyhledávání. Hlavní potíže, s nimiž se vyhledávací stroje při zpracování webových dokumentů potýkají, jsou výstižně shrnuty v [19].

V roce 2001 byla veřejnosti představena myšlenka **sémantického webu** [20]. Současný web je charakterizován neustále rostoucím množstvím webových stránek, v nichž je stále složitější nalézat potřebné informace. Vyřešit tento problém by měla právě postupná přeměna současného webu na tzv. sémantický web.

Sémantický web (Semantic Web) [21] vychází z potřeby dát obsahu webu jasný smysl a význam (sémantiku). To by mělo umožnit, aby obsahu webu porozuměly také počítače (programy). Sémantický web je založen na technologii **Resource Description Framework** (RDF)) [22], která integruje širokou škálu aplikací využívajících syntaktický zápis v **XML** a identifikátory **URI** pro pojmenování. Cílem je zajistit, aby informace (data) v dokumentech na webu měly přesně definovaný význam a bylo možné je strojově zpracovávat. K tomu by měla napomoci konceptualizace dat prostřednictvím **ontologií**. Důležitou podmínkou pro realizaci sémantického webu je standardizovaný popis webových zdrojů prostřednictvím metadat. To by mělo uživatelům umožnit pracovat s webovými zdroji jako s relační databází a dotazovat se na obsah prostřednictvím jazyků podobných SQL. Tím by měla být zajištěna vysoká míra přesnosti a relevance výsledků vyhledávání. Pro vyjádření vztahů mezi metadatovými prvky a schématy by měl sloužit RDF. Sémantika popisovaných dat by měla být zajištěna prostřednictvím klasifikačních schémat a řízených slovníků.

Přestože bylo v posledních několika letech dosaženo značného pokroku v oblasti technologií a jazyků nezbytných pro realizaci sémantického webu, je zřejmé, že jeho vybudování nebude jednoduché. Bude totiž záležet na autorech dokumentů, stejně jako je tomu dosud, zda nástroje, které jim jsou či budou k dispozici, při své práci využijí. A také, jakým způsobem je využijí. Jestliže dnes autoři nepoužívají při své práci relativně jednoduché prostředky XHTML a CSS, nelze předpokládat že budou ochotni zvládat technologie, které jsou mnohem složitější (například systematické vkládání metadat ve formátu RDF do každého dokumentu). Zvlášť když je k tomu vlastně nic nenutí a vyhledávací nástroje jako Google se evidentně na docela solidní úrovni dokáží vyrovnat i s tím, co je dnes na webu k dispozici. Sémantický web je tedy především návratem zpět k původním myšlenkám, které vedly ke vzniku webu. Je orientován hlavně na komunikaci odborných informací. Nepochybně významně ovlivní komerční nakladatele a vydavatele a další oblasti tradičně či nově spojené s šířením informací.

Použitá literatura a WWW odkazy

1. DACONTA. Michael C.; OBRST, Leo J. ; SMITH, Kevin T. *The Semantic web : a guide to the future of XML, Web services, and knowledge management.* Indianapolis : Wiley, c2003. xxii, 281 s. ISBN 0-471-43257-1.
2. HOLZSCHLAG, Molly E. *Cascading Style Sheets : the designer´s edge.* San Francisco : Sybex, c2003. xiv, 274 s. ISBN 0-7821-4184-6.
3. ZELDMAN, Jeffrey. *Designing with Web standards.* Indianapolis : New Riders, c2003. xviii, 436 s. ISBN 0-7357-1201-8.
4. http://www.collectionscanada.ca/iso/tc46sc9/standard/glossry2.htm
5. Semantic Studios | Information Architecture & User Experience, Peter Morville, http://semanticstudios.com/
6. LouisRosenfeld.com, http://louisrosenfeld.com/home/
7. http://www.w3.org/
8. http://www.iw3c2.org/Conferences/index.html
9. http://www.w3.org/MarkUp/
10. http://www.w3.org/XML/
11. http://www.w3.org/Style/CSS/
12. http://www.w3.org/AudioVideo/
13. http://www.w3.org/WAI/
14. http://www.w3.org/International/
15. http://www.w3.org/Graphics/SVG/
16. http://www.w3.org/DOM/
17. http://www.w3.org/TR/xhtml2/
18. http://validator.w3.org/
19. HENZINGER, Monika R.; MOTWANI, Rajeev; SILVERSTEIN, Craig. Challenges in web search engines. *ACM SIGIR Forum,* Fall 2002, vol. 36, issue 2, s. 11-22. Dostupný též na World Wide Web: < http:// citeseer.ist.psu.edu/henzinger02challenges.html>.
20. BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The Semantic Web : a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* [online]. 2001, May [cit. 2004-04-21]. Dostupný na World Wide Web: < http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>.
21. http://www.w3.org/2001/sw/
22. http://www.w3.org/RDF/