XHTML

HTML v XML = XHTML

V poslední době se často hovoří o formátu XHTML, který slouží k vyjádření HTML-dokumentů v XML. Pojďme se tedy blíže podívat, jaké výhody formát XHTML přináší.

Co to je HTML?

Formát nazvaný **HTML** (Hyper-Text Markup Language) byl navržen pro výměnu a prezentaci dokumentů v rámci sítě. To znamená, že dokumenty zapsané v HTML lze číst a zobrazovat pomocí standardních síťových prohlížečů, které jsou schopny tento formát akceptovat a zobrazovat takto zapsané dokumenty do značné míry obdobně, bez ohledu na typ a verzi prohlížeče.

Formát HTML se inspiroval starším a obecnějším standardem **SGML** (Standard Generalized Markup Language – ISO 8879). Dokumenty v HTML jsou správně vytvořené dokumenty dle SGML – HTML je rovněž jazyk používající značky (markups). Formát HTML ovšem překonal původní očekávání a doznal značného rozšíření. V souvislosti s jeho oblibou se vyvíjely další verze; v současnosti se používá verze 4.01, která oproti původnímu formátu obsahuje řadu novinek. Přesto je stále orientována na prezentaci dokumentů – sada značek HTML je pevná a slouží k vyjádření prezentační podoby dokumentu.

Co to je XML?

Formát **XML** (eXtensible Markup Language) je definice vytvořená pracovní skupinou W3C (World Wide Web Consortium) jako formát pro přenos obecných dokumentů. Princip XML je založen na jednoduché myšlence – přenášet spolu s dokumentem i popis jeho struktury (spolu s daty i metadata).

Při návrhu XML využili autoři rovněž podmnožinu standardu SGML. Dokumenty v XML jsou tedy automaticky i dokumenty SGML (XML je aplikace SGML). SGML je ale složitější a komplikovanější, což je pravděpodobně příčina, proč zatím nedošlo k jeho širšímu užití.

Na rozdíl od HTML je XML orientováno nikoliv na prezentační stránku dokumentu, ale na jeho strukturu. Způsobem prezentace se XML nezabývá – prezentaci ponechává XML na prohlížeči, případně jsou popsány transformace XML do prezentačních formátů (včetně HTML). Konsorcium W3C navrhlo rovněž standard **XSL** (eXtensible Stylesheet Language), jako prostředek pro popis transformace XML do prezentační podoby.

Rozdíl mezi XML a HTML

Rozdíl mezi HTML a XML lze přiblížit čtenáři na příkladu tzv. “stylů” u textového procesoru. Textový procesor umožňuje psát text různým písmem. Můžeme tedy např. nadpisy kapitol psát větším písmem a tučně – každý nadpis kapitoly musíme takto systematicky označit. To je způsob odpovídající HTML – vyznačíme, jak by měl dokument vypadat.

Jinou možností je označit všechny nadpisy (stejné úrovně) jedním stylem. Změnou stylu pak lehce změníme prezentaci všech nadpisů. To je způsob odpovídající XML – vyznačíme, co je nadpis. Způsob zobrazení není tak podstatný, rozhodne jej prohlížeč. Uvažme jako příklad tento článek zapsaný v HTML.

```
<HTML>
<HEAD>
<TITLE> HTML v XML = XHTML </TITLE>
</HEAD>
<BODY>
<H1> HTML v XML = XHTML </H1>
<H3> Karel Richta </H3>
<H2> Co to je HTML? </H2>
<P> Formát nazvaný <B>HTML</B> … </P>
<P> Formát HTML se inspiroval … </P>
<H2> Co to je XML? </H2>
<P> Formát <B>XML</B> … </P>
<P> Při návrhu XML využili … </P>
```

```
…
<H2>Literatura</H2>
<OL>
<LI></LI>
<LI></LI>
</OL>
</BODY>
</HTML>
```

Je zde jasně vidět orientace HTML na prezentaci. Totéž, zapsáno v XML, mnohem lépe vystihuje podstatu struktury daného dokumentu – XML dovoluje použít speciální značky pro vyznačení struktury tohoto typu dokumentu.

```
<clanek>
  <nazev> HTML v XML = XHTML </nazev>
  <autor> Karel Richta </autor>
<sekce>
  <nazev> Co to je HTML? </nazev>
  <odstavec> Formát nazvaný <B>HTML</B> … </odstavec>
  <odstavec> Formát HTML se inspiroval … </odstavec>
</sekce>
<sekce>
  <nazev> Co to je XML? </nazev>
  <odstavec> Formát <B>XML</B> … </odstavec>
  <odstavec> Při návrhu XML využili … </odstavec>
</sekce>
…
<literatura>
  <citace></citace>
  <citace></citace>
</literatura>
</clanek>
```

Výše uvedený dokument je správně vytvořen dle pravidel XML – je správně uzávorkován (well-formed). Na rozdíl od HTML však obsahuje nestandardní značky, vyjadřující strukturu přesně tohoto typu dokumentu. V XML můžeme navíc strukturu dokumentu předepsat tzv. definicí typu dokumentu – **DTD** (Document Type Definition). Pokud chceme strukturu dokumentu v XML kontrolovat, je definice struktury dokumentu (v našem příkladu dokumentu typu “**clanek**”) dokonce nutná.

Strukturu článku lze předepsat následující definicí DTD (speciální gramatikou pro články). Tato gramatika stanoví, že dokument typu “**clanek**” obsahuje právě jeden element “**nazev**”, neprázdnou posloupnost elementů typu “**autor**” a “**sekce**”, jeden element “**literatura**” a volitelně i element “**priloha**”.

```
<?xml version="1.0"?>
<!DOCTYPE clanek [
    <!ELEMENT clanek (nazev,autor+,sekce+,literatura,priloha?)>
    <!ELEMENT nazev (#PCDATA)>
    <!ELEMENT autor (jmeno,prijmeni)>
    <!ELEMENT jmeno (#PCDATA)>
    <!ELEMENT prijmeni (#PCDATA)>
    <!ELEMENT sekce (nazev,odstavec+)>
    <!ELEMENT odstavec (#PCDATA)>
    <!ELEMENT literatura (citace+)>
    <!ELEMENT citace (odstavec+)>
    <!ELEMENT priloha (#PCDATA)>
]>
<clanek> … </clanek>
```

Definice typu dokumentu umožňuje libovolnému prohlížeči, či jiné aplikaci, strukturu dokumentu (v našem případě článku) kontrolovat. Navíc je pro XML definován standardní nástroj zvaný XML-procesor, který umí číst libovolné XML-dokumenty a předávat aplikacím jednotlivé elementy. Je-li validující, umí dokonce přímo kontrolovat správnost (validitu) dokumentu podle stanoveného DTD.

Co to je XHTML?

HTML má pevnou sadu značek, kterou však bylo třeba v každé verzi doplňovat. XML má uživatelsky definovanou, a tedy libovolnou sadu značek. Přidávat nové značky není problém. Strukturu dokumentů lze předepsat a kontrolovat. Existují standardní nástroje pro zpracování XML-dokumentů.

Podle odhadu konsorcia W3C se předpokládá, že již v roce 2002 bude cca 75 % dokumentů na internetu v XML. Aby byly jednoduše použitelné i dokumenty v HTML, navrhlo konsorcium W3C formát XHTML, který slouží pro vyjádření HTML-dokumentů v XML. Smyslem je, aby bylo možno HTML-dokumenty zpracovávat stejně jako XML-dokumenty a aby bylo možno je jednoduše doplňovat o nové konstrukty.

XHTML je sada dokumentů (aktuálních i budoucích), které popisují HTML 4 jako aplikaci v XML. Pružnost XML umožňuje snadné rozšiřování možností. Druhou výhodou je interoperabilita dokumentů zapsaných v XML. Dokumenty v XHTML jsou vždy XML-dokumenty a lze je zpracovávat nástroji XML. XHTML 1.0 je první specifikace formátu XHTML (současná verze). Jedná se o reformulaci tří typů dokumentů dle HTML 4 na XML-dokumenty (aplikace XML 1.0).

Striktně konformní dokument v XHTML 1.0 je správně uzávorkovaný (well-formed) dokument v XML 1.0, který je validní proti jedné ze tří definic DTD (Strict, Transitional, Frameset). Navíc musí splňovat následující podmínky:

kořenem XML stromu musí být element **<html>**;

atribut **xmlns** (XML Namespace) tohoto elementu musí mít hodnotu: http://www.w3c.org/1999/xhtml;

před elementem **<html>** musí být v dokumentu stanoveno DTD odkazem na jeden ze tří formátů HTML 4.

Nejjednodušší XHTML-dokument tedy vypadá následovně.

```
<?xml version="1.0" encoding="UTF/8"?>
<!DOCTYPE html
    PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
    "DTD/xhtml1-strict/.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
  <head>
    <title> HTML v XML = XHTML </title>
  </head>
  <body>
    <p> Přesunuto na <a href="http://cs.felk.cvut.cz/">xml.xml</a>.</p>
  </body>
</html>
```

Článek v XHTML by pak mohl mít následující tvar.

```
<?xml version="1.0" encoding="UTF/8"?>
<!DOCTYPE html
    PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
    "DTD/xhtml1-strict/.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
<head>
<title> HTML v XML = XHTML </title>
</head>
<body>
<h1> HTML v XML = XHTML </h1>
<h3> Karel Richta </h3>
<h2> Co to je HTML? </h2>
<p> Formát nazvaný <B>HTML</B> … </p>
<p> Formát HTML se inspiroval … </p>
<h2> Co to je XML? </h2>
<p> Formát <B>XML</B> … </p>
<p> Při návrhu XML využili … </p>
…
<h2>Literatura</h2>
<ol>
<li></li>
<li></li>
</ol>
</body>
```

```
</html>
```

Rozdíly mezi XHTML 1.0 a HTML 4.01

V příkladu jsou vidět některé rozdíly, které nutně musí mezi HTML 4 a XHTML 1.0 existovat. Jeden důležitý rozdíl spočívá v tom, že XML rozlišuje malá a velká písmena (je case-sensitive). Všechny značky XHTML jsou proto povinně malými písmeny.

Další rozdíly vyplývají z toho, že XML vyžaduje, aby dokument byl správně uzávorkován – v HTML se často připouští zkratky (např. konstrukce může být bez koncové závorky, která se automaticky doplní). Elementy se nesmí překrývat, což řada prohlížečů HTML tolerovala. Navíc musí být popsány všechny hodnoty atributů (nelze je zkracovat) a je nutno je vždy uvádět v uvozovkách (i když se jedná o čísla).

Nakonec ještě jeden tip – validaci správnosti dokumentu v XHTML si můžete nechat ověřit na adrese uvedené v následujícím dokumentu.

```
<!DOCTYPE html PUBLIC
  "-//W3C//DTD XHTML 1.0 Transitional//EN"
  "DTD/xhtml1-transitional.dtd">
<html      xmlns="http://www.w3.org/1999/xhtml">
<head>
  <title>Minimal document</title>
</head>
<body>
<p>
  <a href="http://validator.w3.org/check/referer">
            validate</a>
</p>
</body>
</html>
```

Literatura

Bray, T. – Paoli, J. – Sperberg-McQueen, C. M. (eds.): *Extensible Markup Language (XML) 1.0. W3C Recommendation 10-February-1998.* World Wide Web Consortium, 1998, URL: *www.w3c.org/TR/REC-xml*

Clark, J. – Deach, S. (eds.): *Extensible Stylesheet Language (XSL) 1.0. W3C Working Draft 16-December-1998.* World Wide Web Consortium, 1998, URL: *www.w3c.org/TR/WD-xsl*

Clark, J. (ed.): *XSL Transformations (XSLT) 1.0. W3C Proposed Recommendation 8-October-1999.* World Wide Web Consortium, 1999, URL: *www.w3.org/TR/xslt*

Raggett, D. – Hors, A. L., Jacobs, I. (eds.): *HTML 4.0 Specification. W3C Recommendation 24-April-1998.* World Wide Web Consortium, 1998, URL: *www.w3c.org/TR/REC-html40*

Richta, K.: Proč XML? Chip, vol. 2, 2000, str. 98 – 99

www.xml.com

www.ibm.com/developer/xml

www.microsoft.com/xml

Karel Richta