Kapitola 1. Úvod do značkovacích jazyků, základní pojmy, logická a fyzická struktura dokumentu

Obsah

Úvod do XML, odkazy na specifikace

Úvod, definice, motivace, historie,…

- XML je standard (přesněji doporučení konsorcia W3C [http://www.w3.org]) jak vytvářet značkovací jazyky.
- Jedná se tedy o metajazyk.
- Ideově vychází ze zhruba o deset let mladšího standardu SGML (Structure Generalized Markup Lan-

guage).

- Se základním standardem úzce souvisí několik dalších, např. XML Namespaces, XInclude, XML Base.
- Tyto spolu s dalšími standardy (XSLT, XSL-FO, XHTML, CSS...) tvoří "rodinu" standardů XML.

Jazyky rodiny SGML a jejich aplikace (HTML)

- SGML jako standard ISO 8879: http://www.iso.ch/cate/d16387.html
- Stručný "laický" úvod do SGML: http://www.sil.org/computing/noc/156ac.htm
- Další stručný, věcný ale přesný úvod do SGML: http://www.oasis-open.org/cover/naggumWhat.html
- Další Úvod do SGML: http://www.personal.u-net.com/~sgml/sgml.htm
- Srovnání SGML-XML a další zdroje: http://www.oasis-open.org/cover/sgml-xml.html

Extensible Markup Language (XML)

- World Wide Web Consortium (W3C): http://www.w3.org/
- Přehled XML aktivit W3C: http://www.w3.org/XML/Activity - specifikace standardů, konference, odkazy na SW, referenční nástroje, odkazy *(obnovováno podle potřeby)*
- XML Startkabel (EN/NL): http://xml.startkabel.nl - aktuality, odkazy *(obnovováno cca 1x týdně)*
- Zvon: http://zvon.org - asi nejlepší soubor tutoriálů, on-line referencí v mnoha jazycích, místo je hostované v ČR.
- *What is XML?* na XML.COM: http://www.xml.com/pub/a/98/10/guide0.html - jeden z úvodních článků ke XML
- The Extensible Markup Language (XML) USENET newsgroup: news:comp.text.xml - nejznámější USENET news skupina ke XML
- XML-DEV: mailto:xml-dev@xml.org - nejznámější maillist ke xml standardům
- Archív xml-dev: http://lists.xml.org/archives/xml-dev/ - archív předchozího maillistu
- XML: XML Quick Syntax Reference Card [http://www.mulberrytech.com] - výborná stručná referenční karta

Extensible Markup Language (XML) - další odkazy, zejména na software

- IBM DeveloperWorks, sekce XML: http://ibm.com/developer/xml [http://ibm.com/developer/xml/] - články, tutoriály, software atd. na vysoké technické úrovni
- IBM AlphaWorks: http://www.alphaworks.ibm.com - alpha-software fy IBM k volnému vyzkoušení
- O'Reilly XML.COM: http://xml.com - články, tutoriály atd. na vysoké technické úrovni
- Free XML Software (L. M. Garshol): http://www.garshol.priv.no/download/xmltools/ - asi nejlepší kolekce odkazů na nekomerční XML software
- XMLSoftware: http://xmlsoftware.com - asi nejlepší kolekce odkazů na obecný XML software (i komerční)
- *XML Cover Pages* na www.oasis-open.org [http://www.oasis-open.org]: xml.coverpages.org [http://xml.coverpages.org] - denně aktualizovaný soubor odkazů na články, publikace standardů, software, atd. v oblasti XML. Nejlepší zdroj v této kategorii.

Charakteristika a základní zásady XML

Charakteristika XML jazyků

- XML není jeden konkrétní značkovací jazyk; je to specifikace určující, jak mají vypadat značkovací jazyky
- jedná se tedy o "metajazyk"
- konceptuálně jde o zjednodušení SGML standardu, aby se usnadnila práce tvůrcům parserů (analyzátorů) a aplikací

  - například v tom, že každý element musí být uzavřen; pak na přečtení struktury dokumentu nemusíme mít DTD
- XML navazuje na úspěšnou implementaci SGML - jazyk HTML; má podobné charakteristiky z hlediska zaměření na internet
- Momentálně je aktuální specifikací [http://www.w3.org/TR/REC-xml] Extensible Markup Language (XML) 1.0 (Third Edition) W3C Recommendation 4th February 2004, François Yergeau, Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler
- Připravuje se *XML 1.1*, nyní jako *Candidate Recommendation*.
- Vážné diskuse se vedou okolo *binárního XML*, což by měla být rovnocenná reprezentace stejného modelu, jako má "textové" XML.

Základní specifikace: XML 1.0 (Third Edition)

- Momentálně je aktuální specifikací [http://www.w3.org/TR/REC-xml] *Extensible Markup Language (XML) 1.0 (Third Edition) W3C Recommendation* 4th February 2004, autorů François Yergeau, Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler
- Současně s tím je zveřejněna *XML 1.1, W3C Recommendation*, 4th February 2004, François Yergeau, John Cowan, Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler.
- Vážné diskuse se vedou okolo *binárního XML*, což by měla být rovnocenná reprezentace stejného modelu, jako má "textové" XML.

Deset zásad pro specifikaci XML standardů

vyňato z preambule ke specifikaci XML 1.0 (Second Edition)

1. XML shall be straightforwardly usable over the Internet.

   XML bude přímočaře použitelné na Internetu.

2. XML shall support a wide variety of applications.

   XML bude podporovat širokou škálu aplikací.

3. XML shall be compatible with SGML.

   XML bude kompatibilní se SGML.

4. It shall be easy to write programs which process XML documents.

   Tvorba programů zpracovávajících XML bude jednoduchá.

5. The number of optional features in XML is to be kept to the absolute minimum, ideally zero.

   Počet volitelných prvků XML standardu bude málo, optimálně 0.

6. XML documents should be human-legible and reasonably clear.

   XML dokumenty by měly být "lidsky" čitelné a rozumně jednoduché.

7. The XML design should be prepared quickly.

   Návrh XML standardu by měl být rychle hotov.

8. The design of XML shall be formal and concise.

   Návrh XML musí byt formální a správný.

9. XML documents shall be easy to create.

   XML dokumenty bude možné snadno vytvořit.

10. Terseness in XML literal is of minimal importance.

Úspornost XML značkování není podstatná.

Struktura XML dokumentů

Syntaxe XML dokumentů

Základním požadavkem kladeným na *každý* XML dokument je, že musí být *dobře utvořen (well-formed)*.

Toto nastane, právě když:

1. Taken as a whole, it matches the production labeled `document`: `[1] document ::= prolog element Misc*` tj. obsahuje *prolog (hlavičku)* a právě jeden, tzv. kořenový *element*. Dále může před a po kořenovém elementu obsahovat instrukce pro zpracování, komentáře atd. (`Misc`)
2. It meets all the well-formedness constraints given in the specification.

   Musí vyhovovat všem pravidlům pro správné utvoření uvedeným ve specifikaci.
3. Each of the *parsed entities* which is referenced directly or indirectly within the document *is well-formed*.

   Totéž platí pro každou *analyzovanou (parsovanou) entitu* přímo nebo nepřímo odkazovanou v dokumentu.

Podívejte se na tutoriál základů XML v češtině [http://zvon.org/xxl/XMLTutorial/General_cze/book.html]

Rejstřík (glossary) pojmů ke XML [http://zvon.org/index.php?nav_id=173]

Fyzická a logická struktura XML dokumentu

U XML dokumentů rozlišujeme strukturu fyzickou a logickou. Aplikační programátory zajímá většinou jen struktura *logická*, autory obsahu, XML editorů, procesorů, atd. může zajímat i struktura *fyzická*.

Struktura *logická* - *dokument* členíme na *elementy* (jedne z nich je *kořenový* - *root*), jejich *atributy*, *textové uzly* v elementech, *instrukce pro zpracování*, *notace*, *komentáře*

Struktura *fyzická* - jeden logický dokument může být uložen ve více fyzických jednotkách - entitách; vždy alespoň v jedné - tzv. *entitě dokumentu* - *document entity*.

Elementy

Jsou objekty *ohraničené počáteční a koncovou značkou* - *start and end tag*, např.:

```
<tagname ...tag_attributes...>
```

```
  tag_content
</tagname>
```

Příklad 1.1. Příklad elementu s obsahem (neprázdného)

```
<body background="yellow">
   <h1>textový uzel - obsah elementu h1</h1>
   <p>textový uzel - obsah elementu p</p>
</body>
```

Elementy - prázdné

Je-li obsah prázdný (žádné dceřinné elementy ani textový obsah), pak píšeme pouze *značku prázdného elementu - empty element tag*, např.:

```
<tagname tag_attributes/>
```

Příklad 1.2. Příklad elementu bez obsahu (prázdného)

```
<hr width='50%'/>
```

Atributy

- Nesou "dodatečné informace" k elementu - např. jeho ID, požadované formátování - styl, odkazy na další elementy...
- Konceptuálně by bylo možné atributy nahradit elementy, ale z důvodu přehlednosti obvykle používáme obojí.
- Obsah atributu na rozdíl od obsahu elementu není nijak (na úrovni obecných zásad XML standardů) dále strukturován.

  Občas se u některých značkování vyskytne snaha o strukturaci obsahu atributů, to však obecně vede k více problémům, než řeší...

Atributy - zápis

- Atribut je tvořen *jménem* a *hodnotou*.

- Atributy zapisujeme do počáteční značky elementu (který může být i prázdný).
- Hodnota je *vždy* vložena v uvozovkách či apostrofech a od jména ji dělí znak rovnítka (=).
- Pro *názvy* atributů platí stejná omezení jako pro názvy elementů.
- V rámci jednoho elementu *nejsou* přípustné dva atributy se stejným jménem.

Atributy - příklad

Příklad 1.3. Atribut 'width' v prázdném elementu

```
<hr width='50%'/>
```

Příklad 1.4. Atribut 'border' v neprázdném elementu

```
<table border='1'>
   <tr><td>jedna</td><td>dve</td></tr>
   <tr><td>tri</td><td>ctyri</td></tr>
</table>
```

Textové uzly

Nesou textovou informaci.

Např. v následující ukázce je *text* `ahoj!` (nikoli celý element `em`!) textovým uzlem:

```
<em>ahoj!</em>
```

Instrukce pro zpracování

Instrukce pro zpracování (*processing-instruction*) píšeme do značek `<?target content>`

Informují aplikaci o postupu či nastavení nutném pro zpracování daných XML dat. Nepopisují (nepředstavují) obsah, ale *zpracování* dokumentu.

```
<?xsl-stylesheet href="mystyle.xsl">
```

Poznámka

`href` v příkladu *neznamená* atribut; atributy nejsou u instrukce pro zpracování možné!

Notace

Notace (*notation*) píšeme do značek `<!NOTATION name declaration >`

Slouží zejména k popisu binárních (non-XML) entit - např. obrázků GIF, PNG,...

Jde vlastně o *deklaraci způsobu zpracování*.

Komentáře

Podobně jako v HTML - komentář (*comment*) píšeme do značek `<!-- text komentáře -->`

Komentář nebývá obvykle pro zpracování významný, ale záleží na aplikaci - může např. uchovávat značky *Servlet-side Includes (SSI)*

Parsery by proto *měly* komentáře zpracovávat - předávat dále.

Např. *SAX parsery* však tak nečiní!!! (resp. činí až v rozšířené verzi SAX2, v Javě balík `org.xml.sax.ext`).

Entity

Entita je základní jednotkou *fyzické* stavby dokumentu. Odpovídá řetězci, souboru...

Parsery by měly entity zpracovávat tak, aby aplikace nemusela o entitách "nic tušit".

Podrobněji...

Podrobné informace k syntaxi se dozvíme v následující kapitole Standardy rodiny XML [../standards/index.html]