Počítačová lexikografie XML, standardy

Adam Rambousek

XML

- eXtensible Markup Language značkovací (meta)jazyk
- pravidla, jak má vypadat správně vytvořený dokument snadné strojové zpracování a výměna informací
- konkrétní názvy značek určuje uživatel (standardy, vlastní)
- elementy <značka>obsah</značka>
- bez obsahu lze <značka></značka> zkrátit na <značka/>
- atributy <značka atribut="hodnota"/>

XML

- správné zanoření značek
- správně: <a>text
- špatně: <a>text
- speciální znaky se přepisují na entity (např. <)
  - <, >, &

XML

- popis a kontrola obsahu
- DTD (Document Type Definition)
  - seznam elementů a atributů a vztahy mezi nimi
  - nekontroluje obsah
  - <!ELEMENT vyznam (definice, priklad+)>
  - <!ATTLIST vyznam cislo CDATA #REQUIRED>

XML

- XML Schema (XSD, XML Schema Definition)
  - popis obsahu a struktury XML dokumentu, schéma samotné je XML dokument
  - elementy, atributy, struktura
  - možnost určit vlastní typy obsahu (např. opakující se adresa)
  - kontrola obsahu (např. číselný rozsah, regulární výrazy, povolené hodnoty)
  - ```
    <xs:element name="definice">
      <xs:simpleType>
        <xs:restriction base="xs:string">
          <xs:maxLength value="120"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:element>
    ```
    ```

Standardy založené na XML

- web: XHTML
- matematika: MathML
- knihy: EPUB
- grafika: SVG
- dialogové systémy: VoiceXML
- metadata, sémantický web: RDF
- text: TEI

XSL(T)

- eXtensible Stylesheet Language (Transformations)
- převod XML na jiné formáty
  - jiné XML značkování, text, HTML, LaTeX, PDF
- šablony pro části XML dokumentu, postupné procházení dokumentu
- (funkcionální programovací jazyk)

XML databáze

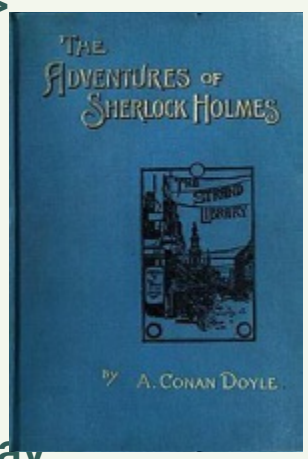
- ukládají se přímo XML dokumenty
- vyhledávání XPath, XQuery
- např. eXist, BaseX, Sedna

TEI

- Text Encoding Initiative
  - http://www.tei c.org/
- TEI Guidelines (aktuálně verze 5 z roku 2007)
  - XML formát pro sémantický popis textových dokumentů
  - velký rozsah značek
  - TEI Lite osekaná verze, "90 % potřeb 90 % uživatelů"
  - romány, poezie, divadelní hry, dokumentace, slovníky, korpusy, grafy, rukopisy, zarovnání, odkazy, změny textu, notové zápisy…
  - nástroje sada XSLT pro převod na LaTeX, docx, EPUB, HTML

TEI

```
<text>
 <front>
  <head rend="italic">Adventures of Sherlock Holmes</head>
  <docTitle>
   <titlePart>Adventure II. —</titlePart>
   <titlePart>The Red  Headed League</titlePart>
  </docTitle>
  <byline>By A. Conan Doyle.</byline>
 </front>
 <body>
  <p>I had called upon my friend, Mr. Sherlock Holmes, one day
      in the autumn of last year and found him in deep conversation
      with a very stout, florid  faced, elderly gentleman with fiery red hair …
  </p>
 </body>
</text>
```



TEI

-
  <head>ACT I</head>

  <head>SCENE I</head>
  <stage rend="italic"> Enter Barnardo and Francisco, two Sentinels, at several doors</stage>
  <sp>
  <speaker>Barn</speaker>
  <l part="Y">Who's there?</l>
  </sp>
  <sp>
  <speaker>Fran</speaker>
  <l>Nay, answer me. Stand and unfold yourself.</l>
  </sp>

TEI

- <q>My dear <rs type="person">Mr. Bennet</rs>, </q>
  said his lady to him one day,
  <q>have you heard that <rs type="place">Netherfield Park</rs> is let at last?
  </q>

- <s n="1">
  <w ana="#NP0">Marley</w>
  <w ana="#VBD">was</w>
  <w ana="#AJ0">dead</w>
  <pc>:</pc>
  <w ana="#TO0">to</w>
  <w ana="#VBB">begin</w>
  <w ana="#PRP">with</w>
  <pc>. </pc>
  </s>

TEI, slovníky

- podoba hesla
- <entry>
  <form>
  <orth>competitor</orth>
  <hyph>com|peti|tor</hyph>
  <pron>k@m"petit@(r)</pron>
  </form>
  <gramGrp>
  <pos>n</pos>
  </gramGrp>
  <def>person who competes.</def>
  </entry>

TEI, slovníky

- významy
- <sense n="1">
  <gramGrp>
  <subc>VP6A</subc>
  </gramGrp>
  <def>turn (a ship) on one side for cleaning, repairing, etc.</def>
  </sense>
  <sense n="2">
  <gramGrp>
  <subc>VP6A</subc>
  <subc>VP2A</subc>
  </gramGrp>
  <def>(cause to) tilt, lean over to one side.</def>
  </sense>

TEI, slovníky

- překlady
- <form>
  <orth>dresser</orth>
 </form>
  <sense>
   <usg type="dom">Theat</usg>
   <cit type="translation" xml:lang="fr">
    <quote>habilleur</quote>
    <gramGrp>
     <gen>m</gen>
    </gramGrp>
   </cit>

TEI, slovníky

- příklady
- <cit type="example">
  <quote>the multiplex eye of the fly.</quote>
  </cit>
- <cit type="example">
  <quote>elle était horrifiée par la dépense</quote>
  <cit type="translation" xml:lang="en">
  <quote>she was horrified at the expense.</quote>
  </cit>
  </cit>

TEI, slovníky

- příznaky
- <form>
  <orth>colour</orth>
  <form>
   <usg type="geo">U.S.</usg>
   <orth>color</orth>
  </form>
 </form>
- <usg type="syn">aube de roue</usg>
- <usg type="dom">Constr</usg>

TEI, projekty

- Oxford Text Archive
- British National Corpus
- FreeDict
- Cambridge University Press
- Chinese Buddhist Electronic Text Association
- Deutsches Textarchiv
- Europeana Regia

Ukázka, SSČ

- <root>
  <h>lov</h>
  <gram> u m</gram>
  <sens>
  <num>1</num>
  <exp>lovení zvěře a ryb</exp>
  <exm><t>lov koroptví</t></exm>
  <exm><t>lov na zajíce</t></exm>
  <exm><t>liška vyšla na lov</t></exm>
  </sens>
  <sens>
  <num>2</num>
  <ref>úlovek<refcateg>syno</refcateg></ref>
  <ref>kořist<refcateg>syno</refcateg></ref>
  <exm><t>mít bohatý lov</t></exm>
  </sens>
  </root>
- přepis

Ukázka, SSJČ

- <root>
  <h>lov</h>
  <norm> u</norm>
  <small>m.</small>
  <small>(</small>
  <small>6. j.</small>
  <norm> u)</norm>
  <bold>1.</bold>
  <ital>stíhání a zmocňování se zvěře</ital>
  <ital>(</ital>
  <ital>nejč. odstřelem); chytání ryb:</ital>
  <norm>
  l. jelenů, divokých kachen, velryb; l. lososů; l. perel; doba lovu; uspořádat l. na medvědy; vyjet na l.; právo lovu; l. odstřelem, chytáním, lapáním; l. lesní, polní, vodní; hromadný l.
  </norm>
- skenováno, OCR

Ukázka, PSJČ

- <h>
  <Cil>lov</Cil>
  <Heslo>lov,</Heslo>
  <Tvar> u</Tvar>
  <Gram>m.</Gram>
  <Vyzn>honba n. lapání zvěře n. chytání ryb.</Vyzn>
  <Dokl>Vrchnost na lovu byla.</Dokl>
  <Pram>Něm.</Pram>
  <Sep>D</Sep>
  <Char>Expr.</Char>
  <Vyzn>chytání, krádež, získávání, shánění čehokoliv.</Vyzn>
  <Dokl>Netopýr na lovu kmitl se kolem.</Dokl>
  <Pram>Baar.</Pram>
- skenováno, OCR, lepší struktura

- různé XML formáty, stejný vzhled (XSLT)

SSJC Slovník spisovného jazyka českého

lov

-u m. (6. j. -u)

1. *stíhání a zmocňování se zvěře (nejč. odstřelem); chytání ryb:* l. jelenů, divokých kachen, velryb; l. lososů; l. perel; doba lovu; uspořádat l. na medvědy; vyjet na l.; právo lovu; l. odstřelem, chytáním, lapáním; l. lesní, polní, vodní; hromadný l. *hon;* liška vyšla na l.; lovu zdar! *(lovecký pozdrav)*

2. expr. *chytání, shánění čehokoliv, vůbec získávání, při kterém se uplatní obratnost a náhoda:* l. vzácného hmyzu; sběratelé se vydali na l. lidových písní; policie podnikla l. na zloděje; expr. to je l.! *šťastný nález, výhodná koupě ap.*

3. *výsledek lovu; úlovek, kořist:* vrátit se s bohatým lovem *s ulovenou zvěří ap.*, přen. expr. *s věcmi získanými obratností n. šťastnou náhodou*

SSC Slovník spisovné češtiny

lov

-u m

1. *lovení zvěře a ryb* lov koroptví, lov na zajíce, liška vyšla na lov,

2. *úlovek (syno) kořist (syno)* mít bohatý lov,
