Jelölőnyelvek és a TEI XML

Vétek Bence,
PIM DBK

Tartalom

- A jelölőnyelvekről általában
- SGML, HTML és XML
- Az XML-ről bővebben

  Feladat: Receptkönyv
- A TEI XML

Jelölőnyelvek

- A szövegek vmilyen céllal való megjelölése metaadatokkal való ellátása már a számítógépes nyelvek kialakulása előtt elterjedt.
  (pl. nyomdászok, szedők, szerkesztők)
- Nem egyenlő a programozási nyelvekkel → nem képes feldolgozásra vagy stílusbeli igazításra
- Metanyelv: egy szöveg (nyelv) elemeinek jelölésére szolgáló nyelv

Jelölőnyelvek

SGML

- Standard Generalized Markup Language = szabványos általános jelölőnyelv
- Általános jelölőnyelv: SGML-ben a jelölések (tag) jelentése nincs meghatározva, ez mindig az SGML-t használó alkalmazás feladata
- 1986: ISO
- Eredeti célja: lehetővé tenni a nagy kormányzati és ipari projektek dokumentumainak közzétételét számítógép által is beolvasható formában
- Szöveges alapú adatbázisok kezelése.
- Első nagyobb alkalmazása: Oxford English Dictionary második kiadás

SGML

```
<QUOTE TYPE="example">
  Valami blabla, benne egy <ITALICS>kiemelt</ITALICS> szakasz.
</QUOTE>
```

- Az SGML lehetővé teszi, hogy a szintaxist az egyes alkalmazásokban testre szabjuk (kis- és nagybetű érzékenység, jelölők határainak megadása, stb.).
- A szintaxist DTD-ben határozzuk meg.
- Egy általános SGML dokumentum csak 7 bites ASCII kódolást használhat.
- Különleges karakterek: SGML-entitások (DTD)

```
<QUOTE TYPE="example">
  Valami sz&ouml;veg, benne egy <ITALICS>kiemelt</ITALICS> r&eacute;sz.
</QUOTE>
```

HTML

- HyperText Markup Language = hiperszöveges jelölőnyelv
- Szöveges formában adjuk meg a weboldal tartalmát és a tartalom kinézetét (nem csak szöveg, hanem kép, videó vagy hangfájl is)
- Az SGML alkalmazása
- Megjelenítési nyelv: a böngésző értelmezi a tageket, és azok alapján megjeleníti a szöveget, képeket stb.

```
<!DOCTYPE html>
<html>
  <head>
    <title>Az oldal címe</title>
    <!--esetleges további fejléc-információk-->
  </head>
  <body>
    <p>első bekezdés</p>
    <p>második bekezdés</p>
  </body>
</html>
```

XML

- A szövegek (online) tárolásának és feldolgozásának alapvető formátuma: **XML** (Extensible Markup Laguage = Kiterjeszthető jelölőnyelv)
- Mind ember, mind gép számára olvasható
- Eszköz- és rendszerfüggetlen
- Különböző adattípusok jelölése
- Kiterjeszthető, nincs fix *tag*-készlete
- Strukturált szöveg és információ megosztása az interneten

XML

- Az XML, akárcsak az elődje, az SGML, lehetővé teszi, hogy az elemeket saját magunk nevezhessük el. A következő példában használhattuk volna a <pers> elemet a <personName> helyett és <townName> elemet a <town> helyett.

```
<personName>Robert Smith</personName>
<street>Castle street</street>
<houseNr>123</houseNr>
<town>Dublin</town>
<telephoneNr>0891232245</telephoneNr>
```

- A fenti példa így is kinézhet:

```
<pers>Robert Smith</pers>
<streetName>Castle street</streetName>
<house>123</house>
<townName>Dublin</townName>
<telephone>0891232245</telephone>
```

XML – A tagek alapszabályai

1. Minden taget le kell zárni!
   - <author>Homer</author>
   - <title>Odyssey</title>
   - <author>Homer
     <title>Odyssey
2. Bármit írhatunk a tagbe, csak maradjunk következetesek!
   - <author>Homer</author>
   - <Writer>Homer</Writer>
   - <theGuyWhoWroteThis>Homer</theGuyWhoWroteThis>
3. Nem lehetnek átfedések!
   - <poem>
     <author>Homer</author>
     <title>Odyssey</title>
     </poem>
4. Mindig egy gyökérelem fogja közre az egész XML-kódot

<poem>
<author>Homer
<title>Odyssey
</poem>
</author>
</title>

XML

Szabályos XML elemnevek	
<_newElement> </newElement>	Az elemnévnek betűvel vagy aláhúzással kell kezdődnie.
<newElement> </newElement>	Az elemnevek kis- és nagybetű érzékenyek, továbbá a kezdő- és zárócímkének egymáshoz illőnek kell lennie.
<my.new_Element-1></my.new_Element-1>	Az elemnevek tartalmazhatnak betűt, számjegyet, kötőjelet, aláhúzást és pontot is.

Nem megfelelő XML elemnevek	
<1Element> </1Element>	Az elemnév nem kezdődhet számjeggyel, kötőjellel vagy ponttal.
< Element /> </ Element>	Az elemnév nem kezdődhet szóközzel.
<newElement> </Newelement>	Az elemnevek kis-és nagybetű érzékenyek, a kezdő- és zárócímkének pedig egymáshoz illőnek kell lennie.
<xmlElement></xmlElement>	Az elemnév nem kezdődhet a következő szavakkal: xml, XML, Xml, stb.
<new Element> </new Element>	Az elemnév nem tartalmazhat szóközt.

XML

```
<anthology>
  <poem>
    <heading>The SICK ROSE</heading>
    <stanza>
      <line>O Rose thou art sick.</line>
      <line>The invisible worm,</line>
      <line>That flies in the night</line>
      <line>In the howling storm:</line>
    </stanza>
    <stanza>
      <line>Has found out thy bed</line>
      <line>Of crimson joy:</line>
      <line>And his dark secret love</line>
      <line>Does thy life destroy.</line>
    </stanza>
  </poem>
<!-- more poems go here -->
</anthology>
```

XML – Attribútumok

- Lehetővé teszi, hogy további információt adjunk egy elemhez

```
<person first-name="Henry" last-name="James" />
```

- Hasonló elnevezési szabályok, mint az elemeknél

Az attribútumok érvényes használata XML-ben	
<newElement attribute1="attribútumérték: 1" />	Az attribútumnév tartalmazhat számjegyeket, de nem kezdődhet számmal. Az attribútumérték tartalmazhat szóközt, írásjelet és alfanumerikus karaktereket bármilyen sorrendben.
<person name='Rob Miller' /> <person name="Rob Miller" />	Szimpla és dupla idézőjel is határolhatja az attribútumértéket.
<address owner="Mary's address" />	Szerepelhet szimpla idézőjel az attribútumértékben, de nem határolóelemként.
<sentence spoken='He said: "go"!' />	Dupla idézőjel is szerepelhet az attribútumértékben, de nem határolóelemként.

XML – Attribútumok

Az attribútumnevek- és értékek nem megfelelõ használata az XML-ben	
<person name=Robert />	Az attribútumértéket idézőjelbe kell tenni.
<person name="Robert' />	Mindkét idézőjelnek azonosnak kell lennie.
<sentence spoken="He said: "go"!" />	Ha dupla idézőjel határolja az attribútumértéket, magában az értékben már nem szerepelhet.
<address owner='Mary's address' />	Ha szimpla idézőjel határolja az attribútumértéket, magában az értékben már nem szerepelhet.
<person first name="Frank" />	Az attribútumnév nem tartalmazhat szóközt.
<person 1stname="Robert" />	Az attribútumnévnek betűvel vagy aláhúzással kell kezdődnie.
<person name="Robert" name="James" />	Egy elemnek nem lehet két attribútuma azonos névvel.

XML

```
<?xml version="1.0" encoding="UTF-8"?>
<Recept név="kenyér" elk_idő="5 perc" sütés_idő="3 óra">
  <cím>Egyszerű kenyér</cím>
  <összetevő mennyiség="3" egység="csésze">Liszt</összetevő>
  <összetevő mennyiség="10" egység="dekagramm">Élesztő</összetevő>
  <összetevő mennyiség="1.5" egység="csésze">Meleg víz</összetevő>
  <összetevő mennyiség="1" egység="teáskanál">Só</összetevő>
  <Utasítások>
    <lépés>Keverj össze minden összetevőt, aztán jól gyúrd össze!</lépés>
    <lépés>Fedd le ruhával és hagyd pihenni egy óráig egy meleg szobában!</lépés>
    <lépés>Gyúrd össze újra, helyezd bele egy bádogedénybe, aztán süsd meg a sütőben!</lépés>
  </Utasítások>
</Recept>
```

XML eszközök

- XPath
- XQuery
- XSLT (Extensible Stylesheet Language Transformations)
- CSS (Cascading Style Sheets)
- Validálhatóság (RELAX NG, DTD, XML Schema)

XML - Feladat

Receptkönyv

Az interneten számos gasztronómiai oldal található tele receptekkel. A legtöbb receptnek hasonló tartalomtípusa és szerkezete van. Ehhez a feladathoz keress egy online, receptekkel foglalkozó weboldalt, és próbáld megtalálni a receptek modelljét.

- Határozz meg legalább öt dolgot, amelyeket címkézni szeretnél a receptekben, és gondolkodj megfelelő elemeken és/vagy attribútumneveken!
- Nyisd meg az XML szerkesztőt és hozz létre egy új XML dokumentumot, amelyet recept.xml-nek nevezz el!
- Másold ki az egyik receptet az általad felkeresett honlapról a recept.xml-be!
- Címkézd fel a recept különböző részeit az elemeiddel és attribútumaiddal!
- Ugyanebbe a fájlba illessz be még két receptet a fenti szabályok szerint, majd zárd le a receptkönyvet.

DTD

- Document Type Definition = dokumentumtípus-definíció
- Jelölési előírások halmaza SGML-típusú (SGML, HTML, XML) dokumentumokhoz
- A DTD tömör formális szintaxist alkalmaz, amely pontosan megmutatja, mely elemek hol fordulhatnak elő az adott típusú dokumentumban, és hogy az elemeknek milyen tartalmuk és tulajdonságaik lehetnek.
- DTD deklarálására két mód áll rendelkezésre: belső vagy külső. A belső deklaráció ugyanabban a fájlban van, amelyben maga a dokumentum. A külső DTD külön fájlban tárolódik.

TEI XML

- Tex Encoding Initiative = szövegkódolási kezdeményezés
- Szövegek digitális megjelenésének standardizálása
- Irányelvek gyűjteménye, amelyek meghatározzák a géppel olvasható szövegek kódolását
- Az egyszerű megjelenítésnél fontosabb a szemantika
- Minden tag jelentése és értelmezése előre meghatározott
- Kb. 500 különböző szövegkomponens (pl. <persName>, <placeName>, <note>)

TEI XML

- 1987-ben három számítógépes nyelvészeti és irodalmi kutatásokkal foglalkozó tudományos társaság alapította:
  - Association for Computers and the Humanities (ACH)
  - Association for Computational Linguistics (ACL)
  - Association for Literary and Linguistic Computing (ALLC)
- irányvonalak kifejlesztése, terjesztése a géppel olvasható szövegek kódolására, közvetíthetőségére, és cserélhetőségére, valamint javaslatok tétele új szövegek kódolására
- ma már rengeteg tag a világ minden részéről, éves konferenciák, munkabizottságok, részterületek (pl. kéziratok kritikai kiadása)

TEI XML

- 2000-től konzorciális keretben működik
- A közös munka eredménye mindig egy DTD, vagyis dokumentumtípus-deklaráció, amely a nyelv jelölőelemeit és egymáshoz való viszonyukat határozza meg.
- szépirodalmi művek, kritikai kiadások, történeti források, élőszöveg átiratok elektronikus feldolgozása
- TEI P5 (2007)
- TEI Guidelines (ajánlások): tartalmazzák a TEI alapelveit, tag-készletét, használati útmutatókat stb.
- Kezdőknek: TEI Lite

TEI XML - alapelvek és előnyök

- Egy olyan csereformátum létrehozásához, amely platformfüggetlen, szükség van egy sajátságos szintaxisra, jól, előre definiált elemekre.
- SGML → XML
- Sem konkrét előírások, sem korlátozások → testreszabható, mégis univerzális
- Az ajánlás a kódolásokat oly módon definiálja, hogy a felhasználó megtudja belőle, mi miért történt a jelölés során.
- Általános célú kódolási forma:
  - szövegek egyidejű, különböző szempontú feldolgozása
  - különböző alkalmazások
  - lefedi tudományos célú szövegkutatások nagy részét

TEI XML

```
<p>
  <s>
   <cl>It was about the beginning of September, 1664,
   <cl>that I, among the rest of my neighbours,
        heard in ordinary discourse
    <cl>that the plague was returned again to Holland; </cl>
    </cl>
   </cl>
   <cl>for it had been very violent there, and particularly at
      Amsterdam and Rotterdam, in the year 1663, </cl>
   <cl>whither, <cl>they say,</cl> it was brought,
   <cl>some said</cl> from Italy, others from the Levant, among some goods
   <cl>which were brought home by their Turkey fleet;</cl>
   </cl>
   <cl>others said it was brought from Candia;
      others from Cyprus. </cl>
  </s>
  <s>
   <cl>It mattered not <cl>from whence it came;</cl>
   </cl>
   <cl>but all agreed <cl>it was come into Holland again.</cl>
   </cl>
  </s>
</p>
```

TEI XML

```
<div type="sonnet">
 <lg type="quatrain">
  <l>Les amoureux fervents et les
  <l> Aiment également, dans leur
  <l> Les chats puissants et doux,
  <l> Qui comme eux sont frileux e
 </lg>
 <lg type="quatrain">
  <l>Amis de la science et de la v
  <l> Ils cherchent le silence et
  <l> L'Érèbe les eût pris pour se
  <l> S'ils pouvaient au servage i
 </lg>
 <lg type="tercet">
  <l>Ils prennent en songeant les
  <l>Des grands sphinx allongés au
  <l>Qui semblent s'endormir dans
 </lg>
 <lg type="tercet">
  <l>Leurs reins féconds sont plei
  <l> Et des parcelles d'or, ainsi qu un sable fin,</l>
  <l>Étoilent vaguement leurs prunelles mystiques.</l>
 </lg>
</div>
```

```
<anthology>
  <poem>
    <heading>The SICK ROSE</heading>
    <stanza>
      <line>O Rose thou art sick.</line>
      <line>The invisible worm,</line>
      <line>That flies in the night</line>
      <line>In the howling storm:</line>
    </stanza>
    <stanza>
      <line>Has found out thy bed</line>
      <line>Of crimson joy:</line>
      <line>And his dark secret love</line>
      <line>Does thy life destroy.</line>
    </stanza>
  </poem>
<!-- more poems go here -->
</anthology>
```

TEI XML

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="https://distantreading.github.io/Schema/eltec-0.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="HU01431" xml:lang="hu">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>Úri muri : ELTeC edition</title>
            <author ref="viaf:22179627">Móricz Zsigmond</author>
            <respStmt>
               <resp>ELTeC conversion</resp>
               <name>Palkó Gábor<ref target="https://viaf.org/viaf/65989506"/>
               </name>
               <name>Fellegi Zsófia</name>
               <name>Vétek Bence</name>
            </respStmt>
         </titleStmt>
         <extent>
            <measure unit="words">72029</measure>
            <measure unit="pages">268</measure>
            <measure unit="vols"/>
         </extent>
         <publicationStmt>
            <p>Published as part of ELTeC</p>
         </publicationStmt>
         <sourceDesc>
            <bibl type="digitalSource">
               <title>Úri muri</title>
               <date>2000-01-08</date>
               <publisher>Magyar Elektronikus Könyvtárért Egyesület</publisher>
               <ref target="http://mek.oszk.hu/01400/01431"/>
               <respStmt></respStmt>
                  <resp>Original Electronic Edition</resp>
                  <name>Somogyi Sándor</name>
                  <name>Gács Daniella</name>
```

TEI XML

```
            <relatedItem type="source">
               <bibl type="digitalSource">
                  <title>Úri muri</title>
                  <author>Móricz Zsigmond</author>
                  <pubPlace>Budapest</pubPlace>
                  <publisher>Móra.</publisher>
                  <date>1982</date>
                  <idno type="isbn-10">963 11 3071 1</idno>
               </bibl>
            </relatedItem>
         <bibl type="firstEdition">
            <title>Úri muri</title>
            <author>Móricz Zsigmond</author>
            <date>1928</date>
         </bibl>
      </sourceDesc>
   </fileDesc>
   <encodingDesc n="eltec-0">
      <p/>
   </encodingDesc>
   <profileDesc>
      <langUsage>
         <language ident="hu"/>
      </langUsage>
      <textDesc>
         <authorGender xmlns="http://distantreading.net/eltec/ns" key="M"/>
         <size xmlns="http://distantreading.net/eltec/ns" key="medium"/>
         <canonicity xmlns="http://distantreading.net/eltec/ns" key="high"/>
         <timeSlot xmlns="http://distantreading.net/eltec/ns" key="T5"/><!--T4-nek 1920-szal van vége. ez 1928-as. ezért T5-->
      </textDesc>
   </profileDesc>
   <revisionDesc>
      <change when="2019-04-18"/>
   </revisionDesc>
</teiHeader>
```

TEI XML

```
69      </revisionDesc>
70    </teiHeader>
71    <text>
72      <body>
73        <div type="chapter">
74          <head>1.</head>
75          <p>A Sárga rózsában csak Borbíró ült egyedül.</p>
76          <p>Ült a spriccere mellett, s nézett a levegőbe. Úgy el tudott ülni hétszámra, hogy
77            egyet se szólott, a világon semmire kíváncsi nem volt, csak ült s nézett. Nézte, hogy
78            a légy hogy mászik a falon, utána nézett, míg el nem repült, akkor megint jött egy
79            másik légy, akkor meg azt nézte, osztán az is elrepült egyszer.</p>
80          <p>Akkor rámeredt a falon függő naptárra, s azt nézte: <hi rend="italic">Június.</hi>
81            Később a számot nézte: <hi rend="italic">7.</hi> Nézte, de nem gondolt hozzá semmit.
82            Azt úgyis tudta, hogy péntek van, s azt is, hogy a Millénnium nagy esztendeje.</p>
83          <p>Hirtelen fölneszelt, Zoltánt látta meg az utcán, Szakhmáry Zoltánt.</p>
84          <p>- Hé, Zóltán, Zóltán! - kezdett el kiabálni - Zóltán, Zóltán!</p>
85          <p>Zoltán odanézett, erre ő visszaült a helyére, s nem nézett többet a legyekre, olyan
86            egyenesre ült, mint a komondor, mikor várja a gazdáját a pincéből. Ellenben
87            észrevette, hogy a vendéglő előtt egy csoport parasztember ácsorog. Biztosan a
88            mérnököt várják a vízszabályozási munkák miatt.</p>
89          <p>Zoltán megjelent az ajtóban.</p>
90          <p>- Gyere mán Zóltánkám, az isten áldjon meg, igyál meg egy pohár sert. Meghalsz
91            szomjan ebbe a melegbe.</p>
92          <p>Szakhmáry Zoltán nem is mosolygott, természetesnek vette ezt a jóságot, amely oly
93            méltó volt e régi templomhoz. A könnyű vinkók és fanyar csigerek áldozati helyéhez. A
94            százéves nagy épület elborulva, elbarnulva áll az alföldi izzó napsütésben, a
95            ráboruló széles porkupola alatt, s minden zuga és minden téglája élő tanúja a magára
96            maradt magyar temperamentum vergődő és verekedő tombolásának.</p>
97          <p>A korcsma volt ez. Ahol az emberek mindig megtalálták, már akik keresték, a
98            vigasztalást, a barátságot és a feledést.</p>
99          <p>Zoltán intett a pincérnek, s az szaladt a sörért.</p>
100         <p>- Nem is tudtam, hogy idebe vagy, azt hittem, odaki vagy.</p>
101         <p>- Bejöttem.</p>
102         <p>- Azír, mer nem vótál idebe.</p>
103         <p>Zoltán kinézett:</p>
```

```
7574            <trailer>.oOo.</trailer>
7575          </div>
7576        </body>
7577      </text>
7578    </TEI>
7579
```

FIN