

# Az XML technológia

## 1. bevezető

*Mészáros Tamás*

<http://www.mit.bme.hu/~meszaros/>

Budapesti Műszaki Egyetem  
Méréstechnika és Információs Rendszerek Tanszék

# A XML előadások áttekintése

- Alapismeretek
  - A technológia áttekintése, az XML előzményei és jelentősége
- Az XML dokumentum és deklarációja
  - Az XML dokumentum felépítése
  - DTD és Schema deklaráció
- Dokumentumok elemzése, feldolgozása
  - Eseményalapú elemzők
  - Struktúraalapú elemzők
- XML transzformációk
  - Dokumentum-elérés: XPath
  - Transzformáció XSLT segítségével
- Strukturált dokumentumok tárolása
  - Natív és hagyományos relációs adatbázis-kezelők használata

## Ebben a részben ...

- megtudjuk, mi is az XML, mire alkalmas
- megismerkedünk az XML nyelvvel, legfontosabb fogalmaival
- áttekintjük az XML technológia elemeit
- összefoglalót kapunk a terület legfontosabb szereplőiről és szabványairól
- röviden megismerjük az XML kialakulásának történetét
- megismerkedünk az XML családfájával és „örökségével”
- a modul végén összefoglaljuk az XML előnyeit és veszélyeit

# Mi az XML?

- Szöveges dokumentum formátum
- Metanyelv
- Szabványcsalád
- Technológia
  
- (*WeschWeb film*)

# Az XML, mint dokumentum formátum

- Dokumentum formátum szabványok
  - de facto: DOC, PDF, RTF, ...
  - de jure: HTML, SGML, XML, ODF, ...
- Szöveges vagy strukturált dokumentumok (megjelenés vagy tartalom)
  - szövegszerkesztés és -megjelenítés
  - információ-elérés (-átvitel, -tárolás)
- Strukturált dokumentum
  - dokumentum definíció (DTD)
  - strukturált dokumentum tartalom
  - dokumentum megjelenés
- Jelölő nyelv
  - a struktúra ábrázolása címkékkel
  - a címkék típusai: procedurális és leíró

*Újrahasznosítható, a számítógép által „megérthető” szöveges tartalom*

# XML példák

```
<név>Alma Mater</név>
```

```
<ember>
```

```
  <név>Alma Mater</név>
```

```
</ember>
```

Hétfőn <ember><név>Alma Mater</név></ember> levizsgázott.

```
<p>
```

Ez egy <i>dőlt betűvel</i> írt szöveg.

```
</p>
```

Hétfőn <ember><név>Alma Mater</név></ember>

**<i>jelesre</i>** vizsgázott.

Hétfőn <ember><név>Alma Mater</név></ember>

**<érdemjegy>jelesre</érdemjegy>** vizsgázott.

Hétfőn <ember><név>Alma Mater</név></ember>

**<i><érdemjegy>jelesre</érdemjegy></i>** vizsgázott.

# Az XML példák nem mindig szépek

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<wsdl:definitions name="DiscoService"
  targetNamespace="http://disco.mit.bme.hu/service" xmlns:wsdl="http://schemas.xmlsoap.org/wsdl/"
  xmlns:tns="http://disco.mit.bme.hu/service" xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/">
<wsdl:types>
  <xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    targetNamespace="http://disco.mit.bme.hu/service">
    <xsd:complexType name="DiscoLanguage">
      <xsd:sequence>
        <xsd:element name="id" type="xsd:int"></xsd:element>
        <xsd:element name="locale" type="xsd:string"></xsd:element>
        <xsd:element name="lang_id" type="xsd:int"></xsd:element>
        <xsd:element name="name" type="xsd:string"></xsd:element>
      </xsd:sequence>
    </xsd:complexType>
    <xsd:complexType name="DiscoLanguageList">
      <xsd:sequence>
        <xsd:element ref="tns:DiscoLanguage" minOccurs="0" maxOccurs="unbounded"></xsd:element>
      </xsd:sequence>
    </xsd:complexType>
    <xsd:complexType name="DiscoTerm">
      <xsd:sequence>
        <xsd:element name="id" type="xsd:int"></xsd:element>
        <xsd:element name="lang_id" type="xsd:int"></xsd:element>
        <xsd:element name="code" type="xsd:string"></xsd:element>
        <xsd:element name="name" type="xsd:string"></xsd:element>
        ...
      </xsd:sequence>
    </xsd:complexType>
  </xsd:schema>
</wsdl:types>
</wsdl:definitions>
```

# Az XML, mint metanyelv

- Metanyelv
  - segítségével nyelvek definiálhatók
  - címkekészlet és struktúra
  - egy deklaráció (DTD, Schema) rögzíti a nyelvet
- XML deklaráció példák

```
<!ELEMENT ember (név)>
```

```
<complexType name="ember">  
  <sequence>  
    <element name="név" type="string"></element>  
  </sequence>  
</complexType>
```



# Deklarációs rendszerek: DTD és Schema

- Dokumentum Típus Deklaráció (DTD)
  - az eredeti XML szabvány része
  - az eredeti (W3C) célnak megfelel
  - korlátozott alkalmazás-építés (nagy deklarációs rendszerek gondja)
  - nincsenek névterek
  - nincs adattipizálás
  - nem XML szintaxisú
- XML Schema szabvány
  - új deklarációs rendszer
  - adattipizálás (programozási nyelvekhez hasonló)
  - objektum-orientált megközelítésből örökölt deklarációépítés
  - finomítható, szűkíthető, bővíthető egyszerű és összetett típusok
  - névtér támogatás
  - XML szintaxis

# Az XML, mint technológia

- technológia: specifikáció (szabvány) és termék halmaz, amely strukturált dokumentumok készítését, feldolgozását és megjelenítését támogatja
- sokféle alkalmazási területen nyújt eszközöket és módszereket a feladatok megoldásához:
  - Web megjelenés (szerver és kliens oldali transzformációk)
  - adatcsere (formátum, transzformáció) – e-üzlet (b2b, b2c)
  - szövegek reprezentációja és feldolgozása
  - Web 2.0
  - szövegszerkesztők dokumentum formátuma (OpenOffice, MS Office)
  - technikai dokumentációk nyelvezete (SGML örökség)
  - szoftverek konfigurálása
  - felhasználói interfészek definiálása (Windows Vista)
  - EU önéletrajzok készítése (Europass CV)
  - ...

# Az XML technológia fontosabb elemei

- Dokumentum formátum és szövegszerkesztés
  - grafikus, kiterjeszhető, testre szabható szövegszerkesztők
- Deklarációs metanyelv, alkalmazás-specifikus nyelvek
  - grafikus metanyelv-szerkesztők
  - nyelvi kötések programozási nyelvekhez (pl. JAXB)
  - ipari szabványos XML nyelvek
- Megjelenítés (és transzformáció)
  - megvalósítási környezettől független módszerek
  - szabványos megjelenítési nyelvek
- Programozott elérés
  - programozási nyelvtől független módszerek
- Tárolás és lekérdezés
  - hagyományos, relációs adatbázisok XML kiterjesztése
  - újszerű, natív XML adatbázisok

# Az XML szabványai

- Jelentősebb szabványosítási szervezetek
  - World Wide Web Consortium (W3C)
  - ISO/IEC JTC-1
  - OASIS: Organization for the Advancement of Structured Information Standards
- Jelentősebb XML alapszabványok
  - Extensible Markup Language (XML) (1998, 2004)
  - XML Namespaces (1999, 2004)
  - XML Schema (2001, 2004)
  - DOM (1998, 2000, 2004)
  - Simple API for XML (SAX)
  - XML Path Language (XPath) (1999, 2007)
  - XSL Transformations (XSLT) (1999, 2007)
  - XHTML (2000, 2001, 2007)
  - XQuery: An XML Query Language (2007)

# Összefoglalás: mi az XML

- Szöveges dokumentum formátum
  - lehetővé teszi a szövegtartalom könnyű gépi feldolgozását
  - keresés helyett lekérdezéssel jutunk az információhoz
- Metanyelv
  - a tartalom leírására (jelölésére) mesterséges nyelvet alkothatunk
  - ezáltal lehetővé válik a jelölt tartalom gépi „megértése”
- Szabványcsalád
  - széles körben elfogadott ipari szabványai vannak
  - gyártói egyetértés
- Technológia
  - sok területen alkalmazhatunk XML-re épülő megoldásokat és eszközöket

# Az XML története

- Az XML története
  - előzményei már a hatvanas évek végén megjelentek (GML, SGML) – kevés sikerrel
  - az alapszabvány 1998-ban született meg
  - kiegészítő szabványai azóta folyamatosan
  - ipari elterjedése már 2000-től jelentős (web, e-üzlet, ...)
- Történeti mérföldkövek
  - Szöveges dokumentumok számítógépes készítése (írógép helyett)
  - A könyvtártudomány információkeresési problémája
  - A mesterséges intelligencia és a természetes nyelvű szövegek megértése
  - Technikai dokumentációk strukturált tárolása
  - A web „információrobbanása”
  - Strukturált dokumentum-kezelés a weben

# Miért jött létre az XML?

- World Wide Web Consortium (W3C): web technológiák fejlesztése
  - bajban van a web (kilencvenes évek, de ma is)
  - dinamikusan növekvő tartalom, kőkorszaki tartalomelérési módszerek
  - a rendszerek a tartalom helyett a megjelenésre koncentrálnak
  - a számítógép nem képes a tartalom megértésére, így nem támogatja a hatékony információelérést
  - „Szemantikus web” vízió – a hatékony webes információelérés álma
- Számítógépes szövegkezelés (-szerkesztés, -feldolgozás, stb.)
  - a számítógép a szövegkezelés szempontjából nem több, mint egy ügyes írógép
  - a természetes nyelvek gépi megértése (mesterséges intelligencia), azaz a hatékony szöveges információelérés egyelőre álom
  - a hagyományos szöveg- (információ-)kereső rendszerek hatékonysága korlátos (a könyvtártudomány már az 50-es évek óta dolgozik rajta)
  - nincs áttörés a szöveges információ kezelésében és elérésében
  - „Strukturált dokumentumkezelés” – segítsük a számítógépet a megértésben

# Strukturált dokumentum-kezelés és az XML

- Strukturált dokumentum-kezelés
  - hatvanas években indult (IBM: GML)
  - nyolcvanas évekre szabványossá vált (ISO: SGML)
  - aztán csend: drága eszközök, eljárások, alkalmazás – csak nagyoknak
- World Wide Web Consortium (W3C)
  - tartalom-orientált web dokumentumok kellene: egy \*ML jó lenne
  - az SGML drága, nehézkes, legyen egy új „webre termett” szabvány
  - az SGML és a webes eszközök előnyeit ötvözve megszületett az XML
- Az XML (Extensible Markup Language)
  - tartalomleírás a számítógépek (feldolgozók, keresőgépek) számára
  - olcsó eszközök, könnyű alkalmazhatóság
  - „szabad kéz” a webfejlesztőknek és tartalom készítőknél
  - a szabványok szigorú betartása (nem HTML)



# Az XML „rokonsága”

- Az XML és az SGML viszonya
  - SGML: kötetlen nyelvtan – bonyolult és drága implementáció
  - XML: az SGML egyszerűsítése, „olcsóbbá tétele”, vele kompatibilis
- Az XML és a HTML viszonya
  - HTML: megjelenítésre koncentráló, nem bővíthető jelölésrendszer
  - HTML: laza (nem szabványos) szintaxiskezelés
  - XML: szigorú szintaktikai ellenőrzés és bővíthető, tartalom-orientált jelölésrendszer
- Az XML célkitűzései
  - legyen egyszerűen használható a webrendszerekben,
  - az alkalmazások széles körét támogassa,
  - legyen kompatibilis az SGML-el,
  - legyen egyszerű XML dokumentumokat feldolgozó programokat írni,
  - ember által olvasható, világos szerkezetű dokumentumok legyenek,
  - legyen egyszerű XML dokumentumokat készíteni.

# Az XML és az információ-elérés I. (adatok)

- Az informatika küzdelme az inkompatibilitással – az XML a kulcs?
  - az XML univerzális, platformfüggetlen, önleíró adatformátum
  - van egy programozási nyelvektől független, erős deklarációs rendszere
  - olcsók és széles körben terjednek az eszközei és megoldásai
  - „automatikus” nyelvi kötés programozási nyelvekhez
- Azonos nyelvet beszélünk?
  - sorban jönnek létre az egyes alkalmazási területek nyelvei
- Jó nyelvet készítettünk?
  - ipari konzorciumok készítik a specifikációkat
  - ISO és W3C szabványok is születtek
- **Az XML átalakítja az informatikai rendszereinket**
  - van gyártói egyetértés (a nagyok között is) – **szabványosítás!**
  - kényszer is van: „ha kimaradsz, lemaradsz”

# Az XML és az információ-elérés II. (szövegek)

- Az univerzális információ-elérés álma
  - a számítógép „megérti” amit tárol és azt adja vissza, amire szükségem van
- Elég az XML?
  - az XML csak egy alapeszköz
  - a W3C Szemantikus Web koncepció sok minden mást is igényel
  - kevés dolog van készen
  - a továbbfejlesztési irányok sem mindig világosak
- Azonos nyelvet beszélünk?
  - a saját tartalmát mindenki a saját nyelvén szeretné leírni
  - a nyelv mellett az értelmezésről is meg kell állapodni
- Jó nyelvet készítettünk?
  - mások igényei
  - jövőbeli igények, alkalmazási lehetőségek

# Az XML előnyei és hátránya

- Az XML lehetővé teszi szövegek tartalom-orientált, strukturált gépi reprezentációját, készítését, tárolását, feldolgozását, lekérdezését és megjelenítését.
- A számítógép képessé válik a tartalom korlátozott megértésére.
- Ezáltal lehetővé válik a tartalom automatikus, gépi ellenőrzése.
- Keresés (web) helyett lekérdezéssel (DB) juthatunk információhoz.
- Könnyen alkalmazható webes rendszerekben.
- Univerzális adatcsere formátumként hozzájárul az üzleti alkalmazások szabványos kommunikációjához
  
- Az eredeti szabvány gyenge deklarációs rendszert tartalmaz (DTD).
- Az időtálló deklarációk készítése gondos munkát igényel.
- A deklaráció tervezési hibáinak javítása igen költséges.
- Sok szabvány, rengeteg technológia, párhuzamosságok.
- Kiforratlan alkalmazói gyakorlat.

# Összefoglalás: az XML alapjai

- Az XML
  - strukturált dokumentum formátum
  - metanyelv (adateleírás, szövegreprezentáció)
  - szabványcsalád (ipari egyetértéssel)
  - technológia (üzleti megoldások széles skálája)
- Alkalmazási célterületei
  - Szöveges dokumentumok gépi „megértése” - hatékony információ-elérés
  - Híd üzleti alkalmazásaink között: szabványos, önleíró adatformátum
- Értékelése
  - Az elmúlt évtized egyik legjelentősebb technológiája
  - Sok területen még kidolgozatlan, kiforratlan alkalmazási gyakorlat
  - „buzzword” (XML, SOA)