

XML bevezető

Jeszenszky Péter

2024. szeptember 9.

Mi az XML? (1)

- Általános célú jelölőnyelv.
- Születése az 1990-es évek második felében.
- Napjaink egyik meghatározó, az iparban széles körben használt technológiája.

Mi az XML? (2)

- **Szűkebb értelemben:** szintaxis strukturált dokumentumok ábrázolására, mely lehetővé teszi azok automatikus feldolgozását (elektronikus dokumentum formátum).
- **Tágabb értelemben:** egy sereg közös töről fakadó specifikációt jelent, melyeket összefoglaló néven **XML család**nak is neveznek.

Előzmény

- Előzménye az SGML:
 - ISO szabvány elektronikus dokumentumok leírására (1980-as évek):
 - *ISO 8879:1986 Information processing – Text and office systems – Standard Generalized Markup Language (SGML)*
<https://www.iso.org/standard/16387.html>
 - Bonyolult, webes környezetben történő széleskörű használatra nem alkalmas.
- Az XML az SGML-en alapul, azzal kompatibilis (részhalmaza), de sokkal egyszerűbben használható.

Jelölőnyelvek

- A jelölőnyelvek szöveg annotálására szolgáló számítógépes nyelvek.
- Lehetővé teszik szövegrészekhez metaadatok megadását a szövegtől jól elkülöníthető módon.
- Példák:
 - AsciiDoc <https://asciidoc.org/>
 - TeX, LaTeX <https://www.latex-project.org/>
 - Markdown <https://daringfireball.net/projects/markdown/>
 - troff (man oldalak) <https://www.gnu.org/software/groff/>
 - XML <https://www.w3.org/XML/>
 - Wikitext <https://en.wikipedia.org/wiki/Help:Wikitext>

Strukturált dokumentum

- Különböző szerkezeti alkotórészek alkotják, mint például címek, fejezetek, szakaszok, bekezdések, megjegyzések, táblázatok.
- Cél az automatikus feldolgozhatóság, ehhez a jelölőnyelvek lehetővé teszik a szerkezeti alkotórészek azonosítását.
- Sokszor nem is dokumentum a szó hagyományos értelmében.

Az XML, mint jelölőnyelv

A szerkezeti alkotórészek azonosítása a dokumentumban úgynevezett címkékkel történik, mint például:

```
<author>Sir Arthur Conan Doyle</author>  
<title xml:lang="hu">A sátán kutyája</title>
```

- Az *Extensible Markup Language* (XML) kifejezés szó szerint jelentése “Kiterjeszhető Jelölőnyelv”.
- Kiterjeszhető, mert nincs előre definiált címkekészlet a szerkezeti alkotórészek azonosítására, hanem egy mechanizmust biztosít ilyen címkekészletek definiálásához.

Meta-jelölőnyelv

- Lehetővé teszi jelölőnyelvek definiálását, ezért meta-jelölőnyelvnek tekinthető.

Az XML és a HTML összehasonlítása

- XML:

- Nincs előre definiált címkekészlet
- Célja adatok leírása
- Adatcsere formátumként használják

- HTML:

- Előre definiált címkekészlet használata
- Célja információ megjelenítés
- Egy prezentációs nyelv
- Tekinthető az XML egy speciális alkalmazásának (XHTML)

Előnyök

- Egyszerűség
 - Az XML dokumentumok közönséges szöveges állományok, melyeket akár a legegyszerűbb szövegszerkesztővel is létre lehet hozni.
- Nyíltság
- Gyártófüggetlenség
- Platformfüggetlenség
- Univerzális adatcsere formátum
- Kiterjedt infrastruktúra
- Az iparban de-facto szabvány

Hátrányok

Mindezek ellenére fontos, együtt kell élni vele:

- Böbeszédű és nehézkesen használható szintaxis
- Nagy tárigény
- Bonyolultság
 - Se szeri, se száma az XML-hez kötődő specifikációknak.

Dokumentum központú XML

- A dokumentumokat jelölésekkel megtűzdelt folyó szöveg alkotja.
- A dokumentumok szerkezete nagy változatosságot mutat.
- Lényeges az elemek sorrendje.
- Az ilyen dokumentumok tartalma elsősorban emberi fogyasztásra szánt.
- Ilyen alkalmazás például az XHTML.

Adatközpontú XML

- A dokumentumokat nagyszámú adatelem alkotja.
- Kevésbé véletlenszerű dokumentum-szerkezet.
- Az elemek sorrendje kevésbé lényeges.
- Az ilyen dokumentumok elsősorban gépi feldolgozásra szántak.
- Ilyen alkalmazás például az SVG.

Alternatíva

JSON (JavaScript Object Notation) <https://www.json.org/>

- Könnyűsúlyú szöveges nyelvfüggetlen adatcsere formátum.
- Strukturált adatok ábrázolására szolgál, esetükben jelenti az XML alternatíváját.
- Nagyjából ugyanazokat az előnyöket kínálja, mint az XML, azonban annak hátrányai nélkül.

- *Extensible Markup Language (XML) 1.0 (Fifth Edition)* (W3C ajánlás, 2008. november 26.) <https://www.w3.org/TR/xml/>
 - Ez az elterjedten használt, de vele párhuzamosan létezik az XML 1.1 szabvány.
- *Extensible Markup Language (XML) 1.1 (Second Edition)* (W3C ajánlás, 2006. augusztus 16.) <https://www.w3.org/TR/xml11/>
 - Az XML 1.0 és XML 1.1 közötti eltéréseket lásd az XML 1.0 ajánlást tárgyaló bemutatóban.
 - Nem elterjedt a használata.

Állomány jellemzők

- Állománynév végződés: `.xml`
- IANA média típus: `application/xml`, `text/xml`
- Sok XML formátumnak saját állomány kiterjesztése és média típusa van.
 - Egy altípus végén a `+xml` strukturált szintaxis utótag jelzi azt, hogy a formátum az XML-en alapul.
 - Példák: `application/xhtml+xml`, `image/svg+xml`, `model/x3d+xml`

Az XML család (1)

- **Magával az XML-lel kapcsolatos specifikációk:**

- Az XML lehetőségeit bővítik.
- Lehetővé teszik XML dokumentumok szerkezetére és tartalmára vonatkozó megszorítások kifejezését (XML sémanyelvek).
- Lehetővé teszik XML dokumentumokból információ kinyerését (lekérdező nyelvek).
- Lehetővé teszi XML dokumentumok más formába alakítását (transzformációs nyelvek).

Az XML család (2)

- **Alkalmazások:** alkalmazási terület-specifikus XML formátumok
 - Digitális tartalomszolgáltatás (például Atom, DocBook, MathML, OSM XML, RSS, SVG, X3D, XHTML)
 - Kommunikáció (például XMPP)
 - Konfigurációs adatok tárolása (például Apache Maven, FXML)
 - Szemantikus web (például OWL, RDF, XMPP)
- **Alkalmazásprogramozási interfészek (API-k):** lehetővé teszik XML dokumentumok feldolgozását programnyelvekből (például DOM, JAXB, JAXP, JDOM, SAX, StAX)

Az XML lehetőségeit bővítő specifikációk

- *Associating Style Sheets with XML documents 1.0 (Second Edition)* (W3C ajánlás, 2010. október 28.)
<https://www.w3.org/TR/xml-styleSheet/>
- *Namespaces in XML 1.0 (Third Edition)* (W3C ajánlás, 2009. december 8.) <https://www.w3.org/TR/xml-names/>
- *XML Base (Second Edition)* (W3C ajánlás, 2009. január 28.)
<https://www.w3.org/TR/xmlbase/>
- *XML Inclusions (XInclude) Version 1.0 (Second Edition)* (W3C ajánlás, 2006. november 15.) <https://www.w3.org/TR/xinclude/>
- *XML Linking Language (XLink) Version 1.1* (W3C ajánlás, 2010. május 6.) <https://www.w3.org/TR/xlink11/>

XML sémanyelvek (1)

- XML dokumentumok szerkezetére és tartalmára vonatkozó megszorítások kifejezésére szolgálnak.
- Az XML sémák XML dokumentumok egy osztályát definiálják, melyeket **példányoknak** (*instances*) is neveznek.
- Adott sémá(k)nak megfelelő XML dokumentumot **érvényesnek** (*valid*) mondunk.
- A sémá(k)nak való megfelelés ellenőrzésének folyamata az **érvényesítés** (*validation*).

XML sémanyelvek (2)

Kortárs XML séma nyelvek:

- Dokumentumtípus-definíció (DTD): az XML specifikáció része
- W3C XML Schema <https://www.w3.org/XML/Schema>
- RELAX NG <https://relaxng.org/>
- Schematron <https://schematron.com/>

Lekérdező nyelvek

- *XML Path Language (XPath) Version 1.0* (W3C ajánlás, 1999. november 16.) <https://www.w3.org/TR/1999/REC-xpath-19991116>
- *XML Path Language (XPath) 3.1* (W3C ajánlás, 2017. március 21.) <https://www.w3.org/TR/xpath-31/>
- *XQuery 3.1: An XML Query Language* (W3C ajánlás, 2017. március 21.) <https://www.w3.org/TR/xquery-31/>

Transzformációs nyelvek

- *XSL Transformations (XSLT) Version 1.0* (W3C ajánlás, 1999. november 16.) <https://www.w3.org/TR/1999/REC-xslt-19991116>
- *XSL Transformations (XSLT) Version 3.0* (W3C ajánlás, 2017. június 8.) <https://www.w3.org/TR/xslt-30/>
- *XQuery 3.1: An XML Query Language* (W3C ajánlás, 2017. március 21.) <https://www.w3.org/TR/xquery-31/>

Invisible XML (1)

- Az *Invisible XML* (“láthatatlan XML”) (ixml) egy módszer nem-XML dokumentumok olyan módon történő kezelésére, mintha azok XML dokumentumok lennének.
 - Lehetővé teszi a szerzők számára, hogy dokumentumokat és adatállományokat az általuk preferált formátumban írjanak, miközben azokat XML-ként teszik elérhetővé olyan folyamatok számára, melyek az XML tartalmakat képesek hatékonyan kezelni.
 - Környezetfüggetlen nyelvekhez használható.
- Webhely: <https://invisiblexml.org/>
- Tároló: <https://github.com/invisibleXML/ixml>

Invisible XML (2)

- Specifikációk:
 - *Invisible XML Specification* (Community Group Editorial Draft, 20 August 2024) <https://invisiblexml.org/current/>
 - *Invisible XML Specification* (Final Community Group Report, 12 December 2023) <https://www.w3.org/community/reports/ixml/CG-FINAL-ixml-20231212/>
- Tutorial:
 - Steven Pemberton. [Invisible XML \(ixml\) Tutorial](#).
- Példák:
 - <https://github.com/invisibleXML/ixml/tree/master/samples>

Invisible XML (3)

Implementációk:

- Aparecium (licenc: GPLv3) <https://github.com/cmsmcq/Aparecium>
- CoffeePot (licenc: MIT License) <https://coffeepot.nineml.org/>
<https://github.com/nineml/coffeepot>
- Markup Blitz (licenc: Apache License 2.0)
<https://github.com/GuntherRademacher/markup-blitz>

Invisible XML (4)

- ixml támogatás bevezetése az XPath 4.0 and XQuery 4.0 szabványokban.
 - Lásd az `fn:invisible-xml` beépített függvényt.
- BaseX támogatás: https://docs.basex.org/main/Invisible_XML

Szerkesztők (1)

Szabad és nyílt forrású szoftverek:

- Visual Studio Code (platform: Linux, macOS, Windows; licenc: *MIT License*) <https://code.visualstudio.com/>
<https://github.com/Microsoft/vscode>
 - Javasolt kiterjesztés: XML <https://marketplace.visualstudio.com/items?itemName=redhat.vscode-xml>
<https://github.com/redhat-developer/vscode-xml>
 - Dokumentáció:
<https://github.com/redhat-developer/vscode-xml/tree/main/docs>

Szerkesztők (2)

Nem szabad szoftverek:

- `<oXygen/>` XML Editor (platform: Linux, macOS, Windows)
<https://www.oxygenxml.com/>
 - Lásd: https://www.oxygenxml.com/xml_editor/xml_editing.html
- IntelliJ IDEA (platform: Linux, macOS, Windows)
<https://www.jetbrains.com/idea/>
 - Lásd: <https://www.jetbrains.com/help/idea/working-with-xml.html>
- XMLSpy XML Editor (platform: Windows)
<https://www.altova.com/xmlspy-xml-editor>