XML Base, XML Infoset

dr. Paller Gábor

XML Base

- Minden dokumentumhoz tartozik egy URI
- Normálisan ez az az URI, ahol a dokumentum fellelhető (pl. webszerver, fájlrendszer ...)
- Lehetséges az u.n. Base URI felülírása az xml:base attribútummal.
- Példa: <doc xml:base=http://example.org/today/ ...>
- Öröklődési szabályok:
  - Egy elemnek van xml:base attribútuma, ekkor azt mondjuk, hogy az xml:base attribútum értéke a Base URI.
  - Az elem szülőjének Base URI-ja, (explicit vagy implicit)
  - A dokumentum Base URI-ja (mindig van, ha más nem, a fellelési hely)
- Relatív Base URI a szülőjéhez képest lehetséges.
- ```
  <doc xml:base="http://example.org/today/">
  ...
    <olist xml:base="/hotpicks/">
  ```
- "doc" base URI-ja: http://example.org/today/.
  "olist" base URI-ja: http://example.org/today/hotpicks/
  ```

XML Infoset

- Még mindig az XML alapspecifikáció része
- XML dokumentumok absztrakt (nem XML formátumú) reprezentációjával foglalkozik.
- Közvetlen haszna csekély, azonban erre épül az XML programnyelv-APIk belső adatábrázolása és az XML dokumentumok bináris kódolása is.
- Minden helyesen formált XML dokumentumhoz létrehozható Infoset, érvényesség nem feltétel.

Dokumentum (document)

- Children list: element, processing instruction, komment, abban a sorrendben, ahogy előfordultak. Ha van DTD deklaráció, az is ebben a listában van.
- XML verzió, karakterkódolás, standalone státusz (a dokumentum standalone státusza "yes", ha nem hivatkozik olyan deklarációra, amely megváltoztatja a dokumentum tartalmát. Pl. ha egy attribútumnak alapértelmezett értéke van és a dokumentumban az az attribútum nem szerepel, a nem validáló elemző az attribútumot nem fogja odaadni az alkalmazásnak (nem is tud róla), míg a validáló elemző tud az alapértelmezett attribútumról és az alapértelmezett értéket fogja odaadni az attribútum értékeként. Az ilyen dokumentum nem standalone. Ugyancsak hasolnló problémát okoznak az entitások.).
- Nem elemzett entitások és a hozzájuk tartozó NOTATION elemek.
- Jelzés arra vonatkozóan, hogy van DTD és az elemző feldolgozta. A DTD nem része az Infosetnek.

Elemek (element)

- Névterület név, lokális név, prefix.
- Children list: element, processing instruction, nem kibontott külső entitás hivatkozás, karakter és komment.
- Attribútumok rendezetlen listája.
- Névterület deklarációk (amelyeket ez az elem deklarál)
- Örökölt névterület-deklarációk (a szülőtől örökölve). Minden örökölt névterülethez egy Namespace Information Item tartozik, amelyik tartalmazza a névterület prefixét és a hozzá rendelt URI-t.
- Base URI
- Referencia a szülőelemre

Attribútum (attribute)

- Névterület, lokális név, prefix
- Attribútum értéke, entitás helyettesítések (u.n. normalizálás) után.
- Alapértelmezett érték jelzője. Jelzi, hogy az attribútum szerepelt-e a dokumentumban vagy attribútum alapértelmezett érték megadásból keletkezett a DTD-ből.
- Attribútum típusa (CDATA, felsorolás, ID, IDREF, IDREFS, ENTITY, ENTITIES, NMTOKEN, NMTOKENS, NOTATION)
- Referencia: ha az attribútum típusa IDREF, IDREFS, ENTITY, ENTITIES, vagy NOTATION, hivatkozik arra az element, unparsed entity, notation Infoset elemre, amely kell az attribútum értékének a használatához.
- Referencia a tulajdonos elemre.

Karakter (Char)

- Minden karakterhez, amely elem tartalmában fordul elő, tartozik egy.
- Karakter unicode értéke
- Jelző, hogy ez a karakter megőrzendő "fehér karakter"-e, amely egy elem tartalomban van. Példa: <elem> </elem> (a két tag között két szóköz van).
  Vegyünk két deklarációját az <elem>-nek
  <!ELEMENT elem EMPTY>
  vagy
  <!ELEMENT elem #PCDATA>
  Az első esetben a tag-ek között levő két szóköz kidobandó a parser által, az "értékes fehér karakter" jelzőjük false. A második esetben a két karakter része az elem tartalmának, tehát az "értékes fehér karakter" jelzőjük true.
- Referencia az elemre, amely tartalmazza a karaktert.

Processing instruction

- Példa az XML alapnyelv leckéből: <?xml-stylesheet href="mystyle.css" type="text/css"?>
- Célalkalmazás (target): esetünkben "xml-stylesheet"
- Tartalom (content): minden a célalkalmazás utántól a PI lezárásáig tehát esetünkben "href="mystyle.css" type="text/css".
- Base URI.
- Notation, ha a PI célalkalmazása hivatkozik egyre.
- Referencia az elemre, amely tartalmazza a PI-t.

Kifejtetlen entitás (unexpanded entity)

- Jelzi, hogy az elemző nem fejtett ki egy entitást.
- Entitás neve
- SYSTEM és PUBLIC referencia.
- Base URI.
- Referencia az elemre, amely tartalmazza az entitást.

Komment (comment)

- Komment szövege
- Referencia a kommentet tartalmazó elemre

Document Type Declaration

- SYSTEM és PUBLIC azonosító
- A DTD elemei sorrendben (nincs további feldolgozás az Infoset részéről)
- Referencia a Document elemre, amely tartalmazza a DTD-t.

Unparsed entity

- Ha emlékszünk, példa az XML alapnyelvből: <!ENTITY hatch-pic SYSTEM "../grafix/OpenHatch.gif" NDATA gif >
- Entitás neve (hatch-pic)
- SYSTEM és PUBLIC azonosító
- Base URI
- Notation neve és referencia a NOTATION elemre

Notation

- Ha emlékszünk: <!NOTATION gif SYSTEM "CompuServe Graphics Interchange Format 87a">
- NOTATION neve (gif)
- SYSTEM és PUBLIC azonosító
- Base URI

Gyakorlat

- Sorolja fel az Infoset elemeket az alábbi dokumentumban!
-
```
<?xml version="1.0"?>
<msg:message doc:date="19990421"

    xmlns:doc="http://doc.example.org/namespaces/doc"
              xmlns:msg="http://message.example.org/">
Phone home!
</msg:message>
```

Megoldás

- Egy darab document objektum
- Egy darab element objektum, neve message, prefix msg, namespace URI: http://message.example.org/
- Egy darab attribútum objektum, névterület azonosító: http://doc.example.org/namespaces/doc, lokális név: date, prefix: doc, érték: 19990421.
- Három névterület objektum: http://www.w3.org/XML/1998/namespace, http://doc.example.org/namespaces/doc és http://message.example.org/ névterületekhez.
- Két attribútum a két névterület-deklarációhoz.
- Két Namespace Information Item a doc és msg névterületekhez.
- 11 karakter objektum.