XML - perusteet

Ctl230: Luentokalvot 4.10.2004

Miro Lehtonen

Johdanto

Mikä on merkkauskieli ?

- Merkkaus (markup): lisätieto dokumentissa
  - Erilaiset kirjasintyylit ja -koot
- Säännöt merkkaukselle
  - Miten merkataan ?
  - Mitä merkataan ?
  - Mitä merkkaus tarkoittaa ?
- Esim. HTML

Käsitteitä 1

Tunnisteet (tag) ja elementit

- <p> on tunniste, jolla merkataan kappaleen alku
- on tunniste, jota seuraava teksti on kursivoitu
- on tunniste, johon kursivoitu teksti loppuu
- Elementti sisältää
  - alkutunnisteen (start tag),
  - elementin sisällön ja
  - lopputunnisteen (end tag).

Käsitteitä 2

Attribuutit

- Voidaan liittää mihin tahansa elementtiin
  - Nimi
  - Arvo
- <tunnistenimi attribuutti=”arvo”>
- Esim. HTML 4.0 <BODY>
  - Mahdollisia attribuutteja: class, id, dir, lang, style, background, link, ...

Extensible Markup Language (XML)

“Laajennettava merkkauskieli”

- Erilainen kuin HTML
  - Ei ennalta määriteltyjä tunnisteita
  - Elementin sisältöä voi kuvailla, ei vain esitystapaa
    - Tunnisteen nimi
    - Attribuuttien nimet ja arvot
  - Itsensä kuvaileva
- Esimerkki: lasku
  - HTML-muodossa
  - XML-dokumenttina

Lasku HTML-muodossa

Ulkoasun kuvaava dokumentti

```
<HTML>
<HEAD><TITLE>Lasku</TITLE></HEAD>
<BODY>
 <H3>Lasku: Miro Lehtonen</H3>
Tilauskoodi: ABC123456
 <TABLE>
  <TR>
   <TD valign=”top”>
    <H4>Laskutusosoite</H4>
    <UL>
     <LI>Miro Lehtonen</LI>
     <LI>Teollisuuskatu 23</LI>
     <LI>00014 Helsinki</LI>
    </UL>
   </TD>
   <TD valign=”top”>
    <H4>Toimitusosoite</H4>
     <UL>
      <LI>Miro Lehtonen</LI>
      <LI>Kotikatu 2</LI>
      <LI>00990 Helsinki</LI>
     </UL>
   </TD>
  </TR>
 </TABLE>

Tuote
 <UL>
  <LI><B>Tuotekuvaus</B> Laite 1A</LI>
  <LI><B>Tuotekoodi</B> 1A2A3A</LI>
  <LI><B>Määrä</B> 20</LI>
  <LI><B>Hinta</B> 10</LI>
 </UL>

</BODY>
</HTML>
```

Lasku XML-dokumenttina

Sisältönsä kuvaileva dokumentti

```
<?xml version=”1.0" ?>
<Lasku
        tilauskoodi=”ABC123456"
        asiakas=”Miro Lehtonen”
        laskutuskatuosoite=”Teollisuuskatu 23"
        laskuspostitoimipaikka=”00014 Helsinki”
        toimituskatuosoite=”Nilsiänkatu 3”
        toimituspostitoimipaikka=”00530 Helsinki”>
   <Ostos koodi=”1A2A3A” määrä=”20" kuvaus=”Laite 1A” hinta=”10"/>
</Lasku>
```

Formaattien vertailu

Mikä oikeastaan on itsensä kuvaileva ?

- HTML - <LI> <B>
  - Fyysinen rakenne
- XML - <Lasku> <Ostos>
  - Looginen rakenne
- Käyttötapauksia
  - Asiakas myyjän www-sivuilla
  - Jakeluverkosto seuraa tilausten toimitusta
  - Asiakaspalvelu tarkistaa palautukset
  - Markkinointiosasto selvittää tuotteiden menekkiä
- Useita alustoja, ohjelmointikieliä

XML-sanastot

Sittenkin ennalta määritellyt nimet ?

- XML-dokumentin tyyppi (DOCTYPE)
  - Rakenne, elementtien keskinäiset suhteet
  - Sanasto, tunnisteiden ja attribuuttien nimet
  - Esimerkissä “Lasku”
  - XHTML: HTML-sanasto, XML-kielioppi
- Tyypin määrittely
  - XML Document Type Definition (DTD)
  - XML Schema-kielet

Monikäyttöinen XML

Sovellusalueita

- XML-tietokannat
- Tekstin tallennusmuoto
  - Tekniset, juridiset dokumentit, määrämuotoinen teksti
  - Räätälöidyt dokumentit
  - Riippumaton julkaisukanavasta ja ulkoasusta
- Rajapinta ohjelmien välillä
  - Riippumaton ohjelmointikielistä ja sovellusalustoista

Historia

Yli 30 vuoden kehitys

- Ensimmäiset ideat 60- ja 70-luvuilla
- SGML: ISO-standardi 8879 vuodelta 1986
  - Osoittautui liian monimutkaiseksi
    - Vaikea tehdä työkaluja
    - Ohjelmistojen tuki vähäistä ja kallista
  - Silti yhä käytössä
- XML 1.0: W3C-suositus 10/2/1998
  - Yksinkertaistettu versio SGML:stä
- XML 1.1: W3C-suositus 4/2/2004

W3C:n sunnittelutavoitteet

Lähtökohdat vuonna 1996

- 1. XML shall be straightforwardly usable over the Internet.
- 2. XML shall support a wide variety of applications.
- 3. XML shall be compatible with SGML.
- 4. It shall be easy to write programs which process XML documents.
- 5. The number of optional features in XML is to be kept to the absolute minimum, ideally zero.
- 6. XML documents should be human-legible and reasonably clear.
- 7. The XML design should be prepared quickly.
- 8. The design of XML shall be formal and concise.
- 9. XML documents shall be easy to create.
- 10. Terseness in XML markup is of minimal importance.

XML-standardiperhe

Muita tärkeitä standardeja

- W3C-työryhmät
  - XML Schema
  - XPath
  - XML Stylesheet Language (XSL)
  - XSL Transformations
  - Document Object Model (DOM)
  - XML Query
- ISO/Unicode
  - Unicode-merkkistandardit, UTF-8, UTF16

Materiaali verkossa

Standardit, yhteisöt, tapahtumat, ohjelmistot, julkaisut jne.

- http://www.w3.org/XML/
- http://www.unicode.org/
- http://xml.coverpages.org/xml.html
- http://www.xml.org/
- http://www.xml.com/
- http://www.xmlsoftware.com/
- http://java.sun.com/xml/
- http://www.xml-finland.org/

Kirjallisuutta

WWW-materiaalin lisäksi

- Neil Bradley, The XML Companion, Addison-Wesley 2001.
- Pinnock, Hunter, ..., Beginning XML, 2nd Edition, Wrox 2001.
- Kirjoja kustantajien verkkokaupoista
  - Prentice-Hall, www.phptr.com
  - Addison-Wesley, www.awl.com/cseng
  - Wrox, www.wrox.com

Kurssin tavoitteet

XML:n perusasiat

- XML:n luku- ja kirjoitustaito
- DTD:n lukutaito
- XML-dokumenttien luominen DTD:n pohjalta
- DTD:n laatiminen
- Dokumenttityypin suunnittelu

XML-merkkaus

Elementit, attribuutit ja puurakenne

Perusrakenteet

Mistä XML koostuu ?

- Esittely (XML Declaration)
- Elementit
- Attribuutit
- Merkkidata (CDATA)
- Viittaukset entiteetteihin
- Prosessointiohjeet
- Kommentit

XML:n esittely XML-dokumentissa

XML-dokumentti saattaa alkaa näin

- <?xml version=”1.0"?> tai
- <?xml version=”1.0" encoding=”iso-8859-1"?>
- Suositeltava muttei pakollinen
- Vain dokumentin alussa
- Myös **declaration** tai **prolog**

Elementit 1

Tärkein osa XML-merkkausta

- Tunnisteet
  - Alkutunniste pakollinen
  - Elementin loppuminen on myös merkattava
    - Lopputunniste
  - Alku- ja lopputunnisteiden pitää olla samat
- Sisältö
  - Tekstiä
  - Elementtejä
  - Sekä tekstiä että elementtejä
  - Ei välttämättä mitään (tyhjä elementti)

Elementtiesimerkkejä

```
<paikka>
  <kaupunki>Helsinki</kaupunki>
  <maa>Suomi</maa>
</paikka>

<uutinen>
  Tapahtumapaikka oli <paikannimi>Helsinki</paikannimi>.
  Julkaistu <pvm/>.
</uutinen>
```

Huomaa tyhjä elementti <pvm/>

Elementit 2

Lukumäärä, dokumentin juuri

- Dokumentissa oltava vähintään yksi elementti
  - Ei ylärajaa
- Juurielementti (root element, document element)
  - Ensimmäinen, uloin, ylimmän tason elementti
  - Dokumentissa vain yksi juuri

```
<EsimerkkiDokumentti>
   Tämä on yksinkertainen XML-dokumentti.
</EsimerkkiDokumentti>
```

Elementit 3

Elementeillä XML-nimet

- Elementin nimen alku
  - Mikä tahansa kirjain
  - Alaviiva (_), kaksoispiste (:)
  - Ei kuitenkaan XML, xml, Xml, xMl, jne.
- Nimen loppuosa
  - Kirjain-numero-yhdistelmä, myös _ - : . käy
  - Ei välilyöntejä eikä rivinvaihtoja

Elementit 4

Sisäkkäisyys

- Kaikki paitsi juurielementti sisältyvät toiseen elementtiin
  - Alku- ja lopputunnisteiden järjestys
- Elementtihierarkia

Oikein:

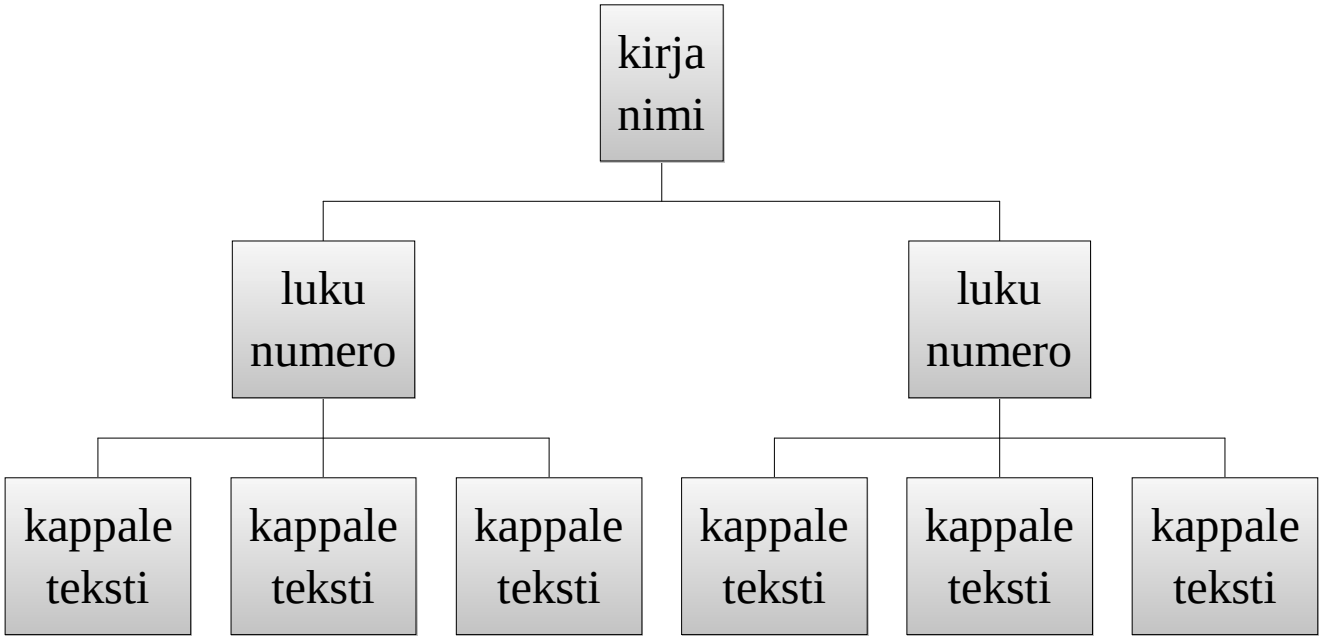
```
<p>
  Tekstiä... <i>kursiivi ja
  <b>lihavointi</b></i>
</p>
```

Väärin:

```
<p>
  Tekstiä... <i>kursiivi ja
  <b>lihavointi</i></b>
</p>
```

Dokumenttipuuesimerkki

Miltä näyttää vastaava XML-dokumentti ?



Puusanasto 1

Erityyppiset solmut XML-dokumenttipuussa

- Elementtisolmut (element node)
  - Juuri (root node), joskus myös dokumenttisolmu (document node)
  - Ei-lehtisolmut (non-leaf node)
  - Lehtisolmut (leaf node)
- Tekstisolmut (text node)
  - Vain elementtisolmuilla
  - Aina lehtisolmuja
- Lisäksi mm. attribuutit, kommentit, entiteetit, prosessointiohjeet

Puusanasto 2

Solmujen väliset suhteet

- Lapsisolmut (child node), äiti-, ja isäsolmut (parent node)
- Jälkeläissolmut (descendant node), esi-isät (ancestor node)
- Sisarsolmut (sibling node)

Puusanasto 3

Solmujen ominaisuudet

- Tyyppi: esim. elementti
- Nimi: esim. elementillä tunnisteen nimi
- Merkkijonoarvo
  - Tekstisolmulla itse teksti
  - Elementtisolmulla lapsisolmujen teksti yhdistettynä
- Attribuutteja vain elementtisolmuilla

Attribuutit 1

Lisätietoa elementteihin

- Nimi-arvo-pareja kuten HTML-attribuutit
- Vain yksi arvo yhdessä elementissä
- Järjestyksellä ei ole väliä
  - Ei omaa paikkaa dokumenttipuussa
- Arvo “lainaus-” tai ‘heittomerkkien’ sisällä

Väärin:

```
<dokumentti kirjoittaja=”A.  B.” kirjoittaja=”C. D.”>
  XML-dokumentti.
</dokumentti>
```

Attribuutit 2

Varatut nimet: kielen määrittely

- xml:lang
  - Dokumentin tai elementin kieli
  - Usein ISO639 tai IANA-numerokoodi
  - Voidaan myös määritellä itse

```
<p xml:lang="en">The quick brown fox jumps over the lazy dog.</p>
<p xml:lang="en-GB">What colour is it?</p>
<p xml:lang="en-US">What color is it?</p>
<sp who="Faust" desc='  leise'   xml:lang="de">
  <l>Habe nun, ach! Philosophie,</l>
  <l>Juristerei, und Medizin</l>
  <l>und leider auch Theologie</l>
  <l>durchaus studiert mit heißem Bemüh'  n.</l>
</sp>
```

Attribuutit 3

Varatut nimet: säilytetäänkö “white space”?

- xml:space
  - Arvona “default” tai “preserve”

```
<p>Taulukko:
   <table xml:space=”preserve”>
eka rivi          toinen sarake
toisella rivilllä vain yksi sarake
kolmas rivi       toka sarake
   </table>
</p>
```

Esimerkki

Elementit, attribuutit, solmut, sukulaissuhteet, tyypit ja arvot

```
<e1>
   <e2 a1=”x”/>
   <e3 a2=”y”> Tekstin alku.
      <e4>
         <e2 a1=”z”/>
         <e3 a2=”o”  a4=”p”>
            <e2 a1=”y”/> Sisältöä.
         </e3>
      </e4>
      <e2 a1=”w”/> Tekstin loppu.
   </e3>
   <e2>
      <e3>Tekstiä</e3>
      <e4 a4=”q”/>
   </e2>
</e1>
```

- Solmu e1
  - Jälkeläiset ?
  - Lapset ?
- Solmu e3
  - Lapsisolmut ?
  - Jälkeläiset ?
  - Solmujen tyypit ?
  - Arvot ?