Luento 1: Johdanto merkintäkieliin

AS-0.110 XML-kuvauskielten perusteet

Janne Kalliola

Johdanto merkintäkieliin

- Merkintäkieliä
  - SGML
  - HTML
  - XML
- XML:n peruspiirteet
- XML-dokumentin rakenne
- XML:n käyttö
- XML-pohjaisia kieliä
- Merkintäkielten yleiset käyttötavat

Merkintäkieli

- Merkintäkieli on tekstipohjainen kieli, jolla voidaan kuvata rikasta joukkoa tietoa hyvin yksinkertaisen säännöstön pohjalta
  - kyseessä on siis eräänlainen hybridi teksti- ja binaaritiedostomuotojen väliltä
- Yleensä merkintäkielen avulla luodaan tiedostoon haluttu rakenne
  - koneet tulkitsevat tätä rakennetta ennalta määrättyjen sääntöjen avulla

Standard Generalized Markup Language

- SGML on kehitetty aikanaan Yhdysvaltain puolustusministeriön tarpeisiin
  - ongelmana useiden eri toimittajien erilaiset dokumentointikäytännöt
  - SGML mahdollisti yhteisen dokumentointirakenteen ja –kieliopin käytön
- SGML on ensimmäinen laajasti käytetty rakenteinen (structural) dokumentoinnin kieli
  - aiemmat ratkaisut olivat sovelluskohtaisia
- SGML on monimutkainen ja raskas kieli

Hypertext Markup Language

- HTML on suunniteltu yhdessä muiden WWW-standardien kanssa CERN:ssä 90-luvun alussa
- Alkuperäisenä tarkoituksena oli sisältötiedon esittäminen tietoverkoissa
  - WWW:n kaupallistuessa kielen ilmaisuvoimaa on laajennettu useaan otteeseen
  - nykyään hybridikieli, jossa sekä sisällön että ulkonäön kuvaamiseen liittyviä elementtejä
  - kielen esitysvoima on rajoitettu ja siinä on ennalta määritelty kielioppi

HTML:n ongelmat

- HTML on ollut alunperin sangen huonosti standardoitu
  - Netscape ja Microsoft ovat lisänneet suuren joukon alkuperäisestä standardista uupuvia elementtejä
    - selaimet eivät ole olleet yhteensopivia
  - selaimet sallivat hyvinkin vajavaisia HTML-dokumentteja
    - dokumenttien laatijat eivät ole kovin tarkkoja dokumentin oikeellisuudesta
    - kunnollisia testaustyökaluja on vähän
    - ei ole syntynyt selkeää prosessia dokumentin käsittelyyn

Extensible Markup Language

- Extensible Markup Language (XML) on World Wide Web Consortiumin (W3C) suositus elektronisen tiedon esitysmuodoksi
  - XML määrittelee ainoastaan tavan esittää tietoa
  - se ei ota kantaa itse esitettävään tietoon eikä myöskään määrittele primitiivijoukkoa, jolla tietoa pitäisi kuvata
- Syntynyt paikkaamaan aukkoa HTML:n ja SGML:n välissä
- XML on metakieli
  - XML:lla voidaan määritellä uusia kieliä, joilla vasta kuvataan itse tietoa
  - puhuttaessa XML-dokumentista yms. tarkoitetaan dokumenttia, joka on XML-määritysten mukainen

XML-määritys

- XML-määritys valmistui vuoden 1998 alussa
  - standardi on suhteellisen nuori verrattuna useisiin muihin Internetissä käytettäviin tiedostomuotoihin
  - siihen ei ole kuitenkaan tarvittu tehdä muutoksia
    - aihetta kypsytelty kauan
  - laajassa käytössä
- XML pohjautuu SGML-kieleen, mutta on sitä yksinkertaisempi ja tiukempi
  - SGML:a käytetään edelleen laajasti teknisessä dokumentoinnissa
  - XML on kuitenkin ajanut SGML:n ohi useilla rintamilla

XML:n peruspiirteet

- XML on rakenteinen kieli
  - dokumentti koostuu elementeistä
  - elementit voivat olla sisäkkäisiä
    - sisäkkäisyys ja peräkkäisyys määräävät dokumentin rakenteen
- XML on metakieli
  - se ei määrittele käytettävissä olevia elementtejä valmiiksi
    - tämä on dokumentin laatijan tehtävä
    - puhutaan XML-kieliopista

XML:n tavoitteet

- Mahdollisuus käyttää dokumentteja Internetissä
- Yhteensopivuus SGML:n kanssa
- Rakenne on formaali ja tiivis
- Vapaavalintaisten ominaisuuksien määrä on minimoitu
- Laaja ohjelmistotuki
- XML:ää käyttävien sovellusten ohjelmointi on helppoa
- Dokumentit ovat lukukelpoisia myös ihmisille
- Dokumentin suunnittelu on nopeaa
- Dokumenttien laadinta on helppoa
- XML:n ei tarvitse olla ytimekästä

Tavoitteiden täyttyminen (1/2)

- XML pohjautuu kokonaan tekstiin
  - lukukelpoista ihmisille
  - dokumenttien laadinta on helppoa
  - helpottaa dokumenttien jakoa Internetissä
  - ei välttämättä ytimekästä
- XML pohjautuu SGML:ään
  - rakenne on formaali
  - melko hyvä yhteensopivuus SGML:n kanssa
    - osa SGML:n ominaisuuksista poistettu
  - laaja ohjelmistotuki

Tavoitteiden täyttyminen (2/2)

- XML:n rakenne on yksinkertainen
  - rakenne on tiivis
  - vapaavalinnaisten ominaisuuksien määrä on minimoitu
  - koneellinen tulkinta on helppoa
  - dokumenttien suunnittelu ja laadinta helppoa

HTML ja XML

- HTML ja XML eivät ole täysin yhteensopivia
  - johtuu XML:n SGML-syntaksista
  - erot lähinnä tyhjissä elementeissä
    - tähän palataan myöhemmin
  - HTML on joustavampaa kuin XML
    - selaimet antavat enemmän anteeksi
- HTML on määritelty uudelleen XML-syntaksin avulla
  - XHTML 1.0 pohjautuu HTML 4.0 -suositukseen
  - seuraavat XHTML-versiot eroavat HTML:sta
  - HTML-suosituksia ei enää kehitetä edelleen

XML-dokumentin rakenne (1/2)

- XML-dokumentti rakentuu peräkkäisistä ja sisäkkäisistä elementeistä
  - elementit merkitään kulmasuluilla: <elementti>...</elementti>
  - elementillä on alku- ja loppumerkintä (start & end tag)
    - elementti voi olla myös tyhjä, jolloin se voidaan merkitä lyhennetysti: <elementti/>
- Elementin alkumerkintään voidaan lisätä attribuutteja (attribute)
  - attribuutti on avain-arvopari, joka täsmentää elementtiä
  - attribuutti merkitään avain="arvo" tai avain='arvo'
    - lainaus- tai heittomerkit ovat pakolliset
  - yhteen elementtiin voi liittää useita attribuutteja
    - attribuuttien nimien, ts. avaimien, täytyy olla yksilöitävissä

XML-dokumentin rakenne (2/2)

- Elementtien välissä voi olla leipätekstiä tai toisia elementtejä
- XML-dokumentin täytyy kokonaisuudessaan olla yhden elementin (juurielementti, root element) sisällä
- Dokumentissa voi olla kommentteja merkkien <!-- ja --> sisällä
  - < ja & eivät saa esiintyä yksinään dokumentissa
  - muut entiteetit täytyy itse määritellä

Esimerkki

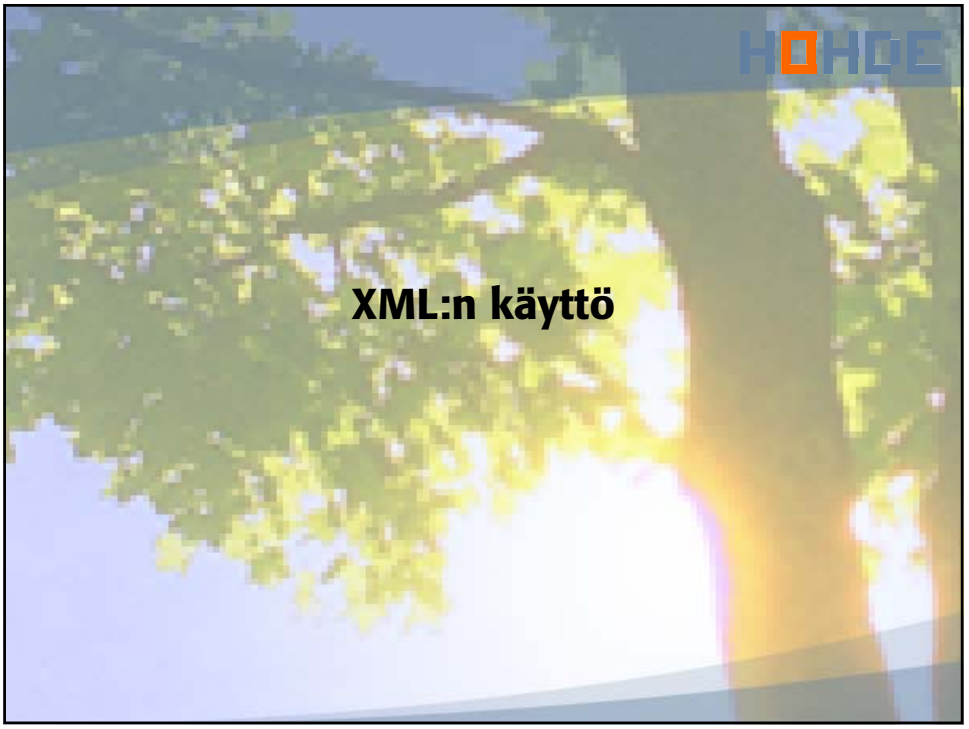
```
<article>
  <meta author="Janne Kalliola"
   date="Oct 5, 1999"/>
  <title>XML example</title>
  <ingress>
      <p>Extensible Markup Language...</p>
  </ingress>
  <paragraphs>
     <p>HTML has reached the end of the road...</p>
     <p>...</p>
     <image href="http://www.hohde.com/img/xml.jpg"
      text="Relationship between XML and HTML."/>
     <p>...</p>
  </paragraphs>
</article>
```

Harvinaisemmat rakenteet

- XML-dokumentti voi sisältää käsittelyohjeita tietylle sovellukselle
  - käsittelyohjeet merkitään merkkien <? ja ?> sisään
  - ohjelman nimi ei saa olla 'xml', joka on varattu XML-standardin sisäiseen käyttöön
- Erikoismerkit voidaan korvata entiteeteillä
  - merkitään &entity;, esimerkiksi > on suurempi kuin -merkki
- Joskus XML-dokumentin sisään täytyy sisällyttää koodia, jossa esiintyy < tai &
  - tällöin koodisirpale sijoitetaan CDATA-lohkon sisään:
  - <![CDATA[<element><subelement/></element>]]>

Rakenteet tarkemmin

- XML:n rakenteet käydään läpi tarkemmin seuraavalla luennolla
  - luennolla käsitellään myös nimiavaruudet
- Samalla pureudutaan dokumenttien kielioppien formaaliin määrittelyyn



XML:n käyttö, perusperiaate

- XML-dokumenttien käyttö on yleensä kaksivaiheista:
  - XML-prosessori lukee XML-dokumentin jostakin lähteestä
  - sovellus saa XML-prosessorilta XML-dokumentin sisällön käyttöönsä
- Näin on pyritty helpottamaan sovellusten laatijoiden työtaakkaa ja samalla pitämään huoli, että XML-dokumentteja käsitellään samalla tavalla sovelluksesta toiseen
  - melkein kaikkiin ohjelmointikieliin löytyy yksi tai useampia rajapintoja XML-dokumenttien käsittelyyn
  - näiden rajapintojen takana on yleensä XML-parseri (tai -jäsennin), joka lukee ja muokkaa XML-dokumentin johonkin ohjelmointikielelle sopivaan muotoon
- XML ei ota kantaa, mitä sovellus dokumentilla tekee

XML:n käyttö

- XML:a voidaan käyttää
  - perustuen valmiiseen kielioppiin, esimerkiksi MathML tai XSLT
  - käyttäen omaa sovelluskohtaista kielioppia
- Sovellus voi olla myös valmis XML-sovellus
  - esimerkiksi XSLT-prosessori lukee sisään sekä XSLT-dokumentin (eräänlainen sovellus) että sovelluskohtaisen XML-dokumentin
- XML:n käyttöön tutustutaan tarkemmin seuraavilla luennoilla

XML-pohjaiset kielet

XML-pohjaiset kielet

- XML-pohjaisia kieliä on syntynyt muutamassa vuodessa runsaasti
  - XML on osoittautunut järkeväksi tavaksi kuvata erilaista informaatiota, jota on ennen kuvattu erikoistuneella tiedostomuodolla
- XML-pohjainen kieli
  - käyttää XML:n syntaksia ja merkintöjä
  - määritellään Schemana, DTD:na tai näiden kokoelmana
    - näihin palataan seuraavassa luennossa
  - XML:n säännöt ovat voimassa
  - Vaatii erillisiä ohjelmia oikeata tulkintaa tai käyttöä varten

MathML

- Mathematical Markup Language
  - XML:n syntaksia käyttävä kieli, jolla kuvataan matemaattisia merkintöjä
  - tarkoituksena mahdollistaa matemaattisten kaavojen jako, vastaanotto ja käsittely tietoverkoissa
  - vaatii erillisen sovelluksen kaavojen visualisoimiseksi
    - tuki on saatavissa Mozillaan erillisenä komponenttina
- Kieli on standardoitu
  - W3C:n suositus MathML 2.0
  - kieltä kehittää W3C Math Working Group

MathML:n suunnitteluperiaatteet

- Kieli kuvaa matemaattisia merkintöjä kaikilla tasoilla
  - sopii opettamiseen
  - sopii tieteelliseen viestintään
- Kieli kuvaa sekä matemaattisia notaatioita että matemaattista merkitystä
- Kieli on laajennettavissa
- Kieli sallii tietyille ohjelmille suunnatut ohjeet
- Kieli on ihmisluettavaa
- Kieli mahdollistaa muunnoksia muihin muotoihin
  - graafiset näytöt, laskentajärjestelmät, puhesyntetisaattorit
  - muut matemaattiset kielet, kuten TeX
  - printtimedia
- Kieli on tehokas pitkissäkin lausekkeissa

Synchronized Multimedia Integration Language

- XML-pohjainen kieli vuorovaikutteisten multimediaesitysten luontiin
- Koostuu moduleista
  - jokainen moduli määrittää yhdellä osa-alueella käytettävät merkinnät
- W3C:n standardi
  - SMIL 2.0
- Käytössä esimerkiksi MMS-viestien kuvaamisessa

Scalable Vector Graphics

- Scalable Vector Graphics (SVG) on XML-pohjainen kieli vektorigrafiikan kuvaamiseen
  - vektorigrafiikassa piirtoprimitiivit eivät pohjaudu pikseleihin, vaan erilaisiin matemaattisiin/geometrisiin muotoihin
    - muodoille asetetaan tiettyjä arvoja (väri, koko, sijainti, yms.), joiden avulla se pystytään esittämään ja tulostamaan
  - ajatusmalliltaan muistuttaa pitkälti XML-dokumenttia
- SVG määrittelee elementit erilaisten primitiivien määrittelyyn
  - attribuuteilla ohjataan primitiivien esitystä
- SVG-dokumentit esitetään erillisellä SVG-ohjelmalla tai SVG-tuki on toteutettu esitysohjelmaan valmiiksi
- SVG:ssa on mukana myös animaatio-ominaisuudet

Resource Description Framework

- Resource Description Framework (RDF) on kieli metatiedon kuvaamiseksi
  - metatieto on tietoa tiedosta
    - määrittää yhtä tai useampaa dokumenttia
- RDF on vain pelkkä viitekehys (framework), jonka avulla voidaan laatia metatietodokumentteja
  - RDF määrittelee ylätason konseptit ja niitä vastaavat XML-kieliopit
  - RDF ei ota kantaa, minkälaista tietoa kuvataan ja kuinka
    - tiedon kuvausta varten määritellään metatietosanasto
    - RDF:ssa sanastoa kutsutaan RDF Schemaksi

RDF Site Summary

- RDF Site Summary (RSS) on kevyt kuvaus- ja syndikointiformaatti
  - RSS-dokumentti kuvaa esimerkiksi tietyllä sivustolla tapahtuneet muutokset tai siellä olevat uutisotsikot
  - RSS-dokumentti voidaan sisällyttää toisiin sivustoihin helposti esimerkiksi XSLT:n avulla
- RSS-syndikointia käytetään esimerkiksi blogien otsikoiden julkaisuun ja siirtoon toisiin blogeihin
- RSS ei ole W3C:n määrittelemä standardi
  - standardi on määritelty osoitteessa *web.resource.org/rss/1.0*

Merkintäkielten yleiset käyttötavat

Merkintäkielet dokumentoinnissa

- Perinteisin – ja helpoiten ymmärrettävä – merkintäkielten käyttötapa on tekstipohjaisten dokumenttien kuvaaminen
- Kielen rakenne sovitetaan vastaamaan dokumenttien tai dokumentointijärjestelmän (tai paremminkin näitä käyttävien tahojen) tarpeita
- Dokumentit laaditaan joko erityisellä ohjelmalla tai sitten yleiskäyttöisellä editorilla
  - dokumenttien jatkokäsittelyyn on olemassa suuri joukko standardeja ja sovelluksia

Kielet tiedonvälityksessä

- Tällä hetkellä voimakkain kehitys tapahtuu koneiden välisessä tiedonvälityksessä
  - perinteisesti koneiden välinen keskustelu on vaatinut erityisten protokollien (yhteyskäytäntöjen) laatimista ja toteuttamista
  - nykyään kommunikointi voidaan suorittaa jotakin standardiprotokollaa (HTTP) käyttäen
    - HTTP ei kuitenkaan tarjoa vastausta tiedon koodaamiseen
    - XML hyvä ratkaisu

Tietorakenteet

- Rakenteisilla kielillä pystytään kuvaamaan – ainakin teoriassa ja joskus hyvinkin tehottomasti – mikä tahansa tietorakenne
  - tietorakenne voidaan esimerkiksi tallentaa tiedostoon ja ladata sieltä myöhemmin käyttöön
    - lataus voi tapahtua toisessa käyttöjärjestelmässä kuin tallennus, myös sovellus voi vaihtua
- Rakenteinen dokumentti voidaan muuntaa aina puuksi (tai metsäksi, jos kieli sallii useat juurielementit)
  - puun avulla on helppo kuvata esimerkiksi lista tai pino ja viittausten avulla myös verkko

Vuot

- Mikäli rakenteista kieltä luodaan ja kulutetaan koneellisesti, ko. dokumentti voi olla käytännössä loputon
  - dokumentti virtaa luontipisteestä kulutuspisteeseen jonkin sopivan protokollan avulla ja siitä on olemassa kerralla ainoastaan hyvin pieni osa
  - dokumenttia ei siis koskaan luoda kokonaiseksi
- Tällainen ratkaisu asettaa tiettyjä rajoituksia dokumentin rakenteelle
  - vältettävä esimerkiksi dokumentin jatkuva syveneminen


Kysymyksiä? Kommentteja?