# XML johdatus: DTD

Jaana Holvikivi

# Dokumenttityypin rakennemäärittely

- DTD = kielioppi esim. XML- esitykselle
- Elementit
- Attribuutit
- Entiteetit ja notaatiot
- Prosessointikomennot
- DTD:n suunnittelu

# DTD

```
<!-- Dokumentin tyyppijulistus (DTD) esimerkki (osa) -->

<!ELEMENT university (department+)>
<!ELEMENT department (name, address)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT address (#PCDATA)>
```

- Dokumentin tyyppijulistus eli rakenne-esittely
- yksi sääntö/elementti
  - nimi
  - sisältö
- dokumentti-ilmentymien "kielioppi"

# DTD:n käyttö

- validoiva jäsennin
  - tarkistaa että dokumentti on DTD:n mukainen
- tunnisteiden johdonmukainen käyttö
- eri sovellusten DTD-standardit
  - yhteinen sanasto

# Hyvinmuodostuneisuus

- XML-dokumentti on **hyvinmuodostettu** (well-formed) jos se
  - sisältää täsmälleen yhden juurielementin ja muut mahdolliset elementit alkavat ja päättyvät saman elementin sisällä eli elementit ovat tasapainossa,
  - vastaa XML-määrittelyn asettamia hyvinmuodostetun dokumentin rajoituksia,
  - ja sen jokainen jäsennetty entiteetti on hyvinmuodostettu

# Validius

- XML-dokumentti on **validi** jos
  - dokumenttiin kuuluu DTD tai skeema
  - dokumentti on hyvinmuodostettu
  - dokumentti noudattaa DTD-määrittelyä
- Validius voidaan tarkistaa validoivalla jäsentimellä
  - dokumentti kokonaisuudessaan ("erä")
  - interaktiivinen

# Dokumenttityypin esittely

DTD ja ilmentymä eri tiedostoissa, rakennemäärittelyn ulkoinen osajoukko, ilmentymään tulee

```
<!DOCTYPE catalog SYSTEM "catalog.dtd">
```

DTD ja ilmentymä samassa tiedostossa:
rakennemäärittelyn sisäinen osajoukko

```
<!DOCTYPE catalog [
... catalog-DTD tähän kohtaan
]>
```

yhdistelmä mahdollinen
sekä sisäinen että ulkoinen rakennemärittelyn osajoukko

```
<!DOCTYPE catalog SYSTEM "catalog.dtd" [
   <!ENTITY % paramodel "#PCDATA | SUB | SUP">
   ...
]>
```

# Dokumenttityypin esittely 2

DTD ja ilmentymä eri tiedostoissa, PUBLIC julkinen standardi ilmentymään tulee

<!DOCTYPE catalog PUBLIC "-//ORG_NAME//DTD CATALOG//EN">

- lippu(-/+) kertoo, ettei standardi ole kovin merkittävä

ISO –standardit alkavat "ISO...

| | |
|---|---|
| ORG_NAME | kertoo DTD:n omistajan |
| DTD | tiedoston tyyppi |
| CATALOG | dokumentin nimi |
| EN | kielikoodi |

tyyppimäärittely voi sijaita XML-prosessorin omassa sisäisessä DTD-tietokannassa

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN">

# Ulkoinen tai sisäinen DTD

- yleinen muoto:

<! DOCTYPE document_element source location [internal subset of DTD] >

- esim. <!DOCTYPE Dictionary SYSTEM ”dictionary.dtd”>

- Jaettu dokumenttityyppi:

  <!DOCTYPE Dictionary PUBLIC
  ”http://www.evtek.fi/DTD/dictionary.dtd”>
- sisäinen:
  <!DOCTYPE MyMessage SYSTEM
  [!ELEMENT MyMessage (#PCDATA)]>
- sisäinen ja ulkoinen:
  <!DOCTYPE MyMessage SYSTEM ”myDTD.dtd”
  [!ELEMENT MyElement (#PCDATA)]>

# Elementtityypin esittely

- **<! ELEMENT country (capital)>**
- elementin nimi
- elementin sisältö
  - sisällön esittely (content declaration), myös sisältömalli (content model)
- erotinmerkit (**<!**, **>**, **(**, **)**) ja avainsana (**ELEMENT**)

# Lapsielementit

- peräkkäiset lapsielementit:
  - <!ELEMENT country (cname, capital, population)>
- vaihtoehtoiset lapsielementit:
  - <!ELEMENT country (cname | official_name)>
- valinnaiset elementit:
  - <!ELEMENT country (cname, capital, population?)>

# Toisto

* toisto: nolla tai useampi, valinnainen

    <!ELEMENT country (cname, capital, city*)>

+ toisto: yksi tai useampi, pakollinen

    <!ELEMENT country (cname, neighbour_country+)>

?  Kertaesiintymä, valinnainen

[none]  kertaesiintymä, pakollinen

Toistuva ryhmä:

    <!ELEMENT country (cname, (city, city_population)*)>

# Tietosisältö

- Dataa

  <!ELEMENT cname (#PCDATA)>

  - "parsed character data"
- Elementtejä
  - alielementtejä (lapsielementtejä)
- Yhdistelmä (mixed content)
  - sekä dataa että elementtejä
  - <!ELEMENT para (#PCDATA | sub | super)*>
  - #PCDATAn on tultava ensin, ja on oltava toistuva ryhmä jossa on vaihtoehtoja

# Tyhjä elementti ja ANY

- <!ELEMENT image EMPTY>
  - ilmentymässä p.o. <image/>
  - ei saa olla <image></image>
- jos tavallinen esittely <!ELEMENT im (...)>
  - ilmentymässä joko <im/>, <im></im> tai <im>...</im>
- <!ELEMENT some ANY>
  - sallii elementin sisällöksi minkä tahansa elementin, joka on esitelty
  - joustava

# Lyhyt esimerkki: Sanakirja

```
<!ELEMENT dictionary (word_article)*>
<!ELEMENT word_article (head_word, pronunciation,
   sense+)>
<!ELEMENT head_word (#PCDATA)>
<!ELEMENT pronunciation (#PCDATA)>
<!ELEMENT sense (definition, example*)>
<!ELEMENT definition (#PCDATA)>
<!ELEMENT example (#PCDATA)>
```

# Sanakirjan XML

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE dictionary SYSTEM "dict.dtd">
<dictionary>
  <word_article>
    <head_word>
      carry
    </head_word>
    <pronunciation>
      kaeri
    </pronunciation>
```

```
<sense>
    <definition>
      support the weight of and move from place to place
    </definition>
    <example>
     Railways and ships carry goods.
    </example>
    <example>
      He carried the news to everyone.
    </example>
</sense>
<sense>
    <definition>
      wear, possess
     </definition>
    <example>
     I never carry much money with me.
    </example>
 </sense>
</word_article>
```

```
<word_article>
    <head_word>gossamer
    </head_word>
    <pronunciation>
     gosomo
    </pronunciation>
    <sense>
      <definition>fine, silky substance of webs made by
       small spiders
      </definition>
    </sense>
  </word_article>
</dictionary>
```

# Sisältömallin moniselitteisyys

- Vältä moniselitteisyyttä, esim.
  - Väärin: (item?, item)
  - Oikein: (item, item?)
  - Väärin ((surname, employee) | (surname, customer))
  - Oikein: (surname, (employee | customer))
- Sisältömalli joka sisältää sekä #PCDATA ja elementtejä
  - rajoitukset estämään moniselitteisyyttä

# Attribuuttiesittelyt

Jokaiselle elementille voidaan esitellä (yksi tai useampi) attribuuttilista

```
<!ATTLIST country
        population NMTOKEN #IMPLIED
        language   CDATA #REQUIRED
        continent  (Europe | America | Asia ) "Europe">
```

- CDATA
  - merkkidataa, mikä tahansa teksti
- Lueteltu tyyppi
  - <!ATTLIST country

    continent  (Europe | America | Asia ) "Europe">
  - voi saada ainoastaan luetellun arvon, myös oletusarvo

# Attribuutin oletusarvot

- #REQUIRED
  - attribuutille on annettava arvo
- #IMPLIED
  - attribuutti voi puuttua
- jos tyyppi on lueteltu, voidaan antaa tietty arvo
- implied/required-attribuuteille ei voi antaa tiettyä arvoa
- Esimerkkejä:

```
<!ATTLIST catalog type CDATA #REQUIRED>

<!ATTLIST catalog type NMTOKEN #IMPLIED>

<!ATTLIST catalog type (phone | e-mail)>

<!ATTLIST catalog type (phone | e-mail) "phone">
```

# Attribuuttiesittelyt

- NMTOKEN
  - name token <country population = "100">
- NMTOKENS
  - name token -luettelo välilyönnillä erotettuina
  - välittävät tietoa prosessoreille/ sovelluksille. Määrittelevät validin nimen

<!ATTLIST DATA AUTHORIZED_USERS NMTOKENS #IMPLIED>

<DATA SECURITY="ON"

AUTHORIZED_USERS = "IggieeB SelenaS GuntherB">

element content

</DATA>

# Attribuuttiarvojen tyypit

- ENTITY
  - <!ENTITY mypicture "123.jpg">
  - <!ELEMENT pic EMPTY>
  - <!ATTLIST pic picfile ENTITY ...>
  - ilmentymässä: <pic picfile="mypicture">
- ENTITIES
  - monta entiteettiarvoa
- ID
  - Tunnus, yksikäsitteinen XML-nimi
- IDREF
  - Viittaus ID attribuutilla määriteltyyn elementtiin
- IDREFS
  - useampia viittauksia

# Attribuutin vakioarvot ym.

- <!ATTLIST country position #FIXED "independent">
- yksi mahdollinen arvo (ainoastaan)
- miksi
  - voidaan kiinnittää tietty arvo tiettyä sovellusta varten
  - voidaan antaa tietylle osalle tunnusarvo

- Elementille voi olla useampi esitelty attribuuttilista
  - jos listoissa samannimisiä attribuutteja, on ensiksi esitelty voimassa
- attribuuttien arvot: pienet kirjaimet eroavat isoista
- varatut attribuutit, esim. xml:lang ja xml:space
  - etuliite 'xml:'

# Elementti vai attribuutti

- Milloin jotain pitäisi merkata elementillä, milloin attribuutilla?
- Elementillä
  - jos on alirakenteita
  - jos se näkyy tulostuksessa
  - mutta sisältöä ei voida rajoittaa samalla tavalla kuin attribuutin yhteydessä
- Attribuutti
  - jos ei ole alirakenteita
  - jos on kysymyksessä "hallinnollisesta" tiedosta joka ei näy tulostuksessa
  - voidaan rajoittaa esim. arvon tyypillä, oletusarvoilla jne.

# Entiteetin esittely

- XML-dokumentti voi koostua useammasta entiteetistä
- päädokumentti = dokumenttientiteetti
  - voi olla ainoa entiteetti
- "alidokumentit" = entiteettejä
  - on oltava nimi
  - viittaus dokumenttientiteetistä ->
  - jäsennin tietää miten dokumentti koostetaan
- Kaikki entiteetit esitellään
  - entiteetin esittely
- viitataan entiteettiin entiteettiviittauksen avulla
- sisäiset ja ulkoiset entiteetit
- jäsennetyt tai jäsentämättömät entiteetit

# Sisäiset tekstientiteetit

- Ennaltamääritelty vakiomerkkijono
  - Nimi ,korvaava merkkijono
- Lainausmerkit vs. heittomerkit
  - <!ENTITY sent 'His foot is 12" long'>
  - <!ENTITY sent "His foot is 12" long">

| Entiteettiviittaus | vastaava merkkijono |
|---|---|
| > | > |
| < | < |
| " | " |
| ' | ' |
| &amps; | & |
| < | < |
| A | A |
| < | < (hexadecimal) |
| ￸ | ... (Unicode) |

# Parametrientiteetti

Lyhennystapa (ainoastaan ulkoisessa) DTD:ssä

```
<!ENTITY % parapart "(emph | supersc | subsc)">

<!ELEMENT paragraph (%parapart | bold)>
<!ELEMENT list (%parapart | item)*>


<!ELEMENT paragraph (emph | supersc | subsc | bold)>
```

# Ulkoinen entiteetti

- Dokumenttientiteetin ulkopuolella
- Järjestelmätunnus
  - <!ENTITY myfile SYSTEM "extra_files/file.xml">
- Julkinen tunnus
  - <!ENTITY myfile PUBLIC "... description...">
  - vaatii luettelon (indeksin)
- Binäärientiteetti
  - Binääritietoa
  - Myös esitysmuoto kerrottava
  - <!ENTITY myphoto SYSTEM "/figures/photo.gif" NDATA GIF>
  - Käyttö:
    Take a look at my photo <picture name="myphoto"/>.
  - Ei viittauksia suoraan tekstistä

# Entiteettiviittaus

- Yleinen tekstientiteettiviittaus
  - ilmentymässä
  - ei DTD:ssä
  - entiteettihierarkia
- binäärientiteetti (jäsentämätön)
  - ei viittauksia suoraan tekstistä
  - ainoastaan attribuutin arvona
- parametrientiteetti
  - DTD:ssä, ei ilmentymässä

# Notaation esittely

<!NOTATION PIXI SYSTEM "">

<!NOTATION TIFF SYSTEM "C:\APPS\Show_tiff.exe">

Notaatioon voidaan viitata entiteetin esitellyssä:
<!ENTITY Logo SYSTEM "logo.tif" **NDATA** TIFF>

Notaatio kertoo sovellukselle, miten jäsentelemättömiä entiteettejä käsitellään (esm. millä ohjelmalla tiedosto avataan)

# Jos DTD puuttuu

- attribuuteilla ei voi olla oletusarvoja
- attribuuttien ainoa tyyppi on silloin merkkijono (CDATA)
- kaikki attribuutit ovat valinnaisia
- entiteettejä ei voida määritellä
- ainoat entiteetit, joita voidaan käyttää ovat ennalta määritellyt entiteetit (', jne.)
- vaikea sanoa mistä elementtien sisältö koostuu
  - elementeistä, datasta vai molemmista

# DTD:n suunnittelu

- Usein XML-järjestelmä vanhan järjestelmän tilalle
- siirryttäessä XML:ään joko voidaan
  - käyttää standardi-DTD:tä ja tehdä mahdolliset muutokset
  - luoda kokonaan uusi DTD
- tukea esim.
  - yrityksen dokumenttimalleista
  - (edustavista) esimerkkidokumenteista
    - sekä niiden tekijöiltä

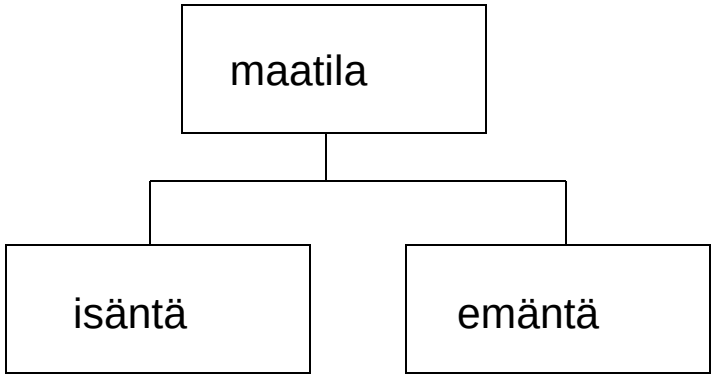
# Dokumentin osien analysointi

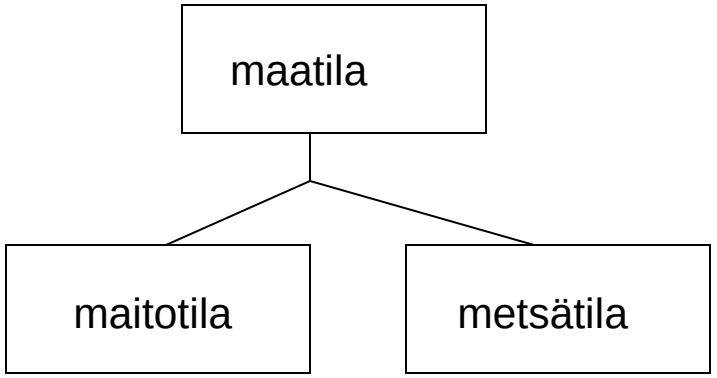
- Dokumentin piirteistä, selvitä
  - onko sillä nimi, vai voiko se puuttua
  - esiintyykö kerran/useamman kerran
  - mikä informaatio edeltää ja seuraa sitä (aina)
  - voidaanko se jakaa osiin (joista voidaan selvittää nämä samat kysymykset)
  - sisältääkö se vakiotekstiä (jota voitaisiin generoida automaattisesti)
- XML-dokumentti (tai sen osa) ehkä generoidaan suoraan tietokannasta
  - silloin voidaan DTD:n kirjoittamisessa käyttää hyväksi mm. mahdolliset tietokantakaaviot

# DTD:n suunnittelu

- standardi-DTD vai uusi?
- yhteensopivuustarpeet ja tiedon vaihto
- prosessointitarpeet
- tulevaisuuden tarpeet
- millaisia nimiä käytetään?
- elementti vs. attribuutti, kumpi valitaan?
- säännöt? sääntöjen järjestys?
- kommentit?
- modulaarisuus?
- nimeämistyyli, lyhyet vai kuvailevat nimet, isot vai pienet kirjaimet?
- elementtien rakenteisuus, järjestys, granulariteetti

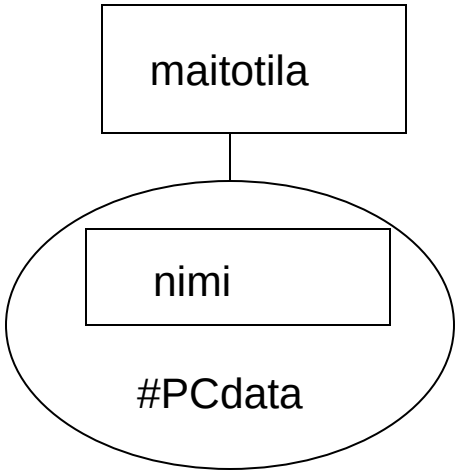
# Puudiagrammit




<!ELEMENT maatila (isäntä, emäntä)>

<!ELEMENT maatila (maitotila | metsätila)>



# Standardi-DTD:t, esim.

- http://www.xml.com/pub/rg/DTDs
- http://www.xml.org/
- http://xml.coverpages.org/
- MathML, CML (kemia), UXF (UML eXchange Format), SMIL (multimedia), RDF (Resource Description Framework), HumanML (luonnollinen kieli), DocBook, jne.
- http://www.ebxml.org/specs/ebBPSS.dtd