



## XML 1.0 (eXtensible Markup Language)

Mikä on XML ja mihin sitä käytetään  
XML:n kielioppi  
Hyvin muodostettu XML-dokumentti

1



## Mikä on XML ja mihin sitä käytetään?

XML on merkkäuskieli, jolla kuvataan tietoa ja sen rakennetta.

XML:ää käytetään tiedonsiirtoon sovellusten välillä, julkaisemiseen sekä dokumenttien tallettamiseen.

2

## Mikä on XML

- ◆ Metakieli, jolla voidaan määritellä omia merkintäkieliä
  - tietoa kuvataan merkintäsanoilla (elementeillä) ja määreillä (attribuuteilla), jotka voidaan päättää itse XML-kieliopin puitteissa => tiedoilla on rakenne
  - merkkkaus kuvaa sitä, mitä dokumentti sisältää, ei sitä miten se esitetään => tiedoille tehdään erikseen esittäminen

```

<henkilo tunnus="mars" esimies="salse">
  <nimi>
    <sukunimi entinen="Lehtonen">Marttila</sukunimi>
    <etunimi>Sirpa</etunimi>
  </nimi>
</henkilo>

```

Diagrammilla on kaksi viivaa: yksi osoittaa elementtiin `<sukunimi entinen="Lehtonen">Marttila</sukunimi>` ja toinen osoittaa attribuuttiin `entinen="Lehtonen"`.

- ◆ Paljon valmiita merkintäkieliä esim. XHTML (web-sivut), RSS: uutiset ja blogit (<http://www.hs.fi/uutiset/rss/>), e-lasku: verkkolasku (<http://www.pankkiyhdistys.fi/verkkolasku/>)

3

## Ja miksi XML

- ◆ Itseen kuvaava merkintä
    - elementtien (merkkkaus) ja attribuuttien (määritteet) avulla kerrotaan, mitä dokumentti sisältää
  - ◆ Sisällön ja esittämisen erottaminen
    - XML kertoo, mitä tieto tarkoittaa ei miten se esitetään
    - samasta tiedosta voidaan luoda useita eri esitysmuotoja eri medioille (esim. selain, kännykkä, paperi)
  - ◆ Avoin, ilmainen
    - tekstitiedosto, unicode merkistö
    - käyttöjärjestelmä- ja laitteistoriippumaton
    - ohjelmisto- ja ohjelmointikieliriippumaton
    - toimittajariippumaton
    - W3:n suositus, jonka kehittämiseen ovat osallistuneet tärkeimmät ohjelmisto- ja laitevalmistajat
- => standardiperustainen koodaus

4

## Mitä standardiperustainen koodaus tarkoittaa?

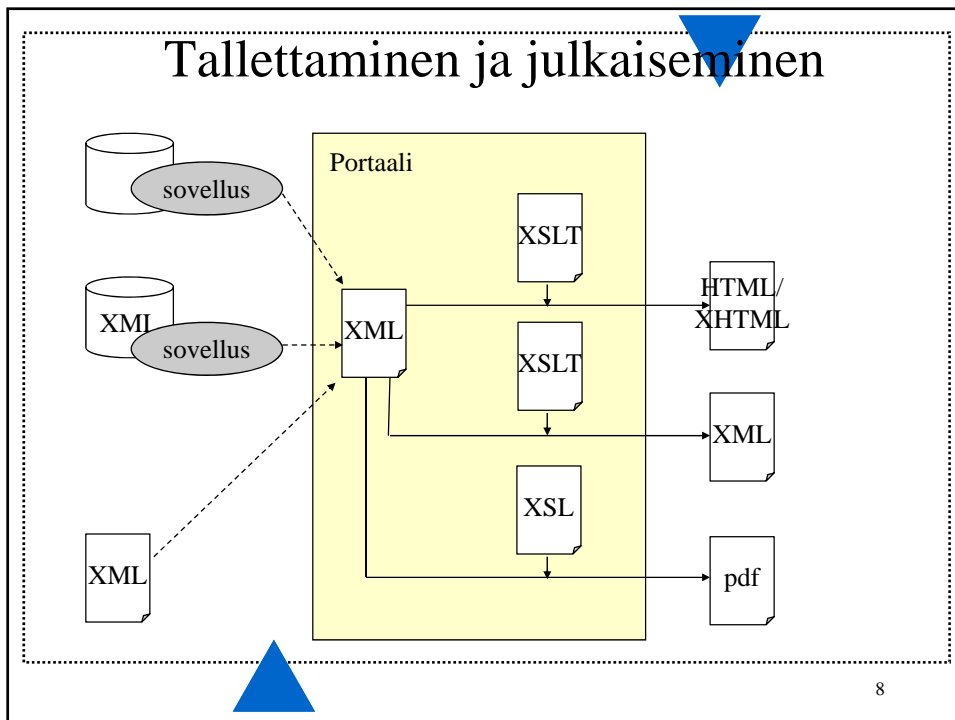
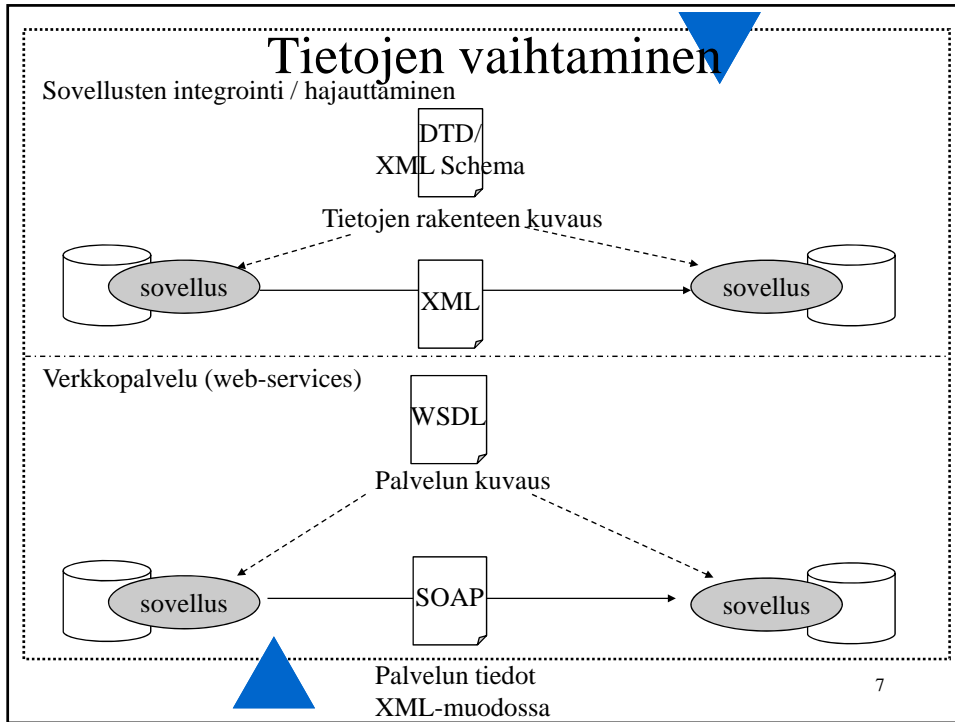
- ◆ XML-kielioppi on standardoitu <http://www.w3.org/>
- ◆ XML liittyy monia liitännäisstandardeja mm.
  - DTD / XML Schema: sanastojen määrittely (rakennemäärittely)
  - XSLT: muunnokset HTML / XHTML / XML
  - XSL: pdf julkaisujen tekeminen
  - WSDL, SOAP: verkkopalvelujen kehittäminen
- ◆ Ohjelmointirajapinnat
  - SAX: tapahtumapohjainen ohjelmointi
  - DOM: puurakenteinen ohjelmointi


5

## Mihin XML käytetään?

- ◆ **Tietojen vaihtoon:** järjestelmien hajauttaminen/integroiminen, tietojen kokoaminen useasta paikasta
  - ⇒ tietojen vaihdon standardi, ohjaa rakentamaan yhtenäisiä rajapintoja
- ◆ **Julkaisemiseen:** tietojen ja niiden esittämisen erottaminen
  - ⇒ tiedon uudelleenkäyttö
  - ⇒ meta-tiedon avulla sisällön tuottaminen (semanttinen web)
- ◆ **Tallettamiseen:** dokumenttien talletusformaatti, juridinen ja tekninen dokumentaatio
  - ⇒ asiakirjojen hallinta ja arkistointi

6





## XML:n syntaksi


Prologi  
Elementti  
Attribuutti  
CDATA  
Prosessointiohje

9

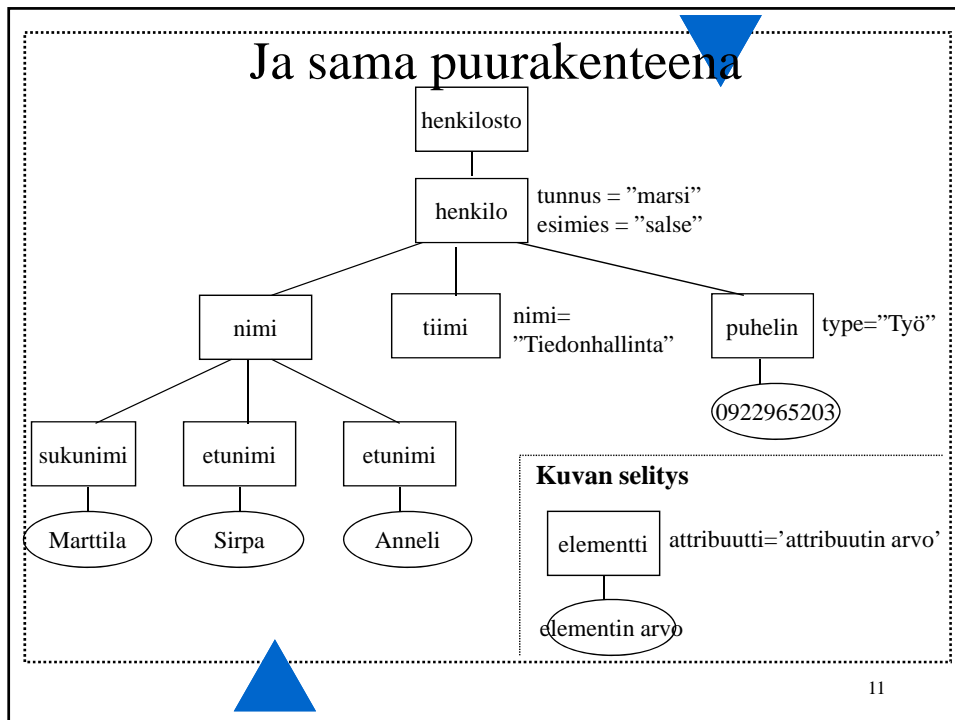


## Esimerkki XML-dokumentista

```
<?xml version="1.0" encoding="ISO-8859-15" ?>  
<henkilosto>  
  <henkilo tunnus="mars" esimies="salse">  
    <nimi>  
      <sukunimi>Martila</sukunimi>  
      <etunimi>Sirpa</etunimi>  
      <etunimi>Anneli</etunimi>  
    </nimi>  
    <tiimi nimi="Tiedonhallinta"/>  
    <puhelin type="työ">0922965203</puhelin>  
  </henkilo>  
</henkilosto>
```



10



- ◆ Rakenne kuvataan merkkaukielellä. Merkkauk kuva dokumentin loogisen rakenteen ottamatta kantaa siihen, miten eri rakenneosat pitäisi esittää.
  - ◆ Merkkauksen voima on siinä, että se ei kerro ulkoasua vaan erottaa muuttumattoman rakenteen mahdollisesti muuttuvasta esitysmuodosta.
  - ◆ Rakenne kuvataan
    - **elementeillä**  
`<puhelin type="työ">09-14890270</puhelin>`
    - elementillä on **sisältö** 09-14890270.
    - elementillä voi olla myös **attribuutteja** type="työ"
  - ◆ Rakenteeseen ei kirjoiteta dokumentin esittämisestä kertovaa tietoa
  - ◆ Esittämisestä tehdään oma dokumenttinsa CSS/XSLT/XSL-kielillä
- 12

## Prologi

```
<?xml version="1.0" encoding="ISO-8859-15" standalone="yes"?>
```

- ◆ Prologi aloittaa XML-dokumentin, suositeltava ei pakollinen.
- ◆ version kertoo dokumentin käyttämän XML-standardin.
- ◆ encoding kertoo käytettävän merkistön. Jos encoding puuttuu, merkistö on UTF-8. ISO-8859-15 sisältää skandit.
- ◆ standalone kertoo, aiheuttaako rakennemäärittely (DTD) muutoksia XML-dokumenttiin ("yes" ei aiheuta muutoksia, "no" aiheuttaa muutoksia). Jos standalone puuttuu, oletusarvona on "yes".

13

## Elementti

```
<henkilosto>  
  <henkilo tunnus="mars" esimies="salse">  
    <nimi>  
      <sukunimi entinen="Lehtonen">Marttila</sukunimi>  
      <etunimi>Sirpa</etunimi>  
      <etunimi>Anneli</etunimi>  
    </nimi>  
  </henkilo>  
</henkilosto>
```

14

- ◆ Elementillä on alku- ja lopputunniste. Lopputunnisteen on vastattava alkutunnistetta.  
`<henkilosto>...</henkilosto >`
- ◆ Elementti voi sisältää toisia elementtejä  
`<henkilosto> <henkilo>...</henkilo> </henkilosto>`
- ◆ Elementti voi sisältää jäsenettyä merkkietoa  
`<etunimi>Sirpa</etunimi>`
- ◆ Elementti voi sisältää sekä toisia elementtejä että jäsenettyä merkkietoa (mixed content)  
`<muistio>Kokouksessa <paiva>1.3.2008</paiva>  
päättiin ...</muistio>`
- ◆ Elementti voi olla tyhjä  
`<tiimi nimi= "Tiedonhallinta" />`

15

- ◆ Dokumentissa on yksi juurielementti, jonka sisällä ovat kaikki muut elementit. Elementtien täytyy olla sisäkkäisiä.
- ◆ Elementin tunnisteet on itse valittavissa.
- ◆ Tunniste alkaa kirjaimella (a-z, A-Z) tai alaviivalla (\_). Tunniste ei voi alkaa  $X\{x\}M\{m\}L\{l\}$ . Myös skandeja voidaan käyttää (ei suositeltavaa).
- ◆ Tunnisteen muut merkit voivat olla kirjaimia (a-z, A-Z), numeroita, alaviivoja (\_), tavuviivoja (-), pisteitä (.) tai kaksoispisteitä (:).
- ◆ Isot ja pienet kirjaimet ovat merkitseviä.
- ◆ Sisältöön voidaan laittaa annetun merkistön (unicode) merkkejä sekä käyttää merkkientiteettejä.
- ◆ Alla olevia merkkejä ei saa käyttää sisällössä.

Kielletty merkki      Sen sijaan käytettävä enititeetti

&	&amp;
<	&lt;
>	&gt;
'	&apos;
"	&quot;

16

## Tehtävä

Mitkä seuraavista ovat väärin ja miksi? Korjaa ne.

- a. `<ikä>21</Ikä>`
- b. `<2006-raportti>Tässä raportissa<2006-raportti>`
- c. `<lasku><paiva>4.3.2006</lasku>`
- d. `<palkka><tuntipalkka>12.00</tpalkka>` on minimi & maksimia ei ole</palkka>
- e. `<kuva />`

17

## Attribuutti

```
<henkilo tunnus="mars" esimies="salse">  
...  
</henkilo>
```

- ◆ Attribuuteilla tarkennetaan elementtien merkitystä.
- ◆ Attribuuttien tunnisteilla on samat nimeämissäännöt kuin elementtien tunnisteilla.
- ◆ Attribuutin arvo laitetaan joko lainausmerkkeihin (") tai heittomerkkeihin (').
- ◆ Elementillä voi olla useita eri attribuutteja, joiden järjestyksellä ei ole merkitystä.
- ◆ Elementillä ei voi olla samaa attribuuttia useaa kertaa.
- ◆ Eri elementeillä voi olla samannimisiä attribuutteja.

18

## Tehtävä

Mitkä seuraavista ovat virheellisiä ja miksi? Korjaa ne.

- a. `<cd bändi="HIM nimi='DARK LIGHT'> </cd>`
- b. `<ehto hinta=">20" ja hinta="<40" />`
- c. `<kirjat nimi="Oy Yritys Ab">  
     <kirja nimi="Muumipeikko ja pyrstötähti"></kirja>  
   </kirjat>`
- a. `<paiva xml:date="3" xml:monh="3" xml:year="2006" />`

19

## Elementti vs. attribuutti

### ◆ Elementti

- antaa tiedolle semanttisen merkityksen
- tallettaa yhden tiedon sekä rakenteita
- yksinkertainen tietosisältö voidaan esittää myös attribuuttina, tietorakennetta ei voida laittaa attribuuttiin
- elementtien järjestyksellä merkitystä
- laajennettavissa

☞ Hierarkkisten rakenteiden kuvaaminen

### ◆ Attribuutti

- kuvaa semanttista merkitystä
- tallettaa yhden tiedon, ei rakenteita
- voidaan esittää aina myös elementtinä
- attribuuttien järjestyksellä ei ole merkitystä

☞ Elementtiin liittyvä meta-tieto

20

## Tehtävä

Mieti, millainen XML-dokumentti voitaisiin laatia omista kurssisuorituksistasi?

Mitkä tiedot kannattaisi laittaa elementteinä ja mitkä attribuutteina?

Voisiko jonkun tiedon laittaa sekä elementtinä että attribuuttina?

21

## CDATA

```
<script>  
<![CDATA[  
  function Laske() { ... }  
]]>  
</script>
```

- ◆ CDATA-lohkoon laitetaan sellaista tietoa, jota ei haluta tulkittavan XML-merkkaukseksi.
- ◆ Jäsentäjä ei jäsennä CDATA-lohkossa olevaa tietoa.
- ◆ CDATA-lohkossa voi olla esim. erikoismerkkejä ja/tai koodia.

22

## Prosessointiohje

```
<?xml-stylesheet href="henkilosto.xml" type="text/xsl"?>
```

- ◆ Prosessointiohjeella (<? . . . ?>) välitetään tietoa dokumenttia käsittelevälle sovellukselle (esim. XSL-jäsentäjälle).
- ◆ Parseri välittää prosessointiohjeen sovellukselle. Parseri ei tee prosessointiohjeen merkitykselle mitään tulkintaa.
- ◆ Jopa XML-dokumentin prologi on yksi prosessointiohjeista.

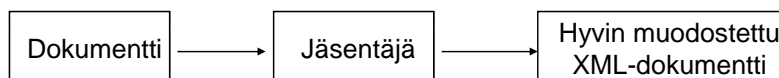
## Kommentti

```
<!-- Kommentti -->
```

- ◆ ei saa sisältää kommentteja eikä -- saa olla kommentin sisässä
- ◆ dokumentti ei voi alkaa kommentilla
- ◆ kommentti ei voi olla elementin sisässä vain elementtien välissä

23

## Hyvin muodostettu XML-dokumentti



XML-dokumentin täytyy olla hyvin muodostettu (well-formed) eli noudattaa XML-kielioppia. XML-jäsentäjä käsittelee dokumentin ja antaa virheilmoituksen, jos dokumentti ei ole hyvin muodostettu.

Dokumentti on hyvin muodostettu, jos

- dokumentilla on yksi juuri elementti
- elementit ovat sisäkkäisiä
- attribuuttien arvot ovat lainaus- tai heittomerkeissä

24

## Työkaluja

- ◆ Editori, jolla dokumentit kirjoitetaan (esim. XML Spy, Oxygen)
- ◆ Parseri/jäsentäjä, joka osaa käsitellä XML:ää
  - ilmoittaa dokumentin muotovirheet
  - ilmoittaa, jos dokumentti ei vastaa DTD:tä/XML Schemaa
  - parseri voi olla
    - ◆ komentoriviparseri
    - ◆ sovellusohjelmaan integroitu (esim. XML Spy, selain)
    - ◆ ohjelmointikielestä käytettävä (esim. Javan API:a käyttäen)
- ◆ Sovellus, jossa merkintäkielelle on käyttöä