XML

I.Savnik, FAMNIT

Ozadje

- 1978 - ANSI (*American National Standards Institute*) - ustanovljena skupina za pripravo standarda jezika za opis besedil.
- 1984 - ANSI in ISO (*International Standards Organization*) - priprava mednarodnega standarda. Osnutek je bil objavljen leta 1985, naslednje leto pa tudi sam standard SGML.

Ozadje

- SGML je bil že med nastajanjem in ob samem začetku podprt z dvema večjima projektoma:
  - 1983-1987 je delovna skupina pri *Association of American Publishers* pripravila v SGML opise zvrsti *knjiga*, *časopis* in *članek* – **EMP** (Electronic Manuscript Project)
  - 1987 *US DoD* zahteva enotno tehnično dokumentacijo vseh naročenih izdelkov – **CALS** (Continuous Acquisition and Life-cycle Support)
- 1990 – nov projekt – **HTML** (Tim Berners-Lee uporabi SGML in zgradi DTD (specifična množica označb) za hipertekst – *podobno lahko obravnavamo LaTeX, pdflatex, html2latex,... kot ~SGML/XML s specifičnim DTD*)

SGML - HTML

- HTML – specifična fiksna množica značk
- HTML – opis prezentacije
- SGML - ločuje med **obliko in vsebino**
- SGML - objavljanje (*publishing*)
- SGML – elektronski dokumenti, arhivi, knjižnice

Zakaj XML ?

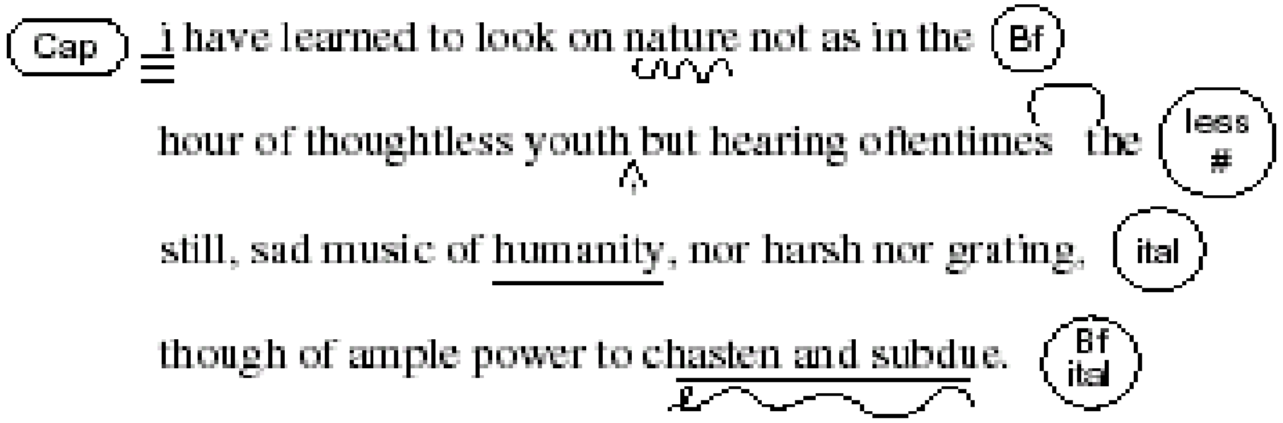
- Kompleksnost SGML
- Definicija strukture dokumenta/podatkov
  - HTML ne omogoča
- Jezik za opis podatkov
- Prenos podatkov med aplikacijami
  (v organizacijah, med organizacijami)
- Shranjevanje/prenos podatkov na spletu

XML – zahteve

- XML mora biti neposredno uporaben na spletu
- XML mora podpirati veliko množico uporab
- XML mora biti skladen s SGML
- Pisanje programov, ki procesirajo XML dokumente, mora biti preprosto
- Število dodatnih (opcijskih) elementov XML naj bo minimalno, idealno nič
- XML dokumenti morajo biti berljivi in jasni
- Priprava modela XML jezika mora biti hitra
- XML mora biti formalen in jedrnat Priprava XML dokumentov mora biti preprosta
- Razumljivost XML dokumentov je bistvenega pomena (in zgoščenost minimalnega pomena)

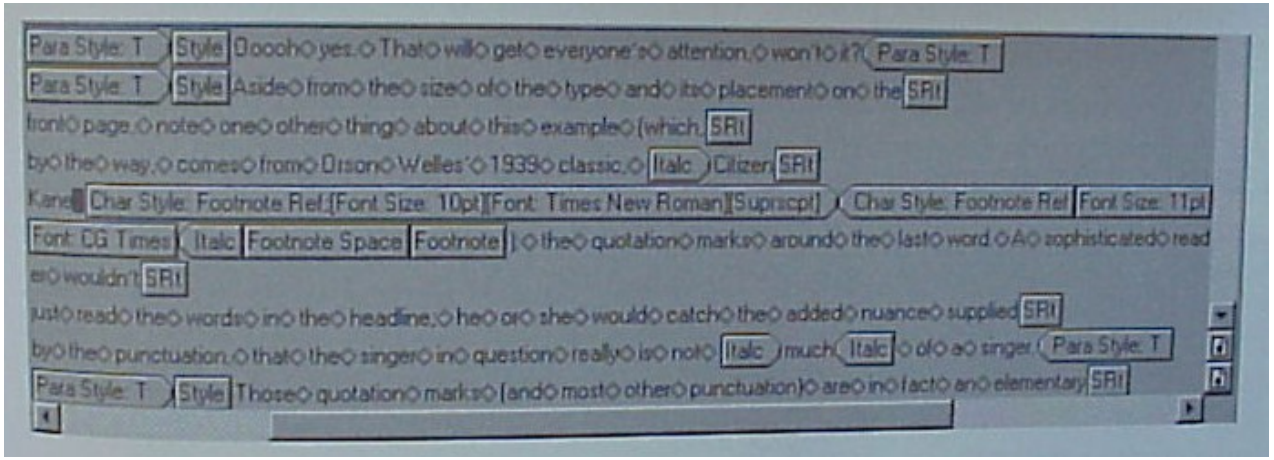
<Tag – značka>

- Očitno: označevalni jeziki uporabljajo značke
- Ideja označevanja je zelo stara: presledki (*white spaces*), označevanje začetka in konca stavka, odstavki, ..., uredniški popravki,



<Tag – značka>

- dandanašnje značke,...



Word *Codes* okno

LaTeX, Word, HTML, ...

<Tag – značka>

- Značke:
  - Postopkovno označevanje: značka določa **kaj se zgodi** z označenim delom besedila
  - Vsebinsko označevanje:
    - opisno – oblikovanje glede na izbran stil ((La)TeX, Word, ...)
    - posplošeno – določitev značilnih sestavin in njihove medsebojne povezanosti – tvorimo *znakovno podatkovno bazo* (SGML, XML)

<Tag – značka>

- SGML navodilo: značka se naj začne z znakom < in konča z znakom >
- Zakaj tekstovne značke? Večja prenosljivost med različnimi okolji! Boljša berljivost!

XML

- XML je namenjen
  - Strukturiranju,
  - shranjevanju in
  - prenosu podatkov
- XML opisuje podatke in zveze med podatki
- XML nima definiranih značk
  - značke definiramo sami - slovnica
  - XML je množica podobnih jezikov
    (prim. kontekstno neodvisni jeziki)
- XML uporablja DTD ali XML Schema za opis strukture podatkov - slovnice

XML – zgradba jezika

- Struktura + Vsebina
- DTD (*Document Type Definition* – zvrst spisa) definira strukturo zvrsti dokumenta
  - slovnica *G* in organizacija značk
- XML dokument, ki uporablja te značke za označevanje vsebine
  - beseda *w*, ki je v jeziku *L(DTD)*

XML – struktura dokumenta

- Uvod (*Prolog*) – ukazi za definiranje razpoznavalnika (*parser*)
  - slovnica, *G*
- *Telo* – vsebina, pomembna za ljudi
  - beseda *w*
- Zaključek (*Epilog*) – zaključni komentarji

XML - jezik

- XML dokument je beseda *w* v jeziku, ki ga definira slovnica DTD $G=(\Sigma, \Gamma, S, P)$:
  - vsak dokument ima korenski element (S)
  - med elementi veljajo odnosi kot jih definirajo produkcije P (drevo izpeljave)
  - simbole ($X \in \Gamma$) imenujemo tudi imenujejo **veje**, črke ($a \in \Sigma$) pa **listi**
  - elementi lahko vsebujejo atribute ali prilastke, ki jim dodajo funkcionalnost (npr. podajajo opise, omogočajo iskanje ipd.)
- Napake v besedilu bodo programsko opremo ustavile
  - prevajanje
  - do napake pride, če $w \notin L$ (DTD)

XML - osnovni gradniki

- Elementi
  - Značke – opisujejo podatkovni objekt
- Atributi (prilastki)
  - Opisujejo lastnosti objektov - elementov
- Entitete
  - Deli skupnega teksta
- Komentarji

XML - primer

```
<?xml version="1.0"?>
<!-- File Name: Inventory.xml -->

<inventory>
        <book genre=“comp” id=“b724”>
                <title>Database Management Systems</title>
                <author><firstname>Raghu</firstname>
                    <lastname>Ramakrishnan</lastname>
                </author>
                <publisher>McGraw Hill</publisher>
                <year>2000</year>
        </book>
        …
</inventory>
```

XML - osnovna pravila

- Vsebuje deklaracijo dokumenta
- Vsaka začetna značka mora imeti tudi pripadajočo končno značko
- Strukture morajo biti pravilno gnezdene
- Vrednosti atributov so navedene v narekovajih
- Na najvišjem nivoju je lahko le en element – korenski element (*S*)

XML - veljaven dokument

- XML dokument je *dobro oblikovan*, če veljajo v njem osnovna pravila
  - Ima pravo obliko in strukturo
  - Preverjanje vsebine (telo XML dokumenta)
- XML dokument je *veljaven*, če:
  - je dobro oblikovan in
  - njegova struktura ustreza slovnici (DTD definiciji)
    - Preverjanje DTD in XML
    - Beseda jezika *L* (DTD)

XML - atribut

- Alternativni način opisa informacij
  - Opis lastnosti elementa
- Vrednost atributa je v narekovajih “”

<book genre=“comp”>

XML - identifikatorji

- Zgolj oblikovni gradnik
- Veljajo znotraj celotnega dokumenta

```
<person id="o555">  <name>Janez</name> </person>
<person id="o456">  <name>Marija</name>
                    <children idref="o123 o555"/>
</person>
<person id="o123" mother="o456"><name>Tone</name>
</person>
```

DTD

- Document Type Descriptor
- Podedovan od SGML
- Podoben DB shemi, čeprav ni zares …
- BNF slovnica (prim. kontekstno neodvisni jezik)
- Definicija elementov specifičnega XML jezika

DTD - primer

```
<!element book (title,author*,publisher,year) >
<!element title #PCDATA >
<!element publisher #PCDATA >
<!element year #PCDATA >
<!element author (firstname,lastname,address?,age?) >
<!attlist book id ID #required >
<!attlist book genre CDATA #required >
…
```

DTD - zgradba

- Elementi
  - <!ELEMENT element-name category>
  - <!ELEMENT element-name (element-content)>
  - E1,E2,... ; E* ; E+ ; E? ; E1 | E2
  - Tipi: #PCDATA, EMPTY, ANY

DTD - zgradba

- Atributi
  - <!ATTLIST el-name attr-name attr-type default-value>
  - Tipi: CDATA, (en1|en2|..), ID, IDREF, ENTITY,…
  - Določila: #required, #implied, #fixed

DTD - zgradba

- Entitete
  - <!ENTITY entity-name "entity-value">
  - <!ENTITY writer "Donald Duck."> -- uporaba: &writer;

DTD+XML vs. Relacije

- XML
  - Primeren bolj za dokumente kot podatke
  - Ime in tip elementov so povezani globalno
  - Ni podpore podatkovnim tipom
    - Ni validacije podatkov
  - Imamo lahko en sam ključ
    - Ni ključev z več atributi
    - Ni tujih ključev (reference na ključe)
    - Ni omejitev za IDREF
  - Ni podpore za ponovno uporabo strukture
    - OO strukture niso podprte
- Ni mogoče ohraniti informacije pri prevajanju iz relacijskega podatkovnega modela v XML

XML – pomembni standardi

- CSS,XSL/XSLT:
  - Prezentacija in transformacija dokumentov in podatkov
- RDF: resource description framework
  - Meta-info, kategorije, semantične mreže, …
- Xpath/Xpointer/Xlink:
  - standard za povezovanje dokumentov in elementov
- Namespaces:
  - Delo z imeni
- DOM: “Document Object Model”
  - Delo z XML dokumenti v prog. jezikih
- SAX: Enostaven API za XML
- XQL,XQuery, XML-QL: poizvedovalni jezik

Zaključek

- Podjetja uvajajo XML kot medij za prenos podatkov med oddelki
- Sistemi (distribuirani) uporabljajo XML kot medij za prenos podatkov
- Konfiguracijske datoteke, logi, … sistemov so pogosto v XML
- Večina DBMS ima XML vmesnik

Viri

- Peter Peer, Andrej Brodnik
  - prosojnice za 2003/04
- Raghu Ramakrishnan
  - Knjiga: Database Management Systems
- http://www.w3.org/XML
  http://www.w3schools.com/