Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Predmet:
Osnove podatkovnih baz

Modul:
Uvod v XML

Gradivo:
v.2016

5.1.
2017

Vsebina

- Kaj je XML
- Primerjava s HTML in prednosti XML
- Sintaksa XML
- XML drevesna struktura
- Ravni skladnosti XML dokumentov
- XML shema
- XML orodja
- XML podatkovne baze
- Podpora XML v MySQL

Kaj je XML?...

- Razširljivi označevalni jezik (eXtensible Markup Language)
- Podoben jeziku HTML

Izsek primera HTML datoteke

```
<html>

<head>
    <title>My First Web Page</title>
</head>

<body>
    <h1>My First Web Page</h1>
    <p><b>Hello World Wide Web!</b></p>
    <p><i>Hello World Wide Web!</i></p>
    <p><u>Hello World Wide Web!</u></p>
    <p>This is my first web page.</p>
    <p>HTML tags can give <b><i>various</i></b>
    <u>looks and format</u> to the content of this web page.</p>
</body>

</html>
```

Kaj je XML?

- Namenjen prenosu, ne prikazu!
- Značke (tags) niso predpisane vnaprej.
- Je samopojasnjujoč.
- Priporočilo organizacije W3C (http://www.w3.org/).

Izsek primera XML datoteke

```
<person id="Big.Boss">
    <name>
        <family>Boss</family>
        <given>Big</given>
    </name>
    <email>chief@oxygenxml.com</email>
    <link subordinates="one.worker two.worker three.worker four.worker five.worker"/>
</person>
<person id="one.worker">
    <name>
        <family>Worker</family>
        <given>One</given>
    </name>
    <email>one@oxygenxml.com</email>
    <link manager="Big.Boss"/>
</person>
```

Razlike med HTML in XML

HTML	XML
vnaprej določen nabor značk	značke definiramo sami
značke namenjene določanju videza dokumenta	značke opisujejo pomen dokumenta
značke lahko izpuščamo	vse značke morajo biti prisotne
strani pogosto nepravilno zapisane – npr. napačne značke	dokumenti morajo biti „ustrezni“ – dobra definiranost

Prednosti XML pred HTML…

- HTML stran vsebuje podatke, ki so zapisani tako, da so razumljivi predvsem HTML prikazovalnikom (brskalniki).

<title> **Google ustavil digitalizacijo starih časnikov (Heise, 20. maj 2011)** </title>

<h1> Omrežja / internet </h1>

 Google je končal digitalizacijo starih časnikov, ki predstavljajo odsev dogajanja po svetu v zadnjih 250 letih. Projekt se je začel leta 2006 z indeksiranjem starih izvodov The New York Timesa, The Wall Street Journala in drugih ameriških časnikov, bistveno pa so ga razširili leta 2008 s skeniranjem mikrofilmov in digitalizacijo vsebine.

Googlov cilj je bil veličasten: želi so najti milijardo časopisnih strani, približno kolikor naj bi jih bilo v zgodovini napisanih, in jih zbrati na enem mestu v obliki, ki omogoča enostavno listanje. Do danes so zbrali več kot 2000 časnikov, ki so predstavljeni na spletu, med katerimi najdemo vse pomembne zgodbe zadnjih dveh stoletij, recimo pristanek na Mesecu.

Projekt je sedaj prekinjen, tako da Google ne bo več dodajal novih vsebin. Digitalizirana dela, vseh skupaj je 60 milijonov strani, bodo ostala prosto dostopna na internetu. Novih funkcionalnosti in vsebin ne bo, tako da bodo skenirana in neindeksirana dela žal ostala v vicah. Namesto tega se bo sedaj posvetil drugim nalogam, kot je recimo Google One Pass za prodajo vsebin založnikov neposredno s svojih strani prek Googla. Predčasni prekinitvi navkljub je Google ustvaril velik digitalni arhiv časnikov, kjer so vsi pomembnejši časniki z večjo naklado.

Matej Huš

HTML ne daje vsebinskih informacij:

- Kdo je avtor zgodbe?
- V katero kategorijo je umeščena?
- Kdaj je bila objavljena?
- …

- Programska obdelava HTML strani je zahtevna ker manjka struktura podatkov.

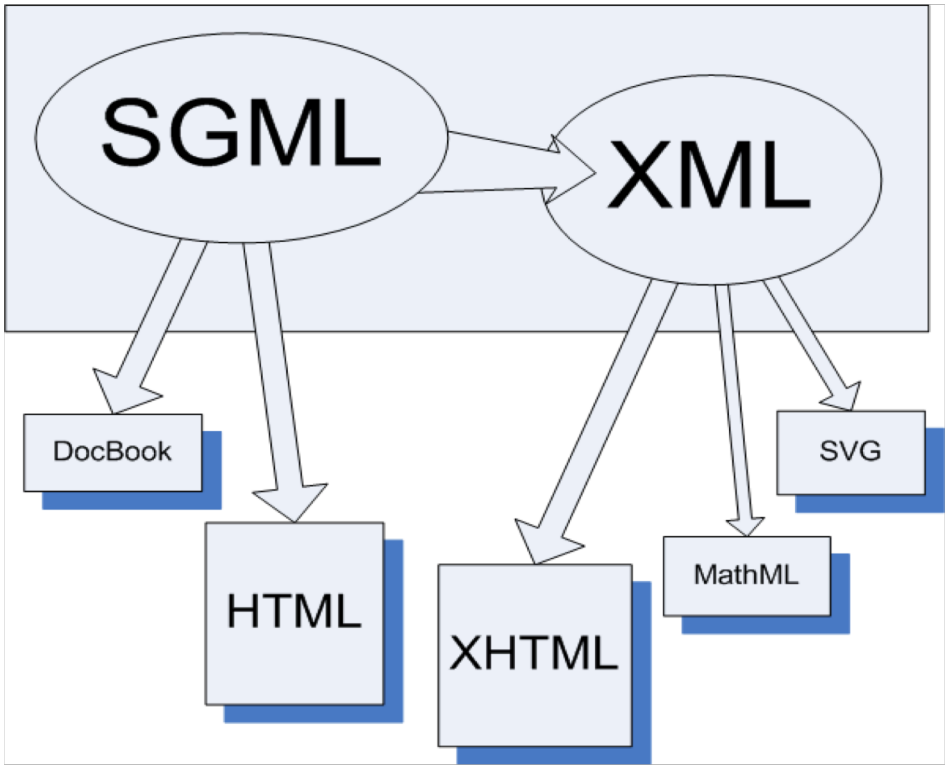
Prednosti XML pred HTML

```
<B>John Q Public</B>
<P>
john.q.public.1@gsfc.nasa.gov<BR>
phone: 301-286-aaaa<BR>
fax: 301-286-bbbb<BR>
Bldg. 23, Rm. 999<BR>
NASA<BR>
Goddard Space Flight Center<BR>
588.0<BR>
Greenbelt, MD 20221<BR>
```


```
<EMPOLYEE>
  <NAME>
        <FIRST>John</FIRST>
        <MIDDLE>Q</MIDDLE>
        <LAST>Public</LAST>
  </NAME>
  <EMAIL>john.q.public.1@gsfc.nasa.gov</EMAIL>
  <PHONE>301-286-aaaa</PHONE>
  <FAX>301-286-bbbb</FAX>
  <LOCATION>
        <BUILDING>Bldg. 23</BUILDING>
        <ROOM>999</ROOM>
  </LOCATION>
  <ADDRESS>
        <ORG>NASA</ORG>
        <CENTER>Goddard Space Flight Center</CENTER>
        <MAILSTOP>588.0</MAILSTOP>
        <CITY>Greenbelt</CITY>
        <STATE>MD</STATE>
        <ZIP>20221</ZIP>
  </ADDRESS>
</EMPLOYEE>
```

Uveljavljena tehnologija

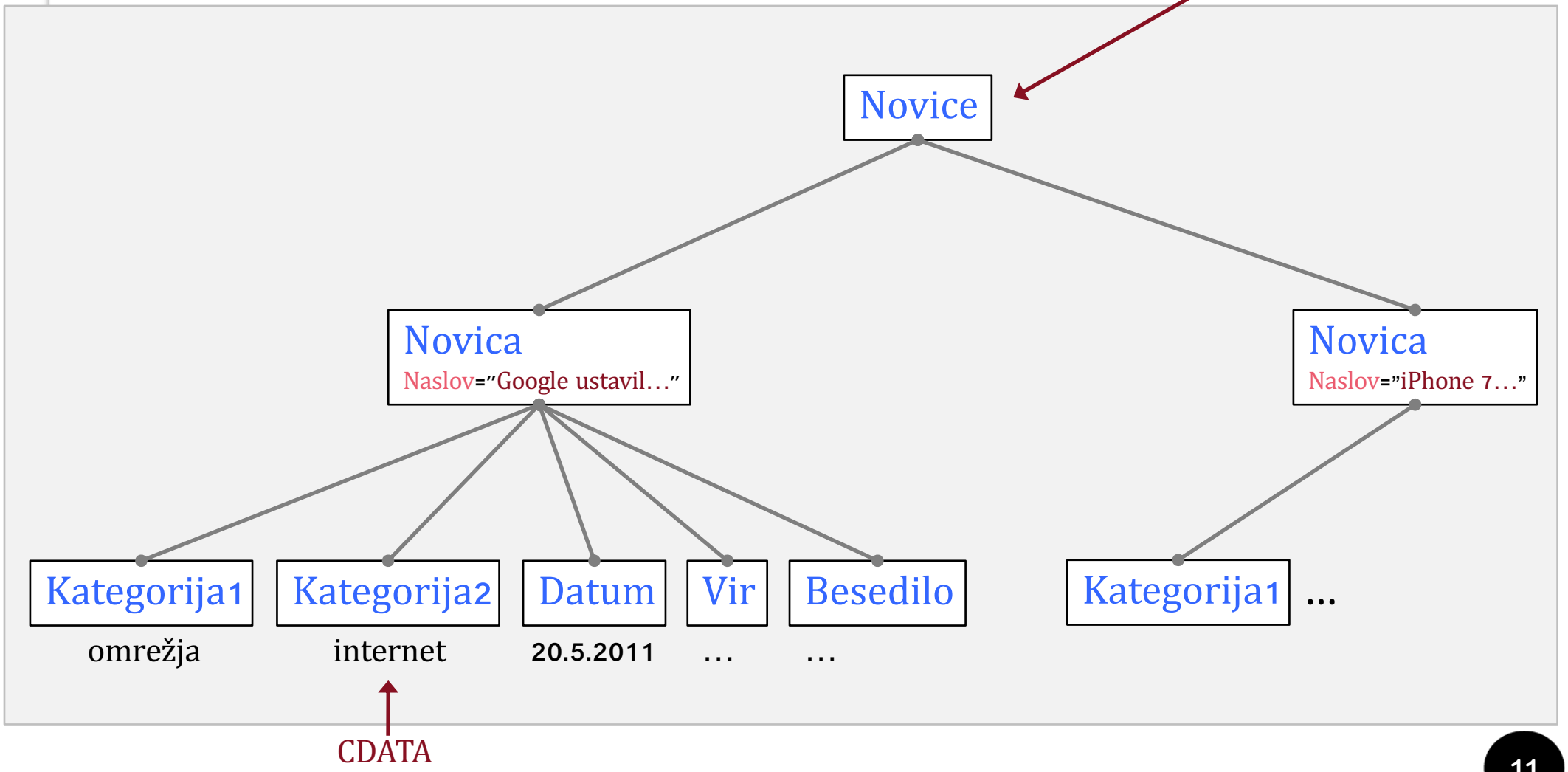
- XML je podmnožica standarda SGML (Standardized Generalized Markup Language), ki je bil definiran že leta 1969



Sintaksa XML dokumenta


```
<?xml version="1.0" ?>
<novice>
   <novica naslov="Google ustavil digitalizacijo starih
      časnikov">
      <kategorija1>omrežja</kategorija1>
      <kategorija2>internet</kategorija2>
      <datum>20.5.2011</datum>
      <vir Heise/vir>
      <avtor>Matej Huš</avtor>
      <besedilo>Google je končal digitalizacijo starih…
      </besedilo>
   </novica>
   <novica naslov="iPhone 7 kmalu v prodaji">
      <kategorija1>telefonija<kategorija1>
      …
   </novica>
</novice>
```

XML dokument in drevesna struktura



Ravni skladnosti XML dokumentov

- Pravilno strukturiran (well formed): zadošča vsem WC3 sintaktičnim pravilom za XML
  - Pravila za poimenovanje, gnezdenje, označevanje atributov…
  - Dobra definiranost je za XML dokumente obvezna

- Veljaven (valid) dokument je dokument, ki je skladen s shemo
  - Formalnost je za XML dokument opcijska;
  - Lahko preverjamo z XML shemo;

Pravila dobro definiranih XML dokumentov...

- XML standard zahteva, da se v dokumentih upoštevajo naslednja pravila:
  - Obstajati mora en element, ki vsebuje vse ostale
  - Značke morajo biti uravnotežene

```
<BOOK>...</BOOK>
<BOOK />
```

  - Gnezdenje značk mora biti izvedeno pravilno.

```
<BOOK> <LINE> to je pravilno </LINE> </BOOK>
<LINE> <BOOK> to </LINE> gotovo ni </BOOK> pravilno
```

  - Tekst značk je občutljiv na velike in male črke

```
<P> različna začetna in končna značka – XML tega ne dopušča </p>
```

Pravila dobro definiranih XML dokumentov

- Atributi v značkah (tags) morajo biti znotraj narekovajev.

```
< ITEM Category="Home and Garden" Name="hoe-matic t500">
```

- Komentiranje je dovoljeno.

```
<!-- They are done just as in HTML… -->
```

- Dokument se mora začeti z ustreznim nizom (XML deklaracija).

```
<?xml version='1.0' ?>
```

- Posebni znaki se morajo pričeti z ubežnim znakom (&)

```
<formula> x &lt; y+2x </formula>
<cd title="&quot; music &quot;">
```

XML shema

- Shema je ločen dokument, ki definira elemente, atribute in strukturo XML dokumenta
- Shemo določimo tako, da
  1. definiramo slovar in
  2. določimo postavitev ter število elementov in atributov v označevalnem jeziku
- Shema definira točno določen razred dokumentov.

XML Shema - Primer

Author.xsd

Shema Author.XSD definira strukturo elementa Author

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="Author">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="FirstName" type="xs:string" />
        <xs:element name="LastName" type="xs:string" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

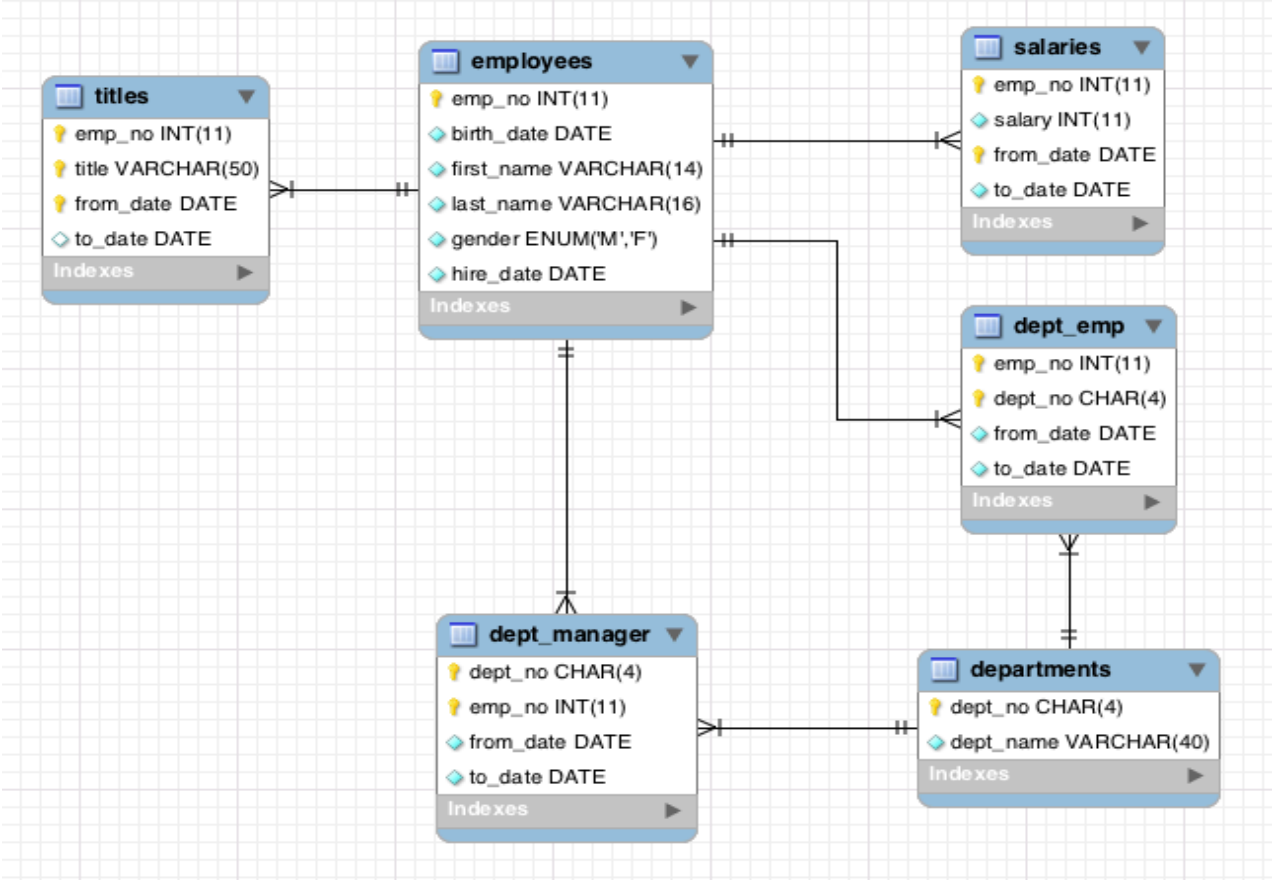
MarkTwain.xml

MarkTwain.XML je instanca oz. primerek zgoraj definirane sheme

```
<?xml version="1.0"?>
<Author xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:noNamespaceSchemaLocation="Author.xsd">
    <FirstName>Mark</FirstName>
    <LastName>Twain</LastName>
</Author>
```

Vaja

- Kako bi podatke, ki ustrezajo prikazanemu podatkovnemu modelu predstavili z XML datoteko?



Pretvorba relacijske sheme v XML…

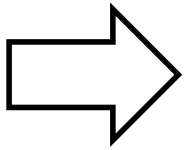
- Povezava 1 : 1

Relacija A

A1	A2
a11	a21
a12	a22

Relacija B

B1	B2	*A1
b11	b21	a11
b12	b22	a12



```
<A A1="a11" A2="a21">
    <B B1="b11" B2="b21"></B>
</A>

<A A1="a12" A2="a22">
    <B B1="b12" B2="b22"></B>
</A>
```

Pretvorba relacijske sheme v XML…

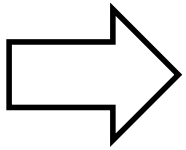
- Povezava 1 : n

Relacija A

A1	A2
a11	a21
a12	a22

Relacija B

B1	B2	*A1
b11	b21	a11
b12	b22	a12
b13	b23	a12



```
<A A1="a11" A2="a21">
    <B B1="b11" B2="b21"></B>
</A>

<A A1="a12" A2="a22">
    <B B1="b12" B2="b22"></B>
    <B B1="b13" B2="b23"></B>
</A>
```

Pretvorba relacijske sheme v XML…

- Povezava m : n

Relacija A

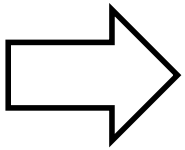
A1	A2
a11	a21
a12	a22

Relacija B

B1	B2
b11	b21
b12	b22

Relacija R

*A1	*B1
a11	b11
a12	b12



```
<A A1="a11" A2="a21" A_id="1"></A>
<B B1="b11" B2="b21" B_id="2"></B>
<R A_idref="1" B_idref="2"></R>

<A A1="a12" A2="a22" A_id="3"></A>
<B B1="b12" B2="b22" B_id="4"></B>
<R A_idref="3" B_idref="4"></R>
```

XML orodja…

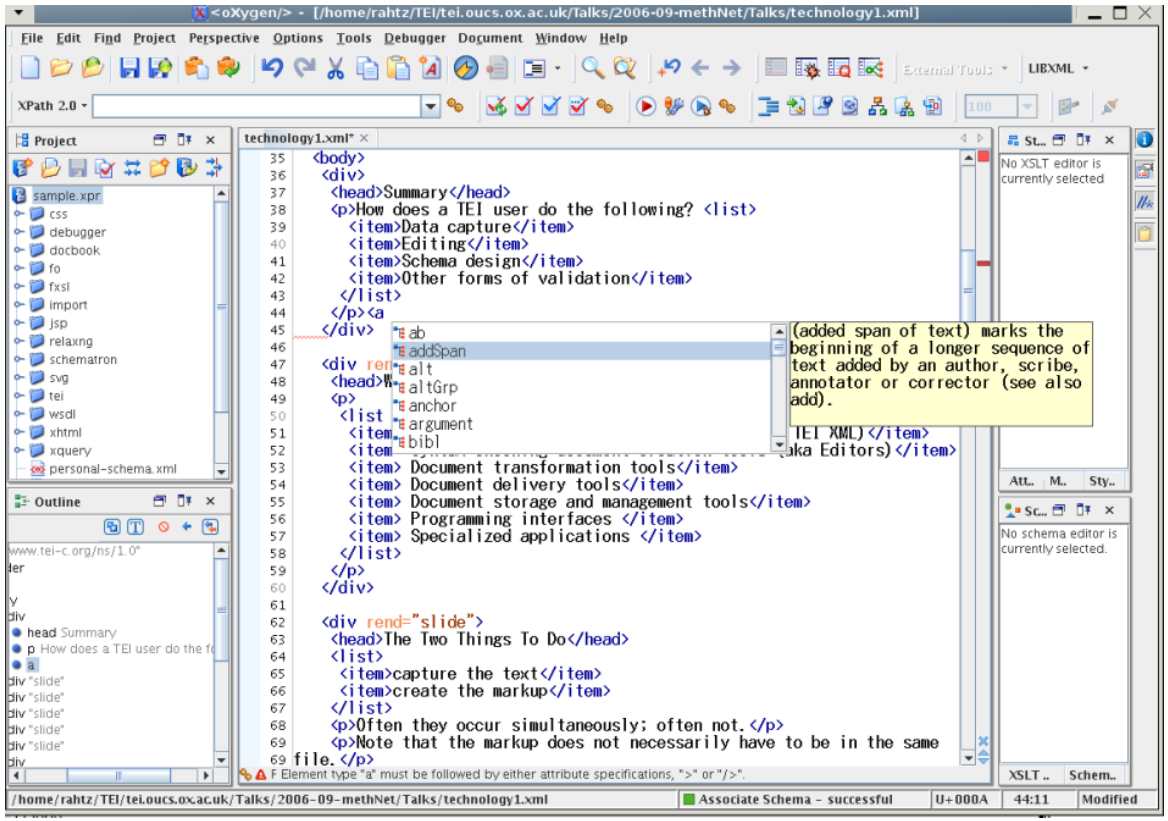
- XML urejevalniki:
  - Orodja, ki omogočajo izdelavo, urejanje in validacijo XML dokumentov.
- Orodja za obdelavo XML:
  - Orodja, ki omogočajo transformacijo XML podatkov (dokumentov) v druge standardizirane oblike… npr HTML.
- Orodja/standardi, ki omogočajo obdelavo XML iz programskih jezikov:
  - XML razčlenjevalci, ki omogočajo razčlenjevanje, validacijo, serializacijo in manipulacijo XML. Npr. Apache Xerces.
  - XML API-ji, ki omogočajo posodabljanje in branje XML podatkov potem, ko so ti že v podatkovnih strukturah nekega programskega jezika (DOM, SAX).

Izdelava XML dokumentov

- Primeri XML urejevalnikov:
  - EditiX – http://www.editix.com
  - XMLSpy – http://www.altova.com
  - Jedit - http://www.jedit.org/
  - Oxygen - http://www.oxygenxml.com/
  - XML Marker - http://symbolclick.com/
  - Sublime text - http://www.sublimetext.com/
  - Notepad

DEMO:

- Oxygen – XML urejevalnik
- http://www.oxygenxml.com/videos



XML podatkovne baze

- XML PB - nudijo upravljanje s podatki v XML formatu:
  - Shranjevanje
  - Poizvedovanje
  - Pretvorbe v različne formate
  - ...
- Tipično temeljijo na dokumentnih bazah.
- Razlog za pojav: XML standard za prenos podatkov. Nepotrebno pretvarjanje med formati...

Vrste XML PB

- PB, ki podpirajo XML dokumente:
  - Pretvarjajo XML v npr. relacije in na izhodu nazaj v XML...
  - Večina novejših SUPB podpira posebne XML podatkovne tipe.
- Naravne XML PB:
  - Interni podatkovni model temelji na XML - XML dokumenti osnovna enota shranjevanja.

PB, ki podpirajo XML

- Navadno ponujajo enega izmed naslednjih pristopov za shranjevanje XML podatkov v relacijsko bazo:
  - XML shranjen v CLOB (Character large object)
  - XML pretvorjen v množico tabel na osnovi sheme;
  - XML shranjen v naravnem XML formatu (XML tip, ISO standard)

- Relacijske PB, ki podpirajo ISO XML tip:
  - IBM DB2;
  - Microsoft SQL Server;
  - Oracle Database;
  - PostgreSQL.

Podpora XML v MySQL…

- Preusmerjanje izhoda poizvedb v XML

```
mysql –u root –p --xml;
mysql> select * from employees limit 1;
```

```
mysql> select * from employees limit 1;
<?xml version="1.0"?>

<resultset statement="select * from employees limit 1;" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <row>
        <field name="emp_no">10001</field>
        <field name="birth_date">1953-09-02</field>
        <field name="first_name">Georgi</field>
        <field name="last_name">Facello</field>
        <field name="gender">M</field>
        <field name="hire_date">1986-06-26</field>
  </row>
</resultset>
1 row in set (0.00 sec)
```

Podpora XML v MySQL…

- Branje XML datotek

person.xml

```
<?xml version="1.0"?>
<list>
  <person person_id="1" fname="Pekka" lname="Nousiainen"/>
  <person person_id="2" fname="Jonas" lname="Oreland"/>
  <person person_id="3"><fname>Mikael</fname><lname>Ronström</lname></person>
  <person person_id="4"><fname>Lars</fname><lname>Thalmann</lname></person>
  <person><field name="person_id">5</field><field name="fname">Tomas</field>
  <field name="lname">Ulin</field></person>
  <person><field name="person_id">6</field><field name="fname">Martin</field>
  <field name="lname">Sköld</field></person>
</list>
```

```
mysql> LOAD XML LOCAL INFILE 'person.xml'
    ->   INTO TABLE person
    ->   ROWS IDENTIFIED BY '<person>';
```

```
mysql> SELECT * FROM person;
+-----------+--------+------------+---------------------+
| person_id | fname  | lname      | created             |
+-----------+--------+------------+---------------------+
|         1 | Pekka  | Nousiainen | 2007-07-13 16:18:47 |
|         2 | Jonas  | Oreland    | 2007-07-13 16:18:47 |
|         3 | Mikael | Ronström   | 2007-07-13 16:18:47 |
|         4 | Lars   | Thalmann   | 2007-07-13 16:18:47 |
|         5 | Tomas  | Ulin       | 2007-07-13 16:18:47 |
|         6 | Martin | Sköld      | 2007-07-13 16:18:47 |
+-----------+--------+------------+---------------------+
6 rows in set (0.00 sec)
```

Podpora XML v MySQL…

- Iskanje po XML datotekah

```
mysql> SET @xml = '<a><b>X</b><b>Y</b></a>';
Query OK, 0 rows affected (0.00 sec)

mysql> SET @i =1, @j = 2;
Query OK, 0 rows affected (0.00 sec)

mysql> SELECT @i, ExtractValue(@xml, '//b[$@i]');
+------+-------------------------------+
| @i   | ExtractValue(@xml, '//b[$@i]') |
+------+-------------------------------+
|    1 | X                             |
+------+-------------------------------+
1 row in set (0.00 sec)
```

Podpora XML v MySQL

- Spreminjanje XML datotek

```
mysql> SELECT
    ->   UpdateXML('<a><b>ccc</b><d></d></a>', '/a', '<e>fff</e>') AS val1,
    ->   UpdateXML('<a><b>ccc</b><d></d></a>', '/b', '<e>fff</e>') AS val2,
    ->   UpdateXML('<a><b>ccc</b><d></d></a>', '//b', '<e>fff</e>') AS val3,
    ->   UpdateXML('<a><b>ccc</b><d></d></a>', '/a/d', '<e>fff</e>') AS val4,
    ->   UpdateXML('<a><d></d><b>ccc</b><d></d></a>', '/a/d', '<e>fff</e>') AS val5
    -> \G

*************************** 1. row ***************************
val1: <e>fff</e>
val2: <a><b>ccc</b><d></d></a>
val3: <a><e>fff</e><d></d></a>
val4: <a><b>ccc</b><e>fff</e></a>
val5: <a><d></d><b>ccc</b><d></d></a>
```

Naravne XML podatkovne baze

- Naravne XML PB:
  - definirajo logični model za XML dokument – omogočajo shranjevanje dokumentov, ki ustrezajo logičnemu modelu in iskanje po njih;
  - Imajo XML dokumente kot osnovno logično enoto shranjavanja (fizično je lahko karkoli).
  - Mnoge ponujajo kolekcije – logične enote skupin dokumentov (collections). PB lahko upravljajo z več kolekcijami na enkrat...
  - Ponujajo vsaj en poizvedovalni jezik: večina XPath, mnoge tudi XQuery.
  - Mnoge ponujajo XSLT za pretvarjanje med formati dokumentov (npr XML – HTML).

DEMO

- Naravna XML podatkovna baza eXists db
- http://exist-db.org/exist/apps/homepage/index.html