XML

Лекција 2

Стеван Гостојић

Факултет техничких наука, Нови Сад

10. март 2022.

Преглед садржаја

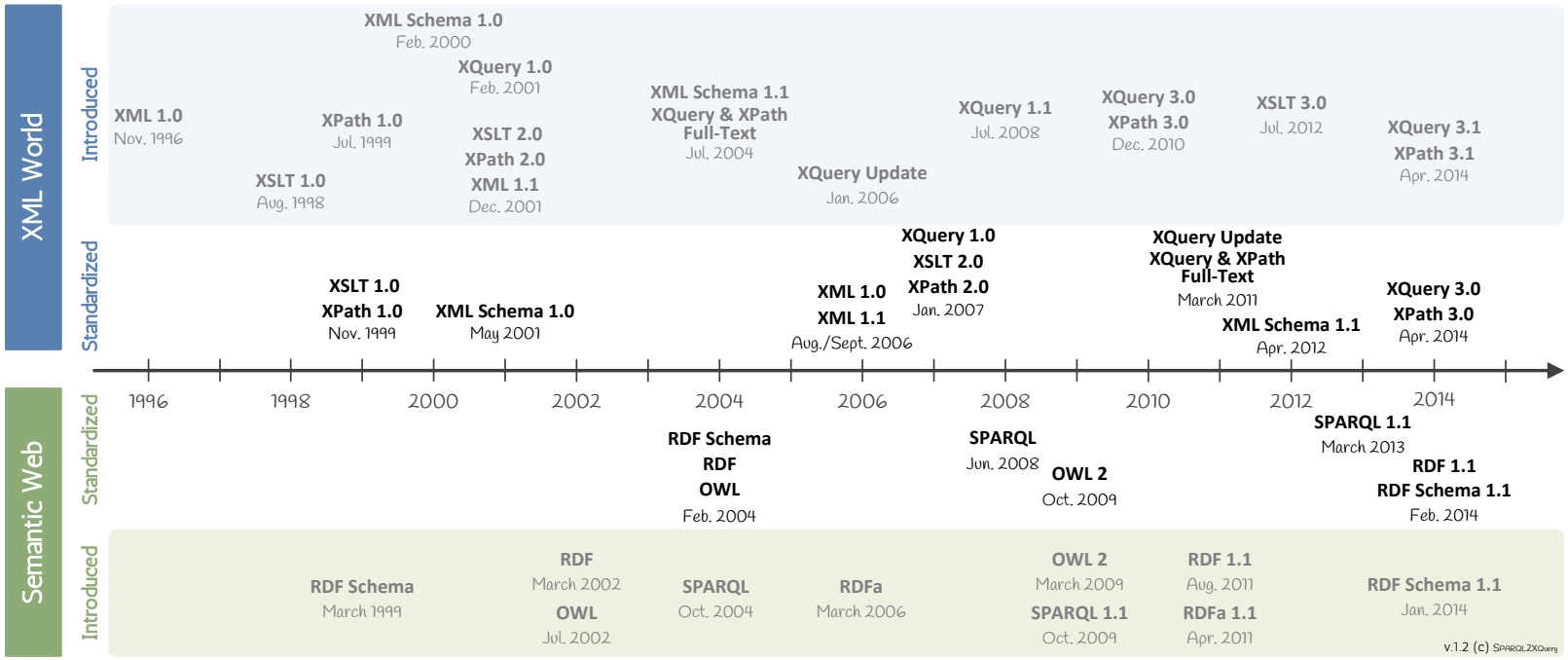
Преглед садржаја

W3C

- W3C (World Wide Web Consortium) је непрофитна организација која се бави стандардизацијом World Wide Web-а
- W3C препоруке (стандарди) су јавно и бесплатно доступни на https://www.w3.org/

W3C Technology стек

W3C временска линија



Ресурси

- Ресурс је било шта што има идентитет (RFC 2396)
  - (електронски) документи
  - сервиси
  - колекција ресурса

Ресурси

- Постоје информациони ресурси и неинформациони ресурси
  - информациони ресурси су ресурси чије се битне карактеристике могу пренети у поруци (обично имају једну или више репрезентација којима се може приступити путем HTTP протокола)
  - неинформациони ресурси су ресурци који нису информациони ресурси (апстрактни ресурси)

Репрезентације ресурса

- Репрезентација ресурса је информација која рефлектује прошло, тренутно или жељено стање ресурса, у формату који може да се лако комуницира преко протокола, и која се састоји од скупа репрезентационих метаподатака и потенцијално неограниченим током репрезентационих података (RFC 7231)

Репрезентације ресурса

- Сваки ресурс може да има више репрезентација:
  - (X)HTML
  - XML
  - RDF
  - JSON
  - итд.

text/xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<Order>
  <BuyersID>GMB91604</BuyersID>
  <BuyerParty>
  <ID>KEEN</ID>
  <PartyName>
    <Name>Maynard James Keenan</Name>
  </PartyName>
  <Address>
    <Room>505</Room>
    <BuildingNumber>11271</BuildingNumber>
    <StreetName>Ventura Blvd.</StreetName>
    <CityName>Studio City</CityName>
    <PostalZone>91604</PostalZone>
    <CountrySubentity>California</CountrySubentity>
    <Country>USA</Country>
  </Address>
  </BuyerParty>
  <OrderLine>
    <LineItem>
      <BookItem>
        <Title>Document Engineering</Title>
        <Author>Glushko and McGrath</Author>
        <ISBN>0262072610</ISBN>
      </BookItem>
      <BasePrice>99.95</BasePrice>
      <Quantity>300</Quantity>
    </LineItem>
  </OrderLine>
</Order>

```

Unicode

- Unicode је стандард за конзистентно кодирање, репрезентацију и руковање текстом
- Садржи више од 120.000 знакова који покривају 129 језика
- Знакови се могу кодирати на више начина (UTF-8, UTF-16 или UTF-32)

Преглед садржаја

XML

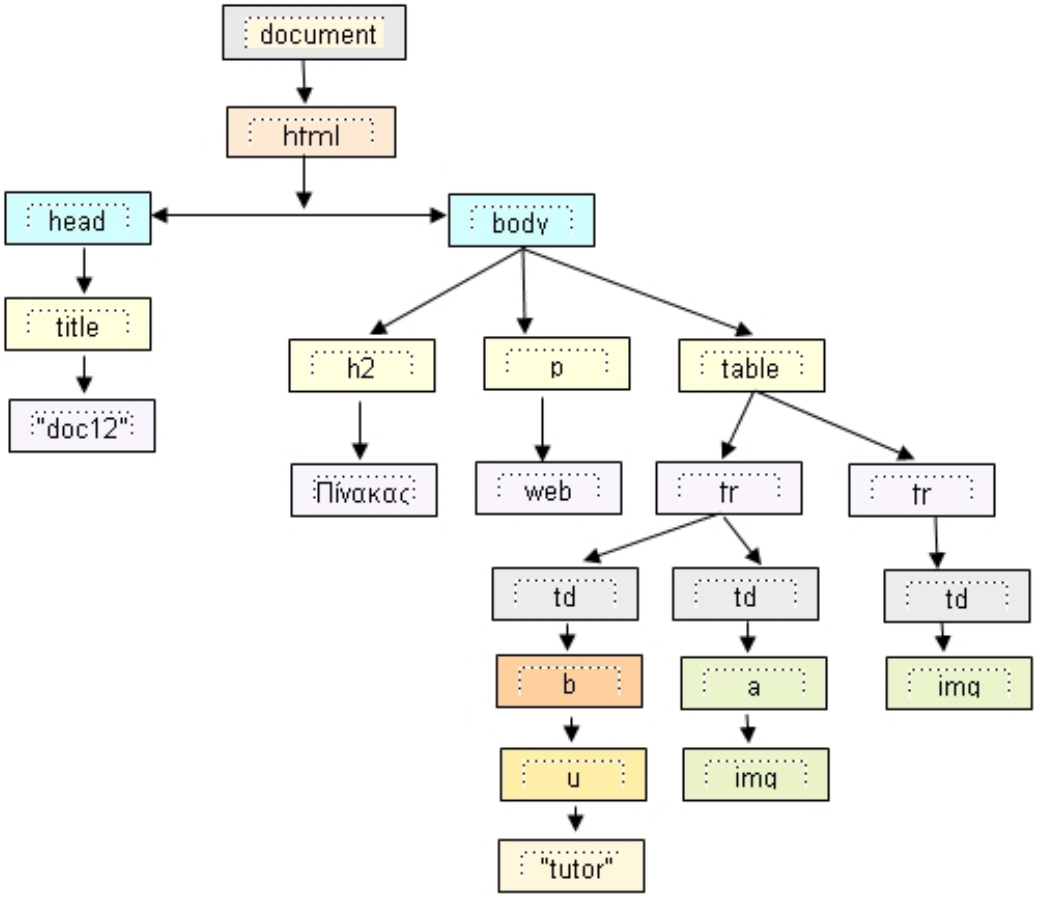
- eXtensible Markup Language (XML) је метајезик (тј. скуп правила за дефинисање конкретних језика)
- XML је (мета)језик за означавање текста (енг. markup language)
- HyperText Markup Language (HTML) се може посматрати као (један од многих) дијалекат XML-а

HTML документ

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html>
 <head>
        <title>doc12</title>
 </head>
 <body>
   <h2>Table</h2>
   <p>web</p>
   <table  width="100%">
     <tr>
       <td><b>
           <u>tutor</u>
           </b>
       </td>
       <td><a href="starweb.html">
              <img src="srat.gif" border="0" />
           </a>
       </td>
     </tr>
     <tr>
        <td><img src="sky.gif" border="0" />
        </td>
     </tr>
   </table>
  </body>
</html>

```

HTML документ



XML

- XML не садржи унапред дефинисан речник (скуп елемената и атрибута) и унапред дефинисану граматику (правила по којима се структурирају елементи и атрибути)
- Речник и граматика се дефинишу за сваки тип XML докумената појединачно (дијалекат)

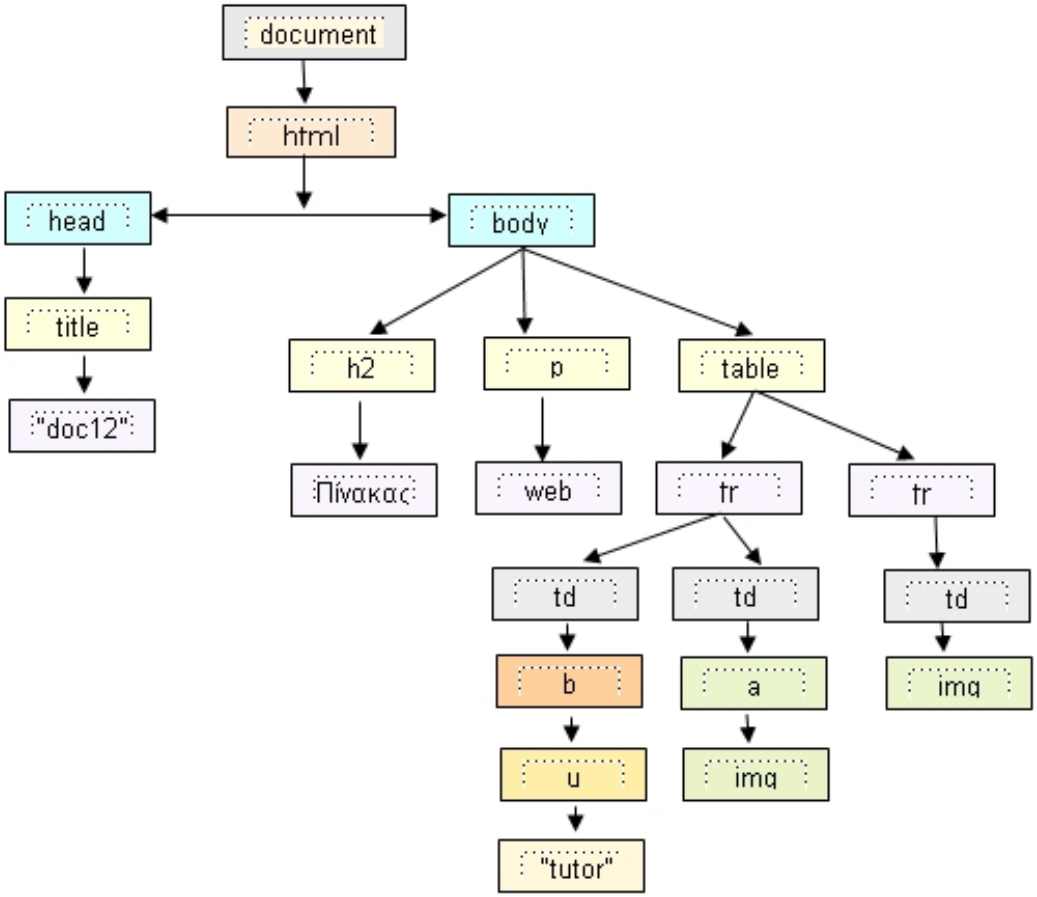
Циљеви XML-а

- Употреба на интернету
- Једнозначност приликом (аутоматске) обраде
- Лако се пишу програми за обраду XML докумената
- Читки и саморазумљиви
- Концизност није битна

XML скуп информација

- XML скуп информација (енг. XML info set) је концептуални модел XML докумената који је основа за друге W3C стандарде
- Елеменити модела су ставке (енг. information items) и везе између ставки
  - Постоји 11 типова ставки (документи, елементи, атрибути, простори имена, текст, коментари, процесне инструкције, референце на ентитете, декларације типова докумената, непарсирани ентитети и нотације)
  - Постоје неколико типова веза (родитељ, дете, предак, потомак, брат или сестра, атрибут, простор имена итд.
- XML скуп информација се може схватити као стабло, а ставке као чворови стабла

XML скуп информација



XML технологије

- XML
- XPath
- XML Namespaces
- XML Schema
- итд.

Структура XML документа

- XML документи имплементирају модел XML скупа информација
- Садрже чворове различитог типа уређене у структуру стабла
- Ознаке (енг. tag) су синтакси конструкт који служи да се хијерархијска структура стабла серијализују у линеарну структуру текстуланог документа (тј. низ знакова)

Типови чворова

- Документи
- Елементи
- Атрибути
- Простори имена
- Текст
- Коментари
- Процесне инструкција
- итд.

Имена елемената и атрибута

- Постоји разлика између великих и малих слова
- Могу садржати само слова, цифре, доњу црту (_), цртицу (-), двотачку (:) и тачку (.)
- Морају почети словом или доњом цртом (_)
- Не смеју почети низом слова "xml"

Елементи и ознаке

- Ознака (енг. tag) је текстуална ознака за почетак и крај елемента
- Садржај елемента налази се између почетне (отварајуће) и крајње (затварајуће) ознаке

Садржај елемената

- (неструктурирани) текст
- поделементи
- мешавина (неструктурираног) текста и поделемената
- без садржаја (празан садржај)

Садржај елемената

```
1  <foo>bar</foo>
2
3
4  <foo>
5    <bar>qux</bar>
6    <baz>qux</baz>
7  </foo>
8
9
10 <foo>
11  bar <baz>qux</baz> bar
12 </foo>
13
14
15 <foo></foo>
16 <foo />
17
```

Структурирање елемената

```
<contact>
  <name>Petar Petrovic</name>
  <address>Dunavska 1, Novi Sad</address>
  <telephone type="mobile">444-333</telephone>
</contact>



<contact>
  <name>
    <first>Petar</first>
    <last>Petrovic</last>
  </name>
  <address>
    <street>Dunavska</street>
    <number>1</number>
    <city>Novi Sad</city>
  </address>
  <telephone type="mobile">444-333</telephone>
</contact>

```

Атрибути

- Елемент може да има један или више атрибута
- Сваки атрибут има име и вредност
- Вредност атрибута је (неструктурирани) текст

Атрибути

```
1 <name first="Petar" last="Petrovic" />
2
```

Структурирање елемената

```
<name>
  <first>Petar</first>
  <last>Petrovic</last>
</name>


<name first="Petar" last="Petrovic" />

```

Коментари

- Нису намењене програмима који обрађују XML документе него људима који га читају
- Низ знакова између ”<!–” и ”–>”

Коментари

```
1 <!-- This is a comment. -->
2
```

Процесне инструкције

- Нису намењене људима који читају XML документе него програмима који га обрађују
- Низ знакова између ”<?” и ”?>”

Процесне инструкције

```
1 <?xml version="1.0" encoding="UTF-8" standalone="no" ?>
2
```

XML декларација

- Врста процесне иструкције (која је обавезна од XML 1.1)
- Ако постоји, мора да се налази на почетку XML документа
- Може да садржи три атрибута
  - "version" (верзија XML стандарда)
  - "encoding" (код текста (обично Unicode, тј. UTF-8))
  - "standalone" (да ли је могуће интерпретирати документ без других докумената (нпр. DTD, CSS итд.))

Референце на ентитете и ентитети

Референца на ентитет	Ентитет
&lt;	<
&amp;	&
&gt;	>
&quot;	"
&apos;	'

Table 1: Референце на ентитете и ентитети

CDATA секције

- Текст који се интерпретира без замене ентитета
- Низ знакова између ”<![CDATA[” и ”]]>”

CDATA секције

```
<p>You can use a default <code>xmlns</code> attribute to
avoid having to add the svg prefix to all your elements:</p>
<![CDATA[
<svg xmlns="http://www.w3.org/2000/svg" width="12cm" height="10cm">
  <ellipse rx="110" ry="130" />
  <rect x="4cm" y="1cm" width="3cm" height="6cm" />
</svg>
]]>

```

CDATA секције

```
<p>You can use a default <code>xmlns</code> attribute to
avoid having to add the svg prefix to all your elements:</p>
<svg xmlns="http://www.w3.org/2000/svg" width="12cm" height="10cm">
  <ellipse rx="110" ry="130" />
  <rect x="4cm" y="1cm" width="3cm" height="6cm" />
</svg>

```

Добро формиран XML документ

- XML документ је добро формиран ако задовољава скуп правила која омогућавају да се машински обради

Добро формиран документ

- Морају да се поштују правила за именовање елемената и атрибута
- Документ мора да има један и само један корени елемент
- Елементи не смеју да се преклапају
- Вредност атрибута мора да буде између једноструких или двоструких наводника
- Елемент не сме да има два атрибута са истим именом
- Коментари и процесне инструкције не смеју да се налазе унутар ознака
- Специјални знаци морају да се нађу само у својој улози

Правила за именовање елемената и атрибута

```
<1st>
  ...
</1st>

```

Један и само један корени елемент

```
<contact>
  ...
</contact>
<contact>
  ...
</contact>
```

Преклапање елемената

```
1 <contact>
2   <name>...</contact>
3 </name>
4
```

Вредност атрибута под наводницима

```
<telephone type=mobile>444-333</telephone>

```

Два атрибута са истим именом

```
<telephone type="mobile" type="work">444-333</telephone>

```

Коментари и процесне инструкције у ознакама

```
<contact>
  ...
</<!-- this is a comment -->contact>

<contact <?ignore?>>
  ...
</contact>

```

Специјални знаци

```
<contact>
  This is a special sign >
</contact>

```

Валидан XML документ

- XML документ је валидан ако је написан у складу са граматиком (нпр. која је специфицирана XML Schema-и)

Преглед садржаја

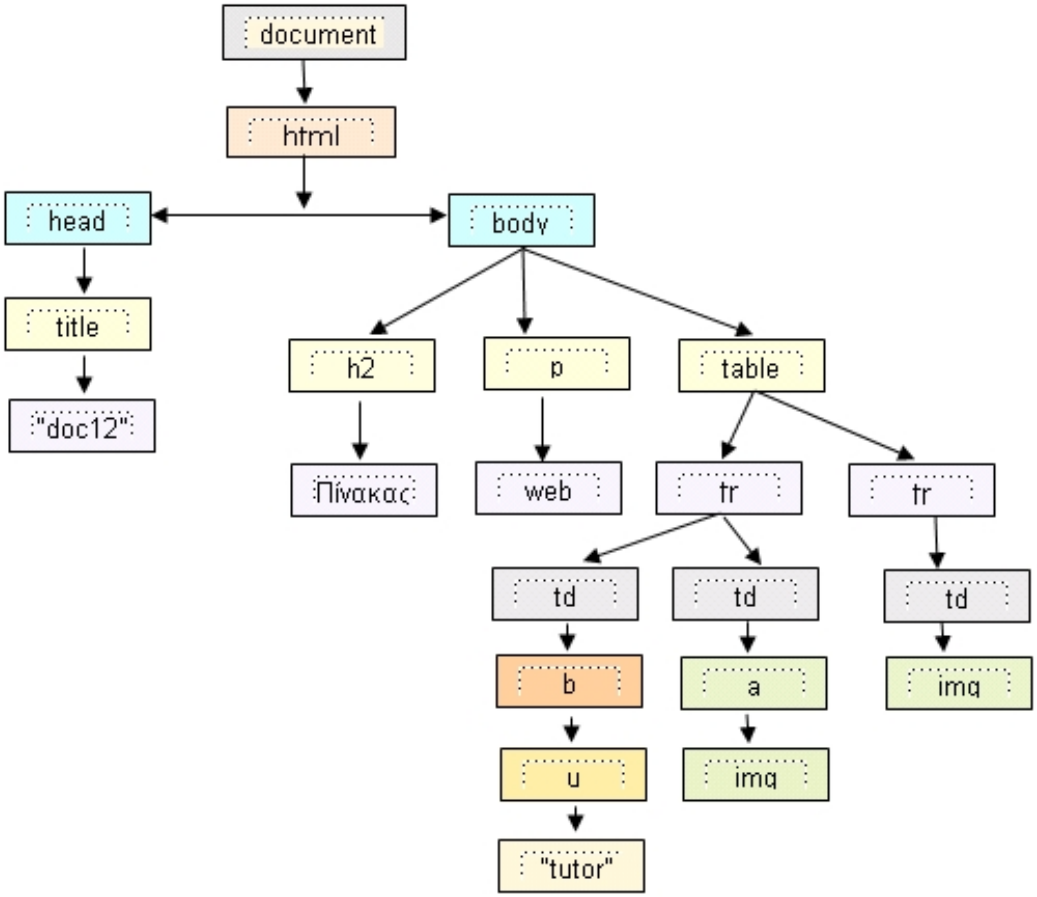
XPath

- XPath је упитни језик за XML документе
- Синтакса XPath израза подсећа на синтаксу путања у систему датотека
- Резултат XPath упита је (листа) чворова XML докумената

XPath израз

- XPath израз се састоји од једног или више корака (који су одвојени знаком "/")
- XPath изрази могу да буду апсолутни (евалуирају се у односу на документ чвор и почињу са знаком "/") и релативни (евалуирају се у односу на текући чвор и не почињу са знаком "/")

HTML документ



XPath израз

```
1 /html/head/title
2
3 head/@title
4
```

Корак XPath израза

- Елементи једног корака XPath израза су оса, тест чвора и предикат

Корак XPath израза

```
1 axis::node_test[predicate]
2
```

XPath оса

- Оса дефинише правац кратања у XML документу (тј. стаблу XML чворова)

XPath оса

Оса	Опис
ancestor	сви преци
ancestor-or-self	сви преци или сам чвор
attribute	сви атрибути
child	сва деца
descendant	сви потомци
descendant-or-self	сви потомци или сам чвор
following	сви чворови после текућег чвора

Table 2: XPath оса

XPath оса

Оса	Опис
following-sibling	сва браћа и сестре после текућег чвора
namespace	сви простори имена
parent	родитељ
preceding	сви чворови пре текућег чвора
preceding-sibling	сва браћа и сестре пре текућег чвора
self	сам чвор

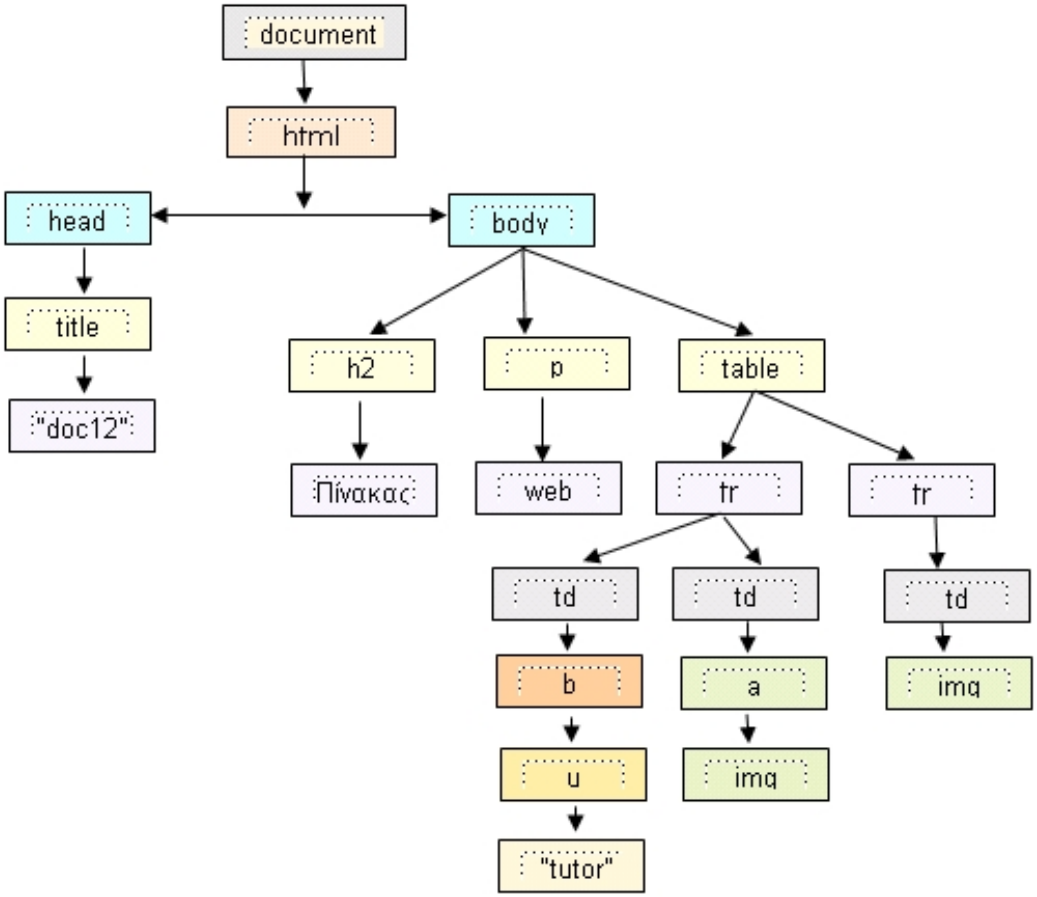
Table 3: XPath оса

XPath оса

Оса	Скраћени облик
child::	подразумевана оса
self::	.
parent::	..
descendant::	//
attribute::	@
following-sibling::	../

Table 4: XPath оса

HTML документ



XPath оса

```
1 html
2 /html
3 ./p
4 ..
5 //p
6 /html/head/@title
7 @title
8 ../
9
```

XPath тест чвора

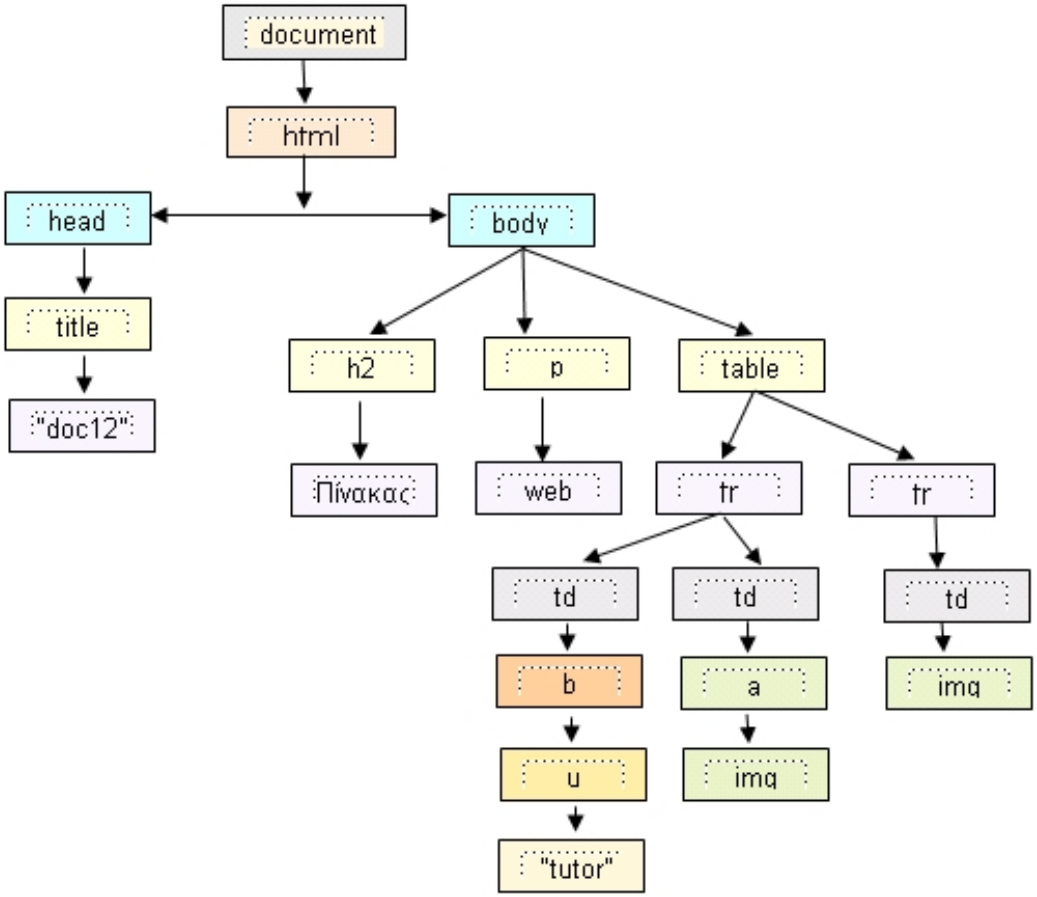
- Тест чвора дефинише назив или тип чвора

XPath тест чвора

Пример	Опис
html	име чвора
text()	тип чвора
*	било који чвор

Table 5: XPath тест чвора

HTML документ



XPath тест чвора

```
1 /html/p
2 /html/comment()
3 /html/*
4
```

Предикат

- Предикат дефинише (логички) услов за филтрирање резултата упита

XPath оператори

Пример	Опис	Пример
|	унија	
+	сабирање	x + y
-	одузимање	x - y
*	множење	x * y
div	дељење	x div y
mod	остатак при дељењу	x mod y

Table 6: XPath оператори

XPath оператори

Пример	Опис	Пример
=	једнако	x = y
!=	различито	x != y
<	мање	x < 0
<=	мање или једнако	x <= 0
>	веће	x > 0
>=	веће или једнако	x >= 0
or	дисјункција	x < 0 or x > 10
and	конјункција	0 < x and x < 10

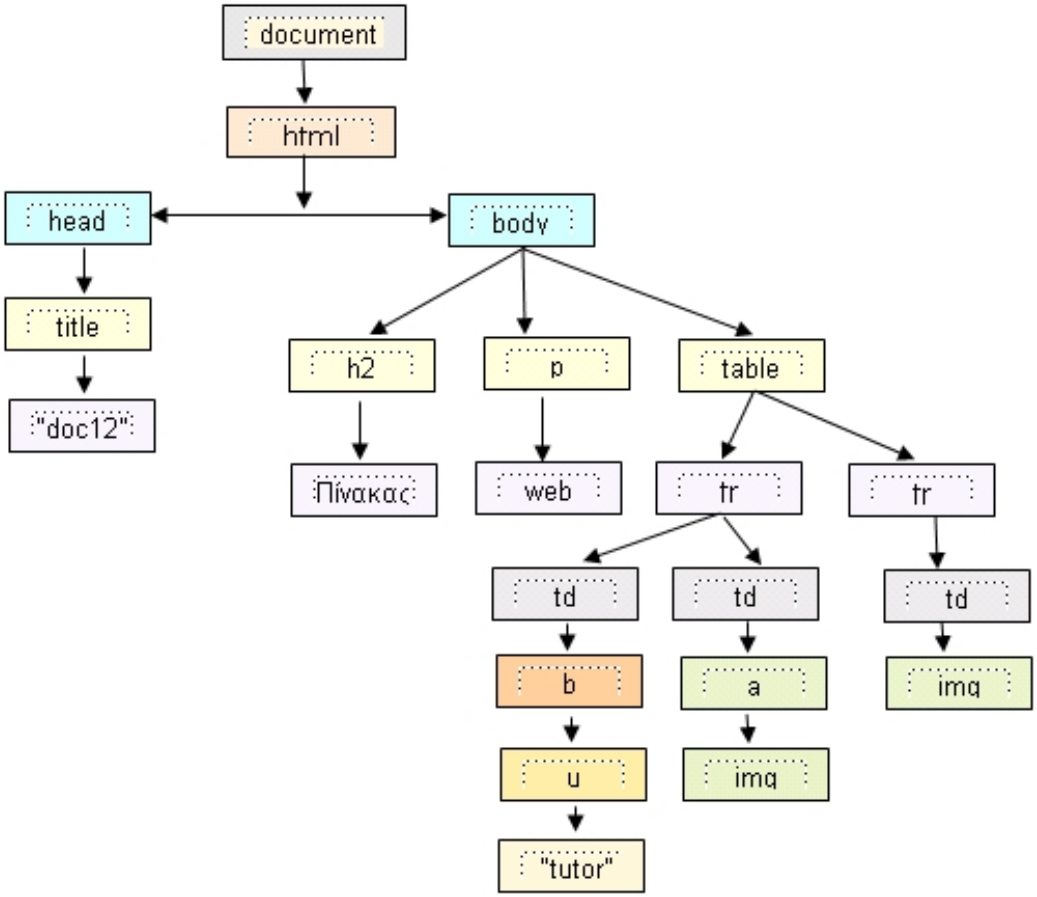
Table 7: XPath оператори

XPath функције

Функција	Функција	Функција
boolean()	id()	starts-with()
ceiling()	lang()	string()
choose()	last()	string-length()
concat()	local-name()	string-length()
contains()	name()	substring()
count()	namespace-uri()	substring-after()
element-available()	normalize-space()	substring-before()
false()	not()	sum()
floor()	position()	translate()
function-available()	round()	true()

Table 8: XPath функције

XPath предикат



XPath предикат

```
1 /html/head[@title]
2 /html/head[@title = "doc12"]
3 /html/p[position() = 1]
4 /html/p[1]
5 /html/p[last()]
6
```

Преглед садржаја

XML простори имена

- XML простори имена (енг. XML namespaces) су скупови елемената, атрибута и типова података који су идентификовани са истим идентификатором
- Решавају проблем конфликта у именима елемената, атрибута и типова података

XML простори имена

```
<table>
  <tr>
    <td>Apples</td>
    <td>Bananas</td>
  </tr>
</table>

```

XML простори имена

```
1 <table>
2   <name>African Coffee Table</name>
3   <width>80</width>
4   <length>120</length>
5 </table>
6
```

XML простори имена

```
1 <h:table xmlns:h="http://www.w3.org/TR/html4/">
2   <h:tr>
3     <h:td>Apples</h:td>
4     <h:td>Bananas</h:td>
5   </h:tr>
6 </h:table>
7
```

XML простори имена

```
<f:table xmlns:f="https://www.w3schools.com/furniture">
  <f:name>African Coffee Table</f:name>
  <f:width>80</f:width>
  <f:length>120</f:length>
</f:table>

```

URL

- URI (Uniform Resource Identifier) је низ знакова који идентификује апстрактне или физичке ресурсе (RFC 2396)
- URL (Uniform Resource Locator) је подскуп URI који идентификује ресурсе преко примарног механизма приступа (нпр. преко локације на мрежи)
  - http://www.ftn.uns.ac.rs/
  - ftp://ftp.is.co.za/rfc/rfc1808.txt
  - mailto:mduerst@ifi.unizh.ch

XML простори имена

- XML простори имена идентификују се са URL
- URL не мора да буде добро формирано XML име
- Имена XML елемената, атрибута и типова података се квалификују са префиксом, а префикс се мапира на URL коришћењем ”xmlns” конструкта
- Један префикс може да буде мапиран на више URL и више префикса могу да буду мапирани на један URL

XML простори имена

```
<f:table xmlns:f="https://www.w3schools.com/furniture">
  <f:name>African Coffee Table</f:name>
  <f:width>80</f:width>
  <f:length>120</f:length>
</f:table>

```

Квалификовано име

- Квалификовано име XML елемента, атрибута или типа података састоји се од идентификатора простора имена и локалног имена

Подразумевани XML простори имена

- Подразумевани XML простор имена (енг. default XML namespace) је простор имена коме припадају елементати, атрибути и типови података који нису квалификовани са префиксом

XML простори имена

```
1 <table xmlns="https://www.w3schools.com/furniture">
2   <name>African Coffee Table</name>
3   <width>80</width>
4   <length>120</length>
5 </table>
6
```

XML простори имена

```
<?xml version="1.0" encoding="utf-8"?>
<pers:person
  xmlns:pers="http://www.ftn.ns.ac.yu/dtds/person.dtd"
  xmlns:xhtml="http://www.w3.org/1999/xhtml">
    <pers:name>
    <pers:title>Sir</pers:title>
    <pers:first>John</pers:first>
    <pers:last>Doe</pers:last>
  </pers:name>
  <pers:position>VP of Marketing</pers:position>
  <pers:resume>
    <xhtml:html>
      <xhtml:head>
        <xhtml:title>Resume of John Doe</xhtml:title>
      </xhtml:head>
      <xhtml:body>
        <xhtml:h1>John Doe</xhtml:h1>
        <xhtml:p>John's a great guy, you know?</xhtml:p>
      </xhtml:body>
    </xhtml:html>
  </pers:resume>
</pers:person>

```

XML простори имена

```
<?xml version="1.0" encoding="utf-8"?>
<pers:person
  xmlns:pers="http://www.ftn.ns.ac.yu/dtds/person.dtd">
    <pers:name>
    <pers:title>Sir</pers:title>
    <pers:first>John</pers:first>
    <pers:last>Doe</pers:last>
  </pers:name>
  <pers:position>VP of Marketing</pers:position>
  <pers:resume>
    <xhtml:html xmlns:xhtml="http://www.w3.org/1999/xhtml">
      <xhtml:head>
        <xhtml:title>Resume of John Doe</xhtml:title>
      </xhtml:head>
      <xhtml:body>
        <xhtml:h1>John Doe</xhtml:h1>
        <xhtml:p>John's a great guy, you know?</xhtml:p>
      </xhtml:body>
    </xhtml:html>
  </pers:resume>
</pers:person>

```

XML простори имена

```
1  <?xml version="1.0" encoding="utf-8"?>
2  <person xmlns="http://www.ftn.ns.ac.yu/dtds/person.dtd">
3    <name>
4      <title>Sir</title>
5      <first>John</first>
6      <last>Doe</last>
7    </name>
8    <position>VP of Marketing</position>
9    <resume>
10     <xhtml:html xmlns:xhtml="http://www.w3.org/1999/xhtml">
11       <xhtml:head>
12         <xhtml:title>Resume of John Doe</xhtml:title>
13       </xhtml:head>
14       <xhtml:body>
15         <xhtml:h1>John Doe</xhtml:h1>
16         <xhtml:p>John's a great guy, you know?</xhtml:p>
17       </xhtml:body>
18     </xhtml:html>
19   </resume>
20 </person>
21
```

XML простори имена

```
1  <?xml version="1.0" encoding="utf-8"?>
2  <person xmlns="http://www.ftn.ns.ac.yu/dtds/person.dtd">
3    <name>
4      <title>Sir</title>
5      <first>John</first>
6      <last>Doe</last>
7    </name>
8    <position>VP of Marketing</position>
9    <resume>
10     <html xmlns="http://www.w3.org/1999/xhtml">
11       <head>
12         <title>Resume of John Doe</title>
13       </head>
14       <body>
15         <h1>John Doe</h1>
16         <p>John's a great guy, you know?</p>
17       </body>
18     </html>
19   </resume>
20 </person>
21
```

Преглед садржаја

Типови XML докумената

- Тип XML документа чини скуп XML докумената са сличним особинама
- Тип XML документа је одређен речником (скупом елемената и атрибута) и граматиком (скупом правила за структурирање документа од елемената и атрибута)
- За дефинисање типова XML докумената могу се користити различити језици, као што су DTD и XML Schema

DTD

- Document Type Definition (DTD) је део XML стандарда који се користи за дефинисање типова XML докумената
- DTD омогућава декларацију елемената, (листа) атрибута, ентитета, нотација итд.
- Постоје многе мане DTD-а (као што су не-XML синтакса, недостатак подршке за просторе имена, недостатак типизације података итд.)

XML Schema

- XML Schema је стандард који се користи за дефинисање типова XML докумената
- Превазилази неке мане DTD-а
- Уводи скуп стандардизованих простих типова података и моделе садржаја
- XML Schema омогућава декларацију елемената и атрибута, дефиницију простих и сложених типова података итд.

Преглед садржаја

Закључак

- XML
- XML информациони скуп
- добро формиран документ
- валидан документ
- простор имена
- префикс
- подразумевани простор имена

Закључак

- XPath
- XPath израз
- XPath оса
- XPath тест чвора
- XPath предикат

Литература

- Fawcett, J., Ayers, D. and Quin, L.R.E. (2012) "Beginning XML". Hoboken, NJ, USA: Wiley.

Хвала на пажњи!