

# XML

## Лекција 3

Стеван Гостојић

Факултет техничких наука, Нови Сад

7. март 2024.

# Преглед садржаја

- 1 Увод
- 2 eXtensible Markup Language
- 3 XML Namespaces
- 4 XPath
- 5 Закључак

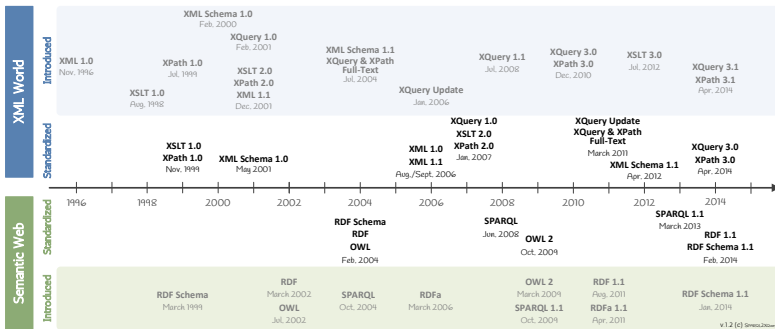
# W3C

- 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



This work is available under a CC BY-SA license. This means you can use it freely (subject to the conditions that you give proper attribution, please do not: distort it, remove it, or otherwise, share it without), and you can share it too. The W3C and Semantic 2004 (October) Technology, Interoperability and Integration: A Survey of the State of the Art in Semantic Web (W3C) (October) Technology, Interoperability and Integration, Springer 2014.

# 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 ресурси 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

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <Order>
3   <BuyersID>GMB91604</BuyersID>
4   <BuyerParty>
5     <ID>KEEN</ID>
6     <PartyName>
7       <Name>Maynard James Keenan</Name>
8     </PartyName>
9     <Address>
10       <Room>505</Room>
11       <BuildingNumber>11271</BuildingNumber>
12       <StreetName>Ventura Blvd.</StreetName>
13       <CityName>Studio City</CityName>
14       <PostalZone>91604</PostalZone>
15       <CountrySubentity>California</CountrySubentity>
16       <Country>USA</Country>
17     </Address>
18   </BuyerParty>
19   <OrderLine>
20     <LineItem>
21       <BookItem>
22         <Title>Document Engineering</Title>
23         <Author>Glushko and McGrath</Author>
24         <ISBN>0262072610</ISBN>
25       </BookItem>
26       <BasePrice>99.95</BasePrice>
27       <Quantity>300</Quantity>
28     </LineItem>
29   </OrderLine>
30 </Order>
31

```

# 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

- 1 Увод
- 2 eXtensible Markup Language
- 3 XML Namespaces
- 4 XPath
- 5 Закључак

# 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

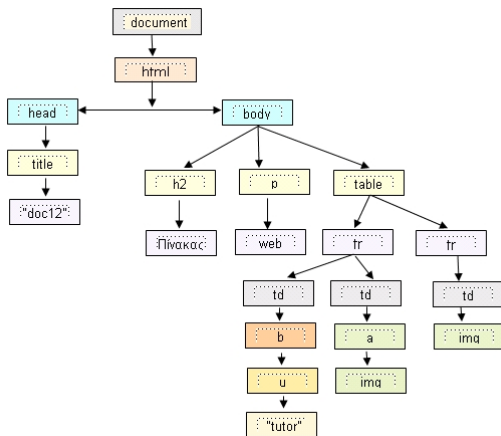
# HTML документ

```

1 <!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
2 <html>
3   <head>
4     <title>doc12</title>
5   </head>
6   <body>
7     <h2>Table</h2>
8     <p>web</p>
9     <table width="100%">
10       <tr>
11         <td><b>
12           <u>tutor</u>
13         </b>
14       </td>
15       <td><a href="starweb.html">
16         
17       </a>
18     </td>
19   </tr>
20   <tr>
21     <td>
22   </td>
23 </tr>
24 </table>
25 </body>
26 </html>
27

```

## HTML документ



# XML

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



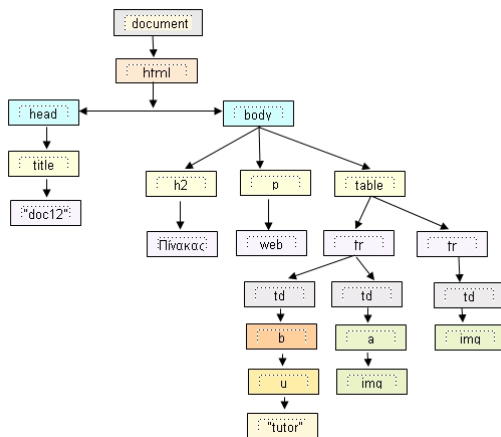
# Циљеви XML-а

- 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
- Елементи 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
- XML Namespaces
- XPath
- 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 серијализује 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

```
1 <contact>
2   <name>Petar Petrovic</name>
3   <address>Dunavska 1, Novi Sad</address>
4   <telephone type="mobile">444-333</telephone>
5 </contact>
6
7
8
9 <contact>
10  <name>
11    <first>Petar</first>
12    <last>Petrovic</last>
13  </name>
14  <address>
15    <street>Dunavska</street>
16    <number>1</number>
17    <city>Novi Sad</city>
18  </address>
19  <telephone type="mobile">444-333</telephone>
20 </contact>
21
```

# 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

```
1 <name>
2   <first>Petar</first>
3   <last>Petrovic</last>
4 </name>
5
6
7
8 <name first="Petar" last="Petrovic" />
9
```

# 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 инструкције (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

```
1 <p>You can use a default <code>xmlns</code> attribute to
2 avoid having to add the svg prefix to all your elements:</p>
3 <![CDATA[
4 <svg xmlns="http://www.w3.org/2000/svg" width="12cm" height="10cm">
5   <ellipse rx="110" ry="130" />
6   <rect x="4cm" y="1cm" width="3cm" height="6cm" />
7 </svg>
8 ]]>
9
```

# CDATA секције

```
1 <p>You can use a default <code>xmlns</code> attribute to
2 avoid having to add the svg prefix to all your elements:</p>
3 <svg xmlns="http://www.w3.org/2000/svg" width="12cm" height="10cm">
4   <ellipse rx="110" ry="130" />
5   <rect x="4cm" y="1cm" width="3cm" height="6cm" />
6 </svg>
7
```

# 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

```
1 <1st>  
2   ...  
3 </1st>  
4
```

# Један и само један корени елемент

```
1 <contact>
2   ...
3 </contact>
4 <contact>
5   ...
6 </contact>
7
```

# Преклапање елемената

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

# Вредност атрибута под наводницима

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

```
2
```

# Два атрибута са истим именом

```
1 <telephone type="mobile" type="work">444-333</telephone>  
2
```

# Коментари и процесне инструкције у ознакама

```
1 <contact>
2   ...
3 </<!-- this is a comment -->contact>
4
5 <contact <?ignore?>>
6   ...
7 </contact>
8
```

# Специјални знаци

```
1 <contact>
2   This is a special sign >
3 </contact>
4
```



# Валидан XML документ

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

# Преглед садржаја

- 1 Увод
- 2 eXtensible Markup Language
- 3 XML Namespaces**
- 4 XPath
- 5 Закључак

# 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

```
1 <table>
2   <tr>
3     <td>Apples</td>
4     <td>Bananas</td>
5   </tr>
6 </table>
7
```

# 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

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

# 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 корићењем "xmlns" конструкта
- Један префикс може да буде мапиран на више URL и више префикса могу да буду мапирани на један URL

# XML простори имена

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

# 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 елементи, 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

```

1 <?xml version="1.0" encoding="utf-8"?>
2 <pers:person
3   xmlns:pers="http://www.ftn.ns.ac.yu/dtds/person.dtd"
4   xmlns:xhtml="http://www.w3.org/1999/xhtml">
5   <pers:name>
6   <pers:title>Sir</pers:title>
7   <pers:first>John</pers:first>
8   <pers:last>Doe</pers:last>
9 </pers:name>
10 <pers:position>VP of Marketing</pers:position>
11 <pers:resume>
12   <xhtml:html>
13     <xhtml:head>
14       <xhtml:title>Resume of John Doe</xhtml:title>
15     </xhtml:head>
16     <xhtml:body>
17       <xhtml:h1>John Doe</xhtml:h1>
18       <xhtml:p>John's a great guy, you know?</xhtml:p>
19     </xhtml:body>
20   </xhtml:html>
21 </pers:resume>
22 </pers:person>
23

```

# XML простори имена

```

1 <?xml version="1.0" encoding="utf-8"?>
2 <pers:person
3   xmlns:pers="http://www.ftn.ns.ac.yu/dtds/person.dtd">
4   <pers:name>
5   <pers:title>Sir</pers:title>
6   <pers:first>John</pers:first>
7   <pers:last>Doe</pers:last>
8   </pers:name>
9   <pers:position>VP of Marketing</pers:position>
10  <pers:resume>
11    <xhtml:html xmlns:xhtml="http://www.w3.org/1999/xhtml">
12      <xhtml:head>
13        <xhtml:title>Resume of John Doe</xhtml:title>
14      </xhtml:head>
15      <xhtml:body>
16        <xhtml:h1>John Doe</xhtml:h1>
17        <xhtml:p>John's a great guy, you know?</xhtml:p>
18      </xhtml:body>
19    </xhtml:html>
20  </pers:resume>
21 </pers:person>
22

```

# 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

- 1 Увод
- 2 eXtensible Markup Language
- 3 XML Namespaces
- 4 XPath**
- 5 Закључак

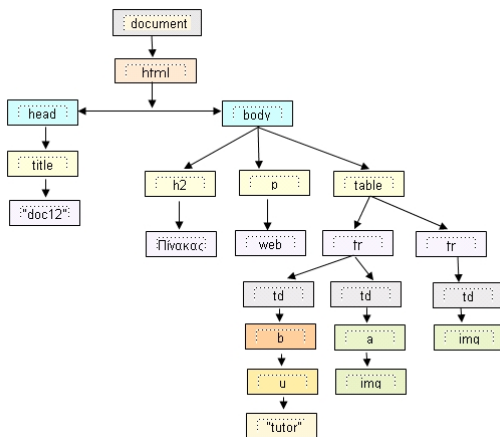
# XPath

- 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
```

# 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 oca

| 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 oca

## XPath oca

| Oca               | 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                               |

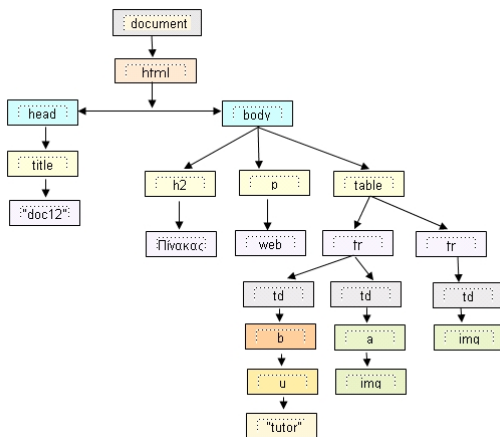
Table 3: XPath oca

## XPath oca

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

Table 4: XPath oca

## HTML документ



# XPath oca

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

# XPath тест чвора

- Тест чвора дефинише назив или тип чвора

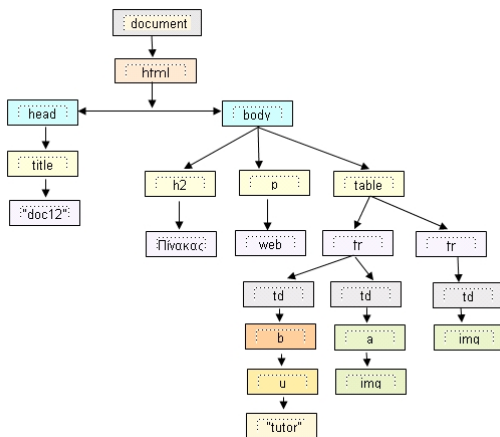
# XPath тест чвора

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

Table 5: XPath тест чвора



## HTML документ



# XPath тест чвора

```
1 /html/body/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            | $x \text{ div } y$ |
| mod    | остатак<br>дељењу | $x \text{ mod } y$ |

Table 6: XPath оператори

# XPath оператори

| Пример | Опис             | Пример                      |
|--------|------------------|-----------------------------|
| =      | једнако          | $x = y$                     |
| !=     | различито        | $x \neq y$                  |
| <      | мање             | $x < 0$                     |
| <=     | мање или једнако | $x \leq 0$                  |
| >      | веће             | $x > 0$                     |
| >=     | веће или једнако | $x \geq 0$                  |
| or     | дисјункција      | $x < 0 \text{ or } x > 10$  |
| and    | конјункција      | $0 < x \text{ and } x < 10$ |

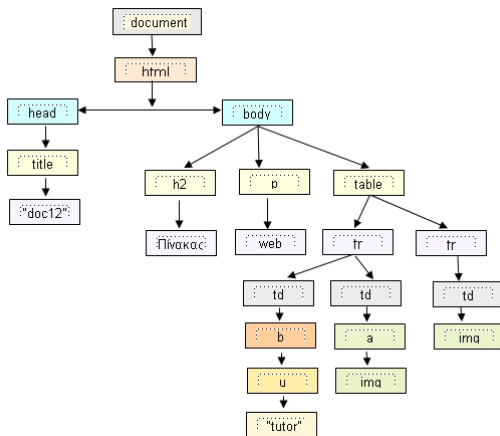
Table 7: XPath оператори

# 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 тест чвора

```
1 /html/body/p[position() = 1]
2 /html/body/p[1]
3 /html/body/p[last()]
4 count(/html/body/p)
5
```



# Преглед садржаја

- 1 Увод
- 2 eXtensible Markup Language
- 3 XML Namespaces
- 4 XPath
- 5 Закључак**

# Закључак

- XML
- XML информациони скуп
- добро формиран документ
- валидан документ

# Закључак

- 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.
  - Chapter 1: What is XML?
  - Chapter 2: Well-Formed XML
  - Chapter 3: XML Namespaces
  - Chapter 7: Extracting Data From XML

# Литература

- Extensible Markup Language (XML) 1.0 (Fifth Edition),  
<https://www.w3.org/TR/xml/>

# Хвала на пажњи!