

# XML

## Extensible Markup Language

Jedinečný **zdroj** obrovského množstva **dát** **zobrazované** a spracovávané v štandarde **HTML**

**SIEMENS**

I Home I Kontakt

Vyberte si divízi:

Hľadaj

Produkty  
Priebeh  
Všetci pôsobíme  
v oblastiach  
em kvality  
neri  
Referencie Award  
Získali  
meho sveta

Spoľnosť E-biznis Press Práca

> Siemens vo  
online shop

Siemens. V znamení inovácie.

Novinky  
**Siemens Artelios je najmenší a najľahší ultrazvukový systém telenoviny**  
16.10.2007 10:56:09 vlnac++  
**OŠADIA Slovensko, a.s. otvorí novú huta na výrobe automobilových zdieľok**  
17.10.2007 10:54:31 vlnac++  
**Siemens reaguje na divíziu Information and Communication Networks**  
16.10.2007 11:56:31

Vyhraj

Pracovné príležitosti

Zahraničné

Ing. Roman Filkorn 2

Dáta **strácajú:**

- [illegible]

Ing. Roman Filkorn 3

Nástroj na **zobrazovanie** (ľubovoľných) informácií  
(s množstvom necností pri práci s nimi)

Nástroj na **zdieľanie** ľubovoľných informácií  
s požadovanými vlastnosťami

- XML

Inq. Roman Filkorn 4

## Fenomén XML (Extensible Markup Language)

Hlavná úloha: **opis dát**

Pridaním meta-infromácie

Textová forma: **značky** (markup, tags)

Metajazyk: jazyk **na definovanie** jazykov

Definovať vlastnú množinu značiek

Definovať ich hierarchiu, ohraničenia a pod.

= definícia typu dokumentu, **DTD**  
**Document Type Definition**

## Príklad

```
<note>
<to>students</to>
<from>lector</from>
<heading>My 1. question</heading>
<body>Is it simple enough?</body>
</note>
```

len strohá informácia - **dáta a štruktúra**

**žiadne** ako zobrazit', čo robiť a pod.

## XML je doplnok k HTML

HTML - **zobrazovanie** informácií

XML:

**popis a prenos** (~ zdieľanie) informácií

hardvérovo aj softvérovo **nezávislý nástroj**

## XML oddeľuje dáta od zobrazenia

Logické rozdelenie - **dáta** (XML) a ich **zobrazenie** (HTML)

- jednoduchosť a prehľadnosť
- manipulovateľnosť

- modularita  
vzor MVC: **model** a **pohľad**
- znovupoužiteľnosť

## XML môže slúžiť pri uchovávaní dát

platformová nezávislosť

fyzické ukladanie dát - textové súbory, databázy, ...

## XML môže slúžiť pri zdieľaní dát

samopopisné dokumenty - XML a DTD

textový, platformovo nezávislý formát

## XML slúži na prenos a výmenu dát

XML - všeobecné pravidlá formátovania a opisu dát

- softvérovo nezávislé
- hardvérovo nezávislé
- jednoducho spracovateľné človekom aj strojom

Výmena informácií - základ informačného veku

- Business To Business (B2B), vedecké aj technické dáta, ...

## XML robí dáta použiteľnejšími

- jednoduché pre človeka aj stroj
- striktné oddelenie dát od zobrazenia
- platformová nezávislosť (hardvér aj softvér)  
uchovávanie, zdieľanie

- jednoduchšie spracovávanie  
(triedenie, filtrovanie, zmena štruktúry, ...)
- rôzne možnosti zobrazenia tých istých dát  
(HTML, PDF, SVG, VRML / X3D, ...)
- širší okruh odoberateľov dát  
(rôzne typy klientov, vyhľadávacie centrály, agenty, ...)

## XML - free and extensible

odporúčanie vo vlastníctve WWW Consortium (W3C)

XML je metajazyk:

- definuje pravidlá **ako** reprezentovať dáta  
ako definovať množinu značiek, ktoré určujú  
význam (~interpretáciu) a hierarchiu dát

### HTML

- jazyk rodiny XML (XHTML)
- definovaná množina značiek a ich sémantika  
(~ zobrazovanie)

## XML - nástroj na tvorbu nových jazykov

definovanie množiny **značiek**, **hierarchie** a **sémantiky**

- WML (Wireless Markup Language)
- XSL (Extensible Stylesheet Language)
- XSLT (XSL Transformations)
- XML Schema
- XML Query
- SOAP (Simple Object Access Protocol)
- X3D, SVG (Scalable Vector Graphics)
- ... a mnohé ďalšie

## XML - flexibilný formát súčasnosti

potomok jednoduchosti zápisu a popularity HTML

jednoduché a škálovateľné využitie

zamerané priamo na dáta

prakticky neobmedzené možnosti nasadenia

*„XML - nástroj informačného veku“*

## XML - plán

1. časť, základy  
názvoslovie a syntax XML  
validácia pomocou DTD  
techniky spracovania
2. časť, spracovanie  
XPath  
XSLT
3. časť, prehľad použití

## 1. časť - základné pojmy a pravidlá XML

- dokument, tag
- element, atribút
- kontrola správnosti, platnosti
- manuálne cvičenia...

## Príklad

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

<note>
<to>students</to>
<from>lector</from>
<heading>My 1. question</heading>
<body>Is it simple enough?</body>
</note>
```

jednoduchá, samopopisná syntax

## Základné pojmy

### Dokument

dáta (text) + doplnkové informácie (markup, ...)

- hlavička XML
- (jediný!) koreň dokumentu (document root)

### Element

značka - tag

- obsahujú ďalšie elementy alebo textové reťazce

### Dáta - textové reťazce

## Značka v XML musí byť ukončená

### HTML

```
<p>
XML Basics:
<ul>
<li>document
<li>element
</ul>
<p>
Following text...
```

### XML

```
<p>
XML Basics:
<ul>
<li>document</li>
<li>element</li>
</ul>
</p>
<p>
Following text...
```

*Poznámka:* výnimky?

## Značky v XML sa rozlišujú veľkosťou znakov (case sensitive)

### HTML

```
<HTML>
<Head>
<TITLE>Welcome</Title>
</HEAD>
<BODY>
Simple HTML page...
</BODY>
</html>
```

### XML

```
<?xml version="1.0"?>
<html>
<Head>
<TITLE>Welcome</TITLE>
</Head>
<BODY>
Simple XML data...
</BODY>
</html>
```

*Poznámka:* vhodné pravidlá...

## Značky v XML musia byť správne “hniezdené” (definujú štruktúru)

### HTML

```
<B><I>Text in bold and italic</B></I>
```

### XML

```
<B><I>Text in B and in I</I></B>
```

*Poznámka:* nevinné HTML

## Hodnoty atribútov musia byť vždy odlišené (úvodzovky, apostrofy)

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

```
<note type='simple' date="2001-10">  
  <to>students</to>  
  <from>lector</from>
```

```
<heading>My 1. question</heading>
```

```
<body>Is it simple enough?</body>  
</note>
```

HTML <TABLE BORDER=1 VALIGN=CENTER>

## “Biele” znaky sú v XML významné

white space - medzera, tabulátor, nový riadok  
súčasť dát (HTML pri zobrazení upravuje)

### Nový riadok - len LF (linefeed)

prežitok písacích strojov a ASCII tlačiarňí  
CR/LF konvertované na LF

## XML ťahák (základná verzia)

Základná syntax XML je veľmi jednoduchá a striktná

Dokument: hlavička a jediný koreňový element  
(element môže obsahovať iné elementy, atribúty, textové dáta)

Element musí byť jednoznačne vyznačený

Značka v XML musí byť ukončená

Značky v XML musia byť správne “hniezdené” (definujú štruktúru)

Značky v XML sa rozlišujú veľkosťou znakov (case sensitive)

Hodnoty atribútov musia byť vždy odlišené (“ ”, ‘ ’)

“Biele” znaky sú v XML významné

## XML ťahák (príklad)

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

<note type='simple' date="2001-10">
  <to>students</to>
  <from>lector</from>

  <heading>My 1. question</heading>

  <body>Is it simple enough?</body>
</note>
```

## Viac o elementoch...

základný prvok - element  
rozšíriteľný  
v kontexte - vzťahoch

pravidlá pre pomenovávanie

## Viac o atribútoch...

deklarácie atribútov  
porovnanie elementov a atribútov

## Elementy sú jednoducho rozšíriteľné

obsah elementu - textové dáta a iné elementy

rozšírenie - pridanie ďalšieho dieťaťa

- jednoduché pridanie do textu
- principiálne neovplyvní doterajšie spracovanie

## Elementy a ich vzťahy

teória grafov

- koreň, uzol, list
- rodič, dieťa, súrodenec, potomok

```
<root>
  <element>
    <childElement>any data</childElement>
    <anotherChild>
      <grandchild>foo</grandchild>
      <grandchild>bar</grandchild>
    </anotherChild>
  </element>
</root>
```

dokument XML - **strom**

## Obsah elementov

všetko od začiatocnej po konečnú značku (vrátane)

- jednoduchý obsah (simple content)
- iba ďalšie elementy (element content)
- zmiešaný obsah (mixed content)  
`<elm>text <sub>+ element</sub> content</elm>`
- prázdne (empty content)  
skrátенý zápis: `<list_stromu/>`

atribúty (dvojica meno - hodnota)

## Pravidlá pomenovávania elementov

- môžu obsahovať písmená, číslice, (iné znaky)
- musia začínať písmenom
- nesmú začínať písmenami xml (XML,Xml,xmI, ...)
- nesmú obsahovať biele znaky

špeciálny význam má ':', nepoužívať '.'

ľubovoľný jazyk: 'poznámka', 'onô', ...

(ale pozor na platformu / softvér)

## Elementy - zhrnutie

- širšie možnosti pomenovávania
- kontext vzťahov medzi nimi (uzly stromu)
- variabilný obsah (dáta, iné elementy)
- jednoducho rozšíriteľné

## Atribúty

dvojica **meno - hodnota**

prípustný len jeden atribút s daným menom

hodnota je textový reťazec

nedefinovaná štruktúra (nič: deti, čiarkou oddelené)

## Atribúty

informácia: • doplňujúca dáta  
• popisujúca dáta

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

<note type='simple' date="2001-10">
  <to>students</to>
  <from>lector</from>

  <heading>My 1. question</heading>

  <body>Is it simple enough?</body>
</note>
```

## Hodnota atribútu povinne odlišená

musia byť použité úvodzovky alebo apostrof

```
<note type='simple' date="2001-10">
  ...
</note>

<gangster name='George "Shotgun" Ziegler' />
```

## Atribúty vs. Elementy

dáta - v elementoch a (alebo) v atribútoch

```
<osoba pohlavie="muž">
  <meno>Roman</meno>
  <priezvisko>Filkorn</priezvisko>
</osoba>
```

alebo

```
<osoba>
  <pohlavie>muž</pohlavie>
  <meno krstne="Roman" priezvisko="Filkorn" />
</osoba>
```

... alebo ešte inak?

## Atribúty vs. Elementy (pokr.)

dátový návrh

- reprezenácia dát
- vnútorná štruktúra (komplexnosť) dát
- kardinalita dát
- vzťahy medzi dátami
- kardinalita vzťahov

spracovanie?

## Atribúty vs. Elementy (pokr.)

nevýhody atribútov:

- nemôžu obsahovať viacero hodnôt
- nedajú sa v budúcnosti jednoducho rozšíriť
- nemôžu popisovať štruktúru

odporúčanie:

**dáta - elementy, metadáta - atribúty**

(ID, typ, ...)

## Správny dokument XML...

...dodržuje syntax XML

“well formed”

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

<note type='simple' date="2001-10">
  <to>students</to>
  <from>lector</from>

  <heading>My 1. question</heading>

  <body>Is it simple enough?</body>
</note>
```

## Platný XML dokument...

... je “well formed” a patrí do triedy dokumentov definovaných (nejako zapísanými) pravidlami

je “valid” voči DTD / XML Schema / ...

```
<?xml version="1.0"?>
<!DOCTYPE note SYSTEM "Note.dtd">
<note>
  <to>students</to>
  <from>lector</from>
  <heading>validation</heading>
  <body>validated against Note.dtd</body>
</note>
```

## Document Type Definition (DTD)

pre elementy v dokumente definuje:

- názvy
- atribúty
- obsah (prázdny, text, hierarchické vzťahy)

norma pre SGML, neXML syntax

## XML Schema

rozširuje možnosti DTD (kontrola typov, ...)

XML dokument (definovaná sémantika, W3C)

## Spracovanie XML dokumentov

človek

- vytvorenie dokumentu
- zobrazenie
- odladenie ("well formed", "valid")

stroj

- parser
- API k dokumentu

## Vytvorenie XML dokumentu

bežná textová editácia

Notepad, Ultraedit, ...

editory s podporou XML

XEddy, Xeena, JBuilder...

## Zobrazenie XML dokumentu

zlepšujúca sa podpora v súčasných (HTML) prehliadačoch

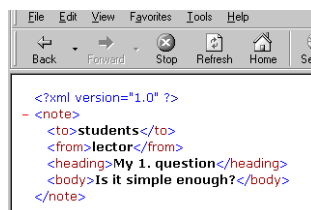
MS Internet Explorer 5.0

zobrazenie štruktúr (+, -)

v budúcnosti podpora  
DTD, XSL

neformátované zobrazenie

vlastné značky => neznáma sémantika



## Odladenie XML dokumentu

XML - po syntaktickej chybe nespracovávať

HTML - samooprava chyby (= nekompatibilita)

MS IE 5.0 - ohlásenie chyby

pravdepodobné  
miesto a príčina



## Formátovanie údajov z XML dokumentu

vlastné značky - neznáma interpretácia

zobrazenie - definovanie predlohy ako formátovať  
XSL - eXtended Stylesheet Language

dva jazyky:

- XSL FO - Formatting Objects  
(parametre zobrazenia na rôznych médiách)
- XSLT - Transformations  
(transformácia XML dokumentov)  
(HTML, inak štruktúrované XML, ...)

## Strojové spracovanie

spracovanie textu - **parser**

- “dekódovanie” postupnosti znakov (well formed)
- poskytnutie info vyšším vrstvám
- platnosť oproti DTD / Schema - **validátor**

Document Object Model, DOM  
objektová reprezentácia stromu

Simple API for XML, SAX  
udalosti počas spracovania prúdu znakov

## Praktické ukážky a ...

príklady v `examples/I`, pozrieť v IE

ukážka tvorby dokumentu XML

... samostatná práca

## Technické spracovanie XML

parser - spracovanie textu v XML dokumente

```
<text>this text is also parsed</text>
```

```
<note><to>parser</to><from>  
author</from><heading>sorry  
</heading><body>  
Can you parse it? :)</body>  
</note>
```

## Technické spracovanie XML (pokr.)

- špeciálne znaky
- znakové sekvencie
- kódovanie (charset)
- hlavička XML dokumentu

## Nedátové prvky dokumentu XML

- komentáre
- inštrukcie (processing instructions)

## Nahradzovanie špeciálnych znakov

`<question>Is 100 <x(ml)?</question>`

**entity reference** - nahradzované reťazce

|                          |                         |                    |
|--------------------------|-------------------------|--------------------|
| <code>&amp;meno ;</code> | <code>&amp;lt;</code>   | <code>&lt;</code>  |
|                          | <code>&amp;amp;</code>  | <code>&amp;</code> |
|                          | <code>&amp;gt;</code>   | <code>&gt;</code>  |
|                          | <code>&amp;apos;</code> | <code>'</code>     |
|                          | <code>&amp;quot;</code> | <code>"</code>     |

`<question>Is 100 &lt;x(ml)?</question>`

## Znakové sekvencie - CDATA

neinterpretovaná časť XML dokumentu

`<![CDATA[ľubovoľné znaky]]>`

```
<code>
<![CDATA[
  if ((a < 100) && ( ! blnContinue))
    return blnDone && blnOk;
]]>
</code>
```

## Kódovanie

“Počítačová Odysea” - fyzická reprezentácia znakov

```
<?xml version="1.0"?>
<otázka>Chcete ešte žiť?</otázka>
```

informácia v hlavičke XML dokumentu

parameter **encoding**

```
<?xml version="1.0" encoding="ISO-8859-2"?>
<otázka>Chcete ešte žiť?</otázka>
```

## Kódovanie (pokr.)

rôzne normy (štandardy)

- ASCII, Latin I, Latin II, kód Kamenických
- ISO-8859-1, ISO-8859-2
- UTF-8, UTF-16

! podpora platformy, softvéru  
windows-1250, Cp1250, ...

## Hlavička XML dokumentu

```
<?xml version="1.0" encoding="ISO-8859-2"?>
```

metainformácie o XML dokumente

- **verzia**
- kódovanie ! v určenom poradí
- standalone (yes/no)

**vždy** na začiatku dokumentu

kódovanie podľa prvých znakov (default **UTF**)

## Komentáre

syntakticky: `<!--text-->`

*text* nesmie obsahovať '--'

komentáre **nie je možné vnárať**

```
<!-- deklarácia <head> & <body> -->
```

*Poznámka:* neinterpretovaný obsah

## Inštrukcie (processing instructions)

inštrukcie pre (vyššie) aplikácie

nie je súčasť dát, parser musí len poskytnúť

syntax: `<?meno param1 param2 ...?>`

meno **nesmie začínať** (nejakou) kombináciou 'xml'

```
<?Import_Type New?>  
<?Language SK?>
```

## XML ťahák, časť 2.

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<?goal review?>
<note>
  <to>listeners</to>
  <from>speaker</from>
  <heading>test</heading>
  <body>
    Are you ready for a test?
    <!-- sure they are :) -->
    What is <![CDATA['<?xml version="1.0">']]>?
  </body>
</note>
```

## Konflikt mien

XML dokument  
dáta uložené v tabuľke

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

XML dokument  
nábytok

```
<table>
  <name>Coffee Table</name>
  <width>80</width>
  <height>120</height>
</table>
```

## Namespace - priestor mien

rozšírenie mena o **prefix** a ':'

```
<f:table>
  <f:name>Coffee Table</f:name>
  <f:width>80</f:width>
  <f:height>120</f:height>
</f:table>

<html:table>
  <html:tr>
    <html:td>Apples</html:td>
```

vytvorené rôzne elementy **table** (a ďalšie...)

## Jazyk - priestor mien

použitie XML - definovanie jazyka (syntax, sémantika)

dokument s dátami z rôznych oblastí

- spracovanie *na vyššej úrovni*

dokument s viacerými jazykmi

- rôzne mená elementov
- konflikt rovnakých mien

príslušnosť značky (mena elementu) k jazyku -  
**priestor mien (namespace)**

## Namespace - priestor mien (pokr.)

definovanie prefixu - atribút **xmlns**

**xmlns:namespace-prefix="namespace"**

platí pre potomkov elementu

```
<html:table xmlns:html="http://www.w3.org/TR/html4/">
  <html:tr>
    <html:td>Apples</html:td>
```

## Namespace - unikátny identifikátor

**xmlns:namespace-prefix="namespace"**

*namespace* - mal by byť **URI**

- zaručená unikátnosť mena
- URL - aj fyzicky dokument  
(nie pre parser)

## Namespace - zjednodušenie

default nastavenie: **<element xmlns="namespace">**

```
<table xmlns="http://www.w3.org/TR/html4/">
  <tr>
    <td>Apples</td>
    <td>Bananas</td>
  </tr>
</table>
```

*Poznámka:* koreň dokumentu

## Namespace - príklad

```
<?xml version="1.0" encoding="iso-8859-2"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">

  <xsl:template match="/">
    <html:HTML xmlns:html="http://www.w3.org/TR/html4/">
      <html:BODY>
        <html:H2><xsl:value-of select="Nadpis"/></html:H2>

        <html:P>
          <xsl:apply-templates select="Vtipy/Vtip"/>
        </html:P>
      </html:BODY>
    </html:HTML>
  </xsl:template>

</xsl:stylesheet>
```

## XML - zhrnutie

Základné pojmy

dokument, element (značky), atribút

Nedátové prvky

inštrukcie, komentáre

Technické prvky

entity, CDATA

Metainformácie

hlavička XML dokumentu, kódovanie

Namespace

## Ďakujem za pozornosť



WWW Consortium (špecifikácie)

[www.w3.org](http://www.w3.org)

OpenSource implementácia parsera (Xerces)

[xml.apache.org](http://xml.apache.org)

Tutoriály, materiály

[www.zvon.org](http://www.zvon.org)

iné

[www.xml.org](http://www.xml.org), ...