

CS 683 Emerging Technologies
Fall Semester, 2004
Doc 8 XML
Contents

XML References	2
XML	3
The XML Universe	7
XML Syntax.....	10
XML Terms	11
Well-Formed XML Documents.....	15
Basic Structure	15
Special Characters	23
Entities	24
Valid XML Documents	25

Copyright ©, All rights reserved. 2004 SDSU & Roger Whitney, 5500 Campanile Drive, San Diego, CA 92182-7700 USA. OpenContent (<http://www.opencontent.org/opl.shtml>) license defines the copyright on this document.

XML References

Sun's XML site

<http://java.sun.com/xml/>

XML Tools - parsers, transformers, SOAP, etc,
Documentation & Tutorials

World Wide Web Consortium (W3C)

<http://www.w3.org/>

XML specifications, Documentation, Tutorials

XML.com

<http://www.xml.com/>

XML news, Articles

Learning XML, Erik Ray, O'Reilly, 2001

Used in creating this tutorial

XML

XML creators wanted

- Flexibility of SGML
- Simplicity of HTML

Key differences from HTML

- Presentation is separate from document description
- Error Checking
- Unambiguous Structure

The Main Point of XML

XML is about

- Document structure
- Describing data

Sample XML

```
<?xml version="1.0" ?>
<CATALOG>
  <CD>
    <TITLE>Empire Burlesque</TITLE>
    <ARTIST>Bob Dylan</ARTIST>
    <COUNTRY>USA</COUNTRY>
    <COMPANY>Columbia</COMPANY>
    <PRICE>10.90</PRICE>
    <YEAR>1985</YEAR>
  </CD>
  <CD>
    <TITLE>Hide your heart</TITLE>
    <ARTIST>Bonnie Tyler</ARTIST>
    <COUNTRY>UK</COUNTRY>
    <COMPANY>CBS Records</COMPANY>
    <PRICE>9.90</PRICE>
    <YEAR>1988</YEAR>
  </CD>
</CATALOG>
```

From <http://www.w3schools.com/xsl/default.asp> example

Developing XML

Defining the XML tags

Creating documents using XML tags

Displaying or processing the documents

Creating XML tags requires thinking about

- What document structures are important
- What data is needed now and later

Document Structure Example

Which is better?

<Paragraph>

This is a short paragraph. It will be used as an XML example

</Paragraph>

<Paragraph>

<Sentence>

This is a short paragraph.

</Sentence>

<Sentence>

It will be used as an XML example

</Sentence>

</Paragraph>

<Paragraph>

<Author>Roger Whitney</Author>

<DateCreated>July 20, 2001</DateCreated>

<Title>XML Example</Title>

<Sentence>

This is a short paragraph.

</Sentence>

<Sentence>

It will be used as an XML example

</Sentence>

</Paragraph>

The XML Universe

Basic Syntax

XML 1.0 spec

XLinks

Linking documents together

Namespaces

Namespaces for tags.

XML Markup Languages

XHTML

XML version of HTML

MathML

Used to describe math equations

The XML Universe Document Modeling

Document Type Definitions (DTDs)
Defines legal tags for a document

XML Scheme
Defines data types in XML

Data Addressing & Query

XPath

XPointer

XML Query Language (XQL)

Presentation and Transformation

XML Stylesheet Language (XSL)

XSL Transformation Language (XSLT)

Cascading Style Sheets (CSS)

Extensible Stylesheet Language for Formatting Objects
(XSL-FO)

The XML Universe Parsing, Programming

Document Object Model (DOM)

Simple API for XML (SAX)

XML Information Set

XML Fragment Interchange

Network Protocols

XML-RPC

Simple Object Access Protocol (SOAP)

XML Syntax Example

```
<!-- A simple XML document with comment -->  
<greetings>  
  Hello World!  
</greetings>
```

XML Terms

Tag

A piece of text that describes a unit of data

Tags are surrounded by angle brackets (< and >)

Tags are case sensitive

The following are different tags

<GREETINGS>

<greetings>

<Greetings>

Attribute

Qualifier in an XML tag

<slide **title="XML Slide"**>

<slide **title="Who's on First"**>

<name **position='First'**>

Value of the attribute must be quoted

Can use either single or double quotes

XML Terms Element

Unit of XML data, delimited by tags

```
<greetings>Hello World!</greetings>
```

```
<name>
```

```
  <firstName>John</firstName>
```

```
  <lastName>Fowler</lastName>
```

```
</name>
```

Elements can be nested inside other elements

Markup

Tags and comments in an XML document

Content

Anything that is not markup in a document

Document

An XML structure in which one or more elements contains text intermixed with subelements

XML Document

```
<greetings>
  <from>
    <nnamnee>
      <firstName>Roger</firstName>
      <lastName>Whitney</lastName>
    </nnamnee>
  </from>
  <to>
    <name>
      <firstName>John</firstName>
      <lastName>Fowler</lastName>
    </name>
  </to>
  <message>
    How are you?
  </message>
</greetings>
```

Issues

Is that a typo or a legal tag?

How would we know?

Levels of XML

Well-formed

XML document that satisfies basic XML structure

Valid

XML document that is well-formed and

Specifies which tags are legal

A Document Type Definition (DTD) is use to specify

- Legal tags
- Correct tag nesting

Well-Formed XML Documents

Basic Structure

Optional Prolog
Root Element

```
<?xml version="1.0" ?>  
<!-- A simple XML document with comment -->  
<greetings>  
  Hello World!  
</greetings>
```

Prolog

For well-formed documents the prolog

- Is optional
- But recommended

Prolog has optional three attributes:

- version
1.0 & 1.1
- encoding
Character encoding
UTF-8 (default), UTF-16, US-ASCII, etc.
- standalone

Is the complete XML document in one file?

```
<?xml version="1.0" encoding='US-ASCII' standalone='yes'  
?>
```

```
<?xml version="1.0" encoding='iso-8859-1' standalone=no  
?>
```

Comments

Comments can be placed nearly anywhere outside of tags

Comments can not come before `<?xml version="1.0" ?>`

```
<?xml version="1.0" ?>
<!-- Another comment -->
<greetings>
  <from>Roger<!-- Legal comment --></from>
  <to>John</to>
  <message>Hi</message>
</greetings>
<!-- Comments at the end -->
```

Root Element

Each XML Document has a single root element

Legal XML Document

```
<?xml version="1.0" ?>  
<greetings>  
  <from>Roger</from>  
  <to>John</to>  
  <message>Hi</message>  
</greetings>
```

Illegal XML Document

```
<?xml version="1.0" ?>  
<from>Roger</from>  
<to>John</to>  
<message>Hi</message>
```

Basic Rules for Well-Formed Documents

- Non-empty elements must have start and end tags

Legal

```
<greetings>  
  <from>Roger</from>  
  <to>John</to>  
  <message>Hi</message>  
</greetings>
```

Illegal

```
<body>  
  <p>Hello world  
  <p>How are you?  
</body>
```

- Empty elements can use one tag

You can shorten

```
<greetings></greetings>
```

to

```
<greetings/>
```

Basic Rules For Well-Formed Documents

- All attribute values must be in quotes

Legal

`<tag name='sam'>`

Illegal

`<tag name=sam>`

- Elements may not overlap

`<center>Bad XML, but ok HTML</center>`

White Space

Are the following the same?

```
<greetings>  
  Hello World!  
</greetings>
```

```
<greetings>Hello World!</greetings>
```

White Space

For some applications white space may be important

XML parsers are to pass white space to applications

The application decides if the white space is important

Special Characters

What happens if we need to use < inside an element?

This is illegal XML

```
<paragraph>
  Everyone knows that 5 < 10 & 1 > 0.
</paragraph>
```

Need to encode the < and & symbols

```
<paragraph>
  Everyone knows that 5 &lt; 10 &amp; 1 > 0.
</paragraph>
```

You can use a CDATA section

```
<paragraph><![CDATA[
  Everyone know that 5 < 10 & 1 > 0. ]]>
</paragraph>
```

Standard element content can not contain: <]]> &

Entities

Predefined

Entity	Character
<	<
>	>
&	&
'	'
"	"

XML allows you do define your own entities

Valid XML Documents

XML document that is well-formed and

Specifies which tags are legal

A Document Type Definition (DTD) is use to specify

- Legal tags
- Correct tag nesting

Example

```
<?xml version="1.0" ?>  
<!DOCTYPE greetings [  
  <!ELEMENT greetings (#PCDATA)>]>  
<greetings>  
  Hello World!  
</greetings>
```