Web Programming

Lecture 6 – Introduction to XML

Meta-markup Languages

- A markup language allows the user to identify individual elements of a document, e.g., what is a paragraph, a heading, or an unordered list.
- Used in combination with a style sheets, data can be properly presented on a web page, a slide show, or any other method that is appropriate for the data.

Meta-markup Languages

- A meta-markup language is a little different; it doesn't specify a document it specifies a *language*.
- SGML (Standard Generalized Markup Language) and XML (eXtensible Markup Language) are examples of meta-markup languages.

SGML

- SGML was based on GML (Generalized Markup Language), which was developed at IBM in the 1960s. SGML was developed in 1974.
- SGML was intended to allow for the sharing of machine-readable documents.
- While it was used in the printing and publishing industry, its complexity kept it from wider use.
- SGML was used as the basis for HTML.

XML

- HTML describes the layout of information but conveys no information about its meaning. This limits the ability to retrieve information from an HTML document automatically.
- One solution to get around HTML's limitation is for groups of users to define and use their own set of tags and attribute and use a meta-markup language to implement them.
- XML is a simpler language than SGML and therefore more useful.

Using XML

- XML is not a replacement for XHTML. It is intended to provide a way to label data in a way that can be analyzed and manipulated automatically.
- XML is normally used together with a style sheet and an appropriate processor to produce a suitable XHTML document based on the XML file and the style sheet.

Syntax of XML

- XML has two levels of syntax:
 - The general low-level syntax within the XML doucment.
 - The higher-level syntax specified by DTD (Document Type Definitions) or XML schemas.
- XML documents can contain:
 - data elements of the document
 - markup declarations (instructions to the XML parser)
 - processing instructions (instructions for an application process that will process the data).

Elements of XML

- All XML documents begin with an XML declaration, which identifies the document as XML, and provides the version number of the XML standard being used and the encoding standard:
- <?xml version = "1.0" encoding = "utf-8"?>
- Comments in XML are the same as in XHTML:
 - <!- This is a comment -->

Names in XML

- XML names are used to identify elements and attributes.
 - XML names must begin with a letter or an underscore and can contain letter, underscores, digits, hyphens and periods.
 - XML names are case sensitive; e.g., Body, body and BODY are three different names in XML.
 - There are no limits to the length of XML names.

Basic Syntax Rules

- Every XML documents defines a root element and that root element's tag must appear on the first line of XML code.
- All other elements must be nested within that element.
- For a XHTML document, the root tag is **html**.
- Every XML element must have a closing tag:
 - <myTag> ... </myTag>
 - <myTag />

Sample XML Document

```
<?xml version = "1.0" encoding = "utf-8"?>
<ad>
    <year> 1960 </year>
    <make> Cessna <make>
    <model> Centurian </model>
     <color> Yellow with white trim </color>
     <location>
          <city> Gulfport </city>
          <state> Mississippi </state>
          </location>
          </do>
```

Another Sample XML Document

XML Attributes

- In XML, attributes can be used to provide additional information about elements in an XML document.
- Example:

```
<file type = "gif"> computer.gif </file>
```

• Attributes must be enclosed in quotation marks (single or double)

```
<file type = 'gif'> computer.gif </file>
is also valid.
```

Attributes Or Nested Tags

- Is it better to add an additional attribute to an element or to define a nested element?
- Sometimes there is no choice an image can only be an attribute (XML only handles text data).
- Nested tags can be added to any existing tag to describe its growing size and complexity – attributes give no information about this.

A Tag With One Attribute

```
<!-- A tag with one attribute -->
<patient name = "Maggie Dee Magpie ">
.... ....
</patient>
```

A Tag With One Nested Tag

```
<!-- A tag with one nested tag -->
<patient>
    <name> Maggie Dee Magpie
    </name>
.... ....
</patient>
```

An Extra Level Of Nested Tags

```
<!-- A tag with one nested tag, which contains three nexted tags -->
<patient>
    <name>
        <first> Maggie </first>
        <middle> Dee </middle>
        <last> Magpie </last>
        </name>
        .... ...
</patient>
```

XML and Auxiliary Files

- An XML document often uses two auxiliary files:
 - One file specifies its tag set and structural syntactic rules. This can be either a DTD or an XML schema.
 - One file contains a style sheet to describe how the document's content is to be printed and/or displayed. This can be either a Cascading Style Sheet or an XSLT Style Sheet.

XML Document Structure

- An XML document consists of one or more entities that are logically related sets of data.
- The document entity describes the document as a whole and is usually subdivided into other entities.
- These other entities may (or may not) be physically located in the same file.
- Entity names can be any length

Entity Names

- Entity names can be any length.
- They must begin with a letter, a dash or a colon.
- The remaining characters can be letters, digits, periods, dashes, underscores or colons.
- Adding an amersand before and asemi-colon after a reference name turns it into a reference.
 - &apple_image is a reference to the entity apple_image.

Character Data Sections

- When a document requires several predefined entities near each other, it becomes hard to read; therefore, we can use a character data section.
- Character data sections are not parsed and appear in an XML document as they are written.
- Character data sections cannot contain tags because they are considered literal text they do not mark up the document.

CDATA

- Their basic syntax is:
 - <! [CDATA [content]]>
- An example:
 - <![CDATA [The last word of the line is >>> here <<<]]>
- This is clearly superior to writing:

```
The last word of the line is > > > > here < &lt;
```

• If I wrote

<![CDATA [The form of the tag is < tag name <]]> I would get:

The form of the tag is < tag name <

Document Type Definitions

- A document type definition (DTD) is a set of rules specifying how a set of elements can appear in an XML document as well as entity declarations.
- While XML documents do not require DTDs, it allows the programmer to check an XML document for validity.
- A DTD can be internal (placed inside the XML document) or external (placed in a separate file that the XML document references).

DTD Syntax

- A DTD is a sequence of declarations:
 - <!keyword ... >
- There are 4 valid keywords:
 - **ELEMENT** used to define tags
 - **ATTLIST** used to define tag attributes
 - **ENTITY** used to define entities
 - **NOTATION** used to define data type notations

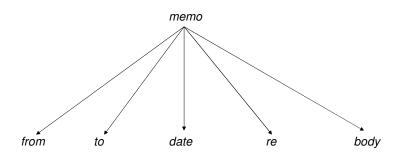
Declaring Elements

- Element declarations are a form that is similar to BNF.
- Each element declaration specifies the structure of <u>one</u> element, containing its names, its constituents (if it has child elements) or the data type of its parent (if it is a leaf).

Declaring Non-leaf Elements

- The general form of an element declaration if there are child elements is:
 - <!ELEMENT ElementName (ChildElementList)>
- An example:
- <!ELEMENT memo (from, to, date, re, body)>





Child Element Specification Modifiers

- Normally, an element specification indicate one occurrence of an element.
- Using a modifier allows the programmer to have multiple occurrences of an element.

Modifier	Meaning
+	One or more occurrences
*	Zero or more occurrences
?	Zero or one occurrence

Declaration With Element Modifiers

- <!ELEMENT person (parent+, age, spouse?, sibling*>
 indicates
- One or more parents
- Age
- Spouse (optional)
- Zero or more siblings

Declaring a Leaf

- The syntax for an element that does not have child elements is:
 - <!ELEMENT element_name (#PCDATA) >
- An example
 - <!ELEMENT year (#PCDATA)>

Declaring Attributes

- Attributes of an element are specified separately from the element declaration.
- An attribute declaration must include:
 - the attribute's name
 - the element to which it belongs
 - its type

Syntax for Attribute Declarations

- If the element has only one attribute it can be declared:
 - <! ATTLIST ElmntName AttribName AttribType [DefltVal]>
- Multiple attributes can be declared separately or in one declaration:
 - <! ATTLIST ElmntNm1 AttribNm1 AttribType1 DefltVal1
 ElmntNm2 AttribNm2 AttribType2 DefltVal2

ElmntNmN AttribNmN AttribTypeN1 DefltValN >

Possible Default Values

<u>Value</u>	Meaning
A value	The quoted value, which is used if none is specified in an element
#FIXED value	The quoted value, which every element will have and which cannot be changed
#REQUIRED	No default value is given; every instance of the element must specify a value
#IMPLIED	No default value is given (the browser chooses the default value); the value may or may not be specified in an element.

Declaring Attributes: An Example

- The declarations:
- <! ATTLIST airplane places CDATA "4">
- <! ATTLIST airplane engineType CDATA #REQUIRED>
- <! ATTLIST airplane price CDATA #IMPLIED>
- <! ATTLIST airplane manufactr CDATA #FIXED "Cessna">
- This is a valid XML element for this DTD:
- <airplane places = "10" engineType = "jet">
 </airplane>

Declaring Entities

- General entities can be referenced anywhere in the content of an XML document.
- Parameter entities can only be referenced in DTDs.
- Syntax:
 - <!ENTITY [%] entityName "entityValue">
- If the percent sign is included, it is a parameter entity.

Referencing Entities

• An entity can be referenced in a declaration by placing an ampersand before and a semi-colon after the name:

<!ATTLIST airport airportName CDATA &jfk; >

External Text Entities

- An entity can be too long to be placed within the DTD; they can be pages long.
- This can be handled by placing it in a different file. These are called external text entities:
- <!ENTITY entityName SYSTEM "fileLocation">

A Sample DTD

```
<!ELEMENT location (#PCDATA)>
<!ELEMENT city (#PCDATA)>
<!ELEMENT state (#PCDATA)>

<!ATTLIST seller phone CDATA #REQUIRED>
<!ATTLIST seller email CDATA #IMPLIED>

<!ENTITY c "Cessna">
<!ENTITY p "Piper">
<!ENTITY b "Beechcraft">
```

Internal DTDs

- DTDs can appear in the same file as the XML document or in a file of their own.
- If the DTD is included in the XML file, it's include in a **DOCTYPE** tag:

```
<?xml version = "1.0" encoding = "utf-8"?>
<!DOCTYPE planes [
    All the declarations go here
    ]>
    The XML document goes here
```

External DTDs

- If the DTD is in its own file, the XML document contains a DOCTYPE tag identifying the DTD file <!DOCTYPE XMLDocumentRootName SYSTEM "DTDFileName">
- In our case, it is
 !DOCTYPE planes_for_sale SYSTEM "planes.dtd">

planes.xml

```
<seller phone = "555-222-3333"> Skyway Aircraft
  </seller>
  <location>
    <city> Rapid City </city>
    <state> South Dakota </state>
  </location>
</ad>
<ad>
  <year> 1965 </year>
  <make> &p; </make>
  <model> Cherokee </model>
  <color> Gold </color>
  <description> 240 hours SMOH, dual NAVCOMs,
          DME, new Cleveland brakes, great shape
  </description>
  <price> 23, 495 </price>
```

Disadvantages of DTDs

- Since DTDs use a syntax different from XML, XML processors cannot analyze them.
- DTDs do not allow restrictions in the form of data that can appear in a tag.
 - DTDs have only 10 data types, none of which are numeric.

Displaying Raw XML Documents

- A browser will not know how to format an XML.
- It will display both the data and the tags, allowing subfields to be collapsed.
- It is important to be able to format XML data using Cascading Stylesheets or XSLT.

Formatting XML Documents

- There are two ways to provide format information to the browser for an XML document:
 - a Cascading Style Sheet (CSS) file
 - Extensible Stylesheet Language Transformations (XSLT)
- While not every browser supports XSLT, it has more power than CSS over the document's appearance.

Displaying XML Documents with CSS

- A CSS style sheet for XML has a list of element names, each followed by the element's attributes (and their values) delimited by braces.
- The only common style property that has not been discussed before is display, which can be inline (the default) or block. These determine if the element is to be displayed inline or in a separate block.
- To establish a connection between the style sheet and the XML file, add the following tag into the XML file:

```
<?xml-stylesheet type = "text/css"

href = "FileName.css">
```

planes.xml

planes.xml

```
<?xml version = "1.0" encoding = "utf-8"?>
<!-- planes.xml - A document that lists ads for
                  used airplanes -->
<?xml-stylesheet type = "text/css" href =</pre>
  "planes.css" ?>
<!DOCTYPE planes_for_sales SYSTEM "planes.dtd">
<planes_for_sale>
  <ad>
    <year> 1977 </year>
    <make> &c; </make>
    <model> Skyhawk </model>
    <color> Light blue and white </color>
    <description> New paint, nearly new interior,
               685 hours SMOH, full IFR King
  avionics
    </description>
```

Namespaces

- An XML namespace is a collection of element and attributes names used in XML documents.
- A namespace's name usually looks like a URI (*U*niform *R*esource *I*dentifier).
- Using namespaces allows you to give tags meaningful names that may conflict with tags that appear in other documents (e.g., HTML tags).

Why Use Namespaces?

```
    Markup for an XHTML table

        Apples
        /td>
        /td>
```

Declaring a Namespace

• When we write:

```
<html xmlns="http://www.w3.org/1999/xhtml">
  we are making the namespace defined at
    www.w3.org/1999/xhtml the default for
  the document.
```

• This tells the browser that the html document that follows (until the </html> tag) uses the tags located at that URI, where they are defined.

Declaring a Namespace With A Prefix

• To use multiple namespaces in the same document, we need to use an optional prefix:

birds

```
xmlns:bd="http://www.audobon.org/names/species">
```

 An element can have more than one namespace declared, but then all but one MUST have prefixes if more than one defines the same identifier:

```
<birds
xmlns:html="http://www.w3.org/1999/xhtml"
xmlns:bd="http://www.audobon.org/names/species">
```

Using Namespaces

XSLT Style Sheets

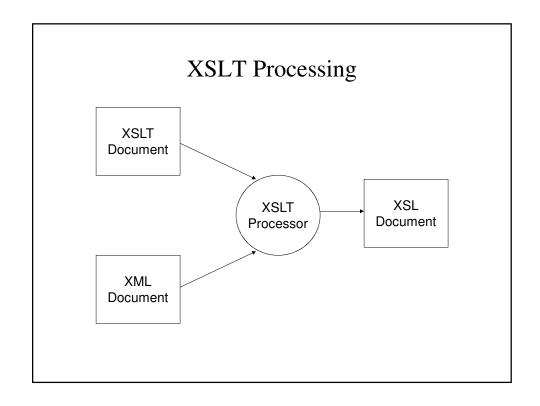
- XSLT is one of three different XSLs (eXtensible Stylesheet Languages) that can be used to transform XML documents into other forms, including XHTML documents.
- XSLT can be used to move, modify, sort XML elements and even to convert them into attrtibutes.
- Since XLST files are XML documents themselves, they can be validate using DTDs or transformed using XSLT stylesheets.

Overview of XSLT

- XSLT is a functional programming language, similar in some ways to LISP or Scheme.
- XSLT has functions, parameters, names to which values can be bound, selection constructs and conditional expressions capable of multiple selection.
- The XSLT processors takes an XSLT as its program, an XML document as input and produces an XSL files as output that can be saved or displayed in the browser.

An XSLT Document

- An XSLT document consists of one or more templates, using XPath (a related language) to describe element/attribute patterns in the XML input document.
- Each template has an associated section of XSLt code, which is executed when there is a match.
- The XSLT processor searches the XML document for template matches and is usually a "template-driven model."



XSL Transformations for Presentation

- This discussion of XSLt will be limited to basic formatting.
- We will assume that we are processing an XML document with an associated XSLT style sheet to produce an XHTML document.
- For an XML document to serve as data for an XSLT document, it must contain a processing instruction:

```
<?x-stylesheet type = "text/xsl"
href="XSLName" ?>
```

Original people.xml

The XSLT Style Sheet

• An XSLT style sheet is an XML document whose root tag is **stylesheet**. This tag must define the namespaces:

```
<xsl:stylesheet
xmlns:xsl = http://w3.org/1999/XSL/Format />
```

person.xsl

```
<?xml version="1.0" encoding="utf-8"?>
<xsl:stylesheet version="1.0"
    xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:template match="person">
    A person
</xsl:template>
</xsl:stylesheet>
```

Rewriting person.xml to Read the Stylesheet

person.xml As Displayed

A person A person

value-of

- Generally, what you want as output is closely related to the text that is in the XML document that provides your input.
- **xsl:value-of** calculates the string value of an Xpath expression and inserts it into the output.
 - The value of an element is the text that it contains after removing the tags and resolving the entity and cahracter references.

Using value-of

```
<?xml version="1.0" encoding="utf-8"?>
<xsl:stylesheet version="1.0"
    xmlns:xsl="http://www.w3.org/1999/XSL/Transform">

<xsl:template match="person">
    <xsl:value-of select = "name" />
</xsl:template>
</xsl:stylesheet>
```

Using value-of: An Example

Alan Turing Richard Feynman

Using apply-templates

- By default, an XSLT processor reads the input XML document from top to bottom, doing a preorder traversal of the document from the root.
- xsl:apply-templates allows the user to make the processing explicit, changing it if one wishes.

Why Use apply-templates

• Using in our earlier example

• We can limit the output to first and last name by writing:

```
<xsl:template match="person">
  <xsl:apply-templates select="name" />
</xsl:template>
```

people.xsl With apply-templates

```
<?xml version="1.0" encoding="utf-8"?>
<xsl:stylesheet version="1.0"
    xmlns:xsl="http://www.w3.org/1999/XSL/Transform">

<xsl:template match="people">
<html lang = "en">
    <head>
        <title> Famous Scientists</title>
        <meta charset = "utf-8" />
        </head>
        <body>
<hl> Famous People </hl>
</table
```

```
<xsl:template match="name">
  <xsl:value-of select = "first_name" />
  <xsl:value-of select = "last_name" />
  </xsl:template>
</xsl:stylesheet>
```

The Output

Alan Turing computer scientist Richard Feynman physicist

Famous Scientists

Alan Turing	computer scientist
Richard Feynman	physicist

xsl:for-each

- The <xsl:for-each> element allows you to do looping in XSLT.
- The XSL <xsl:for-each> element can be used to select every XML element of a specified node-set.

```
mycds.xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml-stylesheet type="application/xml" href="mycds.xsl"?>
<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>
```

```
<cd>
             <title>Greatest Hits</title>
             <artist>Dolly Parton</artist>
             <country>USA</country>
             <company>RCA</company>
             <price>9.90</price>
             <year>1982
      </cd>
      <cd>
             <title>Still got the blues</title>
             <artist>Gary Moore</artist>
             <country>UK</country>
             <company>Virgin records</company>
             <price>10.20</price>
             <year>1990</year>
      </cd>
</catalog>
```

mycds.xsl <?xml version="1.0" encoding="ISO-8859-1"?> <xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform"> <xsl:template match="/"> <html> <body> <h2>My CD Collection</h2>

Title
 Artist

Using for-each



XML Documents and Their Nodes

- There are seven different types of nodes that XML documents can have. These include:
 - Root nodes
 - Element nodes
 - Attribute nodes
- There are built-in template rules for these different types of nodes.
- For attribute nodes, matching an attribute type will lead to outputting the value of the attribute, not its name

people.xsl Checking Attributes

```
<price> 23,495 </price>
  <seller phone = "555-222-3333"> Skyway Aircraft
</seller>
  <location>
    <city> Rapid City </city>
   <state> South Dakota </state>
  </location>
</ad>
<ad>
  <year> 1965 </year>
  <make> &p; </make>
  <model> Cherokee </model>
  <color> Gold </color>
  <description> 240 hours SMOH, dual NAVCOMs,
            DME, new Cleveland brakes, great shape
  </description>
  <price> 23,495 </price>
```