INTERNET & WEB APPLICATION DEVELOPMENT SWE 444

Fall Semester 2008-2009 (081)

Module 4 (I): Data Description and Transformation, XML

Dr. El-Sayed El-Alfy
Computer Science Department
King Fahd University of Petroleum and Minerals
alfy@kfupm.edu.sa

Objectives/Outline

- Objectives
  - Learn the basics of XML
  - Learn to markup data using XML

- Outline
  - Introduction
  - What is XML?
  - How can XML be Used?
  - Components of an XML Document
    - XML Declaration
    - Processing Instructions
    - XML Elements
    - XML Attributes
  - Rules For Well-Formed XML

Introduction

- XML
- XSL / XSLT
- DTD
- DOM
- XSD
- XPath
- XForms

What is XML?

- XML stands for eXtensible Markup Language
- XML is a W3C Recommendation
- XML is a markup language much like HTML
- A portable open (i.e., nonproprietary) technology for data storage and exchange
- Widely supported from the big IT companies
- A meta-language, thus permits document authors to create custom markup for any type of information
  - Can create entirely new markup languages that describe specific types of data, including mathematical formulas, chemical molecular structures, music and recipes
- XML documents are typically files with the .xml extension
- XML documents are readable by both humans and machines
- Built-in internationalization via Unicode
- Built-in error-handling
- Optimized for network operations
- An XML parser is responsible for identifying components of XML documents and then storing those components in a data structure for manipulation

What is XML? (cont.)

- An XML document can optionally reference a Document Type Definition (DTD) or XML schema that defines the XML document's structure
  - XML with a DTD or XML Schema is designed to be self-descriptive
  - DTDs and schemas are essential for business-to-business (B2B) transactions and mission-critical systems.
- Validating XML documents ensures that disparate systems can manipulate data structured in standardized ways and prevents errors caused by missing or malformed data.

Example

```
<?xml version='1.0'?>
<bookstore>
  <book genre='autobiography' publicationdate='1981'
        ISBN='1-861003-11-0'>
    <title>The Autobiography of Benjamin Franklin</title>
    <author>
      <first-name>Benjamin</first-name>
      <last-name>Franklin</last-name>
    </author>
    <price>8.99</price>
  </book>
  …
</bookstore>
```

books.xml

- An XML document begins with an optional XML declaration, which identifies the document as an XML document. The `version` attribute specifies the version of XML syntax used in the document.
- XML comments begin with `<!--` and end with `-->`
- An XML document contains text that represents its content (i.e., data) and elements that specify its structure. XML documents delimit an element with start and end tags
- XML element names can be of any length and can contain letters, digits, underscores, hyphens and periods
  - Must begin with either a letter or an underscore, and they should not begin with "xml" in any combination of uppercase and lowercase letters, as this is reserved for use in the XML standards

Some History

- SGML (Standard Generalized Markup Language)
  - ISO Standard, 1986, for data storage & exchange
  - Metalanguage for defining languages (through DTDs)
  - A famous SGML language: HTML
  - Separation of content and display
  - Used in U.S. gvt. & contractors, large manufacturing companies, technical info. Publishers,...
  - SGML reference is 600 pages long
- XML
  - W3C recommendation in 1998
  - Simple subset (80/20 rule) of SGML: "ASCII of the Web", "Semantic Web"
  - XML specification is 26 pages long

Timeline

- 1986
  - SGML becomes a standard
- 1989
  - Tim Berners-Lee creates the WWW
- 1994
  - W3C established
- 1998
  - XML 1.0 W3C Recommendation
- Jan 2000
  - XHTML becomes W3C Recommendation
  - A Reformulation of HTML 4 in XML 1.0
- Feb 2004
  - W3C XML 1.0 (Third Edition) Recommendation
  - http://www.w3.org/TR/2004/REC-xml-20040204/
- Feb 2004
  - XML 1.1 Recommendation
  - http://www.w3.org/TR/2004/REC-xml11-20040204/
  - updates XML to use Unicode 3

XML vs. HTML

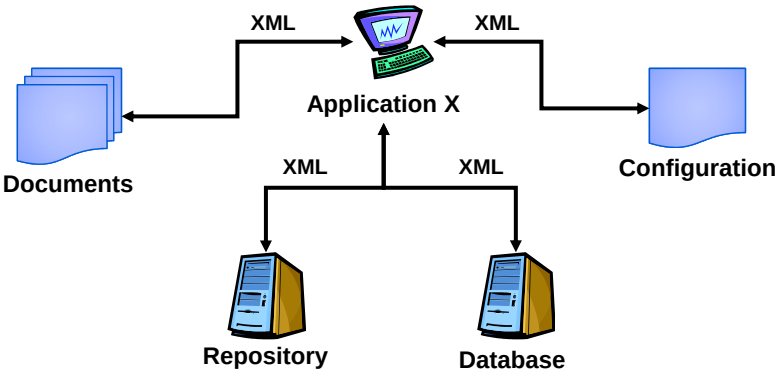
- XML is not a replacement for HTML but a complement to it
- They were designed with different goals
  - XML was designed to structure, store and transport data and to focus on what data is (not how to display)
  - HTML was designed to display data and to focus on how data looks.
    - describes both structure (e.g. <p>, <h2>,) and appearance (e.g.
, <font>, <i>)
- HTML and XML look similar because they are both SGML languages
  - Both HTML and XML use elements enclosed in tags
  - Both use tag attributes
  - HTML uses a fixed set of tags, whereas in XML, you make up your own tags
- HTML is for humans
  - HTML describes web pages
  - You don't want to see error messages about the web pages you visit
  - Browsers ignore and/or correct as many HTML errors as they can, so HTML is often sloppy
- XML is for computers
  - XML is used to mark up data so it can be processed by computers
  - The rules are strict and errors are not allowed
    - In this way, XML is like a programming language
  - Current versions of most browsers can display XML

XML vs. HTML (cont.)

- XML is Free and Extensible
  - XML tags are not predefined
    - XML allows the author to define his own tags and his own document structure
  - HTML uses a fixed, predefined, unchangeable set of tags
    - The author of HTML documents can only use tags that are defined in the HTML standard

XML Future

- XML is going to be everywhere
- XML is a cross-platform, software and hardware independent tool for transmitting information.



Benefits of XML

- Open W3C standard
- Representation of data across heterogeneous environments
  - Cross platform
  - Allows for high degree of interoperability
- Strict rules
  - Syntax
  - Structure
  - Case sensitive

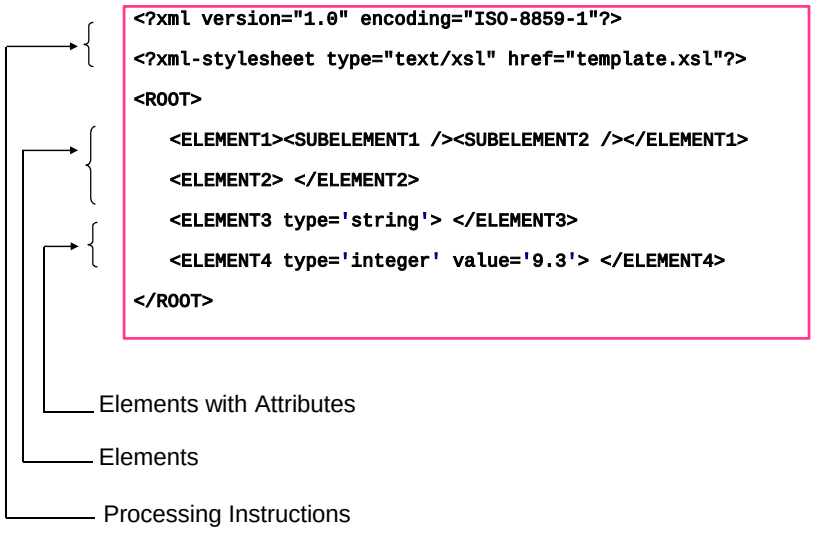
How can XML be Used?

- XML can Separate Data from HTML
  - With HTML, your data is stored inside the HTML document
    - To display dynamic data, it needs a lot of works to edit the html each time the data changes
  - With XML, your data can be stored in separate XML files and with a JavaScript, you can read xml files and update the data content of the HTML document
    - In this way the underlying data will not require any changes to the HTML.
- XML Simplifies Data Sharing
  - XML data is stored in plain text format; this provides a s/w and h/w-independent way of storing data that different applications can share
- XML Simplifies Data Transport
  - XML reduce the complexity of exchanging data between incompatible systems
- XML Simplifies Platform Changes
  - So you can easily upgrade to new systems (operating systems, new applications, or new browsers, etc) without losing data
- XML is Used to Create New Internet Languages
  - A lot of new Internet languages are created with XML such as
    - XHTML the latest version of HTML
    - WSDL for describing available web services
    - WAP and WML as markup languages for handheld devices
    - RSS languages for news feeds
    - RDF and OWL for describing resources and ontology
    - SMIL for describing multimedia for the web

Components of an XML Document

- Elements
  - XML elements form a tree structure that starts at "the root" and branches to "the leaves".
  - Each element has a beginning and ending tag
    - `<TAG_NAME>...</TAG_NAME>`
  - Elements can be empty (`<TAG_NAME />`)
- Attributes
  - Describes an element; e.g. data type, data range, etc.
  - Can only appear on beginning tag
- Processing instructions
  - Encoding specification (Unicode by default)
  - Namespace declaration
  - Schema declaration

Components of an XML Document



XML Declaration

- The XML declaration looks like this:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
```

  - The XML declaration is not required by browsers, but is required by most XML processors (so include it!)
  - If present, the XML declaration must be first--not even whitespace should precede it
  - Note that the brackets are `<?` and `?>`
  - `version="1.0"` is required (this is the only version so far)
  - encoding can be `"UTF-8"` (ASCII) or `"UTF-16"` (Unicode), or something else, or it can be omitted
  - `standalone` tells whether there is a separate DTD

Processing Instructions

- PIs (Processing Instructions) may occur anywhere in the XML document (but usually first)
- A PI is a command to the program processing the XML document to handle it in a certain way
- XML documents are typically processed by more than one program
- Programs that do not recognize a given PI should just ignore it
- General format of a PI:

```
<?target instructions?>
```

- Example:

```
<?xml-stylesheet type="text/css"
href="mySheet.css"?>
```

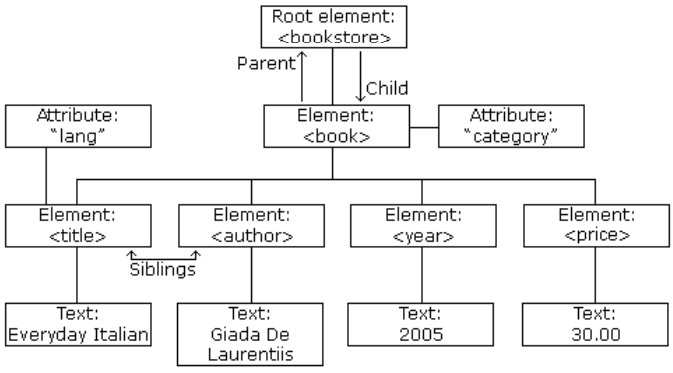

XML Elements

- An XML element is everything from the element's start tag to the element's end tag
- XML Elements are extensible and they have relationships
- XML Elements have simple naming rules
  - Names can contain letters, numbers, and other characters
  - Names must not start with a number or punctuation character
  - Names must not start with the letters xml (or XML or Xml ..)
  - Names cannot contain spaces

XML Attributes

- XML elements can have attributes
- Data can be stored in child elements or in attributes
- Should you avoid using attributes?
  - Here are some of the problems using attributes:
    - attributes cannot contain multiple values (child elements can)
    - attributes are not easily expandable (for future changes)
    - attributes cannot describe structures (child elements can)
    - attributes are more difficult to manipulate by program code
    - attribute values are not easy to test against a Document Type Definition (DTD) - which is used to define the legal elements of an XML document

Example 1



Example 1 (cont.)

```
<bookstore>
<book category="COOKING">
  <title lang="en">Everyday Italian</title>
  <author>Giada De Laurentiis</author>
  <year>2005</year>
  <price>30.00</price>
</book>
<book category="CHILDREN">
  <title lang="en">Harry Potter</title>
  <author>J K. Rowling</author>
  <year>2005</year>
  <price>29.99</price>
</book>
<book category="WEB">
  <title lang="en">Learning XML</title>
  <author>Erik T. Ray</author>
  <year>2003</year>
  <price>39.95</price>
</book>
</bookstore>
```


Example 2

```
<?xml version='1.0'?>
<bookstore>
  <book genre='autobiography' publicationdate='1981'
        ISBN='1-861003-11-0'>
    <title>The Autobiography of Benjamin Franklin</title>
    <author>
      <first-name>Benjamin</first-name>
      <last-name>Franklin</last-name>
    </author>
    <price>8.99</price>
  </book>
  <book genre='novel' publicationdate='1967' ISBN='0-201-63361-2'>
    <title>The Confidence Man</title>
    <author>
      <first-name>Herman</first-name>
      <last-name>Melville</last-name>
    </author>
    <price>11.99</price>
  </book>
</bookstore>
```

Example 3

```
<?xml version="1.0"?>
<weatherReport>
   <date>7/14/97</date>
   <city>North Place</city>, <state>NX</state>
   <country>USA</country>
   High Temp: <high scale="F">103</high>
   Low Temp: <low scale="F">70</low>
   Morning: <morning>Partly cloudy, Hazy</morning>
   Afternoon: <afternoon>Sunny &amp; hot</afternoon>
   Evening: <evening>Clear and Cooler</evening>
</weatherReport>
```


Rules For Well-Formed XML

- There must be one, and only one, root element
- All XML elements must have a closing tag
- Sub-elements must be properly nested
  - A tag must end within the tag in which it was started
- Attributes are optional
  - Defined by an optional schema
- Attribute values must be quoted (i.e. enclosed in " " or ' ')
- Processing instructions are optional
- XML is case-sensitive
  - <tag> and <TAG> are not the same type of element
- White space is preserved
- Comment in XML is similar to that of HTML
- Some characters have a special meaning in XML and using them inside and element generates error

&lt;	<	less than
&gt;	>	greater than
&amp;	&	ampersand
&apos;	'	apostrophe
&quot;	"	quotation mark

Q & A

?

References

- Some useful links with examples and other resources:
  - *Internet and World Wide Web How to Program*, 4/e, H. M. Deitel, P. J. Deitel, and A. B. Goldberg, Pearson Education Inc., 2008. Chapters 13.
  - W3 Schools XML Tutorial
    - http://www.w3schools.com/xml/default.asp
  - W3C XML page
    - http://www.w3.org/XML/
  - XML Tutorials
    - http://www.programmingtutorials.com/xml.aspx
  - Online resource for markup language technologies
    - http://xml.coverpages.org/