Network Europe - Russia - Asia of Masters in Informatics as a Second Competence 159025-TEMPUS-1-2009-1-FR-TEMPUS-JPCR

Sergio Luján Mora
Department of Software and Computing Systems
University of Alicante

XML
What is XML? And XHTML? And XPath, XQuery, XSLT, ...?

Has anybody used?

Contents

- Introduction
- XML applications
- Structure of a document
- Other XML technologies
- Software
- XML 1.1
Introduction

- XML
- Versions
- Advantages
- What do I need to use XML
- My first XML document
Extensible Markup Language (XML) is a set of rules for encoding documents in machine-readable form. It is defined in the XML 1.0 Specification produced by the W3C, and several other related specifications, all gratis open standards.

The design goals of XML emphasize simplicity, generality, and usability over the Internet. It is a textual data format with strong support via Unicode for the languages of the world. Although the design of XML focuses on documents, it is widely used for the representation of arbitrary data structures, for example in web services.
Extensible Markup Language
World Wide Web Consortium (W3C)
Problems:
◦ HTML not flexible, mix of content and presentation
◦ SGML too complex
 Subset of SGML
  GML (IBM, 1969)
  SGML (ISO 8879, 1986)
  XML (W3C, 1998)

<!-- Here's an SGML example file. Format it and print out the source, and 
   use it as a model for your own SGML files. As you can see this is a 
   comment. -->
<article>
<title>Quick SGML Example</title>
<author>Matt Welsh, <tt/mdwacs.cornell.edu/</author>
<date>v1.0, 28 March 1994
<abstract>
This document is a brief example using the Linuxdoc-SGML DTD.
</abstract>
</article>
<!-- Table of contents -->
<toc>
</toc>
<!-- Begin the document -->
<sect>Introduction
<p>
This is an SGML example file using the Linuxdoc-SGML DTD. You can format it 
using the command
<tt>/tscreen>&lt;verb&gt;
% sgml2text example.sgml
</verb></tt>
this will produce plain ASCII. You can also produce LaTeX, and HTML 
and GNU info.
</sect>The source
<p>
Not a language (no predefined tags), XML is a metalanguage:
- Defines tags and attributes
- Defines structural relationships

XHTML: hybrid HTML + XML
- HTML written according to XML (application of XML)
- Substitute of HTML
XML

Metalanguages

- GML
- SGML

Languages

- HTML
- XHTML
- RSS
- ... ...

Versions

- 10/2/1998: XML 1.0
- 4/2/2004: XML 1.0 Third Edition
- 16/8/2006: XML 1.0 Fourth Edition
- 26/11/2008: XML 1.0 Fifth Edition

- The four editions correct errors and clarify/detail the standard, but they don’t define a new standard
Versions

- 4/2/2004: XML 1.1
  - Updates 1.0 to new Unicode standard
  - W3C: “You are encouraged to create or generate XML 1.0 documents if you do not need the new features in XML 1.1; XML Parsers are expected to understand both XML 1.0 and XML 1.1”

  - Clarifies/details the standard, but they don’t define a new standard

Versions

- Extensible Markup Language (XML) 1.0 (Fifth Edition)
  - W3C Recommendation 26 November 2008
  - [http://www.w3.org/TR/2008/REC-xml-20081126/](http://www.w3.org/TR/2008/REC-xml-20081126/)
  - To get the last version:
    - [http://www.w3.org/TR/xml/](http://www.w3.org/TR/xml/)
Advantages

- Provides metadata for data → Improves searches
- Structured data → Allows fine grain updates
- Separates content (data) / presentation:
  - Changes of data / presentation are easier
  - Allows multiples views of the same data
La más grave de las enfermedades mentales tiene un nombre equivocado y provisional. Lo que hoy se denomina esquizofrenia probablemente con varias enfermedades con causas y pronósticos distintas. Pero esto no se sabrá con certeza hasta que no se complete más un enrevesado puzzle en el que hay numerosas piezas que tienen que ver con el desarrollo fetal, los factores ambientales y la genética. Ahora hace falta ponerlos en su sitio para aclarar las causas y mejorar el panorama actual: hoy, como hace milenios, una de cada 100 personas sufre esquizofrenia, y aún hay un 25% que no se recupera ni con tratamiento.

Para estudios en familias de
esquizofrénicos, se sabe que la enfermedad no obedece a un solo gen. La prueba principal la dan los genetistas genéticamente idénticos (monocigóticos): sólo un en menor de la mitad de los casos ambos son
esquizofrénicos. Pero tanto los hijos del genel no afectado como los del genel esquizofrénico tienen un riesgo similar (menor del 20%).
Advantages

- Allows you to create markup languages for specific domains:
  - Chemistry: Chemical Markup Language (CML)
  - Mathematics: Mathematical Markup Language (MathML)
  - Music: MusicXML
  - Monetary information: Open Financial Exchange (OFX)
  - Human resources (job offers, CVs, etc.): HR-XML
Exercise
Find the specification of a markup language based on XML

Advantages

- Self-describing data:
  - Much computer data from the last 40 years is lost because the data format is unknown nowadays
  - XML is a simple data format
  - The removal of bytes does not noticeably corrupt the remaining content
Advantages

- Self-describing data:

  `<persona id="p110" sexo="m">`  
  `<nombre>Pedro López</nombre>`  
  `<direccion>de los Palotes, 120</direccion>`  
  `<fnacimiento>30/06/1970</fnacimiento>`  
  `</persona>`

  `--------
p110;m;Pedro López;de los Palotes, 120; 30/06/1970`  
  `--------`

  0A 11 3C 2E 52 78 90 AA BC EA ...

Advantages

- Improves interchange of data among applications:
  - XML is a non-proprietary format
  - XML is very well documented (no secrets)
  - XML is easy to read and write

- Example: Open Financial Exchange (OFX)
  - XML application for describing and storing financial data
  - Used by: Quicken, Microsoft Money, GnuCash, etc.
What do I need to use XML

Minimum:
- To edit:
  - ASCII standard editor:
    - Microsoft Notepad
  - To display:
    - XML compatible browser:
      - Microsoft Internet Explorer 5
      - Netscape 6

Recommended:
- To edit:
  - Notepad++
  - EditiX
  - XML Spy
- To display:
  - Microsoft Internet Explorer 8 with IE XML/XSL Viewer Tools
  - Mozilla Firefox 3
My first XML document

- File name:
  - Short and easy to remember
  - Don’t use special characters or space character
    - Better only English alphabet
- File extension: .xml
- Editor:
  - Windows Notepad
- Viewer:
  - Microsoft Internet Explorer or Mozilla Firefox

Example: helloworld.xml

```xml
<?xml version="1.0" ?>
<document>
  Hello world!
</document>
```
Exercise
Write an XML document that stores the names of countries and the names of the corresponding capitals

```xml
<?xml version="1.0"?>
<document> Hello world! </document>
```
My first XML document

- countries.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<countries>
    <country name="Kazakhstan" capital="Astana" />
    <country name="Kyrgyzstan" capital="Bishkek" />
    <country name="Spain" capital="Madrid" />
</countries>
```
Error de lectura XML: mal formado
Número de línea 3, columna 36:

<country name="Kazakhstan" capital="Astana"/>

Error de lectura XML: etiqueta sin pareja. Se esperaba: <countries>.
Número de línea 6, columna 3:

</countries>
My first XML document

- But there is often more than one way to organize the data, depending on your needs:

```xml
<?xml version="1.0" encoding="UTF-8"?>
<countries>
  <country>
    <name>Kazakhstan</name>
    <capital>Astana</capital>
  </country>
  <country>
    <name>Kyrgyzstan</name>
    <capital>Bishkek</capital>
  </country>
  <country>
    <name>Spain</name>
    <capital>Madrid</capital>
  </country>
</countries>
```
My first XML document

- Attaching a style sheet to an XML document: helloworld-style.xml

```xml
<?xml version="1.0" ?>
<?xml-stylesheet type="text/css" href="helloworld.css" ?>
<document>
Hello world!
</document>
```
My first XML document

- Style sheet: helloworld.css

document {
display: block;
margin: 10%;
font-size: 36pt;
font-weight: bold;
color: blue;
text-align: center;
}
Hello world!

XML applications
XML applications

- Each specific XML-based markup language is called an XML application
  - Application does not mean a program that uses XML
  - Application means the use of XML to a specific domain
- Application = markup language = semantics and vocabulary

XML applications

- Chemical Markup Language:
  ```xml
  <?xml version="1.0" ?>
  <cml xmlns="http://www.xml-cml.org/schema/cml2/core"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="http://www.xml-cml.org/schema/cml2/core/cmlCore.xsd">
  <molecule title="Water">
    <atomArray>
      <atom id="a1" elementType="H" hydrogenCount="0" />
      <atom id="a2" elementType="O" hydrogenCount="2" />
      <atom id="a3" elementType="H" hydrogenCount="0" />
    </atomArray>
    <bondArray>
      <bond atomRefs2="a1 a2" order="1" />
      <bond atomRefs2="a2 a3" order="1" />
    </bondArray>
  </molecule>
  </cml>
  ```
Mathematical Markup Language (Wikipedia):

Mathematical Markup Language (MathML) is an application of XML for describing mathematical notations and capturing both its structure and content. It aims at integrating mathematical formulae into World Wide Web pages and other documents. It is a recommendation of the W3C math working group.
XML applications

\[ x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \]

XML applications

- RSS:
  - XML application used for content syndication in blogs, newspapers, etc.
  - Useful for any site that wants to provide a continuing feed of new information
  - Normally, an RSS document is associated to a group of HTML pages
XML applications

```xml
<?xml version="1.0" ?>
<rss version="1.0">
  <channel>
    <title>An example of RSS</title>
    <link>http://www.ua.es/</link>
    <description>This is the description</description>
    <language>en</language>
    <copyright>(c) 2011 Sergio Lujan Mora</copyright>
    <item>
      <title>Title</title>
      <description>Description</description>
      <link>http://www.ua.es/</link>
    </item>
  </channel>
</rss>
```

XML applications

- Open XML Format:
  - Microsoft Office 2007 (.docx, .xlsx, etc.)
XML applications
Scalable Vector Graphics:
- Format for describing two-dimensional vectorial graphics
- Many traditional drawing programs can save SVG files just like their native formats
- Requires special display software:
  - Microsoft Internet Explorer 9
  - Mozilla Firefox 3+

```xml
<?xml version="1.0"?>
<svg xmlns="http://www.w3.org/2000/svg"
     width="12cm" height="8cm">
  <title>The pink triangle!!</title>
  <text x="10" y="15">This is SVG!</text>
  <polygon style="fill: pink" points="0,311 180,0 360,311" />
</svg>
```
Exercise

Using SVG make a draw with the following objects: square, rectangle, circle, triangle
Structure of a document

- Logic structure
- Structure of a document
- Definition of a DTD
Logic structure

- XML is based on a containment model:
  - Each XML element (tag) can contain text, other elements, or a mix of both text and other elements
- The first question:
  - What contains what?
  - Which information is a part of which other information?
Structure of a document

- Structure is defined by a DTD (Document Type Definition) or an XML Schema
  - Optional
  - Defines the language: the vocabulary (elements, tags) and the grammar (relationships between elements)

Structure of a document

- “Well-formedness”: follows the rules of XML
- “Valid”: follows the rules of a DTD or XML Schema
Structure of a document

- XML declaration → Processing instruction:
  ```xml
  <?xml
  version="1.0"
  encoding="ISO-8859-1"
  standalone="yes" ?>
  ```
- This declaration must be the first thing in the file
- If this declaration is not included, the default values are:
  - version: 1.0
  - encoding: UTF-8
  - standalone: yes
Structure of a document

Comments:
- All data inside a comment is ignored by an XML processor
- Comments cannot come before the XML declaration
- Comments cannot be placed inside a tag
  ```xml
  <!-- This is a one line comment -->
  ```

  ```xml
  <!-- This is a multiple lines comment -->
  ```

Single root element:
- An XML document has a root element that completely contains all other elements of the document
- Root element = document element
Structure of a document

- Tags → Define an element:
  - Delimited by a start-tag (opening tag) and end-tag (closing tag)
  - Every start-tag must have a corresponding end-tag
    `<BOOK></BOOK>`
- Empty tags (start and end tag together):
  `<BOOK/>`
- Elements may nest but may not overlap:
  `<b><i>Some important text</i></b>`

Structure of a document

- Naming conventions of element name:
  - Made up of one or more characters
  - Begin with a letter or an underscore `_`
  - Subsequent characters may include letters, digits, underscores `_`, hyphens `-`, and periods `.`
  - They cannot include white spaces
  - Important: upper- and lowercase are different!
Structure of a document

- **Naming conventions:**
  - Element names are flexible and can contain any number of letters and digits in either upper- or lowercase
    - `<countries>`
    - `<Countries>`
    - `<COUNTRIES>`
  - It is important to choose one convention and stick to it

- **Attributes:**
  - Elements can have attributes
  - An attribute is a name-value pair separated by an equal symbol =
  - Attribute names follow the same rules as element names
  - The order inside an element is not important
  - The value of the attribute always between single quotes ' ' or double quotes " 
  - Single quotes and double quotes as values:
    - `&quot;` and `&apos;`
Structure of a document

- Predefined attributes:
  - The prefix `xml:` is reserved for XML specification
  - `xml:lang`
    - Identifies the language of element’s content
  - `xml:space`: default | preserve
    - Indicates if the white spaces are significant
  - `xml:id`
    - Unique identifier in the whole document

Example of `xml:lang`:

```xml
<p xml:lang="en">The quick brown fox jumps over the lazy dog.</p>
<p xml:lang="en-GB">What colour is it?</p>
<p xml:lang="en-US">What color is it?</p>
<sp who="Faust" desc='leise' xml:lang="de">
  <l>Habe nun, ach! Philosophie,\</l>
  <l>Juristerei, und Medizin\</l>
  <l>und leider auch Theologie\</l>
  <l>durchaus studiert mit heißem Bemüh’n.</l>
</sp>
```
Structure of a document

- White spaces:
  - Tabulator
  - Line feed
  - Carriage return
  - White space

- Normalization of newline characters:
  - Macintosh CR → LF
  - MS-DOS / Windows CR+LF → LF
  - Unix LF

Structure of a document

- Special characters:

<table>
<thead>
<tr>
<th>Carácter</th>
<th>Codificación</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;</code></td>
<td><code>&amp;lt;</code></td>
</tr>
<tr>
<td><code>&gt;</code></td>
<td><code>&amp;gt;</code></td>
</tr>
<tr>
<td><code>&amp;</code></td>
<td><code>&amp;amp;</code></td>
</tr>
<tr>
<td><code>&quot;</code></td>
<td><code>&amp;quot;</code></td>
</tr>
<tr>
<td><code>'</code></td>
<td><code>&amp;apos;</code></td>
</tr>
</tbody>
</table>
Structure of a document

- CDATA sections: XML processor does not try to interpret the content
- Syntax:
  ```xml
  <![CDATA[
  ...
  ]]>]
  ```
- CDATA section cannot be nested or overlapped
- Useful if you want to include large blocks of text that have a lot of &, <, >, `, and "

Exercise

Algorithm example
Structure of a document

```xml
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
<algoritmo>
a = 5

si a < 5 entonces
    muestra "a < 5"
sino si a > 10 entonces
    muestra "a > 10"
</algoritmo>
```

Exercise

Web page
Use W3C markup validator
(http://validator.w3.org/)
Structure of a document

<?xml version="1.0"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html>
<head>
<title>Una prueba del CDATA</title>
<script type="text/javascript">
  if(a < 5 && a > 1)
    alert("El valor de a no es correcto");
</script>
</head>
<body>
<p>Una prueba del CDATA</p>
</body>
</html>
Exercise
Write an XML document that stores the information about the books of a library: title, author, editor, number of pages, ISBN, etc.

Other XML technologies
The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential. W3C is a forum for information, commerce, communication, and collective understanding. On this page, you'll find W3C news, links to W3C technologies, and ways to get involved. New visitors can find help in Finding Your Way at W3C. We encourage you to read the Procedures and learn more about W3C.

W3C Requests ’905 Patent Re-Examination

2003-10-29: Acting on the advice of the W3C Liaison Officer Group, W3C has presented the United States Patent and Trademark Office with prior art establishing that US Patent No. 5,833,896 (the ’905 patent) is invalid. W3C Director Tim Berners-Lee has written an unprecedented request to US Under Secretary of Commerce for Intellectual Property James E.rogan to take action to remove the patent to allow operation of the Web. Please refer to the mailing (News archive).

W3C Presents W3C Day Japan on 14 November in Tokyo

2003-10-29: W3C Day Japan 2003 (in Japanese) will be held on 14 November 2003 at Keio University Mita Campus.

W3C Holds its First Outreach Event in Mainland China

2003-10-21: The China International Forum on WWWs Development 2003 will be held in Beijing on 12-13 November. Ivan Herman, Philipp Hoehl, Richard Ishida, Shizhong Ji, Judy Brewer, and Matthew May present keynotes and tutorials. Attendees will discuss the future of the Web, accessibility, SVG, the mobile Web, and internationalization. Registration is open. The event is co-organized by the China Computer Federation and the W3C Office in Hong Kong. Read the press release (News archive).

MathML 2.0 Second Edition Is a W3C Recommendation

2003-10-21: The World Wide Web Consortium today released the Mathematical Markup Language (MathML) Version 2.0 (Second Edition) as a W3C Recommendation. The specification has been reviewed by the W3C Membership, who favor its adoption by industry. MathML is an XML application that allows mathematical notation and content to be served, received and processed on the Web. The second edition contains clarifications and some
Other XML technologies

- XML Namespaces
- XML Schemas
- XPath and XPointer
- XLink, XBase, and XInclude
- XQuery
- XSL/XSLT
- DOM y SAX

Other XML technologies

- XForms
- XMI (XML Metadata Interchange)
- XML-QL
- XML Encryption
- XML Signature
- XQL
Software

- Free or pay
- Good place to find programs:
  - http://www.xmlsoftware.com/
Software

- **Notepad++**
- Free (GNU General Public Licence)
- General purpose editor, supports many programming languages
- Support of XML is minimum:
  - Syntax highlight
  - Tag coupling
  - Folding (collapsing):
    - This means that certain lines of your text can be hidden based on certain traits
Software

- Microsoft XML Notepad 2007
- Free
- Very simple and easy XML editor
- It is not necessary to know how to write an XML document
Software

- ezDTD 1.5
- Free
- DTD editor, both for SGML and XML
- Generates documentation of DTDs in HTML format
Software

- **Peter’s XML Editor**
- Free
- XML and DTD editor
- Checks well-formedness and validates XML document against DTD
- Three views of the same document:
  - Source code
  - Tree
  - Internet Explorer
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<DOCTYPE BIBLIOTECA [New source for fulltype...]/>

<BIBLIOTECA>
  <LIBRO>
    <CODIGO>"11"</CODIGO>
    <TITULO>El libro para todos</TITULO>
    <AUTOR>Gonzalo Lozano Mora</AUTOR>
    <AÑO>2001</AÑO>
    <EDITORIAL>Alfa Pressa</EDITORIAL>
  </LIBRO>
  <LIBRO>
    <CODIGO>"12"</CODIGO>
    <TITULO>Libro de ejercicios</TITULO>
    <AUTOR>Maria Lopez</AUTOR>
    <AÑO>1999</AÑO>
    <EDITORIAL>Biblioteca</EDITORIAL>
  </LIBRO>
</BIBLIOTECA>
Software

- **Cooktop**
  - Free
  - XML, DTD, XMLSchema, and XSL/XSLT editor
  - Checks well-formedness and validates XML document against DTD
  - Provides support for XPath and XSL/XSLT
Software

- **EditiX**
  - Free and pay version
  - Checks well-formedness and validates XML document against DTD
  - Special features:
    - DTD ↔ XML Schema conversor
    - SVG viewer
```xml
<?xml version="1.0" encoding="UTF-8"?>
<library>
  <name>Library</name>
  <area>
    <id>1</id>
    <name>Science Fiction</name>
    <books>
      <book>
        <id>1</id>
        <title>The best book in the world</title>
        <author>Dan Brown</author>
        <description>A really good book III</description>
      </book>
      <book>
        <id>2</id>
        <title>War and Peace</title>
        <author>Leo Tolstoy</author>
        <description></description>
      </book>
    </books>
  </area>
</library>
```
Software

- XMLSpy
- Pay
- From company Altova
- Provides support for nearly everything related to XML
- Multiple views of a document
XML 1.1

- XML 1.0 based on Unicode 2.0:
  - Explicitly listed all the characters than could be used in XML names (element names, attribute names, entity names, and processing instruction targets)
  - Characters not defined in Unicode 2.0 are not allowed in names
    - You can’t write XML 1.0 names in Amharic, Burmese, or Cambodian because those languages (and the corresponding scripts) weren’t added to Unicode until version 3.0
    - But you can use whatever you want in PCDATA (text content)
XML 1.1

- XML 1.1 independent of any particular Unicode version (currently 4.0):
  - Allows to use new characters from new languages (e.g., Burmese, Cambodian, Mongolian, etc.)
  - Allows to use a new newline character (Unicode code 133, NEL)
    - NEL is used as a line terminator on some IBM mainframe systems
  - Forbids the direct inclusion of control characters with Unicode code between 128 and 159 (except NEL)
    - There characters can be included as numeric character references such as &#135; or &amp;#xBC;
  - Allows the inclusion of some additional control characters
  - Includes Unicode normalization

Conclusion:
- W3C: “You are encouraged to create or generate XML 1.0 documents if you do not need the new features in XML 1.1; XML Parsers are expected to understand both XML 1.0 and XML 1.1”
- You don’t need XML 1.1 unless you need to write elements (markup) in new languages; otherwise, XML 1.1 makes documents incompatible with the large amount of XML 1.0 software