Characteristics and Merits of XML

Jae Hwa Chang

InfoLumi, Korea

The 2nd Council of Asian Science Editors’ Preconference Workshop I
Hanoi University of Science and Technology, Hanoi, Vietnam
August 20, 2015
What is XML?

- Stands for eXtensible Markup Language
- Not a programming language, but a markup language that is similar to HTML
- Designed to carry data, not to display data
- Designed to exchange and store data
- Tags are not predefined
- Designed to be self-descriptive
XML vs. HTML

- XML is not a replacement for HTML
- XML is no more a programming language than HTML is

- XML was designed to transport and store data
  ◦ Focus on what data is

- HTML was designed to display data
  ◦ Focus on how data looks
Where XML is used

- Used to exchange data
  - Data can be read by different incompatible applications

- Can be used to share & store data
  - Stored in plain text format
    → provide independent way of storing data
    → easy to be shared

- Used to create new internet languages
  - RSS, RDF, and SMIL
Benefits of XML

- Readable and understandable
  - XML tag names are readable and convey the meaning of the data

- Easy to code
  - Data structure follows a noticeable and useful pattern, making it easy to manipulate and exchange the data

- Compatible and portable
  - Any application that can process XML can use your information, regardless of platform

- Extendable
  - Create your own tags, or use tags created by others
<xml version="1.0" encoding="UTF-8"> XML declaration
<library> Root element
  <book> Parent element
    <author> Jacques Barzun </author> Child elements
    <title> On writing, editing, and publishing </title>
    <year> 1986 </year>
    <publisher> University of Chicago Press </publisher>
  </book>
</library>
XML basic rules

- XML document must have a header which tells that it is an XML document
  - `<?xml version="1.0" encoding="UTF-8"?>`
- XML document must have only one root element
- All XML elements must have a closing tag
  - `<head>` Correct `</head>`
  - `<head>` Incorrect `</tail>`
- XML tags are case sensitive
  - `<Body>` Correct `</Body>`
  - `<Body>` Incorrect `</body>`
Element

- Tags that are used to create XML document
  - Opening and closing tags represent the start and end of an element
- An element can contain other elements, text, attributes, or a mix of all of the above
- Elements are extensible
- Elements in an XML document form a document tree

```
<root>
  <child>
    <subchild>......</subchild>
  </child>
</root>
```
Attribute

- Attributes provide additional information about an element
- Attributes must be quoted
- Elements vs. Attributes

```
<book>
  <category>medical</category>
  <title>Surgery</title>
  <author>John Taylor</author>
</book>
```

```
<book category="medical">
  <title>Surgery</title>
  <author>John Taylor</author>
</book>
```
Attributes vs. Elements

- Attributes are not easily expandable for future changes
- Attributes values are not easy to test against a DTD
- Attributes cannot contain multiple value vs. Elements can
- Attributes cannot describe structure vs. Elements can
Entity

Entity references
- Some characters have a special meaning in XML

\[ \text{<note>p-value } < 0.05 \text{ was considered</note>} \]

Predefined entity references
- \&lt; \< less than
- \&gt; \> greater than
- \&\&\& \& ampersand
- \&apos; \' apostrophe
- \&quot; \" quotation mark
DTD (Document Type Definition)

- A set of markup declarations that define a document type
- To define the structure of an XML document
  - Defines the structure with a list of legal elements and attributes
- XML files can carry a description of its own format
- Independent groups of people can agree on a standard for interchanging data
- Can verify that the data you receive from the outside world is valid
Well-Formed XML

- No syntax, spelling, punctuation, grammar errors, etc. in its markup
- Errors can cause XML document to not parse
- XML parser reads XML documents and interprets or parses the code according to the XML standard
Valid XML

- Element structure and markup of the XML document matches a defined standard of relationships
- For the XML document to be valid, it must follow all the rules set forth in the DTD
- Validate using XML validator
- Valid $\Rightarrow$ Also well formed VS.
  Well formed $\Rightarrow$ Not necessarily valid
- Can create XML documents without a DTD VS.
  XML document can't be considered valid without a document type
References
