

Introduction to XML



Shafiq Ur Rahman

Center for Research in Urdu Language Processing
National University of Computer and Emerging
Sciences, Lahore

Overview

- ▶ XML
- ▶ DTD
- ▶ Related Standards

What is XML

- ▶ XML stands for eXtensible Markup Language
- ▶ Set of rules for defining semantic tags to break a document into parts and identify different parts of it
- ▶ Meta-Markup Language

Document

Mrs. Mary McGoon
1400 Main Street
AnyTown, AnyProvince
AnyCountry 12345

XML Document

```
<?xml version="1.0"?>
<address>
  <name>
    <title> Mrs. </title>
    <first-name> Mary </first-name>
    <last-name> McGoon </last-name>
    <street> 1400 Main Street </street>
    <city> AnyTown <city>
    <province> AnyProvince </province>
    <country> AnyCountry </country>
    <postal-code> 12345 </postal-code>
  </name>
</address>
```

Tags

▶ *< Tag_name >*

▶ *Tag_name*

- Starts with letter or underscore (_)
- Subsequent characters include letters, digits, underscores, hyphens and periods

▶ *<name>* *<_8>* *<object.member>*

▶ *<first name>* *<8digit>*

Tags ...

► Types

- Starting Tag

<name>

<address>

- Ending Tag

</name>

</address>

- Empty Tag

<middle_initial/>

Elements

▶ Simple Element

- `<tag> content </tag>`

```
<first_name>
```

```
  Shafiq
```

```
</first_name>
```

```
<first_name> Shafiq </first_name>
```

Elements ...

► Compound Element

- `<tag1> <tag2> content </tag2> </tag1>`

`<name>`

`<first_name> Shafiq </first_name>`

`<last_name> Rahman </last_name>`

`</name>`

Elements ...

► Empty Element

- `<tag> </tag>`
- `<empty_tag/>`

`<middle_initial> </middle_initial>`

`<middle_initial/>`

Attributes

- ▶ Elements may have attributes
- ▶ Name-value pair inside Starting tags and Empty tags
 - `<tag attr-name=attr-value>`

```
<middle_name initial="u" > ur  
</middle_name>
```

```
<IMG width='89' height="36"  
title= "Queen's birthday" />
```

XML Document Rules

1. Must start with an XML declaration
 - Processing Instruction

```
<? xml version="1.0"  
      encoding="UTF8"  
      standalone="yes"  
?>
```

```
<? Xml version="1.0" ?>
```

XML Document Rules ...

2. One element, Root Element, must contain all other elements
 - Tree structured document

```
<?xml version="1.0"?>  
<address>  
...  
</address>  
<address1>...</address1>
```

XML Document Rules ...

3. Non-empty elements must use corresponding start and end tags

```
<first-name> Mary </first-name>
```

```
<first-name> Mary </FIRST-NAME>
```

```
<first-name> Mary </firstname>
```

XML Document Rules ...

4. Use completely nested elements, no overlaps

```
<name>
```

```
  <first-name> Mary </first-name>
```

```
</name>
```

```
<name>
```

```
  <first-name> Mary
```

```
</name> </first-name>
```

XML Document Rules ...

5. Attribute values must be in quotes

```
<img height='36" ' width="96" />
```

```
<img height=36 width=96 />
```

XML Document Rules ...

6. Use < and & to start tags and entities

```
<src>
```

```
  if (x < y)
```

```
</src>
```

```
<img height='>36"' width='96"' />
```

XML Document

- ▶ XML document conforming to these 6 rules is a Well-Formed document
- ▶ Every XML document must be a well-formed document at the least

XML Document

```
<?xml version="1.0"?>
<address>
  <name>
    <title>          Mrs.          </title>
    <first-name>     Mary           </first-name>
    <last-name>      McGoon        </last-name>
    <street>         1400 Main Street </street>
    <city>           AnyTown       </city>
    <province>       AnyProvince  </province>
    <country>        AnyCountry   </country>
  </name>
</address>
```

Additional things

- ▶ Comments:

```
<!-- Here is a comment -->
```

```
<!-- A comment that contains an element  
<first-name> ... </first-name> -->
```

- ▶ It can contain anything except a double hyphen which must occur at the end.
- ▶ Appear anywhere in XML document

Additional things

- ▶ Entity References: these are replaced by character data
- ▶ Five predefined entities:
 - `<`; `<` `&`; `&`
 - `>`; `>` `"`; `"`
 - `'`; `'`

```
<img title='Queen&apos;s mother' />
```

```
<src> if (x &lt; y) </src>
```

What is Markup?

- ▶ Any thing other than character data in an XML document is Markup
 - Processing Instructions
 - Tags
 - Comments
 - Entity References
 - ...

Document Type Definition (DTD)

- ▶ Defines the set of elements, attributes and entity references that may appear in an XML document
- ▶ DTD defines the structure of document
- ▶ DTD defines the schema

XML Document

- ▶ Internal DTD

```
<?xml version="1.0"?>
```

```
<!DOCTYPE address [
```

```
  <!ELEMENT address (name)>
```

```
  <!ELEMENT (title)>
```

```
  <!ELEMENT title (#PCDATA)>
```

```
<address> <name>
```

```
  <title> Mrs. </title>
```

```
</name> </address>
```

XML Document

- ▶ External DTD

```
<?xml version="1.0"?>
```

```
<!DOCTYPE address SYSTEM "abc.dtd">
```

```
<address>
```

```
  <name>
```

```
    <title> Mrs. </title>
```

```
  </name>
```

```
</address>
```

Well-Formed & Valid document

- ▶ An XML document is Well-Formed if it conforms to the XML rules
- ▶ An XML document is valid if, in addition to being well-formed, it conforms to DTD
- ▶ All documents need not be valid

Element Declaration

▶ Simple Element

- `<!ELEMENT name type >`

`<!ELEMENT first-name (#PCDATA)>`

`<!ELEMENT date-of-birth (#PCDATA)>`

- `#PCDATA`: parsed character data

Element Declaration

► Compound Element

- `<!ELEMENT name child-list >`

- *Child-list*: One child

 - `<!ELEMENT address (name) >`

- *Child-list*: Zero or one child (optional)

 - `<!ELEMENT name (middle-initial?) >`

Element Declaration

- *Child-list*: Sequence of children
<!ELEMENT name (title, first-name, middle-initial?, last-name)>
- *Child-list*: Zero or more children
<!ELEMENT address-book (address*)>
<!ELEMENT document
 (chapter-title, chapter)* >
- *Child-list*: one or more children
<!ELEMENT address-book (address+)>

Element Declaration...

- *Child-list*: Choice (one among many)
<!ELEMENT mode-of-payment
 (cash | credit-card | cheque)>

<!ELEMENT article (title, (paragraph |
 photo | sidebar)*, signature?)>

Element Declaration...

- *Child-list*: Mixed content

```
<!ELEMENT parent  
    (child1 | child2 | #PCDATA)* >
```

severely restricts the structure

```
<!ELEMENT article (title, (paragraph |  
    photo | sidebar)*, signature?, #PCDATA) >
```

Element Declaration...

- Empty Elements

```
<!ELEMENT line-break EMPTY>
```

Comments

- ▶ Same as in XML document

```
<!-- address is the root element -->
```

```
<!ELEMENT address (name)>
```

Attribute Declaration

- ▶ `<!ATTLIST element-name Attr-name type def-value>`
- ``
- `<!ELEMENT img EMPTY>`
- `<!ATTLIST img height CDATA "12"
 wight CDATA "48">`
- `<!ATTLIST img height CDATA "12">`
- `<!ATTLIST img width CDATA "48">`

Attribute Declaration

- ▶ Attributes may not have good default values
 - `<!ELEMENT img EMPTY>`
 - `<!ATTLIST img height CDATA #REQUIRED>`
 - `<!ATTLIST img height CDATA #IMPLIED>`
 - `<!ATTLIST img height CDATA #FIXED "12">`
- ▶ CDATA: character data, may not use <

Attribute Types

- ▶ Enumerated
- ▶ ID
- ▶ IDREF
- ▶ NMTOKEN
- ▶ ENTITIY
- ▶ ...

Entity Declaration

- ▶ Declare additional entity references
- ▶ `<!ENTITY name "replacement text">`
 - `<!ENTITY CR05 "Copyright 2005">`
 - `<!ENTITY SR "Shafiq Rahman">`
 - `<copyright> &SR; --- &CR05; </copyright>`
 - `<!ENTITY CR05 "&SR; --- Copyright 2005" >`
 - `<!ENTITY SR "shafiq Rahman &CR05" >`

XML Document

```
<?xml version="1.0"?>
<!DOCTYPE address SYSTEM "example.dtd">
<address>
  <name>
    <title> Mrs. </title>
    <first-name> Mary </first-name>
    <last-name> McGoon </last-name>
    <street> 1400 Main Street </street>
    <city> AnyTown </city>
    <province> AnyProvince </province>
    <country> AnyCountry </country>
    <postal-code> 12345 </postal-code>
  </name>
</address>
```

Complete DTD

- ▶ `<!ELEMENT address (name)>`
- ▶ `<!ELEMENT name (title,first-name, middle-initial?,last-name,street,city, province,country,postal-code)>`
- ▶ `<!ELEMENT title (#PCDATA)>`
- ▶ `<!ELEMENT first-name (#PCDATA)>`
- ▶ `<!ELEMENT middle-initial (#PCDATA)>`
- ▶ `<!ELEMENT last-name (#PCDATA)>`
- ▶ `<!ELEMENT street (#PCDATA)>`
- ▶ `<!ELEMENT city (#PCDATA)>`

Complete DTD

- ▶ `<!ELEMENT province (#PCDATA)>`
- ▶ `<!ELEMENT country (#PCDATA)>`
- ▶ `<!ELEMENT postal-code (#PCDATA)>`

- ▶ `<!ATTLIST country continent
"AnyContinent">`

Other XML-related technologies

- ▶ CSS
- ▶ XML Schema
- ▶ XSLT

- ▶ DOM
- ▶ SAX

Resources

- ▶ IBM dW XML zone
 - www-106.ibm.com/developerworks/xml
- ▶ XML
 - W3.org/TR/REC-xml
- ▶ XML Schema
 - W3.org/TR/xmlschema-0
- ▶ DOM
- ▶ W3.org/TR/DOM-Level-2-core/

شکر یہ



