XML

CS174

Chris Pollett

Oct 27, 2008.

Outline

- Introduction to XML
- Document Type Definitions
- Internal versus External DTDs
- Namespaces
- XML Schemas

Introduction to XML

- Recall that HTML was originally specified as an SGML (Standard Generalized Markup Language) doctype.
- HTML only provides limited semantic information about a document. You can tell if something is in <h1> tags that is probably important, but not much else.
- Starting in 1998 a stripped down version of SGML called XML (extensible markup language) was developed to make it easier to create new tag-based mark-up languages where the tags can be used to carry whatever semantic information is desired.

The Syntax of XML

- A new XML language can be specified in one of two ways:
  - Give a DTD (Document Type Definition) -- this is closer to the SGML way of specifying languages.
  - Give an XML schema -- unlike DTDs such schemas are also XML documents so can easily be parse with XMLParsers. Further schemas can be more detailed.
- In both a DTD and a schema, one specifies:
  - What tag elements exist in the language.
  - What subelements or data a given element is allowed to contain.
  - What attributes an element has and what their values can be.
- Tags are case-sensitive in XML, and every tag must have a close tag, although the abbreviation <element /> works as an implicit close tag.

XML Document Structure and Entities

- Let's look at XML document:

```
<patient type="out of state">
    <![CDATA[This is data that will not be parsed by the XML parser even if it have
    tags in it like this: <tag>]]>
      <name><first>John</first><last>Smith</last></name>
      <insurerID>&kaiser;</insurerID>
</patient>
```

- Two files are usually associated with such a document:
  - one specifies what tags will be used with this document
  - another specifies how those tags should be styled.
- One can abbreviate parts of a document using entities. For example, &kaiser; above might abbreviate some string of numbers.
- Associated with any XML document is a *document entity*. This can be thought of as the root of the tree associated with the tags of the document. This entity thus can represent the whole document or the parts of it which do not change.
- Entities can also be used to reference binary data such as images.

Document Type Definitions

- A DTD is used to specify:
  1. what tags will be used in a document.
  2. what kind of data or subtags these tags can hold.
  3. what attributes the tags can have and what values for these attributes are legal.
  4. what entities can occur in the document.
- DTDs can be *internal* (occur in the XML document itself) or *external* (occur in some separate file).
- A DTD consists of a sequence of declarations enclosed in the block of a DOCTYPE markup declaration.
- Items 1-4 above correspond to the declaration of type <!ELEMENT …> (1 and 2), <!ATTLIST …>, and <!ENTITY …>
- There is also a tag called <!NOTATION …>which is sometimes used.

<!ELEMENT ..>

- Basic syntax of the tag is:

  <!ELEMENT element_name (list of names of child elements) SYSTEM Location NDATA NotationName >

- Some examples:

  <!ELEMENT memo (from, to, date) > ---- CDATA[SYSTEM and NDATA don't have to used

  <!ELEMENT dept_script SYSTEM "dept.php" NDATA "php" >
  <!NOTATION php SYSTEM "/usr/bin/php" > .

  <!ELEMENT person(parent+, age, spouse?, sibling*) > ---- + is one than one, ? is optional, and * is 0 or more.

  <!ELEMENT element_name (#PCDATA) > ---- pcdata = parsable character data can also use EMPTY for no sub-tags or character data or ANY if you want to allow everything .

<!ATTLIST ..>

- The general way to specify a list of attributes is:

<!ATTLIST element_name

attribute_name_1 attribute_type [default_value] … attribute_name_n attribute_type [default_value] >

- For example,

<!ATTLIST airplane places CDATA “4”>

<!ATTLIST airplane engine_type CDATA #REQUIRED><![CDATA[must have the field]>

<!ATTLIST airplane price CDATA #IMPLIED><![CDATA[no default value is given]>

<!ATTLIST airplane manufacturer CDATA #FIXED “cessna”><![CDATA[all instances must have the same value]>

<!ENTITY ..>

- Entities come in two flavors:
  - General entities that can be referenced anywhere in an XML document.
  - Parameter entities which can be referenced only in markup declarations .
- The general form of an entity declaration is:

  <!ENTITY [%] entity_name “entity value” >

  --- % is used when it is a parameter entity.
- For example,

  <!ENTITY cp “Chris Pollett” >

  <!ENTITY cool_pic SYSTEM “/usr/local/cool_pic.jpg”>

Example DTD

<?xml version = “1.0” encoding =“utf-8” ?>
<!-- plane.dtd fragment -->
<!ELEMENT planes_for_sale (ad+)>
<!ELEMENT ad (year, make, model, color, price?, seller) >
…
<!ELEMENT seller (#PCDATA)>

<!ATTLIST seller phone CDATA #REQUIRED>

<!ENTITY c “cessna” >

Internal versus External DTDs

- There are two ways to associate a DTD with an XML document:
- Internally --

<?xml version="1.0" encoding="utf-8" ?>

<!DOCTYPE planes [<!--DTD for planes]>

<!-- The planes document -->

- Externally --

<!DOCTYPE planes_for_sale SYSTEM "planes.dtd" >

Namespaces

- It is sometimes useful to mix and match markup from several XML languages into one file.
- For instance, you might want to have mark-up for HTML as well as mark-up for SVG, a vector graphics language, and MathML, a language for mathematics.
- One could do this with a declaration like:

<div xmlns="http://www.w3.org/1999/xhtml"

xmlns:svg="http://www.w3.org/2000/svg"

xmlns:mml="http://www.w3.org/1998/Math/MathML" >

Now to use an html tag within this div I don't need a prefix.

To use the svg tag rect, I might write <svg:rect …>.

- This solves the problem where one might have two languages both with the same name for a tag. For intances, both with a <table> tag.

XML Schemas

- DTDs are defined in a language which is different from XML.
- We also cannot finely control what kinds of character data can appear within tags.
- The XML schema language was created to address both these issues.

More on Schemas

- The namespace for XML Schema is

http://www.w3.org/2001/XMLSchema.

- xsd is the common namespace to use when dealing with XML schema tags.
- A XML schema is used to define a target namespace of the language we are defining.
- A defining schema can be used to allow or restrict other schemas from being used with the target namespace.
- At its simplest, a schema document consists of a bunch of tags such as <element> <simpleType>, <complexType>, <sequence>, <all>,<restriction>, and enclosed with a <schema> tag.