

Distributed Software Development

XML: Structure and Parsing

Chris Brooks

Department of Computer Science
University of San Francisco

Department of Computer Science — University of San Francisco — p. 1/77

5-2: Outline

- About XML
- Structuring XML documents
- Using CSS to display XML
- Parsing with DOM
- Parsing with SAX

Department of Computer Science — University of San Francisco — p. 2/77

5-3: XML

- XML is a language for describing data
 - Really more of a meta-language
- XML itself provides metadata
 - Data types, relations between data objects, etc.
- Designed to be read, created, and consumed by programs.

Department of Computer Science — University of San Francisco — p. 3/77

5-4: Advantages of XML

- Well-defined, easy-to-manipulate structure
- Human-readable
- Extensible
- Metadata can be included directly with data
- Widely used

Department of Computer Science — University of San Francisco — p. 4/77

5-5: Things to note

- An XML document has two components:
 - tags (metadata)
 - content (data)
- Metadata serves to help an application make sense of the data.

Department of Computer Science — University of San Francisco — p. 5/77

5-6: Example

```
<?xml version="1.0"?>
<book>
  <author> J.R.R. Tolkien </author>
  <title> The Lord of the Rings </title>
  <volumes>
    <volume> Fellowship of The Ring </volume>
    <volume> The Two Towers </volume>
    <volume> Return of the King </volume>
  </volumes>
  <price> 14.95 </price>
  <publisher> Ballantine </publisher>
  <isbn> 0345340426 </isbn>
</book>
```

Department of Computer Science — University of San Francisco — p. 6/77

5-7: XML documents as trees

- An XML document can also be represented as a tree.
- This makes XML very easy to parse.
- The outermost element is the root element, and elements contained within it are children of that element.
- Content is stored at the leaves
- What would the tree for our Tolkien example look like?

5-8: Outline

- About XML
- **Structuring XML documents**
- Using CSS to display XML
- Parsing with DOM
- Parsing with SAX

5-9: Elements

- XML requires that every starting tag have a corresponding closing tag.
- Everything between a starting tag and a closing tag is called an *element*
- For example, `<volume>Return of The King </volume>` is an element
- So is everything between `<volumes>` and `</volumes>`
- As is everything between `<book>` and `</book>`.
- This means that elements must be nested.

5-10: Tags and elements

- Tags form the boundaries of elements, and give processing instructions to parsers.
 - Empty elements: `<coAuthor />` All information is contained in the tag.
 - Container elements: `<price> 14.95 </price>`
 - Comments: `<!-- here's a comment -->`
 - Declaration: `<!ENTITY jrirt 'J.R.R. Tolkien''>`
This provides a way to define variables or constants in a single location.
 - Entity reference: `<author> &jrirt </author>`

5-11: Attributes and Values

- You can also specify that an element has *attributes*
- These attributes can take on *values*
- This is helpful when you want to specify that an object belongs to one of a few types.

```
<book genre="fantasy" size="large"> ...
</book>
```

5-12: Attributes vs. Sub-elements

- We could rewrite the example above using subelements instead of attributes.
- When to use one over the other is largely stylistic.
 - Can always transform one into the other
- If a feature can only take on one of a few values, an attribute might make more sense.
- If we expect to extend the number of genres, a subelement is preferable.
- Also, order is preserved for subelements
 - Semantically, attribute/value pairs are treated as a dictionary.
- So, a list of authors should be done as subelements

5-13: ID attributes

- A particularly helpful attribute is ID - this lets you assign a reference to an element and refer to it later in the document.

```
<volume id="book1"> Fellowship of the Ring </volume>
<volume id="book2"> The Two Towers. Read this book after you've
finished <volumeref idref="book1" />.</volume>
```

- The ref tag refers to a previous volume
- This provides the XML parser with the information that this is a reference to a previous volume with id "book1".

5-14: Document Prolog

- If you've looked at XML that's used by other applications, you've probably noticed a lot of messy-looking stuff at the top.
- This is called the *document prolog*.
- This tells a client that the document is in XML and refers it to other document that indicate which tags are valid.

```
<?xml version="1.0" encoding="US-ASCII" standalone="no">
<!DOCTYPE book
  PUBLIC "-//USF //DTD Book 1.8//EN"
  "http://www.fooobar.com/DTDs/lotr.dtd"
  [
    <ENTITY jrrt "J.R.R. Tolkien">
    <ENTITY elvish-key *elvish.xml">
  ]>
```

5-15: Document Prolog

```
<?xml version="1.0" encoding="US-ASCII" standalone="no">
```

- This is the XML declaration.
- It indicates that the document is XML, the encoding schema, and whether or not the client will need to fetch supporting documents.

5-16: Document Prolog

```
<!DOCTYPE book
```

- This is the document type declaration - it indicates that the root element in the XML document is a book.

5-17: Document Prolog

```
PUBLIC "-//USF //DTD Book 1.8//EN"
"http://www.fooobar.com/DTDs/lotr.dtd"
```

- These lines designate a document type definition.
- Basically, this points to a separate document (called a DTD) that describes what elements books are allowed to have.

5-18: Document Prolog

```
<ENTITY jrrt "J.R.R. Tolkien">
<ENTITY elvish-key *elvish.xml">
```

- These lines declare an *internal subset*. These are sort of like C macros; they give a shorthand for elements that occur repeatedly throughout the document.
- All of the lines in the prolog except for the first are optional.

5-19: Entities

- We could then use our entity definitions later in the document by prepending a '&' to them

```
<book> ...
<description> the Author of The Lord of the Rings is &jrrt; he
invented a grammar and semantics for Elvish, which can be found at
&elvish-key;
</description>
```

5-20: Outline

- About XML
- Structuring XML documents
- **Using CSS to display XML**
- Parsing with DOM
- Parsing with SAX

5-21: Using CSS to display XML

- CSS can also be used to display XML documents.
- Control is limited to laying out a complete XML document.
- If we want filtering or sorting, we'll need to use XSLT.

5-22: An example

- Let's say we have an XML-based CD database:
- We can use CSS to display it in a web browser.
- (see separate examples)

5-23: Outline

- About XML
- Structuring XML documents
- Validating XML with schema
- Using CSS to display XML
- **Parsing with DOM**
- Parsing with SAX

5-24: Parsing XML

- XML also has the advantage of being easy for programs to parse and construct.
- There are two different approaches to parsing and manipulating XML.
- SAX: Simple API for XML
 - Event-driven parser
 - User defines actions to take when an element is found during parsing.

5-25: Parsing XML

- DOM: Document Object Model
 - Tree parser: Entire document is instantiated in memory as a tree.
 - Nice for random-access applications
 - Large documents may consume a large amount of memory
- Most languages provide support for both. We'll start with DOM.

5-26: Libraries

- The DOM model is specified in a language independent way.
- Implementations then follow this specification.
 - This means that they all work very similarly.
- Java
 - javax.xml.parsers built into Java 1.5
 - Apache's Xerces parser provides support for both SAX and DOM.
 - Xerces also has C++ and Perl implementations
 - JDOM is also a popular tool for parsing and creating XML in Java.
- Python
 - Built-in support for SAX, DOM, and minidom

5-27: Libraries

- Perl
 - LibXML provides SAX and DOM functionality.
- C#
 - .NET has built-in support for SAX and DOM

5-28: Parsing a document in Python

- Example:

```
from xml.dom import minidom
doc = minidom.parse('library.xml')
```
- Reads in and parses a document
- creates a Document object.
- toxml() show the XML version.

5-29: Traversing the tree

- childNodes, firstChild, lastChild, parentNode
- childNodes can have childNodes.
- Leaves are text nodes,
 - Respond to 'data', which gives up the data they store.
- This is useful if you need to process an entire document, but annoying if you're searching.

5-30: Finding specific elements

- getElementByTagName finds all elements according to name:

```
eltlist = doc.getElementsByTagName('key')
```
- Can search at any node

5-31: Finding attribute/value pairs

- Nodes have a dictionary-like structure that holds attribute/value pairs:

```
eltlist = doc.getElementsByTagName('key')
node1 = eltlist[0]
attrs = eltlist[0].attributes
keys = eltlist[0].attributes.keys()
```

5-32: An example

- Let's build a simple program for reading and displaying XML

```
#!/usr/bin/python

from xml.dom import minidom
import sys

doc = minidom.parse('./cdcat.xml')
def showCD(cd) :
    for item in cd.childNodes :
        if not item.nodeType == item.TEXT_NODE :
            print '<p>', item.tagName, item.firstChild.data, '<p>'

print '<html><body>'
print 'CDs in my catalog: '
cds = doc.getElementsByTagName('cd')
for item in cds:
    showCD(item)

print '</body></html>'
```

5-33: Outline

- About XML
- Structuring XML documents
- Validating XML with schema
- Using CSS to display XML
- Parsing with DOM
- Parsing with SAX

5-34: Parsing with SAX

- DOM is very convenient to use in many cases, but not all
 - Document is too large to hold in memory
 - Document is malformed
 - Document is being produced (and should be consumed) incrementally
- In these cases, a SAX parser may be more appropriate.

5-35: SAX: Simple API for XML

- SAX is an interface that was developed to provide a uniform way to integrate different XML parsers.
 - Interesting contrast in origin to DOM.
 - SAX developed 'bottom-up' by XML developers
 - DOM developed 'top-down' by the W3C.
- SAX is an *event-driven parser*
 - You define an event handler that is passed to the parser.
 - Describes how to handle particular types of elements.
 - Document is processed sequentially. State must be maintained by hand.

5-36: Using SAX within Python

- (Note: Java looks very similar)
- Most of the work involves creating *handlers*
- For example, to deal with processing content, override the *content handler*

5-37: Using SAX within Python

```
import xml.sax
from xml.sax.handler import *
class CDHandler(ContentHandler) :
    def __init__(self) :
        self.books = []
        self.buffer = ''
        self.inTitle = False

    def startElement(self, name, attrs) :
        if name == 'title' :
            self.inTitle = True
    def endElement(self, name) :
        if name == 'title' :
            self.inTitle = False
            print self.buffer
            self.buffer = ''
    def characters(self, ch) :
        if self.inTitle :
            self.buffer += ch
```

5-38: Using SAX within Python

- To use this, we then register the handler with a SAX parser.

```
parser = xml.sax.make_parser()
handler = CDHandler()
parser.setContentHandler(handler)
parser.parse('cdcat.xml')
```

5-39: SAX comments

- You must keep track of 'where you are' yourself.
 - No access to the enclosing context
 - It's hard with SAX to, for example, print the corresponding artist for each title node.
- SAX has more modest memory requirements than DOM
 - Nodes are discarded after parsing
- More flexible recovery from parsing errors.
- Use the parser that best fits your needs.