XML Primer

1 Introduction

This document explains the bare-bones basics of XML, the extensible markup language of the W3 Consortium which publishes the precise specification[1]. At its heart, XML is a very simple way to describe compound structure. XML is a syntactic device, it says nothing about semantics. It is a document interchange format, not a document format itself.

This document describes only how to write the body of an XML document, the "root element" and XML comments. The biggest omission here is the document type declaration (DTD) and how it is used to check the "validity" of a well-formed XML document.

2 Elements

A well-formed XML document includes exactly one "element." An *element* may be described using the form:

<*T* ...> ... </*T*>

where the first ellipsis gives space to define "attributes" and the second ellipsis is for the "content." Alternatively, if the content is empty, it may be written

<*T* .../>

where we still have the opportunity for "attribute definitions." So, for example the string `<mark/>` is a well-formed element.

The content can be other elements or plain text. For example:

```
<greeting>Hello, world!</greeting>
```

A more complex example is

```
<room> <chair/> <table/>
      <box><paper/><pen/>rather empty</box>
      My messy office
      <phone/>
      Anything else?
</room>
```

As you can see XML permits whitespace between tags.

Attribute definitions are of the form n=v where n is any name (like a C++ identifier, except that colons are allowed and periods and hypens are permitted after the first character) and v is a string constant. String constants can be delimited by either single or double quotes, take your pick. So for example, one could write:

[1] http://www.w3.org/TR/REC-xml

```
<greeting lang="en-US" mood='7'>Hello, world!</greeting>
```

Inside the strings and the plain text, you may not use the special characters `< > " ' &` which have special meaning in XML:

Illegal: `<prove> 6<7 </prove>`

Illegal: `<speech quote="""> Hi! </speech>`

Thus to get these special characters in an XML string or character data, you need to use an *entity* of the form `&`n`;` where n is the name of an entity. Make sure not to forget the semicolon! XML predefines five entities, and we add a sixth:

`<` This means "<" (less than).

`>` This means ">" (greater than).

`&quot;` This means """ (quote).

`&apos;` This means "'" (apostrophe)

`&` This means "&" (ampersand).

`&sp;` This means " " (space).

As an exception, XML does permit one kind of quote to appear in a string using the other kind of delimiter.

3 Comments

XML permits comments of the form: `<!--...-->` to occur anywhere in the contents or at the top level. Comments may *not* appear where attributes are defined. The comment body may not include two hyphens in a row and cannot end in a hyphen. But it may include any other characters.