A Minimalist Introduction to XML

Julia Flanders and Syd Bauman

2010-02-13

Structure?

A good place to start in understanding XML—what it is and why it’s important—is to think about how it as a way of structuring or organizing information.

Structure is what makes digital information useful to us: it helps us find things, it helps us identify them and understand what they are, and it helps us communicate about them to other people. We can think of the difference between structured and unstructured information as being similar to the difference between a compost heap (all very well in its way, but you can’t construct a coffee bean from the individual grains of coffee grounds) and a tool chest, where everything is clearly marked and well-organized.

In that middle image, there is some structure (the papers are baled up somehow so we could grab a bunch of them at once) but the structures aren’t intelligible to us and so they aren’t any actual use.

XML Structures the World

XML provides a very specific way of structuring information and of looking at the world. It asks us to identify individual pieces of information that we are interested in (by naming them), and it asks us to group them together. We can represent these groupings in various different visual ways:

- As a set of nested boxes, where the groupings are carried out by physical placement; if you’ve ever packed a camper, or if you’re a collector of tiny objects, this approach will be familiar and appealing to you. The resulting structure of boxes represents the structure of the collection of objects, whatever they are.
- As a tree, where the groupings are represented by branchings; this is very much analogous to the idea of a family tree, where each individual is grouped with siblings under their common parents, and so forth; the tree structure represents the structure of an entire ancestry
- For purposes of actually editing the XML, we usually find it most useful to use the notation on the right, which represents the XML tree or boxes using a system of tags that mark the boundaries of the boxes and the structural groupings that make up the informational system. Each box or tree node (or element in XML parlance) is represented by a pair of tags that mark its beginning and its end, and these elements can be nested inside one another.

XML in Detail

So let’s look at an XML document in more detail:

- The tags are delimited by special characters (the angle brackets), and the end-tag has a
slash indicating that it marks the end of an element

- We also have these attributes which give more information about an element [ask class to interpret some of the attributes]

A few more things to note as we look at a sample XML document up close:

- Although the earlier examples didn’t reveal this explicitly, an XML representation has order: an XML document is an ordered sequence of elements.

Well-formedness and Validity

The final things you need to know about XML documents: the rules they have to follow!

First, they must observe the basic rules of XML, which we’ve seen in practice in our examples but need to spell out explicitly here:

- The elements of XML documents must all nest neatly inside one another, without overlapping
- They must have all of their delimiters (angle brackets, quotation marks, etc.)

If they obey these rules, they are said to be well-formed

In addition, XML documents may conform to the rules established by a specific XML language