



The Wonderful World of XML

**Presented by Laurie K. Brooks
AML Consulting, Inc.**



XML Precursors

- Hypertext and Multimedia => Hypermedia
- Internet => World Wide Web
- Generalized Markup => SGML and HTML
- Proprietary Systems => Open Systems and International Standards

Internet

- ARPANET plan announced in 1967, with the first node set up at UCLA in 1969
- First International Conference on Computer Communications. October, 1972 (40 nodes)
- ARPANET split into ARPANET and MILNET, 1983
- World Wide Web designed for particle physicists by Tim Berners-Lee in 1989

GML & SGML

- GML was developed in 1969 at IBM to deal with legal documentation
- GML's appeal turned out to be widespread and its concept was adopted by other companies
- Need became apparent to standardize it
- SGML was released as ISO8879 in 1986

HyperText

- Hypercard. Popularized a simple concept of hypertext, 1987
- HTML. Used a simple addressing scheme to link documents across the internet, 1989 (more on HTML to follow)

SGML leads to XML

- At SGML 96, XML draft was released by a working group of the W3C.
- In February of 1998, XML 1.0 became a W3C recommendation (32 pages)
- HTML 4.0 (not XML based) was released April 24, 1998 (367 pages)

XML vs. SGML

- XML is an SGML subset
- Some features left out were “misfeatures”
- Some others were merely conveniences
- Others are useful but can be emulated in other ways
- Bottom line: XML can do everything SGML does, but you may have to reinvent some of SGML’s features

XML vs. HTML

- HTML has *semantics* in addition to *syntax*:
- Each HTML element type means something
 - <A> is universally recognized as a link
 - <P> is a paragraph, etc.
- XML provides extensible syntax but no semantics

XML's Web Heritage

- The dominant markup language on the Web is HTML, which is SGML-based
- XHTML is not a merge of XML and HTML, but HTML based on XML instead of SGML
- HTML has a fixed number of element types
- The data necessary to encode is infinitely complex
- The Web needed SGML's power but not its complexity

XML - Three Fundamental Principles

1. It must be character-based
2. It must be extensible (“a meta-language”)
3. It must have a powerful schema language

Word docs fail all tests

RTF and HTML pass 1

TeX passes 1 and 2

SGML/XML pass all three

XML's Driving Philosophy

- Information is primary, not programs
 - information is a valuable corporate asset
 - applications process information, making it accessible and usable
 - business rules should be coupled to information
- Information should be independent from the applications which manipulate it
 - changing the binding between applications and information should be simple and cheap
- Information exists over time
 - information should be accessible over long, short and medium terms

*XML Benefits

- Information reuse
- Improved lifecycle management
- Increased availability of enterprise information
- Locate and deliver information efficiently
- Configurable information appearance
- Information transport in heterogeneous environments

Information Reuse

- Saves time and money
- All information has same "file format"
 - no transformations required to use in different context
- Information can be modularized
 - optimized modularization to support reuse

Lifecycle Management

- Infrastructure evolution does not affect content
- Old-but-interesting content remains available
- Automated lifecycle management
- Information objects can carry along their own lifecycle attributes
- Systems can query objects about lifecycle status
 - allowing for automated lifecycle stage progressions

Availability of Enterprise Information

- Eliminate duplicate effort and information stove-piping
- Support work with useful information
- Efficient querying makes all information visible to users who might find it interesting or useful
 - even if they did not know of its existence ahead of time
- Transportable information is more useful

Locate Information Efficiently

- Context-sensitive searching can enhance search engine performance by factor of 10 or 20
- Structured markup permits automated document assembly
- Just-in-time assembly eliminates storage and shrinkage costs
- Custom documents can be sold for more money

Configurable Appearance

- Multiple stylesheets permit delivery of same information under more than one brand
- Automated stylesheets force conformance to corporate style guides
- Guaranteed conformance to industry requirements (ATA, AIA, US Clean Air Act)

Information Transport

- Allows for application-to-application data interchange without having to educate the applications about each other
- Permits creation of consolidated information objects from multiple sources
- Encapsulates content and structure for exploitation by receiving system
 - does not tie information to a particular use

Content and Format

- Content: the information (data) contained in a document - usually words and illustrations
- Format: the way words, sentences, and paragraphs are visually presented and distinguished from one another within a document
- Content vs. format depends on your point of view....

Content vs. Format

- Human viewers understand that a title is a title because it is usually at the beginning of a document and it is bigger and bolder than the rest of the text
- XML allows software to distinguish a title because it is tagged or marked as a title:
`<title>Content vs. Format</title>`
- Stylesheets then allow for the title to appear with a certain format on a screen

Full Cycle Example

A Restaurant Review

\$\$ / ★ ★ ★

Moscow's

Seafood

4137 US 90 West, Avondale

436-9942

This old wood-framed roadhouse has been serving unique Creole-Italian food for 2 generations. Enjoy exquisite crab, oyster and shrimp dishes, served dripping in garlic and olive oil. Be sure to get directions or you will never find this place!

The Review, Labeled



The Review's Structure

review

header

restaurant

food.rating
price.rating

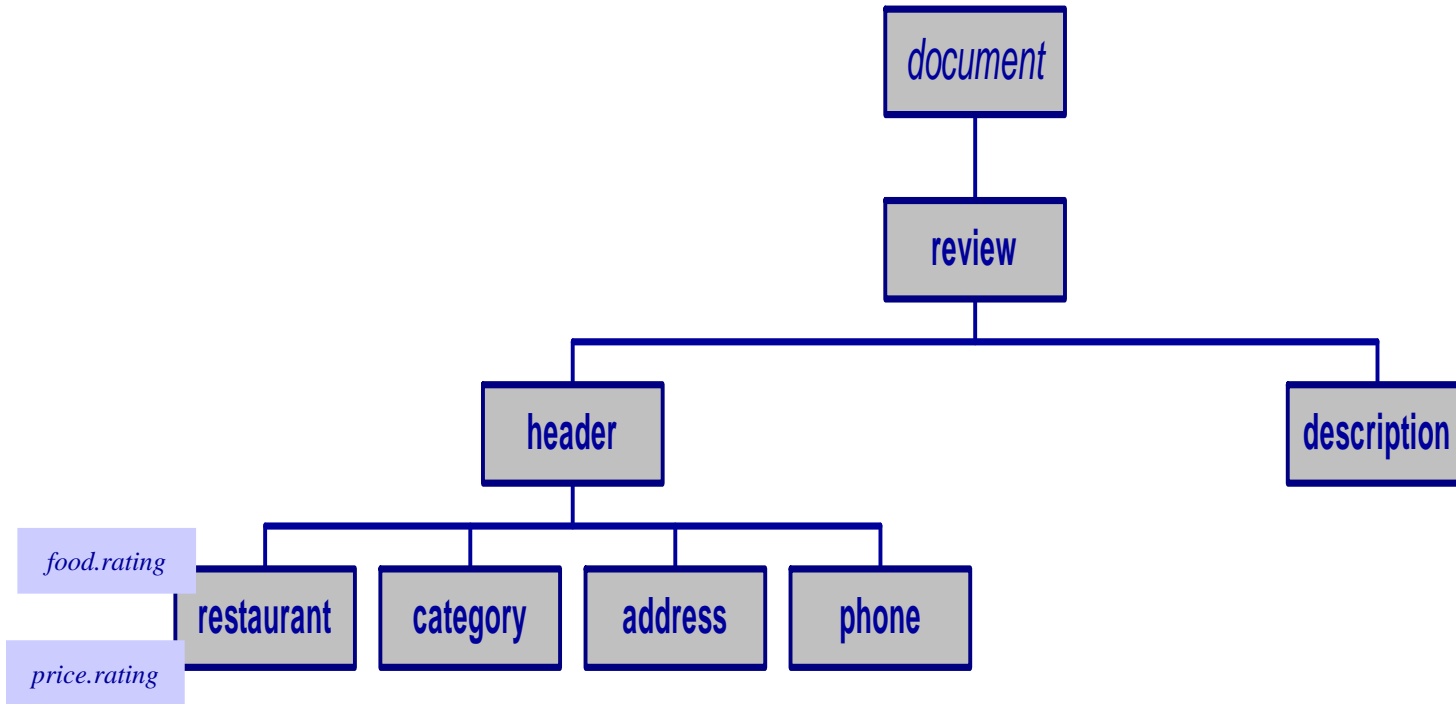
category

address

phone

description

The Review's Structure Tree



Document Markup

- Markup provides information needed to process a document
- Procedural markup
 - instructions for a specific processing system
- Structured markup
 - labels document components according to their meaning
 - processing systems infer what to do based on structure
- Markup and content coexist
 - you must be able to distinguish markup from content

Marked Up Review

```
<review>
```

```
<header>
```

```
<restaurant price="$ $" food="3">Moscow's
```

```
</restaurant><category>Seafood</category>
```

```
<address>4137 US 90 West, Avondale</address>
```

```
<phone>436-9942</phone></header>
```

```
<description>This old wood-framed roadhouse has  
been serving unique Creole-Italian food for 2  
generations. Enjoy exquisite crab, oyster and  
shrimp dishes, served dripping in garlic and  
olive oil. Be sure to get directions or you  
will never find this place!</description>
```

```
</review>
```

Rendering Structured Content

- Note that the review does not contain any clues as to how it should be rendered
- Require rules for rendering
- Based on structural labels and context

Review Rendering Rules

review: new page; typeface Times New Roman
restaurant: print the number of \$\$ signs on the
left, then a /, then the number of stars,
all left-aligned; then a right tab, then
print content of element in bold
category: new line, italic
address: new line, down size 2 points
phone: new line
description: two new lines, up size 4 points

*Note: this is not an actual stylesheet

Rendered Review

\$\$ / ★ ★ ★

Moscow's

Seafood

4137 US 90 West, Avondale

436-9942

This old wood-framed roadhouse has been serving unique Creole-Italian food for 2 generations. Enjoy exquisite crab, oyster and shrimp dishes, served dripping in garlic and olive oil. Be sure to get directions or you will never find this place!

Well-formed vs. Valid

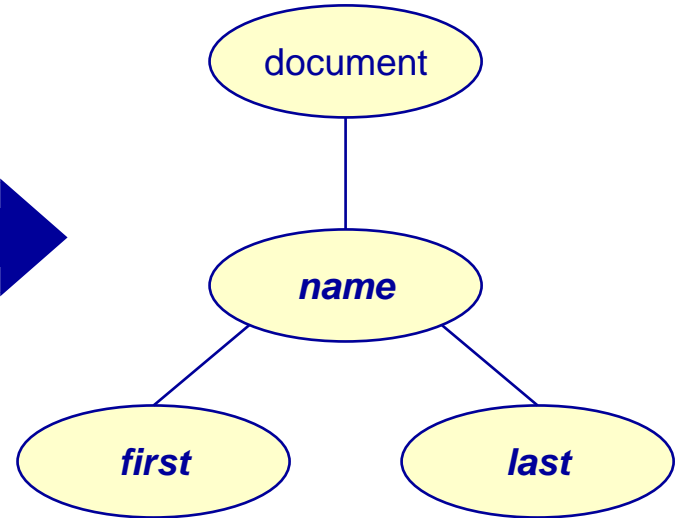
The Value of Well-formed XML

- Well-formed rules guarantee that you can always draw a complete tree for an XML document
 - element structure
 - attributes associated with elements
 - presence of comments and processing instructions
 - location of text strings
- Learning to draw the XML tree accurately is critical
 - all stylesheet processing works on the document tree

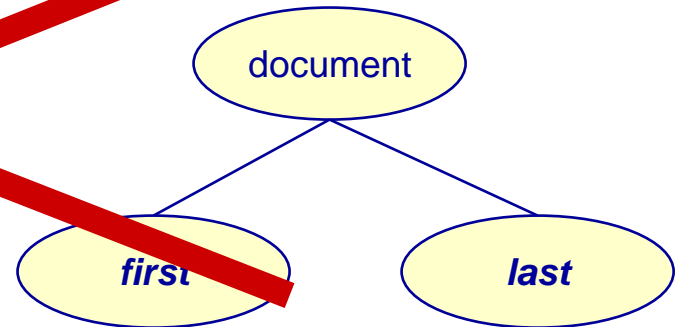
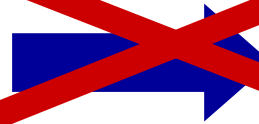
Basic Well-formed Rule

- Single top-level element

```
<?xml version='1.0'?>  
<name><first>Kevin</first>  
<last>Macduff</last></name>
```



```
<?xml version='1.0'?>  
<first>Kevin</first>  
<last>Macduff</last>
```



Well-formed is (Often) Not Enough

- Well-formed documents can be rendered in browsers
 - but more powerful processing is very difficult
- Powerful processing applications need data to be predictable
- A well-formed XML document can have any kind of element in any location
 - with no warning to the processing system that the element is going to be there

What is Valid XML?

- A **valid** XML document has a set of rules concerning the kind of structures it can contain
- These rules include:
 - what elements names are allowed
 - how elements must be organized
 - where data characters are allowed
 - which elements have attributes
 - what attributes are called
 - what kind of values an attribute can have
 - possibly:
 - what data types are allowed where

Advantages of Valid XML

- Predictable tree structure
- Ability to perform data verification
 - based on structural context
- Predictable tree structure
- Tune the processes to your requirements
- Predictable tree structure

DTDs and Schemas

- DTDs have been around for years
 - defined in the XML specification
 - many existing tools which understand and use DTDs
 - compact code
- Schemas
 - provide an alternate to DTDs
 - schema itself is well-formed XML
 - more verbose
 - processable as XML
 - provides additional features, such as data typing

Schemas

- W3C Schema
 - Considered several proposals
 - XML-Data (Microsoft)
 - DCD (XML-Data subset, in RDF syntax)
 - SOX (Commerce One)
 - DDML (Formerly XSchema, XML-DEV)
 - W3C Recommendation 2 May 2001
 - *Part 1: Structures*
 - *Part 2: Datatypes*
 - ...best bet is to read *Part 0: Primer*

Marked Up Review (well-formed)

```
<review>
```

```
<header>
```

```
<restaurant price="$ $" food="3">Moscow's
```

```
</restaurant><category>Seafood</category>
```

```
<address>4137 US 90 West, Avondale</address>
```

```
<phone>436-9942</phone></header>
```

```
<description>This old wood-framed roadhouse has  
been serving unique Creole-Italian food for 2  
generations. Enjoy exquisite crab, oyster and  
shrimp dishes, served dripping in garlic and  
olive oil. Be sure to get directions or you  
will never find this place!</description>
```

```
</review>
```

DTD: Review

```
<!ELEMENT review      (header, description) >

<!ELEMENT header      (restaurant, category,
                        address, phone) >

<!ELEMENT restaurant  (#PCDATA) >
<!ATTLIST restaurant  price          CDATA          #REQUIRED
                        food          (1 | 2 | 3)      #REQUIRED >

<!ELEMENT category    (#PCDATA) >

<!ELEMENT address     (#PCDATA) >

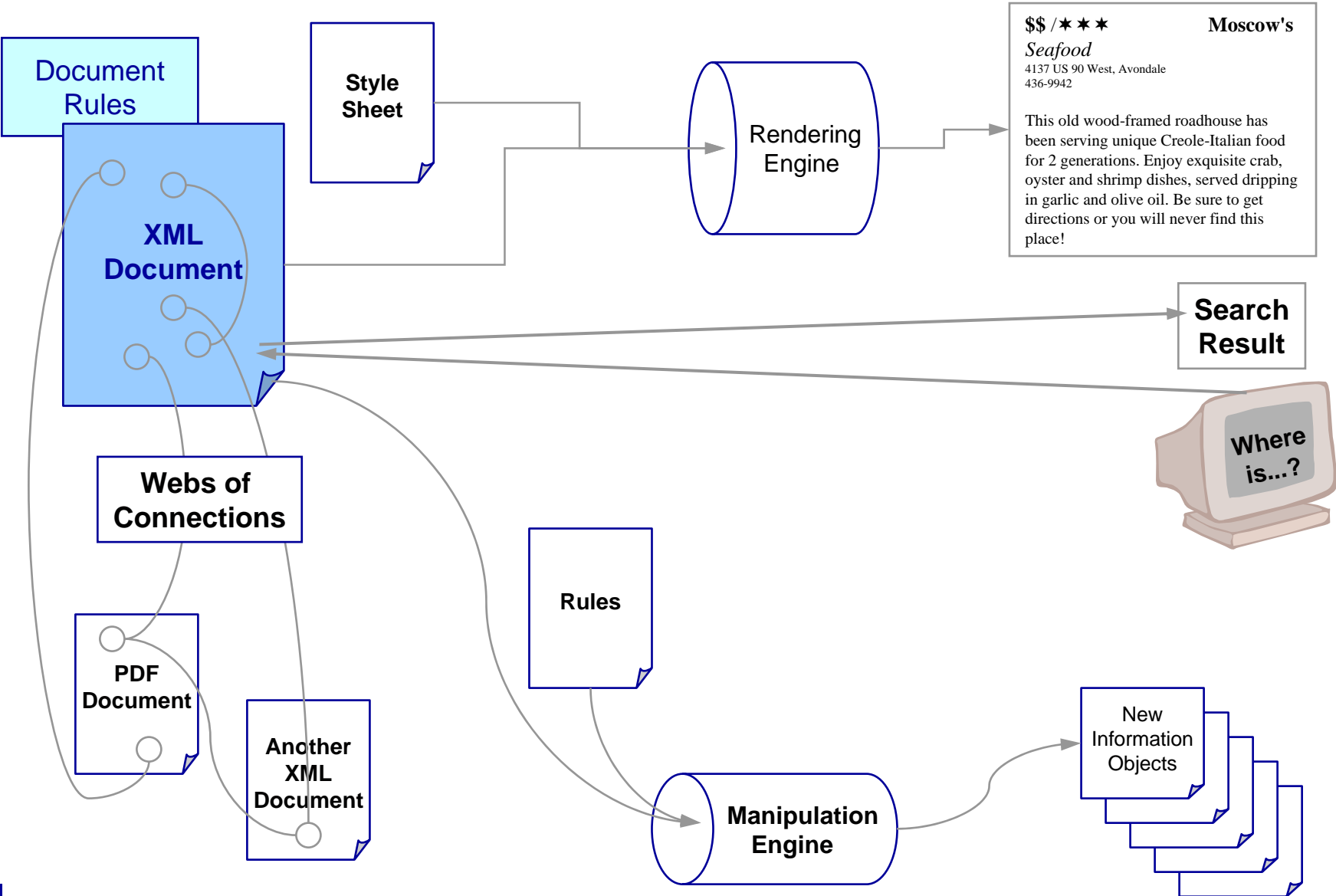
<!ELEMENT phone       (#PCDATA) >

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

Review - Structure, Format and Content

- Documents contain two kinds of information
 - format and content
- Format provides clues to structure
 - helps end users understand the document
 - provides a framework for the content
- Structure answers these questions:
 - what kind of document is this?
 - what are its parts?
 - how are those parts arranged?

The Valid XML Document



What XML is Not

- A binary format optimized for transmission or storage – although it can be used for both
- A programming language - it can be used with any programming language
- A system design methodology - it can be used in radically different systems
- A replacement for HTML - HTML will continue to be used for many documents intended for human consumption
- A vocabulary - there will be an infinite number of XML-based vocabularies

XML Tool Categories

- Creating XML content
- Working with DTDs and schemas
- XML parsers
- XML middleware
- DOM- and grove-based applications
- Content management repositories
- Tools for eCommerce

So You Want to Be an XML Expert

- On-line resources
- Conferences and meetings
- More training
- Books and publications

WWW Resources

- World Wide Web Consortium (W3C)
 - www.w3.org
 - www.w3.org/TR (for W3C specifications)
- The best all-purpose XML web site
 - www.oasis-open.org/cover
- A collection of public domain useful XML stuff
 - www.xml.org
- And dozens of others

Conferences and meetings

- GCA-sponsored events
 - <http://www.gca.org>
- OASIS events
 - <http://www.oasis-open.org>
- Event listing at XML.com
 - <http://www.xml.com>
- XML Users Groups

More Training

- XML courses you might find useful
 - XML Quickstart (Includes DTD Development)
 - Introduction to XSLT
- Tool courses and special interest courses
 - see Robin Cover's website for leads to this kind of thing
 - conferences often have good tutorials

Books and Publications

- There are hundreds of books available
 - and more are released every hour
 - many of them are trying to capitalize on the "coolness" of XML

</course>