An *XML* Tutorial

JTC1/SC32

Victoria, BC Canada

October 2001

Charles E. Campbell Ph.D. (USA)

Why is *XML* an important?

- There is a lot more beneath the surface!
- There is a whale of a lot of stuff that will depend upon *XML* technologies in the future!
- SC32's technologies will all be impacted by *XML* in some way**!**
- *XML* is going to be everywhere and will only become more pervasive with time.

What is XML

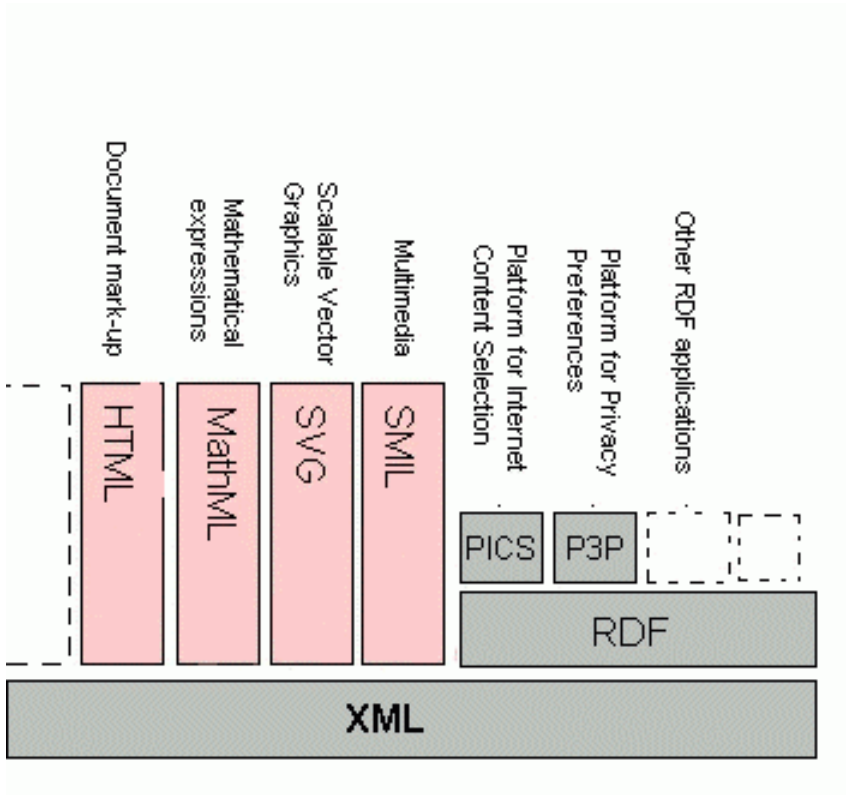
- XML -- A Markup Language
  - It is a protocol for containing and managing data.
  - A family of technologies:
    - Formatting documents to filtering data
  - A philosophy for handling information.

Where XML fits into the picture

- SGML (Standard Generalized Markup Language) as defined by ISO 8879. Not well suited for serving documents over the WEB.
- HTML (Hypertext Markup Language) a W3C Recommendation. Markup Language used to create documents on the WEB. Mixes content and display instructions.
- XML (Extensible Markup Language) a W3C Recommendation, was created so that richly structured documents could be used on the WEB, something neither SGML or HTML were able to provide.

The XML layer

- The XML layer



W3C
World Wide Web Consortium
Domains

- Architecture Domain -- XML
- Document Format Domain
- Interaction Domain
- Technology and Society Domain
- **Web Accessibility Initiative**
- **Quality Assurance Activity**

W3C Process

- Notes
- Workshop
- Charter
- Requirements
- Drafts
- Candidate Recommendations
- Proposed Recommendations
- Recommendations

How the W3C Process Works

- WGs are chartered for a specific time and task
- Communications
  - Face-to-face meetings
  - Weekly teleconferences
  - Email (high volume)
- Consensus driven
- Editor plays large role in creating recommendations
- 3-Month Heartbeat Requirement for publication.
  - All comments are responded to
- W3C is a consortium – not an open body

When did XML Become a Recommendation?

- The W3C published the XML 1.0 Recommendation on 10-February-1998
- A Second Edition was published on 6-October-2000 with the title:

Extensible Markup Language (XML) 1.0 (Second Edition)

XML Activity after XML 1.0

- XML Coordination Group
- XML Schema Working Group
- XML Linking Working Group
- XML Information Set Working Group
- XML Fragment working Group
- The XML Syntax Working Group

Today's XML Coordination Group

- XML Coordination Group – Chairs of XML WGs
- XML Plenary Interest Group
- XML Core Working Group
- XML Query Working Group
- XML Schema Working and Interest Groups
- XML Linking Working and Interest Groups

XML Coordination Group -- Liaison

- XSL Working Group
- DOM Working Group
- CSS and FP Working Group
- XML Protocol Working Group
- XForms Working Group

A Simple XML Document

```
<?xml version="1.0" encoding="UTF-8" ?>
  <greeting>Hello, world!</greeting>
```

A Simple XML Document with internal DTD

```
<?xml version="1.0" encoding="UTF-8" ?>
  <!-- This document is valid -->
  <!DOCTYPE greeting [
     <!ELEMENT greeting (#PCDATA)>
  ]>
  <greeting>Hello, world!</greeting>
```

A Simple XML Document with external DTD, Comment & PI

```
<?xml version="1.0" encoding="UTF-8" ?>
  <!-- Here is a procession instruction
  -->
  <?render bold ?>
  <!DOCTYPE greeting SYSTEM
  "hello.dtd">
  <greeting>Hello, world!</greeting>
```

Elements

A defined piece of an XML Document

<Element Name>Content</Element Name>

<Element Name></Element Name>

<Empty Element Name/>

Attributes vs. Elements

An attribute defines a specific setting or provides additional information about an Element:

```
<team person1="sue" person2="chuck">
```

```
<team>
   <person>sue</person>
   <person>chuck</person>
</team>
```

PCDATA vs. CDATA

- PCDATA is parsed-character data
  - Any character data that should be checked by the XML Processor for entity references.
    - Entity is a name assigned by means of declaration to a chunk of data. [“<” “&”]
- CDATA is non-parsed-character data
  - An entity datatype consisting of non-parsed characters.
    - Used anywhere character data can occur, content not interpreted
    - <![CDATA[10 < 1000, really!!!!]]>

Well-Formed vs. Valid XML

- Well-Formed means that the document conforms to the syntax rules of XML.
  - Has both start tags and end tags and elements don't overlap
- Valid XML means that the document is Well-Formed and conforms to a DTD or ***Schema***.

XML Recommendations of Interest

Specification	Date	Status
XML 1.0	1998-02-10	REC
XML 1.0 (Second Ed.)	2000-10-06	REC
Namespaces in XML	1999-01-14	REC
XBase	2001-06-27	REC
XLink 1.0	2001-06-27	REC
DOM Level 1	1998-10-01	REC
DOM Level 2	2000-11-13	REC
XPath 1.0	1999-11-16	REC
XSLT	1999-11-16	REC
Canonical XML	2001-03-19	REC
XML Schema	2001-05-02	REC

XML Recommendations of Interest

Specification	Date	Status
InfoSet	2001-05-14	CR
XML Fragment	2001-02-12	CR
XSL 1.0	2000-11-21	CR
XInclude 1.0	2001-05-17	LC
XPointer 1.0	2001-01-08	LC
XML Protocol Abstract Model	2001-07-09	WD
SOAP 1.2	2001-07-09	WD

PR – Proposed REC, **CR** – Candidate REC, **LC** – Last Call WD,
WD – Working Draft

Namespaces in XML

- Problem: documents containing multiple markup and vocabularies pose problems with recognition and collision.
- An XML namespace is a collection of names, identified by a URI reference. Provided Scope.
  - No file or content need exist

Namespace Example

```
<?xml version="1.0"?>
   <book:book xmlns:book='urn:loc.gov:books'
              xmlns:isbn='urn:ISBN:0-395-36341-6'>
      <book:title>Cheaper by the
   Dozen</book:title>
      <isbn:number>1568491379</isbn:number>
   </book:book>
```

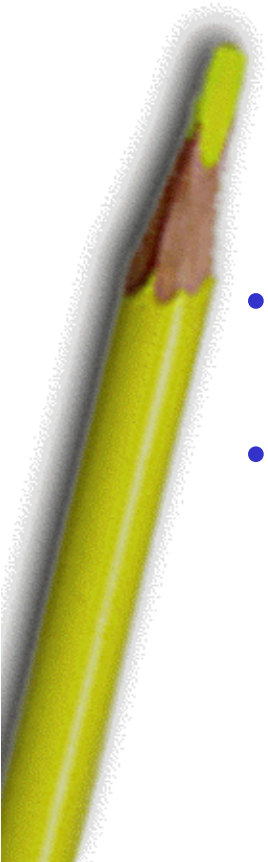
Other XML Recommendations

- **XBase** – for specifying a "base" URL for relative URLs.
- **XLink** – for describing links between resources.
- **DOM** -- **D**ocument **O**bject **M**odel is an API that provides a standard set of interfaces for manipulating an XML Document. Document is modeled in Memory.
- **SAX** – **S**imple **A**PI for **X**ML, non-W3C API for streaming document processing.
- **XPath** – An expression language, not XML, providing a syntax for finding specific parts of an XML document.

Other XML Recommendations

- **XSL** – **E**xtensible **S**tylesheet **L**anguage
  - **XSLT** -- **E**xtensible **S**tylesheet **L**anguage **T**ransformation uses XPath to match nodes for transforming an XML document document into another format i.e.: HTML
  - **FO** – **F**ormatting **O**bject used to format. For example Apache's FOP is used render XSL format object's into PDF.
- **XML Schema** an alternative to a DTD and used to validate XML documents. Unlike DTDs XML Schemas are written in XML and it has Structure and Data Type information.

Other XML Recommendations



- **InfoSet** – An infoset is an abstract model of a well-formed XML document that conforms to the namespace recommendation
  - An infoset consists of information items, each of which has a set of properties.
  - An infoset always contains a single document information item.
- **XML Fragment** – Describes how to split XML documents into pieces for transport across networks.
- **XInclude** – for including existing XML documents or portions of XML documents into another XML Document.

Other XML Recommendations

- **XPointer** – for "pointing" to a documents contents. Built upon XPath and supports addressing into the internal structure of the XML Documents.
- **SOAP** – **S**imple **O**bject **A**ccess **P**rotocol, W3C XML Protocol WG. Provide a framework for expressing application semantics, encoding data and packaging it into modules.
- **XHTML** – is a reformation of HTML 4 as an XML application. The XML DTD defines elements and attributes as they are in HTML 4.01

Custom Markup Languages

- **MathML** – A calculus expression language
- **OpenMath** – Another math language
- **CML** – Chemical Markup Language
- **WML** – Wireless Markup Language
- **GML** – Geographical Markup Language
- **SMIL** – Synchronized Multimedia Integration Language
- **SVG** – Scalable Vector Graphics
- **BML** – Bean markup language
- **X3D** – Extensible 3D language
- **XBRL** – Extensible Business Reporting Language
- **BIPS** – Bank Internet Payments System
- **ebXML** – Electronic Business XML

Custom Markup Languages

- **Visa XML Invoice Specification**
- **cXML** – Commerce XML
- **LegalXML**
- **NewsML**
- **Open eBook Publication Structure**
- **XUL** – Extensible User Interface Language

XML Technologies and Applications

- **DSML** –Directory Services Markup Languages
- **RDF** – Resource Definition Framework
- **XTM** – XML Topic Maps
- **VHG** – Vertical HyperGlossary
- **CDF** – Channel Definition Format
- **ICE** – Information and Content Exchange Protocol
- **RSS** – Rich Site Summary
- **P3P** – Platform for Privacy Preferences
- **BXXP** – Blocks Extensible Exchange Protocol
- **XML Digital Signature**
- **XrML** – Extensible Rights Markup Language
- **XMI** – XML Metadata Interchange

XML Resource Sites

- W3C – www.w3.org
- Oasis – www.oasis-open.org
- Cover Pages –xml.coverpages.org
- XML.org – www.xml.org
- XML.com – www.xml.com