XML Update

Royal Society of the Arts
London, December 8, 1998

Jon Bosak
Sun Microsystems

XML Basics

What is XML?

- Extensible Markup Language
- An activity of the World Wide Web Consortium (W3C) organized and led by Sun Microsystems
- Objective: move the Web to its next stage of evolution by adapting existing ISO standards for markup, linking, and formatting
- Will create new data-centric Web applications
- Will fundamentally change publishing on the web and publishing in general

What made XML necessary?

Two aspects of Web evolution demanded a technology beyond HTML.

- Internationalized electronic publishing
  - Platform-independent
  - Language-independent
  - Media-independent
- New data-centric Web applications
  - Database exchange
  - Distribution of processing to clients
  - Client-side manipulation of views into the data
  - Customization of information by intelligent agents
  - Management of document collections

What's wrong with HTML?

- HTML was optimized for easy learning
  - One tag set for all applications
  - Predefined semantics for each tag
  - Predefined data structures
  - No formal validation
- HTML trades power for ease of use
- HTML is well suited to simple applications, but poorly suited to more demanding applications
  - Large or complex collections of data
  - Data that must be used in different ways
  - Data with a long life cycle
  - Data intended to drive scripts or Java applets

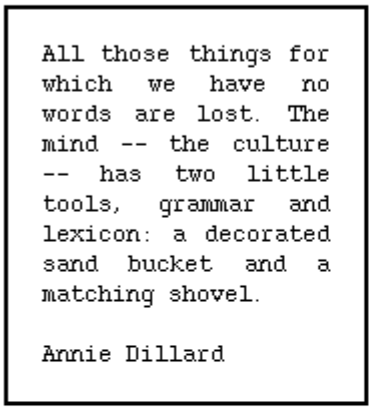
Why did Sun invest in XML?

1. In industry, we knew from electronic publishing experience that HTML would not work for publishing in the general case
2. We also knew that future Web applications would require a method of encoding that could drive arbitrarily complex distributed processes
3. It was clear that if an open standard like XML was not created, HTML would be replaced by a more powerful **binary proprietary format**

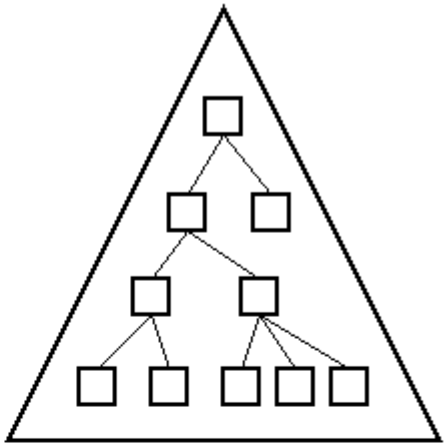
Strategically, we had to have XML in order to keep Web data open and portable. We needed XML to do for data what Java does for programs.

The XML Concept

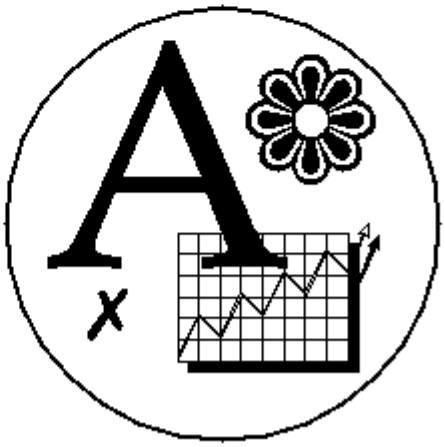
Document components



CONTENT

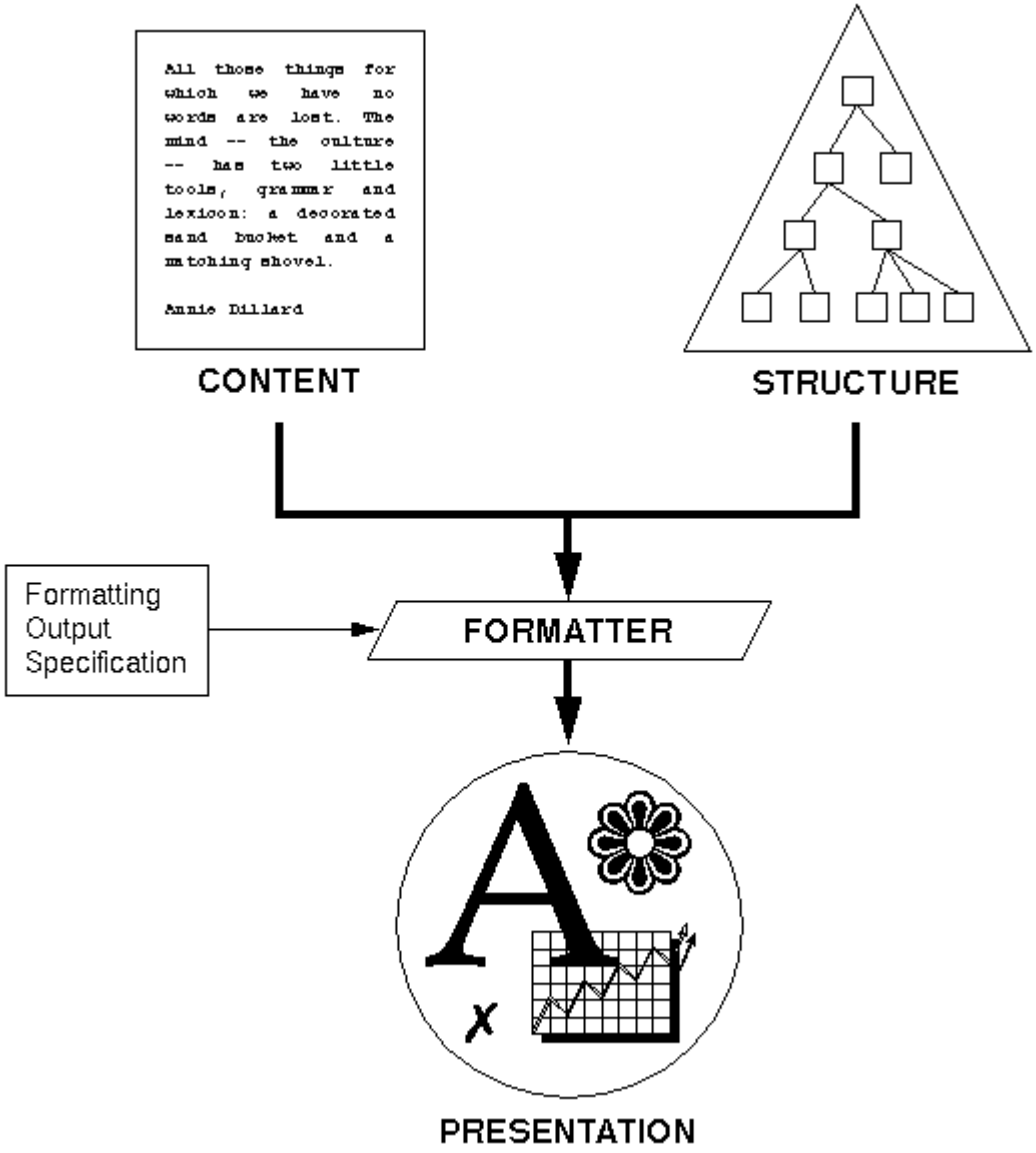


STRUCTURE



PRESENTATION

Structured publishing



XML allows you to specify the content and structure of a document in a way that lets you generate particular presentations as needed.

Proof of concept: this presentation

(These are links in the online version.)

- The XML source from which this presentation was produced
- The optional XML DTD used to validate the XML source
- The DSSSL style sheet for the HTML used in the online version
- The DSSSL style sheet for the RTF used in the printed version
- The Jade DSSSL engine used to produce both the HTML and RTF files
- A PostScript version of this presentation produced from Jade's RTF output
- A PDF version of this presentation produced from the PostScript

Lessons from the proof of concept

- Media-independent publishing works!
- HTML can handle today's online version, but not the print version
- The future output specification (stylesheet) language (i.e., XSL) must support **structural transformation**

Demo: An XML newspaper system for the web

- NetPost is an independent company that specializes in cross-media publishing solutions
- Sun is using NetPost technology to show the power of XML and Java
- This demo shows the central XML concept: automatically generating presentation from content plus "template"
- Arbitrary amounts of content inherit the styles and layout specified in the template
- In this version, a server combines content + template to create an XML output stream designed for a Java-based newspaper browser
- In a later version, binding of content + template will happen in the browser itself
- Download NetPost demo (656050-byte zip file) -- requires Netscape 4.06 or IE 4 to run

What the demo shows

- Reuse of same content in different forms
- Reuse of content in (potentially) different media
- Customized display for different audiences

Things to think about:

- The same system could generate reports from a database
- As displays improve, online formatting will stop being a subset of print formatting
- In fact, online formatting will be recognized as a **superset** of print formatting
- XSL must provide not only powerful transformational ability but also powerful formatting ability

Contact data: Steve Katz (steve@net-post.com)

XML in Context

The XML family

The XML family of languages is a suite of specifications that moves the web to a new level of evolution suitable for electronic commerce and other industrial-strength applications.

- **XML** (Extensible Markup Language): A subset of SGML (ISO 8879) designed for easy implementation
- **XLink/XPointer**: A set of standard hypertext mechanisms based on HyTime (ISO/IEC 10744) and the Text Encoding Initiative (TEI)
- **XSL** (Extensible Stylesheet Language): A standard stylesheet language for structured information based on DSSSL (ISO/IEC 10179) and CSS

XML itself

- A simplified subset of SGML (ISO 8879)
  - Very powerful
  - No limits on namespace or structural depth
  - Easy to implement
  - Small enough for Web browsers
- Internationalized from the beginning
  - Unicode for both content and markup (can mix languages)
  - XML tools **must** support both UTF-8 and UTF-16
- Not a language but a metalanguage
  - Designed to support the definition of an unlimited number of vertical-market languages for specific industries
  - All XML languages can be processed by a single lightweight parser built into every Web browser

XML in isolation

- "Syntax, not semantics"
  - Tags have no predefined meaning
  - XML by itself conveys only content and structure, not presentation or behavior (unlike HTML)
- There are important applications for XML alone: interprocess communication, object serialization, metadata, database exchange
- But associating **presentation or behavior** with XML requires **additional mechanisms**
  - Verbal agreements on the processing of specific tag sets (example: HTML)
  - Embedded programs, applets, or scripts
  - Tag-sensitive components (e.g., JavaBeans) -- best for data-centric applications
  - Stylesheets (XSL or CSS) -- best for applications where data is to be displayed

Example: Just XML

```
<?xml version="1.0"?>
<!DOCTYPE 書籍カタログ [
  <!ELEMENT 書籍カタログ (書籍)+ >
  <!ELEMENT 書籍 (書名,著者,出版社,(定価 | 在庫数)) >
  <!ATTLIST 書籍      xml:lang CDATA #REQUIRED >
  <!ELEMENT 書名      (#PCDATA) >
  <!ELEMENT 著者      (#PCDATA) >
  <!ELEMENT 出版社    (#PCDATA) >
  <!ELEMENT 定価      (#PCDATA) >
]>

<書籍カタログ>
  <書籍 xml:lang="JP">
    <書名>XML入門</書名>
    <著者>村田、門馬、荒井</著者>
    <出版社>日本経済新聞社</出版社>
    <定価>2800</定価>
  </書籍>
  <書籍 xml:lang="EN">
    <書名>Developing SGML DTDs</書名>
    <著者>E. Maler and J. el Andaloussi</著者>
    <出版社>Prentice Hall</出版社>
    <定価>50</定価>
  </書籍>
</書籍カタログ>
```

With stylesheet for reader of Japanese

XML入門

著者　村田、門馬、荒井

出版社　日本経済新聞社

定価　2800円

Developing SGML DTDs

著者　E. Maler and J. el Andaloussi

出版社　Prentice Hall

定価　50ドル

With stylesheet for reader of English

XML入門

Author: 村田、門馬、荒井

Publisher: 日本経済新聞社

Price: ¥2800

Developing SGML DTDs

Author: E. Maler and J. el Andaloussi

Publisher: Prentice Hall

Price: $50

Source files for the bookstore example

(These are links in the online version.)

- The UTF-16 XML source from which the different versions were produced
- The UTF-16 DSSSL style sheet used to produce the version for the reader of Japanese
- The UTF-16 DSSSL style sheet used to produce the version for the reader of English
- The Jade DSSSL engine used to produce RTF files from the source and the style sheets
- The UTF-16 RTF file for the reader of Japanese (font association done in Word 7)
- The UTF-16 RTF file for the reader of English (font association done in Word 7)

Agreeing on meaning separate from behavior

- Meaning must be established by human agreement
- XML documents will be generated from a wide variety of sources
- We need a system for associating meanings with XML components
- XML Namespaces (http://www.w3.org/TR/) is a first attempt at solving this problem

Key sources of information about XML

- **The W3C activity:**

  http://www.w3.org/XML
- **Standards and drafts:**

  http://www.w3.org/TR
- **Markup technology in general:**

  http://www.oasis-open.org/cover/

XML and Open Standards

The XML accomplishment

XML is a **big open-standards victory** for users.

- Freely extensible
  - No tag name limitations
  - No language limitations
- Human-readable
  - Can maintain data using basic text tools like sed, awk, perl, Word macros
- Open standard
  - In theory, XML users can't be held hostage to vendor control
- Easy to implement
  - There will be many powerful, cheap, off-the-shelf commercial XML tools
  - There is already an ever-growing set of free XML tools (almost all of them Java-based)

XML inherits the open-standards agenda of SGML

SGML has always been about **control of content**.

- Freedom from proprietary data formats
- Vendor neutrality
- Platform neutrality
- Language neutrality

XML is about data independence in the same way that Java is about program independence.

Completing XML

- XML schemas: Enhancements to 1.0 DTD functionality
  - Data typing
  - Inheritance
- DTD, schema, and namespace standardization
  - Should be developed by users, not vendors
  - Independent registries (e.g., OASIS: http://www.oasis-open.org)
- XML linking (XLink and XPointer)
  - Fully extensible
  - Addressing by structure
  - Links outside of documents
- XSL stylesheets
  - Media-independent output
  - User-configurable views
  - High-quality typography

W3C XML Working Groups

The original XML Activity has been divided up among six W3C working groups:

- **XML Linking WG:** Developing next-generation hypertext mechanisms
- **XML Schema WG:** Developing a DTD alternative that will support inheritance and data typing
- **XML Fragment WG:** Developing a method for transmitting XML fragments
- **XML Infoset WG:** Defining the XML objects that are available to applications
- **XML Syntax WG:** Investigating XML profiles and style sheet linking
- **XSL (Extensible Stylesheet Language) WG:** Developing an industrial-strength internationalized stylesheet language for XML uses on and off the Web

XML activity coordination

- XML Coordination Group
  - Schedules the work of the XML WGs
  - Coordinates dependencies between XML specifications
  - Interacts with standards efforts outside of W3C
- XML Plenary
  - Union set of participants in the XML working groups
  - Decides major XML policy issues

XML CG and Plenary Chair: Jon Bosak

Opening the process

Organizations outside of the W3C can contribute to the XML design activity in three ways:

1. Join W3C!
2. Establish formal liaison through the XML CG
    - Status: Liaison process not quite in place yet
3. Comment on publicly available requirements documents
    - First public RD for Fragment WG already available
    - Public RDs for the other XML WGs should be available in late January
    - Process includes periodic "checkpointing" and revision in light of comments
    - Process is currently experimental -- XML activity only