XML and Unicode: Implications for Global Publishing

Thirteenth International Unicode Conference

San Jose, September 1998

Jon Bosak
Sun Microsystems

XML Basics

What is XML?

- Extensible Markup Language
- An activity of the World Wide Web Consortium (W3C) organized by Sun to put SGML on the World Wide Web
- Will create 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
- Will fundamentally change publishing on the web and then publishing in general

Who developed XML?

The XML specification was developed from 1996 through 1998 by a wide-ranging group of markup language experts from industry and academia.

- **Computer industry:** Sun Microsystems, Hewlett-Packard, Microsoft, Netscape, Adobe, Fuji Xerox
- **SGML vendors and system integrators:** ArborText, Inso, SoftQuad, Grif, Isogen, Texcel
- **Academic and research community:** Text Encoding Initiative (TEI), NCSA, James Clark
- **Early adopters:** DataChannel, Vignette
- **Recent additions (after XML 1.0):** IBM, Oracle, Omnimark, Lotus, Xerox, SAP, Merrill Lynch, ...

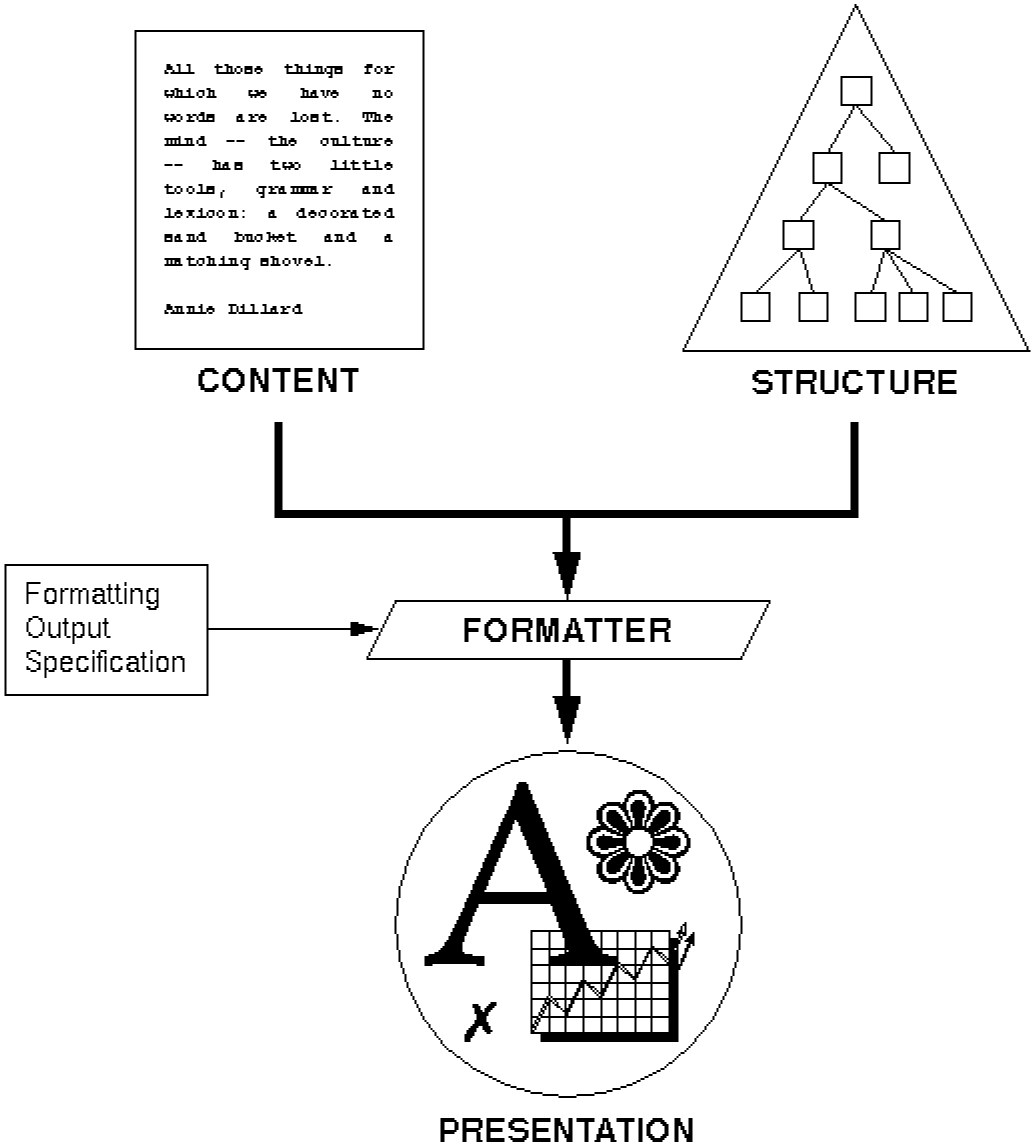
Where is XML going?

It is now generally agreed that Web content will be managed using standards based on XML.

Key predictions for the future:

1. XML will certainly be the basis on which future Web standards are built.
2. XML will probably become the universal serial format for structured data.
3. XML will almost certainly become the basis for international publishing.
4. XML may replace all existing word processing and desktop publishing formats.

The XML concept



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

Why take this approach?

Separating content and structure from presentation and behavior makes possible

- Reusable information
- Media-independent publishing
- One-on-one marketing
- Intelligent downstream document processing
- Large-scale information management

XML Internationalization

XML and Unicode

- XML has been based on Unicode from Day One
  - There is nothing in an XML file but Unicode characters
  - Unicode is used for both content and markup (can mix languages, even in tag names)
  - XML tools **must** support both UTF-8 and UTF-16
- The widespread adoption of XML for data management and electronic commerce will probably make Unicode support universal

Unicode example: an international bookstore

```
<?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)

Lessons from the example

- The catalog example shows that the distinction between data exchange and publishing is ultimately an artificial one (the same source would also be used to create the printed catalog)
- The rendition in each case occurs **on the web client**
- The database owner can publish **a single data stream** to the entire world
- Consider the alternative:
  - Generation of a different HTML output stream for **every possible** user and target platform
  - Much greater load on the server
  - No user autonomy

XML and Standardization

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)

Current W3C XML work

Work on the XML family of standards is taking place in six different 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 Fragments WG:** Developing a method for transmitting XML fragments
- **XML Information Set WG:** Defining the XML objects that are available to applications
- **XML Syntax WG:** Investigating XML simplification 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 and Global Publishing

Beyond a narrow interpretation of XML

1. XML is not just about exchanging data between machines
   - It's also about communication between humans
2. XML is not just about the web
   - It's about information in general
3. XML is not just about technology
   - It's also about the relationship between content creators and software vendors

XML inherits the social agenda of SGML

The social agenda of SGML has always been about **freedom from third-party control of content**.

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

Its foundation on Unicode adds **practical internationalization** to the implicit social agenda.

Next phase of the revolution: XSL

Large-scale cross-platform XML publishing demands that XML deliver on the *display-oriented* promises:

- User-configurable views
- More powerful display-centric client-side applications
- Media-independent publishing
- In particular, printed and online deliverables from the same source
- Asian-language rendering support

XSL is intended to complete the internationalized media-independent publishing story.

Implications of XML+XSL for I18N

- Properly done (like DSSSL), XSL enables XML to serve as an international publishing format
- XML+XSL could allow the whole operation to be based end-to-end on a uniformly processible Unicode text stream
  - Data
  - Stylesheets
  - Formatting objects
  - Instructions to the formatter
- A mass transition to XML+XSL will change the whole translation business
  - Data human-readable and freely interchangeable
  - Tools cheap and freely interoperable
  - Qualified personnel available everywhere
  - A **common infrastructure**

The larger implications of XML+XSL

The combination of XML and XSL can replace all existing word-processing and publishing formats.

- A single format for both print and online publishing
- A single format across different products
- A single format for all languages

What does this mean?

- Users no longer tied to a proprietary format
- A change in the relationship between software vendors and customers
- An end to control of the market by a few big companies
- An end to control of the market by a few big countries

What can go wrong?

The complete implementation of XML and XSL means an end to control of users through proprietary formats.

Companies that have built their business models on proprietary formats can be expected to resist this.

The most obvious ways to subvert the user-empowerment agenda of XML are:

- Vendor control of standard schemas
- Incomplete Unicode and XSL support

Getting to global publishing

- Beware of strategies that limit XML to the role of middleware; it's more than that
- Support the vendor-independent development of standard schemas (DTDs) and namespaces
- Beware of sleazy dodges: nonstandard extensions, incomplete UTF-8/UTF-16 support
- Insist on real XSL support: the ability to render formatting objects, not just HTML tags
- W3C members: work within W3C to make sure that the requirements for international publishing are met
- Support platform-independent tools vendors
- Support the only organization dedicated to interoperable document standards in general: OASIS (http://www.oasis-open.org)

The payoff for XML and Unicode

- Interoperability of both content and style
- Freedom from vendor control of our data
- Creator control of markup depth
- User control of views into the data
- A level playing field for independent software developers
- A common infrastructure for translation
- True international publishing across all media