XML Overview

ebXML Meeting
San Jose 1999.11.17

Jon Bosak
Sun Microsystems

XML Fundamentals

What is XML?

- XML is a document format
- XML is a data format
- XML is 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 international standards for markup, linking, and formatting
- Sources of information about XML:
  - http://www.w3.org/TR/
    - Standards and drafts
  - http://www.oasis-open.org/cover/
    - Markup technology in general

What is XML good for?

- Internationalized electronic publishing
  - Platform-independent
  - Language-independent
  - Media-independent
- 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
  - Business transactions

Why did Sun invest in XML?

Sun organized and led the XML activity, beginning in 1996. Sun continues to lead the XML activity today. Why?

1. HTML was not enough for publishing on the Web.
2. HTML was not enough for distributed applications.
3. Without XML, HTML would have been replaced by a more powerful binary proprietary format.

We developed XML to keep Web data **open and portable.**

XML and users

- Big open-standards victory for users
  - Freely extensible
  - Human-readable
  - Machine-comprehensible
  - Easy to implement
  - Completely internationalized
- "Data the way you want it" (Microsoft)
  - User control of views into data
  - Interoperability of content, style, and behavior

XML in Context

The XML family of standards

XML is one of a **family** of hypertext standards.

- **XML** (Extensible Markup Language): Powerful data structures that are easy to implement
  - Is replacing HTML in industrial contexts
- **XLink/XPointer**: Next generation hypertext linking
  - Will replace HTML linking in industrial contexts
- **XSL** (Extensible Stylesheet Language): A standard stylesheet language for XML publishing
  - Will replace CSS in industrial contexts

XML itself

- A simplified subset of SGML (ISO 8879)
  - Builds on 30 years of research and 14 years of standardization
  - Widely deployed (as SGML) in major industrial settings
  - Powerful data modeling -- no limits on namespace or structural depth
  - But easy to implement and small enough for Web browsers
- Not a language but a metalanguage
  - Designed to support the definition of an unlimited number of languages for specific industries and applications
  - All XML languages can be processed by a single lightweight parser

What XML is not

- XML provides "syntax, not semantics"
  - Tags have no predefined meaning
  - Unlike HTML, XML by itself conveys only content and structure, not presentation, behavior, or meaning
- Associating **presentation or behavior** with XML requires additional mechanisms
  - Industry agreements on the processing of particular XML languages (example: HTML)
  - Downloadable programs, applets, or scripts designed for a particular XML language
  - Stylesheets
- Associating **meaning** with XML requires additional mechanisms
  - Namespaces (identity hooks for meaning)
  - Prose (natural language definitions)

XML in one slide

- Legal XML documents are called **well-formed**
- A well-formed document describes a **logical tree**
- If a well-formed document conforms to an optional set of constraints (a DTD), it is also **valid**

A well-formed XML document:

```
<greeting type="friendly">Hello, world!</greeting>
```

A valid XML document:

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE greeting [
  <!ELEMENT greeting (#PCDATA)>
  <!ATTLIST greeting type (friendly | unfriendly)
                                    "friendly" >
]>
<greeting>Hello, world!</greeting>
```


Documents versus Data

What's a document?

A document is data that you can **read**.

- Documents are a **superset** of data
- Documents are **harder** than data

The basic problem with documents is that we need to display them in lots of **different forms**. This is the problem that XML was originally designed to solve.

Example: This presentation.

- Written in XML
- HTML generated using a stylesheet for online publishing
- RTF generated by the same tool using a stylesheet for print publishing

What does XML do for data?

XML standardizes the **concrete syntax** of data exchange in a **text-based notation** designed to be **obvious** to both people and processes.

Deploying XML creates an **open, distributed information infrastructure.**

1. Standardized parsers for putting data into memory
2. Standardized interfaces (tree-oriented and stream-oriented) for processing the data
3. Standardized ways to display data
4. Standardized ways to query data
5. Standardized ways to link data
6. Standardized training of people in both publishing and data processing

Separation of data from processing

- The XML publishing model decouples data from processing
- This isolates changes in large systems, making them more flexible and reliable (IBM)
- Basing a system on XML makes it well-suited to transactional processing in a heterogenous, asynchronous, distributed environment (like the Web)

Using the document layer

XML uses documents as the transfer mechanism for data. Everything is "documented."

- Easier to track processes
- Creates a large body of data captured in ways that are standard to an industry or enterprise, thus enabling
  - Data mining
  - Long-term retrospective trend analyses
  - Business intelligence (BI)
- XML documents model financial and commercial data very well
  - We've been doing business with documents for thousands of years
  - The document orientation brings our existing legal and commercial frameworks onto the Internet

Documents + data = success

"By the year-end 2002, XML's impact on publishing (in particular) and technology (in general) will be as profound as the Web's impact is today." (Gartner Group)

"XML will revolutionize the way data is stored, processed and retrieved." (BancBoston)

"XML will be like oxygen in the networked environment of the twenty-first century: vital, pervasive, and taken for granted." (Burton Group)

The fork in the road

The world of business is going online. It can go one of two ways.

1. Change 5000 years of business practices to optimize electronic data processing.
2. Structure data processing to model traditional business practices.

A traditional business transaction is an exchange of **documents**.

The move to adopt XML means that the marketplace is choosing ease of data management over ease of programming.

XML and Java

XML and Java are twin pillars of vendor-neutral programming.

- XML = portable, reusable data
- Java = portable, maintainable code

XML works best in Web-like environments.

- Asynchronous
- Loosely coupled
- Heterogeneous

Text formats are **not** the answer to every problem!

The XML trade-off

XML trades off

- Performance
- Centralized control
- Uniformity

in order to get

- Persistence
- Distributed control
- Asynchronicity
- A certain kind of readability (like source code)

XML Standardization

Key XML standards organizations

- World Wide Web Consortium (W3C)
  - http://www.w3.org/
  - Develops core XML technologies
  - Develops other standards related to XML
  - Develops other web technologies not directly related to XML (e.g., P3P)
- OASIS
  - http://www.oasis-open.org/
  - Organization for the Advancement of Structured Information Standards
  - Non-profit corporation dedicated to XML interoperability and conformance
  - Serves as a framework for the development of XML languages for specific applications

The W3C

- Over 350 members
- Key W3C XML working groups:
  - **XML Core WG:** Basic XML syntax
  - **XML Linking WG:** Next-generation hypertext
  - **XML Schema WG:** Next-generation DTDs
  - **XML Query WG:** Standard query language for XML
  - **XML Packaging WG:** Packaging XML components
  - **XSL WG:** XML formatting/transformation language
  - **DOM WG:** Structurally oriented XML API
- Working group decisions are only advisory; **all final W3C decisions are made by the Director**
- XML Activity coordination is provided by the **XML Coordination Group** and the **XML Plenary** (both chaired by Sun)

Why OASIS?

OASIS is the only independent organization solely dedicated to the standardized, open exchange of structured data.

- Works on specific applications of XML (W3C works on core technologies)
- Well-established non-profit corporation
- Open to everyone, even individuals
- Run by an elected board of directors
- Supported by the whole industry: Sun, IBM, Microsoft, Adobe, Oracle, Software AG, SAP, Boeing, Xerox,...

XML Teamwork

The industrial function of XML

XML is a framework for developing an unlimited number of special-purpose data languages.

XML allows people sharing a common data exchange problem to work out an **open solution** to that problem.

- Without interference from third parties
- Without dependence on large software vendors
- Without bindings to specific tools
- Without language restrictions
- In a way that lets anyone with a similar problem use the same solution

XML provides **a standard framework for making agreements about communication**.

XML agreements

- Industry DTDs
- Industry schemas
- Industry namespaces

Schemas and namespaces define a **common language**.

The XML agreements in each industry will define data in that industry for many years.

XML agreements must be very carefully constructed!

The dangers of short-term thinking

Rushing to make money will result in bad XML languages.

- Carelessly designed
- Driven by narrow national interests
- Owned by one company
- Tied to specific products
- Hosted on one vendor's web site
- Written in a proprietary schema language
- Not interoperable with each other
- **Not built for the future**

How do we build good languages?

The process for developing XML industry schemas and namespaces must be **careful, open, and fair**.

- Represents all the stakeholders in a user community
- Uses a democratic process
- Allows real differences to be resolved
- Creates intellectual property that is owned by everyone in the community
- Allows time to test implementations
- Provides a framework for distribution and acceptance

Where can we work together?

OASIS is an **international organization** where **individuals and companies** can build XML industry standards within an **open process**.

- A process for starting new XML standardization efforts
- A process for harmonizing competing XML standardization efforts
  - Electronic commerce
  - Templates for business documents
  - etc.

How can we share our work?

We need a way to publish XML agreements on the Web.

- **Human search:** what existing industry standard meets my needs?
- **Machine resolution:** where is the DTD (schema, namespace, etc.) pointed to in this document?
- **Versioning:** what is the latest version of a standard?

What's needed is an **XML registry and repository.**

Requirements for an XML registry

- Self-supporting
- Universally accessible
- Trusted and vendor-neutral
- Based completely on international standards
- Low cost of entry for individuals and small organizations
- Distributed model

Answer: **XML.org** (http://xml.org/).