Deborah Aleyne Lapeyre
Mulberry Technologies Inc.
17 West Jefferson St.
Suite 207
Rockville MD 20850
Phone: 301/315-9631
Fax: 301/315-8285
dalapeyre@mulberrytech.com

What This Course is NOT About

- Ecommerce, eBusiness, B2B, B2C, new business models
- PPML (Personalized Print Markup Language) or XML for job tickets
- Syndication of content (PRISM et al.)
- Physical interchange of XML and packaging
  - (SOAP, XML-RPC, etc.)
- XML APIs (Sax, DOM, JDOM, et al.)
- XML Tools (XML composition systems, XML editors, XML repositories/databases, XML portals)

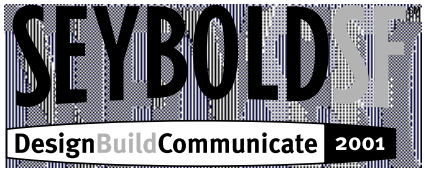
XML for Publishing Managers

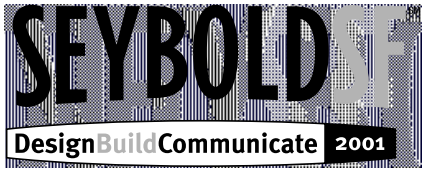
Mulberry Technologies Inc.
17 West Jefferson St.
Suite 207
Rockville MD 20850
Phone: 301/315-9631
Fax: 301/315-8285
info@mulberrytech.com
http://www.mulberrytech.com

August 2001

Mulberry
Technologies, Inc.

XML for Publishing Managers





slide 1

Administrivia

- Start, end, break
- How this will work
- Questions are always in order
- Why this course
- Anything else?

slide 2

Where We Are *Not* Going in This Talk

- Ecommerce, eBusiness, B2B, B2C, new business models
- PPML (Personalized Print Markup Language) or XML for job tickets
- Syndication of content (PRISM et al.)
- Physical interchange of XML and packaging (SOAP, XML-RPC, etc.)
- XML APIs (Sax, DOM, JDOM, et al.)
- XML Tools (XML composition systems, XML editors, XML repositories/databases, XML portals)

slide 3

Where We Are Going Today
XML as Content

XML as your text and your data

- What is XML and how it works
- Where can XML fit?
- XML versus HTML
- XML and PDF
- Living with XML (the changes)
- Information Resources

slide 4

A Quick Poll (Who You Are)

How many of you are (or manage people who are)...

- Book publishers (monographs, reference series, etc.)
- Journal publishers
- Technical Documentation publishers
- System analysts or application programmers?
- Trainers or training publishers (CBT, web, textbook, etc.)

slide 5

What You Know

How many of you:

- Know HTML?
- Use XML or SGML?
- Produce in PDF now?
- Know
  - Quark
  - Microsoft Word Templates?
  - High-powered composition systems such as Miles 33, Penta, 3B2, DataLogics

What is XML?

slide 6

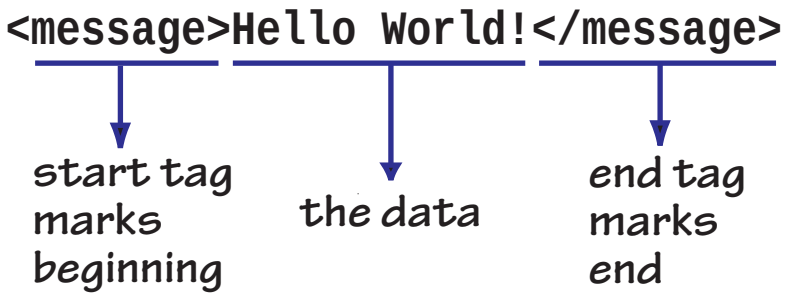
The Word "XML" is Used to Mean:

- An open standard (well ... a W3C recommendation) that provides:
  - A data format
  - A data modeling language
- The use of XML-formatted data in an application (like a browser)
- A metalanguage for creating markup languages
- A set of associated recommendations and specifications
  (link, style, transformation, query, APIs, etc.)

XML Works through Tags

Paired tags:

- Enclose data
- Identify/name the data
- Named component called an “element”



slide 8

XML Documents

- In XML jargon, your data (no matter what form) is called a "document"
- A document is a coherent, ordered collection of information:
  - invoice
  - journal article
  - topic in a help system
  - sales catalog
  - database load file
  - reference book
  - driver code for a hand-held device

slide 9

How XML Looks At Data

Documents

- are made up of *Elements*
- consisting of *Markup* ("tags")
- ... and *Element content*

slide 10

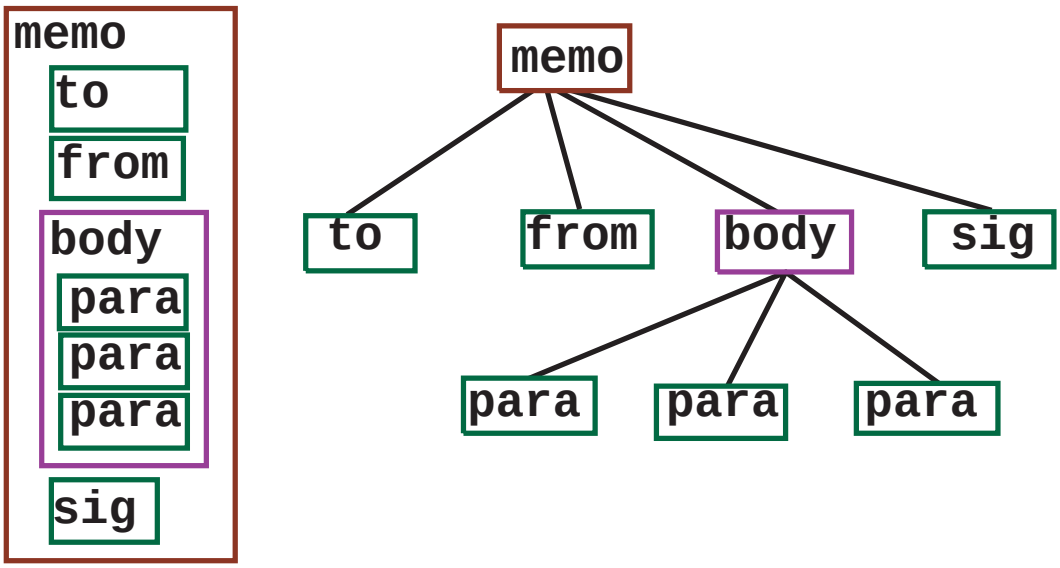
XML Elements

An *element* is an identifiable, named component of a document (payment term, paragraph, part number, title, author, unit price, bulleted list)

- Can have content (data, other elements)
- Can be a pointer to information (hypertext link, table reference)
- Must be contiguous (one start and one end; no holes in the middle)

slide 11

Elements Can Nest



slide 12

Attributes Add Further Description

- Live inside start tags
- Say something *about* the data
- Add information to the our knowledge of the element

```
<phoneNumber type="unlisted" rate="premiumplus"
   assigned="1996-04-01">301/315-9631</phoneNumber>
```

slide 13

XML Isn't:

- A programming language
  Does not replace C++, Java, perl, Python, ...
- A user interface
- A presentation format
- A formatting or processing system
- A standard set of tags
- A recommended set of tags

XML Is a Data Format

"Employee Record" Example

In XML we can separate data content from behavior/presentation.

```
<employee-record type="dog" empno="9">
<name>
<first>Sasparilla</first>
<last>Usdin</last></name>
<affiliation>
<title>Deputy in Charge of Chewables</title>
<company>Mulberry Technologies</company>
<location><city>Rockville</city>
<state>MD</state><zip>20850</zip></location>
<email-name>sassy</email-name>
</affiliation>
<height unit="in">36</height><weight unit="lb">70</weight>
</employee-record>
```

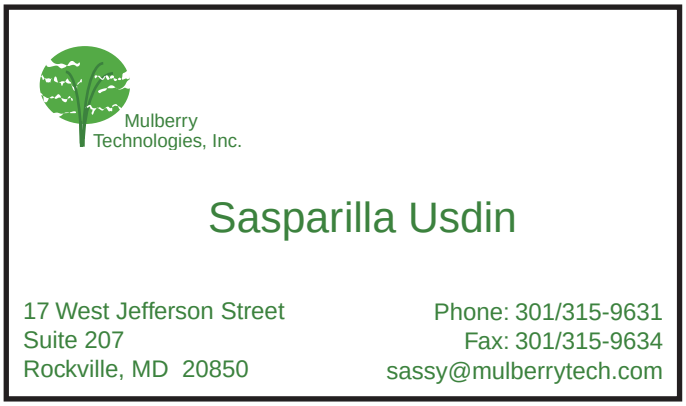
slide 15

View This in a Browser

Convert into HTML (today). Or in an XML browser (tomorrow):

slide 16

A Familiar Print Application



slide 17

Same Data, Different Application

New Employee Announcements

Sasparilla Usdin
has recently joined Mulberry Technologies, Inc.'s Rockville staff as Deputy in Charge of Chewables.

Welcome to the team, Sassy!

- XML elements rolled into "form letter"
- Something (perhaps employee-id) linked to photo

slide 18

Same Source: Load a Database

```
Key: 00095AUS
EMPNO: 009
001:USDIN
002:Sasparilla
008:36
014:70
020:Deputy in Charge of Chewables
```

slide 19

Ultimate Purpose of XML

- Encode (mark up) data only once
- Produce many products from that markup
- Enable semantically complex searching
- Reuse data (in whole or part) many times
- Interchange data freely
- Enable machine-to-machine communication
- Let whole communities agree on data content
- Let data live a long time

slide 20

Bulletin:
XML
DOES NOT DO
ANYTHING!

XML is a data format.

slide 21

You (and Your Software) Can Do a Lot with Markup

- Start/stop behavior
  - Complex behavior (run a script, load a database)
  - Linking
  - Formatting (start bold, end bold)
- Process
  - Extract selected elements
  - Rearrange/resequence content
  - Rename, add content
  - Count how many

slide 22

Types of XML Markup

- Content of the data
- Structure of the document
- Value-added information
  - Location and Navigation
  - Metadata
- Rendering/Processing information (presentation and formatting)

slide 23

Content Markup

What type of information is this?

- Environmental Impact
- City, state, zip code
- Question, answer
- Methodology section
- Part number
- Executive Summary
- Drug Dosage section

slide 24

Structure Markup

What part of the document is this?

- Paragraph
- Title
- Figure
- Chapter
- Table
- Signature block
- List
- Bibliographic header

slide 25

"Location or Navigation" Information

Added to text to make it more functional, useful or manageable

- Hypertext links
- Cross-references
- Indexing terms

slide 26

Metadata (Data About the Data)

- Bibliographic information
- Index or search terms
- Revision or version
- Status and workflow tracking information
- Data source
- Editor's or reviewer's comments
- Abstracts, teasers, cataloging data

slide 27

Rendering/Processing Markup

How text should behave, display, or print

- A complete script to call and perform
- Iconization
- Position of graphics on the page
- Line breaks in titles
- Visual or auditory highlighting
  (sometimes a word is **bold** just because the author said so)

slide 28

XML is also a "Metalanguage"

- A Metalanguage is for defining custom tag sets
- Tags sets get called "languages"
- Languages can be built for
  - Problem domains (journal publishing, textbooks)
  - Applications (like eBusiness, content-management)
  - Information collections (reference works, laws and statutes, biographies, dictionaries)
- Different markup languages for different *information types*

slide 29

New XML Markup Languages

- Not really languages but a set of agreements. May include:
  - Sets of tags
  - Problem and process models
  - Document or message models (DTDs and schema)
  - Vocabularies and dictionaries
- Discipline-oriented like CML (chemistry) and MathML (mathematics)
- Industry oriented like Airlines/aircraft and Semiconductors
- Process-oriented like SVG (Scalable Vector Graphics)

slide 30

There Are Many XML "Languages"

- Vertical market XML "languages"
  (Industry-specific and Cross-industry XML Specifications)
  - Banking and Financial Services
  - Health care
  - Research communities
  - etc. etc.
- XML "languages" for different media
  - Screen display/hypertext (web)
  - Print
  - Handheld devices, heads-up displays, voice, etc. etc.

Parts of an XML Application

slide 31

Logical Components of an XML Application

- XML document (tags and text)
- DTD or Schema (the model)
- Output Specifications (how looks/behaves)
- Transformations (from here to there)

slide 32

Component: XML Document

The tags (markup) and the text (content)

- Two types
  - Well-formed
  - Valid (has a model)
- Usually created
  - using an XML editor (authoring)
  - by a program from
    - a database
    - another type of XML file by transform
    - conversion from another format (like Quark or Word)

slide 33

Component: The Document Model— DTD (Document Type Definition)

The modelling mechanism specified by the XML specification

- Models one type/class of information (a "document") (reference book, bank transfer, journal article, memo, help-topic)
- Is a set of rules describing how documents of that type can be marked up
- Is written in the formal syntax of XML

slide 34

DTDs Express Rules

for example:

- **Journal Article** = *metadata* followed by *article body*, followed by optional *back matter*
- **Purchase Order** = *Order Header* followed by *List of Order Detail*, followed by optional *Order Summary*
- **paragraph** = data characters and may include any of the following: *Person Name*s *URL*s, and/or *Geographic Region*s

slide 35

Why Use a DTD or Schema?

- DTD is a contract between producers and consumers
  (Both can validate to see if they got/sent what they expected)
- Formal specification of information *types* allows consistent downstream processing
- Supports interoperable families of documents
  - Ensure that information conforms to model (validation)
  - Parties don't have to share software or applications

slide 36

To Share Information, Share a DTD

- Publisher communicates to conversion house
- Content provider explains tagging to
  - Compositor for typesetting
  - Web designer for building website
  - Database or repository designer
  - Software vendor for customization

slide 37

XML Schemas Provide New Features

- All the functions of a DTD
- Strong data typing (date, time, integer, string, boolean, etc.)
- Inheritance mechanisms
- Default and required values for content
- Built-in documentation elements
- New relationships among elements

slide 38

Problem: Competing Schema languages

- TRUTH: W3C XML Schemas are very new (official spring 2001)
- Many people (particularly EDI-folks) are using schemas *now*:
  - Microsoft's XDR
  - CommerceOne's SOX
  - Relax (the Japanese alternative)
  - TRex
  - DCD
  - X-Schema
  - Etc., etc. and so forth!

Component: Output Specification/Stylesheet

Formatting: Remember What XML Looks Like!

```
<?xml version="1.0"?>
<RESUME>
<CONTACT.INFO>
<NAME>Heinrich Rudolf Hertz</NAME>
<ADDRESS>Bonn, Germany</ADDRESS>
</CONTACT.INFO>
<OBJECTIVE>To continue researching electrical discharges in
rarefied gases in an academic setting.</OBJECTIVE>
<SUMMARY><PARAGRAPH>Over ten years academic research
studying electromagnetic waves.</PARAGRAPH>
</SUMMARY>
<WORK.EXPERIENCE>
<JOB.BLOCK>
<EMPLOYER><NAME>University of Bonn</NAME>
<LOCATION>Bonn, Germany</LOCATION></EMPLOYER>
<DATES>
<START.DATE>1889</START.DATE>
<END.DATE>1894</END.DATE></DATES>
<JOB.TITLE>Professor of Physics</JOB.TITLE>
<RESPONSIBILITY>Research the discharge of electricity
in rarefied gases</RESPONSIBILITY></JOB.BLOCK>
<JOB.BLOCK>
<EMPLOYER><NAME>Karlsruhe Polytechnic</NAME>
<LOCATION>Germany</LOCATION></EMPLOYER>
<DATES><START.DATE>1885</START.DATE>
<END.DATE>1889</END.DATE></DATES>
<JOB.TITLE>Professor of Physics</JOB.TITLE>
<ACTIVITY>Produced and studied electromagnetic waves
(radio waves), confirming Maxwell&apos;s electromagnetic theory
</ACTIVITY>
<ACCOMPLISHMENT>Established that light
and heat are electromagnetic waves (1887); first to produce
radio waves artificially.</ACCOMPLISHMENT></JOB.BLOCK>
</WORK.EXPERIENCE>
<EDUCATION>
<SCHOOL><NAME>University of Berlin</NAME>
<DEGREE>Ph.D. magna cum laude</DEGREE>
<PROGRAM>Physics</PROGRAM>
<GRANT.DATE>1880</GRANT.DATE></SCHOOL>
</EDUCATION>
<PUBLICATIONS><PARAGRAPH><TITLE>Electric Waves
</TITLE> (1893);<TITLE>Miscellaneous Papers</TITLE> (1896);
<TITLE>Principles of Mechanics</TITLE> (1899)</PARAGRAPH>
</PUBLICATIONS></RESUME>
```

slide 40

XML Separates Content from Format/Behavior

How it looks (16 pt Helvetica Bold) or what it does (starts a javascript)

- Is based on the tagging
- Is the same for every tag *in the same context*
  - NOT one tag per one format
  - Table title may differ from Figure title from Chapter title

slide 41

What We Would Like to See (Print or Screen)

Heinrich Rudolf Hertz
Bonn, Germany

Objective: To continue researching electrical discharges in rarefied gasses in an academic setting.

Summary: Over ten years academic research studying electromagnetic waves.

Experience:

1889—1894	University of Bonn, Bonn, Germany Professor of Physics Research the discharge of electricity in rarefied gasses
1885—1889	Karlsruhe Polytechnic Professor of Physics Produced and studied electromagnetic waves (radio waves), confirming Maxwell's electromagnetic theory. Established that light and heat are electromagnetic waves (1887); first to produce radio waves artificially.

Education:

slide 42

Use an Output Specification to get There

(Frequently called “Stylesheet”)

- Says what XML data will look like or how to behave
  - On screen or paper
  - Or in other media (for example in audible output)
- Defines an appearance or rendition or behavior
  - For each element
  - In each of its contexts within a document

slide 43

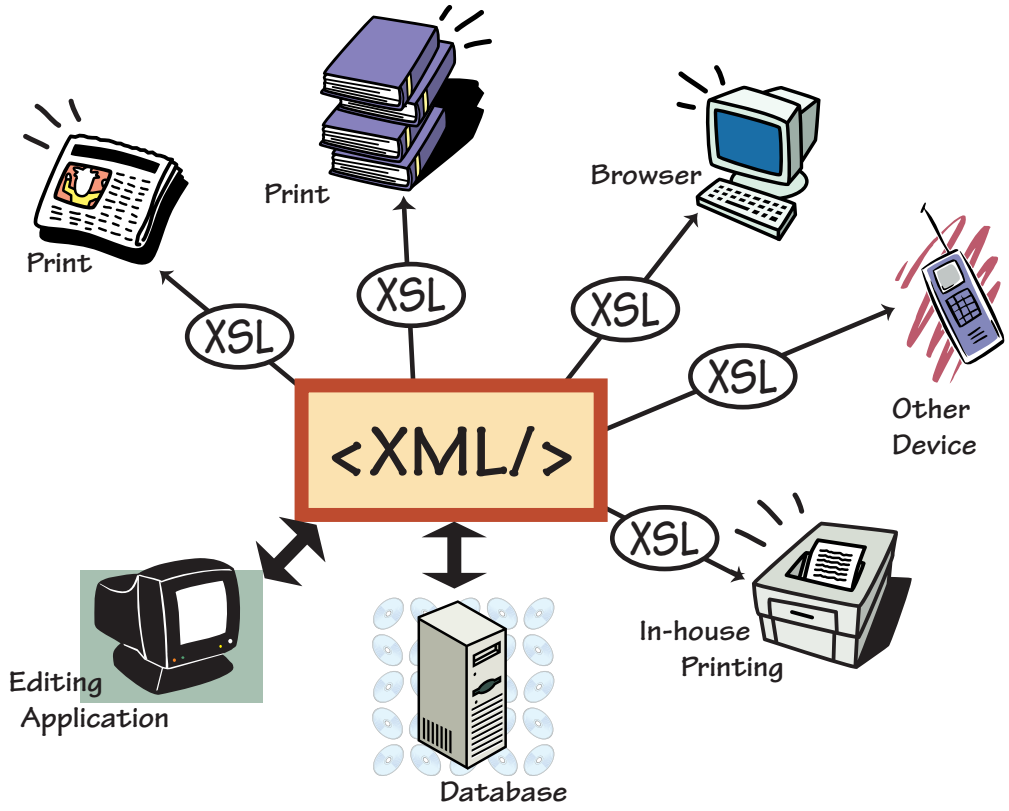
Documents and Stylesheets

- One stylesheet, many documents
  - Maintains consistency of format (“look and feel”) across documents
  - Is easy to develop, maintain, and apply (house style)
- One document, many stylesheets
  - Allows for different media types: print, on-line, etc.
  - Is easy to produce derivative documents: selections, summaries, indexes, catalogs ...

Component: XML Transforms

slide 44

XML and XSL



slide 45

XSL For XML Transformation (XSLT)

Transforms from one set of tags directly into another
Transform XML into

- HTML for browsers
- Other (XML) tag sets for further processing
- Plain text formats (e.g., loader files for databases)
- Non-XML tag sets

slide 46

Why This is Exciting

- Conventional wisdom is that to make XML useful *we must all use the same tags*
- Transformation means maybe not
- Transform my tags into your tags
- We can alias elements and content *when elements are:*
  - Semantically the same but different names
    (purchase-order, po, order, client-purchase)
  - Semantically or functionally “close enough”
    (postal-code, zip-code, parish-number)
  - My elements are recombinations or subsets of yours

slide 47

Different DTDs for Different Purposes

(with transformations between them)

- Authoring DTDs to build document or database
  (very strict rules, enforcing)
- Interchange DTD to exchange data
  (very loose rules, enabling)
- Conversion DTD
  (so loose it's just a tagset, to describe legacy collections)
- Output DTD for information rearrangement or subsetting

Where Does XML Fit in the Publishing Process?

slide 48

XML Files

- Are "plain text" underneath
  - Use any text editor or any word processor that can handle plain text
  - Built on Unicode (represents all major scripts of the world)
- XML could be
  - Native file format for a software package
  - Something you "import" and "export"

slide 49

XML for Print and Web Publishing

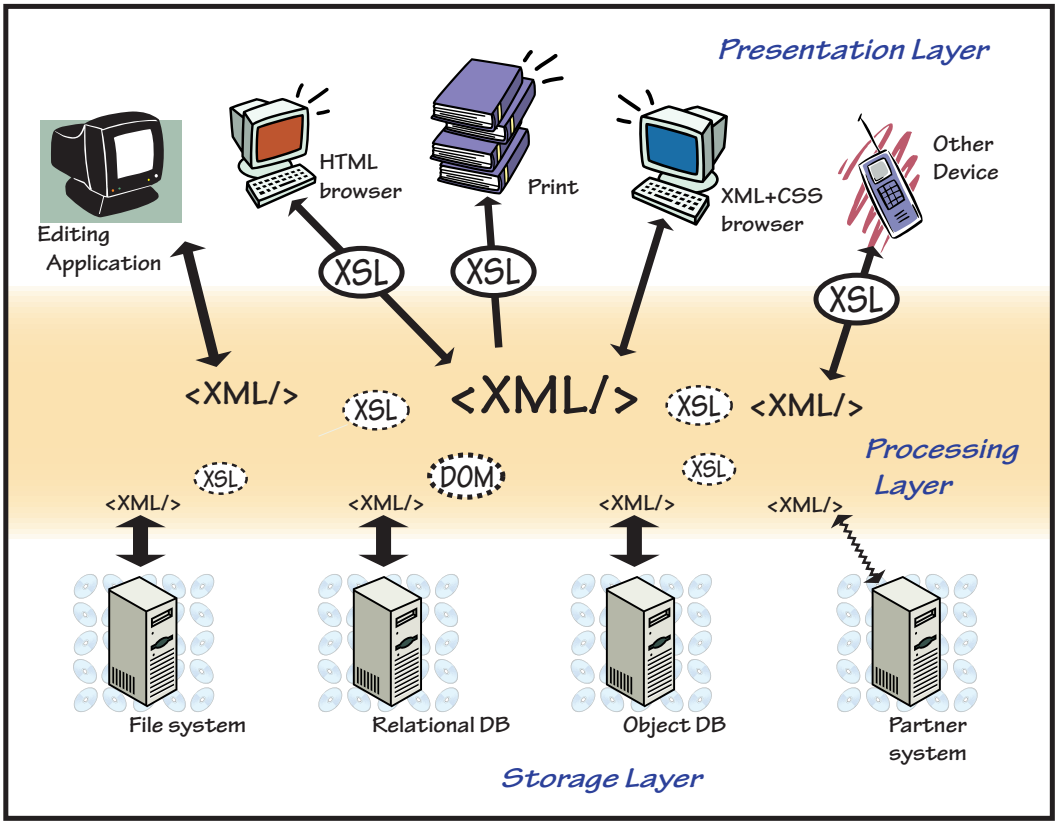
- Many different outputs, one manageable source
  - Many media/device types
    (Web, CD-ROM, handheld PDAs, voice-synthesis)
  - Many styles of print/display
- Different hardware, software, OS for input, manipulation, display
- Publish on demand/Customized output

slide 50

XML Between Application Layers

- In the "Three-tier" system model:
  - Presentation/User interface layer
  - Processing or "business logic" layer
  - Storage or repository layer
- XML used in any of the three tiers, *especially in the middle*
- XSL is used for any processing
  - Within the middle tier, and
  - *Between* tiers

A Typical “3-tiered” Model



slide 52

XML Tagged Data for Information Interchange

- By data aggregators (scientific and journal websites, semiconductor industry)
- Through the life-cycle of a product (among divisions)
- Direct machine-to-machine transfer
- Between rival proprietary formats
- Inter-process communication (IPC)
- Among business partners (E-Business transactions B2B, B2C, EDI replacing EDIFACT or proprietary formats)

XML and Data Repositories

slide 53

XML versus Databases

Not an issue

- XML is being used inside:
  - OO databases
  - Relational databases
  - Object-relational and hierarchical databases
- XML is also the communication
  - Between databases and applications
  - Between applications and databases
  - Between databases and databases

slide 54

Data-Driven Publishing

Creating print-ready output directly from a structured data source

- Used in catalog publishing/parts lists/directory data
- Many many companies (check the exhibit hall) including:
  - Banta Integrated Media
  - B-Bob Infinity
  - DataQuad
  - Enigma
  - IXIASOFT
  - Miles 33
  - LivePage

slide 55

Publishing Relational Data

- Store date in ordinary relational database
- Extract data
- Wrap tags around extracted data producing XML
- Transform or format XML source to produce
  - PDF
  - Postscript
  - Proprietary word-processor format such as .RTF (Word) or .MIF (FrameMaker) or Quark
- Example: directory data, print on demand

slide 56

Content Management Using XML Repository

- Content management at many levels of granularity
- Combine data form many sources
- Reuse and repurposing of data / Electronic slice and dice data
- Increase searching precision
- Feed many applications from one repository
- Customized output
- Enterprise information portals

slide 57

XML Can Improve Search Precision

```
<BugResponse><Reply-to-Customer>
<para><person cust.no="BL432">Mr. A.W. Black</person> has
reviewed the problem list you supplied and he believes the
<partNumber>540A</partNumber> you destroyed may be covered
under our seven-day warrenty period.</para>
...
<para>In addition the <partName>Slo-Mo-52</partName> has
been recalled due to a federal court order and will be replaced
by the <partNumber>Slo-Mo-53</partNumber> at no cost to
you, but is no longer available in black, only in blue or green.
</para></Reply-to-Customer>
...
<Internal-Comment>Send him the usual form letter.  This
is a black day for that marketing campaign.<Internal-Comment>
...
<Address>17 Black Oak Drive...</Address> </BugResponse>
```

slide 58

Aside: Search Effectiveness Problem

- When we search large bodies of information we:
  - Get too much stuff we don't want
  - Don't get all the stuff we do want (or don't know if we got all the stuff we wanted)
  - Have a hard time sorting the stuff we want from the rest

slide 59

This is Not a New Problem

- Measuring quality of information retrieval has been studied since the mid-1960s
- Measuring retrieval quality is very difficult
- The bigger and looser the collection, the harder to measure

slide 60

Definitions

- Swets defined Precision in 1963:
  - Relevance - the proportion of the retrieved stuff you want
  - Recall - the proportion of the relevant stuff is retrieved
  - Precision - the ratio of Relevance to Recall

slide 61

Precision

- Precision is NOT measurable in any real system
  - Cannot measure Relevance well
    (hard to measure how much of what is retrieved is of interest)
  - Cannot measure Recall at all
    (how do you know what you didn't find?)
- We can still aim to improve precision
- System users develop a "Feel" for Precision

slide 62

Improve Relevance Through Grammar

- The way parts of the information are identified
- Roughly, the tags you use
- Design and control to improve Relevance

slide 63

Grammar Can Enable Context Searching

- Markup should identify:
  - Key information (whatever size)
  - Major subdivisions of the information
  - Information-rich portions of information
- The stuff your users want to look for

slide 64

Examples of Markup that Supports Context Searching

Users in various domains might want to find a phrase if it occurs in one of these contexts:

- Conclusions Section
- Goals and Objectives
- Recommendations
- Long-term Implications
- Author
- Policies and Procedures
- Ingredients List
- Problem Description
- Prerequisites
- Typical Applications
- Question
- Purpose
- Answer
- Location

Grammar Can Enable Context Excluding

Markup should identify:

- Unimportant information
  (or information unimportant to some users)
- Secondary or supporting information
- Negated information
  (failed attempts, options not selected, what this info is NOT about)
- Large blocks that can be ignored

slide 66

Examples of Markup that Supports Context Excluding

- Background
- References
- Experimental Design
- Rejected Options
- Methodology
- Acknowledgments
- Introduction
- Veterinary Uses
- History
- Stage Directions
- Examples
- Geographic Names

slide 67

Markup also Supports Navigation

- Metadata
- Linking
- Bibliographic information
- Digital Object Identifiers
- Webs, Nets, Maps, Indices ...

slide 68

Relevance is Improved

- Find “black” in the description of a symptom instead of all occurrences of “black” in a medical database
- Find “shock” in that database except when it occurs in:
  - Names or People or Organizations
  - Documents about nervous or mental disorders
- Find books about Charles Darwin not references to Darwin Jones who lives on Charles Street

slide 69

Improve Recall Through Vocabulary

- The ways subjects are identified
- May be attributes, tags, or content
- Specialized to subject domain
- Designed and controlled to improve Recall

slide 70

Controlled Subject Vocabulary/Indexing

- Natural language is messy; the better the writer, the more variation in vocabulary
- Texts often fail to state the obvious, such as their subjects
- Subject indexing and cataloging were invented to provide subject access to large collections of material (like libraries)
- Structured thesauri help find more and less specific subject matter
- Indexes give access to medium-sized collections (like books)

slide 71

Examples of Controlled Vocabulary

- Equipment type might have a value of
  - Telephone desk-sets
  - Cordless Telephones
  - Cellular Telephone
  - Telephone Switching Equipment
- Subject Codes may be from:
  - Dewey Decimal Classification system
  - Biological Abstracts Header List
  - General Accounting Office Thesaurus

slide 72

Subject Indexing: Expensive and Valuable

- Requires intellectual effort
- Automatic indexing better than none, but not as good as expert human indexing
- If you'd pay for a back-of-the-book index in print, you should pay for the same access to electronic information

slide 73

Recall is Improved

- Find shade-loving plants not shade-intolerant plants in a garden catalog
- Find references to "Miner's Cough "when looking for "Black Lung Disease"
- Find books by Dr. Seuss when looking for books by Theodor S. Geisel

XML for Metadata

slide 74

Metadata

- Is information *about* the data
- Can include:
  - Bibliographic data for formal publishing
  - Application metadata (UML models, specifications, requirements)
  - Content access subect terms
  - Properties of user interfaces
  - Security levels, access, authority
  - Version control, configuration management, tracking

slide 75

Using XML Metadata

- Data repository in XML, metadata in separate database
- Metadata built into XML elements
  - Bibliographic headers of journal articles
  - Index terms/keywords
    - just before content starts
    - mixed with content
- Metadata built into XML attributes
  - Who/what/when/where/why associated with structure
  - Security or access levels
  - Information class or content type associated with structure

XML Compared to Other Formats

XML versus HTML

slide 76

Ultimate Purpose of HTML

- Display pages on the Web
- Easy to use (easy to create content)

slide 77

XML compared to HTML

- HTML is
  - *One (small) set of tags*
  - Tied to display/formatting (combines syntax and semantics)
  - Different for every browser
  - Difficult to scale, integrate, mix with other forms
- XML is
  - *User-defined tag sets*
  - Separate syntax and semantics (layered architecture)
  - Scales and integrates

slide 78

HTML Tags Format
XML *Can* Tag Content

```
<H2>Laptop Computer</H2>
<UL><LI>IBM Thinkpad 560X</LI>
<LI>333MHz</LI>
<LI>8GB</LI>
</UL>
```

versus

```
<COMPUTER CLASS="Portable"><MFR>IBM</MFR>
<FAMILY>Laptop</FAMILY><LINE>Thinkpad</LINE>
<MODEL>560X</MODEL>
<SPEED UOM="MHz">333</SPEED>
<DISK UOM="GB">8</DISK>
</COMPUTER>
```

slide 79

XML Will Replace HTML for:

- Users running into limitations of HTML
  - Retrieval
  - Formatting
  - Richness of encoding
- Users with complex data requirements (semiconductors, airlines)
- Users with complex security requirements
- Users with SGML data

slide 80

XML Won't Replace HTML for:

- Huge amounts of legacy HTML
- Simple display-only pages
- Unstructured data
- Write-once web pages

slide 81

XML versus HTML

- Short term
  - XML will be converted to HTML for display
  - HTML as output product of XML systems
- Longer term
  - Browsers will take XML directly, so who needs HTML?
  - HTML will cease to be mission-critical
  - Kids will still use HTML
  - Nobody will convert the legacy HTML

XML and PDF

slide 82

PDF

- Proprietary Adobe format
- Interchange/display of data *pages*
- Made from postscript or other page format
  (often designed for print, may not be easily read on screen)
- Used for
  - Document delivery
  - Pre-press
  - Online repositories
  - Long-term archiving

Aside: Generated Text (a PDF Issue)

slide 83

Generated Text

Text that is not in the data, but is put in by the display or formatting system, based on the tagging

For example:

- The numbers in a numbered list (1., 2., 3.)
- The bullets in a bulleted list
- The enumerator on a footnote
- Chapter 1.
- Figure 3.4
- (See Figure 3.6: All Cars Eat Gas)

slide 84

Authoring Systems *Should* Generate Text

- Consistency is assured
  - For each document
  - Over time for many documents
- Changes:
  - Automatically appear to all documents
  - Made once appear everywhere
- Author:
  - Can't make it wrong (or better!)
  - Has less work to do
  - Can use GUI while system fills in gory details (cross-references)

slide 85

Display Production Systems *Should Probably* Use Generated Text

- Consistency is assured
  - For each document
  - Across all documents in a collection
- Control in production not with individual authors
- Change one stylesheet not 1000 documents
- Files slightly smaller

(One downside: Display system has more work to do)

Should Archival Systems Use Generated Text?

- Is the stylesheet necessary to know what was actually produced?
- Can you search on “generated” words?
- If the text is not all there, can the document legally be copy-of-record?
- Is there a requirement to reproduce the document exactly as it was on a particular past date, thus requiring knowledge of all stylesheets and the system on that date?
- Is vital information hidden in the generated part?

TAGGED TEXT	**DISPLAY**
`<expn>aaaa</expn>`	`BAD EXAMPLE  - "aaaa"`
`<expg>bbbb</expg>`	`GOOD EXAMPLE - "bbb"`
`<exp time=b">ggg</exp>`	`BEFORE: ggg`
`<exp>fff</exp>`	`AFTER:  fff`

Static Archive Comparison

- PDF
  - There is one piece
  - All text is there
  - Look and feel is as author intended
- XML
  - There may be many parts as well as an external stylesheet
  - Some text may be generated, thus not there for search or raw comparison
  - Look and feel is
    - as stylesheet writer(s) intended
    - lost

slide 88

Distribute PDF When

- Legally, all generated text must be present
- Page layout, design, or location on the page is critical
- When reading is the object — not search, extraction, or reuse
- Source has no repeatable content or capturable structure (many Quark pages)
- Pages are more critical than information

Examples include:

- Read-only archives
- Certain legal documents (for example, materials cited by page and line number)
- Fast proof copies (galley equivalent)

slide 89

Distribute XML When

- More than one look and feel is necessary
- Security issues require complex processing
  - Digital signatures required
  - Documents constructed on-the-fly based on user's clearance
- Further processing/extraction/selection is necessary
- Fine granularity searching is required
- Information is more critical than pages

slide 90

XML *and* PDF

- PDF will still be a popular pre-press/archival form
- XML systems will produce PDF as an output product (as they do now)
- Searchable PDF will make in-roads against XML but be dropped as lesser functionality is recognized.
- It will take pretty pages and snazzy behavior to wean people from PDF, because PDF is
  - easy
  - inexpensive

XML and eBooks

"eBook" (aka e-Book) means many things to many people.

An eBook could be published:

- On the Web
- On CD-ROM
- For an eBook reader
- Any combination of these

The eBook format may be

- HTML
- PDF
- XML
- SGML
- Proprietary electronic publishing format

slide 92

eBooks in HTML (today's most common format)

Advantages

- Familiar
- Easy to create from many authoring and typesetting data formats
- Readable on most user's computers
- Easy to control formatting (within limits)
- Allows hyperlinking

Disadvantages

- Limited formatting options and control
- Variations in browser behavior
- Limited linking capabilities
- Severely limited search capabilities

slide 93

eBooks in PDF (Popular with electronic libraries)

Advantages

- Familiar to most users
- Viewers available for virtually all platforms
- Preserves look and detail of pages
- Allows (limited) linking
- Easy to create from any format that can be printed

Disadvantages

- Limited search capability
- Limited (and awkward) reuse by user
- Pages often designed for print, difficult to read on screen

slide 94

eBooks in XML

Limited, but growing (hot topics and OeB)

Advantages

- Supported by the Open eBook Specification
  More and more eBook readers will accept XML with CSS stylesheets
- High quality search and retrieval possible
- Reuse and re-purposing easy
- Rich linking and hyperlinking

Disadvantages

- Limited software to read
- Unfamiliar to publishers and users
- Requires investment in software and tagging

slide 95

Open eBook Specification

Operational/interchange format for eBooks

- There are two flavors
  - Basic (essentially HTML)
  - Extended (full XML)
- Does *not* include security
- Uses CSS for styling/format

slide 96

Basic Open eBook

- Same tags as HTML
- More rules
  - All end tags required
  - Attributes quoted (`name="value"`)
  - Empty tags (`<br/>` not `<br>`)
  - ... and similar rules

slide 97

Basic Open eBook Formatting

- Uses defined subset of CSS
- Stylesheets for HTML "built in"
- Additional styles may be provided (to change)
  (not all eBooks support this yet)

slide 98

Extended Open eBook

- Any XML tags (well-formed!)
- Must provide stylesheet in defined CSS subset
- Not all eBooks support this yet

XML From Your Content-Provider

slide 99

Industry-Focused XML Solutions

- Vertical market XML "languages"
  (Industry-specific and Cross-industry XML Specifications)
  - Banking and Financial Services
  - Health care
  - Research communities and Societies
  - Training services
- This work already well underway

slide 100

Industry-specific XML Specifications (1)

Banking and Financial Services

- Bank Internet Payment System (BIPS)
- Banking Industry Technology Secretariat (BITS)
- Digital Receipt Alliance: Annotated Digital Receipt
- Interactive Financial Exchange (IFX)
- Financial Services Technical Consortium (FSTC)
- Financial Information eXchange protocol (FIX): FIXML
- Financial Products Markup Language (FpML)
- Infinity Network Trade Model (NTM)
- Investment Research Markup Language (IRML)
- Market Data Markup Language (MDML)
- Open Financial Exchange (OFX) Specification
- Secure Warranted Internet Financial Transactions (swiftML)
- Straight Through Processing Markup Language (STPML)
- eXtensible Business Reporting Language (XBRL)

slide 101

Industry-specific XML Specifications (2)

Health Care

- Health Level Seven (HL7) Initiative
- Interchange98 (XML for Healthcare)
- Phase Forward: Clinical Trail Data Model
- The Open Healthcare Group

Insurance

- ACORD: Property and Casualty Life (XMLife)
- Lexica: iLingo

Telecommunications

- Alliance for Telecommunications Industry Solutions (ATIS)
- Telecommunications Interchange Markup (TIM)
- Wireless Markup Language (WML)
- Wireless Application Protocol Forum (WAP)

slide 102

Industry-specific XML Specifications (3)

Manufacturing

- Automotive Industry Action Group (AIAG)
- Global Automedia: VehicleExport
- MSR: Standards for information exchange in the engineering process (MEDOC)
- Society of Automotive Engineers (SAE): XML for the Automotive Industry (SAE J2008)
- Machinery Information Management Open SYStems Alliance (MIMOSA)
- Manufacturing Data Systems Inc. (MDSI)
- Virtual Instruments Meta Language (VIML)
- Electric Component Information Exchange (ECIX)
- Product Data Markup Language (PDML)

slide 103

Industry-specific XML Specifications (4)

Retail/Distribution

- Association of Retail Technology Standards (ARTS XML)
- BikeXML
- Customer Identity Markup Language (CIML)
- Customer Profile Exchange Network (CPEX)
- FormatData: Document Encoding and Structuring Specification for Electronic Recipe Transfer (DESSERT)
- First Retail Markup Language (FRML)
- Name and Address Markup Language (NAML)
- Retail Enterprise Data in XML (REDX)

slide 104

If Content Provider Has XML Content

- They will already have DTDs or schemas (not necessarily appropriate for publication)
- They will expect you to
  - Accept XML source (minimally)
  - Work in XML (better)
  - Work in *their* XML
- They will expect your compositor/typesettter to
  - Work in XML
  - Provide round-trip services

XML Changes to Workflow, Staffing, and Skills

XML Production Issues

slide 105

The Issues

- Who (in-house versus out)
- When (start of process and all through, versus post-production)
- What it takes (training, skills)
- What it changes

slide 106

Who Tags Your Data

- Authors (in-house or external)
- Copy editors
- Production staff
  - Possibly special “taggers”
  - Production Editors or other Production staff
  - Compositor's production people
- Conversion vendor
- One or more computer programs

slide 107

When is the Data Tagged

- During creation/authoring
- Between author and any edit
- As part of editing
- As part of production
- Post Production
  - As extension to process (for example by compositor before return to publisher)
  - When placed into repository
  - Long after production, possibly by years

slide 108

The Best Place in the Cycle to Use XML

Greatest benefits mean tagging as early as possible in the production cycle

- Pre-repository brings searching benefits
- Pre-production brings electronic reuse and control (XML inside content management repository)
- Pre-editing allows hyperlink editing/checking, snazzy electronic edit features, electronic communication with authors, etc.

If You Take Tagging Responsibility

slide 109

What Tagging Means

- Staff training
- Acquire DTDs or schemas
- Make XML a part of the production process
- Revise procedures and workflow
- More training
- Potential backfile conversion

slide 110

Staff Knowledge Level

- For all, basic XML concepts
- For structural tagging, good language skills (in the appropriate language)
- For bibliographic tagging, reference knowledge (how to tell a journal from a book from a thesis)
- For subject tagging, subject matter expertise
  - Medical
  - Pharmaceutical
  - Legal
  - Chemical

Staff Training

slide 111

Who Needs XML User Training

- Authors
- Content editors
- Copy editors
- Production technicians
- Data entry personnel
- Electronic publishing production staff

slide 112

What is XML User Training

- Principles of generic markup and document modeling
- Tagging basics
  - Format-specific vs. generic
  - Structure vs. content
- Logical structure of their documents
- Use of XML software tools
- Parsing and error resolution

slide 113

Special Case: Math and Chemistry

- Tagging is complex
- Subject expertise is critical
- If authors do not tag,
  extra proofing step is required

slide 114

XML Systems/Transformation Training

For the in-house XML "gurus"

- DTDs/Schemas
  - Reading and writing
  - Internal and external documentation
  - Maintenance strategies
- XML software and systems
- Transformation tools and languages

slide 115

XML Management Training

For Anyone approving budgets, buying tools, developing production schedules or goals, or managing the process

- Basic concepts and vocabulary
- Costs and scheduling
- Risk and opportunities
- Production concerns and personnel factors

DTDs and Schema

slide 116

DTDs and Schemas Required for Content Creation

- Use industry standard
- Use client's
- Use supplier's
- Use ultimate consumer's
- Roll your own (with consultant help)

Be prepared to use many, and transform more

Document Production Changes

slide 117

XML Will Change the Way You Work

- Who does what may alter
- Line between authoring, content edit, and copy edit blurs
- Proofing and checking changes
  - Level changes (less worry about transposed letter and more about missing structures)
  - Specifics change (don't look for comma, look for `<author>`)
  - New possibilities like false color proofs
  - One error may show many times but still be just *one* error

slide 118

Warning: This is a "Paradigm Shift"

- Working in data not working in appearance
- Control of format by controlling tags and attributes (each one *cannot* be different)
- WYSIROTS not WYSIWYG
- Forms and outlines useful

slide 119

Procedures and Workflow

- Job responsibilities may change
- Structured writing may be a new way to think
- Machine do some work once done by people
- People do some work never done before

slide 120

Changes in Staffing/Jobs

- Much rekeying eliminated
- Proofing/rekeying cycles shortened or eliminated (exception *math*! and *chemistry*)
- Changes the nature of the grunt work
- New types of proofing possible
- More time devoted to content instead of format
- Transformation becomes a major activity
  - More post processors
  - More potential output products

slide 121

Warning: XML Does *Not* Reduce Staff

- XML can increase
  - Accuracy
  - Opportunities in sales and new products
  - Customer satisfaction
  - Timeliness (reduce time to market)
- But not likely to be done with fewer people

Common Problems

slide 122

Scale

- XML is relatively simple, one document at a time
  - Proof-of-concepts trivial
  - Demonstration of success encouraging
- Challenge is Scaling for a Large Organization
- Large scale has benefits
  - Reuse content across publications
  - Share tools and customization
- Scaling works if:
  - Coordinate
  - Plan
  - Compromise

slide 123

Added Value Means Added Work

- Added metadata— Repository storage and rights/permissions may need metadata that does not print or display
- Semantic tagging— Value-added tags for searching may require subject knowledge
- Index terms — Adding inline indexing is the same huge effort as a back-of-the-book index
- Writing changes — If structures are to be used in parallel ways; they must be parallel

slide 124

Costs and Benefits Not Equitable

(Warning: Costs in budget of Department A, benefits accrue to Departments C and D or to the company as a whole)

- Some groups will have increased workload, other groups will find production faster and easier
- Expense and hard work are immediate, specific to a group
- Most benefits
  - Increase the bottom line for a different group
  - Are long term
  - Are corporate-wide or sales related
  - Are cumulative (your job takes 2 months longer, but our jobs are: done in half the time, much cheaper, or only possible now)

Conclusion

slide 125

The Bad News: There is no Free Lunch

- Just because it's XML doesn't mean it's good
- Richer tagging means more work
- The XML marketplace is young, growing fast, and unstable
- The benefits of XML are available only to those who
  - Plan
  - Play by the rules
  - Work hard

slide 126

The Good News: You Can Do XML and Benefit

- Many good tools are now available
- Service vendors rapidly becoming XML savvy
- The SGML some of you have is a jump-start to XML
- XML browsers are just around the corner
  - IE5 includes (weak) XML support
  - Netscape 6 promises XML support this month
  - Mozilla almost ready
- Users of electronic products more sophisticated every day, demanding better electronic products

Where to Get More Information

slide 127

The Source for XML and Related Information

- Robin Cover's SGML/XML Web Page: **`http://www.coverpages.org`**

slide 128

General XML Information

- W3C's XML page:**`http://www.w3.org/XML/`**
- XML FAQ (Peter Flynn): **`http://www.ucc.ie/xml/`**
- XML.com: **`http://www.xml.com`** (industry coverage and tools)
- XML.org: **`http://www.xml.org`** (industry coverage and tools)
- XMLinfo.org **`http://www.xmlinfo.org`** (XML tools and development)
- XSLinfo.org: **`http://www.xslinfo.org`** (XSL development and implementation issues)

slide 129

XML E-Business and EDI Information

- The OASIS and UN/CEFACT entry **`http://www.ebxml.org`**
- CommerceOne's entry**`http://www.commerceone.com`** (Common Business Library etc.)
- The Microsoft entry **`http://www.BizTalk.org`**
- RosettaNet consortium **`http://www.rosettanet.org`** (vocabulary and process issues, industry coverage)
- XML.ORG **`http://www.xml.org`**
- Commerce XML (cXML) **`http://www.cxml.org`**
- Oracle's XML material **`http://www.oracle.com/xml/content.html`**
- IBM is heavily into this, too. **`http://www.developer.ibm.com`**[for Business Rules Markup Language (BRML) and Trading Partner Agreement Markup Language (tpaML)

And others too numerous to mention: CommerceNet's eCoFramework, SAIC's Universal Commerce Language and Protocol (ULCP), XEDI.org, etc.

Printed Books on Concepts

- **SGML: the Billion-Dollar Secret**, by Chet Ensign (Prentice-Hall PTR, 1997)
  - Manager level. Written about SGML (XML's parent standard), but almost entirely applicable: excellent on issues of scalable system development.
- **ABCD… SGML**, by Liora Alschuler (Thompson Computer Press, 1995)
  - Written about SGML (XML's parent standard), but change the word "SGML" to "XML" as you read it and it still applies. Talks about work process changes an XML system can bring.
- **XML: A Manager's Guide**, by Kevin Dick (Addison-Wesley Information Technology Series, 2000)
  - Manager level. Solid view, but stays at 10,000 feet up.
- **The XML Companion (2nd Edition)**, by Neil Bradley (Addison-Wesley, 2000)
  - Very good basic technical introduction.
- **Professional XML**, by Richard Anderson, Mark Birbeck and ten more authors. (Wrox Press Ltd.)
  - Light technical level. Each author wrote an introduction and then examples/case study for one technical topic. Introduces the problems of XML and databases, the XML APIs DOM and SAX, server to server XML (XML-RPC, SOAP, etc.) and more.

slide 131

Other Information Sources

- **Markup Languages: Theory and Practice** (a quarterly journal): **`http://mitpress.mit.edu/MLANG`**
- OASIS Home Page (vendor consortium): **`http://www.oasis-open.org`**
  - XML.ORG: **`http://xml.org`** (document model repository and support materials)
  - OASIS XML Conformance Subcommittee: **`http://www.oasis-open.org/committees/xmlconf-pub.html`**
- Graphic Communications Association: **`http://www.gca.org`** (sponsors conferences including XML Europe,Extreme Markup Languages, XML 2001)
- XML.COM: **`http://www.xml.com`**

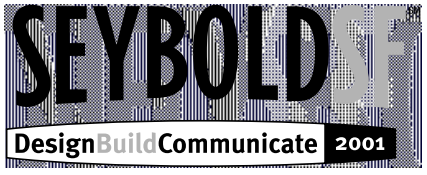
slide 132

Still More Information Sources

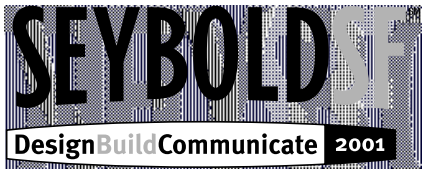
- Basic newsgroup: `comp.text.xml` (also some on`comp.text.sgml`)
- Useful Lists
  - XML-L: **`http:/listserv.heaanet.ie/xml-l.html`** (for newcomers)
  - XML-Developer's List: **`http://www.lists.ic.ac.uk/hypermail/xml-dev`** (heavy technical discussion)
  - XSL-List: **`http://www.mulberrytech.com`**

Appendix 1: Acronym List

AIAG	Automotive Industry Action Group
API	Application Program Interface
ARTS XML	Association of Retail Technology Standards
ATIS	Alliance for Telecommunications Industry Solutions
B2B	Business-to-Business
B2B	Business-to-Customer
BITS	Banking Industry Technology Secretariat
BRML	Business Rules Markup Language
CIML	Customer Identity Markup Language
CML	Chemistry Markup Langauge
CPEX	Customer Profile Exchange Network
CSS	Cascading Style Sheets
cXML	Commerce XML
DESSERT	FormatData: Document Encoding and Structuring Specification for Electronic Recipe Transfer
DOM	Document Object Model
DSSSL	Document Style and Semantics Specification Language
DTD	Document Type Definition
eBook	Electronic Book
ebXML	Electronic Business XML
ECIX	Electric Component Information Exchange
EDI	Electronic Data Interchange
FIX	Financial Information eXchange protocol
FIXML	Financial Information eXchange protocol Markup Language
FpML	Financial Products Markup Language
FRML	First Retail Markup Language
FSTC	Financial Services Technical Consortium
GUI	Graphical User Interface
HL7	Health Level Seven Initiative
HTML	Hypertext Matkup Language
IFX	Interactive Financial Exchange
IPC	Inter-process communication



IRML	Investment Research Markup Language
JSP	Java Server Pages
MathML	Mathematics Markup Langauge
MDML	Market Data Markup Language
MDSI	Manufacturing Data Systems Inc
MEDOC	MSR: Standards for information exchange in the engineering process
MIMOSA	Machinery Information Management Open SYStems Alliance
NAML	Name and Address Markup Language
NTM	Infinity Network Trade Model
OeB	Open eBook Specification
OFX	Open Financial Exchange Specification
OO	Object Oriented (applied to database)
PDA	Personal Digital Assistant
PDF	Portable Document Format
PDML	Product Data Markup Language
POSC	Petrochemical Open Software Corporation
PPML	Personalized Print Markup Langauge
PRISM	Publishing Requirements for Industry Standard Metadata
REDX	Retail Enterprise Data in XML
RPC	Remote Procedure Call (e.g., XML-RPC, SOAP, etc.)
RTF	Rich Text Format
SAE	Society of Automotive Engineers
Sax	Simple Application Profile for XML
SGML	Standard Generalized Markup Langaage
SOAP	Simple Object Access Protocol (Microsoft)
SQL	SEQUEL Query Language
STPML	Straight Through Processing Markup Language
SVG	Scalable Vector Graphics
swiftML	Secure Warranted Internet Financial Transactions
TIM	Telecommunications Interchange Markup
tpaML	Trading Partner Agreement Markup Language
ULCP	SAIC's Universal Commerce Language and Protocol
UML	Universal (Uniform) Modeling Language
URL	Uniform (Universal) Resource Indicator
VIML	Virtual Instruments Meta Language
W3C	World Wide Web Consortium
WAP	Wireless Application Protocol Forum



WML	Wireless Markup Language
WWW	World Wide Web
XBRL	eXtensible Business Reporting Language
XML	Extensible Markup Language
XSL	Extensible Stylesheet Language
XSLFO	XSL Formatting Objects
XSLT	XSL Transformations