Introduction to XML

Mark Eichin, SIPB

IAP 1999

A "new" way to think about text

- structure not formatting
- portable representation
- enhanced content
- Details
- Markup in plain ASCII
- ISO 8879:1986
- 332 definitions; 70 basic
- history: GML
- DTD, Entity, Element

History

- Specific Coding (format-17)
- late 1960’s: Generic Coding (header)
- 1969: GML (IBM, Goldfarb, Mosher, Lorie) - law offices
- Invented 1974 by Charles F. Goldfarb
- 1978: ANSI committee
- 1983: GCA standard
- 1985: final ISO 8879 draft
- 1987: DoD CALS
- 1993: ISO/IEC 10646 (instantiation of Unicode)
- 1994: RFC 1738: Uniform Resource Locators.
- 1996: XML Working Group formed

Acronyms

- **SGML - Standard Generalized Markup Language**
- **DSSSL - Document Style Semantics Specification Lang.**
- **DTD - Document Type Description**
- **CALS - Computer-aided Acquisition + Logistic Support**
- **ISO - is *not* an acronym**
- **XML - eXtensible Markup Language**
- **XSL - eXtensible Stylesheet Language**
- **CSS - Cascading Style Sheets**
- **OASIS - Organization for the Advancement of Structured Information Standards**
- **ECMA - European Computer Manufacturers Assoc.**

The design goals for XML are...

- **... shall be straightforwardly usable over the Internet.**
- **... shall support a wide variety of applications.**
- **... shall be compatible with SGML.**
- **It shall be easy to write programs which process [docs]**
- **The number of optional features in XML is to be kept to the absolute minimum, ideally zero.**
- **[docs] should be human-legible and reasonably clear.**
- **The XML design should be prepared quickly.**
- **The design of XML shall be formal and concise.**
- **XML documents shall be easy to create.**
- **Terseness in XML markup is of minimal importance.**

What you can do with it

- smart editing
- automatic generation
- databases
- typed searching
- output processing
- easy parsing

<tag> </tag>
<tag attr=value>
< & >
<emptytag/>
&

xml-lecture.sgml

```
<?XML version="1.0"?>
<!DOCTYPE slideshow SYSTEM "xslides.dtd">
<slideshow>
  <title>Introduction to XML</title>
  <author>Mark Eichin, SIPB</author>
  <date>IAP 1999</date>
  <slide>
  <header>A "new" way to think about text</header>
  <bullet>structure not content</bullet>
  </slide>
```

slides.dtd

```
<!ENTITY lt     "&#60;">
<!ENTITY gt     "&#62;">
<!ENTITY amp    "&#38;">
<!ELEMENT slideshow (title?, author?, date?, slide*)>
<!ELEMENT (header|title|author|date) (#PCDATA)>
<!ELEMENT slide (header, (bullet|subbullet|url|raw|example)+)>
<!ELEMENT bullet (#PCDATA)>
<!ELEMENT subbullet (#PCDATA)>
<!ELEMENT url (#PCDATA)>
<!ELEMENT raw (#PCDATA)>
<!ELEMENT example (#PCDATA)>
```

XML::Parser

- perl interface to add code to tags

```
sub handle_start {
  my ($expat, $elem) = @_;
  ... } elsif ($elem =~ /^slide$/) {
  &inittag;
  print "%%Page: $pageno\n";
  print "($savetitle) showtitle\n";
  print "($pageno) showpageno\n"; $pageno++; } ...
}
sub handle_end {
  my ($expat, $elem) = @_;
  ... } elsif ($elem =~ /^slide$/) {
 print "finishpage\n";
} ...
$parser = new XML::Parser(Handlers => {
 Start => \&handle_start,
 End => \&handle_end });
```

HTML

- Simple DTD - 70 elements
- historically, no invalid pages
- comment syntax

<-- comment -- -- another one -->

- Reformulating HTML in XML

`http://www.w3.org/TR/1998/WD-html-in-xml-19981205`

Implementations

- SAX - Simple API for XML
- DOM - Document Object Model
- ESIS - Element Structure Information Set
- SX - SGML to XML normalizer
- IBM Alphaworks XML - Java
- XP - James Clark, Java
- LT XML (Edinburgh) - C
- XMLTok (was Expat) - C
- XMLParser class, xmlproc (Python)
- XML::Parser (O'Reilly, ActiveState)

More Implementations

- TclXML, TclExpat, ExCost
- MSXML - Java, C++
- sgrep - structured search
- Microsoft Internet Explorer 5beta
- AXSL - Activated XSL, Java Lobby site

Applications

- XBEL - XML Bookmark Exchange Language (Python)
- Microsoft CDF (Channel Definition Format)
- MicroSoft XML Notepad
- Web Spiders/Robots
- Digital Receipt Consortium DTD
- Mediacenter News Markup Language
- Oracle PLSXML - PL/SQL-based XML utilities
- DrawML
- XQL - XML Query Language
- MARC - MAchine Readable Cataloging

Docbook

`http://www.oasis-open.org/docbook/`

- Davenport Group
- OASIS DocBook Technical Committee
- Software Documentation
- Very rich (400+ tags)
- Industry Consortium

`http://nwalsh.com/docbook/xml/index.html`

- Conversion issues - broader valid set

Style Sheets

- DSSSL
  - R4RS Scheme
  - Document Style Semantics Specification Language
  - ISO/IEC 10179:1996
- XSL
  - ECMAscript (JavaScript) ECMA-262 extension
- Implementations
  - XSLJ - convert XSL to DSSSL
  - XT - James Clark Tree Constructor (Java)

Books

- **XML: The Annotated Specification**

  `http://www.snee.com/bob/xmlann/`
- **The XML Handbook(TM)**
- **XML by Example: Building E-Commerce Applications**
- **Structuring XML Documents**
- **Designing XML Internet Applications**
- **XML: A Primer**
- **XML for Dummies**

  `http://www.xmlbooks.com/`
- **DocBook: The Definitive Guide**

  `http://www.nwalsh.com/docbook/defguide/index.html`

Other Stuff

- XLL XLink/XPointer
- Namespaces
- VRML moving to XML
- Languages
  - Chinese
  - Japanese
  - French

Pointers

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

http://www.mit.edu/iap/iap-sgml.html

ftp://ftp.lysator.liu.se/pub/sgml/psgml-1.0.3.tar.gz

http://www.jclark.com/

http://www.cs.helsinki.fi/~jjaakkol/sgrep.html

- **comp.text.xml**
- **comp.text.sgml**
- **add sgml (jade, nsgmls, stylesheets, psgml)**

http://www.mit.edu/iap/xml/