Creative Commons

Attribution-NonCommercial-ShareAlike 2.0

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:

**BY:** Attribution. You must give the original author credit.

**Noncommercial:** You may not use this work for commercial purposes.

**Share Alike:** If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

This page is also available in the following languages:

[Links to different language versions]

Learn how to distribute your work using this license
Overview
- Standard Generalised Markup Language (SGML)
- Hypertext Markup Language (HTML)
- Extensible Markup Language (XML)

What HTML/SGML/XML have in common
- they are markup languages (as opposed to programming or processing languages)
- they are metalanguages: languages which describe other languages
- all use tags or elements -- special software interprets those tags either for display purposes and/or for search and retrieval
- In the case of HTML encoding is usually used to indicate format --- a browser (Netscape, Internet Explorer) interprets the marked up text:
  `<bold>My Lecture</bold>`
- In the case of SGML or XML, the markup indicates the function of the text:
  `<title>My Lecture</title>`
- markup languages use another language &/or software to render the content for display (CSS/XSL, DynaWeb)
- all use attributes to further delineate specific features of text `<title type="main">My title</title>`
Standard Generalised Markup Language

- the papa language from which HTML & XML are derived
- became an ISO standard in 1986
- developed as a platform & software independent tool to deal with large amounts of text
- some major users are aeronautics, military, text encoding, pharmaceuticals

SGML (Standard Generalised Markup Language)

- HTML
- Pharmaceuticals
- Aeronautics
- Military
- Text encoding

Pharmaceutical documentation written in PharmML

```xml
<NewDrug>
  <name>BrainBooster</name>
  <SideEffects>
    <Effect>
      Makes you mega-intelligent
    </Effect>
    <Effect>
      Turns your hair purple
    </Effect>
  </SideEffects>
</NewDrug>
```

TEI (Text Encoding Initiative)

```xml
<DIV TYPE="poem">
  <HEAD>Straw in the Street.</HEAD>
  <LG TYPE="stanza">
    <L><HI>STRAW</HI> in the street where I pass to\&hyphen;day</L>
    <L>Dulls the sound of the wheels and feet.</L>
    <L>\&rsquo;Tis for a failing life they lay</L>
    <L>in the street.</L>
  </LG>
</DIV>
```

Hypertext Markup Language

- developed by Tim Berners-Lee working for Cern in Switzerland (ISO standard 1991) out of a desire to disseminate scholarly articles amongst colleagues in physics rather than share them via an e-mail type facility
out of SGML developed a simple, relatively small set of 'tags' for marking up the 'physical' features of articles i.e. **bold**, *italic*, underline, green
• how & in what order those tags can be used is determined by a HTML DTD (Document Type Definition)

**Why HTML was a good web start, but a bad web future**
• lack of functionality
• lack of logical markup
• major browsers wanting more rigorous encoding standards
• bad for e-commerce
• too many other languages (javascript, cgi, etc) needed to get things to work

---

**XML**

“.... An extremely simple dialect of SGML... The goal is to enable generic SGML to be served, received and processed on the Web in the way that is now possible with HTML”

**XML**
• became an ISO standard in 1998
• "a simple, very flexible text format derived from SGML (ISO 8879). Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web." [http://www.w3.org/XML/Activity.html](http://www.w3.org/XML/Activity.html)

**XML**
• a simplified SGML rather than a beefed up HTML
• features removed from SGML allows it to be delivered over the web
• a suite or family of languages
• a fledgling technology – many standards are still not in place

**Family of XML Languages**
• XML
• XLink
• XPointer
• XSL
  • XSLT
  • XSL FO
• XML Schema
• [DTDs]
Like SGML . . .

- XML allows users (or communities of users) to create their own tag sets
- uses a stylesheet to display XML encoding
- capability of encoding both logical and physical features of text

Beyond SGML

- a family of technologies
- reusability: one document many publication applications in a variety of media
  - computers
  - mobile phones
  - palm pilots

Features of XML

- Facilitates moving of data from one location to another while ensuring the structure is maintained as content is passed from resource to resource
- separates content from display so that it can be delivered to a variety of devices
- Software independent
- Ability for users or communities of users to develop their own structure of information

Already used to create a variety of standards

- Microsoft Channels (CDF)
- Chemical Markup Language (CML)
- Vector (Graphics) Markup Language (VGML)
- Virtual Reality Markup Language (VRML)
- Synchronized Multimedia Integration Language (SMIL)
The XML Pieces

The Various XML Technologies

- XML Content (.xml)
- XML Rules (.dtd)
  - Schemas
  - DTDs
  - Namespaces (used when you want to combine sets of rules together in a single document)
- Entities (.ent)
  - Reusable data inside a DTD or within markup
- Display (.css & .xsl)
  - Extensible Style Sheet Language
  - Cascading Style Sheets

XML Pieces

- Extensible Style Sheet Language (.xsl)
  - Used for transforming data to another structure
  - Used for Formatting Objects
- Xpath (Technologies used in files)
  - Like &lt;A NAME="XXX"> allows or addressing parts of an XML document
- XLink & Xpointer (Technologies used in files)
  - Like the &lt;A&gt; element in HTML, allows for ways to link in XML

Overview

- HTML, SGML, XML
- DTDs & Schemas

DTDs

- a set of rules indicating which elements can be used where & how many times they can be used
- also indicates how attributes can be used
- uses its own syntax rather than XML syntax

A simple DTD for articles in XML

```xml
<!-- this is an article dtd-->
&lt;ENTITY ss "Susan Schreibman”>
&lt;ELEMENT article (title, author+, pn*, p+, linegroup*)>
&lt;ELEMENT title (#PCDATA )>
&lt;ELEMENT author (#PCDATA | bionote )>*
&lt;ATTLIST author id ID #IMPLIED
  person IDREF #IMPLIED &gt;
&lt;ELEMENT bionote (#PCDATA )>
&lt;ELEMENT p (#PCDATA | note | author )
&lt;ATTLIST p id ID #IMPLIED
  n CDATA #IMPLIED &gt;
&lt;ELEMENT pn EMPTY>
&lt;ATTLIST pn n CDATA #IMPLIED &gt;
&lt;ELEMENT note (#PCDATA )
&lt;ATTLIST note location (foot | end | inline )’foot’
```
**DTDs**

- Can be thought of as an abstraction of document structure
  - What tags and attributes must/can be used
  - How these tags and attributes are structured in relation to each other

**Part of the DTD for PharmML**

```xml
<!Element NewDrug
    (name, SideEffects)>
<!Element SideEffects (Effect)+>
```

**A tiny bit of the TEI DTD in SGML**

```xml
<!ELEMENT name - - (#PCDATA | abbr | address | date | dateRange | expan |
measure | name | num | rs | time |
timeRange | add | addSpan | app | corr |
damage | del | delSpan | gap | orig | reg |
restore | sic | space | supplied |
unclear | distinct | emph | foreign |
gloss | hi | mentioned | soCalled | term |
title | link | ptr | ref | xptr | xref |
anchor | c | cl | m | phr | s | seg | w |
formula | fw | handShift)>```

**XML Schema**

- A way to create rules using XML syntax
- Not backward compatible with DTDs
- Many schema formats
- Allows datatyping
- Allows users to combine schemas (namespaces)