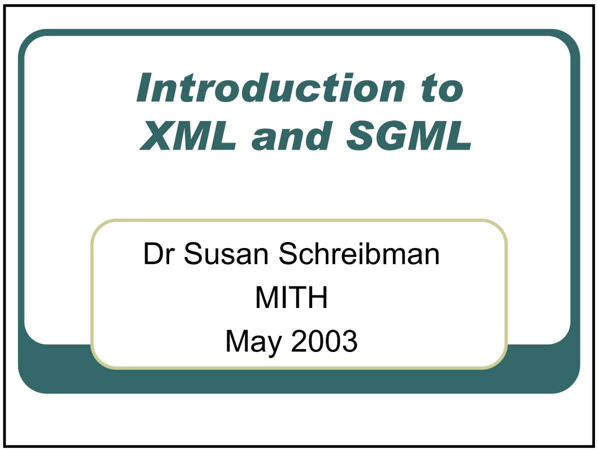
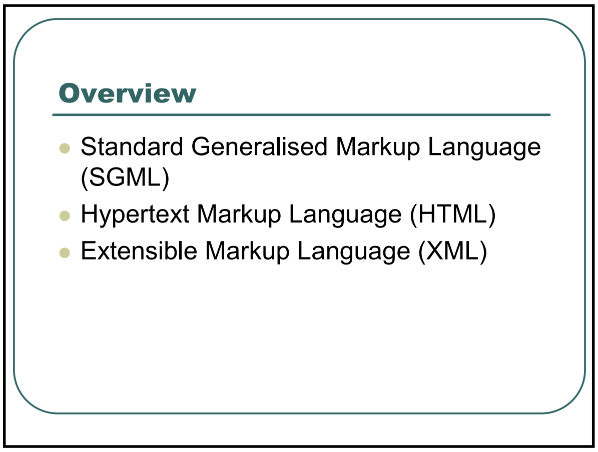


Introduction to XML and SGML

Dr Susan Schreibman

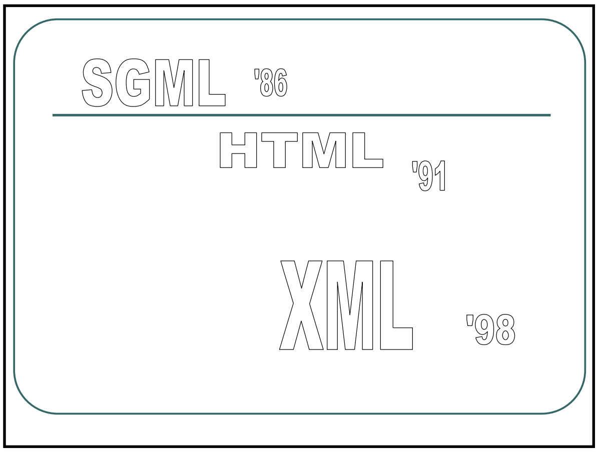
MITH

May 2003



Overview

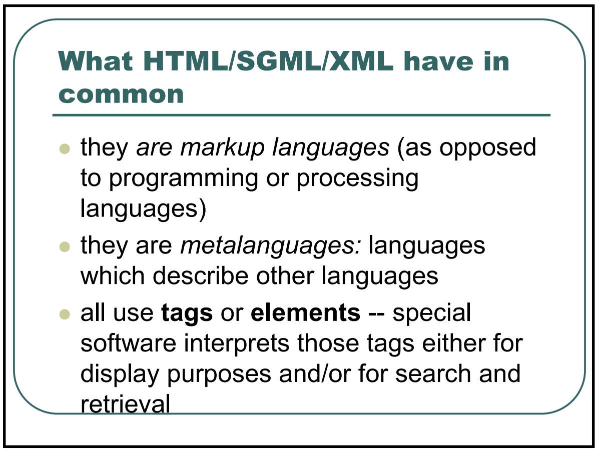
- Standard Generalised Markup Language (SGML)
- Hypertext Markup Language (HTML)
- Extensible Markup Language (XML)



SGML '86

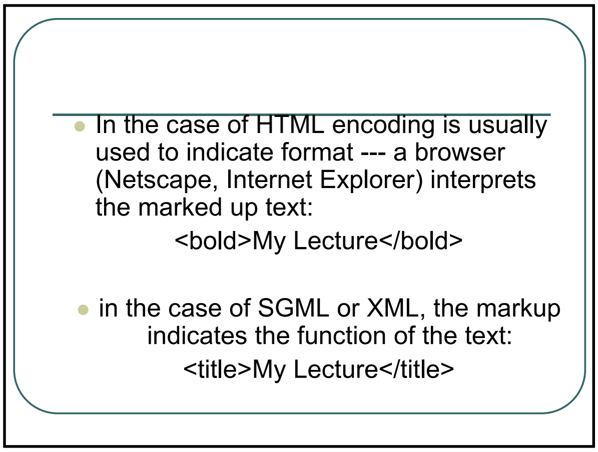
HTML '91

XML '98



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

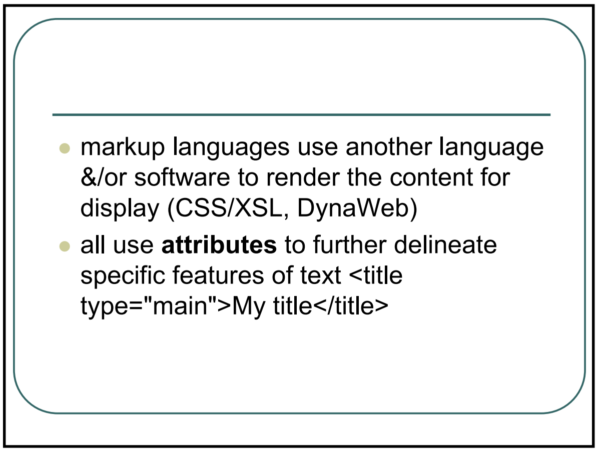


- In the case of HTML encoding is usually used to indicate format --- a browser (Netscape, Internet Explorer) interprets the marked up text:

  <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

- it's
  - huge — potentially comprised of millions of tags
    - allows for users to define and develop their own tag sets
    - extremely difficult to work with syntactically
  - developed in a pre-internet environment so many features difficult to implement via a distributed network
  - yet very powerful in its descriptive capabilities

SGML

Standard Generalised Markup Language

- HTML
- Pharmaceuticals
- Aeronautics
- Military
- Text encoding

Pharmaceutical documentation written in PharmML

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

TEI
(Text Encoding Initiative)

```
<DIV0 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 REND="indent1">Straw in the street.</L></LG>
</DIV0>
```

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

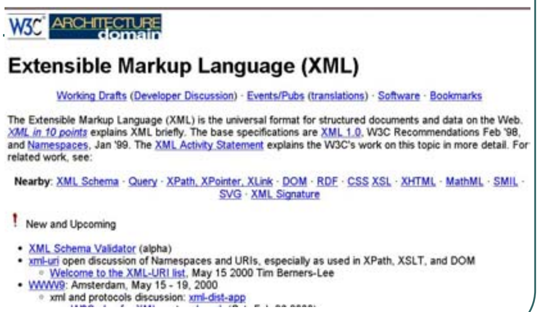
XML

- 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]

http://www.w3.org/XML/



Extensible Markup Language (XML)

Working Drafts (Developer Discussion) · Events/Pubs (translations) · Software · Bookmarks

The Extensible Markup Language (XML) is the universal format for structured documents and data on the Web. *XML in 10 points* explains XML briefly. The base specifications are XML 1.0, W3C Recommendations Feb '98, and Namespaces, Jan '99. The *XML Activity Statement* explains the W3C's work on this topic in more detail. For related work, see:

Nearby: XML Schema · Query · XPath, XPointer, XLink · DOM · RDF · CSS XSL · XHTML · MathML · SMIL · SVG · XML Signature

! New and Upcoming

- XML Schema Validator (alpha)
- xml-uri open discussion of Namespaces and URIs, especially as used in XPath, XSLT, and DOM
  - Welcome to the XML-URI list, May 15 2000 Tim Berners-Lee
- WWW9: Amsterdam, May 15 - 19, 2000
  - xml and protocols discussion: xml-dist-app

Like SGML . . .

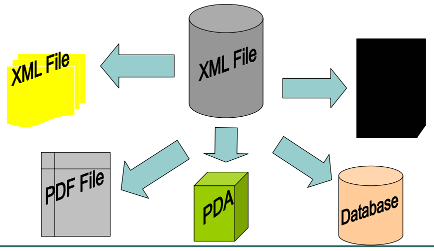
- XML allows users (or communities of users) to create their own tag sets
- uses a stylesheet to display XML encoding
- capability of encoding *both* logical and physical features of text

beyond SGML

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

With XML you can…

Have one XML file that serves up many purposes:



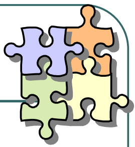
Features of XML

- **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

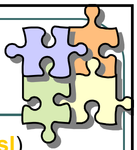
- Microsoft Channels (**CDF**)
- Chemical Markup Language (**CML**)
- Vector (Graphics) Markup Language (**VGML**)
- Virtual Reality Markup Language (**VRML**)
- Synchronized Multimedia Integration Language (**SMIL**)

The XML Pieces
The Various XML Technologies



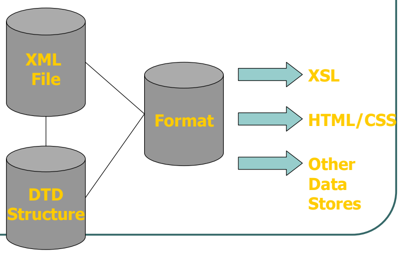
- **XML Content** (.xml)
- **XML Rules** (.dtd)
  - Schemas
  - DTDs
  - Namespaces **(used when you want to combine sets of rules together in a single document)**
- **Entities** (.ent)
  - Reusable data inside a DTD or within markup
- **Display** (.css & .xsl)
  - eXtensible Style Sheet Language
  -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

XML Pieces



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

XML Publishing Process



Overview

- HTML, SGML, XML
- DTDs & Schemas

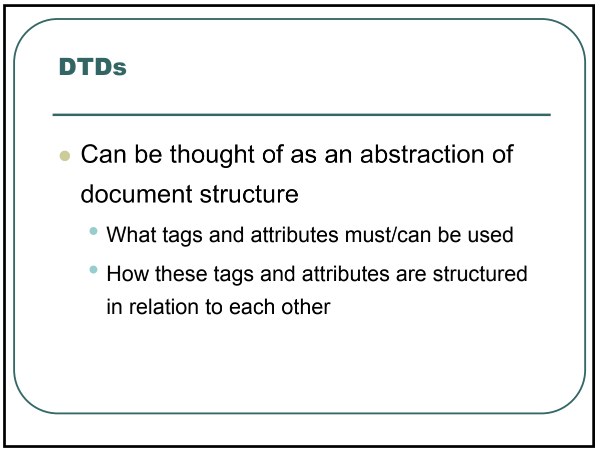
DTDs

- 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

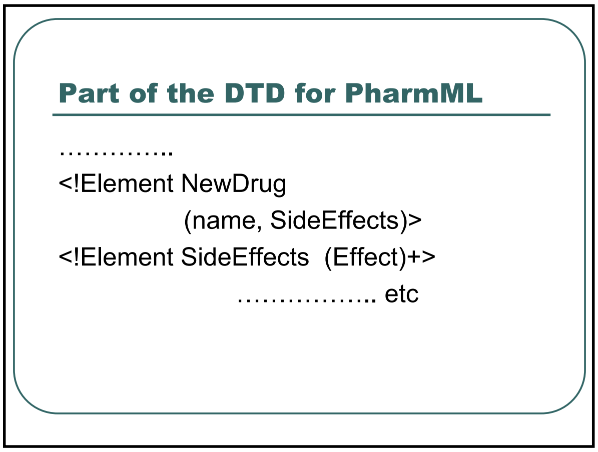
```
<!-- this is an article dtd-->
<!ENTITY ss "Susan Schreibman">

<!ELEMENT article  (title , author+ , pn* , p+ , linegroup* )>
<!ELEMENT title  (#PCDATA )>
<!ELEMENT author  (#PCDATA | bionote )*>
<!ATTLIST author  id     ID      #IMPLIED
                        person IDREF  #IMPLIED >
<!ELEMENT bionote  (#PCDATA )>
<!ELEMENT p  (#PCDATA | note | author )*>
<!ATTLIST p  id ID     #IMPLIED
           n  CDATA  #IMPLIED >
<!ELEMENT pn EMPTY>
<!ATTLIST pn  n CDATA  #IMPLIED >
<!ELEMENT note  (#PCDATA )>
<!ATTLIST note  location  (foot | end | inline )  'foot' >
```



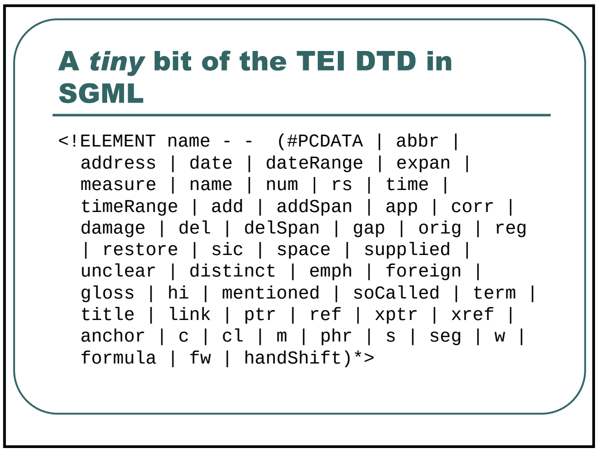
DTDs

- 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



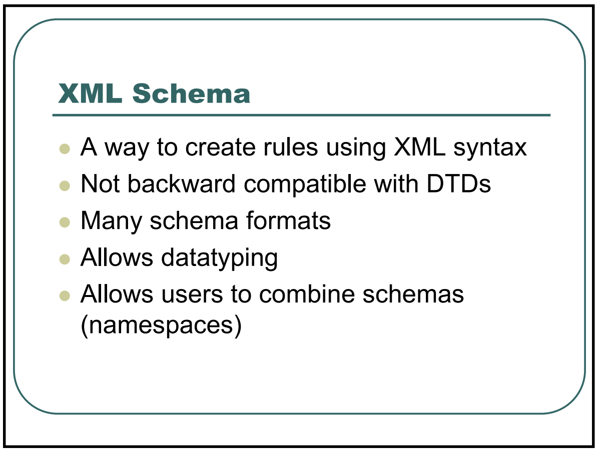
Part of the DTD for PharmML

```
…………..
<!Element NewDrug
              (name, SideEffects)>
<!Element SideEffects  (Effect)+>
                    …………….. etc
```



A *tiny* bit of the TEI DTD in SGML

```
<!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)