

XML

April 12, 2002

Michael Minock

History of XML

- Standard Generalized Markup Language (SGML) - 1986
 - Flexible, Complex, ...
- Hyper Text Mark-up Language (HTML) - Early 90's.
 - SGML compliant - fixed set of tags.
- eXtensible Markup Language (XML) - Berner's Lee and W3C - 1996.
 - Extensibility
 - Structure
 - Validity

Six years of hype - and where is it!

Getting there...

2

XML

A standard that is a tighter specification than SGML

But is more general than HTML

And is able to represent document, semi-structured, and structured data.

It qualifies as a *data-model* – albeit a simple one.

1

What really is XML?

XML gives is the capability to develop Document Type Definitions (DTDs) that specify constraints on recursively embedded expressions.

For our purposes a DTD is a type of *schema*. The constraints in the DTD provide a "tree-like" data model.

A set of 'legal' expressions saved in a file is a valid XML document of the DTD type.

For our purposes an XML document contains a set of *facts* of the corresponding DTD type.

3

DTD Constructors

ELEMENTS must be blocked off by START and END tags.

ATTRIBUTES add name/value pairs to ELEMENTS.

```
<!ELEMENT element-name EMPTY>
<!ELEMENT element-name (#PCDATA)>
<!ELEMENT element-name content-particle>

<!ATTLIST element-name
  attribute-name1 attribute-type1 attribute-default1
  ...
  attribute-namen attribute-typen attribute-defaultn>
```

(0-1) is indicated with a '?'.
(0-X) is indicated with a '*'.
(1-X) is indicated with a '+'.
(-X) is indicated with a '#'.
(X) is indicated with a '1'.

#REQUIRED. The attribute must have a value.

#IMPLIED. The attribute is optional.

#FIXED. The attribute has a value that is defined in the DTD.

4

Movie.dtd

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!ELEMENT movielist (movie*)>
<!ELEMENT movie (title, releaseyear, director, cast?, genre+,
  country, rating)>
<!ELEMENT title (#PCDATA)>
<!ATTLIST title subtitle CDATA #IMPLIED>
<!ELEMENT releaseyear (#PCDATA)>
<!ELEMENT director (#PCDATA)>
<!ELEMENT cast (actor+)>
<!ELEMENT actor (#PCDATA)>
<!ATTLIST actor role CDATA #REQUIRED>
<!ELEMENT genre (#PCDATA)>
<!ELEMENT country (#PCDATA)>
<!ELEMENT rating (#PCDATA)>
<!ATTLIST rating scale CDATA #FIXED "10">
```

6

10 built in data types

- ID. Unique ID only allowed once in the document.
- IDREF. Refers to elements with the given ID.
- IDREFS. The same as IDREF, but more than one ID is used.
- ENTITY. Declares the name of an entity.
- ENTITIES. Declares two or more entity names.
- NMTOKEN. Declares a name token. Is any mixture of name characters.
- NMTOKENS. The same as NMTOKEN but more than one name token is declared.
- ENUMERATION. Enumerated attributes can take one of a list of values declared.
- NOTATION. Declares the name of a notation.
- CDATA. String value.

5

sciFi.xml

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<!DOCTYPE movielist SYSTEM "http://www.cs.umu.se/~c97fno/movielist.dtd">
<movielist>
  <movie>
    <title>2001: A Space Odyssey</title>
    <releaseyear>1968</releaseyear>
    <crew>
      <director>Stanley Kubrick</director>
      <writer>Arthur C. Clarke</writer>
      <producer>Stanley Kubrick</producer>
    </crew>
    <cast>
      <character>
        <actor gender="male">Keir Dullea</actor>
        <role>David Bowman</role>
      </character>
      <character>
        <actor gender="male">Douglas Rain</actor>
        <role>HAL 9000</role>
      </character>
    </cast>
    <genre>Sci-Fi</genre>
    <country>UK</country>
    <rating>8.1</rating>
  </movie>
  .
  .
  .
</movielist>
```

7

XML Query Languages

- XQL - MicroSoft, Texcel, webMethods
- XQuery - XML Query Working Group
- XML-QL - AT&T
- Lorel - Stanford
- ...

XQuery seems to be the choice of W3C

Xquery 1.0 + Xpath 2.0

Check out the W3C website if interested.

8

XML - DTD types

- Meat and Poultry XML (mpXML).
- MathML
- SOAP
- VXML
- UPnP
- SyncML
- Chemical Markup Language (CML)
- Bioinformatic Sequence Markup Language (BSML)
- ...

10

XML Meta Standards

The whole W3C effort has been mired in an alphabet soup of (competing) standards. It looks pretty hopeless. But they are trying now to fuse the whole heap together.

- XSchema - Give name spaces and stronger types to DTD.
- XSL, XSLT - more advanced style sheet language
- XLink - add hyperlinks to other XML documents.
- Xpointer - add links within an XML document
- CSS - style sheet language
- DOM - Document object model. Standard API to manipulating and XML document from within a programming language
- XHTML - the new HTML in XML. Slight changes to HTML.
- RDF - Resource Description Framework

9

Conclusion

XML promises structured data on the web.

A novice query formulator might jump start activity toward this goal.

For now we must wait - hopefully all the dust will settle and all the noise will yield something useful.

11