

## Hypermedia and the Web – Lecture Notes – Introduction to XML

What is XML?

- XML is a data structuring technology, used to design document formats for a wide range of uses.
- XML is a language for creating other markup languages. So, you can use XML to create a specific document format, such as HTML
- XML is a markup language – information about the data is embedded in the document with the data itself. As a result, XML is self-describing. This makes it well suited for data interchange.
- XML is infrastructure – it is the core building block for a wide range of other technologies.
- XML is verbose. Since XML is text-based, it is generally larger than the equivalent binary representation.

Brief history of XML

- Grew out of Standard Generalized Markup Language (SGML), standardized by ISO in 1986 (ISO 8879)
  - Problems of SGML
  - Difficult to write parsers due to tag minimization
  - Specification not freely available
  - Tools were expensive, very few open source tools
- HTML was influenced by SGML, was an application of SGML
- HTML combined content and presentation (things like color, font size, etc.) and made it difficult to encode complex data inside an HTML document in a machine readable way
- circa 1996, SGML community began engaging Web Consortium and key Web vendors to adopt SGML for the Web
- XML grew out of an effort to re-engineer SGML for the Web, generally to make it more simple, and easier to parse
- XML was approved by the Web Consortium in 1998
- Now the basis for a family of standards, including
  - XML Namespaces
  - XML Schema
  - XLink, Xpath
  - XQuery
  - Resource Description Framework (RDF) / Semantic Web
  - RSS (Really Simple Syndication)
  - ... as well as in a host of other specifications

Primary aspects of XML:

data is tree structured, a single-rooted tree

each datum lives in one place in the tree

each node in the tree has a name, properties, and content

in XML-speak, nodes = elements, properties = attributes

For example, document format from the first lab assignment, for representing RFCs:

rfc

+--front  
+--middle  
+--back

front

+--title  
+--author\* (initials, surname, fullname)  
  +--organization  
  +--address  
    +--postal  
      +--street  
      +--city  
      +--region  
      +--code  
    +--email  
+--date  
+--workgroup  
+--abstract

## XML Namespaces

### Goals:

- want to share elements from different XML schemas
- problem: namespace collisions for XML elements having the same name
  - example: different meaning of “day” across domains

### General form

`xmlns:identifier={URL or URN}`

### Example:

```
<?xml version="1.0"?>
<bk:book xmlns:bk='urn:loc.gov:books'
  xmlns:isbn='urn:ISBN:0-395-36341-6'>
  <bk:title>Cheaper by the Dozen</bk:title>
  <isbn:number>1568491379</isbn:number>
</bk:book>
```

### Namespace defaulting

A default namespace is considered to apply to the element where it is declared (if that element has no namespace prefix), and to all elements with no prefix within the content of that element.

```
<?xml version="1.0"?>
<!-- unprefix element types are from "books" -->
<book xmlns='urn:loc.gov:books'
  xmlns:isbn='urn:ISBN:0-395-36341-6'>
  <title>Cheaper by the Dozen</title>
  <isbn:number>1568491379</isbn:number>
</book>
```