# XML ROLE IN DEFINING METADATA USING DTDs FOR THE DIGITAL RESOURCES ON THE WEB

By

**GAYATHRI DEVI S***

## ABSTRACT

*This paper introduces the concept of deficiencies in html and advocates that xml can be a substitute of html for its advantages. Being the web standard, it also accomplishes its application to the library documents as a metatag descriptor. The paper explains this through an example of XML, DTD, CSS, HTML files with few catalogue fields as descriptors. It also focuses on the various recommendations brought by RDF and Dublin Core in the development of metadata.*

---

* Information Scientist, National Law School of India University, P.B. No:7201, Nagarbhavi, Bangalore-72.
E-mail : gayathrimanjunath@usa.net

## 0. Introduction

WWW is an effective source of information of the present day. There is no alternative or substitute to its abundance, utility and value. This ocean of information subjected to retrieval is serving the user community with a big answer pool. The user needs to dig and dig, mine and mine for the appropriate and relevant information. The data mining process is a time consuming, labourious work for a professional in information era. However, this difficulty can be overcome by structuring the web documents in a proper format adhering to certain standards. This helps the web community not only in effective data retrieval but also helps in structured organisation of web documents. The structured organisation paves way for an effective data exchange activity.

The prevailing web documents lack the definition of structure and thereby the process of indexing and categorising is somewhat very minimum. The *present hypertext markup language supports the look of a document but does not call for any structure of the data in a web document.* The extensible markup language makes up for the deficiencies that are with html and supports the web document with clear demarkation of data and markup tags.

However, the WWW consortium is making all attempts in developing a Resource Description Framework which gives a standard for defining the metadata set and its conversion. The evolution of XML is the result of this initiative and recommendations in this direction.

This paper attempts to define metadata for a library document using the DTDs in XML. The library documents that are considered for the construction of DTDs are the books. A few descriptions like author, title, publisher, place, year, etc are considered for coding in XML. . A simple Javascript is written just to display the relevant records pertaining to the selection in the right hand frame.

## 1. Metadata

Metadata or Meta tags define data about data. The Metadata serves as a fundamental solution to improve the access of mass information through better search and retrieval process. The effective use of metadata requires three elements
1) standardized semantics,
2) a definitive syntax
3) framework for exchange.

These three elements provide an architecture for resource description that can work across any subject areas on the web.

## 2. Dublin Core Metadata

The Dublin Core is a collection of elements designed to help researchers find electronic resources in a manner similar to using a library card catalogue.

The Dublin Core MetaData initiative is a mission to make it easier to find resources using the Internet through the following activities :

- Developing Metadata standards for resource
- Search and Retrieval across different subject areas
- Defining frameworks for the interoperability of metadata sets
- Facilitating the development of community or subject specific metadata sets that work within these frameworks.

The Dublin Core Metadata initiative provides not only for general description but also for subject specific extensions. Dublin Core Metadata can be carried in HTML, XML and RDF. The Resource Description Framework builds on the WWW consortium (W3C) efforts to design an architecture for metadata on the web. Resources are described with properties. A property is a specific characteristic, attribute or relationship of a resource. RDF only defines an XML syntax for encoding resource, property, type property values in XML.

The Dublin Core elements include basic cataloguing information, in particular :title, creator, subject, description, publisher, contributor, date(YYYY-MM-DD), type, format, identifier, source, language, rights.

The Dublin Core Metadata Element set has 15 broad categories that are useful in creating simple easy-to understand descriptions for most information resources. The basic

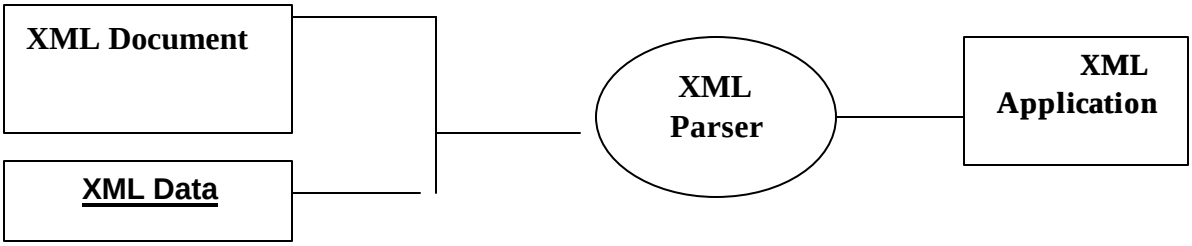
element set is intended to capture most of the fundamental descriptive categories necessary to promote effective search and retrieval.

# 3. Extensible Markup Language (XML)

Extensible Markup Language is a markup language like html which gives the user a flexibility of defining his own tags. The XML defines the data it contains thereby finding its way into the e-commerce solutions like B2B, B2C. XML will be the most common tool for all data manipulation and data transmission. Like html, Xml is also a cross-platform, software & hardware independent tool for transmitting information. XML uses a DTD to describe the data, which can be separately defined or defined within an html document.

## 3.1 XML Authoring

The XML documents are most commonly created with an editor. It could be a simple text editor like notepad, vi, edit(Dos based editor). The Editors like Adobe Framemaker or Jumbo can also be used. The XML document contains content character data marked up with XML tags. The XML documents are little complicated to write than html documents. The XML description expects a particular layout and it validates the document according to that specification. Therefore, the document is always tested for its well formedness or validity by the XML document against the parser. The XML parser checks the XML document against the DTD and splits the document into markup tags and data regions. Once the parser checks for the validity of the document, the XML application can further be processed.




The parser passes the XML document to the browser like Mozilla or IE5. The data are displayed to the user.

The Cascading style sheet rules of html can easily be applied to xml documents. The Extensible Style language XSL is a more advanced style-sheet language specifically designed for use in XML documents. CSS can only change the format of a particular element but XSL can rearrange and reorder elements. XML documents can live on the web like html documents and referred by an URL. XML allows for linking the documents through Xlinks. These Xlinks can be bi-directional, multi-directional or even point to multiple mirror sites from which the nearest is chosen. Xpointers enable links to point not just at a particular document at a particular location, but to a particular part of a

particular document. They can point to ranges or spans. XML provides for the 2-byte unicode character set for all language representations.

### 3.2 Document Type Definition (DTD)

XML being a metamarkup describes markup language tags which are defined through Document Type Definition (DTD). Individual documents can be compared against the DTDs in a process of validation. If the document matches the constraints listed in DTD, then the document is said to be valid, otherwise it is said to be invalid.

DTDs provide a list of elements, attributes, notations & entities contained in a document as well as their relationships to one another. DTDs can be included in the file that contains the document they describe, or they can be linked from an external URL. The external DTDs can be shared by different documents and websites. DTDs provide a means for applications, organisations and interest groups to agree upon a document and enforce adherence to markup standards. DTDs help in data exchange as they follow the structural organisation.

Here is an example of few cataloguing items of the books that is shown in the **xml(Books.xml) and dtd(Books.dtd)** format. A cascading Style sheet file **Books.css** gives the format of how the title and other details have to look in the web browser. The **index.html** is the file which divides the screen into two frames. Depending on the selection of the item in the left frame, the details are displayed in the right frame. The **docselector.html** is the javascript file which defines a loadfile function and calls the appropriate xml file based on selection.

**Books.xml**

```
<?xml version=”1.0"?>
<!DOCTYPE bookcollection SYSTEM “books.dtd”>
<Catalogue>
    <document>
        <Title></Title>
        <Author></Author>
        <Publisher></Publisher>
        <Place></Place>
        <Year></Year>
        <Accno></Accno>
     </document>
    <document>
        <Title>Mastering XML</Title>
        <Author>Ann Navarro</Author>
        <Publisher>BPB Publications</Publisher>
        <Place>New Delhi</Place>
        <Year>2000</Year>
        <Accno>16531</Accno>
```

```
    </document>
</Catalogue>
```

**Books.dtd**

```
<!ELEMENT books (document+)>
<!ELEMENT books (title, author, publisher, place, year,accno)>
<!ELEMENT author (#PCDATA)>
<!ELEMENT publisher (#PCDATA)>
<!ELEMENT place (#PCDATA)>
<!ELEMENT year (#PCDATA)>
<!ELEMENT accno (#PCDATA)>
```

**Books.css**

```
title, author {
display :block;
font-family :Arial, Helvetica;
font-weight: bold;
font-size: 15pt;
color : “red”;
}
publisher, place, year, accno {
margin-left: 25;
display :block;
font-family :Arial, Helvetica;
font-weight: bold;
font-size: 12pt;
color : “black”;
}
```

**index.html**

```
<!DOCTYPE HTML PUBLIC “-//W3C//DTD HTML 4.0 Transitional//EN”>
<html>
<head>
<title>XML BOOKS Selector</title>
</head>
<frameset COLS=”25%,75%”>
<frameset name=”leftFrame” src=”docselector.html”>
<frameset name=”rightFrame” src=””>
</frameset>
</html>
```

**docselector.html**

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional //EN">
<html>
<head>
<script language="JavaScript">
function loadFile()
 {
   Var filename
   Var selectionValue
   SelectionValue=document.forms[0].selectList.selectedIndex
filename = document.forms[0].selectList.options[selectionValue].value
parent.rightFrame.location=filename
}
</script>
</head>
<body>
<H4> XML Books File </H4>
<P>Select the Books Details you wish to see displayed in the right-hand frame.
<FORM NAME="selectForm">
<P>
<SELECT NAME="selectList">
<OPTION VALUE="books.xml">XML
</SELECT>
<P>
<INPUT TYPE="BUTTON" VALUE="Load Document" onClick="loadFile()">
</FORM>
```

## 4. Conclusion

Though the metadata tag in html provides certain degree of indexing, it lacks the hierarchical and structural definitions of data. Therefore, the retrieval is never to the exact matching. The XML providing the hierarchical structure and a greater degree of segregation of markup tags and data, provides for a powerful mechanism of data retrieval. The W3C is making new recommendations which may further enhance the power of XML. XML is still in development and it is bound to change to a great extent. However, the method of enabling a structured definition requires hierarchical organisation of web documents. This development will surely help the search engines to retrieve the documents to a greater proximity. The indexing and categorising can be subjected to greater effectiveness through XML and DTDs. Therefore, the XML with its varied characteristics undoubtedly helps the information users and information professionals in effective search and retrieval process.

## 5. References

1. Paper on "Enhancing Retrieval Effectiveness ; XML Customised DTDs" by

Shalini R Urs and K.S. Raghavan. Seminar on Content Organisation in the new Millenium, 2-4 June 2000, Sarada Ranganathan Endowment for Library Science, Bangalore, 2000.
2. Mastering XML by Ann Navarro, Chuck White, Linda Burman, BPB Publications, New Delhi, 2000.
3. XML in 21 Days by Simon North, Paul Hermans, TechMedia, New Delhi, 1999
4. XML Bible by Elliotte Rusty Harold, IDG Books India, New Delhi, 2000

**Websites**

1. http://www.developer.ibm.com/devcon/rsinnarticle.htm
2. http://www.xml.com/metadata/
3. http://www.w3schools.com/xml/