

METADATA: TOWARDS MACHINE-ENABLED INTELLIGENCE

Vandana Dhingra¹ and Komal Kumar Bhatia²

¹School of Computational Sciences, Apeejay Stya University, Gurgaon, India

vandana.dhingra@asu.apeejay.edu

²Department of Computer Science, YMCA University, Faridabad, India

komal_bhatia1@rediffmail.com

ABSTRACT

World Wide Web has revolutionized the means of data availability, but with its current structure model, it is becoming increasingly difficult to retrieve relevant information, with reasonable precision and recall, using the major search engines. However, with use of metadata, combined with the use of improved searching techniques, helps to enhance relevant information retrieval. The design of structured, descriptions of Web resources enables greater search precision and a more accurate relevance ranking of retrieved information. One such efforts towards standardization is, Dublin Core standard, which has been developed as Metadata Standard and also other standards which enhances retrieval of a wide range of information resources. This paper discusses the importance of metadata, various metadata schemas and elements, and the need of standardization of Metadata. This paper further discusses how the metadata can be generated using various tools which assist intelligent agents for efficient retrieval.

KEYWORDS

Semantic web, Intelligent retrieval, FOAF, RDF, Metadata Schema & Dublin Core

1. INTRODUCTION

Metadata, in general terms is defined as “data about data” or information about information, it is data that describes information resources. Metadata is structured information that describes, locates and makes it easier to retrieve, use, or manage an information resource. The purpose of metadata is ‘to facilitate search, evaluation, acquisition, and use’ of resources [1]. It is basically structured data that allows machine readability and understandability, which is basically the vision for Semantic Web.

Metadata can be used to add information that can be used to describe the document such as:

- Title of document
- Author of document
- Time and date of creation
- Standards used
-

There are three main types of metadata [2]:

1 *Descriptive metadata* describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.

2. *Structural metadata* indicates how compound objects are put together, for example, how pages are ordered to form chapters.

3. *Administrative metadata* provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. There are several subsets of administrative data; two that sometimes are listed as separate metadata types are:

- a. Rights management metadata, which deals with intellectual property rights, and
- b. Preservation metadata, which contains information needed to archive and preserve a resource.

Metadata descriptions present two advantages [3]:

- They enable the abstraction of details such as the format and organization of data, and capture the information content of the underlying data independent of representational details. This represents the first step in reduction of information overload, as intentional metadata descriptions are in general an order of magnitude smaller than the underlying data.
- They enable representation of domain knowledge describing the information domain to which the underlying data belong. This knowledge may then be used to make inferences about the underlying data. This helps in reducing information overload as the inferences may be used to determine the relevance of the underlying data without accessing the data.

The remainder of this paper is organized as follows: In section 2 we have given an overview of metadata schemas and element sets .Section 3 describes the structuring of metadata, including Dublin Core standard .Section 4 describes about the role of metadata in Semantic web. Section 5 covers how we can generate metadata for the existing web pages with the related work and Section 6 concludes the paper.

2. Metadata Schemas and Element Sets

2.1. Metadata Schema

A schema is a set of rules covering the elements and requirements for coding. Metadata schemas are sets of metadata elements designed for a specific purpose, such as describing a particular type of information resource. The definition or meaning of the elements themselves is known as the semantics of the schema.

Some of the most popular metadata schemas include:

Dublin Core
AACR2 (Anglo-American Cataloguing Rules)
GILS (Government Information Locator Service)
EAD (Encoded Archives Description)
IMS (IMS Global Learning Consortium)
AGLS (Australian Government Locator Service)

Examples of schemas in the semantic web include Dublin Core, FOAF (Friend of a Friend), and many others. Metadata schemas specify names of elements and their semantics. Many current metadata encoding schemes are:

- SGML (Standard Generalized Mark-up Language)
- XML (Extensible Mark-up Language).
- XML, developed by the World Wide Web Consortium (W3C.)
- HTML (Hyper-Text Markup Language)

2.2. Metadata Deployment

Metadata may be deployed in a number of ways:

- Embedding the metadata in the Web page using META tags in the HTML coding of the page
- As a separate HTML document linked to the resource it describes
- In a database linked to the resource. The records may either have been directly created within the database or extracted from another source, such as Web pages.

3. Structuring of Metadata

The metadata of each web document has its own structure, so there is need for some structure so that this data can be processed by machines in uniform manner. Different author's will have their own ways to describe a given page, by this each page will have its own unique structure, and it will not be possible for an automated agent to process this metadata in uniform way. So for processing by machines, there is need for some metadata standard. Such kinds of standards are called metadata schemas. Dublin Core is one such standard.

3.1 Dublin Core Standard

It has 13 elements which are called as Dublin Core metadata Element Set. Basic metadata elements indicate the title, author, year of publication and similar simple bibliographic data. Richer metadata structures also cover technical features, copyright properties, annotations and so on[4]. These original 13 core elements were later increased to 15[5]:

- Title: a name given to the resource
- Creator: an entity primarily responsible for making the resource
- Subject: the topic of the resource
- Description: an account of the resource
- Publisher: an entity responsible for making the resource available
- Contributor: an entity responsible for making contributions to the resource
- Date: a point or period of time associated with an event in the lifecycle of a resource
- Type: the nature or genre of the resource
- Format: the file format, physical medium, or dimensions of the resource
- Identifier: an unambiguous reference to the resource within a given context
- Source: the resource from which the described resource is derived
- Language: a language of the resource
- Relation: a related resource
- Coverage: the spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant
- Rights: information about rights held in and over the resource

Figure 1 shows how DC metadata Standard can be used in html web page The Dublin Core metatags can be placed within the head section of the HTML code of web pages. Normally, the Dublin Core elements are preceded by the "DC" abbreviation. The best method to be used

whenever you need to use Dublin Core elements in your HTML code is to have this metadata embedded. The main characteristics of Dublin Core are:

- Simplicity (of creation and maintenance)
- Interoperability (among collections and indexing systems)
- International applicability
- Extensibility
- Modularity

```
<html>
<head>
<title>Example Metadata in Dublin Core</Title>
<meta name="DC.Title" content="Example Metadata in Dublin Core">
<meta name="DC.Creator" content="Vandana Dhingra">
<meta name="DC.Creator.Address" content="vandana@yahoo.com">
<meta name="DC.Date.Created" Content="2012-02-10">
<meta name="DC.Type" Content="Text.Homepage.Tutorial">
<meta name="DC.Format" Content="Text/Html">

</head>
<body>
This is my first tutorial in Dublin Core
</body>
</html>
```

Figure 1: Example of the Dublin Core Metadata Elements could be used in Web content

4. Role of Metadata in Semantic Web

Tim Berners-Lee introduced the term “Semantic Web” foreseeing the concept that we are creating a web which can only be managed when we use intelligent software agents[6]. Tim Berners-Lee's vision for Web is termed the "semantic Web," where semantic Meta data are the building blocks. One such effort towards this is Semantic Search engine is Swoogle[7]. It is an implemented metadata and search engine for online Semantic Web documents. Swoogle analyzes these documents and their constituent parts (e.g., terms and triples) and records meaningful metadata about them [8]. By annotating or enhancing documents with semantic metadata, software programs (agents) can automatically understand the full context and meaning of each document and can make correct decisions about who can use the documents and how these documents should be used. It offers a set of rules for creating semantic relations and RDF Schema can be used to define elements and vocabularies. The relations are defined with a very simple mechanism that can also be processed by machines. In an RDF environment every resource has to have a unique identifier (URI). It can have properties and properties can have values.

Users will not be able to process the huge amount of knowledge available on the web. Semantic Web is based on the fundamental idea that web resources should be annotated with semantic mark-up that captures information about their meaning[9]. The objective is to provide a common framework that allows data to be shared and reused across applications, and which brings structure to the meaningful content of WebPages which results in creating an environment where software agents roaming from page to page readily carry out advanced searches. Hence, the metadata initiatives are important steps towards the realization of the Semantic Web. Following subsections describes various representation languages for describing metadata-

4.1 RDF/XML

RDF provides a rich foundation to describe metadata. RDF (Resource Description Framework) allows multiple metadata schemes to be read by humans as well as parsed by machines. It uses XML (EXtensible Markup Language) to express structure thereby allowing metadata communities to define the actual semantics [10]. RDF allows multiple objects to be described without specifying the detail required. The underlying glue, XML, simply requires that all namespaces be defined and once defined; they can be used to the extent needed by the provider of the metadata.

The Resource Description Framework (RDF), developed by the World Wide Web Consortium (W3C), is a data model for the description of resources on the Web that provides a mechanism for integrating multiple metadata schemes. For metadata, the most popular bindings nowadays are XML, or, more specifically, RDF [11]. In RDF a namespace is defined by a URL pointing to a Web resource that describes the metadata scheme that is used in the description. Multiple namespaces can be defined, allowing elements from different schemes to be combined in a single resource description. Multiple descriptions, created at different times for different purposes, can also be linked to each other.

Example: Description with a single statement, which uses a language-tagged literal value.

“Vandana is the author of the resource”[12]

```
Vandana Dhingra is the creator of the resource
http://vandana.dhingra/home/index.html
can be represented in RDF by
<? xml version="1.0"?>
<rdf: RDF
xmlns:rdf=" http://vandana.dhingra/2000/22 -rdf-syntax-
ns#"
xmlns:s="http://description.org/schema/">
<rdf: Description about="
http://vandana.dhingra/home/index.html ">
<s: Creator>vandana</s: Creator>
</rdf: Description>
</rdf: RDF>
```

The ontology creation language OIL extends RDF Schema and allows you to be much more specific about what sort of thing a person is, the properties a thing needs to have to be a `wn:person` and so on [13].

4.2 FOAF

It is one of the metadata standards for the Semantic Web, Figure 2 represents an example of Web document represented in FOAF. FOAF is all about creating and using machine-readable homepages that describe people, the links between them, and define things they create and do. FOAF defines an RDF vocabulary for expressing metadata about people and the relation among them.

```
rdf:Description>
<dc: title>Tutorial </dc: title>
<dc: creator>
  <foaf:Person>
    <foaf:name>Vandana Dhingra </foaf:name>
    <foaf: homepage
rdf:resource="http://rdfweb.org/people/vandana "/>
  </foaf: Person>
<dc: description>Semantic Web Tutorial
```

Figure 3: Metadata represented in FOAF

5. Generation of Metadata for Existing Web Pages

Metadata generation tools generate metadata from Web resources and store it in RDF for later use. There are various metadata tools to create metadata for already existing Web pages. One such tool for creating metadata is

1. Reggie[14]- is a tool capable of extracting metadata from given Web pages. The user can select any existing schema file or can create his/her own schema files. Reggie extracts the META tags from a given URL and attempts to add them to the most appropriate fields of the chosen schema. It also allows users to create their own metadata schema files.

2. DC-DOT [15] is another tool for metadata extraction. In comparison to Reggie where the user provides the schema file, DC-DOT specifically uses the Dublin Core schema, a metadata element set for description of electronic resources, to extract metadata from a given Web resource. DC-DOT uses the information contained in the META tags of a Web source to generate the RDF model.

Using DC-dot

It will read the page you submit and automatically generate DC metadata for the web page.

UKOLN: DC-dot Dublin Core metadata editor - Mozilla Firefox

File Edit View History Bookmarks Tools Help

UKOLN: DC-dot Dublin Core metadata edi... x Problem loading page x metadata and foaf - Google Search x Metadata - LISWiki

www.ukoln.ac.uk/cgi-bin/dcdot.pl

Most Visited Latest Headlines Customize Links Suggested Sites Web Slice Gallery Home Home

DC dot

Dublin Core metadata editor

Results for URL: <http://ymcaust.ac.in> [summary]

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
<meta name="DC.title" content="YMCA University of Science & Technology, Faridabad" />
<meta name="DC.subject" content="joomla; Joomla" />
<meta name="DC.description" content="Joomla! - the dynamic portal engine and content management system" />
<meta name="DC.publisher" content="#" />
<meta name="DC.date" scheme="DCTERMS.W3CDTF" content="2012-03-22" />
<meta name="DC.type" scheme="DCTERMS.DCHIType" content="Text" />
<meta name="DC.format" content="text/html; charset=utf-8" />
<meta name="DC.format" content="40427 bytes" />
<meta name="DC.identifier" scheme="DCTERMS.URI" content="http://ymcaust.ac.in" />
```

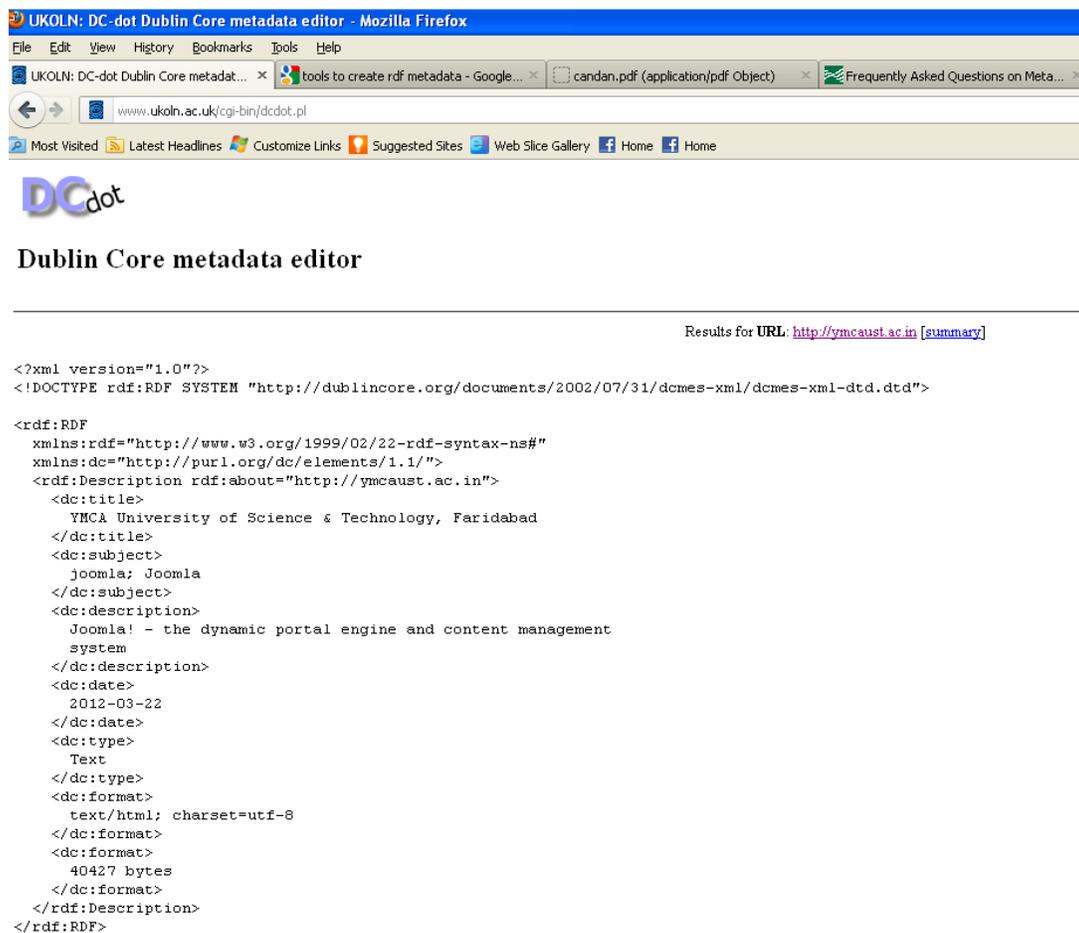
If necessary, edit the values in the boxes below, and

Convert metadata to

Title

Creator (author)

Figure 4: Generation of Dublin Core metadata



UKOLN: DC-dot Dublin Core metadata editor - Mozilla Firefox

File Edit View History Bookmarks Tools Help

UKOLN: DC-dot Dublin Core metadata... tools to create rdf metadata - Google... candan.pdf (application/pdf Object) Frequently Asked Questions on Meta...

www.ukoln.ac.uk/cgi-bin/dcdot.pl

Most Visited Latest Headlines Customize Links Suggested Sites Web Slice Gallery Home Home

DCdot

Dublin Core metadata editor

Results for URL: <http://ymcaust.ac.in> [summary]

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF SYSTEM "http://dublincore.org/documents/2002/07/31/dcmes-xml/dcmes-xml-dtd.dtd">

<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://ymcaust.ac.in">
    <dc:title>
      YMCA University of Science & Technology, Faridabad
    </dc:title>
    <dc:subject>
      joomla: Joomla
    </dc:subject>
    <dc:description>
      Joomla! - the dynamic portal engine and content management
      system
    </dc:description>
    <dc:date>
      2012-03-22
    </dc:date>
    <dc:type>
      Text
    </dc:type>
    <dc:format>
      text/html; charset=utf-8
    </dc:format>
    <dc:format>
      40427 bytes
    </dc:format>
  </rdf:Description>
</rdf:RDF>
```

Figure 5: A Dublin Core description represented in RDF

This metadata generally is read by automated tools. It will read the page for the submitted URL and will automatically generate DC metadata for the user. Figure 4 shows the generated metadata for the submitted URL of <http://ymcaust.ac.in> and Figure 5 shows the generated RDF document with metadata of submitted URL <http://ymcaust.ac.in>.

5. Conclusion

Metadata is a very useful element in organizing content on the Internet and in enabling to find relevant search results effectively and in this research paper we have discussed about the metadata, and how the usage of metadata impacts the searching efficiency for search agents and generation of metadata for existing web pages. But, there are still many problems and issues that remain unsolved. Various open research issues in this area are –

- Interoperability with respect to syntax, semantics, vocabularies, languages, and underlying models.
- Identifying the best metadata models, schemas, and vocabularies for metadata models;

- Ensuring authenticity, and integrity of the metadata by using various metadata extraction tools
- How to design the repositories for storage and management of metadata.
- Lot of efforts is required in human-generated metadata, a large number of research initiatives are required in focusing on technologies to enable the automatic generation of metadata in various domains like text, imaging ,video.

REFERENCES

- [1] IEEE 2001 Draft standard for Learning Object Metadata Draft 6.1, April 2001
<http://ltsc.ieee.org/wg12/index.html>
- [2] N ISO. (2004) Understanding Metadata. Bethesda, MD: NISO Press, p.1
- [3] Sheth, A” Changing Focus on Interoperability in Information Systems: from System, Syntax, Structure to Semantics” Goodchild M.F., Egenhofer, M.J., Fegeas, R., Koffman, C.A. (eds.): Interoperating Geographic Information Systems. Kluwer Academic Publishers, Boston (1999) 5-29
- [4] Erick Duval Journal of Universal Computer Science, vol. 7, no. 7 (2001), 591-601 28/7/01 Springer Pub. Co. Metadata Standards: What, Who, Why?
- [5] Dublin Core Metadata Element Set, Version 1.1 Reference Description
<http://purl.org/dc/documents/rec-dces-19990702.htm>
- [6] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. Scientific American, May, 2001.
- [7] Li Ding, et al., “Swoogle: A semantic web search and metadata engine,” Int. Proc. 13th ACM Conf. on Information and Knowledge Management, Washington D.C., USA, Nov. 2004
- [8] Li Ding, et al., “Search on the Semantic Web,” Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, Baltimore, Tech. Rep. TR CS-05-09 , Sep. 2005
- [9] Steffen Staab ,IEEE INTELLIGENT SYSTEMS Where Are the Rules? 1094-7167/03/\$17.00 © 2003 IEEE
- [10] D. Brickley and R. Guha. Resource Description Framework (RDF) schema specification, 2000.
<http://www.w3.org/TR/RDF-schema>.
- [11] W3C Document Metadata and resource Description, April 2001
<<http://www.w3.org/Metadata/>>.
- [12] Vandana Dhingra,Komal Bhatia” Towards Intelligent Information Retrieval on Web” International Journal on Computer Science and Engineering (IJCSE) ,Vol. 3 No. 4 Apr 2011
- [13] Libby Miller “ Ontologies and Metadata “A Draft Discussion of issues raised by the Semantic Web Technologies Workshop, 22-23 November 2000.
- [14] Resource discovery unit of DSTC, Reggie the Metadata Editor.<http://metadata.net/dstc/SchemaFiles.html>.
- [15] Dublin Core Metadata Editor (DCDOT), 2000.
<http://www.ukoln.ac.uk/metadata/dcdot>

Authors

Ms. Vandana Dhingra, Pursing Ph.d. from YMCAIE University in the field of Semantic Web have done M.Tech (Information Technology), B.E. (Computer Science and Engineering) with First Class with distinction and is having experience of teaching in reputed Engineering colleges for 15 years. She has administrative experience of being Head of Department of Computer for five years. She has published research papers in IEEE Digital Library and various International and National Conferences and currently is working in Apeejay Stya University. Her subjects of Interest include Digital System Design using VHDL Language, Operating Systems and Distributed Operating Systems. Her research interests are Semantic Web, Information retrieval, Crawlers, Search engines and Ontologies.



Dr. Komal Kumar Bhatia received his B.E, M.Tech. And Ph.D. degrees in Computer Science Engineering with Honors from Maharishi Dayanand University. Presently, he is working as Associate Professor in Computer Engineering Department in YMCA University of Science & Technology, Faridabad. He is guiding PhDs in Computer Engineering and has guided many M.Tech Thesis. His subject interests include Analysis and Design of algorithms, Information retrieval. He has almost 80 research papers in referred International Journals, National Journals, and International Conferences to his credit. His research interests include Web Crawlers, Hidden Web, Information Retrieval, Deep Web and Semantic Web.

