UNIT 9 METADATA : MARC21-856 FIELD, DUBLIN CORE, TEI

Structure

9.0 OBJECTIVES

In the previous Unit you have learnt about the record formats for print documents. Presently a substantial portion of information is available in electronic form. In this Unit we will be discussing some major international standards available for record formats of such documents.

After reading this Unit, you will be able to:

- understand the concept of Metadata with special reference to electronic information resources;
- understand MARC21 Field 856 structure and elements;
- identify Dublin Core metadata elements and their applications; and
- know Text Encoding Initiative and its elements.

9.1 INTRODUCTION

The information available in the Internet is so vast and easily accessible that library patrons today expect information services based on it. The library and information profession is traditionally equipped for handling huge information sources so long available in documents of non-electronic form. The absence of professional librarians' skills is well reflected in the indexing approach taken by the general search engines like Alta Vista, Google, Yahoo! etc. which are at the most successful in recall but definitely not in precision. In addition, the output of these search engines definitely lacks uniformity.

In keeping with the changes in form and format of sources of information, the AACR2, Revised edition 1988 uses a highly neutralised terminology in cataloguing.

For example, it uses 'item' instead of document. This trend is also reflected in many Machine Readable Catalogues (MARCs). The CCF uses 'Name of Person' for field 300. The idea is to catalogue any kind of information source, termed as item. The evolution of ISBDs definitely mirrors the changes in electronic information. The introduction of ISBD(NBM – Non-book-material), ISBD(CF – Computer Files), ISBD(ER – Electronic resources) was quite obvious. When Internet became widely popular, hosting a variety of information resources, soon it was realised that the neutralised terminology did not readily accommodate the requirements of describing Internet documents. In fact it implied that more tags are needed for the bibliographic descriptive elements in addition to the already existing huge number of tags.

In essence, there is an absolute need for describing the Internet or web based documents, so that search and retrieval are effective and efficient. The recent trends in providing bibliographic description (metadata) of web documents can be grouped into two. One approach is, embedding the bibliographic description in the HTML documents (or XML) and build search engines that look for the specific tags for indexing purpose. The second approach is to prepare the bibliographic records of web documents in a database management system, as is the case with MARCs. Though first one is the basic approach of Dublin Core, Dublin Core can be applied in the second approach.

The simplest definition of metadata is 'structured data about data'. Metadata is descriptive information about an object or resource – be it physical or electronic. While metadata itself is relatively new, the underlying concepts behind metadata have been in use for as long as collections of information have been organised. Library card catalogs represent a well-established type of metadata that has served as collection management and resource discovery tool for decades. Metadata can be generated either 'by hand' or derived automatically using software. The Unit discusses some standards such as MARC21, Dublin Core and TEI for description of online and electronic resources.

9.2 MARC21 – 856 FIELD

9.2.1 MARC21

The MARC21 format is a standard for the representation and communication of bibliographic and related information in machine-readable form.

A MARC record involves three elements: the record **structure**, the **content designation**, and the data **content** of the record [www.loc.gov/marc/856guide.html].

a) The **structure** of MARC records is an implementation of national and international standards, e.g., *Information Interchange Format* (ANSI Z39.2) and *Format for Information Exchange* (ISO 2709).

b) **Content designation**, the codes and conventions established to identify explicitly and characterise further the data elements within a record and to support the manipulation of those data, as defined in the MARC21 formats.

c) The **content** or data of most data elements is defined by standards outside the formats, e.g., *Anglo-American Cataloguing Rules*, *Library of Congress Subject Headings, etc.*

MARC21 has made provision for electronic resources. In the case of online resources the field 856 is assigned. Field 856 was defined in the USMARC bibliographic and holdings formats to contain information that identifies the electronic location of an item, including enough information to retrieve the item. When it was defined in early 1993, the Uniform Resource Locator (URL) was not an accepted standard and the World Wide Web was in the early stages of development. As institutions began to use the new electronic location field, they

began to discover various applications of it. The specifications for the field 856 as given by the latest MARC21 manual [www.loc.gov/marc/856guide.html] are discussed below:

9.2.2 Field 856 in MARC21

Field 856 in the MARC21 Information formats is used for electronic location and access information to an electronic resource and contains information related to the resource. The field may be used in a bibliographic or holdings record for a resource when that resource or a subset of it is available electronically. In addition, it may be used to locate and access an electronic version of a non-electronic resource described in the bibliographic record, part of the resource, or a related electronic resource. In an authority record it contains the electronic location information about the entity authorized by the record.

9.2.2.1 Content of Field 856

The data in field 856 may be a Uniform Resource Locator (URL), which is recorded in subfield $u, or it may parse the necessary locator information into separately defined subfields. An access method, or protocol used, is given as a value in the first indicator position (if access method is email, ftp, telnet, dial-up, or HTTP) or in subfield $2 (if access method is anything else). The access method is also the first element of the URL.

Repeatability

There are many reasons to include multiple 856 fields in records. Following are the most common examples:

- Different access methods (e.g., a document available through HTTP and from an FTP server);
- Different parts of the item are electronic, using $3 to indicate the part (e.g., table of contents accessible in one file and an abstract in another);
- Mirror sites (the same resource is made available at two different locations, often to facilitate access, perhaps internationally);
- Different formats/resolutions (e.g., the ASCII version of an electronic journal vs. the Web page for that journal; postscript and pdf formats, etc.); and
- Related items, using subfield $3 and second indicator value to specify.

Field 856 contains the following elements:

Indicators

First Indicator (Access Method). The first indicator contains information about access method to the resource and has values defined for Email, FTP, Telnet, Dial-up, and HTTP. Access methods without defined values may contain a first indicator value 7 with the method indicated in $2.

Second Indicator (Relationship). A second indicator is provided to show the relationship between the information in field 856 and the resource described in the record. This may be used for generation of a display constant.

First	Access Method	Second	Relationship
#	No information provided	#	No information provided
0	Email	0	Resource
1	FTP	1	Version of resource
2	Remote login (Telnet)	2	Related resource
3	Dial Up	8	No display content generated
4	HTTP		
7	Method specified in subfield $2		

- *First - Access Method*

  It is a value that defines the access method to the electronic resource. If the resource is available by more than one access method, the field is repeated. When recording a URL in subfield $u, the value corresponds to the access method (URL scheme), which is also the first element in the string.

  - # - No information provided

    Indicates that no information about access method is provided. This value is used when subfield $u contains a URN and there is no URL recorded. When subfield $u contains a URL and the subfield is repeated with a URN, the indicator value for the appropriate access method of the URL is given.

  - 0 - Email

    Indicates that access is through the Mail Transfer Protocol (MAILTP).

  - 1 - FTP

    Indicates that access is through the File Transfer Protocol (FTP).

  - 2 - Remote login (Telnet)

    Indicates that access is through remote login using an application such as Telnet.

  - 3 - Dial-up

    Indicates that access to the electronic resource is through a conventional telephone line (*dial-up*).

  - 4 - HTTP

    Indicates that access to the electronic resource is through the Hypertext Transfer Protocol.

  - 7 - Method specified in subfield $2

- *Second-Relationship*

  A value that identifies the relationship between the electronic resource at the location identified in field 856 and the item described in the record as a whole. Subfield $3 is used to provide further information about the relationship if it is not a one-to-one relationship.

  - # - No information provided

  - 0 - Resource indicates that the electronic location in field 856 is for the same resource described by the record as a whole. In this case, the item represented by the bibliographic record is an electronic resource. If the data in field 856 relates to a constituent unit of the resource represented by the record, subfield $3 is used to specify the portion(s) to which the field applies.

  - 1 - Version of resource
    Indicates that the location in field 856 is for an electronic version of the resource described by the record. In this case, the item represented by the bibliographic record is not electronic but an electronic version is available. If the data in field 856 relates to a constituent unit of the resource represented by the record, subfield $3 is used to specify the portion(s) to which the field applies.

  - 2 - Related resource
    Indicates that the location in field 856 is for an electronic resource that is related to the item described by the record. In this case, the item represented by the bibliographic record is not the electronic resource itself. Subfield $3 can be used to further characterize the relationship between the electronic item identified in field 856 and the item represented by the bibliographic record as a whole.

  - 8 - No display content generated

Subfield Codes:

$a Host name (R)
$b Access number (NR)
$c Compression information (R)
$d Path (R)
$f Electronic name (R)
$g Uniform Resource Name (R)
$h Processor of request (NR)
$I Instruction (R)
$j Bits per second (NR)
$k Password (NR)
$l Logon (NR)
$m Contact for access assistance (R)
$n Name of location of host (NR)
$o Operating system (NR)
$p Port (NR)
$q Electronic format type (NR)
$r Settings (NR)
$s File size (R)
$t Terminal emulation (R)
$u Uniform Resource Locator (R)
$v Hours access method available (R)
$w Record control number (R)
$x Nonpublic note (R)
$z Public note (R)
$2 Access method (NR)
$3 Materials specified (NR)
$6 Linkage (NR)
$8 Field link and sequence number (R)
NR = Non-repeatable sub-field
R = Repeatable sub-field

The most commonly used subfields are as follows:

Subfield $u = [HTTP URL]
Subfield $2 = Access method when first indicator is 7
Subfield $3 = data specifying what URL refers to, if applicable
Subfield $z = Public note.

Commonly used Subfields

- **$u:** Uniform Resource Identifier (URI) provides standard syntax for locating an object using existing Internet protocols. Field 856 is structured to allow for the creation of a URL from the concatenation of other separate 856 subfields. Subfield $u may be used instead of those separate subfields or in addition to them. Subfield $u may be repeated only if one location of the digital object has multiple identifiers (URIs). The field is repeated if the digital object has multiple locations.
- **$3 (Materials Specified)**: Subfield $3 is used to specify to what portion or aspect of the resource the electronic location and access information applies. $3 is used to indicate a portion of the resource is electronic. For example: $3table of contents; $3abstract OR to indicate a related electronic resource which is linked to the record. Example: $3scanned image of photograph

- **$z (Public Note)**: Subfield $z may be used for any additional notes about the electronic resource at the specified location. Examples include subscription information or access restrictions. Example for a mail to URL: $zaccess only through password no guest logins OR $zspecify the desired file format for attachments

Examples

Example 1: Information resource by *remote login to DRTC server* is represented using field 856 as shown below:

8562#$utelnet://drtc.isibang.ac.in$nIndian Statistical Institute, India

Here the descriptive elements are given for a remote login resource where the IP address of the DRTC server is 'drtc.isibang.ac.in' and the actual address is 'Indian Statistical Institute, India'.

Analysis of the description given above:

Field Tag:	856		
Indicators:	2	Access method	telnet
	Blank (#)	Relation	not specified
Subfields:	u	URL	
	n	name of location of the host	

Subfields delimiters: $

Example 2: If the resource is a *downloadable archive* on DRTC server it may be represented as below:

8561#$uftp://drtc.isibang.ac.in/software/downloads/ida.zip$cdecompress with winzip$madmin@ drtc.isibang.ac.in$nIndian Statistical Institute, Bangalore, India$oWindows$s1.52MB

Example 3: The resource *DRTC website* may be described as shown here:

8564#$uhttp://drtc.isibang.ac.in$madmin@drtc.isibang.ac.in$Indian Statistical Institute, Bangalore, India

Example 4: Mirror Sites are sites where information from one is 'copied' or mirrored in the other for ease of access and downloading. Mirror site is one instance where field 856 is a repeated field. An example is shown below:

8561#$uftp://drtc.isibang.ac.in/software/downloads/ida.zip$uftp://isical.ac.in/drtc/pub/ida.zip$madmin@drtc.isibang.ac.in$Indian Statistical Institute, Bangalore, India$oWindows$s1.52MB

8561#$uftp://isibang.ac.in/drtc/pub/ida.zip$madmin@drtc.isibang.ac.in$Indian Statistical Institute, Bangalore, India$oWindows$s1.52MB

Example 5: Link to a subset of the bibliographic item: HTTP URL

8564#$3Table of contents $uhttp://drtc.isibang.ac.in/Seminar/digilib/toc.html

9.2.3 Procedure of Electronic Resource Description

You already have an idea about ISBD (ER). ISBD (ER) forms the basis of electronic resource description for many standard cataloguing codes and content designator schemas. Let us examine encoding of web resources by using MARC

21 content designator standard and on the basis of corresponding AACR 2 rules. We will concentrate only on major areas of description. Note that discussion headings are arranged in the style MARC21 Tag, Description of Tag (Corresponding Rule Number of AACR2).

Fixed Fields

Use the fixed fields for monographs (type a) or serials (type s), depending on the nature of the resource. For most Web sites, use date type "m" (multiple), with an open date to reflect the fact that the resource is subject to updating. Add an 006 secondary fixed field to reflect the "electronic resource" nature of the title and an 007 for the physical characteristics such as sound, color, etc.

Example of fixed field encoding

006 Type m = Electronic resource; Audn = Intended audience (blank = not specified) File = File type (d=document; j=online system or service; m=combination); Gpub = Government publication (f for federal, s for state, etc.)

007 Physical description fixed field
$a category = c, computer file
$b specific material designation= r, remote resource
$d color (most Web pages will be m, multicolor)
$e dimensions (for Web pages use n, N/A)
$f sound (blank=no sound, a=sound)
$g image bit depth (u=unknown)
$h file format (a=one format, m=multiples, u=unknown)
$i quality assurance targets (u=unknown)
$j antecedent/source (u=unknown)
$k level of compression (a=uncompressed, u=unknown)
$l reformatting quality (a=access, n=N/A, p=preservation, r=replacement, u=unknown)

Variable Fields

Before discussing different major variable fields, we should pay attention to sources of information for describing electronic resources. Usually the main source of information will be the title screen; however, according to the 2002 revision of AACR2 Chapter 9, the entire resource serves as the chief source of information for an electronic resource. This means it will not be necessary to bracket information that appears in locations other than the title screen. In addition to the home page screen, look carefully at other pages and at the document's HTML source page to see if there is additional information you need to add to the record. Relevant information can appear anywhere on a Web site. (AACR 9.0B1-2).

1xx Author field. (AACR 21.4B1, 21.1A1)

Many Web sites will not have a known author. If the resource is the official site for a corporate body or government agency, consider it to be the author (21.4B1). Personal creators of Web sites should not be given as authors in a 100 field unless the site being cataloged is that person's own site (21.1A1-variation).

Example

Web site being cataloged: Mississippi Dept. of Vital Statistics home page; Author of site content listed in document source: Sheryl Smith

Main entry: 110 1 Mississippi. $b Dept. of Vital Statistics.

245 10 Mississippi Department of Vital Statistics $h [electronic resource] / $c [Sheryl Smith].

Added entry: 700 1 Smith, Sheryl.

245 Title proper. (AACR 9.1C1)

Usually the title will be taken from the home page screen. Add the general material designation (GMD) "electronic resource" in 245 $h, immediately after the title proper. Transcribe statements of responsibility as indicated in 1.1F. Rule 9.1F2 instructs catalogers to add a word or short phrase to the statement of responsibility if the relationship between the title and the name in the statement of responsibility is not clear.

Example

245 00 Mississippi statistical summaries $h [electronic resource] / $c [maintained by] David Reynolds.

246 Variant title. (AACR 9.7B4)

Examine the document source to see if there is an alternative title that needs to be entered in a 246 field. You may also add a 246 for a title that does not actually appear on the site but which patrons may likely to look under, such as a phrase containing a widely used corporate acronym.

Example: 245 10 Mississippi Department of Transportation $h [electronicresource].
246 3 MDOT home page

256 Type of resource. (AACR 9.3B1)

AACR2 allows three descriptions: Electronic data, Electronic programs, and Electronic data and programs. Web sites consisting primarily of text should be designated Electronic data. Sites that have an interactive component such as a search engine are designated Electronic data and programs. Give extent of resource in parentheses if available (number of files, records, bytes, etc.)

Example: 256 Electronic data (1 file: 13 megabytes)

260 Publication information. (AACR 9.4B2)

Rule 9.4B2 says to consider all remote access electronic resources to be published. The publication information is usually taken from the contact information at the bottom of the home page, where the corporate body gives its physical address. Web pages authored by individuals may not have publication information. If a date is given (e.g. "c2001"), give as an open date in 260 $c (i.e., c2001-) (variation on 9.4F1). In case of multiple copyright dates which apply to various aspects of the resource, use only the latest copyright date (9.4F4).

Example: 260 $a Washington, DC : $b Worldnews Online, $c [1995-

270 Contact address.

Most corporate Web sites include a contact address at the bottom of the home page. Give this address in a 270 field in the following form:

Example: 270 $i Contact address: $a email@host.domain

300 Physical descriptions. (AACR 9.5)

Do not use for Web pages; electronic resources are considered non-physical objects. (9.5, footnote). Some libraries' local practice require giving information such as "1 Web site" or "1 Web resource" in the 300 field.

5xx Notes

516 Nature and scope of the resource. (AACR 9.7B1a)

For Web pages, use the phrase "World Wide Web resource." Information currently entered in the 256 field may eventually be included in the 516.

538 System requirements. (AACR 9.7B1b)

The type of hardware and software needed to access the resource. Include mention of any programs needed to access specific files on the site.

Example: 538 System requirements: World Wide Web browser with Internet connectivity; Adobe Acrobat Reader required for abstracts and full text.

538 Mode of access. (AACR 9.7B1c)

Add another 538 field to indicate mode of access.

Example: 538 Mode of access: World Wide Web.

500 General note: source of title proper. (AACR 9.7B3)

Always give the source of the title proper. For Web pages, also give the date the resource was viewed (9.7B22). These can be combined into a single note.

Example: 500 Title from home page screen, viewed Feb. 19, 2005.

500 General note: title variations. (AACR 9.7B4)

Many title variations can be expressed through the 246 field; give others in a 500 note, e.g.

246 1 $i Title from document source: $a MLA home page
246 1 $i Parallel title in HTML header: $a Ressources minérales du Canada

But for a title that does not apply to the resource you are cataloguing:

500 Companion site: Butterflies of the United States.

500 General note: physical description. (AACR 9.7B10)

Give information about physical details that may affect use of the item.

Example: 500 Includes sound clips for selected text.

521 Audience. (AACR 9.7B14)

Make a note of the intended audience if the information is stated in the resource.

Example: 521 For use by qualified medical practitioners only.

530 Other formats. (AACR 9.7B16)

Use if the content has been issued in another format (e.g., print or microform).

520 Summary. (AACR 9.7B17)

Give a brief summary that indicates the nature of the site and its contents.

Example: 520 Features the Mississippi Dept. of Mental Health. Site includes statement of philosophy, list of state mental health facilities, and links to other mental health resources.

505 Contents. (AACR 9.7B18)

List the parts of a resource. Optional for Web pages, since section headers can change frequently.

Example: 505 0 K-12 requirements – Reading readiness – Remedial courses.

506 Restrictions on access. (AACR 9.7B20)

Use to indicate any restrictions on accessing the site.

Example: 506 Registration and password required to access search engine.

6xx Subject headings.

Include topical headings; corporate or personal headings if the site is about a corporate entity or a person; and, if desired, a genre heading (655) for Computer network resources.

7xx Corporate/personal author added entries.

Give as needed. Remember to check source document for personal or corporate contributors to the site.

856 URL.

For Web sites, the first indicator will be "4," indicating an http address, and the second will be "0," indicating that you are cataloging the resource itself (not a related resource or a reproduction).

Example: 856 40 $u http://www.lib.usm.edu

Full Examples of Online Information Resource with World Wide Web Access according to AACR2 and MARC21 are given below as illustrations:

AACR 2 Examples

Electronic Beowulf [Electronic resource]. - Electronic interactive multimedia. - [Great Britain?] : Electronic Beowulf Project, cop. 1995.
Mode of access: World Wide Web. URL: http://portico.bl.uk/access/electronic-beowulf.html.
Title from title screen.
Digitised images developed by the British Library with Kevin Kiernan and PaulSzarmach.
Summary: Introduction to the Electronic Beowulf Project including images of the manuscript.

Callaloo [Electronic resource]. - Electronic journal. - Baltimore (MD) : Johns Hopkins University Press, cop. 1995-
Quarterly.
18.1 (winter 1995)-
Mode of access: Internet via World Wide Web. URL: http://muse.jhu.edu/journals/callaloo.
Title from title screen.
Also available in a print ed.

MARC21 Example

Sl No.	MARC Tag	**Description of Resource** (along with MARC subfields)
1	040	UOK $c UOK
2	007	$a c $b r $d u $e n $f u
5	245	$a Worldnews online $h [computer file].
6	246	$a World news online
7	256	$a Computer online service.
8	260	$a Washington, DC : $b Worldnews Online, $c [1995-
9	538	$a Mode of access: Internet.
10	500	$a Title from title frame.
11	520	$a "WorldNews OnLine is a service that brings newspapers and news services from around the world to a global community of multi-lingual people who need news from far away places ... our on-line publications have full text of each day's edition on the same day it appears in its local market. The papers may be accessed via any World Wide Web client that supports user authentication."
12	650	$a Newspapers $x Databases.
13	856	$u http://worldnews.net $2 http

Self Check Exercises

1) Explain the elements of a MARC record?

2) What are the contents of the MARC-856 field?

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

.....

9.3 DUBLIN CORE METADATA INITIATIVE (DCMI)

The Dublin Core Metadata Initiative (DCMI) is an organization dedicated to fostering the widespread adoption of interoperable metadata standards and promoting the development of specialized metadata vocabularies for describing resources to enable more intelligent resource discovery systems [DCMI].

The Dublin Core Metadata Element Set (DCMES) was the first metadata standard deliverable out of the DCMI. DCMES provides a semantic vocabulary for describing the 'core' information properties, such as 'Description' and 'Creator' and 'Date'.

Dublin Core metadata is used to supplement existing methods for searching and indexing Web-based metadata, regardless of whether the corresponding resource is an electronic document or a 'real' physical object.

Web pages are one of the most common types of resources to utilise the Dublin Core's descriptions, usually within HTML's meta tags. However increasingly there are many digital archives of physical objects that are starting to make use of the Dublin Core. Dublin Core metadata is often stored as name-value pairs within META tags, which are placed within the HEAD elements of an HTML document. However, it can also be located in an external document or loaded into a database enabling it to be indexed and manipulated from within a proprietary application.

9.3.1 Dublin Core and HTML

The Dublin Core [DC1] is a small set of metadata elements for describing information resources. The following sections explain how these elements are expressed using the META and LINK tags of HTML [HTML4.0].

HTML is not case-sensitive, so it does not matter if you enter the DC elements in either CAPS or small letters. But, it is advisable to be consistent in this matter, in case the metadata needs to be transported to an XML file, since XML is case-sensitive. For example, in XML *'author'* is different from *'Author'*. Though HTML is currently in wide use, once standardized, eXtensible Markup Language [XML] in conjunction with the Resource Description Framework [RDF] promises to become a significantly more expressive means of encoding metadata.

9.3.1.1 META Tag

The META tag of HTML is designed to encode a named metadata element. Each element describes a given aspect of a document or other information resource. For example, this tagged metadata element,

```
<meta name  = "DC.Creator"
        content = "Simpson, Homer">
```

says that Homer Simpson is the Creator, where the element named Creator is defined in the DC element set. In the more general form,

```
<meta name   = "PREFIX.ELEMENT_NAME"
        content = "ELEMENT_VALUE">
```

The capitalised words are meant to be replaced in actual descriptions; thus in the example,

```
ELEMENT_NAME   is: Creator
ELEMENT_VALUE is: Simpson, Homer
PREFIX             is: DC
```

Within a META tag the first letter of a Dublin Core element name is capitalized. DC places no restriction on alphabetic case in an element value and any number of META tagged elements may appear together, in any order. More than one DC element with the same name may appear, and each DC element is optional. The next example is a book description with two authors, two titles, and no other metadata.

```
<meta name  = "DC.Title"
        content = "The Communist Manifesto">
```

```
<meta name  = "DC.Creator"
      content = "Marx, K.">
<meta name  = "DC.Creator"
      content = "Engels, F.">
<meta name  = "DC.Title"
      content = "Capital">
```

The prefix 'DC' precedes each Dublin Core element encoded with META, and it is separated by a period (.) from the element name following it. Each non-DC element should be encoded with a prefix that can be used to trace its origin and definition; the linkage between prefix and element definition is made with the LINK tag, as explained in the next section. Non-DC elements, such as Email from the A-Core [AC], may appear together with DC elements, as in

```
<meta name  = "DC.Creator"
      content = "Da Costa, Jos&eacute;">
<meta name  = "AC.Email"
      content = "dacostaj@peoplesmail.org">
<meta name  = "DC.Title"
      content = "Jesse &#34;The Body&#34; Ventura-A Biography">
```

This example also shows how some special characters may be encoded. The author name in the first element contains a diacritic encoded as an HTML character entity reference – in this case an accented letter E. Similarly, the last line contains two double-quote characters encoded so as to avoid being interpreted as element content delimiters.

9.3.1.2 LINK Tag

The LINK tag of HTML may be used to associate an element name prefix with the reference definition of the element set that it identifies. A sequence of META tags describing a resource is incomplete without one such LINK tag for each different prefix appearing in the sequence. The previous example could be considered complete with the addition of these two LINK tags:

```
<link rel  = "schema.DC"
      href = "http://purl.org/DC/elements/1.0/">
<link rel  = "schema.AC"
      href = "http://metadata.net/ac/2.0/">
```

In general, the association takes the form

```
<link rel  = "schema.PREFIX"
      href = "LOCATION_OF_DEFINITION">
```

where, in actual descriptions, PREFIX is to be replaced by the prefix and LOCATION_ OF_DEFINITION by the URL or URN of the defining document. When embedded in the HEAD part of an HTML file, a sequence of LINK and META tags describes the information in the surrounding HTML file itself. Here is a complete HTML file with its own embedded description.

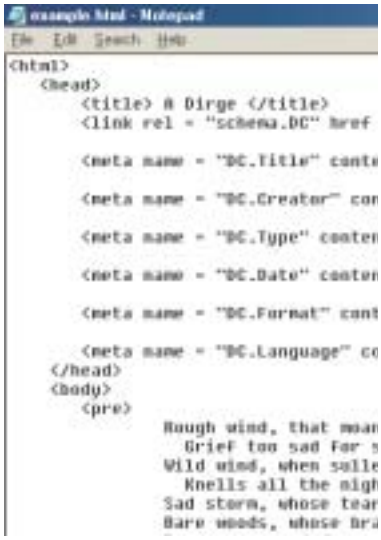

Fig. 9.1: Dublic Core Meta Tags

9.3.2 Dublin Core Elements

The 15 core elements of Dublin core are – Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights. The latest additions are the Audience and rightsHolder elements [DCMI]. These 17 elements are described below:

1) **Title**

	Title
Label:	Title
Definition:	A name given to the resource.
Comment:	Typically, a Title will be a name by which the resource is formally known.

Examples:

<meta name = "DC.Title" content = "Introduction to Philosophy">

<meta name = "DC.Title" content = "Prolegomena to Library Classification">

2) **Creator**

	Creator
Label:	Creator
Definition:	An entity primarily responsible for making the content of the resource.
Comment:	Examples of a Creator include a person, an organisation, or a service. Typically, the name of a Creator should be used to indicate the entity.

Examples:

<meta name = "DC.Creator" content = "Ranganathan, S.R.">

<meta name = "DC.Creator" content = "Russel, Betrand">

3) **Subject**

Subject	
Label:	Subject and Keywords
Definition:	The topic of the content of the resource.
Comment:	Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.

Examples:

<meta name = "DC.Subject" content = "Bamian Buddha">

<meta name = "DC.Subject" scheme = "MESH" content = "Carcenoma">

<meta name = "DC.Subject" scheme = "ddc" content = "201.23">

4) **Description**

Description	
Label:	Description
Definition:	An account of the content of the resource.
Comment:	Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.

Examples:

<meta name="DC.Description" lang="en" content = "The author presents a tutorial introduction to Perl programming examples with extensive examples on regular expressions. He also deals with scripting in Perl">

5) **Publisher**

Publisher	
Label:	Publisher
Definition:	An entity responsible for making the resource available
Comment:	Examples of a Publisher include a person, an organisation, or a service. Typically, the name of a Publisher should be used to indicate the entity.

Examples:

<meta name = "DC.Publisher" content = "Wrox">

<meta name = "DC.Publisher" content = "Dell Computers">

<meta name = "DC.Publisher" content = "MIT Press">

6) **Contributor**

Contributor	
Label:	Contributor
Definition:	An entity responsible for making contributions to the content of the resource.
Comment:	Examples of a Contributor include a person, an organisation, or a service. Typically, the name of a Contributor should be used to indicate the entity.

Examples:

<meta name = “DC.Contributor” content = “Laxman, R.K.”>

<meta name = “DC.Contributor” content = “Knuth, Donald”>

7) **Date**

	Date
Label:	Date
Definition:	A date associated with an event in the life cycle of the resource.
Comment:	Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and follows the YYYY-MM-DD format.

Examples:

<meta name = “DC.Date” content = “1990”>

<meta name = “DC.Date” content = “1990-05-14”>

<meta name = “DC.Date.Created” content = “1990-05-14”>

<meta name = “DC.Date.Available” content = “1998-05-21”>

<meta name = “DC.Date.Valid” content = “1998-05-28”>

<meta name = “DC.Date.Accepted” content = “1998-12-02T16:59”>

<meta name = “DC.Date.Issued” scheme = “W3CDTF” content = “1998-05”>

8) **Type**

	Type
Label:	Resource Type
Definition:	The nature or genre of the content of the resource.
Comment:	Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the DCMI Type Vocabulary [DCMITYPE]). To describe the physical or digital manifestation of the resource, use the Format element.

Examples:

<meta name = “DC.Type” content = “drama”>

<meta name = “DC.Type” content = “software”>

<meta name = “DC.Type” content = “software program source code”>

<meta name = “DC.Type” content = “interactive video game”>

<meta name = “DC.Type” scheme = “DCMI Type” content = “dataset”>

<meta name = “DC.Type” content = “web home page”>

<meta name = “DC.Type” content = “web bibliography”>

<meta name = “DC.Type” content = “painting”>

9) **Format**

format	
Label:	Format
Definition:	The physical or digital manifestation of the resource.
Comment:	Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [MIME] defining computer media formats).

Examples:

<meta name = "DC.Format" content = "text/xml">

<meta name = "DC.Format" scheme = "IMT" content = "text/xml">

<meta name = "DC.Format" scheme = "IMT" content = "image/jpeg">

<meta name = "DC.Format" content = "video/mpeg">

<meta name = "DC.Format" content = "unix tar archive, gzip compressed">

10) **Identifier**

Identifier	
Label:	Resource Identifier
Definition:	An unambiguous reference to the resource within a given context.
Comment:	Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Example formal identification systems include the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).

Examples:

<meta name = "DC.Identifier" scheme= "URI"
content = "http://www.google.com/">

<meta name = "DC.Identifier"scheme = "ISBN" content = "1-56592-149-6">

11) **Source**

Source	
Label:	Source
Definition:	A reference to a resource from which the present resource is derived.
Comment:	The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

Examples:

<meta name = "DC.Source" content = "Bernard Shaw's Saint Joan">

<meta name = "DC.Source" content = "http://abc.org/xyz/">

12) **Language**

	Language
Label:	Language
Definition:	A language of the intellectual content of the resource.
Comment:	Recommended best practice is to use RFC 3066 [RFC3066], which, in conjunction with ISO 639 [ISO639], defines two- and three-letter primary language tags with optional subtags. Examples include "en" or "eng" for English, "akk" for Akkadian, and "en-GB" for English used in the United Kingdom.

Examples:

<meta name = "DC.Language" content = "en">

<meta name = "DC.Language" scheme = "rfc3066" content = "en-US">

<meta name = "DC.Language" content = "es">

13) **Relation**

	Relation
Label:	Relation
Definition:	A reference to a related resource.
Comment:	Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

Examples:

<meta name = "DC.Relation.IsPartOf" content = http://foo.bar.org/abc/proceedings/1998/">

<meta name = "DC.Relation.IsFormatOf"
content = "http://foo.bar.org/cd145.sgml">

<meta name = "DC.Relation.IsVersionOf"
content = "http://foo.bar.org/draft9.4.4.2">

<meta name = "DC.Relation.References"
content = "urn:isbn:1-56592-149-6">

<meta name = "DC.Relation.IsBasedOn"
content = "Shakespeare's Romeo and Juliet">

14) **Coverage**

	Coverage
Label:	Coverage
Definition:	The extent or scope of the content of the resource.
Comment:	Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and that, where appropriate, named places or time periods be used in preference to numeric identifiers such as sets of coordinates or date ranges.

Examples:

```
<meta name    = "DC.Coverage"
      content = "US civil war era; 1861-1865">
<meta name    = "DC.Coverage"
      content = "Columbus, Ohio, USA; Lat: 39 57 N Long: 082 59 W">
<meta name    = "DC.Coverage.Jurisdiction"
      content = "Commonwealth of Australia">
```

15) **Rights**

Rights	
Label:	Rights Management
Definition:	Information about rights held in and over the resource.
Comment:	Typically, a Rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions can be made about the status of these and other rights with respect to the resource.

Examples:

```
<meta name    = "DC.Rights"
      lang    = "en"
      content = "Copyright Acme 1999 - All rights reserved.">
<meta name    = "DC.Rights"
      content = "http://foo.bar.org/cgi-bin/terms">
```

16) **Audience**

Audience	
Label:	Audience
Definition:	A class of entity for whom the resource is intended or useful.
Comment:	A class of entity may be determined by the creator or the publisher or by a third party.

Examples:

```
<meta name = "DC.Audience"
      content = "software developers">
<meta name = "DC.Audience"
      content = "post graduate students">
```

17) **Rights Holder**

	rightsHolder
Label:	Rights Holder
Definition:	A person or organization owning or managing rights over the resource.
Comment:	Recommended best practice is to use the URI or name of the Rights Holder to indicate the entity.

Examples:

```
<meta name    = "DC.rightsHolder"
      content = "Disney Productions">
<meta name = "DC.rightsHolder"
      content = "Warner Brothers Limited">
```

9.3.3 Search Engines Supporting the Dublin Core

Several commercial and non-commercial search engines index Dublin Core META elements with just a little configuration. A recent inquiry on the dc-general mailing list produced this list:

- Ultraseek
- Swish-E
- Microsoft's Index Server
- Blue Angel Technologies MetaStar
- Verity Search 97 Information Server

To get a good overview of what software is available see 'Search Tools' <http://www.searchtools.com/> and 'Search Engine Watch' <http://searchenginewatch.com/>. The well-known 'All the Web' search engines including AltaVista, Yahoo!, HotBot, etc. tend to avoid using the information found in meta elements in their indexing. This is because, unless the pages are from guaranteed 'trusted' servers, the meta information is commonly used by unscrupulous content-providers for spamming, to mislead the indexes into giving web-pages a misleading rating.

Self Check Exercise

3) What is DCMI? How is Dublin Core Metadata different from other cataloguing schemas?

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

9.4 TEXT ENCODING INITIATIVE (TEI)

The Text Encoding Initiative (TEI) was founded in 1987 to develop guidelines for encoding machine-readable texts of interest to the humanities and social sciences. The TEI is an international and interdisciplinary standard that helps libraries, museums, publishers, and individual scholars represent all kinds of literary and linguistic texts for online research and teaching, using an encoding scheme that is maximally expressive and minimally obsolescent [Martin Mueller].

The TEI markup language is a form of SGML or Standard Generalized Markup Language that was developed in the seventies by Charles Goldfarb at IBM, and it has become an international standard (ISO 8879). It became popular when HTML, a subset of it was developed by Tim Berners-Lee at the CERN labs in Geneva and within a few years became the language of the Internet.

Initially, TEI was as a cooperative research effort by three scholarly societies (the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing), and funded solely by substantial research grants from the US National Endowment for the Humanities, the European Union, the Canadian Social Science Research Council, the Mellon Foundation, and others. In December 2000, after a year's negotiation, a new non-profit corporation called the TEI Consortium was set up to maintain and develop the TEI standard.

9.4.1 Basics of TEI Markup Language

There are two different versions of the TEI markup language. The full version of the TEI markup language includes about 450 different elements to satisfy all manner of scholarly needs in the humanities. Teixlite (the XML version of TEI) contains about 150 elements. It is largely based on the experience of marking up the wide variety of electronic texts held in the Oxford Text Archive (OTA) [www.tei-c.org].

All TEI-conformant texts contain (a) a TEI header (marked up as a <teiHeader> element) and (b) the transcription of the text proper (marked up as a <text> element). The TEI header has four parts: a bibliographic description of the machine-readable text, a description of the way it has been encoded, a non-bibliographic description of the text (a text profile), and a revision history.

The text may have an optional *front* or *back*. In between is the *body* of the text, the overall structure looks like this:

```
<TEI.2>
  <teiHeader> [ TEI Header information ] </teiHeader>
  <text>
    <front> [ front matter ... ]  </front>
    <body> [ body of text ... ]  </body>
    <back> [ back matter ... ]  </back>
  </text>
</TEI.2>
```

Example:

```
<TEI.2>
     <teiHeader>[header content]</teiHeader>
     <text>
```

```
<front>[front content]</front>
<body>
  <head>My novel</head>
  <div type="chapter" n="1">
    <head>Chapter One</head>
    <p>It was a dark and stormy night</p>
  </div>
 </body>
<back>[back content]</back>
</text>
</TEI.2>
```

In the above example,

1) The <head> element contains headlines or section titles
2) The element contains sections of a document beyond the paragraph level. elements can nest within each other.

The <p> element contains paragraphs as the basic discursive unit. Paragraphs cannot nest within each other.

9.4.1.1 Grouping Elements

In HTML parlance, there is a distinction between block elements and inline elements. Block elements are the kinds of things that begin a new line, whereas inline elements occur within a line. The practical value of this distinction shows how deeply formatting habits shape our thinking. The full version of the TEI has a formal and quite complex system of classifying element groups, but for the purpose of easy understanding some of the elements are grouped as follows:

a) Major structuring elements: TEI.2, teiHeader, text, body, front, back, div, head
b) Paragraph level or 'block' elements: p, cit, q, l, lg, sp.speaker, nd related elements that mark side text (note stage) or special forms of text segmentation (s seg)
c) Lists, tables, and figures: list, item, table, row, cell, figure, figDesc
d) Phrase level or 'inline' elements: date, emph, foreign, name, num, soCalled, term, title
e) Milestone elements: milestone, pb, lb
f) Referring and linking elements: ref, rs, ptr, xref, xptr
g) Bibliographical elements: bibl, author, editor, publisher, respStmt, resp pubPlace
h) The immediate child elements of the header: fileDesc, profileDesc, revisionDesc
i) The child elements of fileDesc: titleStmt, publicationStmt, sourceDesc
j) Other header elements: langUsage, language.

The nesting rules for these elements are also fairly complex, but the following steps can be followed [Mueller]:

1) The body of a document is segmented by elements
2) The elements are segmented by the elements <p>, <q>, <l>, <lg>, and <sp>

3) <p> and similar elements usually contain plain text, parts of which are marked up with phrase or inline elements.

9.4.1.2 Elements at the Paragraph Level

In paragraphs (the <p> element), the units of a poem are identified as lines or stanzas (<l> and <lg>), a passage is identified as a quotation (<q>), or a speech (<sp>), and so forth. The <note> and <stage> elements identify text units that stand in an oblique relationship to the main text. The sentences in a work are identified by the <s> tag or arbitrary segments by <seg>. These last two elements are not really at the paragraph level, but are usefully mentioned in that context.

The pivotal tag here is the <p> tag, and important constraints of the TEI markup language turn on what can contain or be contained by a paragraph:

a) A paragraph element cannot contain paragraph, line, speech, or div elements

b) A paragraph can contain elements such as <q> or <text> that can contain those elements.

This pivotal position of the <p> tag is recognized by the formal classification of element groups in the full TEI, which distinguishes between

a) **chunk** elements, which can occur between paragraphs but not within them.

b) **phrase** elements, which can occur within paragraphs but not between them

c) **inter** elements, which can occur within and between paragraphs, notably the elements <q> (quotation) and <bibl> (bibliography).

9.4.1.3 Attributes

Attributes are used to add specifications to elements. An attribute always exists as a 'key-value pair' in which the name of the attribute is followed by an equal sign, which is followed by the value of the attribute in quotation marks. The attributes values are declared in the start tag of an element with no punctuation but a blank space between different attributes, as in the following:

```
<div type="chapter" n="1"> . . . . </div>
```

You must not repeat the attribute values in the closing tag.

You cannot just use attributes as you please but must declare them in the DTD in an ATTLIST declaration, which specifies:

a) the name of the attribute

b) the type of data that may appear in the values for the attribute

c) whether it is required, optional, or has a fixed value

In the TEI DTD, every element shares the same four attributes that are known as 'global attributes'. They are id, n, lang, and rend, and they are declared as follows:

```
<!ATTLIST element
id ID #IMPLIED
n CDATA #IMPLIED
lang IDREF #IMPLIED
rend CDATA #IMPLIED
>
```

The *id* attribute is self-explanatory, but its value must be a unique ID value in the document, and the value must begin with a letter of the alphabet. The *n*

attribute may take any alphanumerical value, but it is most commonly used to count lines, stanzas, paragraphs or whatever else you want to number.

The *lang* attribute refers to the language used in an element. Thus, <q lang="FR"> refers to a quotation in French. The value of the *lang* attribute must be an IDREF or reference to an ID that exists in the document. IDREFs are a generic feature of SGML and are used to enforce consistent references. For example, the <sp> element for tagging dialog has a <who> attribute with IDREF values. The parser will return an error message if the value of the attribute is not matched by a speaker ID.

The *<rend>* attribute is used to include information about how a particular element is typographically represented, as in the following example:

<q lang="FR" rend="italics">

9.4.1.4 Lists, Tables and Figures

A list consists of a list element and one or more items as in the following example:

```
<list>
<item>bread</item>
<item> milk</item>
<item>bananas</item>
</list>
```

A TEI table is in some ways like an HTML table, although it uses different element names. The table element contains row elements, which contain cell elements. The coresponding element names in HTML are <table>, <tr> (table row), and <td> (table detail). On the other hand, the content model for a TEI cell element is very similar to that for a paragraph, which means that it cannot contain paragraphs. Unlike an HTML table, a TEI table cannot be repurposed as a layout tool.

The <figure> and <figDesc> elements are used in TEI to refer to and include visual materials in the text.

Example:

```
<TEI.2 lang="en-gb">
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>Alice's Adventures in Wonderland</title>
    <respStmt><resp>Illustrated by</resp> <name>John Tenniel</name></
respStmt>
   <author><name reg="Carroll, Lewis">Lewis Carroll</name></author>
  <editor role="illustrator"><name reg="Tenniel, John">John Tenniel</name></
editor>
  </titleStmt>
 <editionStmt>  <edition n="30">Edition 30</edition>
 </editionStmt>
 <publicationStmt>
  <publisher>Project Gutenberg</publisher>
  <date value="1991-01">January, 1991</date>
  <idno type="etext-no">11</idno>
  <availability>
    <p>This eBook is for the use of anyone anywhere at no cost and
```

```
      with almost no restrictions whatsoever. You may copy it, give it
      away or re-use it under the terms of the Project Gutenberg
      License online at www.gutenberg.org/license</p>
    </availability>
  </publicationStmt>
   <seriesStmt>
    <title level="s">#1 in our series by Lewis Carroll</title>
   <idno type="vol">1</idno>
   </seriesStmt>
    <langUsage>
   <language id="en-gb">British</language>
    </langUsage>
</teiHeader>
<text>
 <front>
  <divGen type="titlepage" />
  <divGen type="pgheader" rend="newpage" />
  <div rend="newdoublepage center">
  <p rend="font-size(150%)">THE MILLENNIUM FULCRUM EDITION 3.0</
p>
  </div> <divGen type="toc" rend="newdoublepage" />
 </front>
 <body rend="newdoublepage">
 <p> Alice was beginning to get very tired of sitting by her
 sister on the bank, and of having nothing to do: once or twice
 she had peeped into the book her sister was reading, but it had
 no pictures or conversations in it, <q>and what is the use of a
 book,</q> thought Alice <q>without pictures or conversation?</q>
 </p>
</body>
</TEI.2>
```

(Source: http://www.tei-c.org/Vault/AI/ai1t03.te1)

9.5 SUMMARY

The electronic information resources, available online, can be effectively described by the Metadata for efficient retrieval. There are a number of metadata schemes that are used to describe the information on the electronic resources. The metadata standards and schemes are developed and/ or recommended by the international agencies to accelerate information exchange and resource sharing over the Internet. Here metadata plays a crucial role in aspects all over the information dissemination process, like, information generation, information storage, information transfer and information retrieval. The Dublin Core Metadata Initiative, the Text Encoding Initiative, and MARC21 – Field 856 – these are the metadata schemes available for describing electronic information resources in general and electronic documents in particular. These three schemes are briefly described in this Unit.

9.6 ANSWERS TO SELF CHECK EXERCISES

1) The MARC21 format is a standard for the representation and communication of bibliographic and related information in machine-readable form.

   A MARC record involves three elements: the record structure, the content designation, and the data content of the record.

- The structure of MARC records is an implementation of national and international standards, e.g., *Information Interchange Format* (ANSI Z39.2) and *Format for Information Exchange* (ISO 2709).
- Content designation, the codes and conventions established to identify explicitly and characterize further the data elements within a record and to support the manipulation of those data, is defined in the MARC21 formats.
- The content or data, of most data elements is defined by standards outside the formats, e.g., *Anglo-American Cataloguing Rules*, *Library of Congress Subject Headings, etc.*

2) Field 856 contains the following elements:

I) Indicators

- First Indicator (Access Method). The first indicator contains information about access method to the resource and has values defined for Email, FTP, Telnet, Dial-up, and HTTP. Access methods without defined values may contain a first indicator value 7 with the method indicated in $2.
- Second indicator (Relationship). A second indicator is provided to show the relationship between the information in field 856 and the resource described in the record. This may be used for generation of a display constant.

II) Subfield codes

The most commonly used subfields are as follows:

- Subfield $u = HTTP URL
- Subfield $2 = Access method when first indicator is 7
- Subfield $3 = Data specifying what URL refers to, if applicable
- Subfield $z = Public note

3) The DCMI, i.e., Dublin Core Metadata Initiative is an organization dedicated to design and development of interoperable and specialized metadata standards for describing resources to enable intelligent resource discovery.

The Dublin Core Metadata Element Set (DCMES) provides a semantic vocabulary for describing the 'core' information properties, such as 'Creator', 'Title', 'Date', etc.

Dublin Core metadata is different from other cataloguing schemas like MARC, etc and it is not meant to replace the existing standards for describing resources. It is used to supplement existing methods for searching and indexing Web-based metadata to increase the interoperability among databases belonging different databases. And also, inspite of the fact that the DCMES is used often to describe Internet resources, it can be used to describe any resource, regardless of whether the corresponding resource is an electronic document or a 'real' physical object.

9.7 KEYWORDS

ANSI : American National Standards Institute is an organisation of American industry groups works with the standards committees of other nations to develop standards to facilitate international trade and telecommunications.

Attribute : Used to describe information which is in some sense descriptive of a specific element occurring but not regarded as part of its content.

Example: <name type="person"> here type is the attribute of the element name.

Attribute Value : The value assigned to a given attribute.

Example: <name type="person"> here type is the attribute of the element name and person is the attribute value

Cataloging record : A bibliographic record (essentially the information on a catalog card), created from following the guidelines in the Anglo-American Cataloging Rules.

DTD : Document Type Definition – The DTD defines the structural rules of a type of document. These rules include a complete list of allowable elements and attributes, special character entities, rules for external files (such as images), as well as the hierarchical structure of all elements. Examples of documents type definitions include TEI, EAD, etc.

Dublin Core Metadata Initiative (DCMI) : Dublin Core Metadata Initiative the body responsible for the ongoing maintenance of Dublin Core. DCMI is currently hosted by the OCLC Online Computer Library Center, Inc., a not-for-profit international library consortium. The work of DCMI is done by contributors from many institutions in many countries. DCMI is a consensus-driven organization organized into working groups to address particular problems and tasks. DCMI working groups are open to all interested parties. Instructions for joining can be found at the DCMI web site under Working Groups (http://dublincore.org)

Element (Dublin Core) : A discrete unit of data or metadata. Example, Title, Creator, Publisher, Date, etc.

Elements (XML) : The technical term used in XML for a textual unit, viewed as a structural component. For example, title, chapter, section, poem, stanza, etc.

Encoding : A process of transferring text and/or data to a searchable electronic medium and organizing it into specific structural and conceptual elements.

HTML : Hypertext Markup Language is the standard text-formatting language for documents on the World Wide Web. HTML text files contain content that is rendered on a computer screen and markup, or tags, that can be used to tell the computer how to format that content. HTML tags can also be used to encode metadata and to tell the computer how to respond to certain user actions, such as a mouse click. For more information, see http://www.w3.org/MarkUp/.

ISO : International Organisation for Standardisation was established in 1947 as a worldwide federation of national standards bodies from some 130 countries.

MARC Cataloguing : The process of recording bibliographic information about an item and then coding that information for the machine. The process of coding cataloged information for machines is called MARC tagging. MARC tagging involves coding the fixed and variable fields, subfield, indicators and tags pertinent to a specific format such as books or maps or serials or computer fields.

Metadata : Information about a publication as opposed to the content of the publication; includes not only bibliographic description but also other relevant information such as its subject, price, conditions of use, etc.

OPAC : Acronym for any Online Public Access Catalog (computer catalog). Often, the O is dropped from OPAC, and computer catalogs are referred to as PACs.

PURL : Persistent Uniform Resource Locator is an approach to the URL permanence problem proposed by OCLC. A PURL is a public alias for a document. A PURL remains stable, while the document's background URL will change as it is managed (e.g., moved) over time. A PURL is created by a Web administrator who is registered as a PURL 'owner' and who maintains a mapping of the PURL to a current and functioning URL.

Retrospective Conversion : The process libraries use to convert their catalogs from card form to machine-readable form, so that bibliographic records can be stored and retrieved in OPACs.

Tag : A Markup Language encoding feature that describes a discrete component of a document, such as a <meta> tag in HTML and <author> tag in XML signifying the author of a document.

XML : eXtensible Markup Language is a subset of Standard Generalized Markup Language (SGML), a widely used international text processing standard. XML is being designed to bring the power and flexibility of generic SGML to the Web, while maintaining interoperability with full SGML and HTML. For more information, see http://www.w3.org/XML/

9.8 REFERENCES AND FURTHER READING

A review of metadata: a survey of current resource description formats. <http://www.ukoln.ac.uk/metadata/desire/overview/rev_ti.htm>

Cataloging Electronic Resources: Olson manual. <http://www.library.cornell.edu/tsmanual/CIRM/Intro.html>

DCMI Metadata Terms. <http://dublincore.org/documents/dcmi-terms/>

Dublin Core Metadata Element Set (DCMES). IETF RFC 2413. <http://www.ietf.org/rfc/rfc2413.txt>

Dublin Core Metadata Initiative (DCMI). <http://dublincore.org>

Guidelines for the Use of Field 856. (2003). <http://www.loc.gov/marc/856guide.html>

Howarth, Lynne C. (2000). *Comments on "AACR2 and its place in the digital world: near-term solutions and long term direction"*. <http://www.loc.gov/catdir/bibcontrol/howarth_paper.html>

Kunze, J. (1999). *Encoding Dublin Core Metadata in HTML*. Request for Comments: 2731, December 1999. (Dublin Core examples are extensively borrowed from the RFC).

MARC21 formats: background and principles. <http://lcweb.loc.gov/marc/96principl.html>

Milstead, Jessica and Feldman, Susan. (1999). Metadata cataloguing by any other name. *Online*, January 1999.

Mueller, Martin. *A very gentle introduction to the TEI.* <http://www.tei-c.org/Sample_Manuals/mueller-main.htm>

Persistent Uniform Resource Locator (PURL). <http://purl.oclc.org>

Text Encoding Initiative (TEI) <http://www.tei-c.org>