METADATA HARVESTING

Sanat Bhattacharya

Abstract

One of the key challenges facing information managers today is the need to inter-relate different sources and types of Information, whether it be in an internet search across a range of resources with different formats, data structures and description standards or an e-commerce system that needs to exchange data between proprietary applications in order to complete a transaction. Understanding the structure of the data allows this to occur and metadata is the means by which this happens. Using metadata to record data about Information sources allows an initial assessment of compatibility and provides an avenue for merging information or for exchanging Information between systems. In other wards the concept of "interoperability" has become a major theme for information managers and ultimately for users.

Keywords : Metadata, Dublin Core

1. Introduction

Metadata is important in the information society. The term metadata probably first appeared in the 1960s and was adopted by the GIS specialists, database developers, statisticians and latterly, in the 1990s, by the web community.

The term metadata 'data about data', i.e. a set of information which remains in same intentional hierarchical relationship with another set of information.[3]So it is a summary of data about some other data. Another concept is that it is machine understandable information for the web. Presently the term refers to any data used to aid the identification, description and location of networked information. For example, a metadata system common in libraries the library catalogue – contains a set of metadata records with elements that describe a book or other library item: author, title, data of creation or publication, subject coverage, and the call number specifying location of the item on the shelf.

The linkage between a metadata record and the resource it describes may take one of two forms:

a) elements may be contained in a record separate from the item, as in the case of the library's catalogue record; or

b) the metadata may be embedded in the resource itself

Examples of embedded metadata that is carries along with the resource itself include the Cataloguing In Publication (CIP) data printed on the verso of a book's title page; or the TEI header in an electronic text. Many metadata standards in use today, including the Dublin Core standard.

1. What metadata is?

Metadata enhances retrieval performance – Metadata can improve retrieval by establishing a context for individual descriptors. For example the word 'Green' in the Creator or Author field indicates the name of an individual, where as 'Green' in the title of a document may be a subject retrieval term.

Metadata provides a way of managing digital objects : Many software packages use metadata as a way of managing electronic resources, whether it is for records retention schedules or for digital preservation.

Metadata can help to determine the authenticity of data : Metadata provides an audit trail to establish ownership and authenticity of a digital object such as an electronic document or image.

Metadata is the key to interoperability : Interoperability depends on the exchange of metadata between systems to establish the nature of the data being transferred and how it should be handled. E-Commerce is one example of interoperating, where several different proprietary systems may need to exchange data.

Metadata is the future : An increasing number of software and systems suppliers are working to metadata standards or are creating their own proprietary standards for metadata. The growth of e-commerce depends on metadata for exchange of data between applications.[8]

2. The purposes of metadata

The purposes of metadata identified by Day [6] are as follows:

1. Resource description: This is particularly important in organizations that need to describe their information assets. For example, In the USA federal agencies have to make information available via the Government Information Locator Service (GILS).

2. Information retrieval: The cataloguing data usually includes a description of the resource, controlled indexing terms and classification headings. This is a metadata resource and may also 'mine' or 'extract' metadata directly from target website or electronic resources.

3. Management of Information resources: The growth of electronic document and records management (EDRM) systems has resulted from the emerging requirements of larger organizations to manage both paper and electronic documentation effectively. EDRM systems need access to 'cataloguing information' about individual documents in order to manage record life cycles. E.g. – authorship, ownership etc.

4. Documenting ownership and authenticity of digital resources- Metadata provides a way of declaring the ownership of the intellectual content and layout of a document.

5. Interoperability – Metadata acts as an enabler of information and data transfer between systems and as such is a key component in interoperability.

3. Need of metadata

The primary aim of metadata is to improve resources discovery.[10]

- Resource identification and location
- Resource documentation
- Resource selection, evaluation and assessment
- Improving the quality and quantity of search results
- Content reuse

- Efficient content development and archiving
- Protecting intellectual property rights
- Electronic commerce to encode prices, terms of pay, etc.

The other functions of metadata that have been identified by C.Taylor are

- Administrative control
- Security
- Personal information
- Management information
- Control rating
- Rights management
- Preservation, etc.

4. Functions of metadata

Metadata is information attached to any resource in the form of keywords or free text. The information contained I a metadata framework is searchable and therefore aids the identification and retrieval of resources. Metadata helps users both to discover the

existence of information objects and to understand the nature of what they found. Information added to a resource will also help the user to evaluate a resource, make a judgment about a resource, compare it with another resource or assess its suitability for the intended use. An effective metadata policy will involve establishing systems for recording and storing information such as title, date of creation, creator, and subject matter and file format to accompany each resource in the collection.

5. There are five types of metadata

- Administrative: Metadata used in managing and administering information resources, e.g. copyright, acquisition information.
- Descriptive: Metadata used to describe or identify information resources, e.g. controlled vocabularies, user annotations.
- Preservation: Metadata related to the preservation management of information Resources, e.g., physical condition of resources, preservation Actions.
- Technical: Metadata related to how a system functions or metadata behave, e.g., Digitization information such as formats, compression.
- Use: Metadata related to the level and type of use of information resources, e.g. Use and user tracking.

6. Tools for creation:

The encoding allows the metadata to be processed by a computer program. Important schemes are – HTML, SGML, XML, RDF, MARC, MIME, LDAP, and Z39.50 etc.

Humans never see some metadata, because it is transient and used for exchanging of data between systems. Visible examples of metadata range from HTML metatags on web pages to MARC records used for exchanging cataloguing data between library management systems. It can be expressed in a structured language such as XML and may follow guidelines or scheme for particular domains of activity.

MARC 21

100	1	sa Pedley, Paul
245	10	sa Essential law for Information professional / sc Paul Pedley
260		sa London: sb Facet, sc 2003
300		sa xviii, 222p; sc 24 cm

7. Metadata Standards

- AGLS (Australian Government Locator Service)
- ANZLIS(Australia New Zealand Information Council)
- CIMI(Consortium for the Computer Interchange of Museum Information)
- DC(Dublin Core)
- EAD(Encoded Archival Description)
- EDNA(Education Network Australia)
- GILS(Government Information Locator Service)
- TEI(Text Encoding Initiatives)
- VRA(Visual Resource Association)

Among these standards DC is most popular and widely accepted due to its compatibility with almost all kinds of E-Sources.

The elements of DC are simple which can be identified easily and describe the physical format and intellectual content of the resources. All elements are optional and repeatable and syntax dependent. Using qualifiers such as thesaurus can modify the elements. The DC consists of 15 core elements, and each of these can be extended by the use of the scheme and type qualifiers.

7.1 Dublin core metadata element set

1. Title – Title of the resource given by the author / creator
2. Creator – Author / Creator person(s), organization(s) for the intellectual content of the resource.
3. Subject- Subject, keyword. The topic, keywords, phrases that describe the resource.
4. Description – Annotation, Abstract, etc.
5. Publisher – Publisher(Person or institution)
6. Contributor – Contributing person or institution
7. Date – The date of the resource made available in the present form

8. Type –Resource type the category such as home page, novel, poem, working paper, report essay, dictionary etc.
9. Format- such as Text, HTML, ASCII.
10. Identifier- Resource Identification URL, URN, ISBN etc.
11. Source – physical and digital form which was presently derived.
12. Language- language of the resource
13. Relation – Relationship to other work
14. Coverage – Geographic
15. Rights – Right management statement, copyright, URL or other URI appropriate.

The DC metadata elements can be broadly categorized into three groups according to the scope of information stored in them.

A. Content of the resource

a) Title
b) Subject
c) Description
d) Source
e) Language
f) Relation
g) Coverage

B. Intellectual property

a) Creator
b) Publisher
c) Contributor
d) Rights

C. Elements related to instantiation of the resources:

a) Type
b) Format
c) Identifer
d) Date

DC characteristics

a) Simplicity of creation and maintenance
b) Commonly understood semantics

The application profiles of schemas consist of data elements drawn from one or more namespaces, combined together by implementers and optimized for local application.

8. Current Issues associated with metadata

Dublin Core and other recognized schemes were intended to support discovery of relatively non-complex resources and were not designed to accommodate the descriptive scope and complexity of audio or visual information. Additional fields and data modeling conventions will be needed to support other structures such as: administration, workflow etc.

Metadata is particularly important for visual resources that might otherwise stand alone without any text, and therefore be virtually irretrievable; users will depend on the information added to the image for accurate searching and retrieval.

9. Metadata Search engines

- Alta Vista search – can index and search metadata and fields from relational database.
- Easy ask – Extracts metadata facets from queries and sends them to relational database
- Endeca –Gathers metadata and fields from structural data
- Hot meta – Gathers the metadata from pages on the Internet or an intranet
- Metastar – Blue angel Technologies – Metadata creation, indexing and searching.[9]

10. Conclusion

The globalisation is the vision for future development of networking aspects. Dc has certainly contributed to this vision, for interdisciplinary and resource discovery. Many people had contributed towards it and everyone began to use DC. The simplicity semantic interoperability (cross domain), international consensus, extensibility and modularity have become the qualities of DC. Cultural heritage and information Professional such as museum registrars, library cataloguers, and archival processor are increasingly applying the term metadata to the valu-added information that they create to arrange, describe, track and otherwise enhance access to information objects, Carefully designed metadata results in the best information management in the short and long term. It is now a viable option and hence found widespread acceptance among the electronic information community.

11. References

1. http://lc web.loc.gov/marc/dc/subtypes- 20000928.html
2. http://metadata.net/ac/draft- iannella- admin- 01.txt
3. Dublin Core Metadata Initiatives. http:// dublincore.org
4. Baker, Thomas, “Languages for Dublin Core”, D-Lib Magazine, December 1998, http://www.dlib.org/dlib/January99/bearman/01 bearman.html
5. DOI, The Digital Object Identifier system, http://www.doi.org/
6. Day, M(2001) Metadata in a Nutshell, Information Europe, 6920,11
7. Dublin Core Metadata Initiative, http://purl.org/DC/
8. Haynes, David, ‘Metadata for information management and retrieval.’- London: Facet, 2004.
9. Library and Information networking naclin 2003 edited By H.K Kaul and B.B.Das.

10. Sudha, V.Vijayan. Metadata Technique in the organization of Digital Resources in Special Library. Content Management in India in Digital Environment:23rd All India Conference. Thiruvanathapuram, 2001. IASLIC, Kolkata, 2001. pp 77-84

About Author

Sanat Bhattacharya, M.A.(Eco.), M.L.I.S., B.ED., C.LIB., and diploma in Hindi, has been in the Library profession since 1994 and is presently working in Visva-Bharati University Library. He has attended two training programmes related to computer application in LIS.

Email : sanatkbh@yahoo.co.in