Metadata – Building Reality from Virtue

Jun Ma, MLIS

Dec 2001

Abstract

This article was written for providing an awareness of current Metadata concepts and standards to librarians and other information specialists who are not familiar with the development of Metadata in information management in the digital world. It tries to cover basic concepts of Metadata and major standards, applications; help readers understand various Metadata application technologies and issues. It also introduces software and tools to support metadata creation, harvesting and migration.

Basic Concept

Library is the traditional information repository to find most information needed. Library cataloguing is the method of organizing bibliographic information of library collections to facilitate their identification, location, access, and use.

But Internet suddenly becomes a major mechanism for users connecting to a tremendous amount of digital resources directly from their own desktops. The Web itself becomes a huge free information database, a virtual library and also a path to acquire resources from other databases within the network.

Coming with this excitement are the phenomena that most of the Internet information is poorly organized, not stable. It is difficult for normal user to search or browse this huge wealth efficiently. There is an increasing demand to have a cataloguing mechanism for the Internet information.

Different types of questions are raised when facing this challenge: are the methods of organizing the digital collections similar to those of organizing traditional library materials? Do we need new describing tools and techniques to catalogue digital resources? What type of standard we deploy? They will be created and maintained by professional librarians or normal users? What skills will be needed for this word and how could users get these skills?

Metadata could be the main if not the only method to answer those questions in the digital world. It helps provide needed descriptions for electronic resources, decrease users' search and discovery time, and what is the most important – increase the recall and precision of retrieval.

So What's Metadata?

The most common definition of **Metadata** is "data about data." ALA Committee on Cataloging: Description and Access lists 27 definitions of Metadata from different groups and projects[1].

It’s very common that most organizations start to create their own digital collections according to the specific types of structure or format of the resources inside. But currently there are strong demands on sharing those digital resources to save the time and investments. For users who want to retrieval and use those resources from different locations and services through a single interface, it’s nearly impossible for them to know the specific retrieval requirements of each type of resource in advance. So there should be a mechanism that removes the barriers between different systems, provides common search terms, access points, and data structures, and also makes those differences transparent to end users.

Metadata is typically designed for describing the bibliographic information and data structure of resources among different systems to distinguish one piece from another, let user easily locate, access and transfer the data between. According to Prairie Village[2], the term "metadata" has been used since the early 1980s by the computer software and systems development community to describe the information required to document the characteristics of information contained within databases. But right now the resource described by Metadata is not limited to the document, it includes text, image, audio, video and any other format.

For a typical Metadata standard, it should have a complete elements set, each of which represents one property of the resource (normally called entity) described, with a structure that how these elements are organized, and a controlled vocabulary to define the name and value of elements.

This sounds very familiar to librarians and cataloguers, since the MARC standard that is wide used in library community can be treated as a sample of Metadata. The concept of Metadata is being spread from library science to other domains such as computer science, and commonly accepted by all. Since the distribution of the first draft version of Text Encoding Initiative (TEI) Guideline in 1990, there have been a number of Metadata standards or schemas created covering a wide range of communities. Some of these metadata schemas are general, such as MARC or the Dublin Core, designed for electronic resources used in various disciplines. Other metadata schemas are dealing with specific discipline or domain, such as Government Information Location Services (GILS) and the FGDC Content Standards for Digital Geospatial Metadata.

The most common feature among these schemas is that they all use a set of defined elements to describe the properties of individual entity. As Vellucci indicated, there are three basic characteristics common to all metadata schemes: (1) syntax, (2) semantics, and (3) structure[3]. To clearly identify each entity within the collection, each element can have one or more qualifiers to provide additional semantics to the values of elements, enhance the authority control to distinguish different entities.

On the other hand, the number of elements of metadata, how they describe the entity, and their structure are quite different. Besides the traditional description function, some schemas also include administrative metadata elements, structural metadata elements to

manage the entity, the link between the entities, the access points and access rights to the entities. As an independent part from the entity described, the metadata elements can be stored together with the entity in the same file, or separately as a document in another physical location.

Issue and problem

Although Metadata has great power and the superiority in digital resources organization and retrieval, it inevitably has a number of issues within this relatively new domain.

Standards

Currently there isn't a universe Metadata standard in the digital world. Most of Metadata schemas were created for specific project or purpose, and named as standards for local usage or within that domain. It's very understandable because each project has specific requirements and data structure. It may not be feasible and economic to accept another metadata schema. In lots of circumstances, "(the project) may start up its own metadata definition effort, and creators are free to use whatever tags happen to come to mind", "any group that is developing metadata sets is free to limit its work to its narrow interests; it need not take a broader view unless it voluntarily chooses to do so"[4].

Currently a variety of Metadata standards and Metadata schemas exist that create a huge barrier for user accessing different systems simultaneously, and sharing the their resources. At the same time, the number of players in this field is continuing increasing. To solve this problem and improve the interoperability of Metadata, setting guidelines for new comers becomes very important and urgent. There are too many factors involved, including technology and policy.

However, the advanced information technologies and Internet are making the dream of sharing digital resources worldwide become true. Some solutions could be:

a. Set up several international standards and guidelines for general usage of Metadata. Dublin Core could be a good sample. After six years development, DC has been used as the module in a variety of disciplines, and different types of elements and qualifiers are added to its fifteen core elements accordingly. The guidelines should be standardized to keep the creation of new schema consistent to the module, but be flexible to ensure the extensions to represent the specific requirements.

b. Develop technology and tools to make it easy to map different metadata schema. Implement the registry and harvest system for the application profile. Resource Discovery Framework (RDF) and XML namespace can be suitable for this purpose. RDF provides a way to encode Metadata, and XML provides a machine-readable and executable syntax for RDF. Further investigations and studies are needed to apply these technology and rules together.

Metadata, embrace or embarrass the search engine?

Internet is the major stage for Metadata to play. There are more than one billion Web pages on the Internet and every day more than one million new pages are created. This exciting number also makes it difficult for users to retrieve the high quality information. Although there are also hundreds of Internet search engines acting inside to help users, everyone has experiences that face the large number of irrelevant hits. This situation becomes worse when multi-format, multi-language resources are created and available there.

World Wide Web Consortium (W3C) has developed dozens of specifications for the Web's infrastructure including HTML, RDF and XML. Starting from HTML 3.2, Meta tags are recommended to be used in the header of a HTML document, which let authors provide the information about a document such as author, description of content, and keywords. The role of Meta tag is providing a method to describe the site. let search engine index the contents of Meta tags and ensure accurate and reliable search result available for users.

Anything has two side edges, and Meta tag could not be exceptional. Internet is open to everybody with limited control. Anyone can contribute any information. You may not get what you expect because not everyone follows the rule.

All the Meta tags used in HTML are optional and depend on user's knowledge of coding and their tendency to do so. We could not force them, or expect everyone to obey W3C's rules. s. Currently there are only several search engines index Meta elements, such as Ultraseek, Microsoft's Index Server. All of them are not very popular and known by most users. Actually it's not hard to index those several Meta tags by search engine. The major problem here is how to prevent spamming from some ugly designed Web pages, which were embedded with inappropriate words, or arrange dozens of same word in the same position to increase their ranking. There is not a good solution yet and most search engines refuse to index Meta elements from Web, or treat Meta elements as normal word, which omits the designed function of Meta tags.

Things are getting more complicated if we expend the scope of Meta elements to include those Metadata schemas such as DC. The challenges are how to harvest the metadata automatically by search engine, map and store different Metadata schemas to the central repository. As Andrew Wood pointed, "The lack of precision in automatically generated metadata makes the Internet too imprecise for rigorous use, not just by people wishing to find information, but for people publishing information."[5]

If all producers of Web-based resources create metadata according to one rich and wide accepted standard and embed this metadata into the HTML header of their document, the "automatic" creation of metadata will become a reality, just as MARC well accepted by library communities around the world. Because librarians are trained to organize all types of information materials with well-designed systems, those traditional cataloging methods are still the main streams of organizing Internet resources to ensure the accurate and fair

of information catalogued. OCLC Cooperative Online Resource Catalog (CORC) is the good sample for organizing, transferring and sharing the catalog records of Internet resources with MARC or DC format. But this type of metadata generation is very labor-intensive, and requires tremendous contributions of time and money. The collection of records could not be free available to library users unless the participating libraries import the records to their local library OPAC.

For the major resources on the Internet, the information creators normally are better candidates at generating metadata than anyone else. Most of them are not information professionals, and lack of knowledge of cataloguing rules. A simple, less elements set, easily to be created and harvested, and for general use Metadata standard is highly recommended on the Internet. Dublin Core may become the best solution for this purpose.

Rights Administration

As the data about an electronic resource, Metadata functions more than describing that resource. It can be used to manage the access right of resource repository, characterize the relationships between resources and users, or between different resources themselves.

The right administration is a new concept of Metadata to protect intellectual property rights in the digital environment. Metadata is produced by publishers or the owner of resource repository to manage the distribution, access and ownership of resources, and effectively control the transfer of such rights to different users. It ensures authorized users to access the whole or part of components of an intellectual "object" (for example, the abstract or full text of an article) according to the terms and conditions assigned.

The INDECS (Interoperability of Data in E-Commerce Systems) project and ONIX (Online Information exchange) standard are two applications of rights control Metadata that combines general metadata schema with copyright control and protection developed by publishers of electronic resources. A detailed introduction of INDECS with DOI (Document Object Identifier) together will be introduced later in this article.

Authority control

As more and more Metadata schemas are created, the number of unfamiliar vocabularies used in those schemas is also rapidly growing. Different Metadata could choose different names for the same meaning, such as author, creator, producer ... all describe the body that creates the entity. How can the searcher know what word or value has been assigned to the resources that he is looking for? This could be much more serious when dealing with the project using special terminology such as Geospacial, chemical, or biological databases.

The traditional subject heading could be a good solution. With RDF namespace, each schema can specify which classification system it uses, and where to find this classification system online. In this way different systems such as LCSH, MESH, or even DDC that only contains numbers can be compared and mapped in one repository. It

enables users to access different resource repositories even they don't know the term or type of elements used in that repository. He/she can type the common terms, and the system will automatically map this name to the correct one using authority file. It not only let users avoid learning different terminology as well as increases the recall and precision of the search, but also provides the convenience for the resource producers to create Metadata with the terms they are familiar with. This could dramatically reduce the workload of information professionals who are currently responsible for making Metadata.

Major Systems & Initiatives

Dublin Core

Dublin Core could be the most famous Metadata standard today. From its first workshop held in 1995, there have been eight Metadata workshops so far and resulted in the current Dublin Core Metadata Element Set (DCMES). The Dublin Core Metadata Initiative is the formal forum responsible for the DCMES, and is "dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems"[6].

The DCMES was designed as the descriptive metadata to support resource discovery. As DCMI director Stuart L. Weibel mentioned, the mission of the DCMI is to make it easier to find resources using the Internet through the following activities[7]:

1. Developing metadata standards for discovery across domains;
2. Defining frameworks for the interoperation of metadata sets;
3. Facilitating the development of community or discipline-specific metadata sets that work within the frameworks of cross-domain discovery and metadata interoperability.

By now the DCMES contains fifteen element sets separated into three groups: Content, Intellectual Property, and Instantiation[8]. These elements could be further refined by qualifiers. A qualifier gives additional information on the values of an element, such as classification or controlled vocabulary or other encoding schemes used, to define or narrow down the meaning of element values in the specific project. A list of qualifiers is recommended within DCMES.

From the beginning of DCMI, the Dublin Core has been heavily influenced by librarians, and seen by them that DC has the similar functions as MARC to organize electronic resources. But DC is not intended to replace those rich and complicated description standards. Instead, it provides a core set of description elements that can be used by normal users who are not familiar with cataloging rules for simple digital resource description.

In September 2001, Dublin Core Metadata Element Set was approved by National Information Standard Organization as an American National Standard ANSI/NISO Z39.85. Besides that, Dublin Core 1.1 has been adopted by CEN/ISSS (European Committee for Standardization /Information Society Standardization System) as part of a CEN/ISS Workshop Agreement (CWA 13874).

The three features of Metadata identified by Vellucci[9] as the new buzzwords in digital information organization: flexibility, interoperability, and extensibility are also reflected in DCMES:

Flexibility
It's a core set of simple elements that could be applied for general digital resources in many domains. All the elements are optional and repeatable. Although each DC element or qualifier would follow a simple structure, there is no formal syntax defined for DC itself. It only deploys W3C's RDF as its application module. The Metadata records could have as many descriptive elements as needed to ensure each entity within the digital collection could be uniquely identified.

Interoperability
There are great demands for embedding more than one Metadata element set in the resource repository. RDF allows a group of different metadata sets to be described by implementing registry within the application. Through registry, the structure and semantics of each Metadata set will be defined, and the way they are recorded and accessed will be described. It will enhance the interoperability of Metadata sets co-exist inside the application and improve the search and exchange ability.

Extensibility
The schema could be modified and extended by choosing part of DC elements set and adding new elements and qualifiers according to the specific requirement of the project within the framework, as long as the creators follow the DCMI Recommendations of schema. This encourages the extension of DCMES application in different domains.

The DCMI Website maintains a comprehensive list of Dublin Core projects around the world, and there are several DCMI work group are studying the particular DC schema for different domains, such as Education, Government, Libraries Working Group and several Special Interest Groups[10]. This also could improve the awareness of DC in most disciplines, prevent duplication by individual project, and enhance interoperability of different applications under DC schema.

Mainly the DCMES focuses on the description of digital resources. Because of its simplicity, it pays little attention on the resource's intellectual property or access rights management. Although the simplicity could enhance interoperability, it does not accommodate the semantic and functional richness supported by complex metadata schemes (Z39.85). Combination of richer metadata schemes with Dublin Core is

recommended by NOSI, including mapping those richer metadata schemes to Dublin Core for export or for cross-system searching[11].

FGDC

The United States Federal Geographic Data Committee (FGDC) approved the Content Standard for Digital Geospatial Metadata (CSDGM) in June 1998. CSDGM was designed to "provide a common set of terminology and definitions for the documentation of digital geospatial data"[12].

The standard is organized in a hierarchy of data elements, compound elements (groups of data elements) and information about the values of the data elements to document a set of digital geospatial data. Through the CSDGM standard, all the participants in geospatial digital resources can make their collections searchable and accessible on the Internet with the implementation software.

The National Geospatial Data Clearinghouse, developed by the FGDC, is a distributed network of geospatial data producers, managers, and users linked electronically. It uses CSDGM to describe field-level of digital spatial data, using ANSI standard Z39.50 for the search. The FGDC has also developed a profile for geospatial metadata, called "GEO," which provides guidance on how to implement FGDC metadata elements within a Z39.50 service[13].

FGDC has developed guidelines for creating extended elements and profiles for CSDGM, if users want to modify the Standard to meet particular requirements for their data sets. "A profile is a document that describes the application of the Standard to a specific user community; contains the Standard, plus modifications to the optionality or repeatability of non- mandatory elements in the Standard. It may also contain extended elements."[14] All the extended elements and profiles must follow these guidelines to ensure the accessibility and exchangeability of other users within the network.

TEXT ENCODING INITIATIVE (TEI)

The Text Encoding Initiative (TEI) is one of the oldest Metadata standard. Started in 1987, TEI becomes "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."[15] The new TEI Consortium was set up in December 2000 to continue maintaining and developing the TEI standard. TEI mainly focuses on the exchange of textual information, but other formats such as images and sound are also addressed.

One of the main characteristics of TEI is that its descriptive metadata is created as the header (known as TEI header) of the encoded file (known as the TEI text proper). The format of TEI encoding scheme is based on the SGML standard. The TEI header has four major parts (<fileDesc>, <encodingDesc>, <profileDesc>, <revisionDesc>), each of which contains a hierarchical set of description using TEI tags. TEI text also includes a set of tags. All the TEI tags are defined in TEI DTD (Data Type Definition).

When TEI was designed, it was expected that the producers of the textual digital data would create the encoding data for the files at the same time. In fact the TEI is very complicated, and the TEI Guidelines encompass some 1300 pages of information[16]. It requires the metadata producer fully understanding the TEI scheme and tags for particular types of text. To make their work easy, a useful basic tag set was selected from the several hundred SGML elements and defined by the full TEI scheme, which contains the most needed elements. Since many texts marked up according to the TEI guidelines are based on printed books for which AACR2R/MARC catalog records exist, in many cases header data is created or reviewed and revised by librarians[17].

There is a widely usage of TEI in the world. For example, The Electronic Text Centers at the University of Virginia and the University of Michigan have used TEI Lite to encode their holdings. TEI Website also maintains a list of projects using TEI encoding scheme.

CDWA and VRA

There are two important Metadata standards designed for visual collections: CDWA and VRA.

CDWA stands for Categories for the Description of Works of Art, a metadata schema designed by of the Art Information Task Force (AITF), to "describe the content of art databases by articulating a conceptual framework for describing and accessing information about objects and images."[18] It was released in February 1996.

This Metadata schema is very extensive and developed for use by art specialists. There are 26 main categories, and each category has its own set of subcategories. All the categories are fit into two groups: "Object, Architecture, or Group" as the information intrinsic to the work, and "Authorities/Vocabulary Control" as the information extrinsic to the work. CDWA also defines several core categories that are necessary to uniquely and unambiguously identify a particular work of art or museum object.

The VRA Core Categories (current vision 3.0) is Visual Resources Association's approach of categorizing visual documents that represent objects of art or architecture. The VAR Core Categories provides a template designed but not limited to visual works collections. As VRA Data Standard Committee pointed, "CDWA was exhaustive in its list of elements needed to describe museum objects, it was not entirely satisfactory for the description of images, and in particular, did not cover all of the elements needed for the description of architecture and other site-specific works."[19] There is a need to expand the concept to non-art objects and visual document for which VRA was created.

Compared with and Benefit from CDWA, VRA Core categories is designed to cover most visual materials. It does not have such comprehensive categories as CDWA. Similar to Dublin Core, It provides a core set of elements, which could be expanded by adding new elements as needed. VRA 3.0 contains 17 categories that can be used to describe both work and representations of the work (defined as images). It also borrows

the concept of Qualifier to help describe and distinguish entities, clear the relationship between works and images, such as a photo of a work, and a digital image of that photo. Because of the characteristics of visual objects, it is recommended that develop additional elements for the local collections, set up controlled vocabularies in categories and use other Metadata sets such as CDWA, MARC for guidance.

GILS

Globe Information Locator Service, is designed as a profile that combines policy, standard, information technology and products to enable people to locate and retrieval information from diverse sources. It involves governments, companies, information providers and different initiatives around the world, led by the Global Information Society initiative, which was organized by G7 countries. Sometimes GILS is called as Government Information Service since it based on the United States Federal Government Information Locator Service initiative project.

The U.S. Government Information Locator Service is an approach to identify, locate, and describe Federal information resources including electronic information resources, and make them available to public. GILS is a decentralized collection of agency-based information locators using information technologies, metadata and standards to let user retrieval Federal Government information resources located in different servers. The U.S. Federal GILS was mandated by the Paperwork Reduction Act in 1995 (U.S. Public Law 44 USC 3511), which requires each Federal agency to establish and maintain an information locator service as a component of the Service.

Besides the United States Federal Government, some U.S. states governments and other countries and provinces' governments including Australian, Canadian Federal Government, Ontario also have set up their GILS standards, or policies. The ISO 23950 / ANSI Z39.50 standard, which was developed for structured search across databases such as bibliographic catalogs in library, is one of the standards within GILS for exchanging electronic information.

Metadata is a key component in GLIS. This is because within GILS there are various agencies and locator services. Each of them may use different data structure for their information, and use different types of Metadata schema to describe those resources. For example, the geospatial data in the U.S. Federal Government agency can use Dublin Core Metadata Element Set, or FGDC Metadata; a library catalog which is also a locator service, uses MARC for bibliographic description. So a Metadata scheme (the GILS Core Elements) with approximately 70 elements is used to map different metadata sets, and describe other electronic resources in a uniform manner for searching and retrieving. In certain way it can be seen as a fairly complex metadata format. This result could be contributed to the heavy influence from the MARC and Z39.50 standard. The Annex B of *GILS Profile* provides the mapping from GILS to USMARC[20].

Encoded Archival Description (EAD) & Metadata Encoding & Transmission Standard (METS)

Both of these two Metadata standards are maintained in the Network Development and MARC Standards Office of the Library of Congress.

The METS schema is a standard using the XML to encode descriptive, administrative, and structural metadata for objects within a digital library. It is being developed as an initiative of the Digital Library Federation, based on the *Making of America II project*. METS attempts to provide an encoding Metadata to manage the digital objects within the digital library, and exchange those digital objects between different digital libraries.
A METS document consists of four major sections: Descriptive Metadata, Administrative Metadata, File Groups, and Structure Maps. Its Website provides the detail function of these sections and samples of using METS Metadata.

Encoded Archival Description (EAD) is a set of rules for preserving the hierarchy and designating the intellectual and physical parts of archival finding aids to help search, display and exchange archives and manuscript collections. The EAD rules are written in the form of a Standard Generalized Markup Language (SGML) Document Type Definition (DTD), because archival description emphasizes intellectual structure and content more than bibliographic description, making SGML, and later XML, a more suitable transport syntax than MARC[21].
The EAD is grouped into two parts: bibliographic header (<eadHeader>) and the marked up finding aid itself (<findAid>). The header only describes the finding aid, and finding aid is used to describe the collection. Both of them contain the EAD elements for the description. The EAD Tag Library contains all EAD data elements defined in the DTD. It serves as a reference tool for archivists who decide which EAD elements to use when designating the content of their finding aids. There are 145 elements in the EAD Tag Library. But only a few are the required elements in EAD DTD. Others are optional, which depend on the requirements of individual collection.

RDF
With more and more Metadata standards or schemas used in a variety of projects, the need of encoding, searching, and exchanging different Metadata sets in one interface becomes much more important. The Resource Description Framework (RDF) is being developed by the World Wide Web Consortium (W3C) to provide an infrastructure for the interoperability of Metadata in a wide range of applications.

Every Metadata must define the syntax, semantics, and structure of its elements for its applications. Using XML (eXtensible Markup Language) as its common syntax, RDF help users define the structure of Metadata elements used the application profile. Those Metadata elements can be from more than one Metadata schemas. Although RDF does not directly define the semantics of Metadata elements, with the namespace feature of XML it specifies the location of schema and its vocabularies so that the semantics of elements of these schemas can be tracked and explained clearly without any ambiguity. Here the vocabularies are defined as the set of properties, or metadata elements for resource description[22].

One of the advantages of this framework is that people do not need to design and create a new Metadata schema for a project. He/she can combine several existing Metadata elements sets together, and choose any element that meet his specific needs from those sets by using XML namespace to indicate which element belongs which schema, and the address of DTD of that schema. This could let the common Metadata schemas reused, avoid the huge investment spent on creating new schema, as well as increase the Metadata interoperability.

Following is a small example that explain how XML namespace can combine different metadata schemas together:

```
<RDF:RDF
   xmlns:RDF = "http://www.w3.org/RDF/RDF/”
   xmlns:DC = "http://purl.oclc.org/DC/">
   <RDF:Description RDF:about = "http://uri-of-object1">
      <DC:Creator>Jack Ma</DC:Creator>
   </RDF:Description>
</RDF:RDF>
```

The first line, which can be interpreted as the element RDF in the context of the RDF namespace, defines a simple wrapper that marks the boundaries of the document using RDF.
The next two lines are XML namespace declaration, which define the namespace prefix ‘RDF’ and ‘DC’, using the uniform resource identifiers (here use two URLs) to indicate which schemas are used in the document.

The next is the RDF “Description” statement designed to group multiple descriptive statements (or elements) for one object (here is the document that specified by its URL http://uri-of-object1) into a “Description” element.

The DC:Creator declares that the “Creator” property element of the object within Description is defined using the property element of DC schema, and its value is “Jack Ma”. The last two lines just close the declaration of Description and RDF.

If needed, qualifiers can be used under element declaration to further define the property of the object. Other metadata schemas can be added into this sample with the similar namespace declaration. Here RDF works like a platform where different Metadata schemas can work together describing different types of property of the same object. Using RDF as a module for Metadata application, a central registry is normally required to support the declaration, and mapping of metadata schemas within the project. Each metadata schema declares the vocabularies of its elements set to provide the ability of unambiguously expressing semantics for encoding, exchanging, and machine processing of metadata consistently.

DOI
The Digital Object Identifier (DOI) was developed and maintained by the International DOI Foundation to describe the digital objects for “intellectual right management”. The

DOI system is made up of four components including Enumeration, Description, Resolution and Policies, each of which depends on others for its value.

Enumeration is the way of DOI system to construct unique identifiers. Each DOI is a unique string assigned to identify one entity (resource). It can be used to identify any of the various physical objects that are "manifestations" of intellectual property; or less tangible manifestations, such as the digital files; or the performance of intellectual property and the "abstractions" that underlie the different manifestations[23]. The DOI can be seen as a persistent identifier, which means that even the ownership of the entity or the rights in the entity has changed, the identification of that entity should not (and does not) change. This is the core of DOI system.

The DOI identifier has two components: prefix and suffix. The prefix is assigned by DOI Registration Agencies (currently are the IDF and CrossRef). Actually the DOI is an implementation of CNRI Handle system that uses “10.” as the start of prefix followed by a number. The suffix can be assigned uniquely by the organization that has applied the prefix. It can be any alphanumeric string but must be unique within that given prefix. Combining prefix and suffix together can get a unique identifier for one entity.

DOI system uses Handle system that developed by Corporation for National Research Initiatives (CNRI) to resolve its identifier to the associated information in the network. The Handle System is a comprehensive system for assigning, managing, and resolving persistent identifiers, such as DOI for digital objects and other resources on the Internet[24].

The Handle System includes a set of protocols, a namespace (prefix and suffix), and an implementation of the protocols. The protocols enable a distributed computer system such as Internet to store handles of digital resources (e.g. DOI) and resolve them into the information necessary to locate and access the resources. This function is very similar to the Domain Name Services (DNS) for resolving the IP address.

Although one resource can be uniquely identified by its DOI identifier, this string only can be manipulated by computer and normal user could not get real meaning from it. It’s necessary to combine metadata with the identifier together to provide a meaningful method for user describing, searching and accessing the entity. Metadata is an essential component of the DOI System.

The DOI system also defines a very small but mandatory Metadata elements set, which only includes six elements: Identifier, Title, Mode, Type, Primary Agent, and Agent Role. Since there is a variety of intellectual property types exist that requires different Metadata elements sets to describe them, DOI system needs to expand the scope of Metadata elements to describe them. It adopts INDECS Metadata Framework as its Metadata model because INDECS schema is also designed for the same domain – intellectual property.

The INDECS (Interoperability of Data in E-Commerce Systems) project was established in1998, and ended in March 2000, which was evaluated as successful by the project

Commission. It aimed to develop specific metadata for electronic commerce that enables the interoperability of different metadata for different media and functions in electronic trade activities. The key point is how to unambiguous identify and describe the rights of using intellectual property, which could not be provided by other Metadata schema such as Dublin Core. It focuses on the "creation, modification, use and 'publishing' of entities, and the conditions to enable these events: transactions, agreements, offers and payments"[25]. INDECS defines the attributes of the entities, the roles they contain, relations, parties, creations, Intellectual Property, IPR (Intellectual Property Rights) Transactions, Assertions and Non-textual Metadata. The INDECS framework contains a Metadata model, and a high-level Metadata dictionary, principles for mappings to other schemas, and a Directory of Parties proposal.

Although the project has been closed, the INDECS partners established a non-profit organization called INDECS Framework Ltd. that continues maintaining and developing the INDECS Framework.

For specific type of entities, the Metadata elements set would be quite different from another set in DOI system. To clearly define the structure of that elements set and its registry, a DOI Application Profile (DOI-AP) is defined to be part of the Metadata.

A DOI-AP is initially defined in terms of a class of Intellectual Property entities and an application (or set of applications). The purpose of a DOI-AP is to enable the implementation of application (or set of applications) in a particular environment. Those applications can range from relatively simple resource discovery to complex e-commerce and rights management applications. Besides the definition of the metadata schema, DOI-AP also contains the commercial and procedural rules (which can include, for example, rules relating to the exploitation of the metadata that is declared), the Registration Agency or Agencies responsible for its application, and DOI User Community responsible for its management[26].

Application tools

Although Metadata provides a good solution to organize a variety of digital resources, and improve the quality of the information retrieval in the giant digital world, its concept is still new to most information users, let alone how to apply it in the traditional information environment, especially there are so many metadata standards and schemas exist. On the other hand, it's inconvenient and uneconomic to use human-generated metadata for each digital entity. The best way for metadata application is let machine generate metadata according to the preset schema or standard. But with current technology, automatically generated metadata could not provide enough precision, which restricts the implementation of metadata in the real world of information retrieval.
Lots of projects have been implemented trying to solve this problem. They can be grouped into several components: Metadata Creation, Metadata Mapping, and Metadata Registry

Metadata Creation:
There are lots of tools or software designed to create metadata for specific digital resources. These include Metadata templates and harvests. Normally templates need user to input the descriptions of a resource's major properties such as title, author and other information to generate the metadata elements according to the predefined metadata schema.

Nordic Metadata Project[27] has a very famous Metadata template for Dublin Core Metadata Element Set creation. The U.S. Geological Survey (USGS) also provides several software to parse or edit Formal Metadata according to the FGDC Content Standard for Digital Geospatial Metadata[28]. This type of software can be simple that only uses fifteen DC elements. It also can be very complicated that use several levels of Metadata elements set, such as GEMCat from U.S. Department of Education Initiative for the Gateway to Education Materials project[29]. It adopts DC elements and qualifiers to describe electronic educational materials on the Internet.

Metadata harvest are those tools that extract metadata description of Internet Website from HTML Meta tags according to Metadata schema chosen (most of them are using DC). Some simple harvest tools only require user type the URL of that Website. They can be seen as a simple web search engine only index Meta tags in HTML headers. They are useful only when most Websites they indexing are using Meta tags for site description. One example is DC-dot from UKOLN[30].

There are some software that combine both template and generator functions together. One example is Reggie, the Metadata editor from The Distributed Systems Technology Centre (DSTC). It can create eight types of metadata format using Java Applet including DC, GLIS, IMS, GEM. User also can input one Website URL to generate specific metadata (if applicable) [31].

Metadata Mapping (Crosswalks)
Mapping is one of the important elements for metadata interoperability. It is the method that indicates the equivalence between concepts in different schema, especially in semantics. By mapping the metadata schemas for different types of digital resources, users could search these resources together using one software or interface without knowing the different characteristics among the resources. What they only need to know is the common concepts. The typical case is that user searches several databases simultaneously, each of which contains different types of digital entity such as text, image, video.

This functionality becomes more and more important since the number of metadata schemas or standards is increasing dramatically, lots of them will co-exist for certain usage. Mapping (sometime called crosswalk) ensures the consistence of semantic interoperability.

Most of Metadata can be mapped to other schemas for certain elements. There always happen that some elements and data are lost when mapped to another metadata. This phenomenon depends on the characteristics of both side of metadata, and the design of the crosswalk. Currently lots of Websites provide different types of crosswalk. Several famous are:

- Dublin Core/MARC/GILS Crosswalk and MARC to Dublin Core Crosswalk, both of which are maintained by Library of Congress.
- UKOLN: The UK Office for Library and Information Networking also maintains a site that lists dozens of crosswalks that map different types of Metadata.
- The Getty Research Institute has a Web page contains several charts that map major metadata standards to one another.

Registry

Registry is very important to the Metadata mapping. To perform the mapping from one metadata to another, both the names (or values) of elements should be contained in a controlled vocabulary, where those values already have been defined. The process of predefine the value of all elements of one Metadata is called registration, and the place (dictionary or authority file) to hold these definitions is registry.

The registry describes the semantics, the structure and the syntax of a Metadata element set it holds. It may also include policies or recommended practice for using the defined terms. By the use of registry, all the Metadata schemas in the system are identified and specified. It avoids the ambiguity of words existing in different metadata, and provides the best mapping between two types of Metadata. UKOLN Website contains several Metadata Registries from different projects, such as DESIRE Metadata Registry, ROADS, and Dublin Core Registry.

ROADS stands for Resource Organization And Discovery in Subject-based services
The ROADS project has several purposes[32]:

1. To produce a software package which can be used to set up subject-specific gateways
2. To investigate methods of cross-searching and interoperability within and between gateways
3. To participate in the development of standards for the indexing, cataloguing and searching of subject-specific resources

All Internet resources with different types of format under one subject domain can be described by different types of Metadata. Those Metadata are created by ROADS templates. These templates consist of a set of simple attribute-value pairs. ROADS use the WHOIS++ protocol to simultaneously search several databases. All the templates ROADS uses are included in Metadata Registry, and each template element is defined inside. Those Metadata cover different types of data objects such as document, dataset, image, and even another Metadata schema, e.g. Dublin Core.

Conclusion

Metadata is still new to most of us. There are lots of questions and uncertaians stand on the way of Metadata development. How to adopt to existing major metadata schema and technology such as Dublin Core and RDF, and expand their potential usage on the digital resources managment require organizations and professionals from different domains cooperate worldwidely. Can the challenges and opportunities coming with new information technologies make metadata another mircle in the digital world? We can not answer right now. But it worth to try.

Reference:

1. ALA Committee on Cataloging: Description and Access. Task Force on Metadata: Final Report. June 2000. Online. Available at: http://www.ala.org/alcts/organization/ccs/ccda/tf-meta6.html.

2. Caplan, Priscilla. International Metadata Initiatives: Lessons in Bibliographic Control. 2000. Online. Available at: http://lcweb.loc.gov/catdir/bibcontrol/caplan_paper.html.

3. Vellucci, Sherry L. Metadata and Authority Control. *Library resources and technical services* Jan. 2000 44(1): 33-43.

4. Milstead, Jessica and Susan Feldman. Metadata: Cataloging by Any Other Name… *Online.* Jan 1999. Available at: http://www.onlineinc.com/onlinemag/OL1999/milstead1.html.

5. Wood, Andrew. Metadata – The Ghosts of Data Past, Present, and Future. 1997. Online. Available at: http://archive.dstc.edu.au/RDU/reports/Sympos97/metafuture.html.

6. DCMI. Dublin Core Metadata Initiative Overview. 2001. Online. Available at: http://www.dublincore.org/about/

7. Weibel, Stuart and Traugott Koch. The Dublin Core Metadata Initiative: Mission, Current Activities, and Future Directions. *D-Lib Magazine*. Dec 2000. Available at: http://www.dlib.org/dlib/december00/weibel/12weibel.html.

8. Milstead, Jessica and Susan Feldman. Metadata: Cataloging by Any Other Name… *Online.* Jan 1999.

9. Vellucci, Sherry L. Metadata and Authority Control. *Library resources and technical services* Jan. 2000 44(1): 33-43.

10. DCMI. Dublin Core Metadata Initiative Workshops. 2001. Online. Available at: http://www.dublincore.org/workshops/

11. NISO. Draft Standard Z39.85-200X: The Dublin Core Metadata Element Set. 2001. Online. Available at: http://www.niso.org/Z3985.html

12. FGDC. Content Standard for Digital Geospatial Metadata (CSDGM). 2000. Online. Available at: http://www.fgdc.gov/metadata/contstan.html

13. FGDC. Geospatial Data Clearinghouse: Questions and Answers. 2001. Online. Available at: http://www.fgdc.gov/clearinghouse/background.html.

14. FGDC. Guidelines for Creating a Profile for the Content Standard for Digital Geospatial Metadata. 1998. Online. Available at: http://www.fgdc.gov/metadata/csdgm/profile.html.

15. TEI. What is the TEI consortium? 2001. Online. Available at: http://www.tei-c.org/Consortium/index.html.

16. Sperberg-McQueen, C. M. Textual Criticism and the Text Encoding Initiative. 1994. Online. Available at: http://www.tei-c.org/Vault/XX/mla94.html

17. Caplan, Priscilla. International Metadata Initiatives: Lessons in Bibliographic Control. 2000.

18. Baca, Murtha and Patricia Harpring. Categories for the Description of Works of Art: Introduction. 2000. Online. Available at: http://www.getty.edu/research/institute/standards/cdwa/

19. VISUAL RESOURCES ASSOCIATION Data Standards Committee. The Core Categories for Visual Resources – Introduction. Online. Available at: http://php.indiana.edu/~fryp/coreintro.htm.

20. National Institute of Standards and Technology. Version 2 of "APPLICATION PROFILE FOR THE GOVERNMENT INFORMATION LOCATOR SERVICE (GILS)". 1994. Online. Available at: http://www.gils.net/prof_v2.html.

21. Caplan, Priscilla. International Metadata Initiatives: Lessons in Bibliographic Control. 2000.

22. Miller, Eric. An Introduction to the Resource Description Framework. *D-Lib Magazine*. May 1998. Available at: http://www.dlib.org/dlib/may98/miller/05miller.html.

23. International DOI Foundation. What is a Digital Object Identifier? *DOI handbook*. 2001. Online. Available at: http://www.doi.org/handbook_2000/what_is_a_doi.html

24. International DOI Foundation. Appendix 5: The Handle System. *DOI handbook*. 2001. Online. Available at: http://www.doi.org/handbook_2000/appendix_5.html.

25. International DOI Foundation. Appendix 6: An Introduction to the <indecs> metadata framework. *DOI handbook*. 2001. Online. Available at: http://www.doi.org/handbook_2000/appendix_6.html.

26. International DOI Foundation. Appendix 3: Defining metadata schemas for DOI Application Profiles: an outline guide and template. *DOI handbook*. 2001. Online. Available at: http://www.doi.org/handbook_2000/appendix_3.html.

27. Koch , Traugott and Mattias Borell. Dublin Core Metadata Template. 1997. Online. Available at: http://www.lub.lu.se/cgi-bin/nmdc.pl

28. USGS. Formal metadata. Online. Available at: http://geology.usgs.gov/tools/metadata/

29. The Gateway to Educational Materials. Cataloging Workbench. Online. Available at: http://www.geminfo.org/Workbench/Cataloging/index.html

30. Powell, Andy. Dublin Core metadata editor. Online. Available at: http://www.ukoln.ac.uk/metadata/dcdot/

31. The Distributed Systems Technology Centre. Providing Digital Resource Management Solutions. Online. Available at: http://www.dstc.edu.au/RDU/

32. The Metadata Group of UKOLN. What is ROADS? Online. Available at: http://www.ukoln.ac.uk/metadata/roads/what/