ISSN:0975-9646

METADATA STANDARD HARVESTING

G.Sivaraman[1], Dr.K.Thangadurai[2]
[1] *Department of Computer Science, M.G.R.College,Hosur,TN, India*
2Department of Computer Science, Government Arts College(Men) , Krishnagiri- 635 001, India
E-mail: sivaphd_guru@yahoo.co.in,ktramprasad04@yahoo.com.

ABSTRACT
The rapid growth of Internet resources, digital collections and libraries are constructed with the help of metadata schemas. Each metadata schema has been designed based on the requirements of the particular user community, intended users, type of materials, subject domain, the depth of description, etc. Problems arise when building large digital libraries or digital information resource with metadata records prepared according to related schemas. Most of the users do not know or understand the underlying structure of the digital collection; but in reality, they are experiencing difficulties in retrieval. The challenge will be overcome through metadata harvesting. This paper is reviewing this harvesting with example.

Keywords: Metadata, Metadata standard, Harvesting, Crosswalk, Interoperability, Harmonization

1. Introduction

Information retrieval from heterogeneous resources is quite difficult. Because of the information holding follow different material administration and different metadata implementation techniques**.** There is two or more type of metadata standards are used in same subject domain or in same type of resource. In building a large digital library or digital collection, an issue often encountered is that the resource may have used different schemas and description methods to create their metadata records. Users want to retrieve information through one search what digital objects freely available from a variety of collections rather than searching each collection individually. User community can be developed to attain harvesting it will be possible to facilitate the exchange and sharing of data prepared according to different metadata schemas and to enable cross-collection searching. This article analyzes some of the methods currently used to achieve harvesting in a broader context, that is, among different metadata schemas and applications.

2. Harvesting

Harvesting refers to the gathering together of metadata from a number of distributed repositories into a combined data store. In other words, harvesting is a technique for extracting metadata from individual repositories and collecting it in a central catalog.

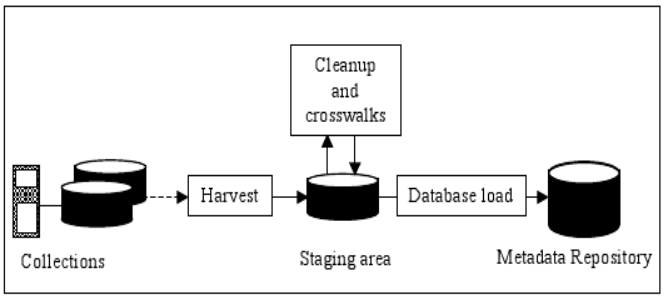

Figure 1: Collection of record store into metadata repository

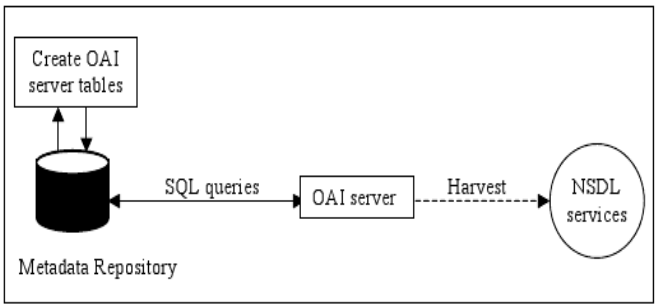

Figure 2: Process of Harvesting

Resources across the network can be searched more flawlessly using defined metadata standards and shared transfer protocols between these standards. Different metadata standards are available like the Dublin Core, LOM etc. For the accurate retrieval of information using metadata, the different metadata standards should be able to operate between themselves. Hence the concept of metadata harvesting arose.

Harvesting achieve in three ways, Crosswalk, Interoperability and Harmonization.

3. Crosswalk

A crosswalk is a specification for mapping one metadata standard to another. Crosswalks provide the ability to make the contents of elements defined in one metadata standard available to communities using related metadata standards.

A crosswalk is defined as a mapping of the elements, semantics, and syntax from one metadata scheme to those of another. The predominant method used is direct mapping or establishing equivalency between and among elements in different schemas. Equivalent fields or elements are mapped in order to allow conversion from one to the other. Most of the crosswalk effort to date has been in the form of mapping between two metadata schemas; mapping among multiple schemas has also been attempted.

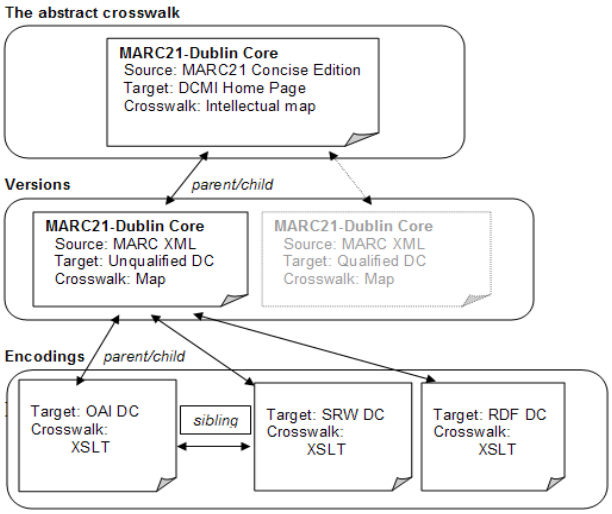

Figure 3: Levels of Metadata Abstraction

There have been a substantial number of crosswalks. Some examples are:

- MARC21 to Dublin Core
- MARC to UNIMARC
- VRA to Dublin Core
- ONIX for books to MARCXML
- FGDC to MARC
- EAD to ISAD(G)
- ETD-MS to MARCXML
- Dublin Core/MARC/GILS
- ADL/FGDC/MARC/GILS
- MARC/LOM/DC
- Etc., etc., etc.

The crosswalk approach appears to be more workable when mapping from complex to simpler schema. An example is the crosswalk between the Dublin Core and MARC. Because of different degree of depth and complexity, crosswalk works relatively well when mapping MARC fields to Dublin Core elements but not vice versa, because MARC is a much more complex schema. One of the problems identified is the different degrees of equivalency: one-to-one, one-to-many, many-to-one, and one-to-none. Also, while crosswalk works well when the number of schemes involved is small, mapping among multiple schemas is not only extremely tedious and labor intensive but requires enormous intellectual efforts. For example, a one-way crosswalk requires one mapping process (A-->B), and a two-way crosswalk requires two mapping processes (A-->B and B-->A).When the process becomes more and more cumbersome the more schemas are involved. For example, a crosswalk involving three schemas would require six (or three pairs of) mapping processes, a four-schema crosswalk would require twelve (or six pairs of) mapping processes and a five-schema crosswalk would require twenty mapping processes.

4. Interoperability

Interoperability means that the compatibility of two or more systems such that they can exchange information and data and can use the exchanged

information and data without any special manipulation.

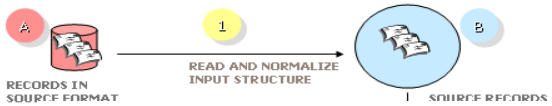

Figure 4: Process of Interoperability

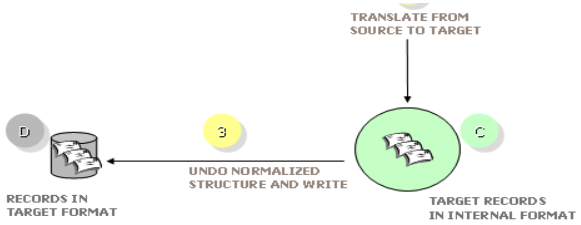

Figure 4: Process of Interoperability

In recent years, numerous projects have been undertaken in the information community to achieve interoperability among different metadata schemas. Some of these efforts are outlined below.

- Uniform standard
- Application profiling/adaptation/modification
- Derivation
- Switching schema
- Lingua franca

4.1. Uniform Standard

In this approach, all participants of a consortium, repository, etc., use the same schema, such as MARC/AACR or the Dublin Core. However, although it is a conceptually simple solution, it is not always feasible or practical, particularly in heterogeneous environments serving different user communities where components or participating collections contain different types of resources already described by a variety of specialized schemas. This method is only viable at the beginning or early stages of building a digital library or repository, before different schemas have been adopted by different participants of the collection or repository. Examples of uniform standardization include the MARC/AACR standards used in union catalogs of library collections and the Electronic Thesis and Dissertations Metadata Standard (ELD-MS) based on the Dublin Core used by members of the Networked Digital Library of Thesis and Dissertations (NDLTD).

4.2.Application Profiling/Ad aptation / Modification

In the heterogeneous information environment, different communities manage information that has different characteristics and requirements. There often is no one metadata schema that meets all needs, that is, “one-size-does-not-fit-all. “To accommodate individual needs, in this approach, an existing schema is used as the basis for description in a particular digital library or repository, while individual needs are met through specific guidelines or through adaptation or modification by:

- Creating an application profile (a set of policies) for application by a particular interest group or user community.
- Adapting an existing schema with modification to cater to local or specific needs, that is, a DTD of an existing schema

4.3. Derivation

In a collection of digital databases where different components have different needs and different requirements regarding depths, an existing complex schema such as the MARC format may be used as the “source” or “model” from which new and simpler individual schemas may be derived. This approach would ensure a similar basic structure and common elements, while allowing different components to vary in depth and details. For example, both the MODS (Metadata Object Description Schema) and MARC Lite are derived from the MARC21 standard, and the TEI Lite is derived from the full Text Encoding Initiative (TEI).

4.4. Switching Schema

In this model, an existing schema is used as the switching mechanism among multiple schemas. Instead of mapping between every pair in the group, each of the individual metadata schemas is mapped to the switching schema. This model reduces drastically according to the number of mapping processes required. The switching schema usually contains elements on a fairly broad level. Examples of using switching schemas include the Picture Australia project and the Open Archive Initiative (OAI). Both use the Dublin Core as the switching schema.

4.5. Lingua Franca

If no existing schema is found to be suitable for use as a switching schema, an alternative is the use of a lingua franca. A lingua franca acts as a superstructure, but is not a "schema" in itself. In this method, multiple existing metadata schemas are treated as satellites of a superstructure (lingua franca) which consists of elements common or most widely used by individual metadata schemas. This model facilitates cross-domain searching but is not necessarily helpful in data conversion or data exchange. However, the lingua franca model allows the retention of the richness and granularity of individual schemas.

The lingua franca superstructure is built from a set of core attributes that are common to many or most of the existing schemas used by participants in a digital library or repository. An example is the ROADS template, which uses a set of broad, generic attributes.

5. Harmonization

Harmonization is refers to the ability of different systems to exchange information about resources. Metadata created in one system and then transferred to a second system will be processed by that second system in ways which are consistent with the intentions of the metadata creators (human or software).
Different forms of Harmonization,

5.1. Extensibility

The ability to create structural additions to a metadata standard needs application-specific or community-specific. Given the diversity of resources and information, extensibility is a critical feature of metadata standards and formats.

5.2. Modularity

The ability to combine metadata fragments adhering to different standards. Modularity metadata extensions from different sources should be usable in combination without causing ambiguities or incompatibilities.

5.3. Refinements

The ability to create semantic extensions, i.e., more fine-grained descriptions that are compatible with more coarse-grained metadata, and to translate a fine-grained description into a more coarse-grained description.

5.4. Multilingualism

It has ability to express, process and display metadata in a number of different linguistic and cultural circumstances. One important aspect of this is the ability to distinguish between what needs to be human-readable and what needs to be machine - processable.
Harmonization then refers to the ability to use several different metadata standards in combination in a single software system. The rest of the deliverable will analyse the different groups of standards and try to find obstacles to harmonization.

6. Conclusion

In the open, networked environment enable multiple user communities using a multitude of standards for description of digital resources, the need for harvesting among metadata schemas is over-riding. Currently, mapping metadata schemas still require enormous effort even with all the assistance computer technology can provide. If the

information community is to provide optimal access to all the information available across the board of digital libraries and depositories, information professionals must give high priority to the task of creating-and maintaining-the highest feasible level of exchange methods among schemas and new information services.

7. References

[1] ALCTS/CCS/Committee on Cataloging: Description and Access Task Force on Metadata. (last updated 2004). Summary Report.< http://www.libraries.psu.edu/tas/jca/ccda/tf-meta3.html>

[2] Guenther, Rebecca, and Sally McCallum. (2002). New metadata standards for digital resources: MODS and METS. ASIST Bulletin, 29(2)

[3] Heery, Rachel M , Andy Powell, and Michael William Day. (Mar. 1998). Metadata: CrossROADS and interoperability [computer file]. Ariadne (Online) no. 14

[4] Gillman, D.; M. Appel; and S. Highsmith. Building A Statistical Metadata Repository, Second IEEE Metadata Conference 1997: http://computer.org/conferen/ proceed/meta97/papers/dgillman/ dgillman.html

[5] LaPlant, W. P. Jr., Lestina, G. J. Jr., Gillman, D. W., and Appel, M. V. (1996), "Proposal for a Statistical Metadata Standard", Census Annual Research Conference, Arlington, VA., March 18-21, 1996.

[6] Consortium for the Computer Interchange of Museum Information (CIMI): <http://www.cimi.org/downloads/ CIMI_profile/profile.htm>.

[7] Baker, T. & Dekkers, M., (2002), CORES Standards Interoperability Forum Resolution on Metadata Element Identifiers. http://www.cores-eu.net/interoperability/cores-resolution

G. Sivaraman is a Faculty member in the Department of Computer Science at M.G.R. College , Hosur. He received his Master's Degrees from Madurai Kamaraj University, India in 2000. He received his M.Phil in Computer Science from Bharathidasan University, Tiruchi in 2004.

DR. K. Thangadurai is a Faculty member in the Department of Computer Science at Government Arts College (Men), Krishnagiri. He received his dual Master's Degrees from the Bharathidasan University, India in 1996 and 1999 respectively. He received his M.Phil in Computer Science from M.S University, Tirunelveli in 2002 and received Ph.D in Computer Science in Vinayaka Mission's University, Salem, India in 2009. His research interests include Software Engineering, OOAD and cloud computing.