Published in Homayoun, Behrous and other; ed. Workshop Proceedings of *First Euro-Asian Conference on Advances in Information and Communication Technology (EURASIA ICT 2002),* Shiraz, Iran. October 29-31, 2002: Pp 487-491.

Ontological Metadata Approach for Accessing Distributed Web content: A Proposed Model for Bibliographic Databases

Yatrik Patel
Scientist - B
(Software R&D Group)
yatrik@inflibnet.ac.in

J K Vijayakumar
Scientific Technical Officer
(Database Dev. Group)
vijay@inflibnet.ac.in

Badnapuri Ramesh
Scientist - B
(Web R&D Group)
ramesh@inflibnet.ac.in

Information and Library Network (INFLIBNET) Centre / UGC
Gujarat University Campus, PB 4116, Navrangpura, Ahmedabad – 380 009, Gujarat, India
Phone 91-79-630 4695 / 630 5971 Fax 91-79-630 0990

Abstract

Content in Distributed servers, located at various locations, mostly non homogenous in nature with different platforms, different OS, different format becoming a big challenge for information professionals. This paper proposes a model for how to map relational bibliographic databases such as library databases/catalogues and bring these content in to a single homogenous platform irrespective of its multiple platforms and provides access through a single (gateway or portal) and what are the tools and techniques to be used. A detailed discussion on tool for data mining from various RDBMS, a common pool for data mapping, interfacing non homogenous data mines at distributed locations through ontological Metadata approach and display through XML etc are presented in this paper. The implementation of the proposed model by data mining, mapping the bibliographical relations, pool for interfacing, and mechanism of accessing the content through are also described.

0 Introduction

The huge volume of content available in the web currently is distributed nature and not extracted semantically and cognitively. More meaningful content creation and mapping features are necessary for tackling the problems. Over the past 5-10 years we have seen dramatic improvements in information accessibility, speed, usability, and fact-finding. Expectations have been raised concerning what is possible. However, advancements in the ability to synthesize knowledge from data have not kept pace with the ability to amass information. Content, technology and integration, business and economic, and people issues challenge us in increasing the use of these tools. This paper describes an ontological based metadata approach model of an interface which can be made available on the Web for browsing any knowledge base including relational bibliographic databases, particularly heterogeneous in nature and distributed over a number of servers located different places with different syntax and formats.

1 Descriptions of the Terms

1.1 Metadata

Metadata is a primary tool and the key part of the information infrastructure necessary to help create order in the chaos of the Web, infusing description, classification, and organization to help create more useful stores of information. *Metadata* - data about data - or more commonly "descriptive information about Web resources". The use of standardized descriptive metadata can substantially improve the discovery and retrieval of relevant networked resources. Yet there is much confusion about how metadata should be integrated into information systems, particularly for heterogeneous and distributed systems. Content-related meta-data plays an important role in intelligent information systems. Especially on the Web, meaningful meta-data describing the contents of a web site is the key to intelligent retrieval and access of information. Meta-data description standards like RDF and RDF schema have been developed and work in progress addresses the use of ontologies to provide a logical foundation for meta-data. The need to improve the effectiveness of searching for information resources on the World Wide Web has prompted the development of simplified

metadata standards which could be used by authors or others at relatively low cost.

Dublin Core is being developed as a generic metadata standard for use by libraries, archives, government and other publishers of online information. The Libraries recognises that this standard may be applied broadly to citation and full text descriptions, and may support interoperability between a range of schemas, including US MARC. The Dublin Core standard is still under development, OCLC is responsible for development and maintenance of this standard. The Dublin Core element set defines a set of metadata elements for cataloging library items and other electronic resources. Such items are known as "resources", and there exist certain relationships (Dublin Core calls these relationships "elements") between resources and other resources or data. In this paper an attempt has been made to proposes a strategy which generate metadata based on Dublin Core, for any type of content irrespective of its format, location, type etc [1].

1.2 Ontological Approach for Metadata

Ontologies constitute the foundation for very many intelligent systems nowadays. They have gained popularity for very different needs of groups like the World Wide Web community, the database community and the machine learning community. The idea of ontology facilitates computational reasoning about the data. Ontologies are more than a particularly elaborate approach to the description and classification of information. They can be used to support the operation and growth of a new kind of digital library, implemented as a distributed, intelligent system. In artificial intelligence, “ontology” is a set of vocabulary definitions that expresses a community’s consensus knowledge about a domain. This knowledge is meant to be stable over time, and reused to solve multiple problems. Note that this concrete and utilitarian approach is quite different from “ontology” in philosophy, which concerns the abstract nature of reality apart from human endeavor.

Loom is a language and environment for constructing intelligent applications. The heart of Loom is a knowledge representation system that is used to provide deductive support for the declarative portion of the Loom language.

1.3 Formats Available

From traditional libraries point of view, cataloguing codes, such as, the Anglo-American Cataloguing Rules (AACR2) and the Classified Catalogue Code, provide guidelines the choice, rendering and description of the elements for conventional documents and other medias, so as to ensure consistency, and facilitate data exchange. There are also the International Standard Bibliographical Description (ISBD), the Common Communication Format (CCF) and others with associated documentation for the choice, rendering and description of bibliographical elements in a database. For content description on the Internet / Web, metadata formats, such as, the Dublin Core, MARC, and Text Encoding Initiative, and also markup languages, such as, HTML and XML are available.

1.4 Bibliographic databases

A bibliographic database can be a Library catalogue or a database of dissertations and theses, or a database of research papers published in technical journals or conferences etc. In bibliographic databases, the data stored comprises inputs of bibliographic details of a document for identification storage and retrieval purposes, under bibliographic records. A bibliographic record may comprise fields like document number, title, author(s), ISBN, publisher, year, imprints, source reference, abstracts, full text, indexing words or phrases, citation, local information such as classification number, book number, collection number, location etc. The organization of record data elements in a particular record is called Record structure for entering the information and display output in a particular database.

1.5 Storage techniques

Presently data is being stored as flat file structure or as in RDBMS structure depending upon volume of data as well as available resources in terms of computers as well as expertise.In flat filestorage peoples are using their own structure and interfaces to managedata,This is very useful for small to medium amount of data but while quantity is more one has to go for RDBMS as flat file structures becomes more and more complex. For RDBMS peoples are using their own table structures designed according to their convenience but still there is more flaxiblity and easy managablity is being supported by RDBMS like MS SQL, Oracle , and Sybase. As well as they are on multiple OS platform so, Interplatform migration becomes easy.

2 The problems

The holy grail for content providers and users alike is a seamless, integrated, transparent network that allows searchers to link quickly and painlessly to any document/record they seek. Users want easy access to content, may be full text, and content providers are continually breaking new ground in their attempts to give users what they want. This creates an exciting variety of options, but a variety that can confuse both information professionals and end users. Then too, when trying to find the full text of journal articles, the promises and advertisements of aggregators and publishers often seem inflated.

The big question is that how to bring content from a heterogeneous bibliographic databases to a homogeneous window.

3 The Proposed Model

The bibliographic databases may be available in different platforms with lot of variations in its syntax, formats, tags and their interpretation by individuals. The technology that emerged to process and manipulates data of various kinds is broadly termed as database management systems (DBMSs). The bibliographic relations in various databases are hidden in redundant data and in the meaning of natural language notes. In the knowledge base, bibliographic relations are explicit, labeled links amenable to manipulation by computers. Here the idea of Relational Database Management Sysytems (RDBMS) can be incorporated with the help of mapping technology and metadata. The complexity of bibliographic relations and the immensity of the task of recreating metadata in a new and relatively elaborate format are the big challenges here.

Mapping of records from one to another standard is the first step in this model. Here we take MARC as the standardized format for bibliographic information and assume content is available in distributed systems in other MARC formats, CCF formats etc. The steps are coded in Figure 1.

3.1 Conversion to MARC

Fields and tags used in a particular format need to be identified, their repeatability needs to be taken care with reference to main record and then using simple file conversion and string manipulation functions it can be easily transferred to the MARC format.

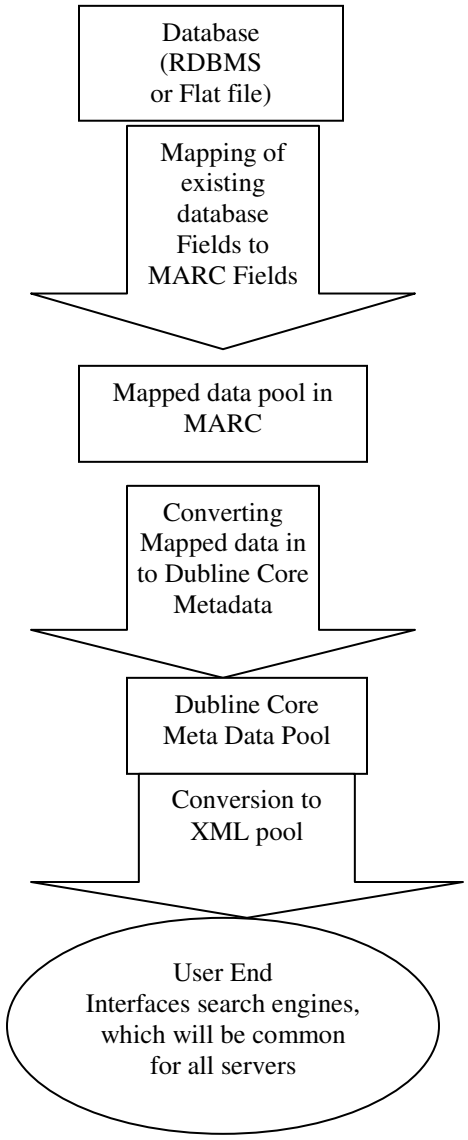

Figure.1

Flat file data.

1. Take a particular record, which contains all fields filled from the flat file. This can be retrieved by examining field structure/ directory structure (skeleton of flat file).
2. Bifurcate this record in to different chunks (packets) according to fields definition (i.e. precisely distinguish

each component (field) from flat file record.

3. Map each of above component according to MARC standard tags.
4. By above process now you will be having all fields tagged according to MARC, now only things remains is to place it into MARC structure this can be easily done by reverse process of step 1 using MARC skeleton /file structure/tag structure.

Now to convert record from RDBMS one can easily follow following steps

1. Study the database structure , Normalisation and relations,repeatability of RDBMS table structure.
2. Retrieve each and every associated field from table structure for a particular record. This will be easier compared to above (flat file) case as every thing is relational (provided proper relations are maintained while defining table structure).
3. Map these fields according to MARC
4. Put these fields according to MARC structure.

3.2 MARC to Dublin Core

We have to analysis and identify the entities that the descriptions in MARC data and with Dublin Core. It involves mapping data from MARC to the ontology, and reasoning about the data to establish relationships in the following steps:

1. Convert MARC data to tagged text. Then map MARC fields and values to ontology concepts with a set of control files that identify the MARC fields and codes, and establish priorities and conditions to resolve cases where several MARC values produce one ontology value, or where one MARC fields can produce several ontology values.
2. Extract coded attributes and values from natural language comments in the tagged text.
3. Convert the tagged text to Looms statements. At this stage every value is treated as a distinct object.
4. Reason about the data to establish relationships. First, merge matching conceptions, expressions, and manifestations. Then, based on transcribed clues and other information, deduce relations between works. When a preceding work is not in the collection, generate knowledge-base objects that reflect what we do know about it.

3.3 Encoding of Dublin Core in XML

Document structures help humans read documents, but they do little to help computers find the critical pieces of data they need without human assistance. On the other hand, XML data structures reflect the content directly, with little concern for where they appear in the overall document structure. Since XML allows for the creation of tag sets that are content savvy, an XML data structure can serve as a road map to information. XML allows users to search for and manipulate data found on Web sites to suit their own needs. XML is able to accomplish these structuring tasks through the language's ability to associate an XML document with a document type definition (DTD). The DTD is where the structured tags are actually declared and attributed to certain values. DTDs can be included at the beginning of the XML document, or maintained as an external file.

This section describes step by step, this method of how to create a document for the DCMES in XML.

1. XML declaration
   Any well-formed XML document should include a statement of the version of XML used (and content encoding). At present, the only valid version of XML is 1.0, It is therefore **strongly** recommended to include the statement
2 Referencing the XML DTD
   The DTD is where the structured tags are actually declared and attributed to certain values. DTDs can be included at the beginning of the XML document, or maintained as an external file.
3 Declaring the use of RDF
   It is necessary to declare that RDF is being used so that software can recognise this is an RDF/XML application. This declares the outer RDF containing tag with its XML namespace and the XML namespace for the DCMES elements.
4 Describing the resources
   The format can describe multiple resources therefore for each one, they must be

enclosed in a container element - a pair of rdf:Description tags. Resources may have no, one or several identifiers and some of these may be URIs. If a resource has at least one URI,

This can be repeated for all other DCMES elements that are needed with the standard Dublin Core guidelines - all elements are repeatable and optional. Note that there is no requirement on applications consuming this document to preserve the order of elements in the container and therefore you should not expect this to be preserved.

5 Language and character encoding

XML provides an xml:lang attribute that can be used on any element. This provides a way to describe the language used for the *content* of the element. The DCMES provides a *Language* element which is used to describe the language of the *resource*.

The values of the elements and attributes will need to be encoded using the rules of XML when there are special characters in the value.

4 Conclusion

The authors expect that, if this model is a successful one, the problems faced by librarians and users for accessing bibliographic information about the holding of other libraries. Towards the global bibliographic information exchange will be somewhat easy, even in digital environement.

Acknowledgement

The authors are very thankful to Dr T AV Murthy, Director and Mr. S M Salgar, Scienitst G INFLIBNET Centre, for their encouragements and supports.

Published in Homayoun, Behrous and other; ed. Workshop Proceedings of *First Euro-Asian Conference on Advances in Information and Communication Technology (EURASIA ICT 2002),* Shiraz, Iran. October 29-31, 2002: Pp 487-491.