Building a Consortium of Digital Libraries Using Open Source Protocols

D.S. Bhilare
Head, Computer Centre Devi Ahilya University, Indore,

Abstract

Digital libraries are growing across the world exponentially, particularly after the availability of open source software. These libraries are being built and utilized in isolation. It is proposed that depending on interest groups a cooperative organization structure of these libraries may be formed. This would minimize the duplication of the resources and enable effective utilization of the resources among member organizations. For an end user these libraries will appear as a single library. An architecture has been proposed to enable this kind of environment, where information can be retrieved and shared. Our architecture enables participation of existing digital libraries, in the cooperative digital libraries environment. This is achieved by generating and managing metadata, which can be harvested to the service provider for satisfying end user queries.

KEYWORDS: Digital library, federated searching, harvested searching, cooperative library

INTRODUCTION

Digital libraries are growing very rapidly, particularly in the academic institutions, where most of the universities have already built or in the process of building their own digital libraries. There is a need for a suitable architecture supported by an open source protocol, which can permit access to all these digital libraries treating them as a single large integrated library. There should be a common interface across the libraries from users point of view i.e. a query submitted to any library should fetch and present the information from all the libraries, which are member of the cooperative and follow the proposed architecture. In order to include existing libraries, which were built before the introduction of the proposed metadata and harvesting standards, the proposed architecture enables these libraries also to join the cooperative of digital libraries by generating the metadata compliant to existing standards. However, there are potential problems when it comes to group searching which mainly consist of [5]:

- Finding the most appropriate databases from which the user request can be successfully satisfied and
- Integration of all the results into an agreed upon global format, that can be consistently presented to the end user.

Existing practices can be broadly categorized into the following two categories:

- Federated searching
- Harvested searching

Federated searching

In a federated search, a query fired by a user is sent to different search engines corresponding to different digital libraries. Each search engine then executes the query, processes it and fetches the possible outcomes from its database server. The results are then integrated and passed on to the integrator. The set of all the results from all the digital libraries is provided to the integrator module for further processing. This set can then be passed on to the client end and presented according to the agreed global format.

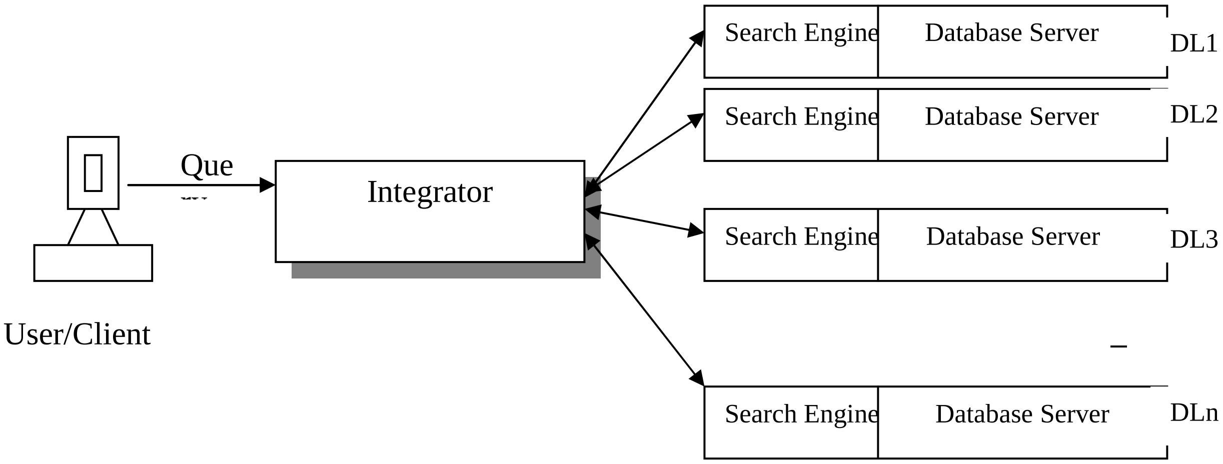

Fig. 1: Federated Searching

In federated search, handling issues related to the management of integrator module are critical. The key issues are identification of the optimal digital library and integration of all the results from each digital library. Other issues include the query processing time and duplication of results at the integrator module. The main problems can be summarized as follows:

- Identification of the optimal digital library
- Integration of results from various member libraries
- Query processing time may differ depending on the size of the database of that digital library
- Reliability of individual search engine
- Duplication of results

The integrator module has to first analyze the fired query, establish a connection to the different search engines and send the query to these search engines. Each search engine has to first confirm the connection request, then convert the query into a form acceptable by it, produce the results and then send them to the integrator. The integrator has to integrate the results and present it to the client.

Harvested searching

The proposed architecture uses harvested searching to establish cooperative digital library model. In a harvested searching, generally called harvesting, all the metadata from all the digital libraries present at distinct remote places is gathered at a common location. This metadata information then plays a vital role to find out the digital object requested by the user.

With harvesting, we have less network traffic as the metadata from all the digital libraries is presented well in advance to the central common place and thus depending on the fired query it can be decided which digital library to be searched for finding the results, hence less network traffic. Moreover, harvesting is a robust light-weight protocol.

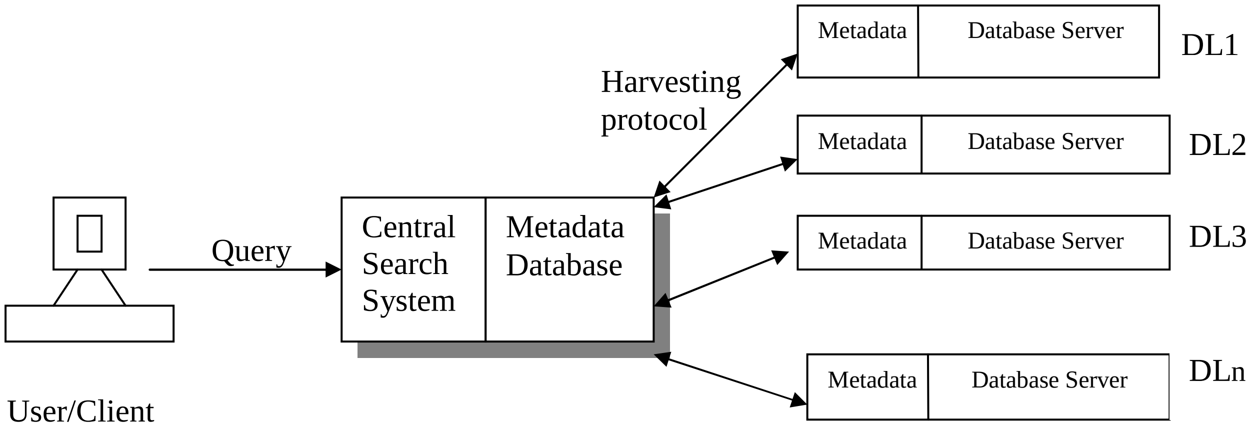

Fig. 2 Harvested Searching

RELATED WORK

The need for having cooperative digital libraries has been felt for quite some time. An ad hoc program over HTTP or Z39.50(Version 3) is one most common approaches for constructing the integrator module. The Z39.50 is an information retrieval service protocol proposed in 2005 to have search and retrieval of information from various digital libraries. In Z39.50 there are no service providers, there are only data providers.

In October 1995, OAI abbreviated for Open Archive Initiative was established for building an interoperability framework across digital archives [1]. An OAI service provider “Arc” is developed for cross-archive searching. In Arc harvester design, the OAI protocol is used to harvest digital archives.

The collaboration in the digital library initiative has included work with the University Libraries, the departments of Electrical Engineering and Technical Communications in the College of Engineering, and the School of Library and Information Science [7]. Faculty in the School of Library and Information Science are interested in partnering in digital library development and contributing to our understanding of retrieval issues [6].

The OAI protocol has separate data provider and service provider. The data provider implements the OAI protocol on archive to allow external access to data whereas service provider uses the OAI protocol to access external archives and provide services like searching and linking on their metadata. Unique identifiers for each record, date-time stamp for each record when last modified, created or deleted and HTTP server with scripting facility are the basic requirement for OAI protocol.

This paper presents a vision of the start of a collaborative, digital academic law library, one that will harness our collective strengths while still allowing individual collections to prosper. It seeks to identify and answer the thorniest issues - including copyright - surrounding digitization projects. It does not presume to solve all of these issues. It is, however, intended to be a call for collective action, to stop discussing the law library of the future and to start building it [8].

A standard protocol called OAI-PMH, a protocol for metadata harvesting is being used now a days for accessing metadata archives. This protocol is used in an open source digital library software called Dspace developed by MIT-libraries and HP(Hewlett Packard) [3]. Dspace system is used by almost all the universities and educational organizations for having their digital repository. Hence OAI-PMH is becoming the de-facto standard for accessing metadata archives [4].

SYSTEM ARCHITECTURE

For a harvested searching, it is required that the digital libraries should expose the metadata related to the digital contents and the metadata standard used by the central search system. With OAI-PMH, each client is having a software called harvester which helps to issue the OAI-PMH request for knowing the metadata of digital library. This request is first sent to the service provider, which then passes it on to the identified digital library based on the parameters in the request. There is a module called data provider at the digital library end, which exposes this metadata to the harvester via the same service provider.

Hence in OAI-PMH it is the data provider that is solely responsible for exposing the metadata to the harvester. As all the digital libraries do not confirm to a particular metadata standard (like Dublin Core (DC), Metadata Encoding and Transmission Standard (METS) or XMLMARC), there are problems to include these digital libraries in the cooperative infrastructure [2].

The digital libraries that want to be a member of the cooperative digital libraries group has to first ensure that they have some mechanism to expose their metadata information to the central search system. Thus all the digital libraries that were developed before the evolution of the standards in the digital library creation and management need to have a module to help them harvest their metadata. This module should be easily integrated with the existing framework of the current digital library.

The proposed module called “Metadata Management Module (M3)” for mapping metadata, manages the mapping relationships between the information present in the digital library to the standard form like DC or METS. The M3 aims at identification of all the metadata and transforming all such metadata into the standard metadata format like DC or METS. Hence the digital libraries which are already in existence from a long time but do not follow any metadata standard can also participate in the cooperative digital libraries by continuously mapping data to desired metadata standard. Thus, these existing non standard digital libraries could also be able to harvest their metadata to the harvester.

The proposed architecture mainly aims at continuously generating metadata from the digital library by always keeping an eye on the type and format of the contents present therein. This is accomplished by a watch-dog program. The architecture has two modules:

- **Metadata Generation Module(MGM)**
- **Transformation Module(TM)**

The watch-dog program is present in the MGM. This program is continuously analyzing the database for any additions, deletions or updates in the database. The watch-dog program periodically interacts with the generator module and thus helps it generate the metadata for the complete database. The main purpose of it is to maintain a log of all the information in the digital library so that it can be used by the administrator to decide the policies for the management of the digital library. With all such information the administrator can effectively maintain and update the digital library. The generator module is mainly responsible for metadata generation. This metadata is generated depending on the contents of the digital library. This is carried out by checking how frequently a term is used in the digital object. The mapping is thus completely dependant on the contents of the digital object. Hence the watch-dog program has to continuously interact with the generator module to make it available the information about the added, deleted or updated digital objects.

The intermediate metadata thus generated by the MGM is forwarded to TM as raw data. This metadata need not have any of the elements from the metadata standard formats like DC or METS. The TM module accepts raw metadata as input and filters it to the standard format say DC. The other metadata is maintained in a separate database so that it can be used by the watch-dog program in the generator module to avoid generating this metadata. However this metadata information can also be used by local search engines to handle the local requests.

Hence the TM besides generating the standard metadata is helping to satisfy local queries by maintaining remaining metadata in a database. The standard metadata is also maintained in a database and can be exported in XML format to the data provider.

Design of the Metadata Management Module(M3):

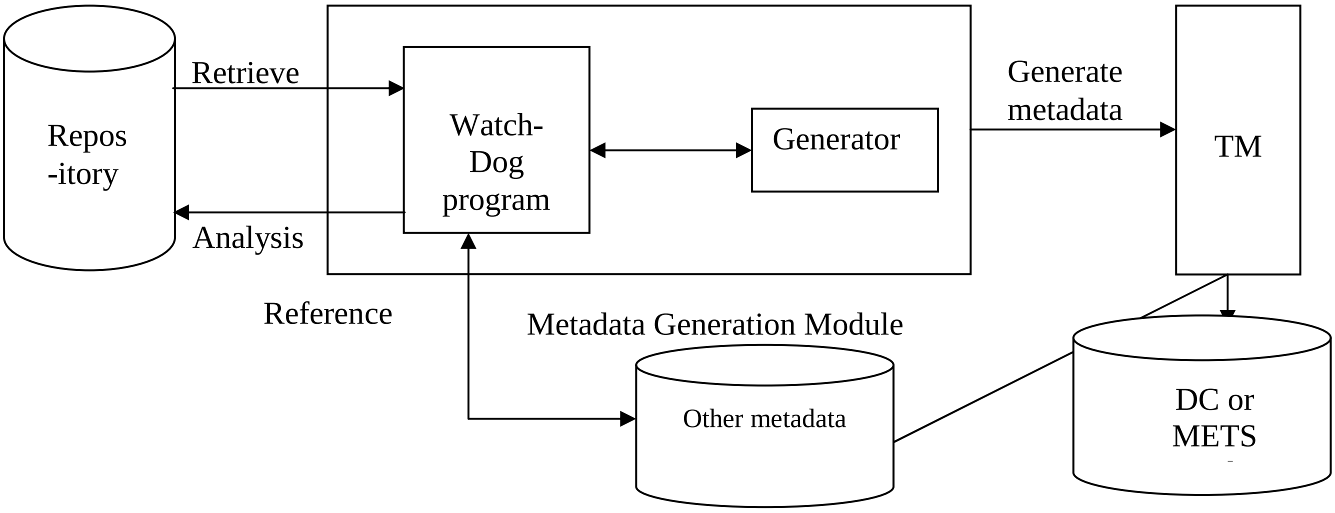

Fig 3: Design of Metadata Module (M3)

The harvesting approach with the use of M3: The proposed Extended OAI-PMH protocol for cooperative digital libraries allows any digital library to use harvesting protocol

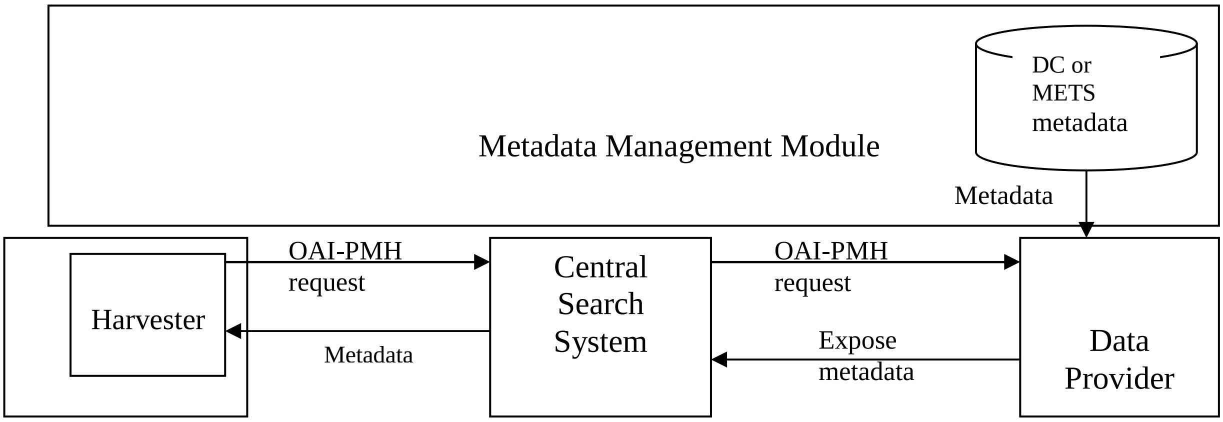

Fig. 4: Extended OAI-PMH Protocol

The proposed architecture ensures the participation of any digital library into the cooperative digital libraries infrastructure that use OAI-PMH harvesting protocol for metadata harvesting. This helps any digital library to supply the information asked by any client if that digital library has that information. To supply the information the metadata related to the information is generated and harvested to the central search system which actually satisfies the client requests. The generation and management of metadata is carried out by the proposed module called M3(Metadata Management Module) which works at the corresponding digital library present at remote location. The efficiency of M3 mainly depends on the database size of that digital library and the processing speed of the server wherein this module is running. So each M3 module at corresponding digital library works at its own speed to generate the metadata. However the data provider has to expose the metadata as soon as a OAI-PMH request comes to it. If the metadata is not available at that time the request could not be satisfied. But it may happen that the metadata becomes available the immediately after saying no to the service provider.

This problem can be eliminated by deciding global rules in addition to the local rules while adopting the M3 module. The local rules are the strategies explored to solve the issues related to the corresponding digital library. The global rules ensures the management of M3 at a higher level in hierarchy by describing the abstract classes and interfaces to be used in M3.

CONCLUSION

Thus the idea of having cooperative digital libraries can be accomplished by either federated searching or harvested searching. In a federated search, we have studied various issues involved in the management of the integrator module for searching the databases from different digital libraries. In harvested searching, the proposed architecture allows any digital library to participate in cooperative digital libraries infrastructure. It is specified that how metadata can be generated and send to the data provider for exposing it to the service provider.

REFERENCES

1. The Open Archive Initiative. http://www.openarchives.org
2. Dublin Core Metadata Initiative. http://purl.org/DC/
3. Dspace digital library. http://www.dspace.org
4. Van de Sompel, H. and Lagoze , C. The Open Archives Initiative Protocol for Metadata Harvesting. Open archives Initiative, 2001. Available at http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html
5. Lagoze., C., and Davis, J. R. Dienst – An architecture for the distributed document libraries,
   in Commun. ACM
6. Twidale, M., Chaplin, D., Crabtree, A., Nichols, D.M., O'Brien, J. and Rouncefield, M., " Collaboration in Physical and Digital Libraries", British Library Research and Innovation Report No. 64
7. Greg Zick, Geri Bunker, "Collaboration as a Key to Digital Library Development**"**, D-Lib Magazine
   March 1999, Volume 5 Issue 3 ISSN 1082-9873
8. Michelle M. Wu, "Building a Collaborative Digital Collection: A Necessary Evolution in Libraries", 2011, *Law Library Journal, Vol. 103, p. 527-551, 2011*