Unified Access To Heterogeneous Resources In A Distributed Library Consortia Environment: Challenges And Opportunities

M. Paul Pandian

Scientific Officer (Library)
Institute of Mathematical Sciences,
CIT Campus, Taramani, Chennai-600 113, India
Tel: 0091-44-2541856
Email: pandian@imsc.res.in

C.R. Karisiddappa

Chairman and Professor
Department of Library and Information Science
Karnatak University, Dharwad 580 003, India
Tel: 0091-0836 – 747121
Email: karisiddappa@yahoo.com

ABSTRACTS

Library consortia in today's digital age are quite different from that of library networks in yester years. The main reason is that the resources that are shared in today's consortia environment are predominantly in electronic form such as electronic journals and databases. Hence the technology and associated tools to support sharing the electronic resources are also important components for the success of any library consortia. It is essential that each participating libraries of a consortium is equipped with necessary and sufficient technology to support sharing the resources across. Unified access to heterogeneous resources is one of the greatest challenges that library consortia face. The paper looks at the unified access system to search heterogeneous resources in a distributed library environments. The paper discusses the challenges and opportunities of unified access systems by looking at some of the major international initiatives.

KEYWORDS

Unified access; heterogeneous resources; library consortia ; international initiatives; opportunities ; challenges

INTRODUCTION

Libraries have always striven to share resources through union catalogues, indexing and abstracting services and interlibrary loan, but the Internet has provided opportunities for unmediated access to distributed resources in ways not dreamt of until a few years ago. As an increasing proportion of information is made available in digital form, libraries are seeking new system solutions to the problem of providing a coherent view of the range of electronic resources available to their users. These include resources freely accessible on the Internet as well as subscribed CD-ROMs and commercial online services. In addition, many libraries have their own online collections acquired through digitisation programs. As the number of electronic resources that a library provides access to grow, so does the number of different interfaces from varying sources a user must learn and attempt to navigate.

On one hand, libraries subscribe to many types of database retrieval systems that are produced by various providers. The libraries build their data and information systems independently. This results in highly heterogeneous and distributed systems at the technical level (e.g., different operating systems and user interfaces) and at the conceptual level (e.g., the same objects are named using different terms). On the other hand, end users want to access all these heterogeneous data via a union interface, without having to know the structure of each information system or the different retrieval methods used by the systems. Libraries must achieve a harmony between information providers and users.

The term integration may have part of its origin in the library world as an application and concept called integrated library systems. As a further use of the word integrated, Carol Tenopir has written that “in an integrated reference environment, a common interface leads users seamlessly to the best resource for their needs. It may not be obvious to the user whether the database is on the university’s resident computer, at another university,

at an online vendor's office 2000 miles away, or on the other side of the world at an Internet site .

In an increasingly complex and global information environment, an integrated library is of vital importance in enabling end users to search through large quantities of information. An integrated library means creating intelligent search, retrieval and presentation tools and interfaces for users; it means incorporating new information types, metadata and document encoding schemes. It also means new hardware and software systems which are capable of interpreting users' requests, including selecting from multiple databases. In the integrated environment a user interface should be a simultaneous gateway to the electronic and traditional collection of the library and to all central resources. An end user will not have to install different software clients and he will not be assigned different usernames and passwords.

In many digital library projects the access to heterogeneous information resources is a major issue, as users prefer a unified interface to the available information and staff will be relieved of too many routine instructions to individual databases. These access systems encompass four major components: access organisation, access control system, pricing module and billing/clearing/charging system. The so-called unified access means access to heterogeneous types of resources via HTTP and Z39.50 protocols. The users will have authorized access via a unified, integrated and comfortable interface. Most access systems consist of some type of metadata database access to the (linked) heterogeneous information resources.

LIBRARY CONSORTIA IN THE ELECTRONIC AGE

Library consortia in today's digital age are quite different from that of library networks in yester years. The main reason is that the resources that are shared in today's consortia environment are predominantly in electronic form such as electronic journals and databases. Hence the technology and associated tools to support sharing the electronic resources are also important components for the success of any library consortia. It is

essential that each participating libraries of a consortium is equipped with necessary and sufficient technology to support sharing the resources across. And ideally, these technology tools must be integrated into the library automation software that the libraries are already using. Though the resources are accessible on the internet, the formats in which the resources available are different for different information providers. Each one has their own resource discovery system or search engines, the content display, the download options etc. When a member library subscribes to various sources under consortia through many different information providers, the end user in the library will have to repeat the search for the information, which he is looking for in every resource discovery system of the various information providers so as to get a comprehensive list of search results. This will be time consuming and laborious.

UNIFIED ACCESS

Unified access refers to a process in which a user submits a query to numerous information resources. The resources can be heterogeneous in many aspects: they can reside in various places, offer information in various formats, draw on various technologies, hold various types of materials, and more. The user's query is broadcast to each resource, and results are returned to the user. The development of software products that offer such simultaneous searching relies on the fact that each information resource has its own search engine. The simultaneous searching product transmits the user's query to that search engine and directs it to perform the actual search. When the simultaneous searching software receives the results of the search, it displays them to the user. Simultaneous searching is also known as integrated searching, metasearching, cross-database searching, parallel searching, broadcast searching, and federated searching.

User Level:

Consider a user of a participating library who wishes to search and find out all relevant information from any available sources that might satisfy his or her particular information need in a distributed library consortia environment. The user is most often concerned

about the relevance and timeliness of information, not about, which source the information comes from, which data model it adheres to, or which query interface was used to retrieve the information. The user should not need to be aware of the technical underpinnings of the system, nor be limited as to what type of information he or she can access or from what type of sources. Ideally, there should be a uniform interface for expressing common queries for multiple information types and a single, consolidated set of results consistently ordered regardless of the particular scoring mechanism used by each source. The user should also be allowed to select, compare, cluster, and otherwise analyse information sources at a meta-level.

System Level:

Unified access at system level will provide:

- A single point of access
- Unified login (including one user ID)
- One common user interface, i.e. one presentation structure
- One uniform user-friendly retrieval system
- Direct access to electronic media and unified request service
- Patron-initiated online requests of resources and interlibrary loan facilities

UNIFIED ACCESS – CHALLENGES

Most libraries today are, in fact, already hybrid libraries -- they own and subscribe to a range of resources and services which are supplied in a variety of formats and media: print monographs and serials, electronic journals, abstract and indexing services on CD-ROM, music CD-ROMs, etc. Many electronic resources are accessed on remote servers. An increasing number of end-users are also accessing these services from outside the "home" institution: users and services are both distributed. However, there is currently no uniform way of managing and providing integrated access to these hybrid resources.

Users are forced to interact with each service individually and waste time in repeating the same steps to search different systems. At the same time, using different interfaces also increases the risk of inefficiencies -- such as failure to discover relevant resources because of unfamiliarity with one service's idiosyncrasies. In a discover/locate/request cycle, the user will also be forced to re-enter the same data when he or she moves from one stage to the next.

To support these user requirements, a system for accessing heterogeneous information sources needs mechanisms for hiding the differences between sources, for identifying sources likely to contain relevant information, and for combining results. System scalability, extensibility and customisability are also important because the system needs to adapt as the environment or the user's needs change. Scalability is essential for handling the exponentially growing number of available information sources, as is the case on the WWW. No single WWW crawler indexes the entire WWW and no single corporate database or search engine provides one-stop access to global information. Extensibility is critical because new information sources and new interfaces, protocols, and formats emerge constantly. Users and system administrators should be able to extend the system with minimum manual effort. Ideally, the system should be able to discover new information sources automatically and know how to talk to them. A system should also be able to adapt to a changing environment because information sources, especially external ones, typically enjoy autonomy and may change unexpectedly. An Unified access system should therefore not depend on other systems' data models, query languages and vocabularies, or the protocols and interfaces they support. Internally, an unified access system therefore needs to employ a data model and query language that are rich enough to subsume the data and query representation capabilities of other systems, and it should provide a flexible abstraction mechanism which hides the interface details of an information source. Customisability is important because it is unrealistic to expect a single generic system to be able to handle all application domains and information seeking tasks as competently as a tailored system. Therefore, an architecture in which individual components can be customised to a particular application domain is desirable..

Much of what described above requires rich meta-information. Meta-information is critical to the success of an extensible heterogeneous system and should preferably be extracted automatically. Since the quantity and quality of meta-information that can be automatically extracted from a source vary, the system should be able to cope with missing or incomplete meta-information as well. Ideally, meta-information about a source would include all of the following: the identity of a source (e.g. its name, type and location); service parameters (e.g. cost, response time, reliability, availability, etc.); content descriptions (e.g. major topics covered in a text collection); retrieval functionality supported (e.g. truncation patterns and stopword lists); and schema information, i.e. the structure of the information types supported by a source (e.g. names of information structures, names and types of attributes, etc.). Schema information has particularly important uses in a heterogeneous information system. First, it guides the system in merging structured, disparate pieces of information. Second, it allows document retrieval to be more targeted and document attributes to be matched with query conditions according to their type.

Barriers to access to heterogeneous resources in a consortia environment include multiple and confusing authentication and registration procedures. Users will not return to services that are time-consuming and difficult to use. In the print environment issues of ownership and access are very clear to our users, this is not true in the electronic environment where access may not necessarily include the ability to browse, download and print. Our users may also not have the necessary knowledge and skills to utilise the diversity of resources, and they may simply not be aware of the range of resources available.

The success of the unified access to heterogeneous resources will be heavily dependent on the use of agreed standards implemented in agreed ways. Standards will:

- Provide interoperability and communication across resources and services
- Provide for the consistent description of electronic data

- Provide consistent, unambiguous interpretation of queries and results
- Provide consistent return and display of results and data

However, at the moment there is an absence of standards in key areas. One of the main influences here is the fact that commercial interests are controlling the developing technologies such as the web, distributed processing, broadband networks etc.

The distributed library environment presents a wide range of service challenges. It is easy to make the assumption that the primary challenges are technological but this is to ignore the impact of a technologically dynamic and diverse environment on users.

UNIFIED ACCESS – OPPORTUNITIES

At the lowest level of integration, the World Wide Web now enables heterogeneous information services to be presented to a user through a simple menu-driven interface. In this model, the developer's role is to select appropriate services to list and to ensure the links to online services remain current. This is an appropriate way of building an information map to existing services where standards or interfaces are not yet in place to provide a higher level of integration; or where there is no business need for a higher level of integration. The services may be so heterogeneous that it is always appropriate to search them through a separate interface.

One way of providing unified access is to use broker architectures to integrate access to the library catalogue and the library's digital collections through standard protocols. Another is to build a central set of indexes for resource discovery purposes. These are appropriate solutions for integrating access to disparate collections. Within a single library, however, there is a need for a collection management architecture that can provide full system support for the "hybrid library": collecting, storing, managing and delivering access to information resources regardless of format. This architecture may consist of separate modules, including a digital collection management module, which

can be integrated through the use of appropriate protocols. Such architectural model may have the following major components, which facilitates building robust, scalable and interoperable heterogeneous distributed library systems:

- Client Desktop
  - World Wide Web browser
  - Telnet client
  - Application-specific desktop client
- Service-specific functionality
  - User profiles
  - Application services
  - Searching
  - Analysis
  - Result set management
- Integrating components
  - Citation linking
  - Search management
  - Format & display normalizing
- Server tools
  - Middleware (web/database integration)
  - Database
  - Custom tools
  - Mediators and converters
- Protocols and infrastructure
  - Transport
    - HTTP
    - TCP/IP
    - Telnet
    - Storage Request Broker

    - Information management
        - Z39.50
        - ODBC, Corba
        - SDLIP
    - Directory services
    - Security
        - X.509 authentication
        - SSL, Kerberos, PGP, SSH...
- Object metadata (EAD, Dublin Core, Open Archives, CDL, A&I, FGDC...)
    - Descriptive
    - Structural
    - Administrative
- Persistent identifiers
- Digital objects
    - Structured text (SGML, XML)
    - Semi-structured text (HTML, TeX)
    - Unstructured text
    - Images
    - Proprietary objects
- Storage
    - HPSS
    - AFS/NFS
    - Local file systems

EXISTING INITIATIVES – INTERNATIONAL

When we look at some of the heterogeneous information systems proposed in the literature, the common to many of these systems is that they can be described using three-

tier mediator architecture. In this model, a user or application communicates with a mediator (sometimes referred to as a broker or agent) which retrieves and consolidates information from multiple sources. The mediator interacts with information sources via wrappers (also called proxies, adapters, or converters) which constitute an abstraction mechanism and thus contribute to the extensibility of the system. WWW search engines like Google can be viewed as information integration tools since they distribute a user's query to multiple individual search engines and consolidate the results by removing duplicate result items and normalising the relevance scores.. Outside the realm of WWW meta search engines, few heterogeneous retrieval systems aim to provide a consistent notion of relevance across different information sources.

According to Fang, there are three other approaches such as bibliographic control, database navigation system and union search platform for integrating distributed information of different types into one union system, and these three ways coexist within many library services. One substantive approach to metasearch (search across heterogeneous data) is to create a new application that integrates multiple search requests into a union search platform. He referred to two possible ways (core metadata based method and web based method) to meet the requirements for this approach. As Fang finds both strengths and weaknesses in both the methods, he proposed a framework of heterogeneous resources integration and retrieval system—MUSP. In MUSP, multiple metadata forms in one system.

Some of the initiatives in providing unified access to heterogeneous resources in a distributed environment are:

AARLIN (Australian Academic and Research Library Network) uses portal technology to provide a whole host of services for researchers, including:

- accessing a uniform search interface which will permit distributed searching of multiple electronic databases, websites, online library catalogues and other electronic information resources using a single search syntax;
- populating a web-based form with appropriate metadata and generating a document delivery request, if required;
- accessing a range of appropriate or extended services (including deep linking to full-text where available) using context sensitive reference software via the OpenURL framework;
- pushing to researchers the relevant 'information landscape' or suite of information resources as determined by their authenticated user profile;
- permitting users to personalise and refine their search 'environment' making it possible for them to suppress or expand the 'information landscape' pushed to them as a default;
- allowing users to receive literature alerts informing them of newly available material matching their specified search profile on a regular basis.

The AARLIN Service model is built around a national portal framework, which is linked to the local authentication systems of the participating universities. Both SOAP (Simple Object Access Protocol), which is an XML-based protocol for exchanging information in a decentralised or distributed environment and LDAP (Lightweight Directory Access Protocol), which is a protocol for accessing online directory services (usually X.500-based) were used for authorization system. The model provides push facility (matching user profiles with relevant information resources) that reduces the amount of effort required by users to access relevant information.

Another major component of the portal is the common user interface, which allows parallel searching of a diverse range of databases, information resources and websites using multiple protocols. These protocols include Z39.50, HTTP, and SQL. Thus, it would be possible for a user with a single search query to search across multiple citation

and full text databases, online library catalogues, Internet search engines, websites and subject gateways, and get a uniform search outcome from this parallel search.

ENCompass for Resource Access (a commercial solutions from Endeavor Information systems) integrates access to licensed and free resources such as:

- A & I databases
- e-journals
- e-books
- relevant web sites
- the local OPAC

ENCompass for Resource Access is intuitive to search across resources, via:

- HTTP searching for web-enabled databases
- XML searching for structured requests and receipt of information
- Z39.50 gateway searching for resources enabled with this protocol

Metalib of Ex libris' is a library portal application. It enables the user to access the institution's e-collections, to obtain relevant services, and to maintain a personalized environment. The following search functionality is supported:

- Z39.50 gateway and server. MetaLib can also access resources supporting ZING, Z39.50 International Next Generation specifications.
- HTTP and XML
- API (Any database with a defined API for searching, retrieving the number of hits and retrieving the resulting records can be searched via MetaLib).

The UIAS (Unified Information Access System) at California State University is defined as a single, easy to use, integrated, and coherent computer-based user interface which provides direct online access to or delivery of:

- print resources described in CSU Libraries' Online Public Access Catalogs and described in catalogs of libraries beyond CSU;
- print resources described in other bibliographic/abstract databases such as periodical indices;
- digital resources, including text, image, video, and multi-media;
- Internet-based resources including those on the World-Wide Web;
- guidance in the use and evaluation of information resources including access to self-paced information competence instruction.

The Agora hybrid library project is implementing many of the concepts developed by the MODELS (MOving to Distributed Environments for Library Services) project. MODELS is a UKOLN project, which receives funding from eLib and the British Library Research and Innovation Centre. It has been exploring ways of more effectively managing access to distributed heterogeneous information resources. The project has discussed technical and organisational issues with leading stakeholders via a series of focused workshops. Issues and requirements identified have led to the development of the MODELS Information Architecture (MIA), a technical framework based on a three-layer model, for talking about service components with a common language. It provides a tool which helps library managers to lever development, and guides systems developers.

DECOMATE II – a project of European Commission is to develop an end-user service, which provides access to heterogeneous information resources distributed over different libraries in Europe using a uniform interface, leading to a working demonstrator of the European Digital Library for Economics.

OCLC SiteSearch is a more or less complete set of software tools for developing large-scale digital libraries. It provides parallel searching and browsing of remote and local databases, tools for building local databases and defining tailored user interfaces. SiteSearch is comprised of a Database Builder and a WebZ gateway. The Database Builder provides easy-to-use software tools for building and maintaining local databases.

The records can be created and edited with a standard web browser. The set of specialised templates allows the builder to take advantage of multiple metadata standards and formats, SGML and MARC. These formats can be indexed flexibly, ie nothing is predefined. Data needs to be converted to ASN.1/BER notation for indexing. Conversion programmes exist for MARC and SGML. Access to the databases is based on Z39.50 as implemented in the WebZ component. In indexing and searching the Newton search engine is used, developed by OCLC and it is also used in their own database services such as FirstSearch. Boolean searching, proximity operators and wildcards, plus truncation, are available in the search engine. The size of the database is not limited. The WebZ provides a toolkit for building a web-based interface for electronic resources available on the Internet. Simultaneous searching and browsing of multiple databases is possible, along with result-set merging and weeding of duplicate records. The WebZ toolkit includes access control methods, allowing the library to decide which resources to make available to all, which to protect with password and IP-based methods. There is also a module for vocabulary- assisted searching to assist the users in their information searching tasks. SiteSearch is available both for Windows NT and selected Unix operating systems. WebZ can be integrated as part of a web server to provide the speed and security that is needed in this kind of application.

CONCLUSION

Numerous issues face libraries today. Libraries of all types are challenged to provide greater information access and improved levels of service, while coping with the pace of technological change and ever increasing budget pressure. The growth in the number and volume of electronic resources has created a new problem: how can the average library user identify the best resources to search for desired information and learn to navigate the disparate user interfaces to make effective queries?. In many cases, when libraries have invested significant sums to make commercial databases available to their patrons, they have experienced very low utilization of those resources. Faced with large numbers of

potential sources and interfaces, many users revert to the ease of using well-known Internet search engines – even though the quality of information returned is usually poorer than that available through specifically focused subscription databases.

A major library technology trend is the desire to integrate all library resources and services behind a single Internet presence with personalization features, allowing patrons a customized view into the library. The promise of a truly integrated environment in a heterogeneous world may not yet be a reality, but with the active involvement of all the stakeholders, significant progress has been made. Just a few years back, metasearch systems seemed like a dream; today they are already a building block in the information resource environment serving the academic and research community. No doubt, more and more distributed and heterogeneous information retrieval systems will be produced. So the differences of system, syntax, semanteme and structure among these retrieval systems will continue to exist for the long term. In this environment, improving interoperability (at the technical level and at the conceptual level) becomes urgent. The benefit of interoperability is that it makes it possible for libraries to produce more effective, flexible search platforms to integrate heterogeneous resources. Interconnectivity and interoperability are essential if any "seamlessness" is to be achieved in information provision. The role of information professionals in facilitating the creation of navigable information landscapes is now also central

REFERENCES

1. Corcoran, Miriam. The Hybrid Library: Revolution or Evolution. Retrieved in July 2004 from http://lirgroup.heanet.ie/events/Lirseminar140203_files/TextMostly/document.html

2. Edward H T Lim and Earle Gow. A new model for collaborative library service: the AARLIN project. Retrieved in July 2004 from http://conferences.alia.org.au/online2003/papers.html
3. Fang, L. A Developing Search Service, Heterogeneous Resources Integration and Retrieval System. Retrieved in July 2004 from http://www.dlib.org/dlib/march04/fang/03fang.html
4. Girke, Thomas. The one-stop shop: a single end-user interface for search and discovery across digital library collections. Retrieved in July 2004 from http://conferences.alia.org.au/online2003/papers.html
5. Hylton, Jeremy. Access and Discovery Issues and Choices in the Design of DIFWICS. Retrieved in July 2004 from http://www.dlib.org/dlib/march96/03hylton.html
6. Informia: a Mediator for Integrated Access to Heterogeneous Information Sources. Retrieved in July 2004 from http://www.ubs.com/informia
7. Lopata, Cynthia L.(1995). Integrated Library Systems. ERIC Digest. Retrieved in July 2004 from http://www.ericfacility.net/databases/ERIC_Digests/ed381179.html
8. Metalib (ExLibris). Retrieved in July 2004 from http://www.exlibrisgroup.com/metalib.htm
9. OCLC Sitesearch. Developing and delivering a virtual electronic library. Retrieved in July 2004 from http://opensitesearch.sourceforge.net/docs/helpzone/sitesearch/servover.htm
10. Russell, Rosemary. Agora: building the technological infrastructure. Retrieved in July 2004 from http://hosted.ukoln.ac.uk/agora/dissemination/articles/nral.html
11. Tenopir, Carol (1995). Integrating electronic reference. Library Journal. v.120, Issue 6. 39.
12. ENCompass. Retrieved in July 2004 from http://encompass.endinfosys.com/whatis/whatisENC2.htm

13. UIAS Status Report. Retrieved in July 2004 from http://uias.calstate.edu/UIAS_Status.html

Brief Biography of Author

M. Paul Pandian is Scientific Officer (Library) at Institute of Mathematical Sciences, Chennai. He was earlier the Librarian, Indian Institute of Management, Indore. He holds Associateship in Documentation and Information Science (DRTC, ISI). His research interests are digital libraries, library consortia, standards and protocols, web based information services, and digital copy rights.

Brief Biography (Co- author)

Dr. C. R. Karisiddappa is Chairman and Professor at Department of Library and Information Sciences, Karnatak University, Dharwad. Dr. Karisiddappa served as a convenor of the UGC Subject Panel and recipient of the Motiwale Best Teacher Award for the year 2000.

Dr. C R Karisiddappa is elected as the President of Indian Library Association, (2002-2004) a premier association committed to the cause of Library Movement and Development.