Article

Federated Search & Discovery Services : A Standards and Protocols Perspective

Dinesh Ranjan Pradhan and Miteshkumar Pandya, Scientists - B (LS)

Abstract

With advent of information and communication technology, drastic changes can be seen in the services offered by the libraries. One of the primary objective of the library is to serve their users in a time brand manner. In this age of information explosion there are lots of printed as well as digital documents that are stored in a library are available online either in open access or on subscription and it is very difficult to find relevant one in time. Federated search and discovery tools facilitate integrated searching for all types of materials from a single search box. The article describes some of the standards and protocols that can be used to streamline the operation of developing federated search engine and discovery tools.

1. Introduction

Today's libraries have evolved and expanded from repository of books, journals and other non-print materials in physical format to the hybrid libraries having access to vast collection of electronic content ranging from subscribed full-text e-journal and e-journal databases, bibliographic databases, e-books, e-theses, digitized manuscripts, images and photographs, artifacts of cultural heritage, e-learning materials, scientific datasets, etc, in addition to printed documents. One of the major challenges that the modern libraries faces is to find ways and means to select and offer seamless and coherent access to growing collections of heterogeneous resources to the users in a fashion that enriches learning and research experience, provides timely and convenient access to relevant and appropriate resources, exposes potentially valuable resources that otherwise might have been overlooked and enables users and the library to focus on a fruitful use of collections rather than dealing with different aspects of access, navigation and manipulation of the result sets. To resolve this issues tools like federated search / meta search tools and discovery services have come up.

The primary expertise of libraries is in the management of content as well as in creation of content. As such, most libraries depend on the expertise of third party for providing the unified search solutions. Adoption of standards and protocols are essential for a library for providing such services. If support for the search protocol is built into the library's tools, then the library can use it with a small code or convert the library's existing tools to support the new protocols. Librarians need to understand these standards and protocols to provide better interoperability among various services of the library and help in providing unified service to its users. Similarly the the developers of the library's tools will find it easier to provide services if common protocols were used.

2. Standards & Protocols related to Federated Search and Discovery Services

2.1. Z39.50 : Z39.50 is an ANSI/NISO standard for information search and retrieval. It is a protocol which specifies

data structure and interchange rules that allow a client machine to search from databases on a server machine and retrieve records that are identified as a result of such a search. Z39.50 protocol allows to communicate multiple computer systems on a network for the purpose of information retrieval. Z39.50 protocol enables users to search various databases located on different servers over networks or Internet through single search interface. It is used for search and retrieval of bibliographic records from multiple databases of the library. This protocol is not used by the Internet search engines. It is more complex, comprehensive and powerful than searching through http. Z39.50 has been extended to allow system feedback and inter-system dialogue. Z39.50 works in a client-server environment and acts as a common language that enables all the networked systems to understand the command. The name Z39 came from the ANSI Committee on Libraries, Publishing and Information Services which was named Z39. NISO standards are numbered sequentially and Z39 is the 50th standard developed by the NISO. The current version of Z39.50 was adopted in 1995 superseding earlier versions adopted in 1992 and 1988. Z39.50 provides rich vocabulary of search terms for search queries from the multiple databases with single search box. Z39.50 protocol facilitates the following features:

v   Sort : It allows users to sort and refine the searched results as per the requirement of the user.

v   Delete : It allows to delete unwanted results from the searched records, either entirely or for specified records.

v   Browse : It provides search and browsing facility through index lists of items such as subject terms, titles, author names, etc. It also provides flexibility to browse single database at a time or multiple databases.

v   Access Control : Z39.50 protocol allows administrator to access control through user ID and password.

This protocol is commonly used by many library catalogues and also some publishers provide the federates search services using this protocol.

2.2. SRU/SRW : SRU (SRU: Search/Retrieve via URL) is a web service protocol supported over both SOAP and REST for client-server based search. It was originally developed as a replacement for the NISO Z39.50 protocol. The first version 1.1 was released in 2004 and the latest version of SRU is SRU 2.0. The SRU protocol defines a request message (sent from an SRU client to an SRU server) and a response message (sent from the server to the client). This transmission of an SRU request followed by an SRU response is called a SearchRetrieve operation. Other than the SearchRetrieve Operation the SRU protocol defines two other operations namely, Scan and Explain. The Scan, operation supports index browsing, to help a user formulate a query. The Explain specification describes a server's Explain file, which provides information for a client to access, query and process results from that server.

The Abstract Protocol Definition of SearchRetrieve operation describes the data model, processing model and the result set model. The protocol defines the request parameters and the response parameters. The request parameters defined in the earlier version of SRU are query, startRecord, maximumRecords, recordXMLEscaping, recordSchema, resultSetTTL, Stylesheet and Extension parameters. The SRU2.0 added additional parameters, i.e.: queryType, sortKeys, Facet Parameters, RenderedBy, httpAccept, responseType and recordPacking.

The response parameters defined in SRU are numberOfRecords, resultSetId, record, nextRecordPosition, diagnostics, echoedSearch and RetrieveRequest. The default query language for SRU is the Common Query Language (CQL). The 'recordSchema' parameter in request defines the format of the record returned in the 'record' parameter of SRU.

SRW is a variation of the SRU. In case of SRU the request and response are conveyed via a URL over HTTP where as in SRW the request and response are made via an XML over the HTTP (webservice) as per the SOAP protocol.

The SRU/SRW services are commonly used by journal publishers for search services for their content. It can also be applied to subject gateways, library portals for exposing the content to federated search services.

2.3. NISO Metasearch XML Gateway (MXG) : MXG is proposed as an alternate to Z39.50 protocol and is based on the SRU protocol. The NISO MXG is a low-barrier-to-entry method for content providers to expose their content to metasearch application.

The MXG protocol is a simple message / response based on the NISO-registered Search and Retrieve URL (SRU) protocol. The metasearch provider send a query to MXG URLs via HTTP (Hypertext Transport Protocol) and in return gets a MXG compliant XML formatted response.
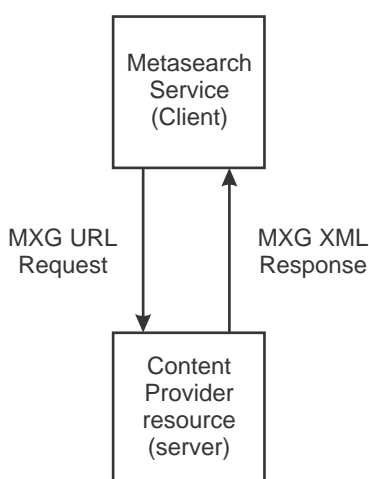

Figure : MXG Protocol Model (source : NISO Metasearch Initiative http://www.niso.org/publications/rp/RP-2006-02.pdf)

The MXG works in 3 steps as given in above model:

1. The Metasearch Service (MS) transmits an MXG URL request that incorporates the user's search query to a Content Provider's (CP) resource, usually a database.

2. The CP server receives and interprets the MXG URL and conducts the specified search query on its database.

3. The search query results are packaged into an MXG XML response, which is transmitted back to the MS.

The MXG implementation is recommended in 3 levels and the implementation at 3rd level is compliant with the SRU protocol. In level 1, the Meta-search service has to convert the user input into content providers search language. In level 2, the content provider gives an XML-formatted record that includes information about the resource, such as its host name and port and the database name. Optionaly, it may include human-readable information about the database(s) on the server such as its name, a database description, information about the indexes available for searching the database, and the schema that can be used to display returned records. The level 2 is extended by level 3 with additional feature to support a standard query grammar: CQL (Common Query Language).

The MXG recommends to have atleast on XML schema, although multiple schema for different Metasearch

Providers or different communities of users. It recommends the Dublin Core as one possible choice and suggests for adoption of additional schema for specific type of content like highly bibliographic content may find the MODS (Metadata Object Description Standard) schema useful or an e-learning community may use the LOM (Learning Object Metadata) schema.

The MXG protocol takes input of version and query as the request parameter with optional parameter of number of maximum records and starting record number. The response provides the version and number of records found along with record data which contains the schema and the record data for each record.

2.4. OpenSearch : It is a way for websites and search engines to publish search results in a standard and accessible format suitable for syndication and aggregation. It was created by A9.com, an Amazon.com company and made available under the Creative Commons License for open participation. The OpenSearch is built on XML and helps search engines and search clients to communicate by introducing a common set of formats to perform search requests and syndicate search results. So the search results are easy to process and display by a federated search service.

The OpenSearch consist of components mentioned below:

1. The Description elements which provides the details to identify and describe a search engine. It also describes the language, example query, encoding for the input and output content.
2. The OpenSearch URL template format describes how to query a search engine through the URL. It lists the parameters that can be used for making query.
3. The OpenSearch parameters for making a search query. It allows to query for the search term including count for number of results to display, starting page, startindex, language of output, input encoding and output encoding.
4. The query element used for defining the specific search request.
5. The response element that can be used by search engines to augment existing XML formats with search-related metadata. The most commonly used format are RSS 2.0, Atom 1.0 which provide list based results.

OpenSearch is commonly used by many libraries to provide search box facility in browsers for library catalogues or similar services of library. This facility is also provided by some publishers.

2.5. NISO Open Discovery Initiative : With the latest trend in information discovery, libraries are adapting to the new technology of index-based search in place of the traditional federated search for better search and discovery functionality.

Since the discovery services were widely accepted by libraries, NISO made an attempt to create "best practice" guidelines for discovery services named as 'Open Discovery Initiatives(ODI)' for creating transparency in discovery service. The ODI Group was established with a goal to:

v   Create ways for libraries to assess the level of content providers' participation in discovery services;

v   Help streamline the process by which content providers work with discovery service vendors;

v   Define models for fair or unbiased linking from discovery services to publishers' content; and

v   Determine what usage statistics should be collected.

The NISO ODI group made a survey of all the stakeholders of discovery service in 2012 which made the following findings:

v   Libraries are more interested in finding specific journals that are included in the discovery index than the publisher's packages or databases;

v   Libraries prefer the result lists to be resolved directly to a publication's original interface when the record comes from a primary, full-text content provider (where they have access);

v   Prefer link resolver as playing a pivotal role in linking and prefer libraries should have primary control of link resolution settings in discovery services, in order to manage access and prioritization of primary vs. secondary resources;

v   Prefer receiving usage metrics on total number of searches, list of search query terms, and URLs referring end users into the discovery service to better understand the user needs; and

v   Content providers to the discovery interfaces lack in providing open access indicators which will allow more free content available to users.

Based on the findings of the survey, NISO released the Recommended practice in June 2014. The recommendations are made for content providers as well as for discovery service providers are as follows:

a)   Content providers were advised to provide the content to the discovery service providers based on the KBART format with additional metadata for subject, abstract/description, content type and content format. Content providers are also advised to disclose their level of participation in discovery services to the library subscribers. It was also recommend that the transfer of data from content providers to discovery service providers should make use of existing standards where applicable. Most applicable standards are the OAI-PMH and the KBART. Also ResourceSync is being established by the NISO which strong potential to provide mechanism for the data transfer from the content providers to the service providers.

b)   It recommends the Discovery Service Providers to provide sufficient information about the content of their respositories to ensure an adequate evaluation of the content against the users needs. It recommends to provide the metadata of the content covered with publication title, date of coverage in case of journal or date of publication in case of monographs, standard identifiers (ISSN, ISBN), content type and format, and the depth of the coverage in the index. It recommend to use the KBART format in providing the metadata information. Discovery service providers are recommended to use fair linking policy in linking to the contents. The discovery service providers should ensure proper mechanism to handle primary metadata encoding formats, e.g., MARC, MODS, METS, VRA, DC, KBART, EAD, ONIX.

It also recommends the service providers to provide usage statistics to content providers as well as libraries. The library should be provided with the usage metrics for total number of searches per month, unique visitors per month, click-through per month, top 500 search queries, top referring URLs, etc. It is recommended to incorporate these metrics in future COUNTER code of practice.

It also provides two checklists (one for the content providers and one for the discovery service providers) which can be used by libraries for evaluation purpose.

3. Conclusion

Providing unified search and discovery of the resources to the users is one of the major goals of a library. It helps the library to maximize the usage of the library. To help achieve this goal, libraries need to make the most possible ways to create inter-operability among its services and resources. Various search standards and protocols like Z39.50, SRU/SRW, NISO MXG, OpenSearch help libraries in achieving the goal. The NISO ODI is a step towards standardization of the new index-based search landscape in the library community. The new ResourceSync Framework of NISO will also help libraries to create an automation in this new discovery environment. The libraries and librarians need to come forward to adopt these standards and protocols/guidelines and also ask the service providers to adopt these standards to create an easy to use inter-operable environment for the libraries.

4. Reference

v   Walker, J, (2006). New resource discovery mechanisms. The e-resources management handbook, 78-89

v   LeVan, R. (2013). OpenSearch and SRU: a continuum of searching. Information Technology and Libraries, 25(3), 151-153.

v   OpenSearch Specifications. Available at http://www.opensearch.org/Specifications/OpenSearch/1.1. Last accessed 27th Oct 2014

v   Morgan, Eric Lease. (2004). An Introduction to the Search/Retrieve URL Service (SRU).Ariadne 40. Available: http://www.ariadne.ac.uk/issue40/morgan/. Last accessed . 27th Oct 2014

v   NISO RP-2006-02, NISO Metasearch XML Gateway Implementers Guide. Available: http://www.niso.org/publications/rp/RP-2006-02.pdf. Last accessed 27th Oct 2014.

v   NISO, Z39.50: A Primer on the Protocol. Available: http://www.niso.org/publications/press/Z3950_primer.pdf. Last accessed 27th Oct 2014.

v   NISO, ResourceSync (Resource Synchronization). Available at: http://www.niso.org/workrooms/resourcesync/

v   NISO, NISO ODI Survey Report: Reflections and Perspectives on Discovery Services. Available at http://www.niso.org/apps/group_public/download.php/9977/NISO%20ODI%20Survey%20Report%20Final.pdf

v   NISO RP-19-2014, Open Discovery Initiative: Promoting Transparency in Discovery. Available at http://www.niso.org/workrooms/odi/publications/rp/rp-19-2014.