Journal of Geomatics Vol 10 No. 2 October 2016

Comparative study of distributed search protocols used in distributed geospatial systems

Sreeja Suresh and Sameer Saran
Geoinformatics Department, IIRS, ISRO Dehradun
Email: sureshsreeja.2015@gmail.com, sameer@iirs.gov.in

(Received: Feb 23, 2016; in final form: Jul 15, 2016)

Abstract: Terabytes of data are available in the present day world. Ever-growing amount of data is accumulating daily in both national and global databases. This data is generated in different formats and gets stored and maintained in various databases. There is a need to do geospatial data retrieval from distributed sources and accumulate data in an integrated environment for ease of access. This case study helps in understanding geospatial data retrieval and management systems which help in searching, harvesting and storing data in an integrated environment. Distributed search protocols namely, Z39.50, OAI-PMH are discussed in detail. The advantages and disadvantages of these protocols along with their use cases have been studied. A case study on GeoNetwork has been done to explain these protocols. The comparison of these protocols is made to facilitate appropriate usage when the distributed systems are being set up.

Keywords: Z39.50, OAI-PMH, GeoNetwork, Client-server, Distributed search, IBIN

1. Introduction

Present day world is generating data at exponential speed owing to wider usage of internet and access to mobile devices. Many applications are being made and used which is actively producing both geospatial and non-spatial data. Significant amount of scattered and heterogeneous geospatial and non-spatial data are produced, which lack integration. Hence, a user has to access each data collection individually to meet requirements and needs. This overhead results in wastage of time and makes it harder to get work done. Even though there is data, true information system to integrate Indian bioresources data is not yet available. Many approaches have evolved to provide solutions to various challenges in information access and data retrieval.

Data retrieval systems developed were primarily from scientific publications and library records. Later, it extended to different other forms of content useful to lawyers, doctors, etc. Major evolution of data systems have occurred in these contexts and current practices of data retrieval deals with giving access to both structured and unstructured geospatial and non-spatial data in many corporate and governmental domains (Manning et al., 2008). In earlier days, data retrieval systems typically looked for a strict match, that is, the check was done to confirm presence or absence of data in a file. Current data retrieval systems may or may not do a test for presence or absence of data. Here the emphasis is on the partial matching of items and selecting most optimal match among the retrieved data.

Data is exchanged from system to system using mechanism like client-server communication. The client-server architecture is used to understand and implement storage and retrieval systems like Indian Bioresources Information Network (IBIN). Servers are powerful systems or processes dedicated to accomplishing tasks like printing, file handling, taking care of network traffic, etc. Clients are simpler systems on which users can run applications. Clients rely on servers for resources, such as files, devices and even processing power. Clients send requests to the servers and can display the results the server sends back. Servers handle the requests received from clients and respond to them. Ideally, a server provides an abstract, standardized and transparent interface to clients so that clients remain unaware of the hardware and software implementation that is providing the service. Today clients are often placed on standalone workstations or computers of the user while servers are distributedly located in any part of the globe connected and are to the network. Servers are hence more powerful machines than clients.

Retrieval and harvesting of data began much before the growth of the web. Nevertheless, in recent years, a primary driver of such systems has been the World Wide Web (www), which is releasing content at the rate of millions per month. This ever-growing outburst of published data would be wasted if information gathering would be difficult. This problem may be due to difficulties in finding, interpreting and analyzing the data. Data becomes valuable only if every client or user can rapidly find information that is both significant and inclusive of their needs. Data and metadata, both play a crucial role in understanding, analyzing and extracting information.

Metadata refers to data about data. Geospatial metadata contains data about geospatial data. This metadata provides information regarding the geographic location of the data. Harvesting of Geospatial metadata refers to aggregating metadata into a central storage for ease of access. Metadata provides a description of the date and time of creation of data, authors, the extent of the data (if geospatial), etc. It contains the description of the data and context at the time of acquisition of data.

1.1 Search architectures

There are various distributed search mechanisms which are available in the current scenario. The two principal methods are centralized search and distributed search. Centralized or aggregated search refers to the existence of all distributed data in a central repository. This repository handles all storage, retrieval and management of data. All the primary processing like data storage, maintenance, computing and retrieval is done at the central level. Indexing and storage are also done centrally. Some of the disadvantages include exponential increase in data, greater storage requirements, maintenance of indices become much more challenging and machines are less fault tolerant and scaling up gets harder.

Federated or distributed search refers to the mechanism of performing a parallel search across all distributed databases simultaneously. This mechanism ensures that user need not have a central storage for storing all data beforehand. Search is done dynamically upon user request. Here data processing and storage are done on systems which are connected to an interconnected network, which provides a single system appearance to the user. This technique involves searching across distributed databases simultaneously. Here there is no central storage of data, but rather the systems are loosely coupled (decentralized) here. The distributed search mostly works by aggregating (harvesting) metadata which gives information about where information is stored and what data it holds. Aggregated storage is continuously updated at regular intervals for maintaining integrity and consistency.

Advantages of distributed search include first, enhanced processing levels and speed due to increased number of systems sharing processing. Secondly, scalability is better in this case. Thirdly, network bandwidth also shared across systems. Hence, downloading becomes faster. Lastly, fault tolerance is improved as a failure of one machine does not affect other machines.

Both centralized and distributed architecture can make use of client-server architecture for search. One of the most popular centralized systems of geospatial data is United States Department of Agriculture's (USDA) Geospatial Data Warehouse (GDW). GDW is made centralized to achieve continuous availability of data. Data is broken into small subparts or subsets called data marts and thus improves the efficiency of search. This architecture ensures access, availability and fitness-for-use of the spatial data (Fisher and Reed, 2005). IBIN is a distributed system, which is a national level portal for the integration of biodiversity data.

1.2 Open Geospatial Consortium (OGC) initiatives

OGC has been successfully developing publically available interface standards. OGC provides various standards which can be used to provide interoperable solutions by geo-enabled web, wireless and location based services. OGC's Interoperability Program (IP) has been utilized to promote OGC standards. It has various advantages like reducing technology risks, mobilizing new technologies, providing cost effective methods for stakeholders and expanding the market and improving choice of stakeholders. OGS's Catalog Service for the Web (OGC CSW), OpenSearch, etc. are standards which are used to search data in an interoperable manner.

Z39.50 is an international (ISO 23950) standard describing a protocol for information retrieval, used by networked computers to retrieve, store and manage data, based on client-server architecture. Z39.50 facilitates a user in a single isolated system to search and retrieve data from other systems located spatially anywhere in the world, without being aware of the search syntax. Such systems should be compatible with the Z39.50 protocol. Z39.50 standard-based clients are used to send requests and Z39.50 based servers receive requests and send responses back to the client. This mechanism makes use of federated search where all databases are searched simultaneously for the data, and the responses are returned to the client.

Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) (Muller et al., 2015) describes a mechanism for harvesting records containing metadata from repositories. Data from distributed databases are centrally aggregated into a repository and then services can be provided by making use of this repository. The implementation is based on Extensible Markup Language (XML) over Hyper Text Transfer Protocol (HTTP) requests and responses. It requires the use of simple Dublin Core (DC) format as essential means to provide interoperability but also supports the use of other metadata schemas like Machine Readable Catalogue (MARC) etc.

Open Geospatial Consortium Catalog Service for the Web (OGC CSW) provides the ability to publish and search metadata, services, and related data objects. CSW is one of the accepted standards for geospatial data retrieval and search. CSW is a mandatory service for supporting discovery and binding to registered data resources. Metadata gathered in catalogs can be further used by both humans and machines for representing resource characteristics which can be queried and given for evaluation. CSW is implemented as a web service and can be consumed to maintain a relationship between client and server to access metadata.

2. Background

Z39.50 grew from Linked System Project (LSP), which was a major initiative in the 1980s to standardize searching in the main bibliographic databases. National Information Standards Organization (NISO) and American Nation Standards Institute (ANSI) were working in parallel with LSP initiative to standardize information retrieval protocol. In 1988, Z39.50 got approval as version 1 ANSI/NISO standard for information retrieval. Library of Congress is designated as official Maintenance Agency and Registration Authority. Z39.50 Implementers Group (ZIG) got the

primary role in ongoing development work of the Z39.50. ZIG developed version 2 in 1992 and version 3 in 1995 of the protocol. The current version, Z39.50-2003 is a compatible superset of Z39.50-1995. International standard ISO 23950 is based on versions 2 and 3 and presumes that releases 1 and 2 are identical (National Information Standards Organization (U.S.) and American National Standards Institute, 2003).
The Santa Fe Convention was the first initiative of OAI-PMH. The focus of the Santa Fe Convention was to optimize the discovery of e-print repositories. The first version 1.0 of the OAI-PMH was released in January 2001 for the unqualified Darwin Core (DC) element set as the base for metadata interoperability. This version used metadata harvesting model for metadata interoperability, built on top of Hyper Text Transfer Protocol (HTTP) GET and HTTP POST requests and Extensible Markup Language (XML) response. The current version 2.0 is a major revision of version 1.0, released in June 2002 based on W3C XML standards. Version 2.0 is a stable protocol, and subsequent versions would be backward compatible unlike version 1.0 (Lagoze et al., 2002).

GeoNetwork (Ticheler and Hielkema, 2007) is a project of Spatial Data Catalogue System for Food and Agriculture Organization (FAO), United Nations World Food Programme (WFP) and United Nations Environment Programme (UNEP). It is a part of OSGEO. It provides instant search on both local and distributed geospatial data.

3. Parameters used in distributed search mechanisms

Distributed search mechanisms are characterised by various parameters as mentioned below:

3.1 Content acquisition
This parameter deals with the procedures used for acquiring of the data, its submission into the system, managing acquisition and management workflow like sending an email notification to users about the status of the submission.

3.2 Content management
This category involves functionalities related to incoming and outgoing data into the systems and understanding regarding various versions and supported document types.

3.3 Metadata
Metadata is the most important component of a system for content indexing, storage, availability, access and durability. The system should have the capability to add and delete user-specific metadata fields and real-time updating and indexing of accepted content.

3.4 Search support
Searches can comprise metadata search, full-text search and hierarchical subject based browsing. Search is one of the most sought facilities in information retrieval systems and hence becomes the critical parameter.

3.5 Interoperability
Interoperability refers to the interaction of one system with other homogeneous or heterogeneous system in local or distributed environment.

3.6 User interface
This category includes the ability to customize user interface to suit needs of different system implementations.

3.7 Standard compliance
Standards is the most important factor to be considered for sharing of digital content and a permanent preservation.

4. Working of Z39.50

Z39.50 (Clifford, 2015) is a pre-web technology and this has resulted in the need to update it regularly to adapt to modern technology. Z39.50 International: Next Generation (ZING) is one of the most dedicated working groups and pursues various strategies. Searches are performed on attributes based on use, relation, position, structure, truncation and completeness. Even complex queries are permitted in the syntax of the Z39.50 protocol. The Z39.50 syntax is abstracted from the structure of the underlying database. Based on this structure, each database can define its mechanism of searching using the indices. This feature strengthens the use of the formulation of Z39.50 queries without much knowledge about the corresponding databases. This characteristic results in obtaining different result sets for databases. This variation might be coming because one server might hold one index and other might be holding a different index, and another may have no suitable index and hence return an error. Z39.50 operates in synchronous mode. This differentiates it from a harvester. The client queries are run directly on the remote server and the results are sent immediately.

Z39.50 framework consists of the following components:

1. Web browser like Internet Explorer, Google Chrome, etc.
2. Web server, which accepts client request in HTTP format.
3. Common Gateway Interface (CGI) which helps to pass the request to Z-Client.
4. Z-client is client module which is implemented using Z39.50 standard.
5. Z-server is server module which accepts the request in Z39.50 recognizable format. Distributed databases contain actual data.

4.1 Process flow for Z39.50 protocol
The following steps are involved in the process flow as represented in Figure 1.

1. User demands query to web server through web browser.

2. Web server passes the query to CGI in HTML format from where it is passed to Z39.50 Client.
3. The client then converts the query into Z39.50 standard format and passes to Z39.50 server.
4. The Z39.50 server sends a request in Z39.50 recognized format to various databases.
5. Queries are run on the individual databases and results are sent in the form of responses to the server.
6. The server returns the response to the Z-client.
7. Client converts responses into user requested format and sends to the Web Server through CGI.
8. Web server displays the results to the user.

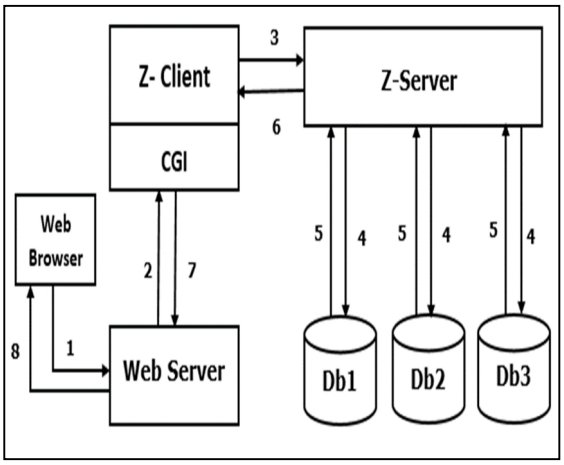

Figure 1: Z39.50 framework

The Z-client converts the result back into a format requested by the user and returns to the web Server. This protocol has been implemented on a small scale with the help of Java Application Program Interface (API).

System configuration:

Operating System	Windows 8.1
RAM	8GB
OpenSource technology	Java API, XML
Library used	yaz4j
Architecture	Client-Server
Database	Voyager

Using yaz4j library which is widely accepted Java API for establishing client-server communication between Z-client and Z-server.

Steps involved in the process are:

1. Establish connection with distributed server database which follows Z39.50 standard.
2. Read records from the Voyager database.
3. Display the records to the client.

The Z39.50 protocol analyses user requests and converts the request into Z39.50 standard format and searches in local Z39.50 servers. Results obtained are then converted to user required formats like XML or JavaScript Object Notation (JSON) and sent. XML and JSON are two formats used to provide interoperable data.

5. Facilities or queries supported by Z39.50

Z39.50 emerged as a best possible solution for searching, browsing and other services. These features are discussed in this section. The user provides these queries when retrieving data. The results are given to the user based on the below-mentioned queries or facilities. These queries can be run for both geospatial and non-spatial data.

5.1 Sort
Provides means to sort resultsets based on given criterion.

5.2 Search
This feature enables search across one or more databases with the help of structured query written in standard format.

5.3 Initialization
This property is used at the first establishment of communication between client and server.

5.4 Browse
This characteristic is used to browse index terms and other database fields using browse window.

5.5 Access control
Provides proper authentication services for each session and additionally for each operation also.

5.6 Resource control
Provides means for the cancellation of a search or other requests in the middle of an operation while still maintaining the session.

5.7 Maintaining multiple search results
This feature helps in an extended form of searching even on historic result sets.

5.8 Extended services
Provide database maintenance operations, mostly Create, Read, Update, Delete (CRUD) operations. In addition to these facilities, it also provided a better user interface and web-based browsing.

6. Working of OAI-PMH

OAI-PMH is primarily used for metadata harvesting in repositories containing records. It uses HTTP GET and POST requests and XML responses, due to which web-based harvesting is made possible. The most used metadata standard here is Dublin Core (DC). Metadata from various repositories can be collected and stored. This characteristic is the core concept of harvesting or aggregation.

There are two logical roles used in the context of OAI-PMH: Data providers and service providers. Data

providers are the depositors of data. They create, publish and maintain data in repositories which are available for harvesting resources. Service providers are users for metadata provided by data providers for actual harvesting. Using this metadata, they provide various services like searching, reviewing, etc. One agency can act as both data and service provider simultaneously. Data providers only need machine interfaces, while Service Providers need user interfaces for clients. If an organization wants to act as both kinds of provider, it will need to set up both the interfaces.

OAI-PMH requests include Identify, ListMetadataFormats, ListSets, ListRecords, ListIdentifiers, GetRecord and responses include General Information, Metadata formats, Set structure, Record Identifier, Metadata. All requests have to follow these formats for querying purposes and responses can be given in user requested formats.

OAI-PMH framework consists of the following components:

1. End user application, which initiates a request in XML over HTTP format.
2. Data providers are modules which are used to expose data, store data like repositories.
3. Service providers are modules which harvest metadata and store them in the aggregated data repository.
4. Aggregated metadata repository is the storehouse of metadata from where all the data is stored and maintained for use.
5. OAI-Harvester is the application which aggregates data from distributed databases.
6. Distributed databases are storehouses of heterogeneous data available anywhere globally.

6.1 Process flow for OAI-PMH protocol

OAI-PMH framework is represented in Figure 2. The steps involved in the process flow are described below:

1. End user sends HTTP request to the service provider.
2. The service provider searches the aggregated metadata repository.
3. The harvester aggregates data from distributed data providers.

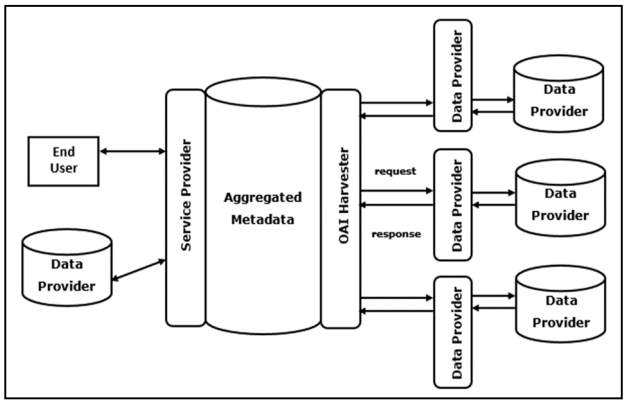

Figure 2: OAI – PMH framework

4. The distributed data is searched and returned as HTTP responses and send back to the client.
5. The databases may include image files, XML files, Audio files, etc.

6.2 Process flow for client-server communication in OAI-PMH

Steps involved in the process are:

1. Establish connection with any number of distributed databases which follow OAI-PMH standard.
2. Read MARC records from these databases.
3. Search various records from distributed databases.
4. Display the MARC records to the client.
5. Style the records as per user request using XSLT.

System configuration:

Operating System	Windows 8.1
RAM	8GB
OpenSource technology	Java API, XML
Library used	marc4j
Architecture	Client-Server
Database	Voyager

This protocol has been implemented on a small scale with the help of Java API. Using marc4j library which is widely accepted Java API for establishing client-server communication between OAI-PMH client and OAI-PMH server.

7. Systems utilizing Z39.50

Z39.50 protocol has been used in many distributed systems for the purpose of achieving searching across distributed databases. (Cao et al., 1998) Describes one such application called CLIB which describes the use of Z39.50 for there is a major focus on the integration of multilingual bibliographic data collections. This characteristic has been strived for to reduce the complexities of using distributed and heterogeneous databases. The system developed uses Z-client to allow multi-language queries to the user. This search query is processed by the Z search servers which performs the distributed searching. ISite is another such work done to search and retrieve multiple geographic footprints (Nebert and Fullton, 2015). In ISite, geographic coordinates are made searchable within ANSI Z39.50 standard. Thus, data can be searched and retrieved based on their geography. ZDSR profile is another usage of Z39.50 protocol for distributed search and ranked retrieval based on Stanford Protocol for Internet Search and Retrieval (STARTS) project Annon. 2015). Next section describes how IBIN uses GeoNetwork in order to perform distributed search using these protocols.

8. Sequence flow of search in IBIN

This section discusses the process flow in the actual search happening in IBIN for species data. The client sends an ‘init’ request query and this enables a session

from the server side and a confirmation is sent back as response. Upon receiving this response, an 'explain' request is sent to the server, which returns server configuration regarding the available repositories for search. A 'search' request is then initiated by the client for example single word like 'santalum' or a phrase can also be given like 'caccia tora' and server translates this query into database understandable ones and performs search. The results are stored on the server in the form of 'Result Sets'. The client then uses 'sort' and 'browse' facility to customize the results. When client sends 'present' request, the results are sent to the client from the stored 'Result Sets'. Finally, a 'delete' query is used to remove stored data from the server.

In case of OAI-PMH, client sends an 'Identify' request to the server. Server sends the details regarding the available repositories. Client then sends 'ListIdentifiers' and 'ListRecords' query to retrieve information regarding the collection of records and identifiers for the records. To retrieve set structure of the repositories, 'ListSets' request is given. The available metadata formats can be accessed through 'ListMetadataFormats'. To retrieve the records, 'GetRecord' query is used. When user sends 'GetRecords', various species which match the user query are sent as response in the user requested metadata format.

8.1 Technical specifications for IBIN system

System configuration for IBIN is given below:

Operating System	Windows 2003/2008 Enterprise Edition
Web Server	Apache tomcat
Map Server	MapServer/GeoServer
Data Cataloging	Geonetwork
Programming Environment	JAVA, PHP, Javascript
Search protocols	Z39.50
Harvesting Protocol	OAI-PMH

Hardware and software configurations play an important role in the implementation of these protocols. These protocols are highly dependent on the speed of the network available. In IBIN, Bioresource Information Centres (BRIC) nodes play the role of distributed data providers. These BRICs include University of Agriculture (UAS), Bangalore, Institute of Himalayan Bioresource Technology (IHBT), Palampur, and so on. The availability of data in these repository nodes when the actual search is performed also is a major concern.

8.2 GeoNetwork: Case study

Formerly Z39.50 and OAI-PMH were used mostly in library catalogs and information retrieval systems, but both are used in GeoNetwork (Carboni, 2006) for harvesting and searching of geospatial metadata catalogs. IBIN utilizes GeoNetwork for integration of bioresource species data in a single portal. Z39.50 protocol is used for distributed search across geospatial metadata catalogues in the IBIN using Geonetwork. The Z-server has to be enabled and a port has to be provided to search for data using Z-protocol. The user can query and retrieve geospatial metadata from multiple distributed Z39.50 servers using the harvester in GeoNetwork. Z39.50 servers are used to search databases and users can select which servers have to be searched. The Z39.50 query can be provided to sort, search, etc. as discussed in section 5 above. Additional features like providing icons for metadata can be done in GeoNetwork, which will be displayed in the search results. Furthermore, scheduling options like what time to run the query: run-at, scheduling when to repeat the query: will run again every (provide some days as input), etc. is also provided. The results are styled using Extensible Stylesheet Language (XSLT) and sent back to the user. XSLTs are used to style the XML responses in either user requested or default format (in case the user does not specify).

Thus, if the OAI-PMH harvests metadata in Dublin Core (DC) format and then the response can be sent back to the user requested style by applying the corresponding XSLT. OAI-PMH uses the concept of harvesting metadata in GeoNetwork also. The various OAI servers which have to perform harvesting mechanisms have to be selected by the user. Metadata can be provided multiple sets (categories in GeoNetwork), which will enable faster search in the system. OAI-PMH also supports the scheduling options like run at, will run again every (provide the number of days), etc. Output formats like Dublin Core, MARC, etc. can be provided so that response can be returned in that particular format.

Z39.50 uses Media Access Control (MAC) address, which is a unique identifier assigned to each network component. When new data is supplemented to the database, synchronization is performed by checking the unique id, which is a combination formed by appending MAC address with last updated date-time stamp. Harvesters store records in a date range, and this can be used in DateSearch operations. Search is performed by seeing the Temporal Extent provided in metadata records or using the Modification date stored in the database, which is the last modified date of the record.

In short, harvesting of metadata is done with the help of a unique identifier created with a combination of MAC address, date/time stamp of last modification and a random number. This unique identifier is called Universally Unique Identifier (UUID). UUID is unique both locally and globally. Thus, last update date/time stamp of all records is stored and used to synchronize local record with the changes made to any remote record. For example, if a record is updated remotely, and when a synchronization occurs, there would be a change in the last updated date/time stamp locally. Thus, same changes are made in the updated records locally also.

The conditions to be considered during synchronization are:

1. UUID must be unique; else harvesting may result in the corrupted metadata.
2. Both remote and local metadata schema must match so as to harvest successfully.

9. Discussions

Various comparisons of Z39.50 and OAI-PMH are made in Table 1. Z39.50 is a remote harvesting and searching protocol that is regularly used to permit search and harvest of metadata. Although the protocol is often utilized for library catalogs, significant geospatial metadata catalogs can also be searched using Z3950 (e.g. the metadata datasets of the Australian Government organizations that participate in the Australian Spatial Data Directory (ASDD)). This harvester permits the user to state a specific Z39.50 query and retrieve metadata records from one or more Z39.50 servers.

Discovery systems use several technologies to harvest information to provide a display of information to the users. Z39.50 is frequently used to capture data from library catalogs, OAI-PMH services are used for Open Archive repositories, and OpenURL is used for article and journal databases that require patron authentication for access. For harvesting content from websites, some technologies are employed including OpenURL, SearchRetrieve via URL (SRU), OpenSearch, and, when necessary, screen-scraping.

There are various advantages of using Z39.50 protocol for distributed searching. It helps in searching across many databases, concurrently, using the same query. This property being one of the widely accepted standards helps in easier integration with many organizations which deal in library environments, bibliographic reference software, etc. Furthermore, inter-library catalog searches can also be performed using Z-Queries. The Contextual Query Language (CQL) or previously known as Common Query Language, is also based on the semantics of Z39.50. Z-client typically searches across structured data and hence preserves the semantics of user's query. Though there are various advantages as well as disadvantages of using Z39.50 protocol. Poor implementation of Z39.50 systems is a major consideration while using the protocol. There are specific details of implementation provided by the protocol, resulting in only a part of functionalities of the protocol implemented in various commercial and non-commercial users. Extensive work is required in implementing all the features of Z39.50 and implementation costs are also high. Network bandwidth requirement for using Z protocol is very high as it supports real-time searching.

There are advantages and disadvantages in using OAI-PMH for metadata harvesting. Benefits include the existence of support for this protocol in many distributed repositories. The only disadvantage being the requirement of conversion of metadata to standard protocols like Dublin Core.

Table 1: Comparison of distributed search protocols

Compare	Z39.50	OAI-PMH
Content Type	Heterogeneous	Heterogeneous
World View	Bibliographic, Geospatial	Bibliographic, Geospatial
Data Stores	Distributed	Distributed
Contrast	Z39.50	OAI-PMH
Searching Technique	Distributed or Federated	Centralized or Aggregated
Search performed by	Data provider	Service provider
Data Supported	Only Z39.50 compatible data	Z39.50 and non-Z39.50 data
Scalability	Less Scalable	More Scalable(Lightweight)
Usage	Real-time searching of data	Does harvesting of metadata
Implementation	Hard	Easy
Cost	High	Low

Using customized harvesting mechanism as an alternative to OAI-PMH can result in problems like a dependency with underlying database structure and so on. One of the major disadvantages of harvesting is its resource intensive nature. This nature of OAI-PMH makes the whole set up expensive, although setting up OAI-PMH environment itself is cost effective. As a result, OAI-PMH is using HTTP over Z39.50 to perform a search, so as to perform real-time search along with harvesting. GeoNetwork uses the features of both Z39.50 and OAI-PMH simultaneously to achieve data retrieval and storage. Implementations similar to GeoNetwork has been very powerful to achieve real integration of data.

10. Conclusion

As quoted by Sebastian Hammer and John Favaro, "The essential power of Z39.50 is that it allows diverse information resources to look and act the same to the individual user."

If the future usage used structured metadata across World Wide Web, the actual power of Z39.50 search clients would become evidently known. Z39.50 is thus, a language used to communicate between to computers located in different geographical areas. It can search multiple distributed databases simultaneously and retrieve records. It maintains a high level of abstraction to the users, by pretending like it is searching local machines for data while it searches distributed data. Z39.50 is used in the library catalog and bibliographic references, and even for geospatial data search in applications like GeoNetwork, EndNote, etc. It has to be specially noted that Z39.50 is not a client interface or a search engine. It is just a protocol used for communication and translation of user's query into Z standard and back to translate results to user requested format. Clients have to be well chosen while implementing the protocol in order to achieve best results. Results are provided in XML or JSON format to achieve interoperability. DC format is mostly used to represent metadata in OAI-PMH. MARC format is also used to describe metadata.

OAI-PHM has been used for metadata harvesting in various applications like Wikimedia for sharing Wikipedia feeds. OGC CSW has been widely used to publish and bind data on registered information. The case study has been implemented to do a comparative study of three different client-server based Information Retrieval systems. Thus, harvesting of metadata can be done using OAI-PMH framework wherein the searched data is very easily aggregated into an integrated environment. From the aggregated metadata, using Z39.50 format a search can be easily made to retrieve records. OGC CSW can be used for cataloging this data and also for providing publishing and binding of data both geospatial and non-spatial. Z39.50 and OAI-PHM have been utilized in the case study to retrieve records in XML and Marc format. Furthermore, OAI-PMH is not searching protocol but rather metadata harvesting protocol. The current version of OAI-PMH is not backward compatible, hence creates more overheads during interoperable operations. Future releases of OAI-PMH are to be made backward compatible.

Indian Bioresource Information Network (IBIN) is making use of the GeoNetwork to integrate bioresource data in a single portal. Facilities provided by IBIN include species search, spatial and non-spatial web services like Web Map Service (WMS), Web Feature Service (WFS), etc. to provide open data retrieval in biodiversity domain. The power of Z30.50 and harvesting mechanisms of OAI-PMH has been well realized in this system. Further works are being done in IBIN to increase the potential of species search in the portal. As Z39.50 protocol is not suited to support a scalability requirement of more than 100 nodes at a time, there is need to implement other distributed mechanism like a creation of a light-weight version of Z39.50 protocol or mechanisms which support NoSQL like mechanisms which will support scalability to a larger extent. NoSQL is schema-less and hence offers wider support to heterogeneous data presently being created and maintained in various heterogeneous storages. Future works can be done to integrate NoSQL like mechanisms with IBIN to achieve scalability and performance.

References

Annonymous (2015). ZDSR profile: Z39.50 profile for simple distributed search and ranked retrieval. 1-1.

Cao, Ling, Mun-Kew Leong, Ying Lu and Hwee-Boon Low. (1998). Searching Heterogeneous Multilingual Bibliographic Sources. Computer Networks and ISDN Systems 30 (1-7): 612–15. doi:10.1016/s0169-7552(98)00063-4.

Fisher, Brandon, and Carl Reed (2005). White Paper: Server Architecture Models for the National Spatial Data Infrastructures (NSDI). Open Geospatial Consortium. Version: 1.1.

https://www.loc.gov/z3950/gateway.html. Gateway to Library Catalogs Z39.50. (Last accessed 21 August 2015).

http://geonetwork-opensource.org/docs.html. GeoNetwork Architecture and Technologies. (Last accessed 01 January 2015).

http://www.niso.org/standards/resources/Z3950_Resources.html. The NISO web site. (Last accessed on 3 December 2015).

http://lcweb.loc.gov/z3950/agency. The Z39.50 Maintenance Agency web site at the Library of Congress. (Last accessed on 17 August 2015)

http://geonetworkopensource.org. GeoNetwork open-source community website. (Last accessed 13 January 2015).

Lagoze, C., H. Van de Sompel, M. Nelson and S. Warner (2002). Open archives initiative-protocol for metadata harvesting-v. 2.0.

Lagoze, C., H. Van de Sompel, M. Nelson and S. Warner (2002). Implementing Guidelines for the Open Archives Initiative for Metadata Harvesting: Guideline for Harvesting Implementers. (Last accessed 04 July 2015).

Lagoze, C., H. Van de Sompel, M. Nelson and S. Warner (2002). Implementing Guidelines for the Open Archives Initiative for Metadata Harvesting.

Lynch, Clifford, A. (1997). The Z39.50 Information retrieval standard - Part I: A strategic view of its past, present and future. D-Lib Magazine, April 1997, ISSN 1082-9873.

Manning, C.D., P. Raghavan and H. Schütze (2008). Introduction to information retrieval. Cambridge University Press, New York.

Muller, U., A. Powell, P. Cliff, H. Van de Sompel, C. Lagoze, M. Nelson and S. Warner (2015). OAI for Beginners – the open archive forum online tutorial. Last accessed 01-07-2015.

National Information Standards Organization (U.S.), American National Standards Institute, (2003). Information retrieval (Z39.50): application service definition and protocol specification: an American national standard. NISO Press, Bethesda, Md.

Nebert, Douglas D and James Fullton (2015). Use of the ISite Z39.50 software to search and retrieve spatially - referenced data. 1–10.

Ticheler, J. and J.U. Hielkema (2007). Geonetwork open source internationally standardized distributed spatial information management. OSGeo Journal, 2(1).