Introduction to a regional document sharing platform in China

Guo Yi-Ping, Wang Liang

Library of Huazhong University of Science and Technology, Wuhan, P.R. China 430074

Corresponds author: E-mail: gyp@hust.edu.cn, Tel: (+86)027-87557701

[Abstract] This paper introduces the Hubei science and technology document sharing platform from the aspect of technology, mainly including the requirement analysis, technical solution, services and some innovation. The future development of such platforms is also briefly analyzed. It could give the reference for the construction of similar platform.

[Keywords] Science and technology document，resources sharing, digital library

1 Introduction

In April 2005，China 'National science & technology infrastructure platform' project launched. The corresponding project in Hubei Province also began in 2006. Our group is charge of science & technology document sharing platform in this province. The main aim of our project is to share the science and technology documents among different universities and public libraries. Providing convenient information service for public uses based on these resources is another important mission of this project. This paper gives the detailed introduction for the technical solution of this platform.

2 Preliminary investigation and analysis

At the beginning of this project, we deploy an investigation for the similar platforms in other provinces. The document resources distribution of our province is also inspected.

There are some same features in these available platforms. Their resources cover metadata of some commercial databases, the MARC data of bibliography and the full recorders of some self-construct information resources. The search service is the main services of platform. Some full text could be directly obtained by users, if their organizations have purchased the corresponding databases. If not, they could obtain the related recorders by document transfer system. Such platforms also provide the information consult and other conventional library services.

But we also find some common problems in these platforms. First, they normally buy some very conventional resources. But some participants have purchased these resources. Second, how to deal with the relation between platform and local library is still a difficult problem. Obviously, if the platform can't bring the benefits for the local libraries, they will lack the impetus to participate the work of sharing platform.

In the resources investigation of our province, we find the universities are possessed of 80% science and technology documents of whole province. But there is only few resources shared among these universities. These universities also don't provide information service for public users.

According to these conditions, we decide the basic principle of our projects. First, we should decide the clear system architecture to ensure the harmony between sharing system and local libraries. Second, the metadata and standard search service should be the main sources of sharing

system, but not the full text resources. The users may obtain the full text by document transfer system if there is no copyright problem.

3 Technical solutions

3.1 system architecture

According to the analysis above, we decide two layers architecture for this information sharing system, which contains a central node for resources sharing and some local nodes in participant libraries. This architecture is very common in current resources sharing system like Fedora[1]. We also have some theory research in this topic [2][3]. These works mainly apply the metadata/OAI to integrate the data resources and Web services/UDDI to integrate the information services.

The central node and local nodes are all resources integration system, but have different scale. The central node stores the metadata of some commercial database and the metadata from participant libraries. The standard search service description data is also index by this node. The local node is the resources integration system of local library. The data between local nodes and central node could exchange by standard protocols. For users, they can search the resources in local library first. If there is no the records they want or lack the related records. The local nodes could automatically recommend the search link to the sharing system in central node. The whole system architecture is shown in fig.1.

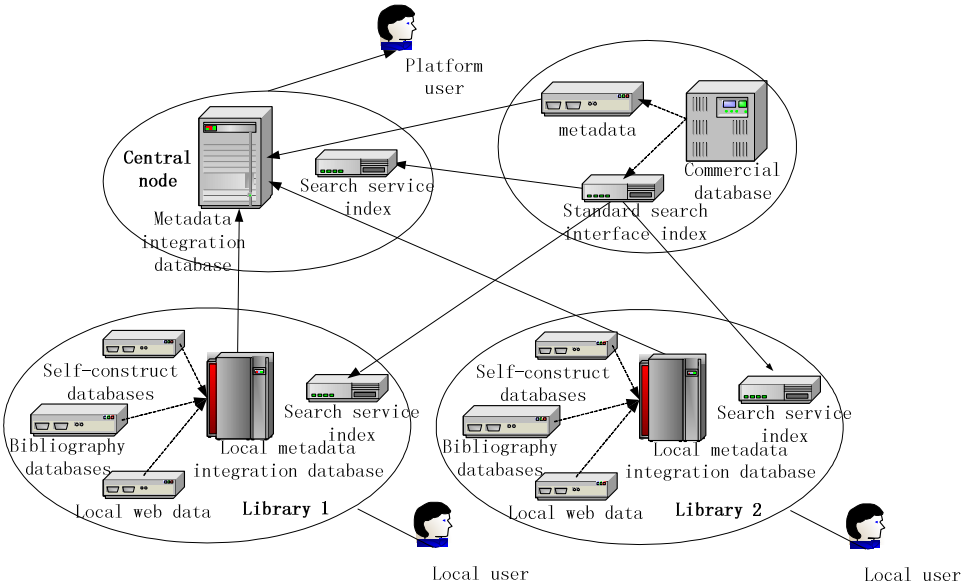

Fig.1 whole system architecture

3.2 resources of platform

The total quantity of resources decides the quality of service of sharing system. Now our platform mainly includes these resources:

(1) Some metadata of commercial databases. For the domestic databases, we could easily purchase their data or make some agreements to freely obtain their metadata. But there is still no an effective way to arrange with foreign databases providers to get the metadata. Most of them just provide very simple metadata contains article title and author. So we could only extract few latest metadata by RSS resource or web spider technology. Now we have obtained most Chinese journals, meetings and thesis metadata, about 1million digital books item, 3000

multimedia records.

(2) The papery bibliography data of participant libraries. We mainly harvest the MARC data of these libraries. The current platform contains about 1 million bibliographies from eight universities in Hubei Province.

(3) The metadata of some self-construct database of participant libraries. This platform includes the metadata of mechanism, geology and optoelectronics subject databases.

(4) If some resources can't provide the metadata, we wish they could provide the standard search interface like ZING. Because there are only few databases providing such interface in current progress, we use some web search interface to build the union search system, just like Meta search system.

To integrate these metadata, we convert them into the standard DC format and then store them into the common relational database like MS-SQL. Here we only select 10 field of DC to store the metadata from different databases, which could ensure most search request. These fields are shown in Table.1.

Table.1 Metadata format of platform

	Name (DC)	Meaning in platform
1	Title	The title of document
2	Creator	Author of document, normally contains several authors
3	Description	The abstract of document
4	Publisher	The publishing company or organization of document
5	Format	The storing format, PDF,JPG, Word, etc
6	Subject	The keywords of document
7	Identifier	The URL or DOI of document
8	Date	The publishing data of document
9	Language	The language of document
10	Source	The document provider

As a data integration platform, the amount of data reaches the 10 million items level and still increases. We find the search response is very slow if the parallel users increase, which is mainly caused by the bottle-neck of database system. So we use the Lucene, a full text index component to re-index the SQL data and provide the search service by this new index. The search response problem is well solved. Now most data of platform is still manually updated through FTP. In the next step, we wish apply OAI and Z39.50 to automatically update the metadata of platform.

3.3 Service of platform

3.3.1 Information retrieval services

Information retrieval service is the kernel service of central platform. The users can search all the resources from a union search system. In system background, the data may come from the metadata databases, the HTTP search interface or Z39.50 providers. But users only face one web search interface. Now the system only applies some simple algorithms to combine and rank the search results from different resources. Some latest research achievement like 'universal search' will be added to this system in the future [4]. The main function of this search system is described

as follows:

- It could provide the simple search and advanced search options, just like most current academic databases. The user could also choose the special subject to search in part of resources, which is automatically selected by system.
- Most search results contain the full text link or the link to document transfer system.
- The system could analyze the use condition.

The users have two search approaches. First, they could search in the local nodes. If there is no related result or lacking the records, the system will give the suggestion link to central node. And of course, the user could also directly search in central node. But part of search results will link to their local databases. We recommend the first search method, which accords with the common search habit of users.

3.3.2 Use management

As a resources sharing system, it includes the users from participant library, the public users and subsequent participant library, etc. The system is also managed by different organizations. So the user role management is an indispensable component of system. According to the user organization and other related information, we set the different roles, which are managed by a unified identity authentication system 'Shibboleth' [5]. The role of platform is shown in Table.2:

Table.2 the role in platform

User role	Title	Description
Common user	Public users	They can browse and search the database and web information, but could only get the full text of part self-construct resources. They need submit the document transfer form to get the full text of other resources.
	Common federation users	Mainly the users of participant library. They can browse all the information in the central nodes. They could also visit part of full text by IP authentication.
	Advanced federation users	The federation could register in central to update to this kind of role. Besides the right of common federation users, they could also directly submit the document transfer request in the web and obtains some personalized services.
Management user	Participant library management	Contractor person of participant library. They are in charge of data management, publishing information and coordinating the document transfer work.
	Data management	The data manager of central nodes, they maintain all the data of sharing system.
	System management	The top role in system, they maintain all the users of system, the web site of sharing system, etc.

3.3.3 Document transfer system

For users who can't obtain the full text, they could get the text by this system, the digital text or papery form document. When users find a record but no full text visiting right, the system will

give a link to this document transfer system. After they submit the transfer order, the system will distribute the order to a participant library or other document transfer system. The related fee is also counted and paid through this system.

3.3.4 Participant library system

The kennel system of local system is also a union search system which includes all the resources of local library. The user can search all the local data in this system or get a recommendation to search in central node.

Another important component is designed for the management of participant library. They need extract the data from the local database, and then submit it to central nodes or maintain a unload interface like OAI. They could also submit a request to download the data from central node to develop their own application.

4 Summary and further work

There are two main functions of this document sharing platform. First, it expands the search scales of users. Second, it also provides the efficient method to obtain the full text for users. For some public resources, the user can directly download. This platform also provides the interface to document transfer system to help user get the full document.

The amount and scale of information resources is the key to the success of such document sharing platform. The self-construct resource is only a small part of platform. So how to cooperate with database providers to get more metadata or standard search interface is the main mission of our further plan. Collaborating with other resources federation is another way to deal with this problem.

There are also two trends for the services of such platform. First, it should provide 'simpler' search service. We design a web interface only containing a search box like Google. It's very useful for some preliminary search applications. But how to combine the different kinds of results is still a difficult academic problem. Second, some users also need more special services. For example, we build an 'Optoelectronics' database based on the data of this platform. Some Web2.0 technologies like 'Dspace' are also introduced in this system. This system is also very welcome for related subject users.

Reference

[1] Thornton S,Ross W,Sandra P. The fedora Project:An Opensource Digital Object Repository System.D-Lib Magazine,2003(4)

[2] Guo Yi-ping, Wang Liang, Domain Resources Integration System and its Protocol Series, Journal of academic libraries, 2005, (6) 40-46

[3] Wang Liang , Guo Yi-ping, Web search engine based on DNS, Journal of Network and Computer Applications, 2007,30 466-478

[4] Anish Das Sarma, Xin Dong,Alon Halevy. Bootstrapping pay-as-you-go data integration systems. Proceedings of the 2008 ACM SIGMOD international conference on Management of data

[5] About Shibboleth. http://shibboleth.internet2.edu/about.html