Development of an Institutional Repositories Network in Chinese Academy of Sciences

Zhongming Zhu, Jianxia Ma
State Key Laboratory of Frozen Soil Engineering
CAREERI, CAS
Lanzhou, China
e-mail:{zzm,majx}@lzb.ac.cn

Linong Lu, Wei Liu and Denglu Wu
Department of Information Technology
Lanzhou Branch of NSL, CAS
Lanzhou, China
e-mail:{luln, liuw, wudl}@llas.ac.cn

***Abstract*—This paper introduces the current practices of building the network of institutional repositories in Chinese Academy of Sciences (CAS), which called CAS IR Grid. Firstly, it presents and discusses its architecture. Then it discusses issues related to its implementation and promotion activities, which covers mainly the development of IR software package based on open source software DSpace, establishment of long-term working mechanism of promotion IR service, achievements of spreading activities, and the development of CAS IR Grid service portal. Finally, it puts forward some considerations on future development of CAS IR Grid.**

Keywords-Institutional repository; institutional repositories network; architecture; implementation strategies; Chinese Academy of Sciences; knowledge management

I. INTRODUCTION

Institutional repository (IR) was first described by SPARC [1] as a "digital collection that captures and preserves the intellectual output of a single university or multiple institutions". Now, it is regarded much more as a set of services that serve the purposes of management and dissemination of scholarly research outputs created by the institution. The value proposition made by IR come mainly from three perspectives [2]. Firstly, from the view of institution, it will be a knowledge asset management platform to help the institution set up a centralized space to collect, organize, showcase, preserve and disseminate its knowledge artifacts in a systematic manner, thus to avoid the most likely risk of losing them, which now are scattered among various groups and administered by researchers themselves. Secondly, for individual researchers and research community, it will maximize the accessibility, availability, and impact; enable the discoverability, increased functionality, long-term storage and curation, and other potential benefits of scholarly research outputs. Thirdly, IR provides research funders a new venue to see and share the outcomes of their publicly funded work. Currently, IR has become a popular knowledge management practice in universities and research institutions. Latest statistics from OpenDOAR [3] shows a steady growth and development of IRs worldwide. As research distribution strategy [4] has become a hot topic in recent two years, IR is destined to get more attention than ever.

Chinese Academy of Sciences(CAS) has more than one hundred institutes in 28 cities across China and more than 25,000 researchers and 40,000 graduate students (50 percent are doctoral) [5], it plays a very important role in Chinese science research and is a major producer of STM information in China. So as a pragmatic strategy for advancing knowledge management practice and advocating open access activities in CAS and China, a network of institutional repositories named CAS IR Grid was brought forth by National Science Library of CAS (NSL) in 2007. It envisaged helping each institute establish its own local repository as a node of the Grid, and NSL will construct a centralized metadata repository via harvesting and aggregating metadata of academic resources stored in distributed institutes' IRs, and will keep an integrated interface for the aggregated resources and provide other value-added services. At present, CAS IR Grid is progressing smoothly, there are nearly half of institutes having started IR service, and a pilot portal for CAS IR Grid has been launched as well.

II. ARCHITECTURE OF CAS IR GRID

As shown in Fig. 2, CAS IR Grid uses a three-layered architecture comprised of the content layer, the aggregation layer and the interface layer.

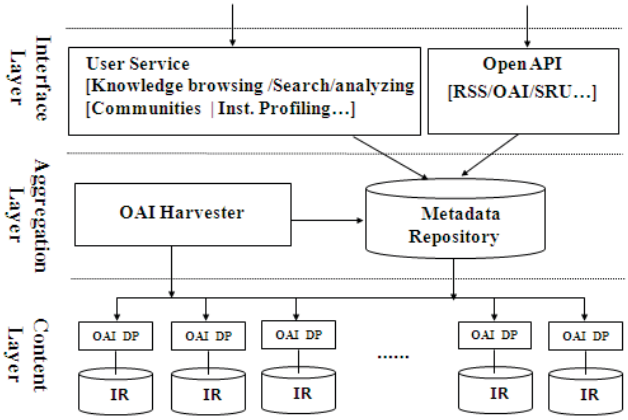

Figure 1. Architecture of CAS IR Grid

The content layer consists of IRs distributed in institutes. They are the nodes of the Grid, and each has a built-in OAI-PMH data provider interface, which exposes IR's metadata

978-1-4244-5265-1/10/$26.00 ©2010 IEEE

with default OAI DC metadata format or more applicable qualified DC metadata format to be used in central metadata repository in aggregation layer.

The aggregation layer has main components of the OAI-PMH harvester and the central metadata repository. The OAI-PMH harvester is mainly an implementation of OAI-PMH service provider, which harvests and aggregates metadata records from OAI-PMH data providers, in here are institutes IRs. Harvested records are aggregated and stored in the metadata repository, which serves as a base on which various tools and services can be developed subsequently.

The interface layer provides not only user-oriented services but also application-oriented services. User services will include basic and conventional services such as browse and search service. In addition, it will support advanced services like dynamic creation of knowledge communities, institute profiling, generation of CAS-level knowledge inventories, and cross-subjects or cross institutes data linking, integration and discovery service. Application oriented services are standard-based interoperability interfaces, which support machine-to-machine or application-to-application data exchange or service interactions. For example, RSS service interface will publish newly added records as RSS feeds to support RSS alert service; OAI-PMH interface is used to re-exposing harvested metadata records to enable cooperative data sharing among relevant repositories or applications in a wider scope; SRU interface will offer a standard and professional means of embedding IR search service in a application easily.

III. IMPLEMENTATION AND PROMOTION

A. Development of IR Software Platform

In general, we think using OSS solutions are cost saving and highly efficient in construction of IR platform. After some comparisons and evaluations, we eventually choose DSpace [6] as a prototype system to develop CAS IR platform via a way of extending and optimizing. Primary reasons for choosing it are as follows:

- It has the largest community of users and developers worldwide.
- It can theoretically manage and preserve all content types.
- It has critical core functions we need.
- It has a layered architecture and clearly defined API, by which can it be easily customized and extended.
- Using Java platform makes it very convenient to integrate plentiful OSS packages or tools, when needed, in Java world.

In a concrete case, DSpace 1.4.2 was selected as our prototype system. To make it well adapt to meet our needs and requirements, we have done many optimizing and extending development work. Following are some major customizations and extensions:

- It has been highly localized to support Chinese application, not only providing Chinese interfaces, but also handling names of people, order of sorting, search queries, etc to conform to Chinese customs.
- The default metadata schema has been extended to meet the needs and requirements for describing special content types, for instance, patent.
- Additional quick submission workflow is extended to be default alternative to replace its built-in complicated submission workflow. Now it contains just two-steps (pages) to finish one submission in a few minutes, therefore, making submission an easy-doing job.
- Browse and search facilities are enhanced. In particular, while an item is reached, an integrated service bar will offer a set of subsequent operations related to it, including bookmarking, recommending, viewing usage statistics, social bookmarking in Connotea, CiteUlike, Digg, etc., providing reference citation in normal style, starting associative search with extracted information in external academic search systems such as Google Scholar, Scirus, CSDL Cross Search, etc.
- The import function is improved to support bulk import particular formatted data, for example, those legacy data accumulated in excel format.
- An extended function is implemented to support knowledge asset statistics. It can generate knowledge inventories at institute-level, research unit-level, or individual-level, and with combined conditions of researchers' position, status, and publication date of item. The statistics results can be exported with an excel-formatted file to be saved or printed.
- The default implementation of OAI-PMH data provider is also improved to allow flexibly defining crosswalks between internal metadata format to standard metadata formats, for example, the compulsory OAI DC metadata format, or qualified DC metadata format. This will well support the construction of harvest and aggregation service.
- The default installation of DSpace is based on a compiling method containing many steps to be executed manually. This makes the deployment of DSpace-based application a laborious work, and usually needs some technical skills. For deploying an IR application in many institutes easily, we improve its installation process and create one-click installation package, therefore, any one with little computer skill can install the IR package in a few minutes.
- There are other major or minor optimizations, enhancements, or extensions having been done concerning user interfaces, browse and search, knowledge organization, user management, access control, etc.

Fig. 2 is home page of IR of Institute of Mechanics, CAS [7], which is a real application based on extended version of DSpace, which is internally called CAS IR package.

Fig. 3 gives an exemplar page of showing integrated service bar related to an item.

Figure 2. Home page of IMECH IR

Figure 3. An exemplar page of showing integrated service bar

For more about extended work we have down, consult above IR or any other delopments in CAS.

B. Promoting IR Service in Institutes

In order to effectively promote IR service, a long term work mechanism is established. Fig. 4 presents an overall organization structure for promoting IR service in CAS.

The consultation board consists of leaders from NSL. Its main tasks are to guide the direction of promotion service, make important decisions involved in overall promotion arrangements, and provide regular consultation service on major issues concerning the promotion activities.

The coordination team is mainly responsible for overall organization, coordination and collaboration tasks involved in the process of promoting IR service. They include proposing a set of policies templates to be referenced by institutes; capturing, analyzing and allocating efforts to meet service needs from institutes; and establishing work mechanism for promotion of IR service.

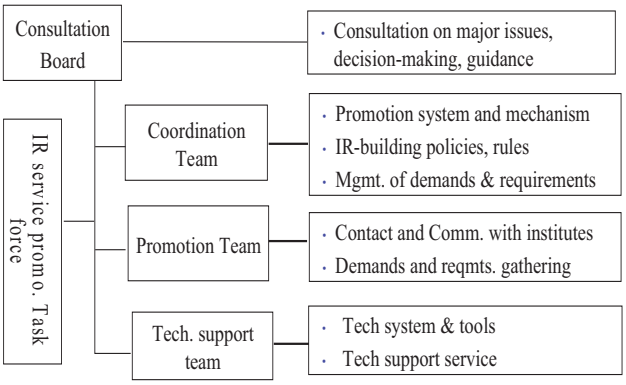

Figure 4. Organizational structure of IR promotion

This team is composed of members from NSL and institutes. Those from NSL are usually head of coordination groups of NSL. They coordinate efforts to meet the requirements that need their support. In addition, members from institutes are usually directors of libraries of institutes as representatives of institutes to propose requirements.

The promotion team is composed of subject librarians from department of subject information service of NSL, and is responsible for concrete IR promotion tasks. Each member of this team will take charge of several institutes' IR promotions. They are required to provide following services:

- Pre-consultation service: it may cover any issues concerning IR building.
- Help the institutes plan an IR service.
- Help institutes formulate an adapted policy framework based on what policies templates proposed by the coordination team.
- Help institutes work out feasible work procedures for running IR service.
- Help institutes install CAS IR package.
- Provide training in using and maintaining IR system.
- Help the institutes prepare legacy content and transfer them into IR system.

The technology support team consists of IT engineers from IT department of NSL. Their responsibilities concerning following two aspects:

- Develop software package and tools in support of building IR. We need to customize, improve and extend DSpace to be a more suitable application for us. In addition, we also need to add new features constantly to keep up with the development of IR itself.
- Provide technical support service involved in IR promotion and application activities. Regular technical support services may covers a range of services, such as installation of IR package or tools, training, guiding the use of application, customization, distribution of software patches or updates as required to resolve specific

issues/problems, providing technical advices or resolving any other specific inquires or technical problems. These services usually are required to be provided in a timely manner via telephone, email, instant messages, support site, or remote desktop technology.

Organizational structure clearly defines division of responsibilities and sets up communication and collaboration mechanisms among working teams. However, in order to conduct IR promotion service methodically, we have established following official working procedures to guide the smooth running of the promotion.

- First, with the help of responsible subject librarian from NSL, the institute library on behalf of the institute submits an application request to NSL, namely consultation board of the project. This application should include clear-cut definition of strategic goals of building IR service, definite team organization, commitments to establish effective policies and supply supportive conditions and facilities, such as funds, basic hardware and software, etc.
- When consultation board receives an application, it performs a routine assessment of its feasibility, checks whether required conditions are met, and finally draws a conclusion if or not the application is passed through.
- If an application is approved, NSL assigns an agreement with the institute. As an official engagement between two sides, it prescribes bilateral responsibilities and obligations and is taken as a starter of building IR service in the institute.
- Finally, we enter procedure of bilateral practical cooperation. Those teams of NSL and the team of the institute will work together to push the implementation of IR service in the institute. During the course of this procedure, subject librarians will get much involved in consulting and helping institute work out a detailed implementation plan and finish constituting an adaptable policy framework; then the institute perform preparations such as hardware and software environment and other required pre-conditions to get ready for installation of IR package; technical support team help institute complete tasks such as remote or on-spot or self installation of IR package, post-configuration, training, trial running, official lunching, and other technical ones.

Until now, over 50 institutes have assigned building agreements with NSL, and nearly 40 institutes have launched IR service.

Among them, some have gradually established sustainable support mechanism for the development of IR service. For instance, the Institute of Semi-conductors has incorporated the developing IR service into its informationization development, and established mechanism of combining content submission with assessment and reward of graduate students, researchers, and research units, to stipulate and encourage individual participation in IR building. The Institute of Software, the Dalian Institute of Chemical Physics and others have also set up similar mechanisms.

Concerning the content collection building, peer-reviewed journal articles, conference papers, theses and dissertations are the most concerned scholarly content. Many IRs also have included presentations, books, book chapters or sections, patents, project achievements, and other materials. Moreover, research data are of most future concerned research outputs to be included in IRs.

In respect of content recruitment, it is still in an initial stage overall. Except some IRs have reached a several thousand-level, many of them have totally accumulated less than one-thousand items.

C. Development of CAS IR Grid Portal

We continue to follow our OSS strategy in development of harvest and aggregation components of CAS IR Grid. After a preliminary investigation and evaluation, two candidates were left, they are D-NET [8] and OAI ORE patch for DSpace [9].

Actually, we much prefer D-NET. It has powerful and attractive service oriented architecture, and is not just a harvester system; in fact, it is a open-sourced software package with much more complex functionalities to support running of a repositories network. Moreover, it is support platform of DRIVER Repository network. From the point view of ongoing collaboration between DRIVER and CAS IR Grid, it is a best selection for using it as basic platform for CAS IR Grid. Thus, as part of collaboration between two sides, DRIVER installed a test D-NET application in our local servers.

However, current D-NET 1.1 as an interim version is still under development, so it is not so stable to use it in a production environment. Moreover, from the point of view of using DNET in China, we need additional Chinese interfaces. But current version of D-NET has no built-in l10n/i18n support, so if we want to use it, we have to hack source codes to gain a localized version beforehand. Apparently, doing like this is time-consuming and not sustainable as well.

Therefore, we temporarily use alternative solution of building CAS IR Grid service portal, i.e. using OAI ORE patch for DSpace. This patch can function as a basic OAI-PMH harvester and an OAI ORE harvester, and we just use it as the former. It is relatively easy to be integrated into DSpace. Based on the integration, we have set up a test CAS IR Grid service portal (see figure 5). However, the patch is just an implementation of OAI harvester; it is worth looking forward to new release of D-NET, because we are informed that the new release will have updates we need.

IV. FUTURE CONSIDERATIONS

Developing CAS IR Grid service is a long-term task, we will continue to try our best to push it forward. Our future concerns may exist in following aspects:

- Develop successive enhanced version of IR package. It is expected to support wide-expansion of knowledge content types, provide further search and discovery features, enhance capturing context and

relationship between knowledge objects, and advance its composability and interoperability to integrate or to be integrated with other services easily.

- Spread IRs in most institutes across CAS. Currently, less than half of institutes have launched IR service; we need to help rest institutes join our IRs community in next years.
- Upgrade CAS IR Grid service continually. We will keep it harvesting IRs across CAS as many as possible, and develop advanced search and discovery functionalities.
- Make concrete progress in metadata sharing with DRIVER. In addition to current collaborations, which are particularly in the area of technology, we both sides need to promote bilateral collaboration to metadata exchange level.

Figure 5. CAS IR Grid portal

ACKNOWLEDGMENT

The authors would like to thank for contributions of Xiaolin Zhang, Tan Sun, and Dongrong Zhang from NSL. We also thank DRIVER, in particular, Dale Peters, Marek Imialek for their help in providing consultation and support services involved in evaluating D-NET. This work was supported by the Knowledge Innovation Program of the Chinese Academy of Sciences and in part by the West Light Foundation of The Chinese Academy of Sciences.

REFERENCES

[1] R. Crow, "The case for institutional repositories: A SPARC position paper," ARL: A Bimonthly Report, no. 223, August 2002. [Online]. Available: http://www.arl.org/resources/pubs/br/223br.shtml. [Accessed: Dec. 18, 2009].

[2] K. Weenink, L. Waaijers and K. van Godtsenhoven, A DRIVER's Guide to European Repositories, Ed. Amsterdam: Amsterdam University Press, 2007. [E-book] Available: http://dare.uva.nl/document/93898. [Accessed: Dec. 18, 2009].

[3] University of Nottingham, The Directory of Open Access Repositories – OpenDOAR, "OpenDOAR charts – worldwide," December 2009. [Online]. Available: http://www.opendoar.org/find.php?format=charts. [Accessed: Dec. 18, 2009].

[4] K. Hahn, C. Lowry, C. Lynch and D. Shulenberger, "The university's role in the dissemination of research and scholarship — A call to action." Research Library Issues, no. 262, pp. 1-5, February 2009. [Online]. Available: http://www.arl.org/resources/pubs/rli/archive/rli262.shtml. [Accessed: Dec. 18, 2009].

[5] X. L., Zhang, X. W. Liu and L. Li, "NSL OA implementation: promoting the development of sciences," presented at Sino-German Symposium on Development of Library and Information Services, Beijing, China, 2009. [Online]. Available: http://conference.las.ac.cn/Sino-German/2009/DOC/Session5/14.pdf. [Accessed: Dec. 18, 2009].

[6] DuraSpace, "DSpace Open Source Software," [Online]. Available: http://www.dspace.org. . [Accessed: Dec. 18, 2009].

[7] Institute of Mechanics, CAS, "Institutional repository of Institute of Mechanics, CAS," [Online]. Available: http://dspace.imech.ac.cn. [Accessed: Dec. 18, 2009].

[8] DRIVER, "D-NET v1.0," [Online]. Available: http://www.driver-repository.eu/D-NET_release. [Accessed: Dec. 18, 2009].

[9] DSpace Foundation JIRA, "OAI-PMH & OAI-ORE harvesting support," [Online]. Available: http://jira.dspace.org/jira/browse/DS-289. [Accessed: Dec. 18, 2009].