In-depth Customization and Extension of DSpace: Case Model of Developing an Institutional Repository System for Institutes in CAS

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

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

***Abstract*—This paper describes a practice model of developing institutional repository system based on open source toolkit DSpace for institutes in Chinese Academy of Sciences (CAS). It introduces the requirements gathering by using a method of patchwork prototyping with DSpace, then it discusses the in-depth customizations and extensions of DSpace, which covers expansion of metadata and supported types of content, bulk import enhancement, additional optimized submission process, introduction of associative integrated services bar, extension of statistical functionality, enhancement of interoperability, etc. Applications and practices show that this model or approach is effective and successful in getting a customized repository system in a quick and easy way.**

Keywords-institutional repository; case study; DSpace; customization and extension; Chinese Academy of Sciences

I. INTRODUCTION

Planning and building an institutional repository (IR) service has become a popular knowledge management practice in universities and research institutions in recent years. OpenDOAR[1], which is an authoritative and dynamic directory of academic open access repositories, shows that there has been a steady growth and development of IRs worldwide, and the overall number of quality-assured IRs registered so far has reached 1,310. IR can benefit both institutions and researchers, and even general public in many folds: to preserve the intellectual output of the institution; to remove barriers to access; and, to begin to address the scholarly communications crisis[2]. While there are emerging serious concerns and call to action by universities and research institutions to make their scholarly resources and content to be as usable and broadly accessible as possible, it is critical for IR to serve valuable role in establishing the institution's research distribution strategy[3-4]. In summary, it is identified as essential infrastructure for scholarship in the digital age[5].

The Chinese Academy of Sciences (CAS) is a leading academic institution and comprehensive research and development center in natural science, technological science and high-tech innovation in China. It has more than 100 institutes and institutions located in 23 provinces and municipalities throughout China. Therefore, when we began to introduce institutional service into CAS in the second half of 2007, a two-staged strategy was considered and followed. At the first stage, main tasks are building IR service at institute-level, which is to set up IRs and spread IR service in each or most institutes. At the second stage, work will be done on implementing an IR service at CAS-level, which is conceived to be a network of distributed and independent IRs and is realized via a way of harvesting metadata from those IRs.

With regard to the first stage, one of most important tasks should be prototyping and development of an IR system for institutes. Considering the proliferation of open source repository toolkits in the community, we adopted a method of in-depth customization and extension of DSpace, which is a famous open source digital asset system, to have a localized and suitable IR system to meet our intended needs and purposes.

The rest of the paper is organized as follows. Section 2 discusses how requirements are identified and gathered by a method of rapid prototyping using DSpace. This is followed by a detailed customization and extension of DSpace in section 3. We conclude in section 4.

II. REQUIREMENTS GATHERING

A. Method

In order to identify and determine requirements quickly and reliably, we partly adopted the idea of patchwork prototyping method proposed in [6]. It is an improved rapid prototyping approach to requirements gathering. The method takkkk advantages over conventional rapid prototyping in that encouraging developers to rapidly build and evaluate a high-fidelity prototype in real-world situations by combining multiple open source applications. Such prototype can span the breadth of a horizontal type and the depth of a vertical prototype within a single system, and avoid the drawbacks of typical methods of either developing the prototypes that do not have high enough fidelity, or prototypes too expensive to develop in a limited time framework.

B. Requirements gathering

Following steps given by patchwork prototyping method[6], we first made an investigation of typical IR systems and applications around the world and basically got better knowledge and understanding of what the target system might look like.

Then detailed evaluations and comparisons of open source repository software were carried out to select appropriate toolkits as a prototype system for requirements elicitation and validation, and as an evolutionary prototype

that would grow into a full-blown production-scale system. Finally DSpace[7] was chosen for the following reasons:

- It has the largest community of users and developers worldwide.
- It comprises most desired functionality.
- 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.

Steps were taken to roughly localized and customized DSpace to have Chinese interfaces and deploy it in Institute of Mechanics, CAS, which was selected to as our direct user to test the prototype, solicit feedback and set example application for others to follow.

After a several turns of feedback and reflection, we had achieved a level of formalizing and forming a functional requirements framework, which comprises main points that DSpace is overall providing most of functionality to meet institute's needs and requirements, but there are principal gaps as summarized below:

- Existent supports for content types are needed to be deliberately optimized, enhanced or even re-factored, and extra specific types should be added in.
- The submission process contains too many steps or pages to take submitters much time in finishing one submission, which will definitely frustrate researcher's interest and original enthusiasm in self-submission.
- Functionality of bulk import is desired to translate legacy data in specific format such as Microsoft excel format, or from legacy or current applications, into IR.
- Creative services are demanded to change too conventional browse and search facilities.
- More advanced functionality for exercising usage and content statistics are expected to allow researchers and institute to know current usage and impact of content in a timely manner, create knowledge inventories at individual or institute level at any time, and even do some analytical work to support research assessment activities.
- IR system should be well embedded within institute's information environment and interoperate with other research information systems easily.

In terms of such an approach, the requirements are rapidly identified and clarified, and as a result, a detailed specification is defined in a speedy manner.

III. CUSTOMOIZATION AND EXTENSION

As stated in above section, we used DSpace not just as a prototype to observe and capture requirements; we envisaged that it would evolve into a real-life system. To make it a reality, the key tasks were to fill the gaps in needs and requirements that we just identified.

A. Expansion of Metadata and Supported Types of Content

DSpace provides a default descriptive metadata schema based on DC-Library Application Profile (DC-Lib) [8] and uses default "dc" namespace. It is suitable for simple description of resources such as journal articles, books, book chapters, etc., and has limitations in creation of description for ETD, conference proceedings, and patent resources. To meet desired requirements of providing a comparatively description of listed resources as journal paper, conference paper, thesis and dissertation, monograph, book and book sections, research report and other types of content, we expanded DSpace metadata schema to introduce more elements to meet of description of various types of content. The expansion process was guided by the principle of metadata application profile [9], but was forced to introduce some local elements which, out of consideration of implementation, are used in the name of "dc". Expanded elements are listed in Table 1.

TABLE I
EXPANDED METADATA ELEMENTS LIST

dc.contributor.author	dc.identifier.callnumber
dc.contributor. advisor	dc.identifier.isbn
dc.contributor. editor	dc.identifier.issn
dc.contributor. correspondent	dc.identifier.patentnumber
dc.contributor. patentee	dc.identifier.other
dc.contributor.other	dc.identifier.url
dc.coverage.spatial	dc.identifier.uri
dc.coverage.temporal	dc.language.iso
dc.date.accessioned	dc.publisher
dc.date.available	dc.publisher.place
dc.date.application	dc.relation.ispartof
dc.date.copyrighted	dc.relation.ispartofseries
dc.date.issued	dc.relation.isversionof
dc.date.modified	dc.relation.isbasedon
dc.degree.level	dc.relation.haspart
dc.degree.grantor	dc.relation.hasversion
dc.degree.place	dc.rights
dc.description	dc.rights.rank
dc.description.abstract	dc.source
dc.description.indexed	dc.source.conferencename
dc.description.patentpriority	dc.source.conferenceplace
dc.description.pctapplication	dc.source.issue
dc.description.pctpublication	dc.source.pages
dc.description.project	dc.source.volume
dc.description.provenance	dc.subject
dc.description.sponsorship	dc.subject.discipline
dc.description.tableofcontents	dc.subject.keyword
dc.format.extent	dc.subject.classification
dc.format.mimetype	dc.title
dc.identifier.certificatenumber	dc.title.alternative
dc.identifier.applicationnumber	

B. Additional Optimized Submission Process

Submission process in DSpace is composed of six steps (pages) and can be customized to include workflows for editing or checking metadata, approving submission before public availability. Although this will bring substantial advantages to metadata quality control, but it is time-consuming and liable to hinder researchers from

depositing content into IR. Therefore, we designed a new simplified quick submission process, which contains only two steps or pages: step one provides a page for description of object to be submitted; step two is a page for checking previous inputs and granting the deposit license. Moreover, the description page is backed by content type aware templating techniques that can present appropriate description template according to submitter's selection of content type, and in each kind of template, only minimal required elements are displayed, other optional elements are activated and unfolded at submitter's will. Further, this new process can also be compatible with existent workflow framework.

C. *Improvement of Bulk Import*

DSpace has a built-in import utility. It is a command line tool that can be used to ingest data conforming to DSpace Simple Archive Format[10], which is a structure of a directory full of items, with a subdirectory per item. Each item directory contains a file for the item's descriptive metadata, and the files that make up the item.

```
archive_directory/
   item_000/
         dublin_core.xml
         contents
         file_1.doc
         file_2.pdf
        ……
   item_001/
      dublin_core.xml
      contents
      file_1.png
      ...
```

The dublin_core.xml has the following format, where each metadata element has its own entry within a <dcvalue> tagset. There are currently three tag attributes available in the <dcvalue> tagset: <element> (the Dublin Core element), <qualifier> (the element's qualifier), <language> - (optional ISO language code for element).

```
<dublin_core>
      <dcvalue element="……"
               qualifier="……"
                language="……" >
          ……
      </dcvalue>
      ……
</dublin_core>
```

Such that the extension of bulk import of EXCEL or other formatted data were implemented through following steps:

- A configurable XML template was introduced for mapping legacy data fields to DC metadata elements.
- Then a helper class was designed for conversion of specific formatted data to a structure of DSpace Simple Archive Format.
- Last, existent ItemImport class was reused to import the formatted data to be installed as items in IR.

For the purpose of convenient use, this extended import functionality was integrated into browser-based administration toolkit with a combination of designing corresponding jsp pages and servlets. The Administrator, therefore, can do bulk import in a way of visually defining maps and import them easily.

D. *Introduction of Associative Integrated Services Bar*

This extension was meant to enhance browse and search facilities. We designed such an integrated services bar that will be accompanied with each item reached during the course of browse and search. The bar will offer a set of subsequent operations related to current item, 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. As seen in Fig. 1, all services related to an item are exhibitted in a single location and at user's fingertips to use freely.

Figure 1. An exemplar page of showing integrated services bar

E. *Extension of Statistical Functionality*

Content and usage reporting are most useful functionality for both researchers and repository managers.

In the aspect of content statistics, we implemented it mainly as a tool to support knowledge asset auditing. It can generate knowledge inventories at different levels of institute, research units, or individual researchers. Combining with other conditions of researchers' position, status, date of publication or submission, etc., reports, summaries or bibliographies that are more specific are possible to be created. Those items listed in these inventories in HMTL presentation are clickable to link to their metadata and content pages. Moreover, the automated generation of lists and reports can be exported to an Excel spreadsheet to be saved or printed for reuse. Fig. 2 is the page of dimensions of content statistics. From top to down are statistics type (by author, by submitter), submission date, institute, laboratory or department, positional title, member type (faculty, visiting

scholar, student, etc.), indexed-by (SCI, EI, etc.), authorship rank (first author or other), openness (public access, embargoed), author or submitter name, publication date. Thus, we can exercise complex statistics through combination of those dimensions.

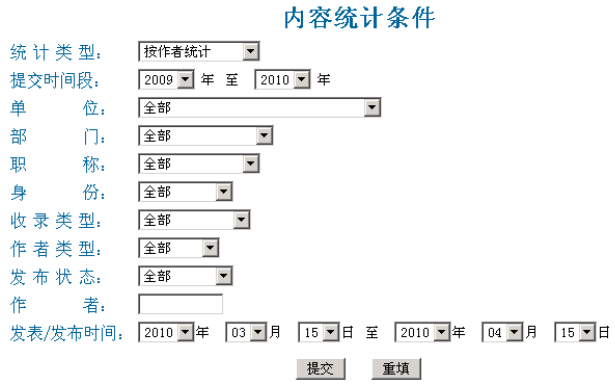

Figure 2. Dimensions of content statistics

Fig. 3 is one of representation pages of showing statistics result, which is a general summary of knowledge assets in a repository level listed by members and items count corresponding category types of research output. As stated, cells with name or numbers are clickable to show detailed bibliographies and the overall summary list can be exported to an Excel spreadsheet.


内容统计(按用户)(注：点击姓名列可查看相关统计直方图，点击表内数字可查看相关统计详情。)

统计时间段: 1950-05-28--2009-06-28

统计结果导出

姓 名	期刊论文	会议论文	学位论文	专著	文集	专著章节/文集论文	研究报告	演示报告	其他	总计
总 计	1254	72	3	0	3	15	35	32	84	1498
韩红	8	0	0	0	0	0	0	0	0	8
李苑	2	0	0	0	0	0	0	0	0	2
杨玉洁	2	0	0	0	0	0	0	0	0	2
高士雷	0	0	0	0	0	0	0	0	0	0
羅琳	6	5	0	0	0	0	0	0	0	11
李燕	7	0	0	0	0	0	0	0	8	15
李春旺	19	0	0	0	0	2	0	1	0	22
陈朝晖	17	6	0	0	0	1	0	0	0	24
....										
马建霞	11	0	0	0	0	0	0	0	0	11
乐小虬	1	0	0	0	0	0	0	0	0	1
徐雯	0	0	0	0	0	0	0	0	0	0
本页总计	197	12	1	0	0	3	0	1	13	227

[首页][上页][下页][末页] | 共有信息：256条 | 共有9页 | 当前为第1页 | 转到第 1 页 确定

Figure 3. A demonstration page of presenting a summary list

As for usage statistics, we adopted the Tasmania statistics package [11], which is a porting package of ePrintsStats for DSpace. The integration was easy and slight customizations were made to generate usage reports at item and repository level. Researchers can use it to find out who is using their work; repository managers can get a visual view of access and usage report of whole IR in any time, showing status quo of browsing, downloading, ranking, users distribution by country or domain, etc.. As a consequence, this value adding functionally is going to facilitate uptake of IR and encourage researchers' confidence in deposit to IR.

F. Enhancement of Interoperability

As a new kind of emerging application and service, IR should avoid going to be an information silo and bear mind to provide good support for interoperability. According our informal investigation, there are many kinds of information applications in institutes, which mainly covers research management system, office automation system, electronic theses and dissertations system, journal publication system, online meeting platform, library automation system, e-portfolio, data repositories, institute portal. So as to effectively and efficiently capturing and recording relevant knowledge content, IR should try best to establish reliable mechanisms of interoperating with these systems to enrich content recruitment channels. To this end, we had carefully optimized and extended open standard-base interfaces.

First of all, supporting OAI-PMH [12] has almost become a default implementation of all open repositories, which provides a standard for support metadata harvesting and sharing among open archives, and unqualified DC metadata schema is required to be used as default format to expose metadata. DSpace has an implementation of OAI-PMH data provider, but with hard coded exposure of internal metadata as unqualified DC metadata. This is usually problematic in keeping metadata semantics. Therefore, we improved its coding of OAI-PMH data provider to allow flexibly defining crosswalks between internal metadata format to standard metadata formats, including the compulsory unqualified DC, qualified DC, and METS metadata format. This will well meet the needs of building harvest and aggregation services.

We also added support for SRU based on OCLC SRW/SRU Open Source package [13]. SRU-Search/Retrieve via URL [14]- is a standard XML-focused search protocol for querying web-based databases and returning results, utilizing CQL (Contextual Query Language), a standard syntax for representing queries, and using the HTTP GET for message transfer. Thereby IR with an implementation of SRU will own a lightweight RESTful search service interface; any internal or external application could machine-to-machine-search of IR in a unified way.

Combining altogether the interfaces of OAI-PMH, SRU, inbuilt RSS, customized XML import, the enhanced IR system can be easily embedded or integrated into institute's research information environment as depicted in Fig., and will truly become indispensible component of a research infrastructure.

G. Other Customizations and Extensions

There are other major or minor optimizations, enhancements, or extensions done concerning user interfaces, browse and search, knowledge organization, user management, access control, etc.

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.

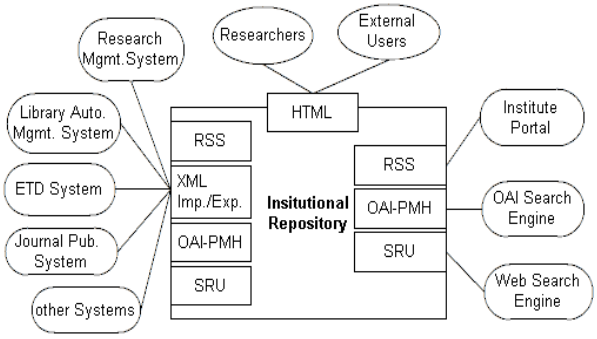

Figure 4. IR with related systems

Fig. 5 is home page of IR of Institute of Mechanics, CAS[15], which is a real application based on extended version of DSpace.

Figure 5. An exemplar page of showing integrated service bar

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

IV. CONCLUSION AND FUTURE WOKR

By adopting patchwork prototyping method to realize rapid requirements gathering and mannement, and using open source software as an evolutionary prototype system to obtain a production scale system, we had successfuly released an IR system. Currently, this extensively extended system has been deployed in more than 50 institutes across CAS.

However, developing an IR system for number of institutes in CAS involved many challenges. In particular, while institues' information and knowledge environment are becoming much more complicated than ever, IR system must be ready to face those constantly emerging problem and demand. Our future concerns may exist in aspects such as supporting multiple metadata schema coexisting for allowing various kinds of complex content objects to be well managed in IR, applying semantic web technolgies in IR system to build and provide semantic relation based discovery service.

ACKNOWLEDGMENT

The authors would like to thank for contributions of Tao Zhu, Yihong Xu from Institute of Mechanics. 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] University of Nottingham, OpenDOAR charts–worldwide, OpenDOAR. [Online]. Available: http://www.opendoar.org/find.php?format=charts[Accessed: Jun. 10, 2010].

[2] J.-G. Bankier and C. Smith, "Digital Repositories at a Crossroads: Achieving Sustainable Success through Campus-wide Engagement," in Proc. VALA 2010: Connections, Content, Conversations, Melbourne: VALA.

[3] K. Hahn, C. Lowry, C. Lynch and D. Shulenberger, "The University's Role in the Dissemination of Research and Scholarship—A Call for Action," EDUCAUSE Review, vol. 44, no. 2, Mar./Apr. 2009, pp. 6-7.

[4] J.-G. Bankier, C. Smith and K. Cowan, "Making the Case for an Institutional Repository to Your Provost," Research on Institutional Repositories (IRs), 2009. [Online]. Available: http://works.bepress.com/ir_research/28[Accessed: Jun. 10, 2010].

[5] C. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," ARL, no. 226, Feb. 2003, pp. 1-7.

[6] M. Jones, I. Floyd and M. Twidale, "Patchwork Prototyping with Open Source Software," in Handbook of Research on Open Source Software: Technological, Economic, and Social Perspectives, K. St. Amant and B. Still Eds., Hershey: IGI Global, 2007, pp. 126-140.

[7] DuraSpace, DSpace Open Source Software. [Online]. Available: http://www.dspace.org. [Accessed: Jun. 10, 2010].

[8] DC-Library Application Profile(DC-Lib), DCMI Standard, 2004.

[9] R. Heery and M. Patel, "Application profiles: mixing and matching metadata schemas," Ariadne, no. 25, Sep. 2000. [Online]. Available: http://www.ariadne.ac.uk/issue25/app-profiles/intro.html[Accessed: May 12, 2010].

[10] R. Tansley, M. Bass, M. Branschofsky et al., DSpace System Documentation, 2007.

[11] K. Suzuki, The Tasmania Statistics Software for EPrints to DSpace, Oct. 2008. [Online]. Available: http://www12.ocn.ne.jp/~zuki/Japanization/others/es-stats.html [Accessed: Apr. 9, 2010].

[12] The Open Archives Initiative Protocol for Metadata Harvesting, OAI Standard, 2002.

[13] OCLC, SRW/SRU Server, 2006. [Online]. Available: http://code.google.com/p/oclcsrw/[Accessed: Apr. 9, 2010].

[14] SRU-Search/Retrieve via URL, Library of Congress Standard, 2002.

[15] IMECH, Home Page of IMECH IR. [Online]. Available: http://dspace.imech.ac.cn[Accessed: Jun. 12, 2010].