CSpace – A More Practical and Customizable Repository Platform Serving Local Needs

Zhongming Zhu, Wangqiang Zhang, Wei Liu, Xiaona Yao, Linong Lu,
Lanzhou Branch of National Science Library, Chinese Academy of Sciences
Lanzhou, P.R. China
e-mail:{zhuzm, zhangwq, liuw, yaoxn, luln}@llas.ac.cn

1. Introduction

CAS IR Grid [1] is an overarching repository infrastructure that connects the institutional repositories (IR) of member institutes throughout Chinese Academy of Sciences (CAS). It aims to advance scholarship and promote research infrastructure development within but not limited to the scope of CAS. The primary goals are to first establish sustainable IRs in most institutes as a practical and strategic means for them to build partnerships with their researchers to ensure capture, preservation, and the widest possible sharing of all types of their research outputs and; then, these IRs should gradually grow to be essential and embedded infrastructures in research environment, upon which overlay services can build and layer.

In support of its construction and helping achieve above goals, the software toolkit CSpace is developed based on popular OSS DSpace package [2] and was officially open sourced in October last year. With successive customization and extension by modifying existing, or adding new components or modules, it offers functionalities and services that are more practical and well fitting for Chinese language settings and responds to constantly expanding local needs and concerns. Moreover, facing the evolving context of digital content in research environment, CSpace allows for creating or tailoring content type aware templates and associated rules in a non-programmable way, to adapt them to different and changing needs of content management. A range of other useful customization options are also available to enable that it to be easily deployed within certain local settings as well.

2. Extended Functionalities and Services Portfolio

There was general agreement that successful advocacy for repositories is dependent on being able to offer services that people valued [3-6]. Accordingly, taking DSpace's services as a basis and taking into account of deeper knowledge of specific local needs and concerns, a comprehensive range of extended services are optimized and introduced to support the implementation of institutional knowledge asset management. The fundamental designs of these services are to make them real incentives to attract different stakeholders.

Collection Building Services: A variety of modules and tools are implemented to support capturing content in various formats through multiple channels. They fall into categories as follows:

● Optimized self-submission: despite complex steps of submission helping enhance metadata quality, investigation showed that cumbersome and time-consuming submission procedures are a major barrier, and that every effort needs to be made to minimize the amount of work faculty must do to submit their work to the institutional repository, and to maximize the benefits [7]. Moreover, a simple, low-barrier-to-submission is much of the point of setting up the repository in the first place [8]. Therefore, we add a default quick submission workflow, which is highly simplified and optimized for just containing two steps or pages (one for description and another for verification) for completing one deposit in the minimum time of less than 5 minutes. The description page is document-type aware, displaying minimal required description fields and fine-grained rights control options (see below paragraph on dissemination and rights management) at its start, unfolded optional fields can be activated and unfolded to be filled at submitter's will. Here to be highlighted are during the course of submission, a submitter can query publisher's submission policies via integrated ROMEO [9] service by journal title; and after the submission, a background subject indexing process will be activated to assign subject headings automatically based on a third-party OpenKOS [10] service, and a conversion of a non-pdf formatted text document to include a formal pdf version as well for the purpose of online browsing. To keep up with the trend of standardized deposit interface to digital repositories [11], the SWORD based submission module is integrated to support remote deposit via a SWORD clients or a word processor add-in [12-13].

● Revised bulk import: there are needs and requirements for IR to collect and integrate related content from various existent or legacy systems or external applications to speed up the rate of content recruitment. Thus, bulk import has become common features for most repository software [14]. However, there are shortcomings in practice due to the lack of easily used web-based means rather than merely providing command line interfaces. We revise the data import utilities to have web-based interfaces in the first place. A set of fully extensible XML- and EXCEL style templates are adopted to adapt the import process to our needs of ingesting various kinds of content types. For example, a predefined exel_journalPaper template, modifiable in effect, can be followed to organize a batch of journal papers both their metadata and full content information as an Excel sheet to be imported. There are other predefined templates to support import data exporting from Endnote, SCI, CNKI, and other sources.

● Automated ingestion: in an increasingly open and interoperable digital environment, there are opportunities and concerns for automatic capturing content and populating IRs [15-17]. In our settings, we have a CAS-wide pre-existing ETD database that containing data about ETDs from the institutes of CAS. In addition, we have ARP (Academia Resource Planning) systems, i.e. research management systems, deployed in every institute, including data concerning research outputs. To avoid duplicate work, the modules capturing content from these internal applications are added. For capturing related data from external sources, the utility allowing for looking up of Web of Science source record information against Web of Science via its web services API [18], using the institute's subscription entitlements to get metadata from Web of Knowledge. While some publishers are beginning to provide automatic article deposit service [19-20], a SWORD based automatic remote deposit service is also implemented to support such using scenario. For example, CAS has now signed an agreement with BMC to cooperate to accept papers published in BMC publications to be automatically deposited into CAS IRs. Meantime, by incorporating an OAI-PMH harvester module with an extension of harvesting content objects function, it can harvest metadata records and content objects from related repositories via their OAI-PMH data provider interfaces, if applicable.

Dissemination and Rights Management – Although IRs are advocated to be open accessed, they are often faced the complicated content rights management requirements because of diversity of content types and demands for rights and interests. In practice, multi-level and fine-grained access control mechanisms are designed and implemented to support such scenarios. They include defining item embargo period, designating content types aware distribution policies, imposing IP addresses based full content access control, selecting access scope(public, institute, community, individuals, etc.), watermarking the content or not, controlling access types (just allowing online browsing, just allowing online browsing and/or downloading watermarked version, etc.), monitoring and blocking malicious downloading, managing complaints, etc.

Author identification & authorship claim – Uniquely and unambiguously identifying the authors and authorship in repositories are critical to reliably cluster materials related to a specific author. Drawing dome ideas from author-registration approach of name authority control service [21], we work on a mechanism of combining alias control with authorship claim to establish defining authorship relationships between the authors and the article and form a reliable base for clustering related articles by authors. An author or a researcher manages his or her various aliases based on researcher profile service, and all names variants of an author are all point to a unique author identifier similar to researcher ID. While an individual submission is completed and its authorship status is not clearly specified, the service will be automatically activated to check the author/creator values of the submission against alias database. The found matches in forms are taken as possible authors and the emails are sent to them calling for claiming authorship. It is then easy and simple for authors received emails to identify whether or not an item belongs to him or her, and whether or not to follow links contained in the messages to make a claim confirmation. Note that during the course of making a claim confirmation, the author is asked to decide not only the authorship but also the authorship order.

Multi-faceted content use and reuse – Besides offering usual faceted browse and search services, it also provide a facet of clustering results based on dynamically evaluated DDC classifications on OpenKOS service. Google-like auto-suggestion and auto-completion features are also incorporated into CSpace to improve user interactions and experiences. Moreover, while browsing an item, a set of integrated associated services will be displayed. They include bookmarking, recommending, commenting, viewing usage statistics, social bookmarking in Connotea, CiteUlike, Digg, etc., providing reference citation in normal style, showing citation counts in Web of Science, starting an outside search with extracted information in external academic search systems such as Google Scholar, Scirus, CSDL Cross Search, etc., exporting as Endnote-readable data, etc. Moreover, an online pdf version of the item's full content for convenient browsing is available. Thus, all services related to an item are gathered together and made available at the user's fingertips to use or reuse related content freely.

Researcher knowledge profile - Researchers can use this service to create and maintain their research profiles based on IR. A researcher can set up and maintain his or her personal information including personal photos, blogs, educational and academic background, research interests and projects, etc. Based on these data, a personal homepage and knowledge inventory list can be created automatically, and too, be exported as Excel spreadsheet for later use. An alias management module is also provided to help researchers manage their various forms of names, and combination with authorship claim service to lay a foundation for clustering researcher's works correctly.

Usage statistics - It supports analyzing usage impact of IR in a customizable way. Each time, an analyzing process can be executed in combination of various parameters such as different content object levels (site, community, collection, item), different time interval levels (year, month, day, custom time period), different access styles (robot access, intranet access, repeated clicks), different countries or regions, etc. The result can be presented in variety of forms such as histograms, ranking lists, Excel spreadsheets, etc.

Knowledge asset audit - The basic objective of this service is to provide researchers and research managers with tools to perform personalized knowledge asset auditing. It can support performing knowledge asset auditing in combination of a range of dimensions including content organization levels (institution, community, collection, and individual), content types, time spans, etc. The generated results are organized and visualized in various maps. Its highlights of flexible personalization are as follows:

- The overall set of audit conditions can be dynamically defined and configured. The repository manager can prescribe any statistically meaningful metadata element into audit conditions set, if needed.
- Each time of audit process can be customized according to the audit needs and requirements, based on prescribed audit conditions set.

- The audit results manifestations can be customized to display as knowledge inventory lists, histograms, line graphs, pie charts. In addition, the knowledge inventory lists can be exported as an Excel spreadsheet for later use.
- The columns of items appeared in a knowledge inventory list also can be adjusted as desired.

Knowledge mapping – Currently, it supports mapping of co-authorship network with combination of dimensions of all or specified community, research output types, and publication year. It can also mapping IR knowledge domains based on subject categories.

Open Interfaces and interoperability – In an increasing linked and integrated research information environment, it is necessary for an IR to provide open interfaces and support interoperability to avoid it from being information silos. Thanks to contribution from repository community, in particular, DSpace community, CSpace have integrated and implemented many open standard based interfaces such as OAI-PMH, SRU, SWORD, OpenSearch, RSS, and XML Sitemaps for SEO. They really guarantee CSpace based IRs can be easily integrated with other systems or embedded in related research environment.

3. Customization Capabilities

The fundamental and key customization capability is grounded in its extensible and adaptable content management framework, which is comprised of two indispensable components - extensible metadata schema and adaptable content templates. In addition, such extension and adaptation can be easily performed in non-programmable way.

Extensible metadata schema – There is a general agreement that a fixed metadata schema are not accommodating of various forms of content management. Therefore, most popular repository software support extensible metadata schema for the implementation of descriptive and technical metadata [14]. However, their extension is usually operated in technology-savvy way. We provide web form based interfaces to allow for easily changing its QDC-like metadata schema. For example, repository administrator can introduce a new metadata term by defining its basic properties such as element, qualifier, title_en (title in English), title_cn (title in Chinese), scope_note, etc., and its usage related properties such as display_on_submission (whether applicable for submission usage or not), display_on_browse (whether available for browse purpose or not), display_on_stat (whether applicable for content statistics usage or not), edit_allowed (whether allowing for editing or not), etc. Of course, we can also select and modifying properties of existing terms, if applicable. Moreover, if needed and applicable, some terms can be deleted as well. In practice, we see this metadata schema as a dynamic common pool for keeping all metadata terms. Among them, some may be used or reused for several content types and while others may be just suitable for specific content types; still some may be introduced from a standardized metadata name spaces and others may be local defined. Anyway, although this rather mixed metadata schema is peculiarly for local use, the exposure to public for interoperability could be guaranteed by establishing and maintaining an appropriate configurable mapping between inner schema with standard metadata schemas for OAI-PMH compatibility.

Adaptable content templates – Different types of content have different description (in submission) and display (in browse) requirements. In addition, in Chinese context, even for a specific metadata field may have different commonly accepted names or labels. For example, the contributor of a paper or a patent could be both labeled as “contributor” without any disambiguation in English. But in Chinese, while the contributor of a paper should be appropriately labeled with “作者” and the contributor of a patent might be better labeled with “发明人” (in case of patent for invention) or “设计人”(in case of patent for utility). Therefore, content type templating techniques are used to address those specific needs and issues. We define a content type is an aggregation of a set of metadata terms (fields), description and browse forms, and related settings for specific content class. And a content template is created following steps: (1) selecting appropriate candidate metadata terms (fields) from overall metadata schema, and extending new metadata terms if needed; (2) determining the metadata terms (fields) order displayed in submission form, editing form and browse form; (3) defining how to use each metadata term (field), including its display name, input style (e.g. textbox, dropdown list), default value, multiple_value(whether accepting multiple values or not), requiredness, repeatability, hidden(whether or not hidden in folded optional fields list in submission form); (4) defining citation format for this type of content items; (5) creating or assigning related distribution and rights policies. Once a new content type template is created, it can immediately become effective. Those existing template can also be modified and deleted, if applicable. Together with extensible metadata schema, CSpace possesses built-in flexibility and extensibility to manage any existent or new emerging types of content.

Other useful customization options – They include creating and customizing configurable XML import templates to import text data in any format, and creating and customizing content type related EXCEL import templates to import corresponding types of data. Moreover, most of parameters or options needed to be localized during deployment and running are collected together into one web based form to be adjusted or customized conveniently. The skin change functionality is also added for customizing the style of the IR site according to local deployment needs.

4. Uptake and Future Development

CSpace has been deployed over 100 institutes of CAS. By the end of March this year, all of these IRs have reached a significant size of accumulating more than 400,000 research items, having a spectrum of coverage of content types of journal papers, ETDs, conference papers, books, patents, reports, awards, presentations, etc. Of all items, about 76 percent contain open accessed full content. The access usage statistics shows that these IRs have more than 20 million total views and nearly 4 million total downloads. Thus, CAS IR Grid has developed to be the largest and most influential IR network in China. Outside CAS, the Academy of Military Medical Sciences, Shihezi University and other several universities and institutions are also adopting CSpace to establish their IR applications.

There are also good signs that increasingly many institutes and researchers have realized the value of IRs. Based on IRs' open interface, several institutes such as IMECH(http://www.imech.ac.cn/), YIC(http://www.yic.ac.cn/), IC(http://www.ic.cas.cn/), etc., either directly link their research attainments columns in web sites to their IRs, or automatically and dynamically capture related metadata via IRs' SRU api and embed them in corresponding columns in their web sites. In addition, about 20 research laboratories repositories are dynamically linked with their institutes IRs via the SRU interfaces to capture and exchange related data. Meantime, although there have not been concrete statistical numbers, we really know that many researchers have set up their personal researcher profiles or web sites based on their institutes IR platform.

For the future development of CSpace, we will primarily focus on extending its capabilities in aspects of non-textual content management, in particular research data management; automatic metadata extraction and text mining, micro-services based repository infrastructure, and Semantic enhancement services. Moreover, as open source software, we will spare our efforts to develop CSpace community to include contributions from community members and better ensure its future development.

ACKNOWLEDGMENT

This work was supported by the Knowledge Innovation Program of the Chinese Academy of Sciences and the West Light Foundation of The Chinese Academy of Sciences.

References

[1] CAS IR Grid. http://www.irgrid.ac.cn
[2] DuraSpace. DSpace Open Source Software. http://www.dspace.org.
[3] Foster, N. F., Gibbons, S. 2005. Understanding faculty to improve content recruitment for institutional repositories. D-Lib Magazine , 11: 1-12.
[4] Ferreira, M., Rodrigues, E., Baptista, A. A., et al. 2008. Carrots and sticks: Some ideas on how to create a successful institutional repository. D-Lib Magazine, 14: 3.
[5] Walters, T. (2006). Institutional Repositories and the Need for" Value-added" Services, CNI Task Force Meeting. Georgia Institute of Technology.
[6] Salo, D. (2009). Innkeeper at the roach motel. Library Trends, 57: 98-123.
[7] Westrienen, G., Lynch, C. A. (2005). Academic institutional repositories: deployment status in 13 nations as of mid 2005. D-Lib Magazine 11:9.
[8] Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age, ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC, no. 226.
[9] Sherpa/Romeo. http://www.sherpa.ac.uk/romeo/
[10] National Science Library, CAS. OpenKOS for digital knowledge resources environment. http://openkos.whlib.ac.cn
[11] Allinson, J., François, S., and Lewis, S. (2008). SWORD: Simple Web-service Offering Repository Deposit. Ariadne, Issue 54. http://www.ariadne.ac.uk/issue54/.
[12] SWORD client. Http://swordapp.org/category/clients/
[13] Microsoft Office Add-ins for Scientists. http://research.microsoft.com/en-us/collaboration/tools/officeaddins.aspx
[14] Masrek, M., Hesamedin, H. (2012). Evaluation of Three Open Source Software in Terms of Managing Repositories of Electronic Theses and Dissertations: A Comparison Study. Journal of Basic and Applied Scientific Research, 2(11):10843-10852.
[15] Proudfoot, R. E., Sharma-Oates, A., Middleton, M. M., & Shipman, B. (2009). JISC Final Report: IncReASe (Increasing Repository Content through Automation and Services).
[16] Ponomareva, N., Gomez, J. M., & Pekar, V. (2010). Air: a semi-automatic system for archiving institutional repositories. In Natural Language Processing and Information Systems (pp. 169-181). Springer Berlin Heidelberg.
[17] Symplectic. Symplectic Elements. http://www.symplectic.co.uk/
[18] Thomson Reuters. Web of Science Web Services. http://wokinfo.com/products_tools/products/related/webservices/
[19] Nature Publishing Group. Manuscript Deposition Service. http://www.nature.com/authors/author_resources/deposition.html
[20] Biomed central. Automated Article-Deposit. http://www.biomedcentral.com/libraries/aad
[21] Cruz, J.M.B., M.J.R. Klink and T. Krichel. 2000. Personal data in a large digital library. In Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, 127. http://openlib.org/home/krichel/phoenix.a4.pdf.