Building An Institutional Repository With DSpace

Juli Thakuria

Abstract

Paper deals with open source institutional repository software specially DSpace. After defining the terms, it discusses implementation of DSpace as an institutional repositories. DSpace has developed a model that allows users to use the system, submit and use content, and administrators can organize and configure the system. In order to be more usable to different types of users, the software provides a configurable submission and workflow process that can be fit to any organization's information needs.

Keywords: Open Source, Institutional Repository, DSpace, Open Source Software

1. Introduction

Repositories now represent potentially rich sources of information, data, images and valuable research results. The movement is new and the time it takes to plan, formulate policies, and bring institutional communities to consensus can make it a slow process. The Institutional Repositories are powerful systems that allow institutions to store and maintain their digital documents and allow for interaction and collaboration among users in the organizations. There are a number of digital library software available as "Open Source" as well as in "Proprietary format". Open source software helps libraries mainly in lowering initial and ongoing costs, eliminating vendor lock-in and allowing for greater flexibility. The main advantage of open source software is that it is generally available in free. DSpace is a groundbreaking digital library system to capture, store, index, preserv and redistribute all scholarly research material in digital formats.

2. Definition of Open Source

Open source doesn't just mean access to the source code. The distribution terms of open source software must comply with the following criteria:

2.1 Free Redistribution: The license shall not restrict any part from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale.

2.2 Source Code: The program must include source code, and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well- publicized means of obtaining the source code for no more than a reasonable reproduction cost preferably, downloading via the Internet without charge. The source code must be the preferred form in which a programmer would modify the program. Deliberately obfuscated source code is not allowed. Intermediate forms such as the output of a preprocessor or translator are not allowed.

2.3 Derived Works: The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.

2.4 Integrity of the Author's Source Code: The license may restrict source-code from being distributed in modified form only if the license allows the distribution of " patch files" with the source code for the purpose of modifying the program at build time. The license must explicitly permit distribution of software built from modified source code. The license may require derived works to carry a different name or version number from the original software.

2.5 No Discrimination Against Persons or Groups: The license must not discriminate against any person or group of persons.

2.6 No Discrimination Against Fields of Endeavor: The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a business, or from being used for genetic research.

2.7 Distribution of License: The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.

2.8 License Must Not Be Specific to a Product: The rights attached to the program must not depend on the program's being part of a particular software distribution. If the program is extracted from that distribution and used or distributed within the terms of the program's license, all parties to whom the program is redistributed should have the same rights as those that are granted in conjunction with the original software distribution.

2.9 License Must Not Restrict Other Software: The license must not place restrictions on other software that is distributed along with the licensed software. For example, the license must not insist that all other programs distributed on the same medium must be open- source software.

2.10 License Must Be Technology- Neutral: No provision of the license may be predicated on any individual technology or style of interface.

3. Open Source System

Open source system may be classified under the following heads:

- Open Source Library Software: Library Manager, Library Management System, GPL Library Loan Management System, Greenstone3, Librarian DB, NewGenLib.
- Content Management Software: Drupal, Joomla
- Institutional Repository Software: DSpace, Eprint
- Other OSS Resource: Open Source System for Libraries (OSS4LIB), National Research

Centre for free/ Open Source Software (NRCFOSS), Open Source Software Repository (Sourceforge), UNESCO Free & Open Software Portal (Link)

4. What is an Institutional Repository?

An Institutional Repository is an online locus for collecting, preserving, and disseminating information in digital form for the intellectual output of an institution.

"A university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution." (Source: Clifford A.Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age" ARL, no. 226 (February 2003): 1-7.)

An institutional repository may contain work of which the author or institution owns copyright, or for which permission has been obtained to include a copy of the work in the repository. Thus for example - a repository might contain the text of a journal article with the agreement of the author or as a condition of an employment contract. A repository may also contain a copy of the formatted publication with the agreement of the publisher, and authors may be encouraged by their institutions to ensure that a publisher's copyright agreement allows for this possibility. It follows that an institutional repository should not contain content for which suitable copyright or licensing arrangements have not been made.

4.1 Why institutional Repositories?

Repositories provide services to faculty, researchers, and administrators who want to archive research, historic, and creative materials. The open access and open archives movement, the need for changes in scholarly communication to remove barriers to access, and the increasing awareness that universities and research institutions are losing valuable digital and print materials have begun driving the establishment of institutional repositories. Using open archive models, established metadata standards, and digital rights management; new and important information sources are seeing the light of day and becoming more generally available.

While the main purposes of institutional repositories are to bring together and preserve the intellectual output of a laboratory, department, university, or other entity, the incentives and commitments to change the process of scholarly communication have also begun serving as strong motivators. Computers have been ubiquitous on campuses since the late 1980s. Students and faculty are comfortable with the power of online communication. Faculty teachers and researchers want to archive their own materials and have them available on personal or institutional Web sites, these articles, along with the development of the Internet and more powerful search engines, have enabled people to think in practical terms about the establishment of central facilities for storing, archiving, preserving, and making scholarly and artistic materials available. Repositories may be limited to one field, one department, one institution, or a consortium of several institutions. Collaboration through a consortium reduces costs for each member through resource sharing while expanding access to digital materials.

Scholarly societies may establish discipline-based repositories to preserve the history and literature of a particular subject area. However, these societies have a serious dilemma. They publish journals to disseminate research about their fields. If the societies establish open access repositories, they could experience reduced or zero publishing profits, which might in turn affect their ability to pay overhead expenses and to provide enhanced member services. The loss of revenue could place these societies in the position of having to ask members to pay more of the cost of member services.

The increasing demand for scholarly information, especially in science, will probably increase the pressure on scholarly societies and universities. Digital publishing, global networking, more research, and increased communication among communities of scholars are driving the demand for broader access. The idea of the invisible college nurtured by meetings and preprints of journal articles has been replaced by global, discipline-or project-based online communities.

Governments and its agencies may use repositories in the same ways as universities are doing.. Some agencies will find repositories useful for storage and access to technical reports, white papers, hearings, and other documents.

5. DSpace

DSpace (www.dspace.org) is a Digital Repository Software, created as a joint project of MIT Libraries and the Hewlett-Packard Company, and publicly released in November 2002 as Open-Source Software.

The DSpace Digital Repository software is freely available as open source software from SourceForge (www.sourceforge.net/projects/dspace) under the terms of the BSD distribution license. Open source software DSpace is available for anyone to download and run at any type of institution, organization, or company (or even just an individual). Users are also allowed to modify DSpace to meet an organization's specific needs. The specific terms of use are described in the BSD distribution license.

DSpace is one of the open source software platform to store, manage and distribute the collections in digital format. As much of the world's content is now being developed and disseminated in digital format, the DSpace software supports next-generation digital archiving that is more permanent and shareable than current analog archives. DSpace can support a wide variety of artifacts, including books, theses, and 3D digital scans of objects, photographs film, video, research data sets and other forms of content.

DSpace was developed in response to expressed faculty needs for an easy to use, dependable service that could manage, host, preserve, and distribute faculty materials in digital formats. It offers faculty the advantages and convenience of web based submission and dissemination. DSpace can accommodate a variety of genres like: documents, datasets, and images and formats like: txt, pdf, doc, and jpg. It manages and distributes digital items, made up of digital files (or "bitstreams") and allows for the creation, indexing, and searching of associated metadata to locate and retrieve the items. It is also designed to support the long-term preservation of the digital material stored in the repository.

DSpace is also well suited to housing digitized historic collections to enhance the contextual reference for newly submitted works. For the submission of research materials in DSpace, the self-defined, depositing Communities determines who may have access to archived works, with options ranging from a worldwide audience to a select few. There is no charge for submitting to or viewing digital material in DSpace.

DSpace provides a way to manage research materials and publications in a professionally maintained repository to give users greater visibility and accessibility over time.

5.1 Top Reasons to Use DSpace

5.1.1 Largest Community Of Users And Developers Worldwide

DSpace has over 250 institutions that are currently using the DSpace software within their organization in a production or project environment. The most common use is by research libraries as an institutional repository, however there are many organizations using the software to manage digital data for a project, subject repository, web archive, and dataset repository.

A census of Institutional Repositories in the United States was done by CLIR in 2007 and found that DSpace was the preferred Institutional Repository system software of the 446 participants in the survey.

5.1.2 Completely Customizable To Fit One's Needs

Some of the key ways that can customize the DSpace application to suit one's needs are as follows:

User interface -

One can fully customize the look and feel of DSpace website so it will integrate seamlessly with their own institution's website and can be more intuitive for their users. This is possible by using the Manakin extension, which is now to release 1.5.

Ability To Customize the Metadata -

Dublin core is the default metadata format within the DSpace application, however one can add or change any field to customize it for application. DSpace currently supports any non-hierarchical flat name space. However, it is possible to ingest other hierarchical metadata schemes into D Space such as MARC and MODS. This requires using tools such as crosswalk and having some technical capability to map the transfer of data. DSpace is OAI-PMH compatible.

Ability to configure Browse and Search - One can decide what fields would like to display for browsing, such as author, title, date etc. on DSpace website. One can also select any metadata fields would like included in the search interface. All of the text within a given item and metadata associated with the item, are indexed for full text search if desired.

Configurable database -

One can choose either Postgres or Oracle for the database that DSpace manages items and metadata.

Ability To Choose the Default Language -

The DSpace web application is available in over twenty languages so if English is not the local language one can customize.

5.1. Can Manage and Preserve All Types of Digital Content

The DSpace application can recognize and manage a large number of file format and MIME types. Some of the most common formats currently managed within the DSpace environment are PDF and Word documents, JPG, MPG, TIF files. In the near future files ingested will be recognized and validated via the Global Digital format registry (GDFR) and JHOVE application tool.

The system provides bit integrity checking through MD5 checksum reporting. A History system provides an audit trail of all changes to an item and associated metadata.

5.1.4 Used By Educational, Government, Private and Commercial Institutions

The platform is not only used by higher education institutions, who the platform was initially developed for, but also additionally the software has broader appeal. Museum, state archives, journal repositories, consortiums, and commercial companies to manage their digital assets have used the software.

5.1.5 Can be Installed out of the box

DSpace comes with an easily configurable web based interface, where any system administrator can install on a single Linux or Windows box to get started.

6. Planning and Implementing an Institutional Repositories

Each DSpace implementation is unique. While the technology is fairly easy to install and setup, designing and building institutional repository service with DSpace requires planning upfront, before we build the technology platform and launch service. To help plan and build DSpace implementing, offer planning tools and content focused on each stage of DSpace project:

6.1 Defining DSpace Service Offering

DSpace is a flexible and powerful digital repository system. Before build the technical infrastructure of a system, it's important to define exactly how to plan to use the system and what type of service will offer. For example, MIT uses DSpace as a repository for digital research and student theses and has plans for DSpace to serve as the digital archive for all of MIT digital course materials. Cambridge University also will use DSpace for research and theses but will explore the possibility of using DSpace to manage university administrative records. Others might use DSpace as a publishing platform, or simply as a pre-print archive (or a combination of all these uses).

6.2 Creating Service Support Infrastructure

Just as a technical staff will assemble the DSpace technical infrastructure; need to assemble the infrastructure of the DSpace service as well. Building a DSpace service requires input and planning from several sectors of a research institution: library staff and administrators, faculty, and institution leaders.

6.3 DSpace Object Model (Building Collection and Communities)

DSpace is designed to make participation by depositors easy. The system's information model is built around the idea of organizational "Communities"—natural sub-units of an institution that have distinctive information management needs. In the case of MIT (a large research university) , "Communities" are defined to be the schools, departments, labs, and centers of the Institute. Each Community can adapt the system to meet its particular needs and manage the submission process itself.

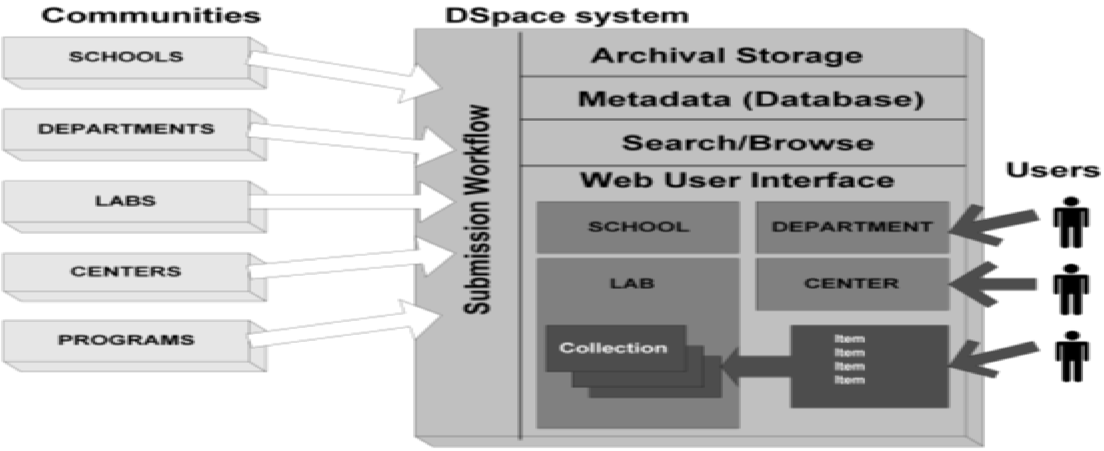

Figure 1: DSpace information model

Items are organized in a hierarchy in which similar items are grouped and submitted into Collection of similar content. Communities are the highest level of content organization. As such a collection can be in more than one Community. Each item stored in a DSpace repository is made up of a bundle of bit streams; so as many files can be stored in a single digital object as needed. Bitstreams adhere to the Bitstream Formats that the system knows about, and DSpace behaves in different ways with different types of objects- e.g., images may have their thumbnails displayed while browsing the system.

6.4 Metadata

DSpace uses a qualified Dublin Core metadata standard for describing items intellectually (specifically, the Libraries Working Group Application Profile). Only three fields are required: title, language, and submission date, all other fields are optional. There are additional fields for document abstracts, keywords, and technical metadata and rights metadata, among others. This metadata is displayed in the item record in DSpace, and is indexed for browsing and searching the system (within a collection, across collections, or across Communities). For the Dissemination Information Packages (DIPs) of the OAIS framework, the system currently exports metadata and digital material in a custom XML schema while we work with the METS community to develop the necessary extension schemas for the technical and rights metadata about arbitrary digital formats.

6.5 User Interface

DSpace's current user interface is web-based. There are several interfaces: one for submitters and others involved in the submission process, one for end-users looking for information, and one for system administrators.

The end-user or public interface supports search and retrieval of items by browsing or searching the metadata (all fields for now, and specific fields in the near future). Once an item is located in the system, retrieval is accomplished by clicking a link that causes the archived material to be downloaded to the user's web browser. “Web-native” formats (those which will display directly in a web browser or with a plug-in) can be viewed immediately; others must be saved to the user's local computer and viewed with a separate program that can interpret the file (e.g., a Microsoft Excel spreadsheet, an SAS dataset, or a CAD/CAM file).

6.6 Technology Platform

DSpace was developed to be open source, and in such a way that institutions and organizations with minimal resources could run it. The system is designed to run on the UNIX platform, and comprises other open source middleware and tools, and programs written by the DSpace team. All original code is in the Java programming language. Other pieces of the technology stack include a relational database management system (PostgreSQL), a Web server and Java servlet engine (Apache and

Tomcat, both from the Apache Foundation), Jena (an RDF toolkit from HP Labs), OAICat from OCLC, and several other useful libraries. All leveraged components and libraries are also open source software

6.7 System Architecture

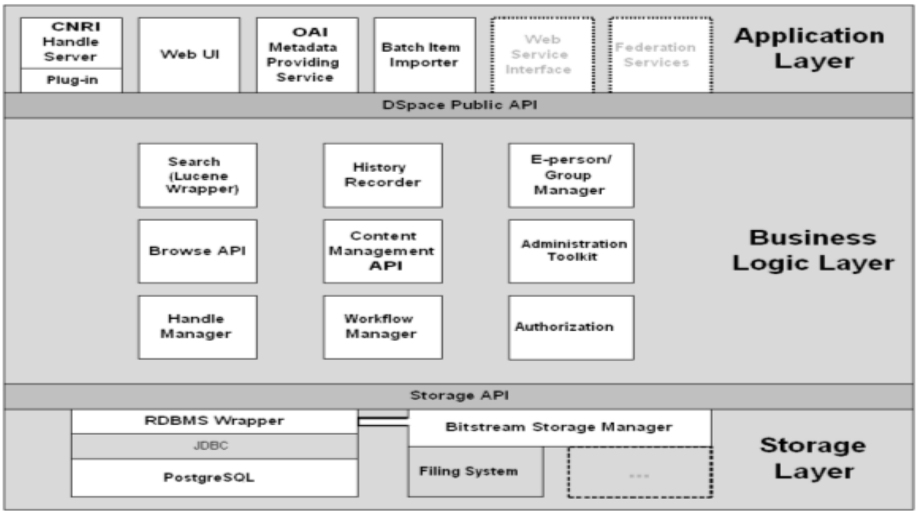

Figure 2: DSpace technical architecture

As detailed in the given figure 2: the DSpace architecture is a straightforward three-layer architecture, including storage, business, and application layers, each with a documented API to allow for future customization and enhancement. The storage layer is implemented using the file system, as managed by PostgreSQL database tables. The business layer is where the DSpace-specific functionality resides, including the workflow, content management, administration, and search and browse modules. Each module has an API to allow DSpace adopters to replace or enhance that function as desired. Finally, the application layer covers the interfaces to the system: the web UI and batch loader, in particular, but also the OAI support and Handle server for resolving persistent identifiers to DSpace items. This is the layer that will get much of the attention in future releases, as we add web services for new features (e.g., to support interoperation with other systems) and define Federation services across the range of institutions adopting DSpace.

6.8 DSpace Ingestion

DSpace is the software that serves as a repository and stores digital content. In a system with such a goal, perhaps the most critical aspect of the system is how that data enters the system. This occurs mainly in two ways in DSpace. The web based UI for the software allows a user to submit items to collections as long as they are logged in as a registered user. When users log in, they go through a configurable workflow where they upload and describe their submissions.

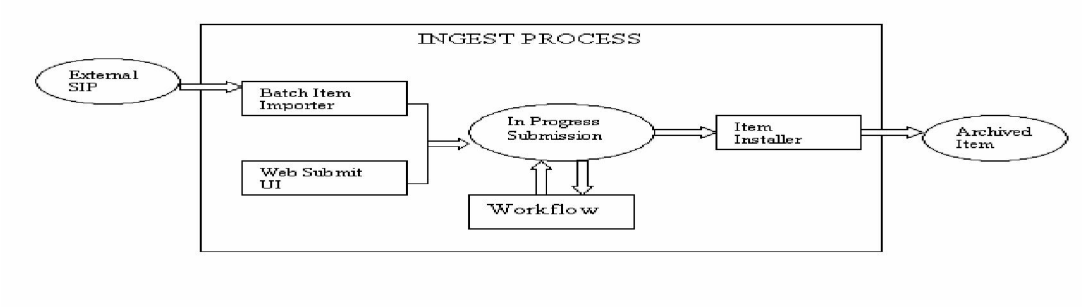

Figure: 3 DSpace Ingest Process

Alternatively, DSpace administrators who have a large amount of content to be batch imported may take advantage of the import/export functionalities of the system. The Item Importer is a command line tool that comes bundled with the system and allows users to import collections of content into the system.

```
archive_directory/
    item_000/
        dublin_core.xml  — qualified Dublin Core metadata
        contents           — text file containing one line per filename
        file_1.doc        — files to be added as bitstreams to the item
        file_2.pdf
    item_001/
        dublin_core.xml
        contents
        file_1.prg
        ......
```

Figure: 4 DSpace's simple archive format for importing and exporting

The Item Importer uses DSpace's simple archive format, which is a simple directory structure that holds items for import into the system. (An example of a simple archive is given above in Figure4). A top-level archive directory contains uniquely named directory, each of which contains everything necessary to import a single item. Each sub-folder is required to contain two files, in addition to the actual content to be imported. The required file “dublin_core.xml” contains an XML representation of qualified Dublin Core element names and the textual content that contains metadata records, including author, title, and so on. A plain text “content” file has one line containing the filename of each file that will be included in that digital object. Once this structure is put in place, the import command can simply be run and all content will be imported into the repository. The tool provides a “map file” after being run, which details all items that were imported and their new location within the system—this file can be helpful in future for exports or removal of imported contents.

6.9 DSpace Workflow

The DSpace submission workflow system is a critical part of the DSpace architecture that allows submission, processing, and final addition of content to the live repository. DSpace’s underlying model includes E-People, users who have registered with the system and have certain authorizations, roles, rights, and privileges that translate abilities to complete certain tasks within the DSpace system. A typical submission begins with the system asking the user a couple of questions about digital document to be added in the repository and number of files involved in the submission. Then the system guides the user through the different steps, which are outlined in the following Figure: 5

Workflow Step	Description
1. Describe	User enters metadata about the document (s) they are submitting, including but not limited to author, title, keywords, and a description.
2. Upload	The user selects and uploads the files on their local machine that they like to upload as part of the submission. Each file’s type is identified by the system and the user verifies the file type
3. Verify	An overview of all details of the submission is given including a summary of the entered metadata and the files involved in the submission.
4. License	The user is shown and must agree to the license the system administrator has assigned to submit content for this collection.
5. Complete	The user’s actions in the submission process are complete. Based on the workflow steps set for the collection, the item may immediately be added to the collection or have to be reviewed by system administrators before its addition to the collection.

Table: 5 DSpace submission workflow overview

6.10 Disseminate

The items submitted and archived into the DSpace digital library repository can be disseminated and accessed by the users through search and browse. DSpace offers users the capability to search DSpace for items of interest both simple and advanced. From the DSpace home page, users can browse all items in DSpace by title, author, or issue date.

6.11 Observation

DSpace provides a way to manage research materials and publications in a professionally maintained repository to give them greater visibility and accessibility over time. It helps to:

- Getting research results out quickly, to a worldwide audience
- Reaching a worldwide audience through exposure to search engines such as Google
- Storing reusable teaching materials that one can use with course management systems
- Archiving and distributing material would currently put on personal website
- Storing examples of students' projects (with the students' permission)
- Showcasing students' theses (again with permission)
- Keeping track of own publications/bibliography
- Having a persistent network identifier for work, that never changes or breaks

7. Conclusion

One of the leading uses for DSpace is as an institutional repository. DSpace followed the librarian's inclination to create a system that would be as easy as possible to implement and use, rather than push strictly in the direction of digital library research from which a more flexible system might have emerged. DSpace, therefore, was designed as an open source application that institutions and organizations could run with relatively few resources. The intention to support interoperability (with DSpace implementers at other institutions, for example) led to the adoption of the Open Archives Initiative Protocol for Metadata Harvesting (OAIPMH). The OAI Registry includes DSpace, making its Dublin-Core-formatted metadata available to compatible harvesting code. In addition, DSpace chose to implement CNRI handles as the persistent identifier associated with each item to insure that the system will be able to locate and retrieve documents in the distant future.

References

1. Sharma Hari Prasad . ýÿ Moving Beyond Library Automation: Role of E- Resources in Academic Librariesýÿ In *University News*, 2008, 46(34) pp.6-10.

2. www.dspace.org/introduction/index.html

3. http://www.sparceurope.org/Repositories

4. http://www.infotoday.com/searcher/may04/drake.shtml

5. http://www.sherpa.ac.uk/documents/SP_iisc-cni_020626.ppt

6. http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf

7. http://www.dlib.org/dlib/january03/smith/01smith.html

8. http://www.dspace.org/mit/plan.html

9. www.dspace.org/implement/leadirs.pdf

10. www.dspace.org/introduction/index.html

11. www.soros.org/openaccess/software/may04/drake.shtml

About Author

Ms. Juli Thakuria, Librarian, Dr. Birinchi Kumar Barooah College, Puranigudam, Nagaon, Assam.