Development of an Institutional Repository using DSPACE

Anjali Maisal

Librarian, Bethune College

Abstract: Institutional Repository represents a historical and tangible embodiment of the intellectual life and output of an institution. It is a visible manifestation of emerging importance of knowledge management within higher education. This paper discusses the creation of an Institutional Repository through Dspace. It aims all individuals who have some familiarity with digital library services and technology and who require more detail than other collateral can provide.

Introduction: Academic institutions of higher learning being coupled with rigorous research in the field of Science, Technology, Social Science and Humanities, have their obvious interest in research output collection, preservation and making them accessible among its peers both at domestic level and also at the National/International arena too. This act of intellectual resource sharing enhances its academic quality and reflects its academic status both at National and International level in comparison to other similar institutions. Traditionally, academicians usually publish the articles (i.e. research output) in their preferred print version of journals and presently some of the publishers are also bringing out simultaneously an electronic media of the same. But the publisher concerned in terms of subscription both again restricts the accessibility to such information for the individual as well as for the institution. At this juncture of time, development in the field of "Digital Library" (DL) revolutionise the library services by way of digital information (i.e. data; articles...) collection, repackaging, and online distribution starting from LAN, WAN, to INTERNET. This DL is not only simple complementing the need of a Conventional Library Service system, rather it has opened a most convenient, rational and a democratic platform for the academicians to contribute and share their intellectual research output through an Institutional Repository (IR), among the peers within the host organization and the outside without having any kind of hassle from the publishing houses.

What is an Institutional Repository?

An Institutional repository consists of formally organized and managed collections of digital content generated by faculty members, staff and students at an institution. The content of these repositories can be available for integration with on-campus library and course management systems, and can also be made available to colleagues and students at other institutions, as well as to the general public.

When we use the term "repositories" today, we can be speaking about one of many different technologies that support the storage and distribution of digital content:

1) Collection-based digital repositories which are managed by library professionals
2) Course management system and associated file stores;
3) Collection of research data and reports managed by academic departments;
4) Institutional file storage system;
5) Digital asset management workflow systems;
6) Web content management systems used by institutions or departments to store and stage web content.

While many of these components can play roles in capturing and managing digital content, an institutional repository is a more specific concept- a centrally managed collection of institutionally-generated digital objects designed to be maintained in perpetuity. An institutional repository will be capable of indexing and serving a wide range of static and moving images, and will be seamlessly visible from course management system, integrated library systems, administrative workflow systems, and via public portals. Institutional repositories can be viewed "as a natural extension of academic institutions' responsibility as generators of primary research seeking to preserve and leverage their constituents' intellectual assets; and as one potentially major component in the evolving structure of scholarly communication."

What is DSpace?

A Software is required to develop the digital repository. There are many open source software available , such as, Eprint, Fedora, Greenstone Digital Library Software, iTor, ROADS, iVia, Phronesis etc. Among these Dspace Software is the best to develop the Digital Library.

Digital Preservation Using DSpace

- Dspace provides long term physical storage and management of digital items in a secure, professionally managed repository including standard operating procedures such as backup, mirroring, refreshing media and disaster recovery.
- DSpace assigns a persistent identifier to each contributed item to ensure it is retrievable far into the future.
- DSpace provides a mechanism for advising content contributors of the preservation support levels they can expect for the files they submit.

DSpace means different things to different people and constituencies. Sometimes it refers to technology for DSpace is an application. 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 adopt 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(Corporation for National Research Initiatives) handles as the persistent identifier associated with each item to ensure that the system will be able to locate and retrieve documents in the distant future. DSpace was also designed with a batch load submission feature to ease the loading of existing collection cut costs.

DSpace the technology, then, is the software- the open source computer application that drives and manages submission, storage, and retrieval processes. But DSpace has non-technological aspects of importance, which this narrative will highlight for the benefit of those who no longer need to develop the application.

DSpace - Origin

In March 2000, Hewlett-Packard Company (HP) awarded $1.8 million to the MIT Libraries for an 18- months collaboration to build DSpace, a dynamic repository for the intellectual output in digital formats of multi-disciplinary research organizations. HP Labs and MIT Libraries released the system worldwide on November 4, 2002, under the terms of the BSD (Berkley Software Development) open source license, one month after its introduction as a new service of the MIT Libraries. As an open source system, DSpace is now freely available to other institutions to run, modify and extend as they require to meet local needs.

Contents of DSpace: Documents for DSpace Repository

DSpace accepts all forms of digital materials including text, images, video and audio files. Possible content includes the following :

- Documents, such as articles, preprints, working papers, technical reports, conference
- Books
- Theses
- Data Sets
- Computer programs
- Visualizations, simulations and other models
- Multimedia publications
- Administrative records
- Web pages
- Published books
- Overlay journals
- Bibliographic datasets
- Images
- Audio files
- Video files
- Reformatted digital library collections
- Earning objects

DSpace is designed for ease-of-use with a web-based user interface that can be customized for institutions and individual departments.

Features

- DSpace supports OAI - PMH (Open Archives Initiatives Protocol for Metadata2 Harvesting) Resumption tokens.
- DSpace also includes batch tools to import and export items in a simple directory structure.
- DSpace exposes the Dublin Core metadata for items that are publicly (anonymously) accessible.
- DSpace uses the CNRI Handle System for creating identifiers.
- DSpace supports uploading and downloading of bit streams as-is. This fine for the majority of commonly used file formats such as PDFs, Microsoft Word documents, Spreadsheets.
- Document discovery and retrieval.
- Digital preservation

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, 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).

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.

Workflow : DSpace is the first open source digital repository system to tackle the complex problem of how to accommodate the differing submission work flows needed for a multidisciplinary system. In other words, different DSpace Communities, representing different schools, departments, research labs and centers, have very different ideas of how material should be submitted to DSpace, by whom, and with what restrictions. Who is allowed to deposit items? What type of items will they deposit? Who else needs to review, enhance or approve the submission? To what collections can they deposit material? Who can see the items once deposited? All of these issues are addressed by the Community representatives, working together with the Libraries' DSpace user support staff and are then modelled in a work flow for each collection to enforce their decisions. The system models "e-people" who have "roles" in the work flow of a particular Community in the context of a given collection. Individuals from the Community are registered with DSpace, then assigned to appropriate roles ; i.e. a department may choose to have two collection : one for working papers and another for datasets. They may then decide that any member of the faculty can deposit items to either collection directly and that any member of the general public can have access to these collections. In this example the work flow is very simple and the only "role" is that of submitter.

In a more complex example, the same department may have a working paper collection that requires tight editorial control by the head of the department. In this case, they may choose to again designate all faculty as "submitters", but also designate a small group of people as "reviewers", an administrative staff person as a "metadata editor", and the head of the department as the final " coordinator". An item deposited by a faculty member would then go through a process of review, clean up and approval before finally being deposited to the relevant DSpace collection. Each person with a role to play in this process is notified of the new submission and goes to a personal work space in the system to perform their assigned task. Items that do not make it through the process are not archived in the system.

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 like 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), OAUCat from OCLC, and several other useful libraries. All leveraged components and libraries are also open source software. The system is available on SourceForge, linked from both the DSpace informational website and the HP Labs site.

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 functionally resides including the work flow, content management search and browse modules. Each module has an API to allow DSpace 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.

Persistent Identifiers (Handles) : One goal of persistent digital repositories is that it will be possible to find and retrieve deposited items far into the future. In particular, it is considered crucial that citations to archived material, whether found in printed articles or online, remain valid for long periods.

Handle resolution can be done using a special client, or handles can be packaged in the form of URLs and a proxy server used to resolve these into the handle form, which is, in turn resolved to the local system location for the item. This second approach is the one we have taken in DSpace. The main alternative to using handles is to use persistent URLs with HTTP redirection to allow items to move around over time. The long-term viability of these alternatives is not yet sufficiently understood.

How DSpace Works ?

DSpace is offered as a web-accessible service. Users access the service via a web browser. DSpace supports virtually any browser for dissemination. To access submit and manage / administer functionality, web browsers must support web forms and file upload. A DSpace site is specific installation of the DSpace software, upon which services are offered that are backed by the commitment of a host institution. A DSpace Host Institution is that institution to whom used rights for submitted material are granted and which stands behind the commitments that are made during the course of offering services atop the DSpace software platform. In the first instance of the DSpace site at Department of Library and Information Science, Jadavpur University, the host institution is DLISc, Jadavpur University. For example, submitters offer distribution rights for submitted content to the host institution (DLISc, Jadavpur University). The host institution (DLISc, Jadavpur University) makes some commitment regarding storage and perservation of the submitted materials and that it will use such materials in accordance with the rights granted.

User (e-Persons) : A DSpace User is an individual who uses the DSpace system, by visiting a DSpace site with their web browser. DSpace users can be at any given time either unknown or credentialed to the system to some degree, for example via user name / password or IP-address-based network presence. DSpace keeps some basic information for registered users (email address, name, credential information), so that they can take advantage of all of the systems functionality (for example : Submission, My DSpace).

Group of Users : DSpace administrators can organize DSpace users into groups, which may be used to define participants in a role within a collection's submission process (e.g., "approver"). Policy statements can also refer to groups of users (e.g. allow users in group "thesis-administrators" to edit the collection metadata for the thesis collection). While it is true that there may be some organizational or socio-political group of people that correspond to an overall social "community", DSpace functionality is concerned with defined groups for specific roles within the community; e.g., who can edit the community's home page ? Who can add collection to the community ? Who can submit items to a collection's submission process ?

Community, Collection and Item : A DSpace Community is a convenient entry point of "portal" into the corpus of material in the repository. A Community consists of a configurable home page for the community, a set of collection referred to by the community and a group of users with management and administrative responsibility for the community. Because communities must be administered, DSpace communities typically correspond - at lease initially - to an organizational entity, for example, a school, department, laboratory or research centre.

A DSpace Collection groups together a set of DSpace items that are related in some way. A DSpace collection consists of a configurable home page for the collection, a set of items referred to by the collection, a configurable submission process for content entry into the collection and a group of users with management and administrative responsibility for the collection. Users who submit items to DSpace can choose a collection to submit to. Further, DSpace administrators can re-organize items into another collection or even multiple collections after their initial submission. Collections typically contain items that are similar in some dimension (e.g., source, purpose, existing series or audience, subject matter, research topic). Administrators can also use collections to organize a submission process for consistency of submitted content (e.g., with respect to scope of content, metadata requirements, required bitstream formats, etc.).

A DSpace Item is a logical grouping of a useful set of content and metadata that are related in some way. Examplex of DSpace items include : a working paper, a conference presentation, a monograph (book), an annotated series of images, a video clip, materials for a course lecture, a research paper with auxiliary material (e.g., dataset, extended bibliography, rich media images).

Browse : DSpace users may browse the contents of DSpace in the following ways :-

Browse Communities and Collection : DSpace administrators organize DSpace items into collections and include collections in communities. Users can browse the structure of communities and collections, in an outline view. Each outline entry includes text describing the contents and / or purpose of the corresponding community or collection. Users can select an entry from the outline view to access the home page of the selected community or collection from which further bounded browse or search can be performed.

Browse Items : From the DSpace home page, users can browse all items in DSpace by title, author or issue date. From a community or collection home page, users can initiate a bounded browse within that community or collection. DSpace supports bounded browse by title, author or issue date. All browses support paginated results displays, with links for previous and next page, as well as short-cuts that allow the user to jump to a particular location within the browse results (for example, browse authors beginning with 'S').

Browse by date and title display one entry for each item. The displayed entry is linked to the corresponding item overview>

Browse by author collapses all items by same author into a single entry, which links to a list of all items having authors with that name. The user may sort this list either by title or by date. Entries from the browsed author's item list link to the corresponding item overview. Future authority control services for authors may allow browse to refrain from collapsing works created by different authors having the same name.

Search : DSpace offers users the capability to search DSpce for items of interest. DSpace offers the following search features through its web-based user interface :-

- Search all of DSpace
- Bounded search, within all of a specified community's collections.
- Bounded search, within a specified collection
- Simple search : Searches fields : author, title, keywords
- Case insensitive search : All searches are case insensitive.
- Truncation, Constraints

a) Input : Program*
Output / Result : Program,
Programme,
Programming

b) Input:?i?er
Output / Result : diver, wiper

c) Input: + gone + wind
Output / Result : "gone with the wind" (but not "here today, gone tomorrow" or "the cold north wind")

- Word Stemming
  Searches for "processed" match any of [process, processing, processor]

- Stop words
  Common words (e.g. "a", "an", "the") are Omitted from the search

These query features are available for DSpace baseline metadata. Other metadata submitted with the item (i.e., in bitstreams within the item) is currently neither indexed nor searchable by DSpace.

Search Results : Once the user specifies a search, DSpace performs the search and produces a result set. DSpace displays the result set, including a short description for each item in the results. From the short item description displayed in the search results, the user may select a desired item to view its Item.

Conclusion

We are likely to see the concept of institutional repositories develop in both convergent and divergent ways over the next few years. Not every institution will develop a formally managed institutional repository along the lines of DSpace. But every institution that is utilizing course management systems, library catalogue systems and student portfolio systems will see increased "repository-like" functionality in their products. The open source movement, coupled with greater network collaboration among researchers, should give rise to discipline-specific federated repositories hosted by institutions, research projects or professional associations.

Institutional repositories are visible manifestation of the emerging importance of knowledge management within higher education. Paradoxically, "scholarly respect for knowledge and a desire to ensure academic freedom make most institutions reluctant to manage knowledge of any sort". The long-term impact of institutional repositories is likely to change many of the basic assumptions about how intellectual output is managed by individuals, their colleagues and the academy and how research itself is conducted.

References

1. Lynch, C.A. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. ARL Bimonthly Report, 226, February 2003.

2. Brainin,Joseph. Institutional Repositories. Encyclopedia of Library and information Science. Marcel Dekker, 2004, pp. 1-11.

3. Johnson, Richard K. Institutional Repositories: Partnering with Faculty to Enhance Scholarly Communication. D-Lib Magazine, November 2002.

4. SourseForge.net, <http://sourceforge.net/projects/dspace/>.

5. DSpace, <http://dspace.org/>.

6. Open Archives Initiative Protocol for Metadate Harvesting (OAI-PMH), <http://www.openarchives.org/OAI/openarchivesprotocol.htm>.

7. OAICat can be found at <http://www.oclc.org/research/software/oai/cat.shtm>