DSpace Summer of Code Ideas 2008

This is an archived listing of DSpace project ideas for Google Summer of Code in 2008. To see which projects actually took part in GSoC 2008, please visit the Past DSpace Summer of Code Projects page

Contents

Ideas for Potential Google Summer of Code projects 2008

Storage and Database

Refactor DSpace to use a common data abstraction layer (HP and Others)

DSpace code is currently fairly closely tied to the underlying database on which it runs. Refactoring the data access layer to use something like Spring or Hibernate would open up the possibility of providing better support for more databases, rather than the current options of PostgreSQL or Oracle (with MySQL support available through a patch).

A good project would be to build a version of DSpace that uses Hibernate as the storage layer, removing all database platform-specific dependencies.

 Several efforts already exist in this area on the 2.0 roadmap, but there may be room for participation --Mark Diggory 10:56, 18 March 2008 (EDT)

This is already being worked on in the sandbox repository (a hibernate prototype) --James Rutherford

Themes and Aspects for XML-UI

Manakin Themes

- Using the SIMILE timeline tool or Exhibit

- Adapting css themes from http://www.csszengarden.com/ and http://alistapart.com/ to showcase how easily Manakin can be skinned to create dramatic brandings.

Manakin Aspects

- External Search/Link-back Service Aspect capable of serving up links to your favorite aggregator/search engine from within DSpace Communities, Collections, Search Results and Items (Google Scholar, SFX, MIT Document Services, RefWorks)
- Subject Overlay Aspect - Subject-centric overlay that will provide a subject centric Search/ Browser through a set of communities/collections in DSpace/Manakin. Includes extensions to theme'ing to support different branding over subjects (instead of just communities/collections).

DSpace UI + AJAX

Creating (or integrating) some AJAX tools for the DSpace Web UI. Some specific examples could include: auto-complete/listing already entered values (in search and metadata entry for subjects, author names, etc), simple spell check (in metadata entry and search), highlight matching terms (in search results), etc.

Search, Browse, Discovery and Semantic Web

I believe that the Manakin UI is an excellent starting point to Semantic Web enable DSpace. That by making some simple adjustements to our internal metadata formats and crosswalks and exposing those to the Manakin Aspect transformation chain, we can begin a process to make all DSpace isntances better participants in the Semantic Web world (and likewise, lay a foundation for introducing up-incoming standards like ORE into DSpace). --Mark Diggory 11:29, 10 March 2008 (EDT)

Port DSpace Crosswalk API to be a SAX/XSLT pipeline rather than JDOM

History repeats itself... the Cocoon folks learned long ago that Pipelines would be both more efficient and more flexible if they were SAX driven rather than DOM driven. DSpace could learn a lesson from that play-book and re-implement the Crosswalk API to Be a suite of SAX XMLReaders That take DSpace Objects and serialize them to XML. This would leverage the existing work done in the Manakin DSpace Adapter API and bring it into a new Addon or dspace-api directly making it available as a much more efficient mechanism for getting DSpace Objects into XML.

Replace DIM with RDFXML as default "DSpace Internal Metadata" format

Replace DIM with RDFXML as default "DSpace Internal Metadata" format for embedding RDF into METS manifests in Manakin, OAI, or Packager

This is a project that would begin to solidify the use of RDF within DSpace as a descriptive metadata solution. It would allow Manakin to begin working with RDFXML out of the box and standard adapters could be written which work with the RDF Crosswalks to render descriptive metadata. Ideally, RDF would be a replacement for DIM and allow us to begin utilizing RDF at the core of DSpace.

Embed RDFa and Microformats into Manakin

Utilize above work to introduce RDFa into the Manakin Site, Community, Collection and Item dri2html theme templates. So that Semantic Web tools can directly glean descriptive metadata out the xhtml directly.

http://www.w3.org/TR/xhtml-rdfa-primer/http://en.wikipedia.org/wiki/RDFa

Ideally this will also form a foundation for introducing ORE embeded directly into the rendered page.

This should almost certainly be for both the JSPUI and XMLUI --James Rutherford(*GSoC Projects shouldn't be forced to have such a requirement IMO --Mark Diggory 12:30, 5 April 2008 (EDT)*)

Customize Manakin Aspects to render formats other than DRI

Enhance Manakin Aspects to support multiple "named" pipelines (RDFXML, DRI, etc) Expose RDFXML in a cocoon pipeline (or Manakin style Aspect) which can be customized by the implementor to inject new RDF into UI for rendering. An RDFXML SAX generation library has already been created and integrated (external to Aspects) by MIT Libraries and can be the foundation for further integration into Aspects.

Proposed by: Mark Diggory. DSpace Systems Manager, MIT Libraries, email: mdiggory@mit.edu

Search - Revise search to use Solr instead of plain Lucene

Solr is an Apache project that extends Lucene to provide (perhaps most notably of the list of features) faceted search and the ability to index and search specific fields.

Metadata Tracings

Enable tracings on such metadata fields as Author, Subject, etc. to launch a repository search for the metadata value from within another item or search results list (i.e. clicking 'Albert Einstein' in the author field display of one item would search the repository for all occurrences of 'dc.contributor.author=Albert Einstein' in the repository).

I think DSpace JSPUI already has this feature? In XMLUI this would be xslt development and in general Manakin/XMLUI could use some more extensive development in this area --Mark Diggory 12:22, 5 April 2008 (EDT)

Metadata Registry Service

Stackable/Plugable service that can be used to query disparate CV, Ontological, Naming registries, via a shared query syntax (XQuery? Sparql?) returning XML/RDF. Our first goal being the ability to plug these services into Fields in the Customizable Submission workflow pages to populate suggested values for fields. (Sources: LDAP, JDBC, DNS, XCat/OCLC/Barton, Google Scholar etc, GFR, other Metadata Registries).

Stretch it even further... DSpace itself is a managed registry of Metadata. A properly architected infrastructure that would allow the meshing of DSpace and other registry metadata in a "Semantic Web, Linked Open Data" way would establish a foundation for exposing DSpace instances as LOD, providing for SPARQL endponts on both ends of the design means that DSpace instances could become both producers and consumers of LOD. --Mark Diggory 12: 28, 5 April 2008 (EDT)

LinkOut

The metadata of a document can be very useful to propose services to the user around the document.
For instance:

- an author name can be used to send an e-mail (if it is a DSpace user) or to make Google Scholar Search
- an ISSN can be used to access publisher home page
- the title can be used for a citation search
- a CAS can be used to search a chemical database (or any ID of a gene, plant, etc. can serve to make a database search).
- an institutional Id. can link to applications around this Id providing some service
- etc.

 I would propose to be able to parameterize a link from any given metadata field to a "linked services" page (services provided by external applications OR by DSpace itself, for instance documents of the same author, on the same subject, etc.).
This is a generalisation of existing features proposed by different patches. Christophe Dupriez 14:50, 17 March 2007 (EDT)

Build, Installation, Testing and Running DSpace

Create a DSpace RPM/Debian package/Windows installer

Maybe a better idea to create a Java based installer that can work across platforms?

Open source installer generators

Sanity Checking Framework

- checking for prerequisites and rights
- are all components installed in running
- system diagnostics
- other dependencies
  like code and content
  metadata registry, input-forms and Messages.properties
- amount of bitstreams in db and assetstore

Build an automated testing system (unit tests, kinda)

- Create some automated tests to detect bugs in DSpace code, particularly in the org.dspace.content package. It does not need to be particularly fast. The test rig could load some test data into DSpace, check that it's there, perform various manipulations and check that the expected results appear. Authorisation would also be an important factor to test.

Unit tests that cover a lot of this have been written for 1.6 (though they aren't yet in trunk). What I would like to see is a mechanism for the construction of 'sandbox' environments for running such tests in. For example, proper testing would require a test database, test asset store, etc, which is all quite messy to do by hand, and tricky to automate. --James Rutherford

A useful precursor to unit tests, IMHO, would be an eclipse project file setting. This would allow developers to more easily build eclipse without them having to do all the IDE setup themselves. Then you could use the builtin JUnit support that eclipse has. Plus, of course, eclipse would help do development generally.

Profiling and Load Testing

Get some real scalability figures

With some given hardware, load large numbers of objects, large objects, lots of users etc. etc into a DSpace and get some real, empirical data about where the bottlenecks are, at what point the system starts to fail, how many concurrent users the system can cope with etc.

Administration and MyDSpace

Administrative and Collection Management Reports using Exhibit

Exhibit is an excellent and easily embeddable report explorer that could be utilized when exploring small (<2000) sets of items. --Mark Diggory 11:29, 10 March 2008 (EDT)

Exhibit: http://simile.mit.edu/exhibit/

Inline example: http://members.porchlight.ca/bower/simile.html

Embed Exhibit with inline JSON directly into new Aspects to do things like "My Submissions", "Manage Collection Submission" and/or "Manage Collection Workflows" Where the Collection Manager would have the ability to browse existing Workflow Items filtering down to the criteria exposed in facets and take /give/remove ownership of these tasks.

Proposed by: Mark Diggory. DSpace Systems Manager, MIT Libraries, email: mdiggory@mit.edu

DSpace Export/Import of citations from/to Bibliographic Software

Add a export feature so anyone can export the metadata in DSpace to bibliographic software such as EndNote and RefWorks. Second, develop an import feature so administrators can import citations from from bibliographic software such as EndNote and RefWorks to create records in Dspace.

Exporting citations is a request of many faculty and a positive selling feature for Librarians/Project Managers to help convince faculty to deposit their work into Dspace. Importing citations will help Dspace administrators get records into Dspace faster and easier. It's all about getting content into local instances of Dspace.

Some generic bibliographic file formats like RIS, which is widely adopted by both
reference managers (Endnote, RefWorks, ...) and digital library systems
(IEEEXplore, ACM, ...) and BibTeX should be offered as as a minimum.

Manage Documents

Users should be able to manage documents of the IR in their DSpace. They should be able to create structures to manage documents. Could be done as a kind of mapping to MyDSpace. Furthermore there should be an option to "automap" new documents based on the alerting mechanism.

This is closely related to being able to ex/import different bibliographic formats.

Cart Functionality

Add a mechanism to allow for selection of multiple individually-selected items, queried search results, and filtered search results into a 'temporary' holding space for a later activity. Ideally, this holding space would be persistent for a logged-in user of DSpace, perhaps as an expansion of the present MyDSpace functionality. At least, allow this to be available for the length of the session. Possible (current or future) uses for the cart contents might include:

- Input data for an ad-hoc report (e.g. statistics, metadata field, etc.)
- To perform a repository maintenance activity ((e.g. create 'virtual' community for cross-discipline research)
- Export/e-mail citations in EndNote format for these records
- Use contents to generate citation analysis report
- Set RSS feed for all authors represented by these items
- Use content as input for data visualization tools such as Exhibit's maps and time-lines

Expand Admin UI to include (most) configurations

Create an Admin UI which can allow uses to edit most simple configurations (from dspace.cfg, input-forms.xml, etc) without actually having to dig into the configuration file (and restart Tomcat to refresh the cache).

 this will require a significant rewrite of the ConfigurationManager, IMHO the UI should know nothing about what the serialization of a configuration property looks like and an infrustructure for managing editable Configuration properties would need to be added to the core of Dspace. --Mark Diggory 12: 34, 5 April 2008 (EDT)

User Management - Enhancement

- User self management
  - delete ones account
  - reminder of account
  - reminder of unfinished tasks and submissions
- User contact (for alerts, feedback, ads)
  - all active users
  - all registered users
  - users of a group
  - users with special rights (i.e. all submitters)
- manage non-active users
  based on set of rules
  - reminders sent
  - deletion
- check for invalid accounts
- validate unsendable emails (registration, alert and so on)

Item Mapper

The item mapper should be redone. Item should be mappable from the item itself. At the moment you must start a search (author only) to get the items. This would among other things include a "navigationable select collection" functionality, which would be usefull for other stuff like starting a submission from My DSpace.

Workflow Mechanisms

Define different workflow types like

- peer reviewing
- pipe through a plagiarism finder tool

Long Term Preservation

Format migration UI

Create plug-in interface (a la PluginManager; maybe using MediaFilter?) that can convert from BitstreamFormat X to BitstreamFormat y. This could wrap tools like OpenOffice (Tim Donohue's work), ps2pdf, latex2html etc. etc. Then a simple admin UI could allow administrators to initiate conversions of particular items, collections, or the whole site.

NOTE: The first part of this is already somewhat "completed" in my work on the OpenOffice.org MediaFilter. To create such a custom MediaFilter, I had to change the *old* MediaFilter abstract class into a Java *interface*, so that you can create MediaFilters which are *also* Plugins (a la PluginManager). I released a patch a while back (SF#1589429, perhaps wrongly named "Named MediaFilters"), which does exactly this and is required for the OpenOffice.org work I've done. *That being said, I like the simple admin UI idea to potentially allow for*any *MediaFilters to be kicked off (and perhaps even configured more easily) via an admin UI.* - TimDonohue

- NOTE: Anyone looking into this area should take note of the "Transparent Format Migration" work of the LOCKSS team (Rosenthal et.al.) [-JohnErickson
- NOTE: This should be based on the BitstreamFormat+Renovation] work Completed by Larry Stone earlier this Year. --Mark Diggory 11:24, 18 March 2008 (EDT)

New ideas

Research Trend Analysis using Institutional Repositories

Institutional repositories gather an organization's scholarly content and buttress knowledge sharing and dissemination of intellectual output. The communities and collections in a repository are designed according to an institution's distribution of research centers, schools and divisions. Each collection that represents a school, division or research centre holds erudite contents facilitated by respective academic projects mentored at those centers. By performing Co-Word analysis on each document collection, the individual research strength of that division or school can be found out. The same principle can be propagated to sub-collections as well as communities in a repository. This extrapolates the research strength of a particular division /school and in general can be extended to determine the profound research strength of an institution. In addition to this, a qualified variation or trend in the research strength of individual divisions/schools can also be found out by applying Co-Word analysis over a predetermined period of time. The above described feature can be extended as an add-on to the existing framework of DSpace and would facilitate the burgeoning representation of DSpace as a research platform.

Proposed by: Jayan C Kurian, Research staff, National University of Singapore, Singapore. email: Jayan@comp.nus.edu.sg, jayanntu@gmail.com

Potential Student: Ashly Markose, Post-graduate student, National University of Singapore, Singapore email : xxxxx@nus.edu.sg

Content Submission in DSpace supported by Domain Directory Services

Individual content submission strategies need to be highly supportive for facilitating potential digital resource submissions to repositories. One of the factors affecting content submission is authenticated sign-on to repository collections. This can be supported by a single sign-on authentication mechanism for content submissions by authorized users. Users in a domain directory service are identified by a fully qualified user context. A user context of the form cn= Jayan ou=Users, ou=SCI, dc=staff, dc=ntu, dc=edu, dc=sg will specifically identify the active directory group to which a user belongs. Each collection (e.g. Schools/Divisions) in a DSpace repository can be assigned with an authorization group using the DSpace in-built functions. By extracting specific features from a user context, a user can be added to a pre-defined authorization group corresponding to respective collections. Thus, when an authorized user authenticates against a DSpace instance, the user would be automatically embedded with content submission privileges only to privileged collections. This also eliminates the task of user selecting an appropriate collection to submit and displaying collections at large. In addition to this, specialized collection access privileges can be given to content submitters who share common active directory context profiles. The authentication strategy would be tested on a Windows environment supported by Windows Active Directory Services.

Proposed by: Jayan C Kurian, Research staff, Nanyang Technological University, Singapore. email: Jayan@ntu.edu.sg

Potential Student: Sunil Thomas, Post-graduate student, National University of Singapore, Singapore email : thomas.sunil@nus.edu.sg

Adaptive Question Answering System based on the DSpace Mailing List Knowledge Base

Recent years have witnessed the tremendous usage of repository software since majority of scholarly content are published in digital form with no exception to the proliferation of DSpace instances. Popular software does manage user queries through mailing list supported by dedicated committers and contributors. A good number of questions asked in a mailing list would have been responded previously. In this case, a Question Answering (QA) system would help users by answering their questions, if it has been responded earlier or would suggest related answers encompassing the subject asked. For this, information available on the DSpace mailing list knowledge base can be extracted using template based extraction techniques or a rule based system. Once extracted, this can be classified according to a taxonomical structure (i.e. Functional Overview, Installation, Upgrading, Configuration, Customization, Architecture, and Versions) that represents the DSpace system documentation. The keywords automatically generated from the message text improve the adaptive retrieval of relevant information in this QA system. A test-bed for this QA system can be build using the DSpace platform and the taxonomical structure can be facilitated by the in-built controlled vocabulary feature.

Proposed by: Jayan C Kurian, Research staff, National University of Singapore, Singapore. email: Jayan@comp.nus.edu.sg, jayanntu@gmail.com

Potential Student: Sunil Thomas, Post-graduate student, National University of Singapore, Singapore email : thomas.sunil@nus.edu.sg

An Add-On to facilitate the existing DSpace Batch Import Procedure

Efficient content acquisition strategies make it easier to import scholarly information into repositories. DSpace supports batch content acquisition through the ItemImport procedure. This procedure requires digital resources to be represented in a Submission Information Package (SIP). The lead time required for preparing this format can be facilitated by encoding document metadata and digital resource location in a spreadsheet. This has been implemented at The Nanyang Technological University (Singapore), The Institute of Scientific and Technical Information of CNRS (INIST-CNRS, France), The University of Calgary Library, National Informatics Centre (India), and The Lanzhou Branch of Chinese Academy of Sciences (China). Few recent requests include The University of Waikato Library, The University of Sydney Library and the NITLE (U.S.A). Although the current implementation on Windows environment looks promising for the user community, there has been considerable request (New York University Library, Raman Research Institute Library (India).....) to make this development compatible with the UNIX environment. The proposal describes the following additions. (1) A GUI interface to facilitate the SIP preparation. (2) Compatibility with the UNIX environment. (3) Utility to automatically bridge the SIP generation and ingestion into specified collections. (4) Exploring the feasibility of template based document metadata extraction from digital resources into spreadsheet. It's anticipated that this add-on would facilitate content acquisition in DSpace installations.

Proposed by: Jayan C Kurian, Research staff, Nanyang Technological University, Singapore. email: Jayan@ntu.edu.sg

Potential Student: Sunil Thomas, Post-graduate student, National University of Singapore, Singapore email : thomas.sunil@nus.edu.sg

Potential Student: Blooma Mohan John, Research Student, Nanyang Technological University, Singapore email : bl0002hn@ntu.edu.sg

Changed AIP to SIP from the OAIS recommendations, AIP are more of an internal abstraction of an IP, SIP and DIP (Dissemination Information Package) are IP's that are transported between systems. --Mark Diggory 12:16, 17 March 2008 (EDT)

Integrate JA-SIG Central Authentication Service

JA-SIG Central Authentication Service (CAS) is an open source authentication system originally created by Yale University. The single sign on authenticates the user to access all the applications he or she has been authorized to access. It eliminates future authenticaton requests when the user switches applications during that particular session. It is the most popular single sign-on solution for universities. For details about JA-SIG CAS, please visit http://www.ja-sig.org/products/cas/. JA-SIG CAS has been deployed by (at least) 80 universities worldwide, and the user base is over 1 million!

My idea is to create a CAS plug-in for DSPACE to allow single sign-on with CAS. It will allow university users to use their campus-wide ID/password to use DSPACE. This plug-in will do more than just single sign-on. It will integrate with DSPACE's built in access control so that administrator can set access to particular user group, for example: faculty only, staff only, or students in Computer Science only, etc.

DSPACE and CAS are popular open source applications that both have a large education user base. The integration of the two will make DSPACE more appealing.

Proposed by Minghui Yu, 4th year Computer Science student, University of British Columbia, Vancouver, Bc, Canada. Email: minghuiy AT interchange DOT ubc DOT ca.
Personal website: http://www.ugrad.cs.ubc.ca/~s3p5/

This has already been done. See http://sourceforge.net/tracker/index.php?func=detail&aid=1601221&group_id=19984&atid=319984 --James Rutherford

Explore integration of DSpace and Fedora

Fedora is a popular Middle Tier solution for IR that utilizes SOAP/REST for accessing and modifying stored content. Explore options to integrating Fedora with DSpace at the Assetstore level and with integration into the Community, Collection, Item, Metadata areas of DSpace.