Open Repositories 2009

This is a summary of my experiences at Open Repositories 2009 which was held at Georgia Tech in Atlanta, Georgia. The Conference was held for two days followed by two days of User Group Meetings.

Some highlights from the Conference:

1. Global Registries Initiative – Jeremy Frumkin (www.ockham.org) working with the University of Manchester and the Australian National Data Service is working on creating a federated network of research data collections. Collections and services records are harvested locally and aggregated nationally & globally. An example can be found at http://registry.ockham.org.

2. Many Lightweight Views into Complex Repository Content: Enabling Rapid Application Development for Fedora Repositories – Matt Zumwalt (Media Shelf) is working on agile development. Ruby on Rails that floats on top of Fedora. This gives more views of data in a single repository which results in the appearance of having different repositories. Many lightweight views into complex content is called "active Fedora".

3. Eliciting Faculty Requirements for Research Data Repositories – Michael Wiit (Purdue University). Sponsored by IMLS, investigators from Purdue and the University of Illinois studied which researchers are willing to share data, when, with whom, and under what conditions. They developed 10 questions to begin a conversation with your faculity about data curation:

    a. What is the story of your data?
    b. What form/format is your data in?
    c. What is the expected lifespan?
    d. How could data be used/reused/repurposed?
    e. How large is your dataset and it's rate of growth?
    f. Who are the potential audiences for the data?
    g. Who owns the data?
    h. Does the dataset include sensitive information?
    i. What publications/discoveries resulted from the data?
    j. How should the data be made accessible?

4. Research 2.0: Evolving Support for the Research Landscape – Make Leggott (University of Prince Edward Island). The University of PEI has developed a Drupal/Fedora data repository and collaborative web environment to accommodate a wide range of research

requirements. The Repository is hidden and the user logs into a Research Space. This is called "Islandora" with Drupal front-end and collaborative editing layer. Fedora handles the data assets, metadata and policies.

5. DataONE (Observation Network for Earth): Envisioning a New Distributed Organization and Cyberinfrastructure to Enable Science – John Kunze (California Digital Library). DataONE is a distributed collection of science data (especially data on global change). Cornell is a partner in this effort. CDL is imagining a non-repository. It is thinking of things in terms of micro-services. These would be unbundled alternatives to monolithic systems with a single archival "culture". Micro-services would provide a low barrier and low commitment tools. They would be decoupled in design and recoupled in deployment. The web is the "de facto" distributed filesystem. In this process "WGET" is the automated client. Things are arranged using the pairtree hierarchy-based collection (for example, id/en/ti/fi/er/identifier). The work used the HTTP URL Mapping Protocol, THUMP – based on the commands that happen after the "?". CDL will be hosting iPRED 2009 in San Francisco (Oct. 5-6).

6. The Data Conservancy: A digital Research and Curation Virtual Organization – Sayeed Choudhury (Johns Hopkins University). Data curation is not an end, but rather a means to collect, organize, validate and preserve data to address grand research challenges that face society. Data Conservancy focuses on the connection of systems into infrastructure through a program informed by user-centered design and research, sustained through a portfolio of funding stream and managed through a shared governance structure.

7. DuraSpace – the new organization that combines the DSpace Foundation and the Fedora Commons. Headed by Sandy Payette and Michele Kimpton, the DuraSpace will maintain Fedora and Dspace and offer new technologies and services (such as DuraCloud). There were several presentations by Sandy and Michele. They are combining their staffs and work is progressing on new releases of the DSpace and Fedora. DSpace 2.0 (planned in 2010) promises to use Fedora as a repository and to be more modularized. The idea of communities and collections will give way to the idea of "entities".

8. Naming, Branding and Promoting the Institutional Repository: A Social Marketing Approach from a Canadian Perspective – Wayne Johnston (University of Guelph Library). The purpose of social marketing is to affect behavioral change to benefit the individual. The Four P's of marketing are:
    a. Prudent
    b. Price
    c. Place

d. Promotion

Branding is important.

Strategies that need to be imployed:

a. Events and presentations
b. Liaison with librarians
c. Brochures
d. Web Content
e. Articles in the campus newspaper
f. Seeding the repository
g. Incentives
h. Usage feedback

9. Secrets of Success: Identifying Success Factors in Institutional Repositories – Elizabeth Yakel (University of Michigan)
    a. Success Measures
        i. Content recruitment - success can be measured only when you meet a broad approval by communities
        ii. Services – without a set of services, it is not a good reason to have an IR.
            1. Search
            2. Discovery
            3. Promotion
            4. Preservation
        iii. Sustainability
            1. integrated into Institutional planning
            2. funding
            3. relationship to other IRs on campus
            4. interoperability
            5. Documentation/measurement
    b. Impact Measures
        i. Outcome versus outputs
        ii. Internal vs external indicators of success
        iii. Unintended consequences
        iv. Longterm view
    c. Content
        i. Library as publisher
        ii. Dealing with more and different types of content
        iii. Library as curator
        iv. IR was seen as a way to build an infrastructure and provide stewardship
    d. Technology
        i. Building technological competence

ii. Experience with new and different technologies
iii. Digital preservation experience

e. Role
   i. Library as Publisher
   ii. Library as participant in discussion & answers about content curation throughout campus
   iii. Library as function not as a building
   iv. Getting into Scholarly workflows

f. Mission
   i. Framing the IR
   ii. Changing the message from an IR to author rights was a key for getting faculty by-in at Michigan.

g. Use – from a scholar's point of view:
   i. Citation
   ii. Ranking
   iii. Access to promotion of materials
   iv. Preservation

10. Reusing Open-Access Content Using Authoring Tools – Lieven Droogmans (@mire). @mire has created a plug in for Word and Powerpoint that allows the user the ability to submit from the tool bar into their repository.

11. Adding OAI-ORE Support to Repository Platforms – Alexey Maslov (Texas A&M University). Developing code which will be incorporated into DSpace, Texas A&M has developed a arvester using ORE and PMH protocols to develop a statewide ETD repository. This code can synchronize DSpace instances.

12. Cloud Computing – dynamic capacity (elastic) with high availability
    a. DuraCloud (from DuraSpace) – provides trust and reliability in a cloud. Can be used to replicate to multiple storage providers (2nd Qtr 2010)
       i. Cloud issues:
          1. Security
          2. Transparency
          3. Data lock in
          4. Service Level Agreement
          5. Trust
       ii. DuraCloud provides:
          1. Replication to up to 3 providers
          2. Web based "dashboard"

3. Data integrity checking & monitoring
4. Can push content from DSpace & Fedora repository platforms via plug-ins
5. Pay-per-use
6. Initial compute services on content
7. Services:
    a. Search
    b. Aggregration
    c. Streaming service for videos
    d. Migration
    e. Hosting repository

b. Cloud Task Replica – Towards a Preservation Strategy – Richard Rogers (MIT). MIT has been working on using the Cloud as a Preservation tool. This has all been theoretical. They look at the Cloud as having reliable messaging; enabling asynchronous handling; coordination of work; access control; and being cheap. They view roles involved as
    i. Archive – content home
    ii. Replicator – manages policy
    iii. Auditor – implements policy
c. From Desktop to the Cloud: Leveraging Hybrid Storage Architectures in Your Repository – David Tarrant (University of Southampton). Part of Eprint 3.2, Eprints has a storage controller that decides where to put a file. It uses a rule based policy defined by a simple XML configuration file.

13. Mounting Books Project – Steve DiDomenico (Northwestern University). Northwestern has developed a software system and workflow that takes books from the Kirtas scanner and moves them through quality assurance and structure build process. The metadata is stored as METS files. Handles are generated, and the books can then be "uploaded" into Voyager with an 856 field. A viewer is provided to allow the person to read the book.

14. DSpace User Group Meeting Highlights
    a. DSpace 1.6 planned for Fall 2009 – stepping stone to DSpace 2.0. Promises better statistics, embargo facility & batch metadata editing
    b. Source code for DSpace moved to OSU Opensource Lab
    c. Developers for DSpace and Fedora will be reporting to the same Management structure
    d. @mire has some DSpace plug-ins that could prove useful to us (if we stay with DSpace):
        i. Reporting Suite – automated generation of listing and reports
        ii. Audiovisual Streaming Module

iii. Image Zoom Module
iv. Document Streaming Module