

Archive and Preservation Repositories: Activities & Directions

Tyler O. Walters
Library and Information Center
Georgia Institute of Technology

Sun PASIG, November 19, 2008

Georgia Tech Library and Information Center

"What's Going On?"

- Today's talk… Subjective, personalized view
  - "The World according to TW"

- Meant to be a summation, an overview…

  … get the Convo started…

Themes for Today – Repositories and…

1. Exchange (Harvesting & Interoperability)
2. Infrastructure
3. Synergies

Georgia Tech Library and Information Center

Exchange: Harvesting & Interoperability

- Open Archives Initiative Object Reuse and Exchange *(OAI-ORE)*

- Defines standards for description and exchange of aggregations of Web resources (these are compound objects, i.e. text, images, data, video)

- Goal = expose rich aggregated content to applications that support authoring, deposit, exchange, visualization, reuse, and preservation

- ORE is it! Metadata is important, but source content is what its all about!

Georgia Tech Library and Information Center

Exchange: Harvesting & Interoperability

- OJS/OCS to DSpace and Fedora (JISC)
  - SWORD (Simple Web-service Offering Repository Deposit)
  - APSR, ANU
- FCLA project (IMLS)
  - "Towards Interoperable Preservation Repositories"
  - Cornell, NYU Libraries
- MetaArchive: LOCKSS - iRODS work (NARA/NHPRC)
  - Integrating technologies to pass collections from PLNs to grid-based systems
  - GTRI ITTL, GT & Emory Libraries

Georgia Tech Library and Information Center

Exchange: Harvesting for Preservation

Harvesting Repositories for Distributed Preservation via Private LOCKSS Networks

- Seek enhanced DSpace export, harvester/preservation support
- Export (METS) via web interface
  - Similar to dsrun/export via the command line
  - Allow harvesting of items, collections, or communities (e.g. DSpace)
  - Directories not viewable, indexable via standard UI/bots

- Directories LOCKSS friendly, complete with auto-generated manifest page
- Suggest LOCKSS plugin settings based on export
- Expand other IR's with same functionality (Eprints, Fedora)
- Have LOCKSS support SWORD to facilitate "crash" recovery of IR

*Slide contents contributed by Chris Helms, Georgia Tech

Georgia Tech Library and Information Center

Infrastructure – Format Management

FM = identify/validate, emulate, migrate, convert

Integrate or Modularize FM tools w/ Repositories?

- E.g. Format Conversion in DSpace using Open Office
  - Tim Donohue's work @ UIUC
- @Mire: Information Conversion
  http://atmire.com/infocon.php
- Automatic Data Migration @ Sun (OpenSolaris)
- http://opensolaris.org/os/project/adm/
- http://opensolaris.org/os/project/adm/WhatisADM/;jsessionid=7CC1B5D83D54E52AC61D2310FC8A28EB

Georgia Tech Library and Information Center

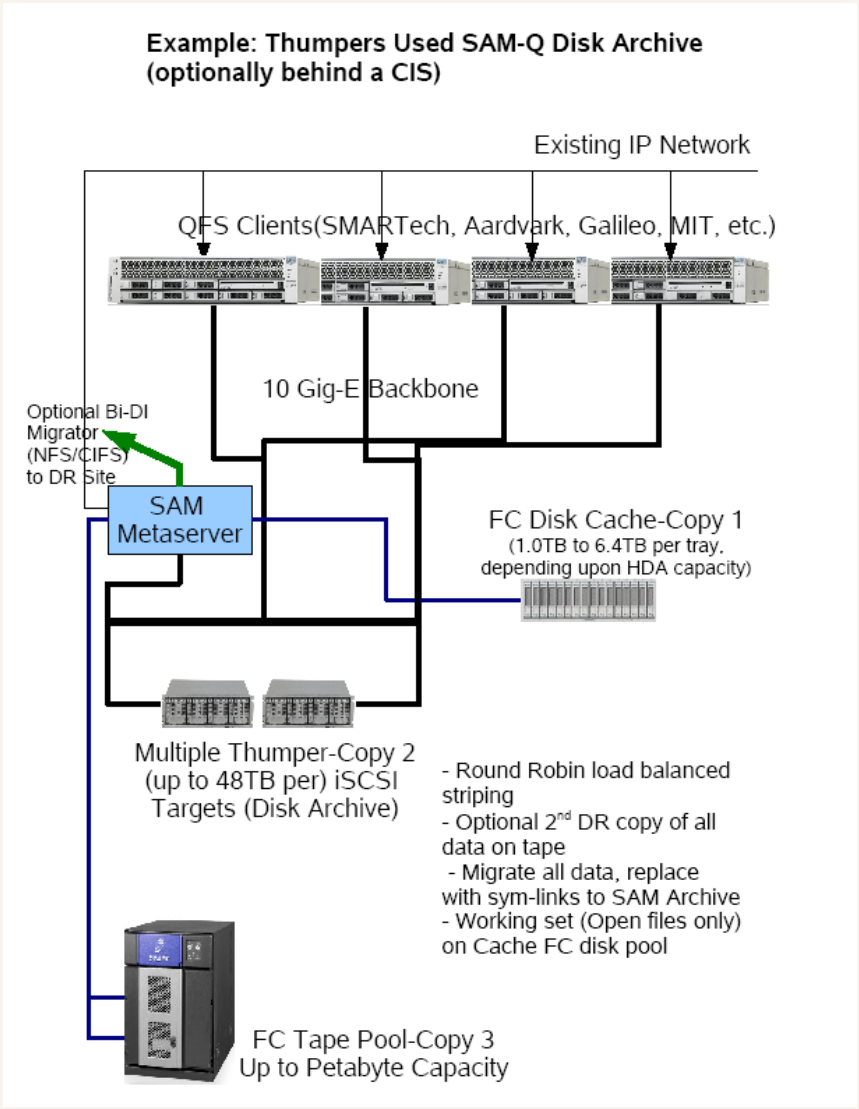
Format Management, cont'd

Identification / Validation / Registries:

- New JHOVE2 developments (Nov. 6 announcement)
  - http://confluence.ucop.edu/display/JHOVE2Info/Home
- PRONOM-ROAR (web-based preservation service uses DROID)
  - http://trac.eprints.org/projects/iar/wiki/Profile
- GDFR -- Use Cases and Requirements work in 2008
- DSpace Bitstream Format Renovation (Larry Stone/MIT)
  - Improve how DSpace bitstream registry works. Makes outside services more useful.

Georgia Tech Library and Information Center

Infrastructure: Storage - Tiered



Storage Layer:

Abstracting the storage layer from the content-organizing features of the repository

Georgia Tech Library and Information Center

Infrastructure: Cloud Computing & Storage

- Google
- Amazon (EC2 and S3)
- IBM
- Microsoft
- Higher Education
  - NSF TeraGrid
  - Biomedical Information Research Network (BIRN)
  - Open Science Grid
  - Open Grid Forum

Problem

- Understanding how the brain works is a major scientific challenge
- Globally, over 100,000 neuroscientists are working on problem
- Data that forms the basis for this work is rarely shared, difficult, and expensive to produce

Solution

- Developing scalable cloud architecture to enable data sharing, integration, and analysis supported by metadata
- Expandable range of services is provided in the cloud to extract value from data
- Promotes sharing of analysis services as well as data, and allows services to execute close to data on which they operate
- This is essential to avoid having to ship vast quantities (TBs) of data out of the cloud to the user's machine for analysis

Repositories Using The Cloud

(http://www.carmen.org.uk)

CARMEN is an e-Science Pilot Project funded by the Engineering and Physical Sciences Research Council (UK)

Core Set of Services:

1. Data repository for file and structured data
2. Metadata repository to allow users to locate & interpret data
3. Service repository with dynamic deployment onto computing resources
4. Workflow enactment engine, and security infrastructure

Synergies

Convergence, e.g. DSpace & Fedora Collaboration

"Move toward common architecture for modularization of core software; evolve to a plug-in approach where modules are sharable by both DSpace and Fedora"

Georgia Tech Library and Information Center

DSpace/Fedora Synergies

Short-Term Projects:

#1 Shared Storage Abstraction Layer

- Evaluate Akubra Project, a collaborative project of Fedora and Topaz developers. Move beyond plug-ins of file systems and local storage to accommodate cloud storage and other external providers.
- Develop stand-alone software component that can run under Fedora, DSpace, other and future systems

#2 Common Repository Exposure for the Web

- Demonstrate moving objects back and forth among DSpace and Fedora
- Technical Approach: Atom Publishing Protocol (read/update/delete); SWORD (deposit), and ORE for serialization of objects aggregations

Georgia Tech Library and Information Center

DSpace/Fedora Synergies

Short-Term Projects:

#3 Integration of Repositories with Common Authoring Tools

- Zotero: design plug-in strategy for repositories
- Technical approach: Web protocols: SWORD, Atom Publishing Protocol
- Serialization: OAI-ORE, Atom
- Consider challenges of fitting a generic approach with particular authorizing tools, both in terms of what standards these tools do or do not want to support.

DSpace/Fedora Synergies

Short-Term Projects:

#4 DSpace Running on Fedora

- Run DSpace application and workflow on top of Fedora
- High priority for Universities and Libraries interested in both systems
- Model on Google Summer of Code prototype that is underway
- Fedora Commons will participate in new DSpace 2 data model work

What's left to talk about?

- Metadata generation and management…
- User interface design tools…
- Data curation…
- Semantic technologies…
- Workflow and business processes…
- Security…
- Authenticity…
- Persistent / Alternate identifiers…
- Disaster recovery…
- Ingest…
- More…

Georgia Tech Library and Information Center

That's it… but Let's Continue…

- Tyler Walters

- 404-385-4489 voice
- Tyler@gatech.edu email
- TyWalters1 - AIM/Skype/Gmail/Facebook

Georgia Tech Library and Information Center