

Enterprise Repositories and Federated Archives:

The MetaArchive Approach to Distributed Digital Preservation

Tyler O. Walters

Associate Director, Technology & Resource Services

Georgia Tech Library & Information Center

Georgia Tech Library and Information Center

Sun Microsystems PASIG

San Francisco, CA -- May 2008

Briefing on MetaArchive Cooperative

http://www.metaarchive.org

- **Project Summary:**
  - Eight partner institutions:
    - Emory - Georgia Tech - Florida State – Library of Congress
    - Virginia Tech – Auburn – Louisville – Univ. of Hull (UK)
  - Collaborate w/ LC/NDIIPP – $1.2M initial effort to develop cooperative for preservation of digital content, 2004-2009
- **Goals:**
  1. Distributed preservation network infrastructure (LOCKSS)
  2. Conspectus of digital content held by the partner sites
  3. Harvest a body of most critical content to be preserved (4 TB)
  4. Cooperative charter model for collaboration and sustainability

Networks: 1) Southern Culture, 2) ETDs, 3) Trans-Atlantic Slave Trade

Georgia Tech Library and Information Center

Preservation Network Design Precepts

1. Distributed Preservation Infrastructure
2. Peer-to-Peer Network Architecture
   - Each node communicates with all other nodes
   - All nodes are joint custodians
3. Flexible Organizational Model
4. Formal Content Selection Process
5. Capability for Migrating Archives
6. Dark Archiving Strategy (no public access to MA network content)
7. Low Cost to Deployment
8. Self-Sustaining Incentives

Effective digital preservation succeeds by distributing copies of content in secure, distributed locations over time

Georgia Tech Library and Information Center

Advantages of Adapting LOCKSS Software

- **Supports "distributed digital replication" approach:**
  - Data integrity checks
  - Rigorous security checks
  - Focused web crawls to gather/ingest content
- **Advantage:** many preservation efforts mix high accessibility online with long-term access (preservation). High accessibility = high costs
- Network content discoverable via metadata search. Conspectus will link to live sites designed for access
- Originally designed for minimal expenditures
  - Low barriers to adoption
  - Inexpensive computers / modest systems administration

Georgia Tech Library and Information Center

University of Louisville

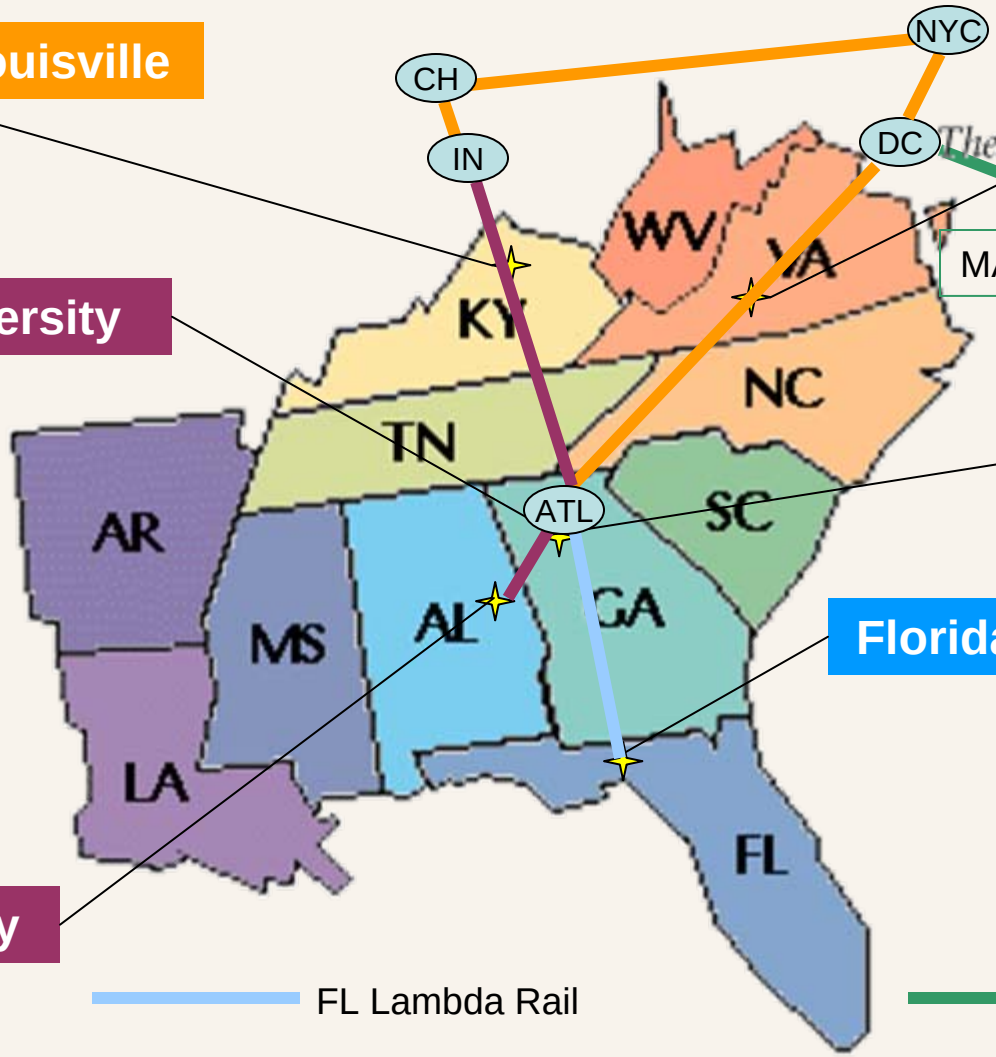
Va Tech

Emory University

Ga Tech

Florida State University

Auburn University



The LIBRARY of CONGRESS

MAX Connection to Va Tech

FL Lambda Rail

MAX Network

Abilene Network

SOX Network

Georgia Tech Library and Information Center

Current MetaArchive Technology

Server:

- Dell PowerEdge 1850
- 2x 3.0Ghz/1MB Cache, Xeon 800Mhz / 2Gb Memory

SAN: (could easily be Sun or other hardware)

- PetaBox PowerStore PS4000
- AMD Athlon 64 X2 Dual Core Processor 4600+
- 2GB Memory / 4x1TB SATA HD

- Dell/EMC AX100 Array (single processor)
- 2 TB Storage (12x250 7200 RPM Serial ATA)

OS and LOCKSS:

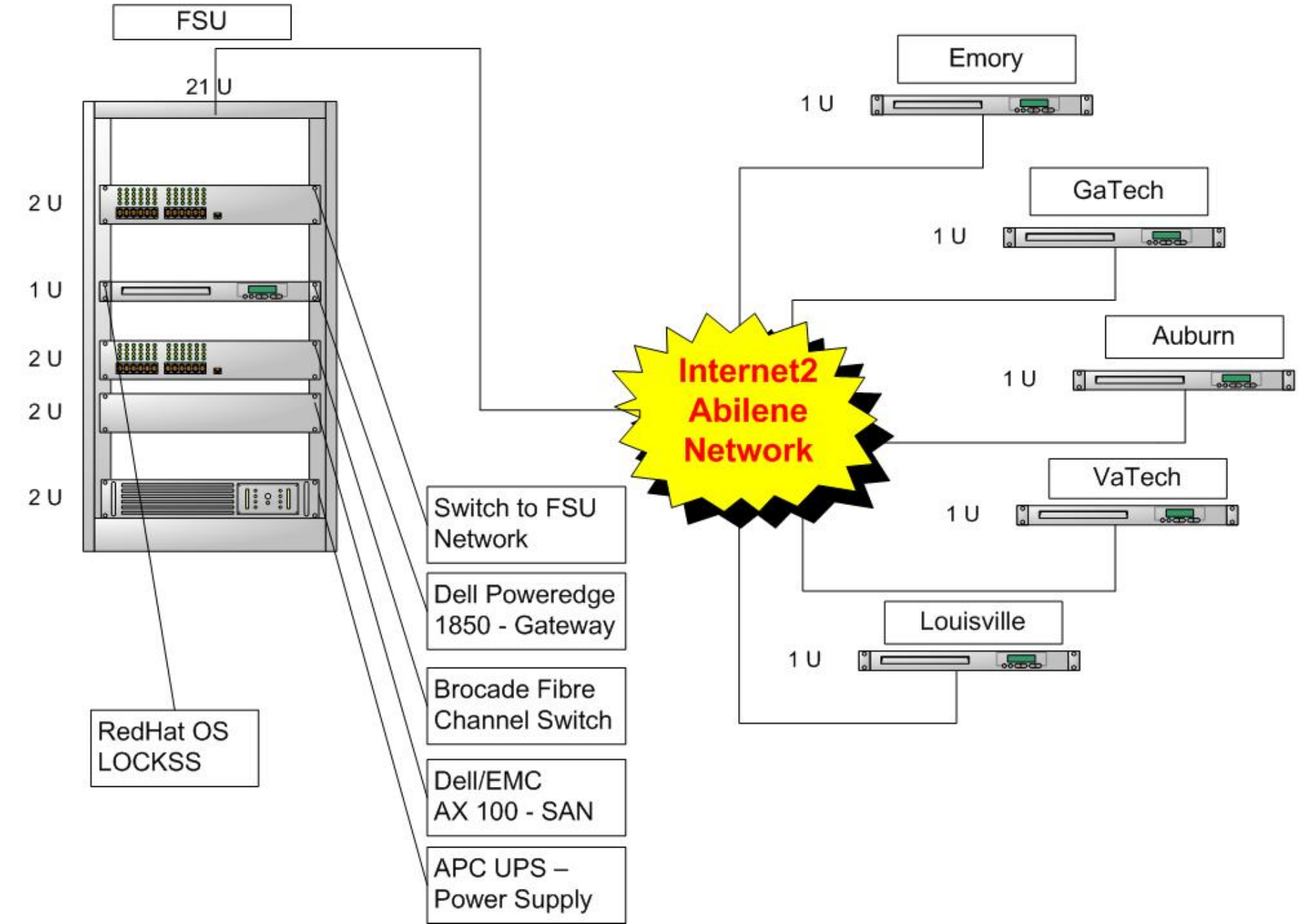
- Red Hat Advanced Server 4 (release 6) / (Fedora Core 8 – new node)
- Hw/Sw firewalls, access control lists / LOCKSS daemon 1.29.4

Database Driven Conspectus

- MySQL/PHP Interface – Integrated w/LOCKSS Plugin Directory
- Manages Collections within LOCKSS
- Network Administrative Management Tool

Georgia Tech Library and Information Center

MetaArchive Preservation Network



Collections Replication

Online Digital Collections

Auburn Yearbooks

Emory Southern Spaces

FSU ETD Collection

LOCKSS for MetaArchive

Emory Node

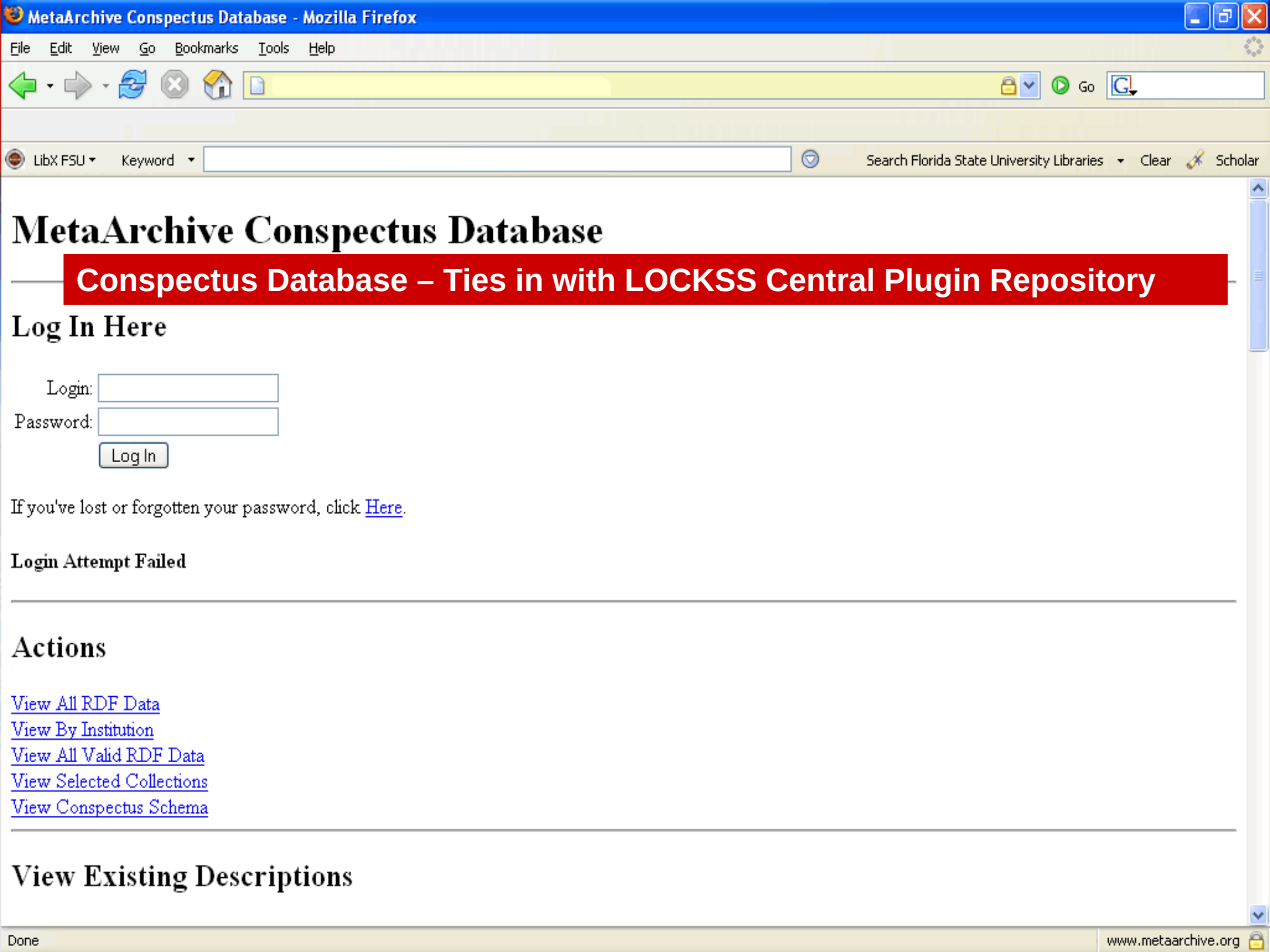
Auburn Node

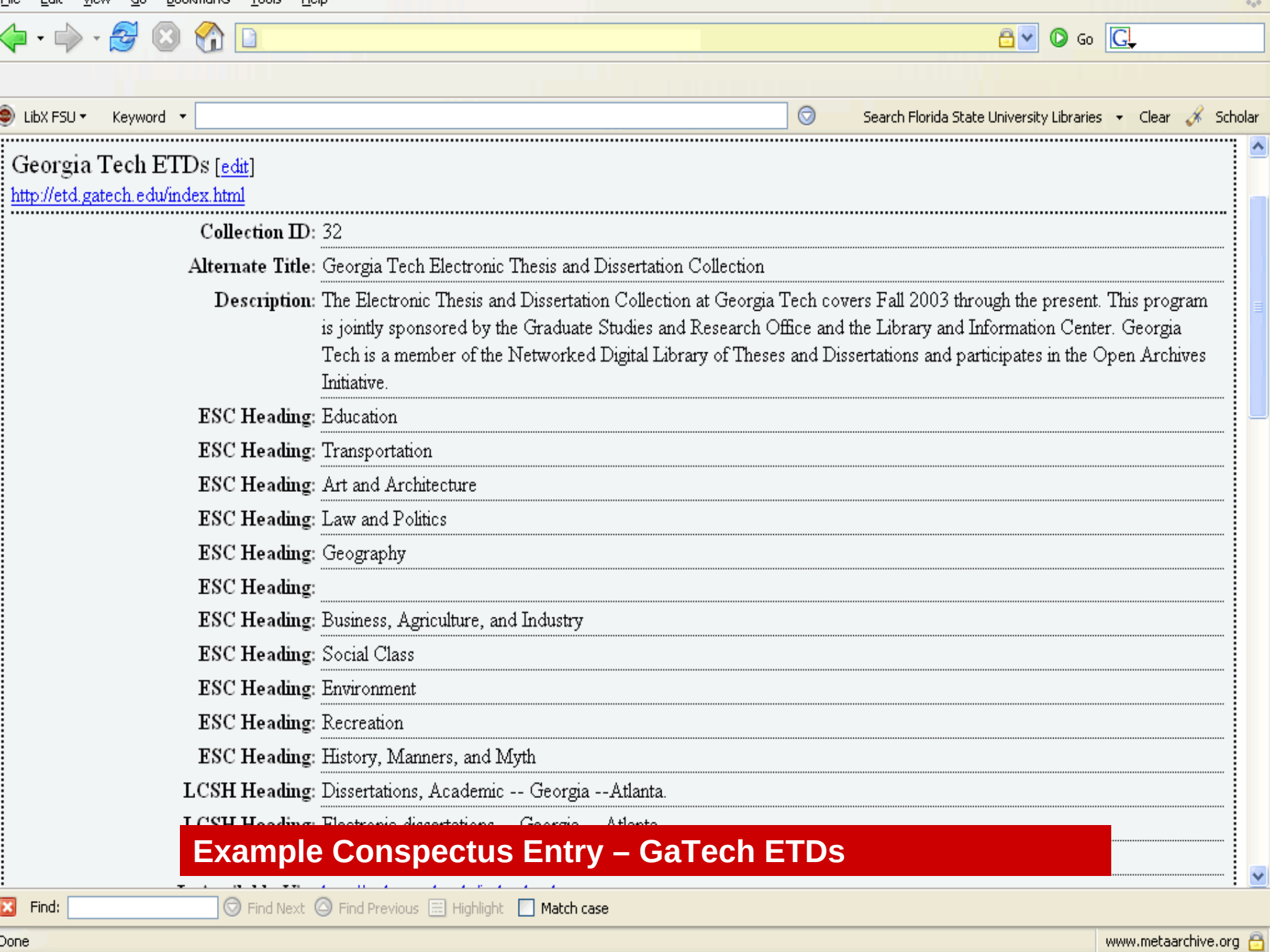
GaTech Node

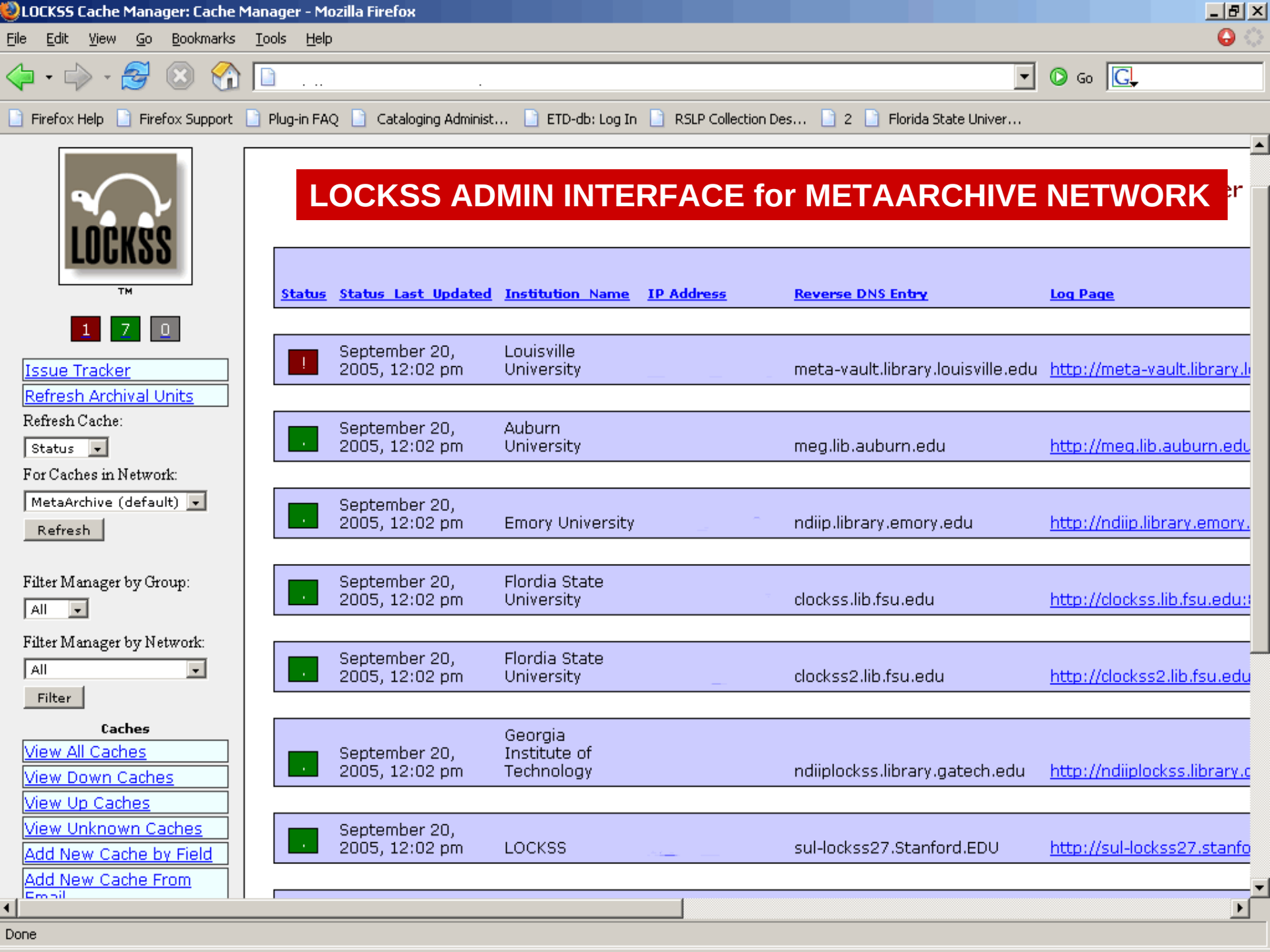
FSU Node

Admin Interface

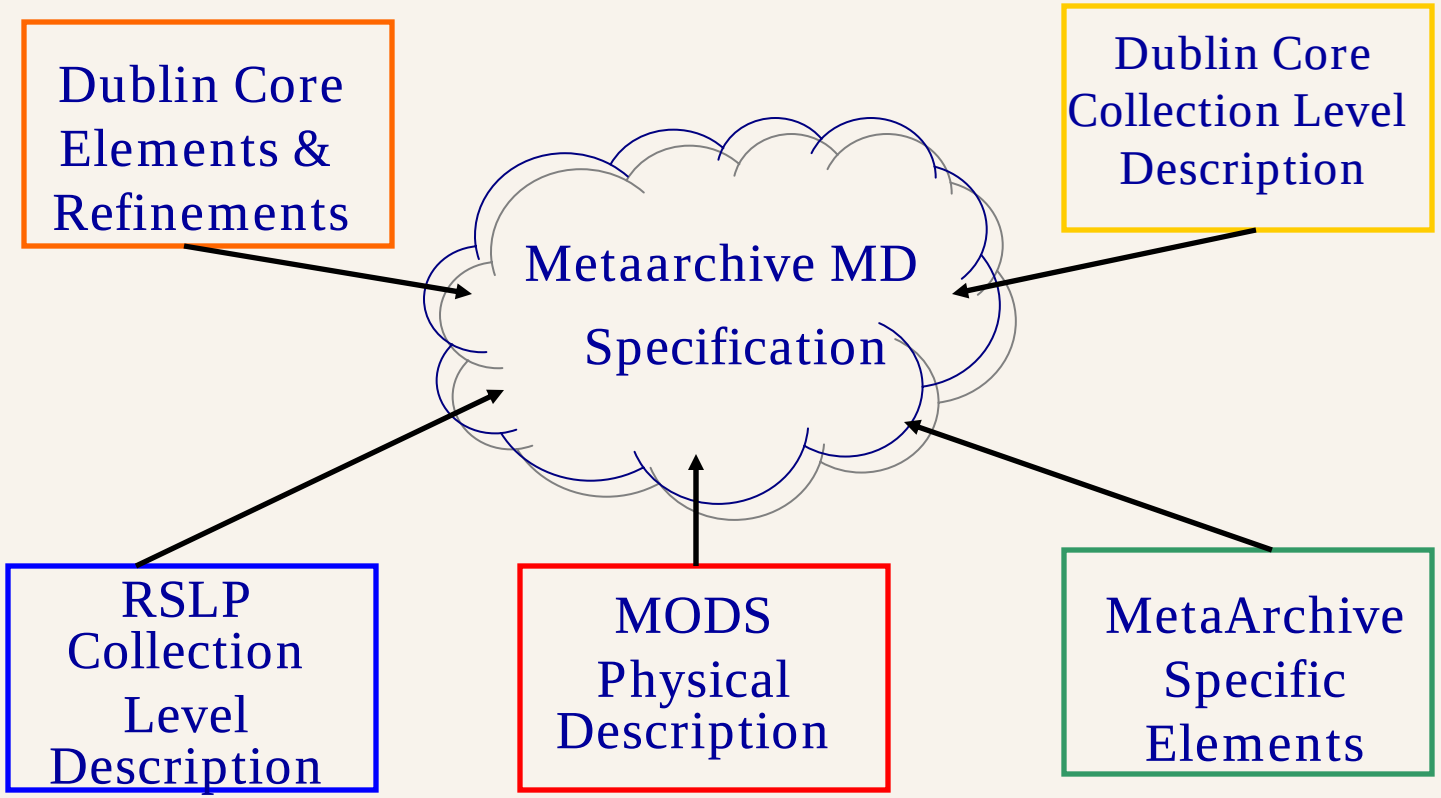
Georgia Tech Library and Information Center


MetaArchive Conspectus Database - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
LibX FSU
Keyword
Search Florida State University Libraries
Clear
Scholar
MetaArchive Conspectus Database
Conspectus Database – Ties in with LOCKSS Central Plugin Repository
Log In Here
Login:
Password:
Log In
If you've lost or forgotten your password, click Here.
Login Attempt Failed
Actions
View All RDF Data
View By Institution
View All Valid RDF Data
View Selected Collections
View Conspectus Schema
View Existing Descriptions
Done
www.metaarchive.org


Georgia Tech ETDs [edit]
http://etd.gatech.edu/index.html
Collection ID: 32
Alternate Title: Georgia Tech Electronic Thesis and Dissertation Collection
Description: The Electronic Thesis and Dissertation Collection at Georgia Tech covers Fall 2003 through the present. This program is jointly sponsored by the Graduate Studies and Research Office and the Library and Information Center. Georgia Tech is a member of the Networked Digital Library of Theses and Dissertations and participates in the Open Archives Initiative.
ESC Heading: Education
ESC Heading: Transportation
ESC Heading: Art and Architecture
ESC Heading: Law and Politics
ESC Heading: Geography
ESC Heading:
ESC Heading: Business, Agriculture, and Industry
ESC Heading: Social Class
ESC Heading: Environment
ESC Heading: Recreation
ESC Heading: History, Manners, and Myth
LCSH Heading: Dissertations, Academic -- Georgia --Atlanta.
Example Conspectus Entry – GaTech ETDs


LOCKSS Cache Manager: Cache Manager - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
Firefox Help Firefox Support Plug-in FAQ Cataloging Administ... ETD-db: Log In RSLP Collection Des... 2 Florida State Univer...
LOCKSS
TM
1 7 0
Issue Tracker
Refresh Archival Units
Refresh Cache:
Status
For Caches in Network:
MetaArchive (default)
Refresh
Filter Manager by Group:
All
Filter Manager by Network:
All
Filter
Caches
View All Caches
View Down Caches
View Up Caches
View Unknown Caches
Add New Cache by Field
Add New Cache From
LOCKSS ADMIN INTERFACE for METAARCHIVE NETWORK
Status Status Last Updated Institution Name IP Address Reverse DNS Entry Log Page
! September 20, 2005, 12:02 pm Louisville University meta-vault.library.louisville.edu http://meta-vault.library.l
September 20, 2005, 12:02 pm Auburn University meg.lib.auburn.edu http://meg.lib.auburn.edu
September 20, 2005, 12:02 pm Emory University ndiip.library.emory.edu http://ndiip.library.emory.
September 20, 2005, 12:02 pm Flordia State University clockss.lib.fsu.edu http://clockss.lib.fsu.edu:
September 20, 2005, 12:02 pm Flordia State University clockss2.lib.fsu.edu http://clockss2.lib.fsu.edu
September 20, 2005, 12:02 pm Georgia Institute of Technology ndiiplockss.library.gatech.edu http://ndiiplockss.library.c
September 20, 2005, 12:02 pm LOCKSS sul-lockss27.Stanford.EDU http://sul-lockss27.stanfo
Done

Metadata Specification



Georgia Tech Library and Information Center

Conclusion / Thank You!

- www.metaarchive.org
- www.digitalpreservation.org
- www.lockss.org

- Tyler Walters
- Tyler@gatech.edu
- 404-385-4489

Georgia Tech Library and Information Center