BUILDING A COLLABORATIVE DIGITAL PRESERVATION NETWORK

Caroline Arms
Office of Strategic Initiatives, Library of Congress

Robert H. McDonald
Associate Director of Libraries for Technology & Research
Florida State University

Lizabeth B. Nicol
Digital Library Coordinator
Auburn University

Tyler O. Walters
Associate Director for Technology and Resource Services
Georgia Institute of Technology

OVERVIEW

- Introduction to the NDIIPP Partnerships
- The MetaArchive Partnership
  - Auburn University
  - Emory University
  - Florida State University
  - Georgia Tech
  - University of Louisville
  - Virginia Tech
- The MetaArchive Metadata Strategy
- The MetaArchive Technical Architecture

INTRODUCTION TO NDIIPP

- Federal legislation in December 2000
- LC to work with public and private sector to support preservation of significant "born-digital" content that is at risk
- $25 million + another $75 million if matched, for potential total of $175 million.
- Started with planning period
  - consultation with stakeholder groups
  - commissioned surveys and reports
  - plan approved December 2002

OVERALL NDIIPP GOALS

- Help identify and preserve at-risk digital content
- Support development of improved tools, models, and methods for digital preservation
- Work with industry, concerned federal agencies, libraries, research institutions and not-for-profit entities
- Develop a national digital collection and preservation strategy

NDIIPP PORTFOLIO

- Plan calls for LC's actions to be:
  - Catalytic, collaborative, iterative, strategic
- General approach
  - Find smart, willing collaborators. Learn by doing.
- Three areas of focus
  - Network of preservation partners
    - Digital Preservation Partnerships
    - Working with state libraries, archives, CTOs, etc.
  - Architectural framework for preservation
  - Digital preservation research
    - Funding DIGARCH program through NSF

DIGITAL PRESERVATION PARTNERSHIPS

- Competition, cooperative agreements
  - 8 awards announced in September 2004
  - Partners collect/preserve content, collaborate with LC and each other
  - 3 year term, LC to report to Congress
- Primary outcomes for partnerships:
  - Identify and preserve significant at-risk content
  - Leverage resources & experience via collaboration
  - Promote standards and best practices
  - Learn how to build and sustain partnerships

PARTNERSHIPS DIFFER

- In content scope
  - Public television programs (high-definition digital)
  - Dot-com era business records
  - Social science datasets
  - Geospatial information (2 projects)
  - Heterogeneous content
    - harvested from web
    - for which partners are already responsible
- In nature of partnership
  - Partners playing different roles
  - Group of peers

ACTIVITY ACROSS PROJECTS

- LC is providing resources and leadership
  - Individual LC staff as liaison to each project
  - Meetings twice a year
- 'Affinity groups' on cross-cutting issues
  - Selection and Collection – appraisal & tools
  - Rights – copyright and privacy
  - Technical Architecture
  - Economic Sustainability - costs and incentives
- Connections to other NDIIPP activities

METAARCHIVE PARTNERSHIP

Project Summary:

- Six partner institutions:
  - Emory - Georgia Tech - Florida State
  - Virginia Tech – Auburn - Louisville
- Collaborate with LoC – 3-yr $1.4M effort to develop a cooperative for preservation of digital content.
- Content focus is s*outhern culture and history*.

MetaArchive Project Goals

- Create a conspectus of digital content within the subject domain held by the partner sites
- Harvest a body of most critical content to be preserved (3 terabytes, w/ capability to expand)
- Develop a model cooperative agreement for ongoing collaboration and sustainability
- Distributed preservation network infrastructure based on the LOCKSS software

Governance & Structure

- Committees:
  - **Steering**: coordination, communication, reporting (original six univs.)
  - **Content**: organize, develop, select content
  - **Preservation**: content retention/transfer, acquisition practices, metadata maintenance, text/image structures, migratability
  - **Technical**: server architecture, software development

Governance & Structure

- Membership Type:
  - **Development partner**:
    Testing and development of hardware, software, networking, and design of Network features. Carry out activities of preservation partner sites as well.
  - **Preservation partner**:
    Network participation -- maintain a node, ingest collections from partners or content contributors. Network development is optional.

Cooperative Agreement

- Develop a simple, flexible agreement as a model for other institutions seeking to cooperate in digital preservation
  - Membership criteria (and member withdrawal)
  - Roles and responsibilities – joint and equal custodians of content harvested
  - Sustainability plan (over time)
  - Ensure broad applicability

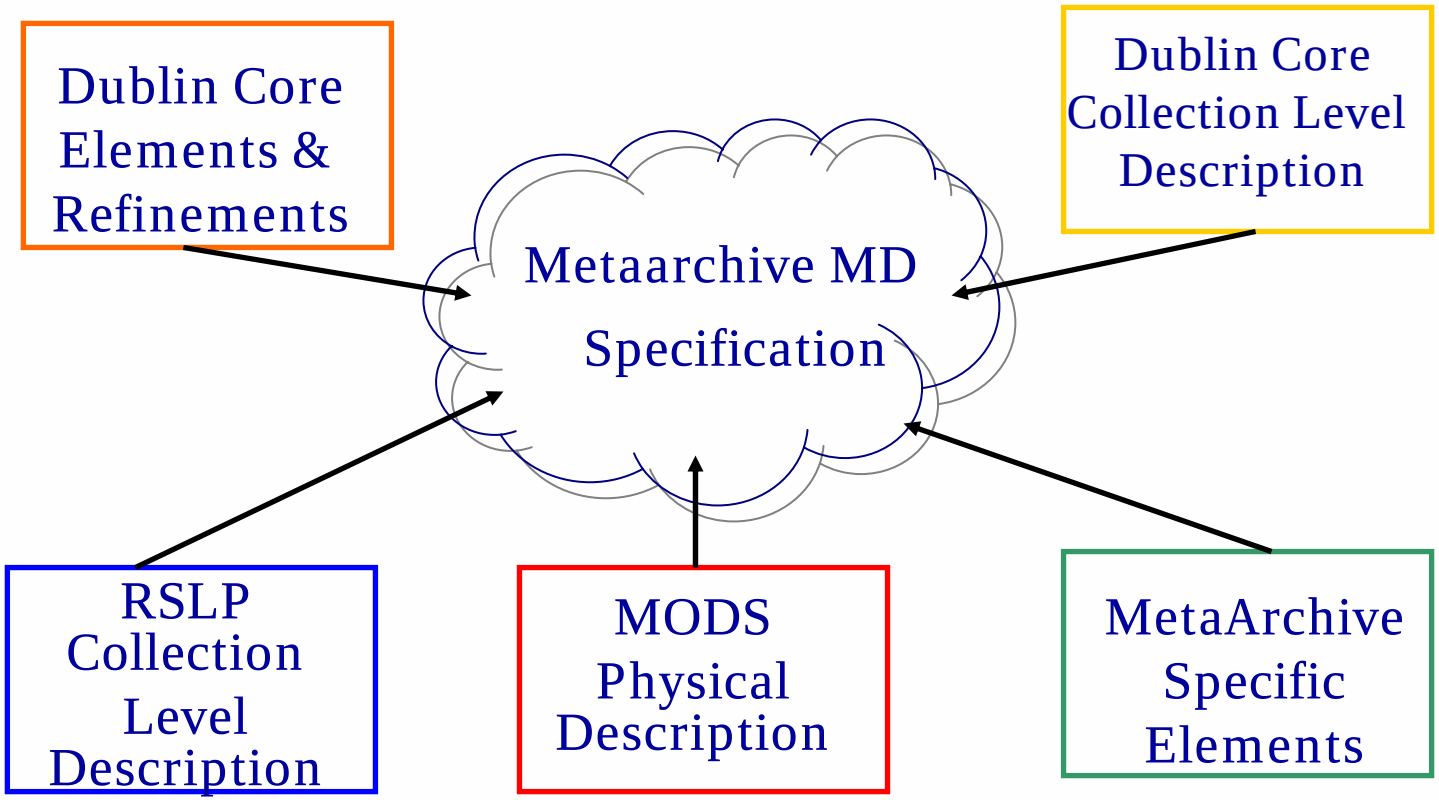
Cooperative Agreement

- Issues to Address:
  - New members: by invite only? by application?
  - 3rd member type: content contributor?
  - LOCKSS Alliance membership and fees
  - Central administration vs. decentralized
  - Financial sustainability (need central funds?)
  - Memo of agreement between institutions – detailing what members will do

METADATA OVERVIEW

- The MetaArchive Conspectus Database contains metadata elements that not only describe the collections that are to be collected, but also provide information that will be necessary for storage estimates, format migration, accrual rules, location, ownership and LOCKSS specific elements.
- The Conspectus Database is archived along with the collections.

GENESIS OF MD SPECIFICATION



METADATA SCOPE

- Intellectual content of the collection(s) including subjects, spatial and temporal coverage
- Format of contained items - and extent of file sizes and formats
- Relation to other collections
  - Accrual rules (periodicity, open/closed)
  - Rights management rules
  - LOCKSS manifest pages and plugin information
  - Risk assessment

METADATA ELEMENTS

- Multiple name spaces utilized:
  - Dublin Core Elements
  - Dublin Core Refinements
  - Collection Level Description
    - RSLP (Research Support Libraries Programme)
    - MODS (Metadata Object Description Schema)
    - MetaArchive defined terms
- MetaArchive Metadata Specification
  - http://metaarchive.org/pdfs/conspectus_md_2005.html

COLLECTION LEVEL DESCRIPTION

- DC Collection Description Application Profile
  - Accrual Periodicity [cld:accrualPeriodicity]
  - Accrual Policy [cld:accrualPolicy]
  - Contents Date Range [cld:dateContentsCreated]
  - Is Available Via [cld_gen:isAvailableAt]
  - Spatial Coverage [cld:spatial]
  - Temporal Coverage [cld:temporal]
- MODS
  - Manifestation [mods:physicalDescription] (1/3 of element definition)
- RSLP
  - Accumulation Date Range [rslp:created]

METAARCHIVE SPECIFIC

- Cataloged Status [ma:catalogedstatus]
- LOCKSS Manifest page [ma:manifest]
- MetaArchive Collect. Identifier [ma:collectionid]
- OAI Provider [ma:oaiprovider]
- Recommended Harvest Proc. [ma:harvestproc]
- Risk Rank [ma:riskrank]

TECHNICAL ARCHITECTURE

- Off-the-Shelf Strategy
  - Dell/Intel Based Hardware
    - Could easily be HP or SUN Intel Based Hardware etc.
    - Could be old desktops w/large hard drives.
  - New Low Cost SATA SAN
    - EMC AX100
      - $4.00 per GB (already dropping in price)

TECHNICAL ARCHITECTURE

- Operating System
  - RedHat Linux Enterprise AS v. 3/4
    - Ease of update management and experience w/OS
      - Could easily work on other versions of Linux
    - JAVA SDK
- LOCKSS Content Ingestion/Replication
  - LOCKSS Daemon 1.8.3 – 6-8 week updates w/RPM
- Conspectus Database
  - MySQL/PHP Interface – Integrated w/LOCKSS Plugin Directory
- MetaArchive Collection Description Metadata Schema

TECHNICAL ARCHITECTURE

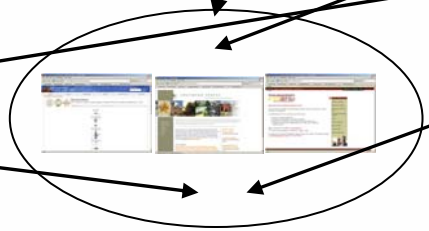
Online Digital Collections

Auburn Yearbooks

Emory Southern Spaces

FSU ETD Collection

LOCKSS for MetaArchive



Emory Node

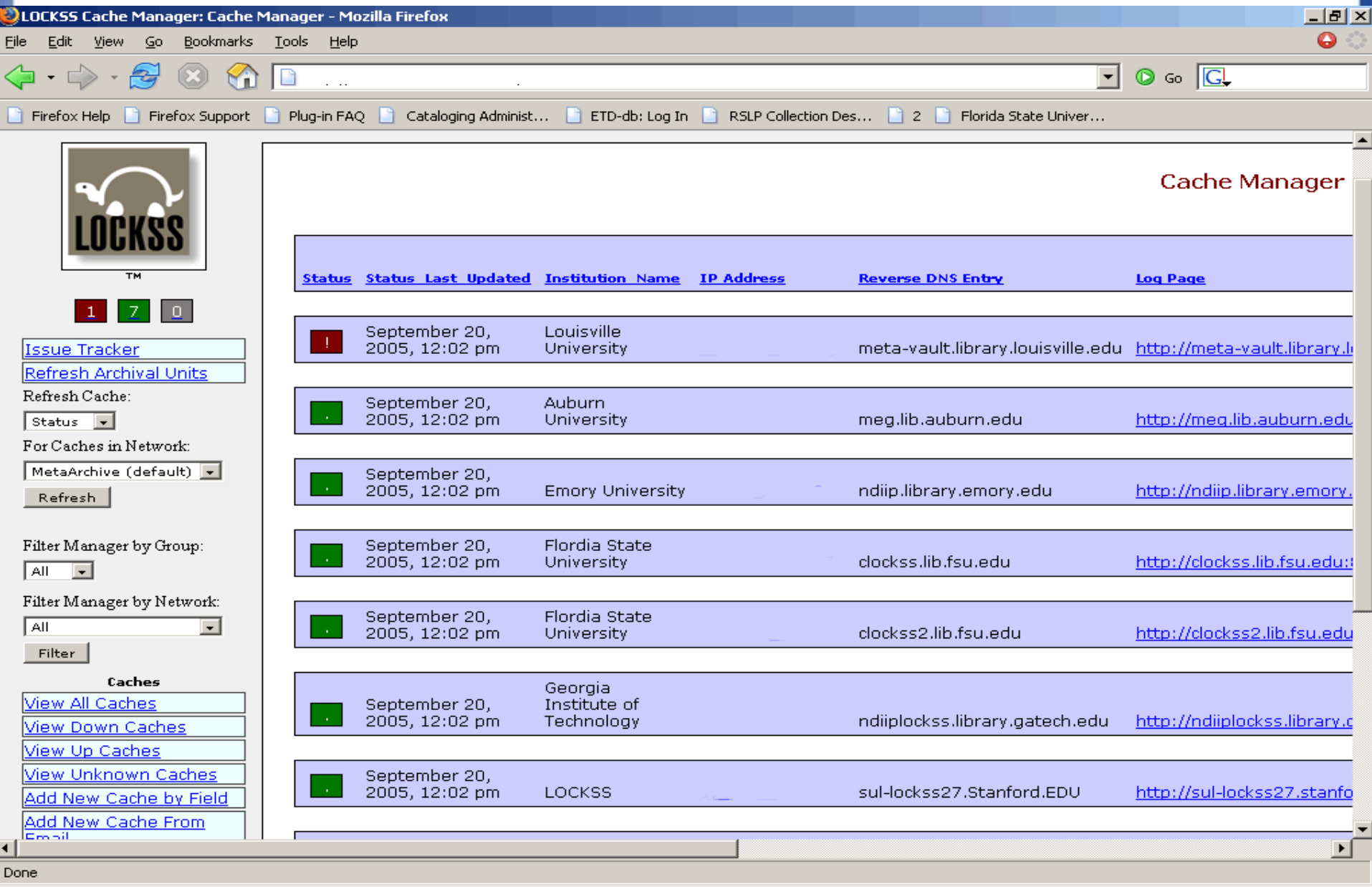
Auburn Node

GaTech Node

FSU Node

Admin Interface

LOCKSS ADMIN INTERFACE



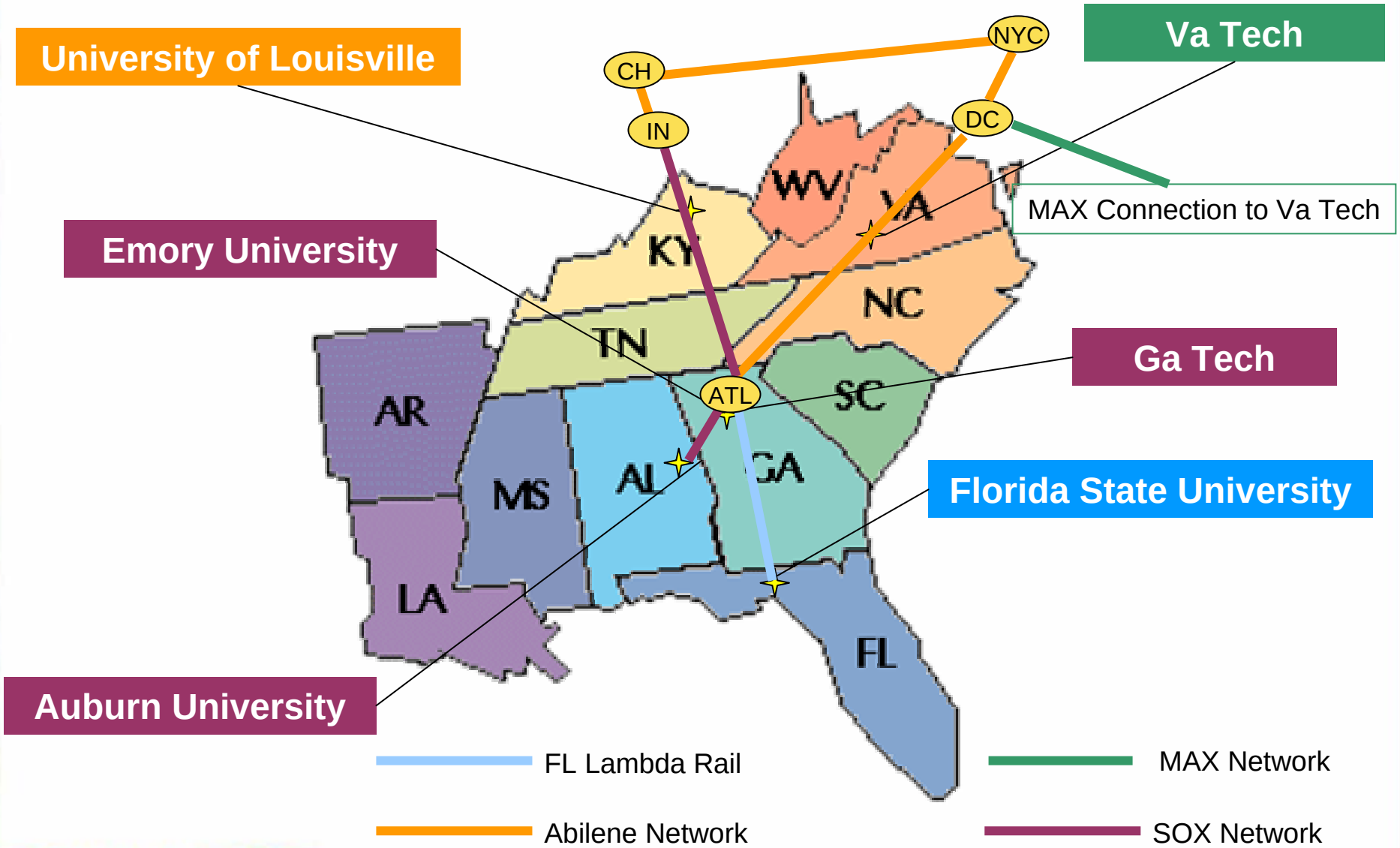
TECHNICAL ARCHITECTURE

- STANDARDS
  - OAIS Reference Model
    - LOCKSS Compliance
      - See http://arxiv.org/abs/cs.DL/0509018
  - OAI-PMH 2.0 (Submission Information Package)
    - Using as alternative to current LOCKSS AU strategy w/ETDs – VaTech, GaTech, FSU
  - MetaData
    - Based on Known Collection Level Namespaces
      - http://www.metaarchive.org/pdfs/conspectus_md_2005.html

TECHNICAL ARCHITECTURE

- COLLABORATION
  - Kickstart Installations for Linux Servers
    - Easy to setup all hardware exactly the same.
  - Efficiency of Replication
    - Kickstart can be used with production system as well as with any Intel based machine.
    - Currently running several test machines (old desktops) to trigger test LOCKSS quorums.
  - Communication Strategies
    - Phone Conference, Video Conference I2 Commons, Wiki (MoinMoin), PhpCollab, iVocalize Chat/VOIP Room

MetaArchive NDIIPP Network via I2



QUESTIONS

- MetaArchive Web
  - http://www.metaarchive.org
- NDIIPP Web
  - http://www.digitalpreservation.gov
- Contacts
  - Caroline Arms – caar@loc.gov
  - Robert H. McDonald – rmcdonal@mailer.fsu.edu
  - Lizabeth B. Nicol – nicollb@auburn.edu
  - Tyler O. Walters - tyler.walters@library.gatech.edu