

New Developments in Private LOCKSS Networks for Digital Preservation

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

Sun PASIG - Malta, June 24, 2009

Georgia Tech Library and Information Center

New Developments:

1) The Proliferation of PLNs (7 in 5 years)
2) Cloud Computing Usage
3) New LOCKSS Cache Manager (consider PLNs)
4) LOCKSS – Grid Data Exchange
5) PLN to PLN Data Exchange
6) PLN Harvesting of DSpace

Current PLNs:

PeDALS Persistent Digital Archives & Library SYSTEM

A project of Arizona State Library, Archives and Public Records, in collaboration with State Library and Archives of Florida, New York State Archives, New York State Library, South Carolina Department of Archives and History and the Wisconsin Historical Society

Data-PASS
DATA PRESERVATION ALLIANCE for the SOCIAL SCIENCES

CLOCKSS

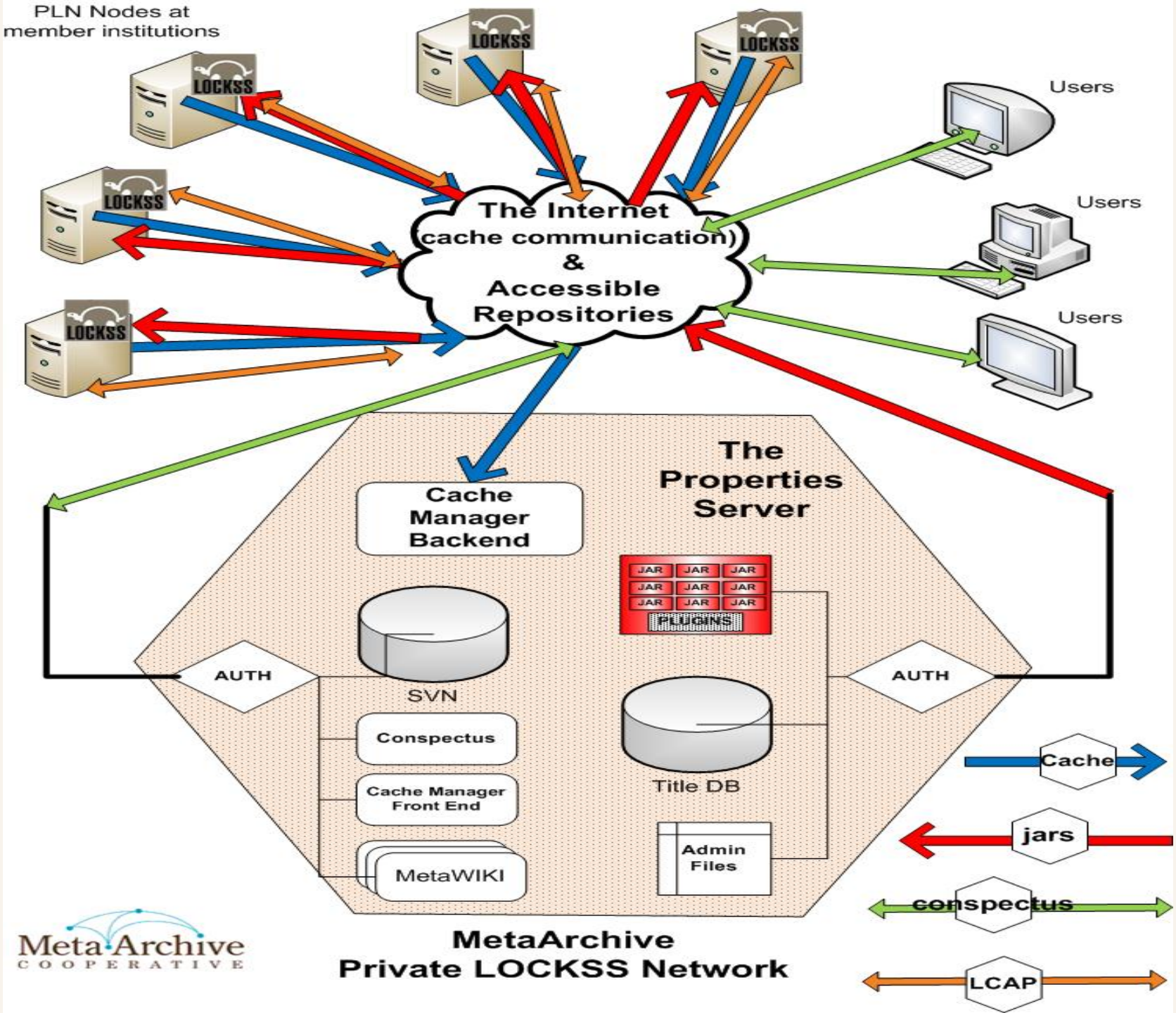
COPPUL

US Gov Docs

THE ALABAMA DIGITAL PRESERVATION NETWORK
PRESERVING ALABAMA'S DIGITAL RESOURCES

MetaArchive COOPERATIVE

Georgia Tech Library and Information Center


PLN Nodes at member institutions
LOCKSS
LOCKSS
LOCKSS
LOCKSS
LOCKSS
Users
Users
Users
The Internet
cache communication)
&
Accessible
Repositories
The Properties Server
Cache Manager Backend
JAR
JAR
JAR
JAR
JAR
JAR
JAR
JAR
JAR
Fetchers
AUTH
SVN
AUTH
Conspectus
Cache Manager Front End
MetaWIKI
Title DB
Admin Files
Cache
jars
conspectus
LCAP
Meta Archive
COOPERATIVE
MetaArchive
Private LOCKSS Network

MetaArchive Core Properties – System Architecture

- Conspectus DB
- Cache Manager
- LOCKSS Title DB
- Plugin repository
- SVN repository for plugins
- MetaArchive public web site
- MetaArchive internal wiki
- MetaArchive LOCKSS software
- Keystore for plugins
- Admin files & scripts
- TRAC ticketing system

Cloud Computing Services
(Amazon EC2)

The MetaArchive Properties Server:

(core systems of LOCKSS network to functional properly)

- Don't keep this in the infrastructure of any one member

- Lease hardware, physical space, power, storage, network bandwidth all in one package
- Move to independent location / shared ownership, access
- Immediate access to tools needed to prepare new content for ingest

Georgia Tech Library and Information Center

New LOCKSS Cache Manager

- Jointly developed by LOCKSS & MetaArchive Cooperative
- Cache Manager = Network monitoring tool
- Four main components to the Cache Manager:
  - Caches
  - Collections
  - Archival Units
  - Disks
- Sample desired improvements:
  - Sort AUs by institution (not a normal LOCKSS feature)
  - Improving collection/AU name displays
  - Reporting features by Archive, by Institution
- Tools for a private LOCKSS network, not the general/public network

Georgia Tech Library and Information Center

Exchange between PLNs and Grids

BagIt

- A format for packaging and transferring digital content while facilitating inventory validation

1. There is no software to install
2. Consists of base directory with manifest file & subdirectory w/ content
3. Manifest file has a row for each content file with:
    1. Full path in content directory
    2. A checksum for file

Holey BagIt

1. Has additional 'fetch.txt' file in base directory & empty content directory
2. URLs for each content file are listed in fetch.txt file.
3. Can reduce transfer time by fetching content in parallel

http://www.digitalpreservation.gov/library/resources/tools/docs/bagitspec.pdf

Chronopolis BagIt Ingest

- Transfer to SDSC SRB resource; use parallel threads if holey bag
- Validate content against manifest; md5checksum
- Register in SDSC SRB
- QAQC data in SRB
- Replicate to NCAR, UMIACS SRB zones using Replication Monitor https://wiki.umiacs.umd.edu/adapt/index.php/Replication:Replication_Monitor_2.0
- Monitor data integrity using ACE at SDSC, NCAR, and UMIACS https://wiki.umiacs.umd.edu/adapt/index.php/Ace:Main

Metadata / Content Exchange between PLNs

Technical Information Representation Schema

(In XML, common to multiple PLNs)

- Combine parts of MetaArchive & DataPASS schemas to start, with elements of LOCKSS title DB
- Data description of whole network / preservation system
  - Replications expected
  - Characterization of machines / network parameters
- Need to capture status information vs. policy info
- **In sum, need a PLN technical metadata set (together with descriptive metadata) – could be put together in a METS package (i.e. "core operational data")**

Georgia Tech Library and Information Center

PLN Harvesting of Repository Systems – DSpace

- LOCKSS is designed to harvest static web sites
- Harvesting database-driven, dynamic content is different

Objectives:

- Harvest directly from DSpace by LOCKSS (not an intermediary site)
- Harvest content and associated metadata
- Reliable method for reconstructing repository from harvested data

Potential Solutions:

- Outside product: UIUC's Hub and Spoke or CDL's 7Train
- A custom application, interact with LOCKSS deamon & DSpace APIs
- **DSpace's native OAI and METS export functionality**

Georgia Tech Library and Information Center

PLN Harvesting of Repository Systems – DSpace

Methods:

- Get data from DSpace in format that LOCKSS daemon understands
- OAI requests for METS XML works / MTS delivers URIs for associated content as well as metadata
- Teach LOCKSS daemon to understand the data delivered, crawl rep.
- Design flexible AU structure, work with OAI/METS delivery system
- Design mechanism to restructure repository from LOCKSS data
  - Future: LOCKSS support SWORD to facilitate "crash" recovery of IR

Georgia Tech Library and Information Center

PLN Harvesting of Repository Systems – DSpace

Progress:

- DSpace delivers METS XML well
- Generic LOCKSS plugin reads DSpace OAI/METS output
  - Works in our test region / not used with production systems yet
  - Might need custom plug in Java – haven't determined yet
- Experimenting currently with new, flexible AU structures for growth

Unresolved:

- Getting LOCKSS daemon to parse METS XML (in progress)
- Design a recovery plan (in early stages)
- Dealing with "odd" file types, streaming media, datasets, binary data
- LOCKSS recrawl rates to not stress the IR

Georgia Tech Library and Information Center

Thank you…

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

Georgia Tech Library and Information Center