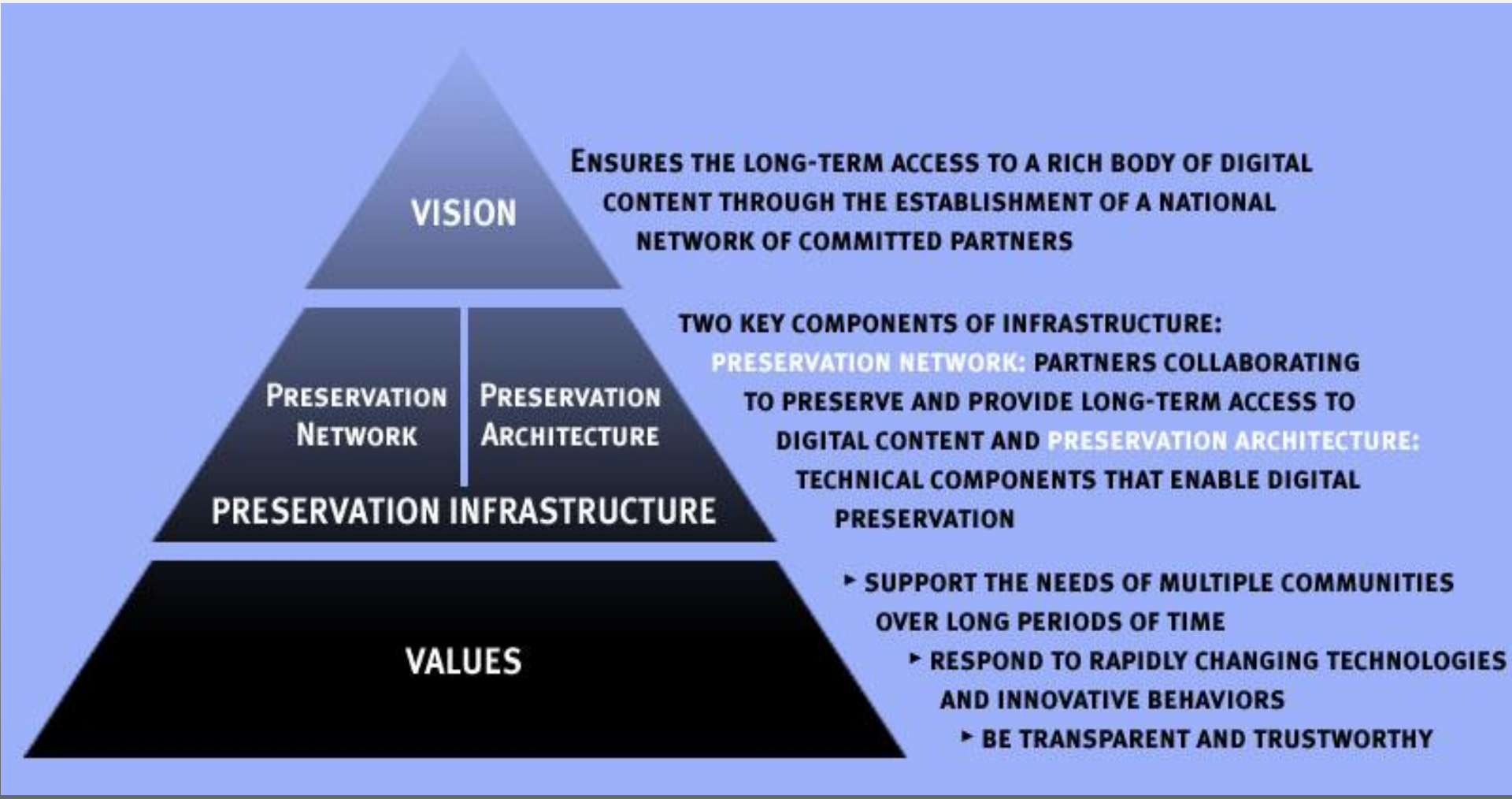
The National Digital Information Infrastructure and Preservation Program: Initiatives in Cloud and Distributed Computing

Leslie Johnston
Manager of Technical Architecture Initiatives, NDIIPP
Office of Strategic Initiatives
Library of Congress

NDIIPP: A Network of People, An Architecture for Preservation



To ensure access over time to a rich body of digital content through the establishment of a national network of partners committed to selecting, collecting and preserving at-risk digital information

Learn by doing

Areas of Work and Focus

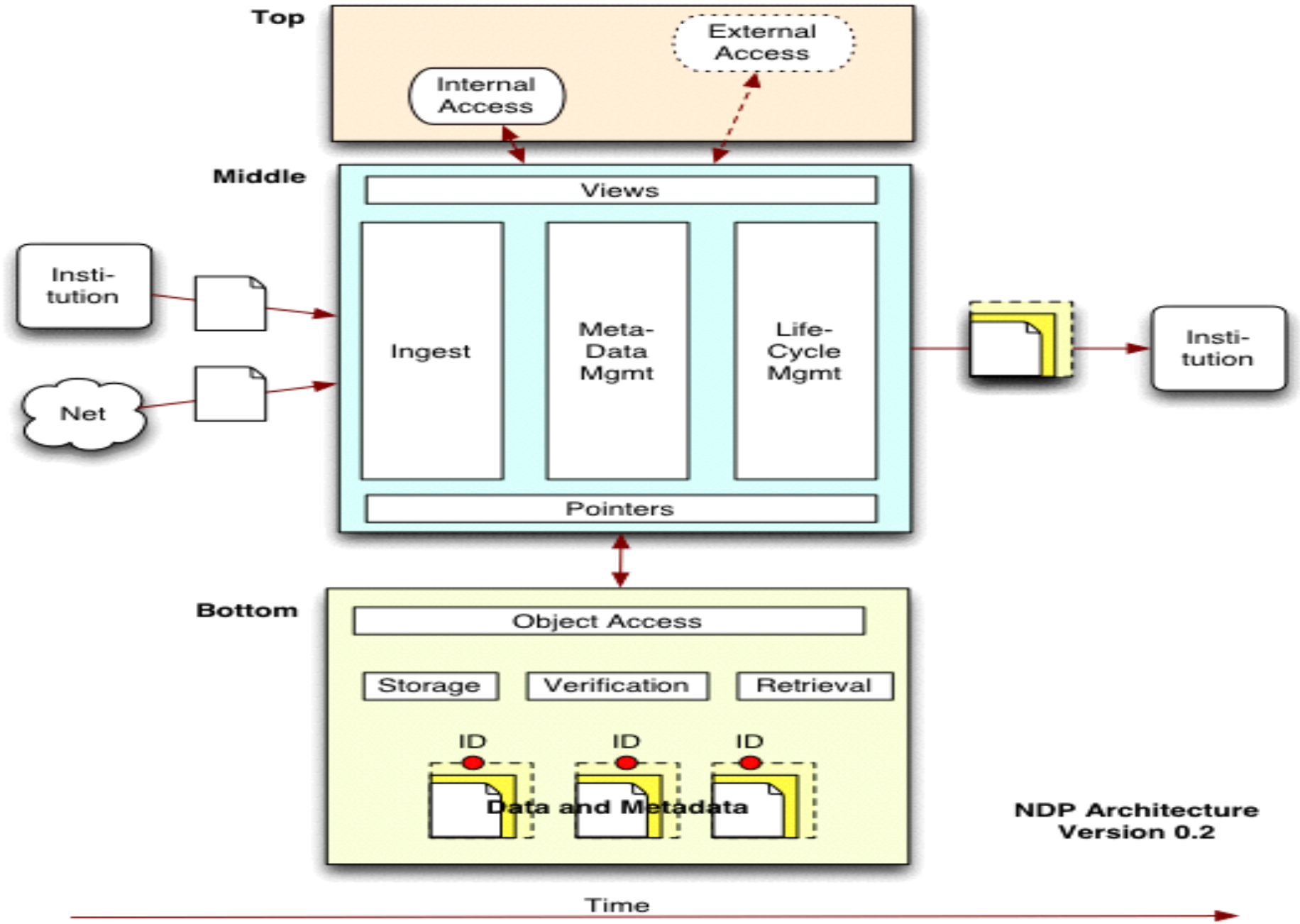
Content

Network

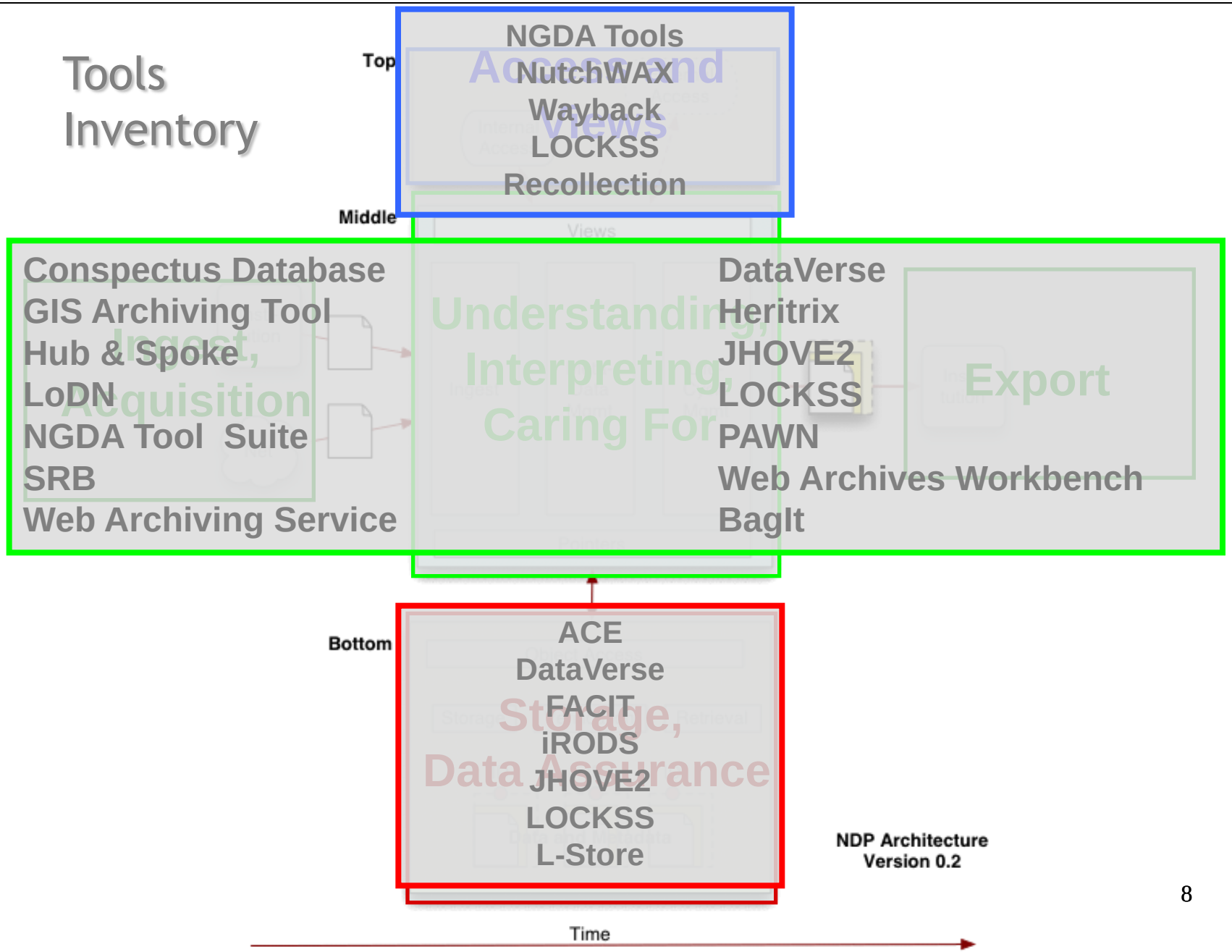
Technical Infrastructure

Technical Architecture: How is preservation happening?

An Architecture of Cooperation



An Architecture of Cooperation

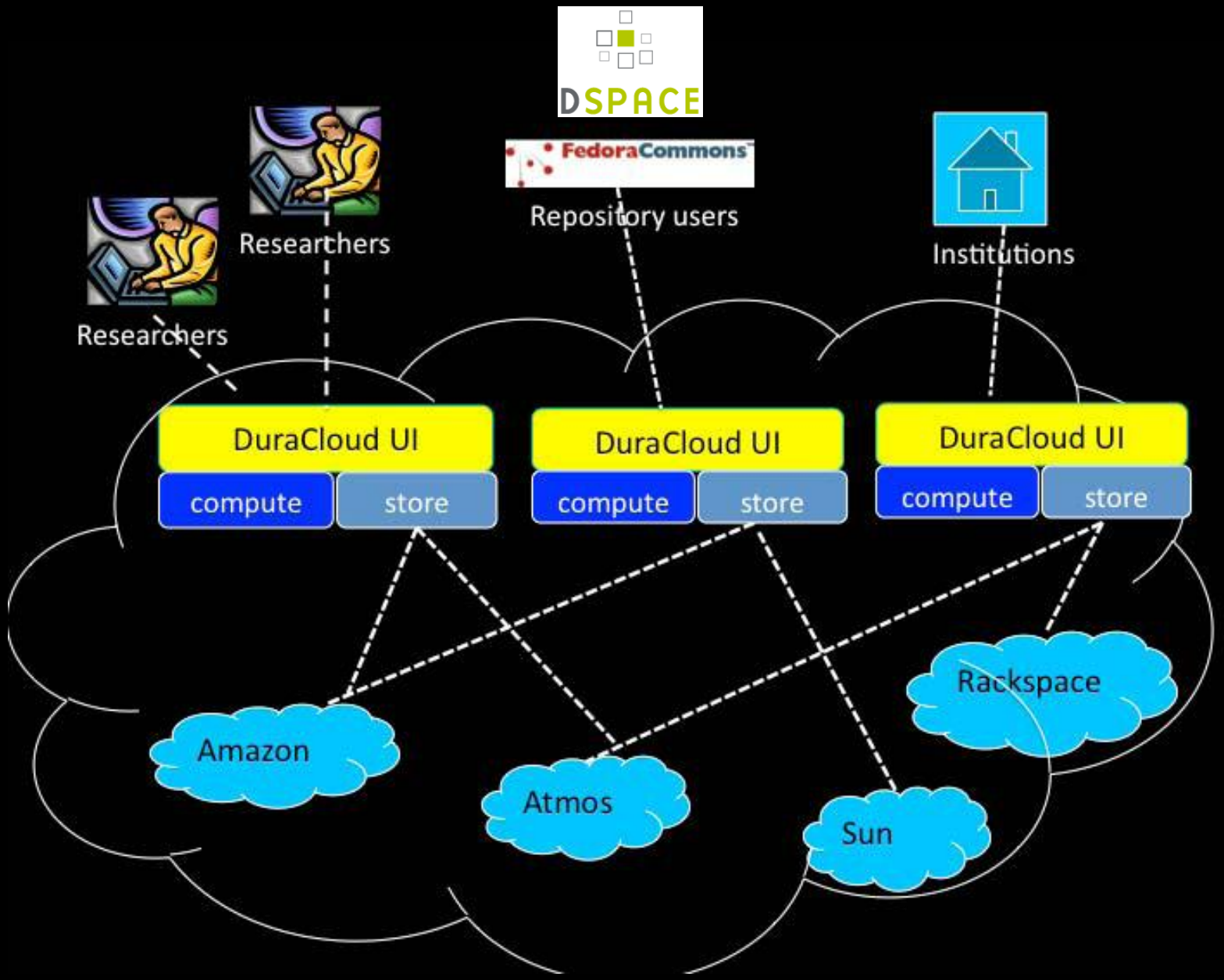


Some Current NDIIPP-Supported Initiatives in Cloud and Distributed Computing

DuraCloud

DURASPACE™

- A hosted service and open technology to help organizations and end users effectively utilize public cloud services.
- Built upon existing cloud infrastructure: Leave the basics of pure storage to those who do it best.
- Provide baseline functionality to replicate and distribute content across multiple cloud providers.
- Add value by enabling the deployment of services for access, preservation, reuse, and sharing of content stored in the cloud.


DSPACE
FedoraCommons
Repository users
Researchers
Researchers
Institutions
DuraCloud UI
compute
store
DuraCloud UI
compute
store
DuraCloud UI
compute
store
Amazon
Atmos
Sun
Rackspace

DuraCloud Pilot

DURASPACE™

- The outcome of a successful pilot program will be the public launch of the service, which will include a set of preservation support tools and services:
  - Ability to replicate content to multiple cloud providers through a single web interface.
  - Independent, on-demand bit integrity checking.
  - Synchronization with local Fedora or DSpace repositories.
  - Implementation of discovery tools enabled through the DuraCloud web interface, initially simple browse and search.
  - Ability to add metadata and tags to items and spaces defined within DuraCloud.

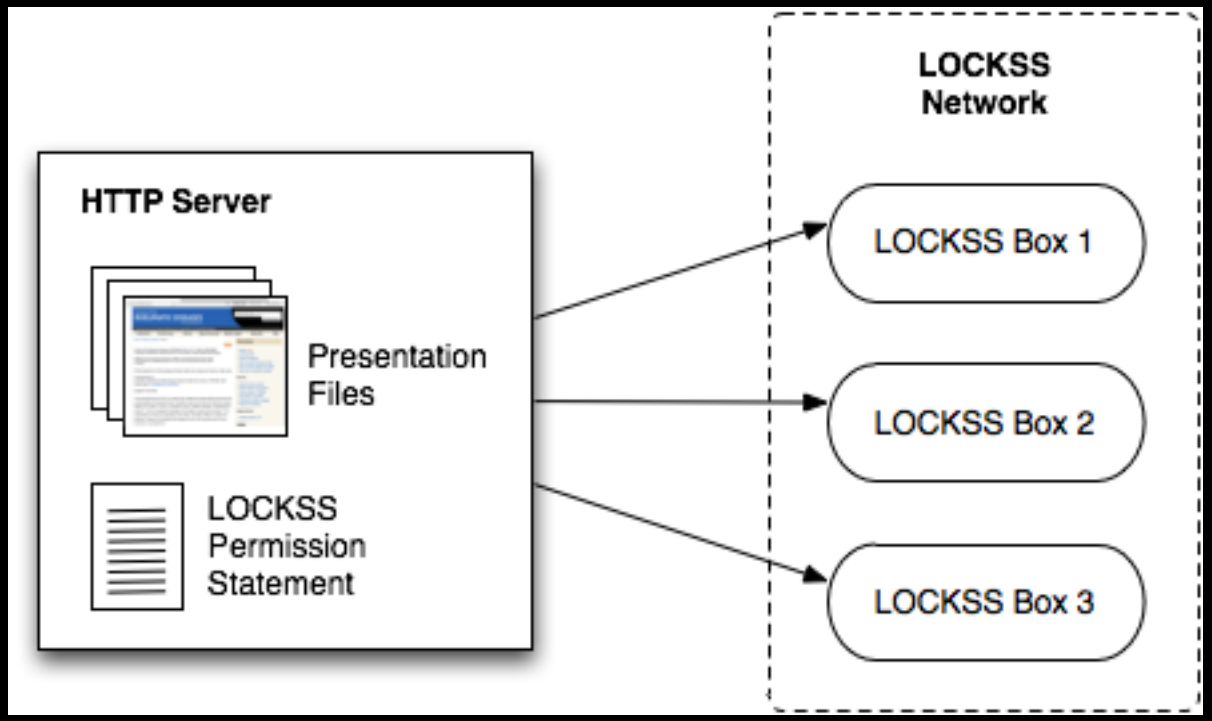
DuraCloud Pilot

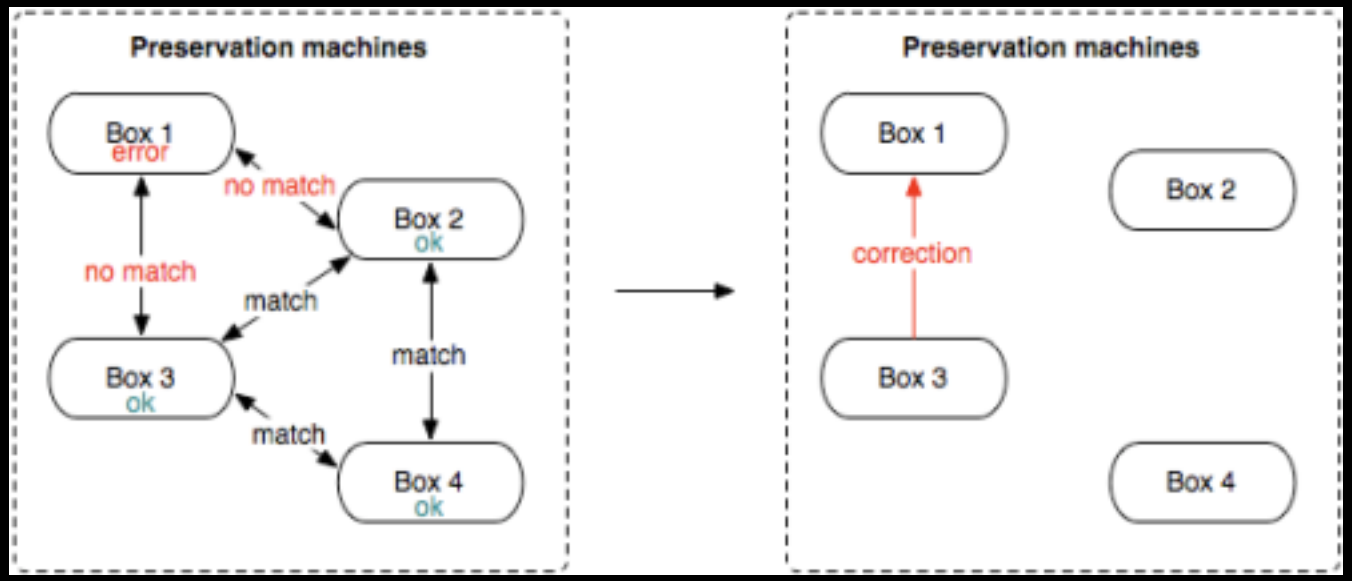
DURASPACE™

- The pilot also includes cloud content services based on open source applications:
  - Kaltura open source software for video streaming, editing, and possibly transcoding.
  - ImageMagick open source software for file format transformation.
  - Djatoka open source software for serving and viewing JPEG 2000 images.
  - “Taxonfinder" service to demonstrate the use of data mining tools.

LOCKSS in the Cloud

- A library uses LOCKSS software to turn low-cost hardware into a digital preservation appliance called a LOCKSS Box that performs the following four functions:
  - It collects content from the target web sites using a web crawler.
  - It continually compares the content it has collected with the same content collected by other LOCKSS Boxes, and repairs any differences.
  - It acts as a web proxy or cache, providing browsers in the library's community with access to the publisher's content or the preserved content as appropriate.
  - It provides an administrative interface for library staff to target new journals for preservation, monitor the state of the journals being preserved, and control access to the preserved journals.


LOCKSS
Network
HTTP Server
Presentation
Files
LOCKSS
Permission
Statement
LOCKSS Box 1
LOCKSS Box 2
LOCKSS Box 3


Preservation machines
Box 1
error
no match
Box 2
ok
no match
match
Box 3
ok
match
match
Box 4
ok
Preservation machines
Box 1
Box 2
correction
Box 3
Box 4

LOCKSS in the Cloud

- New LOCKSS development will provide the capability to run a LOCKSS box in the cloud under the DuraCloud umbrella:
  - For each cloud service provider, a "canned" version of their compute instances with the LOCKSS software installed will be made available, so LOCKSS is already correctly configured when the instance is activated.
  - It will be also possible to set up a LOCKSS box in a cloud compute service with the preserved content already held in the compute instance.

LOCKSS in the Cloud

- The LOCKSS audit and repair protocol is being re-packaged as a library that other digital preservation services could use for mutual audit and repair between multiple copies held in different storage technologies.
- These audits would not involve transferring the content being audited between copies, nor would they involve trusting the storage service, the compute service or the repository code holding the copies.
- LOCKSS will also add instrumentation to the audit and repair library to collect and report reliability information in a usable way, and measure how reliable the storage services in use actually are.

MetaArchive

- The creation of the MetaArchive Cooperative was a direct result of an NDIIPP award to develop a digital archive and cross-institutional scholarly portal service. Given the ease and minimal cost, combined with its ability to ensure the integrity of copies, this project found LOCKSS to be the best fit for its goals.
- MetaArchive Cooperative member institutions identify collections that they want to preserve – not just ejournals.

MetaArchive

- Using a technical framework that is based on the PLN software, these collections are ingested into a geographically distributed network where they are stored on secure file servers in multiple locations.
- A public LOCKSS network holds material of general interest to a wide community. Private LOCKSS networks (PLNs) hold material for smaller communities.
- MetaArchive is working with LOCKSS and Chronopolis to initiate cloud services and improved distributed services.

Chronopolis

- NDIIPP has been supporting the Chronopolis Project, a joint digital preservation data grid framework initiative between the University of California, San Diego (San Diego Supercomputer Center (SDSC)/UCSD Libraries), the National Center for Atmospheric Research (NCAR), and the University of Maryland.
- Chronopolis federates three repositories, each of which manages an independent metadata catalog, storage systems, authentication and authorization environment, and supports access, replication, and preservation services.
- The initial implementation was based on the Storage Resource Broker (SRB) grid system; the service is moving to iRODS, the Integrated Rule-Oriented Data System.

Chronopolis Digital Preservation DataGrid

Administration for Policy and Outreach

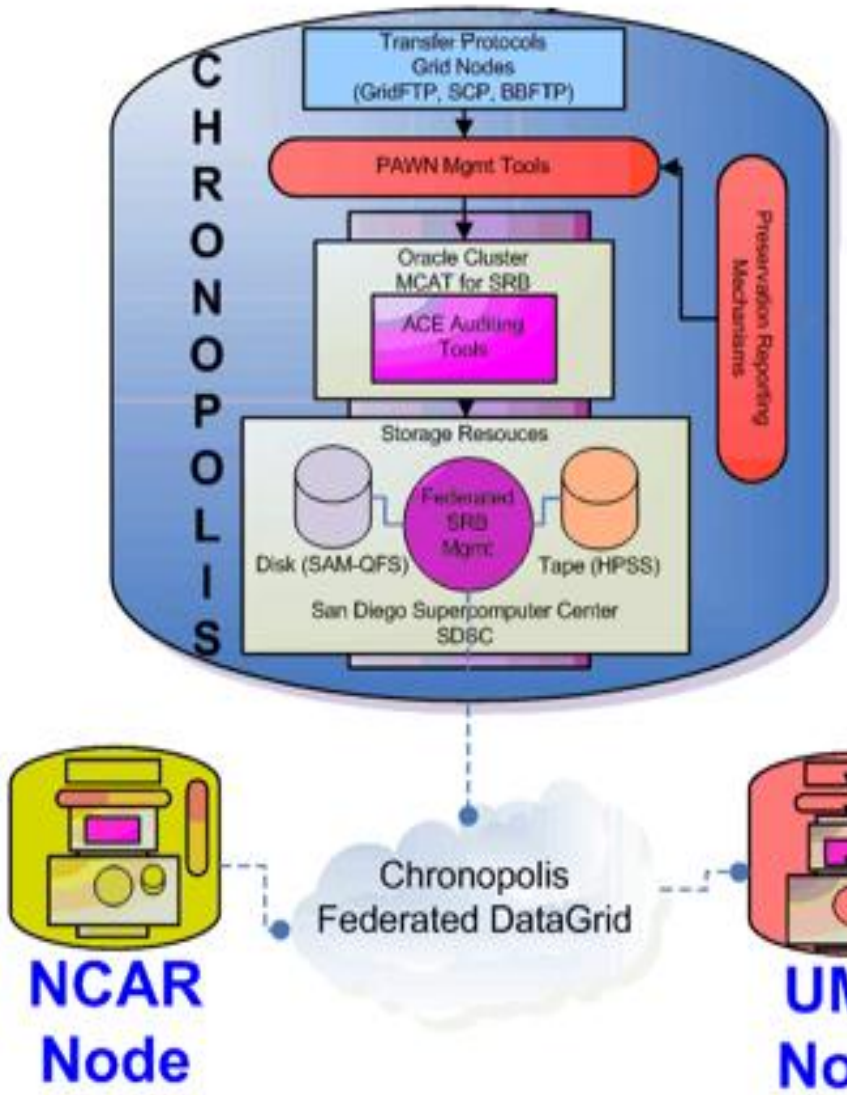
(Supports the overall partnerships and mgmt for preservation services and works as a liaison with Chroonpolis partners and other regional and national preservation programs)

Research and Development

(Research and development for rules-based preservation mgmt and technology forecasting for continual technology migration and mgmt)

Production Digital Preservation

(Long-term preservation with geographic replications and preservation services)



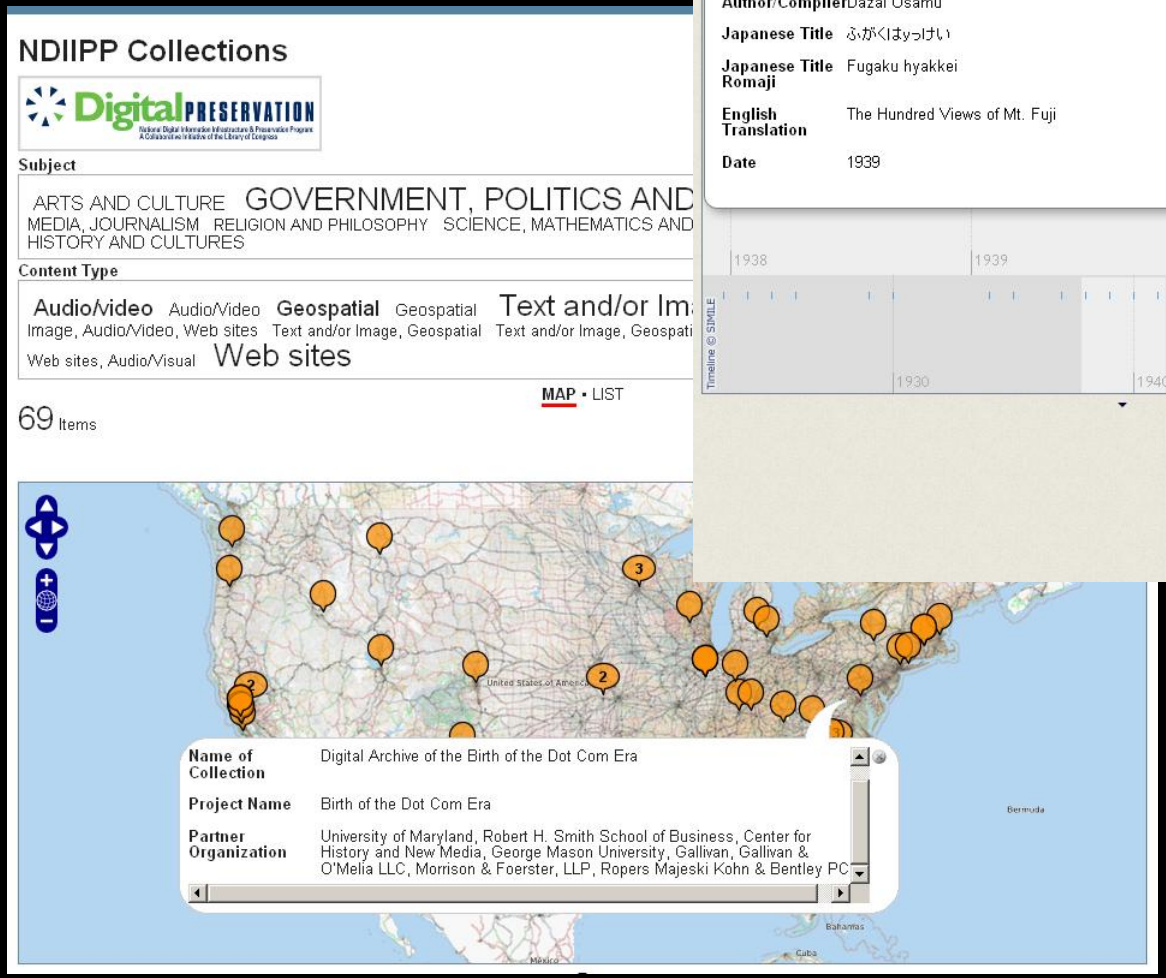
- Federated Replication
- Verification Tools
- Integrated Management

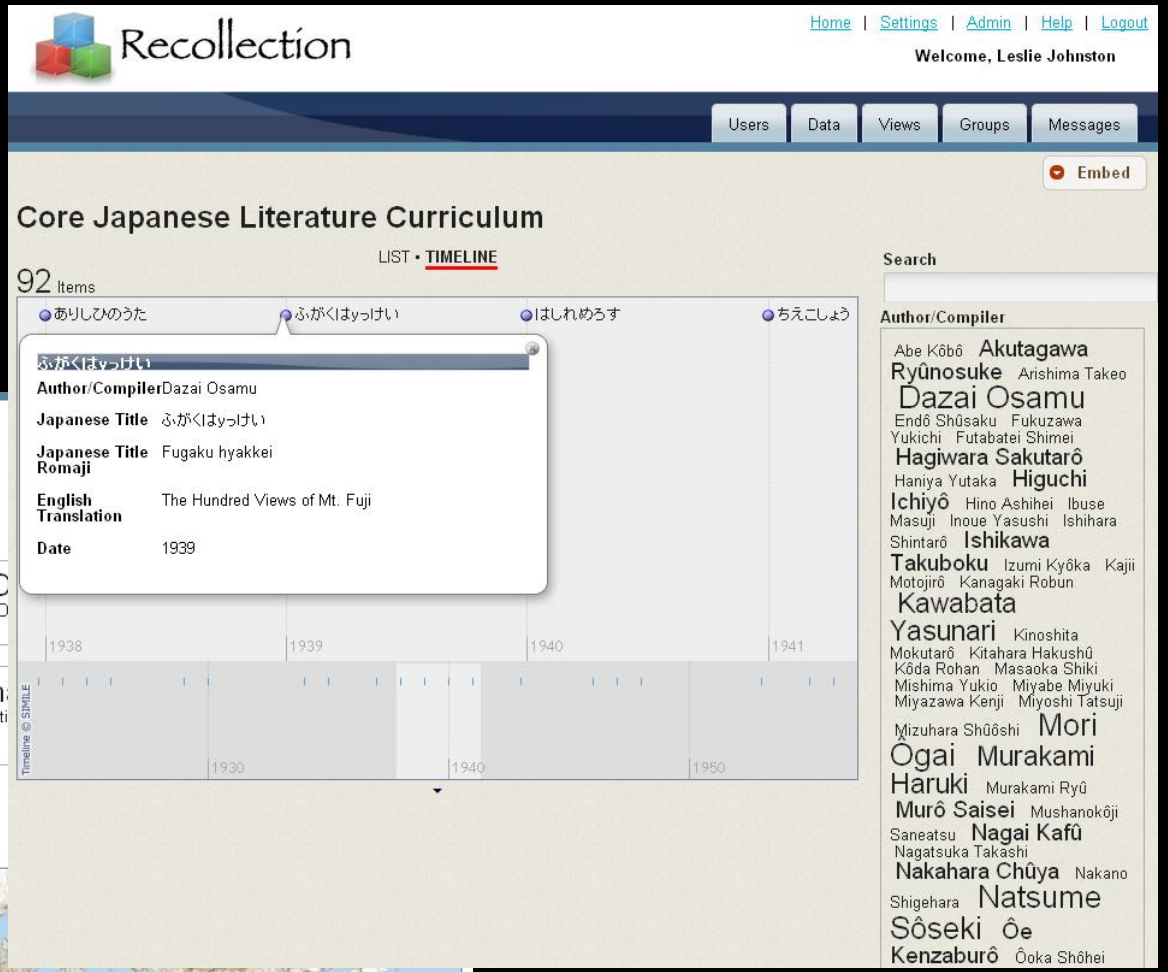
Chronopolis Tools

- SRB Replication Monitor
  - Watches registered directories and ensures that copies exist at designated mirrors.
  - The monitor stores enough information to know if files have been added or removed from the master site and when the last time a file was seen.
  - Every action is logged.
- Auditing Control Environment (ACE)
  - Software to protect the integrity of digital assets in the long term.
  - Underpinnings are based on rigorous cryptographic techniques.
  - Scalable and cost-effective, can interoperate with any archiving architecture.
- Both are product of UMIACS at the University of Maryland.

Recollection

- NDIIPP wanted to develop an environment that can be used to collect and explore information about digital collections, sharing diverse content across partners' collections.
- Partnering with the company Zepheira to create Recollection, a tool for creating new interfaces, relationships, and access points for digital resources.
- Recollection can be used to expose, visualize, and manage data uploaded from Excel spreadsheets or XML MODS (currently).
- You can view this data in a list or a table, on a map, or in timeline, scatter plot or pie chart, and you can embed the views you create in other websites.


NDIIPP Collections
Digital PRESERVATION
Subject
ARTS AND CULTURE GOVERNMENT, POLITICS AND
MEDIA, JOURNALISM RELIGION AND PHILOSOPHY SCIENCE, MATHEMATICS AND
HISTORY AND CULTURES
Content Type
Audio/video Audio/Video Geospatial Geospatial Text and/or Im
Image, Audio/Video, Web sites Text and/or Image, Geospatial Text and/or Image, Geospati
Web sites, Audio/Visual Web sites
MAP • LIST
69 Items
Name of Collection
Digital Archive of the Birth of the Dot Com Era
Project Name
Birth of the Dot Com Era
Partner Organization
University of Maryland, Robert H. Smith School of Business, Center for History and New Media, George Mason University, Gallivan, Gallivan & O'Melia LLC, Morrison & Foerster, LLP, Ropers Majeski Kohn & Bentley PC


Recollection
Home | Settings | Admin | Help | Logout
Welcome, Leslie Johnston
Users Data Views Groups Messages
Embed
Core Japanese Literature Curriculum
LIST • TIMELINE
92 Items
ありしひのうた
ふがくはゃっけい
はしれめろす
ちえこしょう
ふがくはゃっけい
Author/Compiler Dazai Osamu
Japanese Title ふがくはゃっけい
Japanese Title Romaji Fugaku hyakkei
English Translation The Hundred Views of Mt. Fuji
Date 1939
1938
1939
1940
1941
1930
1940
1950
Timeline © SIMILE
Search
Author/Compiler
Abe Kôbô Akutagawa Ryûnosuke Arishima Takeo Dazai Osamu Endô Shûsaku Fukuzawa Yukichi Futabatei Shimei Hagiwara Sakutarô Haniya Yutaka Higuchi Ichiyô Hino Ashihei Ibuse Masuji Inoue Yasushi Ishihara Shintarô Ishikawa Takuboku Izumi Kyôka Kajii Motojirô Kanagaki Robun Kawabata Yasunari Kinoshita Mokutarô Kitahara Hakushû Kôda Rohan Masaoka Shiki Mishima Yukio Miyabe Miyuki Miyazawa Kenji Miyoshi Tatsuji Mizuhara Shûôshi Mori Ôgai Murakami Haruki Murakami Ryû Murô Saisei Mushanokôji Saneatsu Nagai Kafû Nagatsuka Takashi Nakahara Chûya Nakano Shigehara Natsume Sôseki Ôe Kenzaburô Ôoka Shôhei

Recollection

- Recollection is a cloud service, hosted at Rackspace.
- We envision this is as a a data sharing service, a tool that lets the community share data, combine data, and add create customized views.
- Recollection is still in development; selected NDIIPP partners will be working with it this summer.

Discussion?

Leslie Johnston
lesliej@loc.gov
http://www.digitalpreservation.gov/