NDIIPP Preservation Architecture: Archive Ingest and Handling Test Interim Report

Digital Library Federation
October 2004, Baltimore, MD

National Digital Information Infrastructure and Preservation Program Goals

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

NDIIPP Focus Areas

- Network of preservation partners
- Preservation architecture
- Digital preservation research

What is the Preservation Architecture?

- A *conceptual framework* for supporting the technical functions and developing tools required for cooperative, distributed preservation of digital content

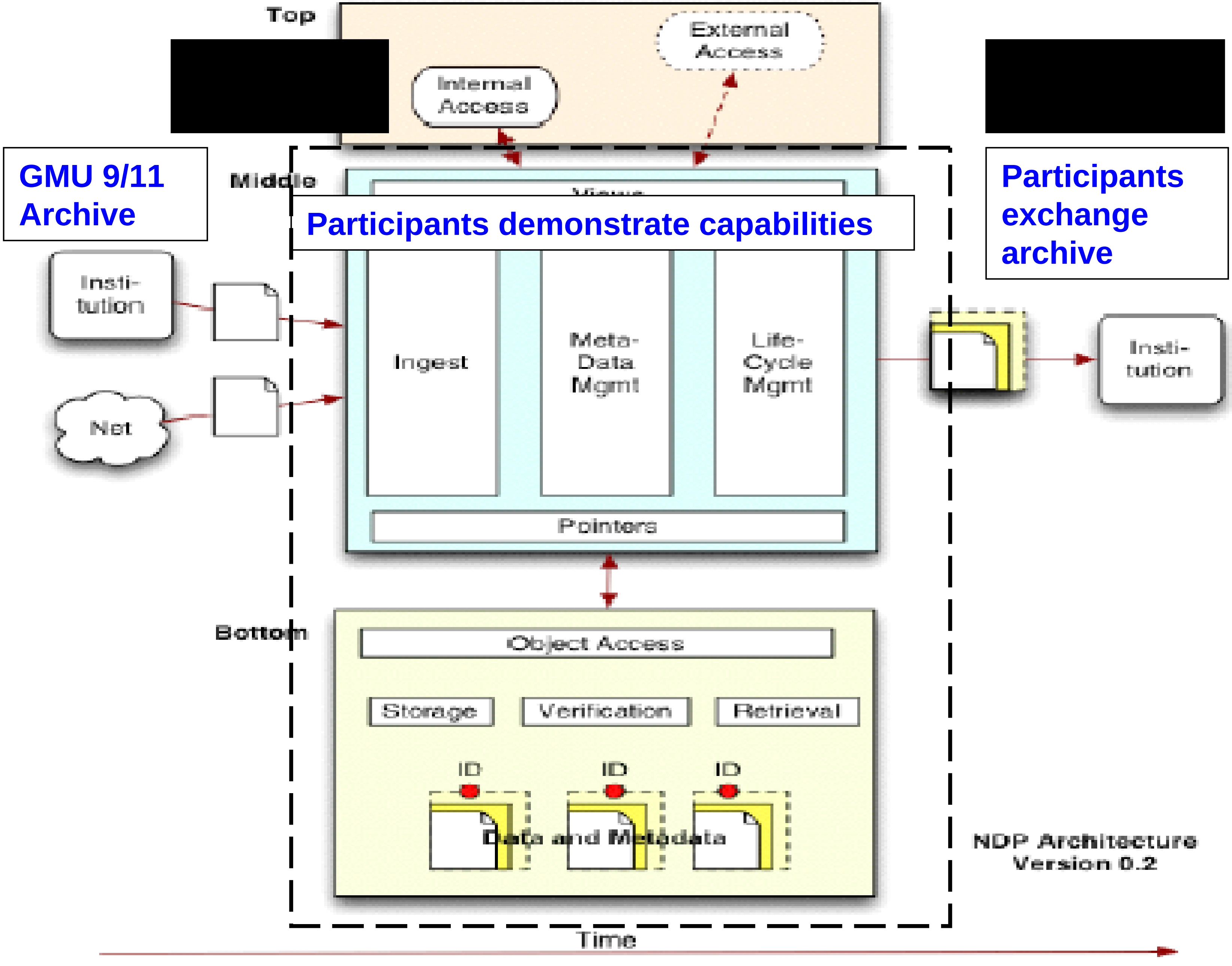
- It must
  - support relationships between institutions.
  - allow questions of preservation to be handled separately from questions of public access.
  - be built modularly, using existing technology and efforts wherever possible.
  - be able to be assembled over time.
  - be specified using broadly adoptable protocols.

Goals of Architecture

- Evaluate Systems
- Look for Areas of Interoperability
- Encapsulate Institution-specific Goals
- Generalize Interfaces
- Provide View Towards Federation

Archive Ingest & Handling Test

- AIHT is a first test of proposed preservation architecture
  - Leveraging existing systems and research
  - Uses a common data set, the George Mason University 9/11 Archive
- Phase I tests transfer from donating archive and data handling within local systems
- Phase II tests export and import between test participants



[redacted]

[redacted]

GMU 9/11 Archive

Participants demonstrate capabilities

Participants exchange archive

Participants

- Harvard University Library
- The Johns Hopkins University, Sheridan Libraries
- Old Dominion University, Department of Computer Science
- Stanford University Libraries & Academic Information Resources
- The Library of Congress, Office of Strategic Initiatives

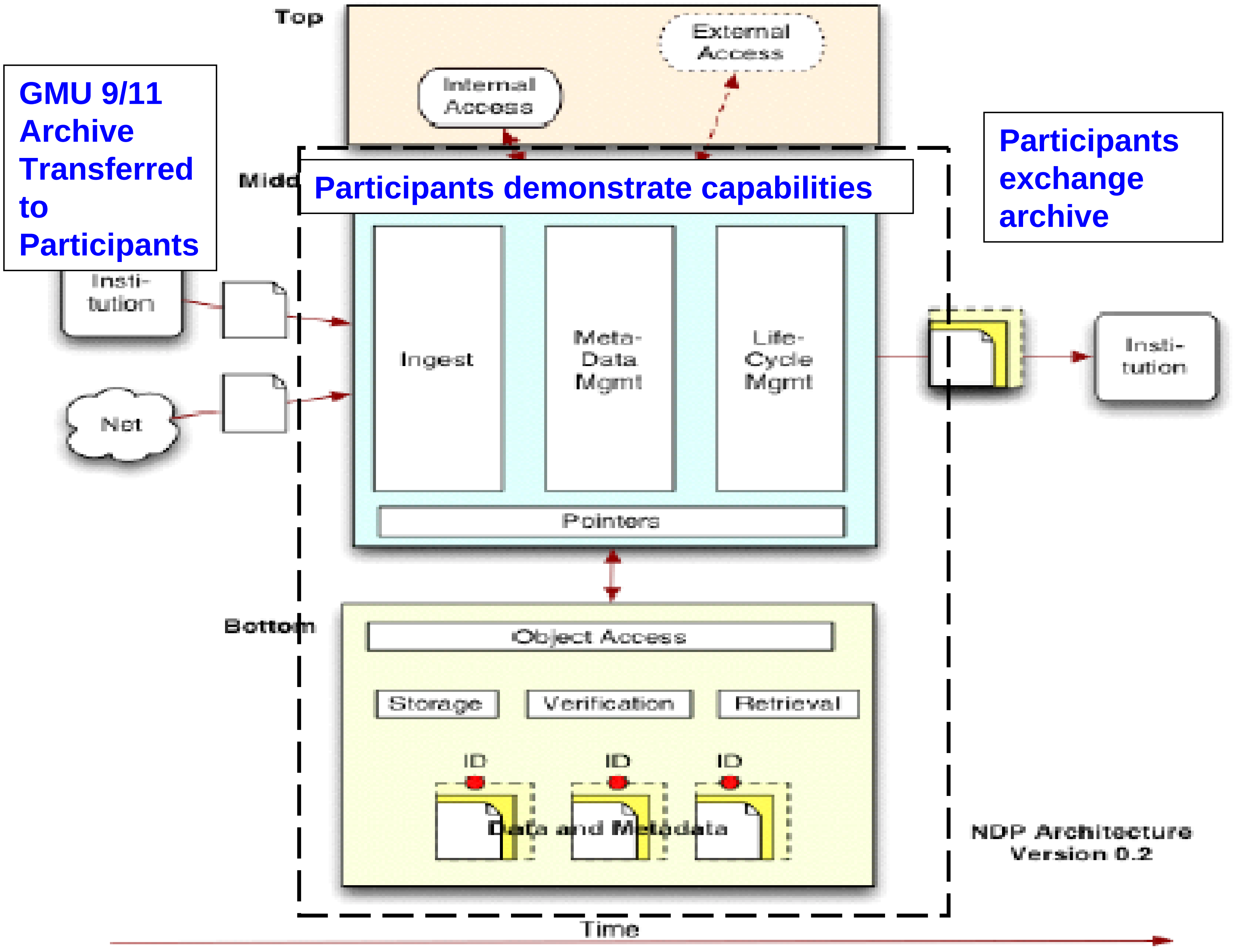
Design of AIHT

Give a moderately complex archive to several institutions and have them:

- Describe it
- Mark it up
- Ingest it
- Transform it
- Share it

Goals of AIHT

- Gain practical experience with multiple institutions
- Document transfer and ingest processes for multiple systems
- Determine next set of tasks for developing interfaces between layers and institutions



GMU 9/11 Archive Transferred to Participants

Participants demonstrate capabilities

Participants exchange archive

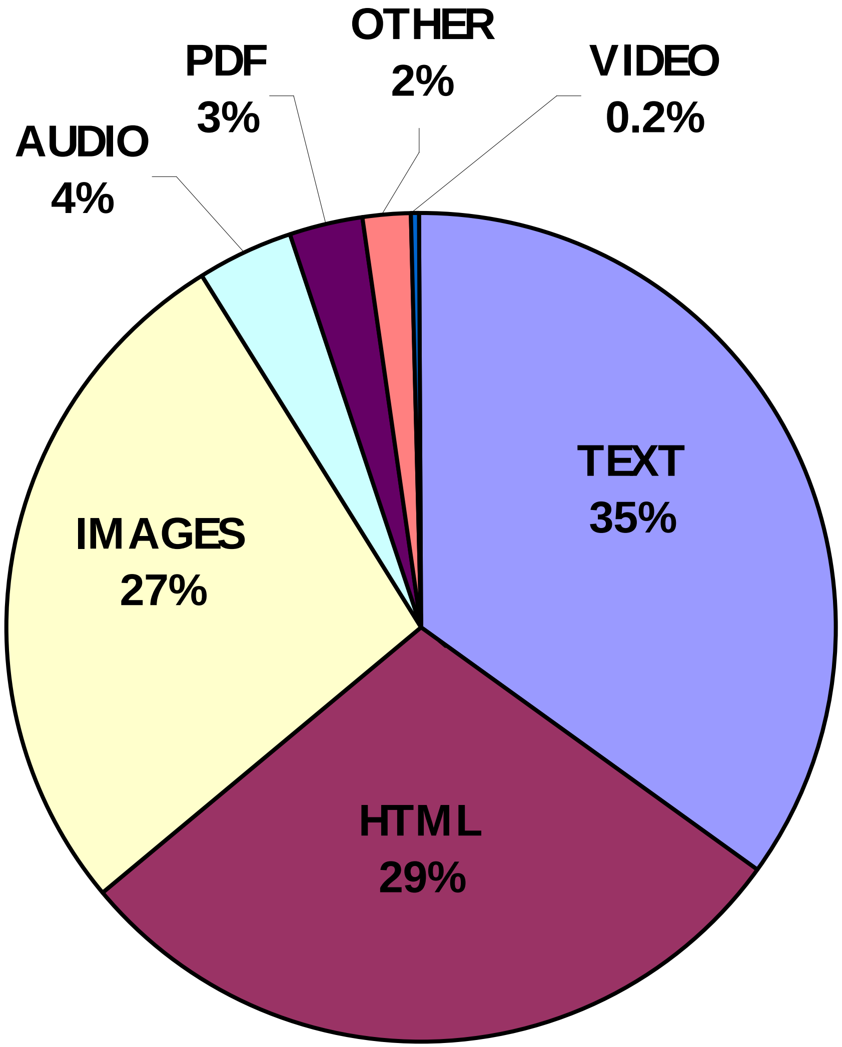
GMU 9/11 Archive

- Physically small (~12Gb)
- Conceptually large (~57K files, many types)
- Messy (Amateur contributions, various naming schemes)
- No solid meta-data
- No access to original sources
- As inconsistent as real life

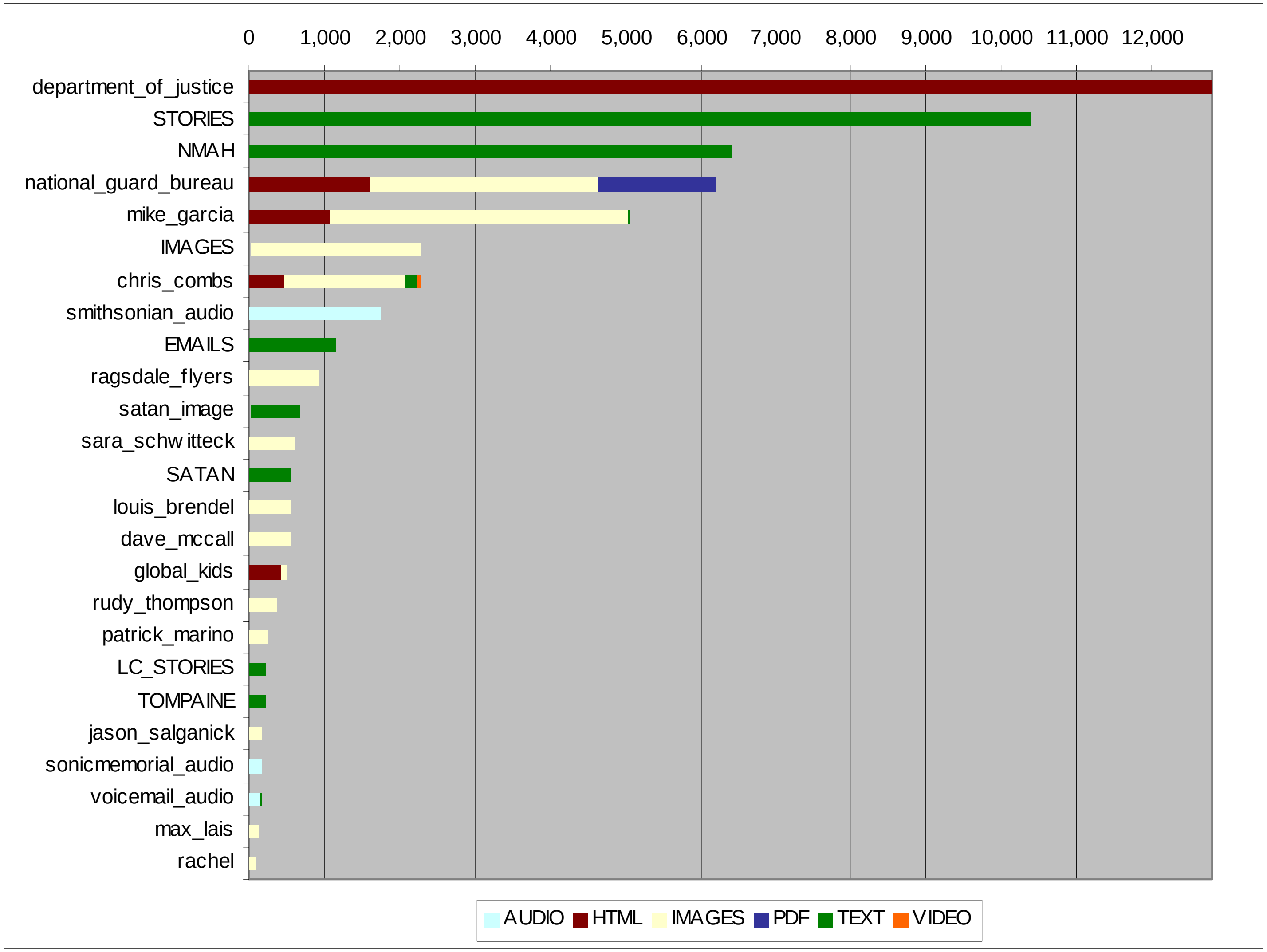
Inconsistent Descriptions

	GMU Document	GMU DB	GMU TMD	LC Inspection
Size	12GB			12GB
File Count	57,442	57,492	57,540	57,540
“Collections”	2,105	171		
“Sub-Collections”		1,934		
“Contributors”		17,504		170

Imbalanced Disposition of Content



Great Breadth of Contribution



Current Issues

- Simple Receipt
- Physical Storage and Naming
- The Issue of Uniqueness
- Triage and the 80/20 Rule
- Markup as Forensics

Harvard University

- Background statement
  - Current policy of the library repository limits deposit to objects created by approved workflows, in a small set of formats, and accompanied by preservation metadata. As this policy evolves towards that of an institutional repository, AIHT presents the opportunity to investigate issues surrounding deposit of arbitrary content of unknown provenance.
- Project Approach
  - Use JHOVE to provide enriched technical metadata
  - Build tools to generate SIP packages automatically
  - Investigate TIFF-to-JPEG 2000 transformations
  - Enhance metadata model to record PREMIS-like provenance information
  - Add export functionality to repository API

Harvard University

- Project team
  - Dale Flecker – Principal investigator
  - Stephen Abrams – Project manager
  - Stephen Chapman – Reformatting analyst
  - Sue Kriegsman – Project administration and reporting
  - Gary McGath – Developer
  - Germain Seac – Operations
  - Robin Wendler – Metadata analyst
- Technologies
  - Digital Repository Service (DRS) – Oracle (metadata), Java API, RAID (content), Solaris, XML-based SIP package
  - JHOVE for extraction of encapsulated technical properties
  - Automated SIP creation tools

Harvard University

- Observations
  - JHOVE can process 97% of the 57,000 files
    - ASCII/UTF-8, HTML, JPEG, WAV, TIF, PDF, GIF, AIFF, XML
  - The PREMIS event model is very flexible, but it is difficult to determine the best way to capture provenance metadata
  - Data manipulation issues:
    - You can FTP 13GB as one file in 3 hours; to FTP it as 57,000 files takes 35+ hours
    - Some FTP clients do not like 0 length files
    - Some ZIP tools have a file size limitation
    - Some network appliance file servers have a file size limitation
  - The data does not include any infected files!

Old Dominion University

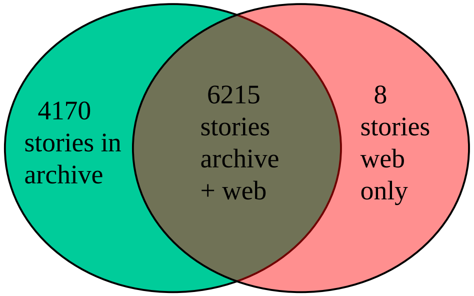
- Background
  - experiment with alternate archive architectures
  - create self-preserving digital objects
- Project Approach
  - build ingestion tool to test individual file validity
  - use JHOVE, unix “file”, Fred, and other tools to generate technical metadata
  - perform preservation analysis comparing the archived version with the versions that are available on the open Internet
    - original site, Google, Yahoo, IA, etc.
  - create an MPEG-21 DIDL that contains:
    - preservation analysis, technical metadata, original tar file, current tar file and “deltas” (cf. diff/patch semantics) for intermediate versions
  - store DIDLs in self-contained, mobile archivelets (“buckets”)

Old Dominion University

- Project Team
  - professors
    - Michael L. Nelson, Johan Bollen
  - graduate students
    - Giridhar Manepalli, Rabia Haq
- Technologies
  - Bucket 3.0 Digital Objects
  - MPEG-21 DIDL
  - JHOVE, file, Fred
  - various locally developed ingestion / conversion scripts

Old Dominion University

- Observations
  - significant learning curve for MPEG-21 DIDL
    - hoping to incorporate MPEG-21 Rights Expression Language (REL) in the AIHT testbed
  - conversion utilities (e.g. ImageMagick) are assumed to:
    - exist outside of the archive
    - be transient services
  - significant discrepancies between archived and live web site:



Stanford University

- Background: The Stanford Digital Repository's ingest infrastructure was originally focused on highly normative bibliographic digital objects. The AIHT project provided the opportunity to develop our capabilities for real-world non-normative digital collections.
- Project Approach – Develop tools (such as the Stanford Empirical Walker™) and integrate others (such as JHOVE) to automate the process of digital collection assessment, including technical metadata harvesting, structural description, and preservation risk assessment.

Stanford University

- Team:
  - Richard Anderson – Programming
  - Keith Johnson – Project Management
  - Hannah Frost – Preservation Methodologies
  - Nancy Hoebelheinrich – Metadata
  - Jerry Persons – Information Architecture
  - Cathy Aster – Reporting and Financial Management
- Technologies:
  - Solaris, Windows, Java, METS, Harvard METS Toolkit, JHOVE, PREMIS

Stanford University

- Observations:
  - Expected preservability status:
    - 70% HIGH
    - 27.5% ACCEPTABLE
    - 2.5% MINIMAL
  - A large file collection generates a very large METS file, and large XML files require lots of memory and processing power
  - Keeping metadata in parallel file hierarchy judged potentially more efficient that collecting all into a single XML file
  - User-supplied metadata can be messy and difficult to transform to a standard format
  - PREMIS data elements/model looks very promising for storing preservation status and methodologies

The Johns Hopkins University

- Background: Johns Hopkins University Sheridan Libraries has been investigating multiple repositories. AIHT provided a digital preservation use case.
- Project Approach: Large-scale ingestion with a repository agnostic design

The Johns Hopkins University

- Team: Mark Patton (developer), Sayeed Choudhury (PI), Tim DiLauro (tech lead), Jacque Gourley (project manager), Ying Gu (student), David Reynolds (metadata), Jason Riesa (student)
- Technologies: DSpace, Fedora, METS, Mac OS X, Java

The Johns Hopkins University

Observations:

- Where possible there should a high degree of coordination and agreement between the content provider and the archive recipient
- Design metadata from established standards, instead of attempting to shoehorn
- Currently there is not a seamless way to ingest to multiple repositories without developing a repository agnostic layer
- Bulk ingestion of a complex archive is a good way to stress test repository interfaces

Early Conclusions

- Every Choice Matters (e.g. hard drive)
- Low-level Tools Work Best (e.g. tar file)
- Almost no support for archive-level transfer (Transfer Metadata a key early format)
- Poor support for file inspection (LC developing pluggable software)
- Numbers are meta-data

Next Steps: Phase I

- Collate experiences of participants
- Next revision of TMD format
- Work on inspection tools
- Draft recommendations for naming and file and MIME types
- Explore format registry

Next Steps: Phases II and III

- Transform data formats
- Destroy and backup data
- Export/import entire archive
- Observe and report results