

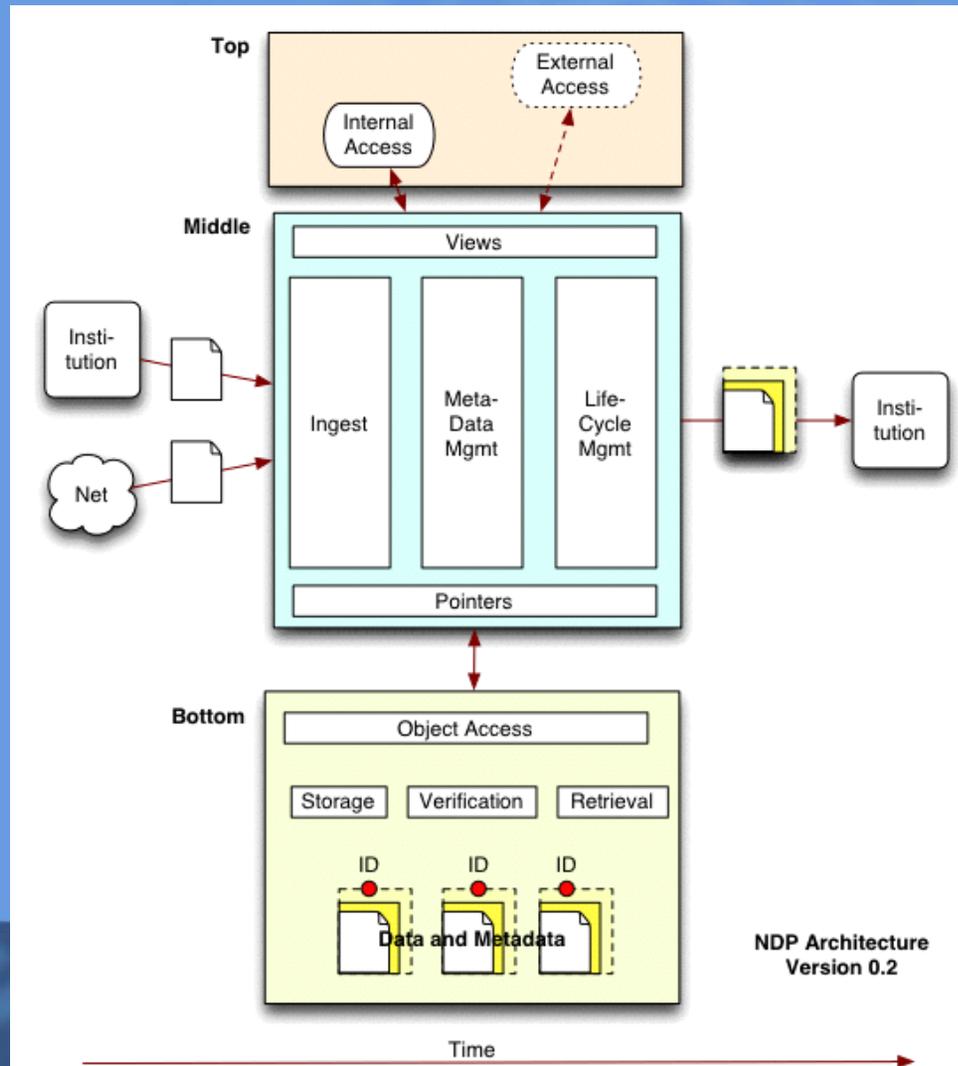


NDIIPP  
Technical/Preservation  
Architecture Partners  
June 28, 2004

# Preservation Architecture Partner Activities

- Federated Preservation: *Archive Ingest and Handling Test*
- Technical Infrastructure and Tools: *Los Alamos National Lab Research Library*
- Prototype Digital Archive System: *NASA Langley's Atmospheric Sciences Data Center*

# Architecture



# Things We've Learned

- There Are Too Many Questions to Get One Right Answer
- Great Minds Don't Think Alike
- Metadata is Worldview
- No Matter Who You Are, Most of the Smart People Work for Someone Else

# 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

# AIHT Participants

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

# Design of AIHT

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

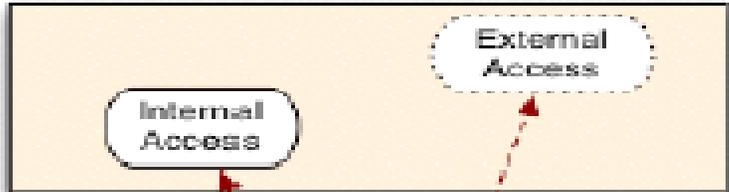
- Describe it
- Mark it up
- Ingest it
- Transform it
- Share it

# 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**

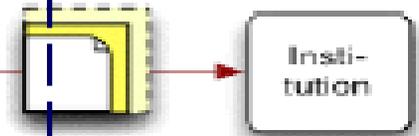
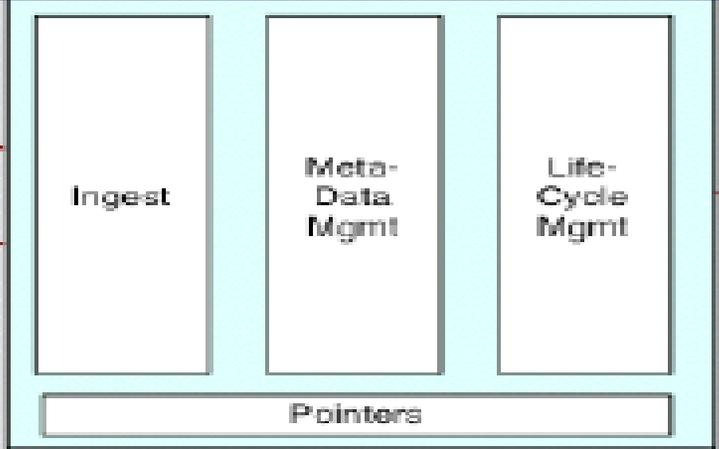
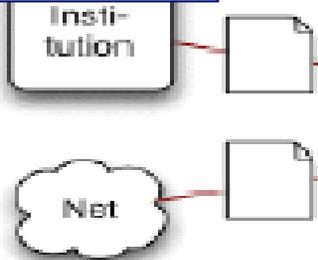
Top



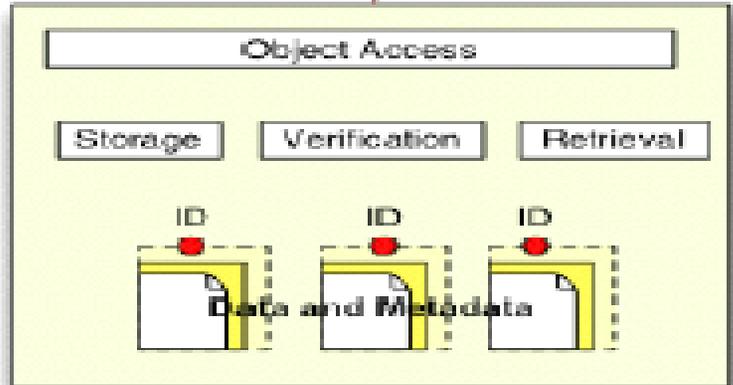
Middle

**Participants demonstrate capabilities**

**Participants  
exchange  
archive**



Bottom



NDP Architecture  
Version 0.2

Time



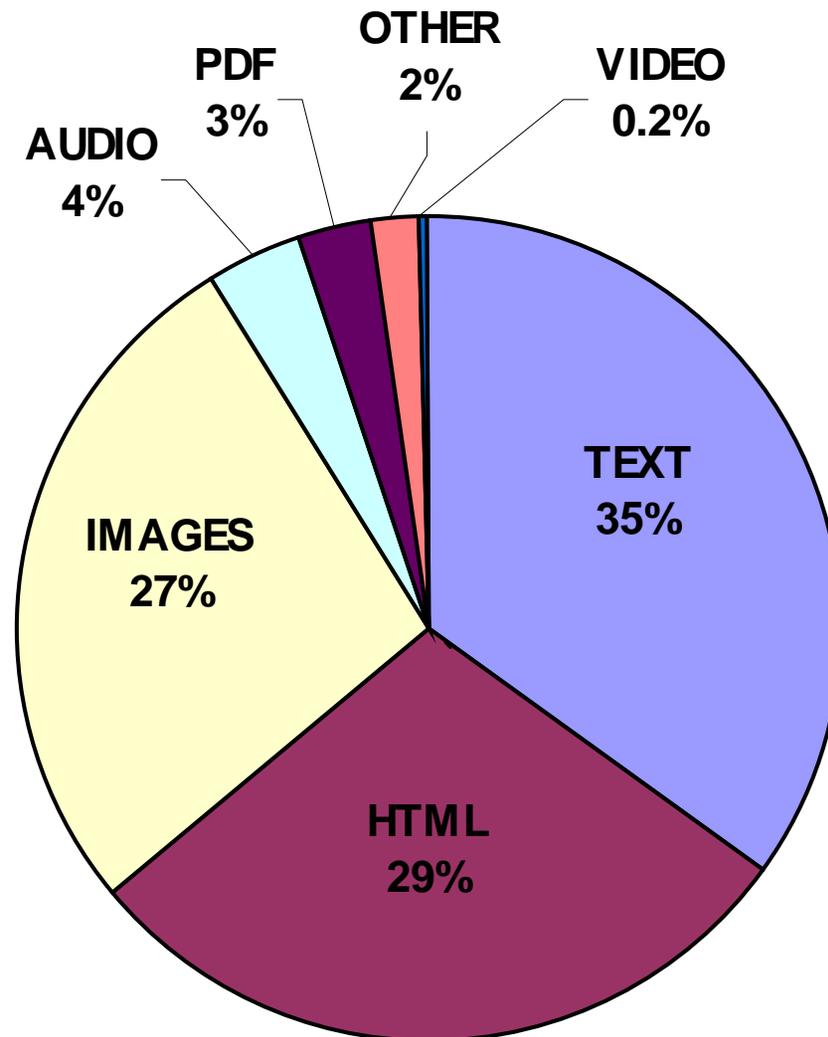
# GMU 9/11 Archive

- 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

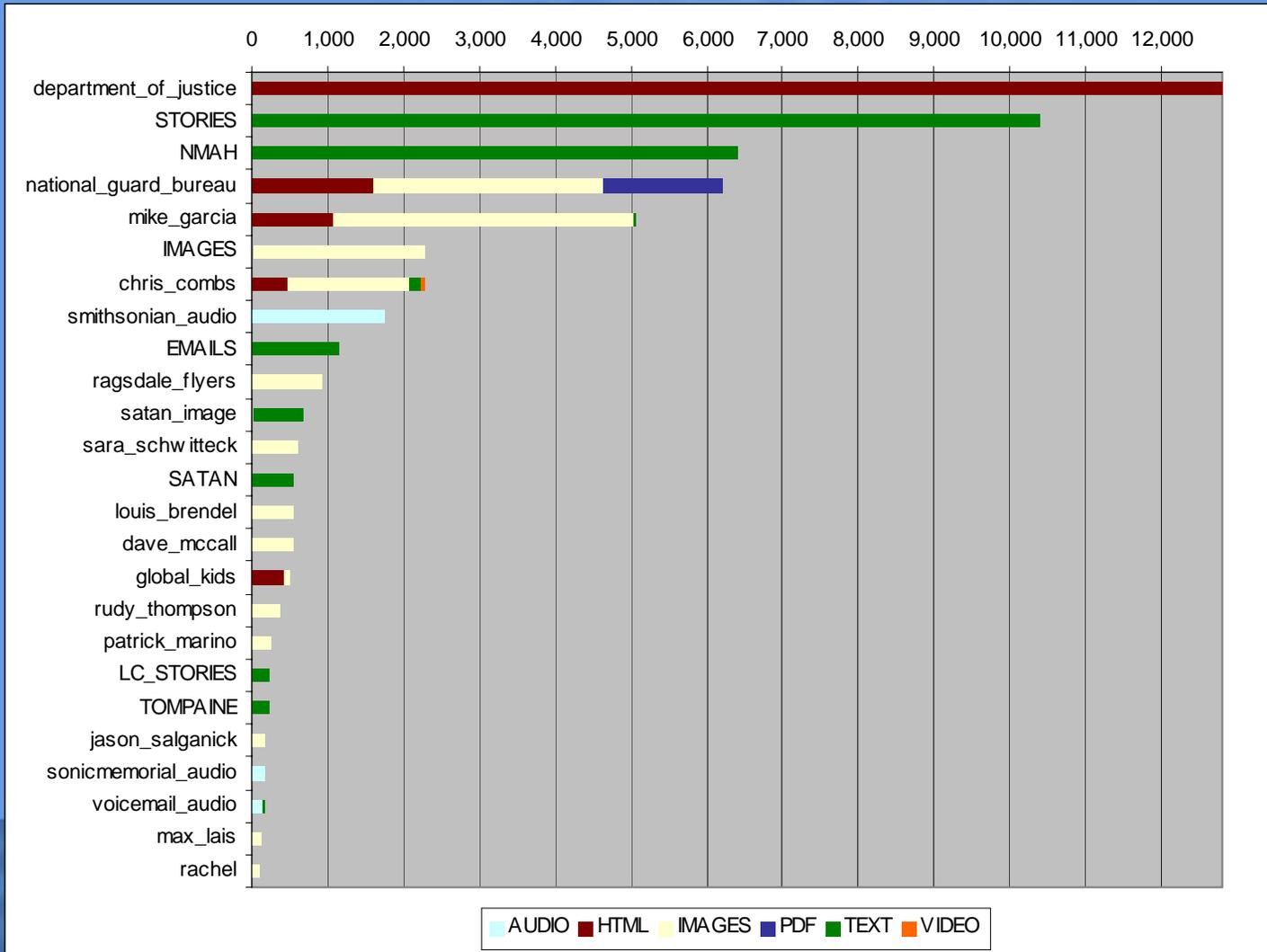
# Inconsistent Descriptions

	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

# 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: Phase II

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

# Technical Infrastructure and Tools: Los Alamos National Laboratory Research Library

- Three sub-projects focusing on content from the following sources:
  - Electronic journals
  - Web crawls
  - American Memory
- Approach: To develop a common technical infrastructure using XML, MPEG-21 DIDL and OAI-PMH technologies

# Prototype Digital Archive:

## NASA Langley's Atmospheric Sciences Data Center

- Architecture based upon a Open Archive Information System (OAIS) Reference Model
- Proposes to extend current ASDC archive to a grid-enabled Linux cluster
- Applicable to a high speed- high volume data flow such as video, satellite feed or web capture