Tools and Services for the Long Term Preservation and Access of Digital Archives

Joseph JaJa, Mike Smorul, and Sangchul Song
Institute for Advanced Computer Studies
Department of Electrical and Computer Engineering
University of Maryland, College Park

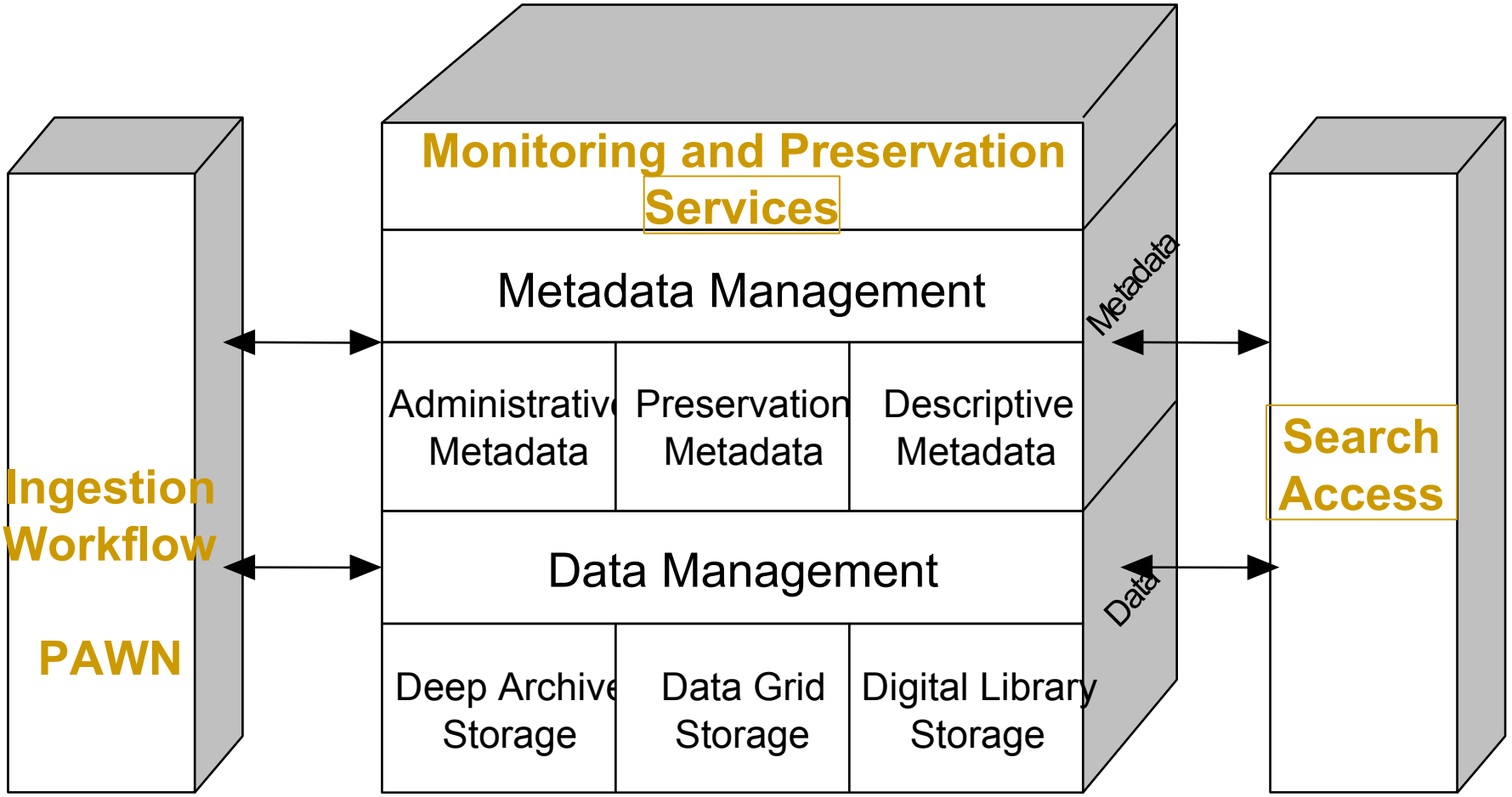
Background

- Started as an ERA project focusing on setting up and testing a distributed archiving infrastructure.
- Evolved into the development of archiving tools and services that are scalable and platform independent.
- In addition to the continued NARA support, the work has been supported by NSF, Library of Congress, and the Mellon Foundation.

Main Tools Developed

- Flexible software environment for ingestion and for handling producers – archive interactions: PAWN.
- Tools to ensure the long term integrity of digital holdings based on rigorous cryptographic methodologies: ACE.
- Methods to ensure compact storage and fast retrieval of archived web contents: PISA.
- Tracking and Monitoring tool of the digital holdings of an archive.

Software Developed and Tested on TPAP:



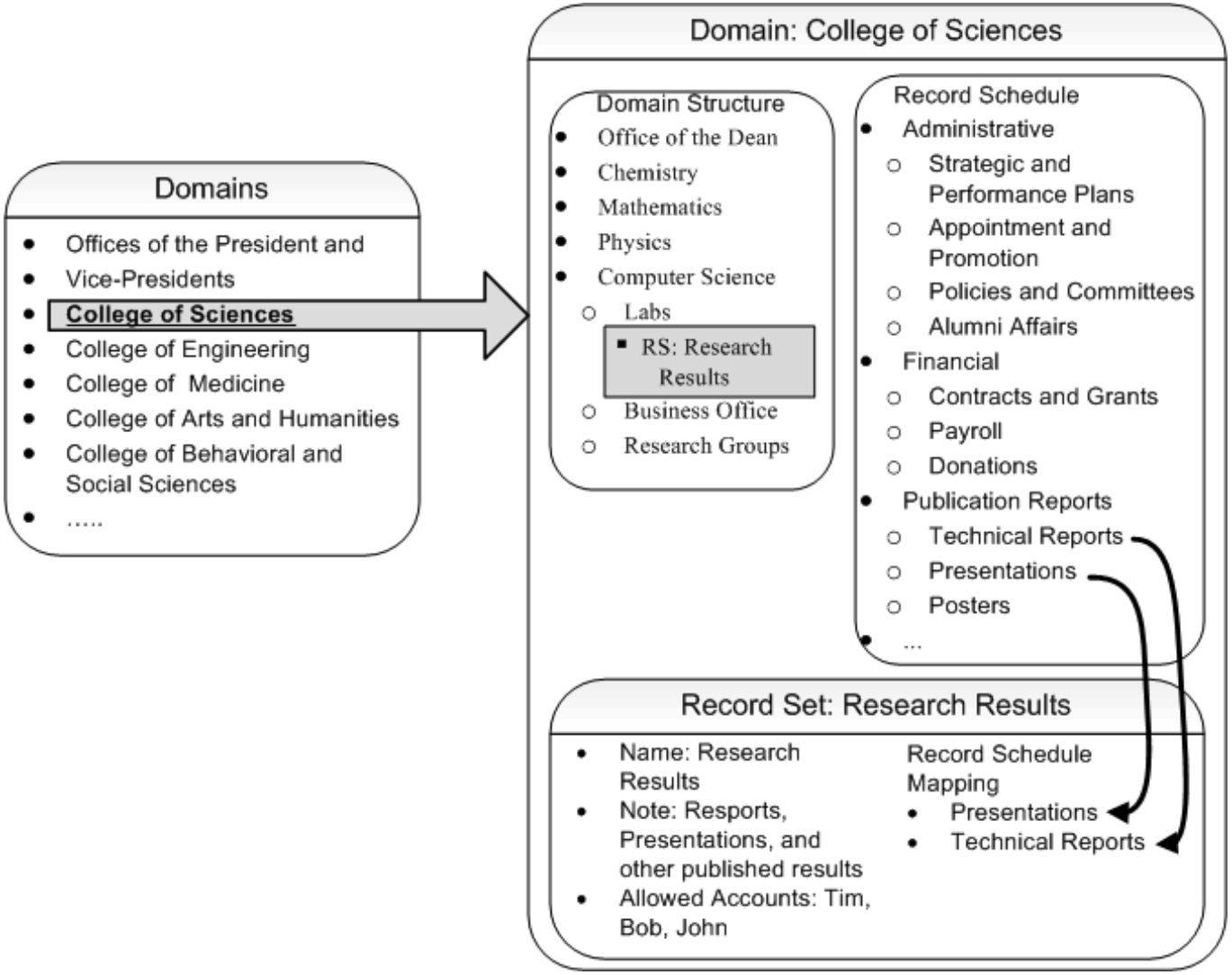
PAWN – Producer Archive Workflow Network

- Software that provides a flexible and customizable ingestion framework
- Handles the process in a reliable and secure fashion:
  - From package assembly
  - To archival storage
- Simple interface for end-users
- Flexible interface for archive managers
- Designed for use in multiple contexts

Overall Organization

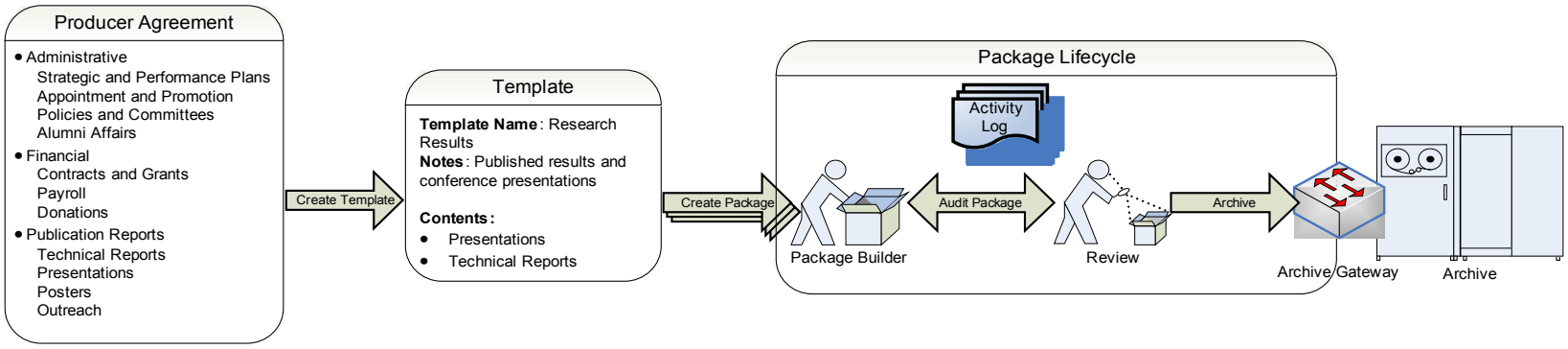
- Producers organized into domains, each domain contains a transfer agreement negotiated with the archive.
- Each domain contains a hierarchical organization of data grouped into record sets/templates (convenient groupings from the transfer agreement).
- An end-user operates within a domain with record sets associated with the account.

Producer-Archive Agreement



Package Workflow Overview

1. Create Producer-Archive Agreement and client package template.
2. Create package based on template
3. Optionally, review submitted items
4. Invoke publishing processes.



Customizable Components

- Definable Roles
  - Actions in PAWN can be grouped to create arbitrary types of users
- Flexible Approval Requirements
  - Signature requirements can be placed on parts of a package.
- Automated Processing
  - API for creating processes to validate, transform, approve, or publish items in a package
  - Processes can be invoked manually or automatically
  - Processes may have dependencies on item approval

Sample Submission

1. Client ingests image data
2. First process chain: Validators check image format and marks ‘good’ files as approved.
3. Files that are rejected (misc mp3’s, etc..) are held for manual processing
4. Second Chain: push approved files into DSpace/Fedora/whatever

PAWN Summary

- Flexible environment to handle ingestion between many producers and an archive.
- Very little effort for producers to push their data into the archive.
- Granular workflow definition.
  - Fully automated to completely manual.
- Easy to include new standards (metadata, packaging, …).
- Tested in a number of environments (including the NARA TPAP testbed and the Library of Congress).

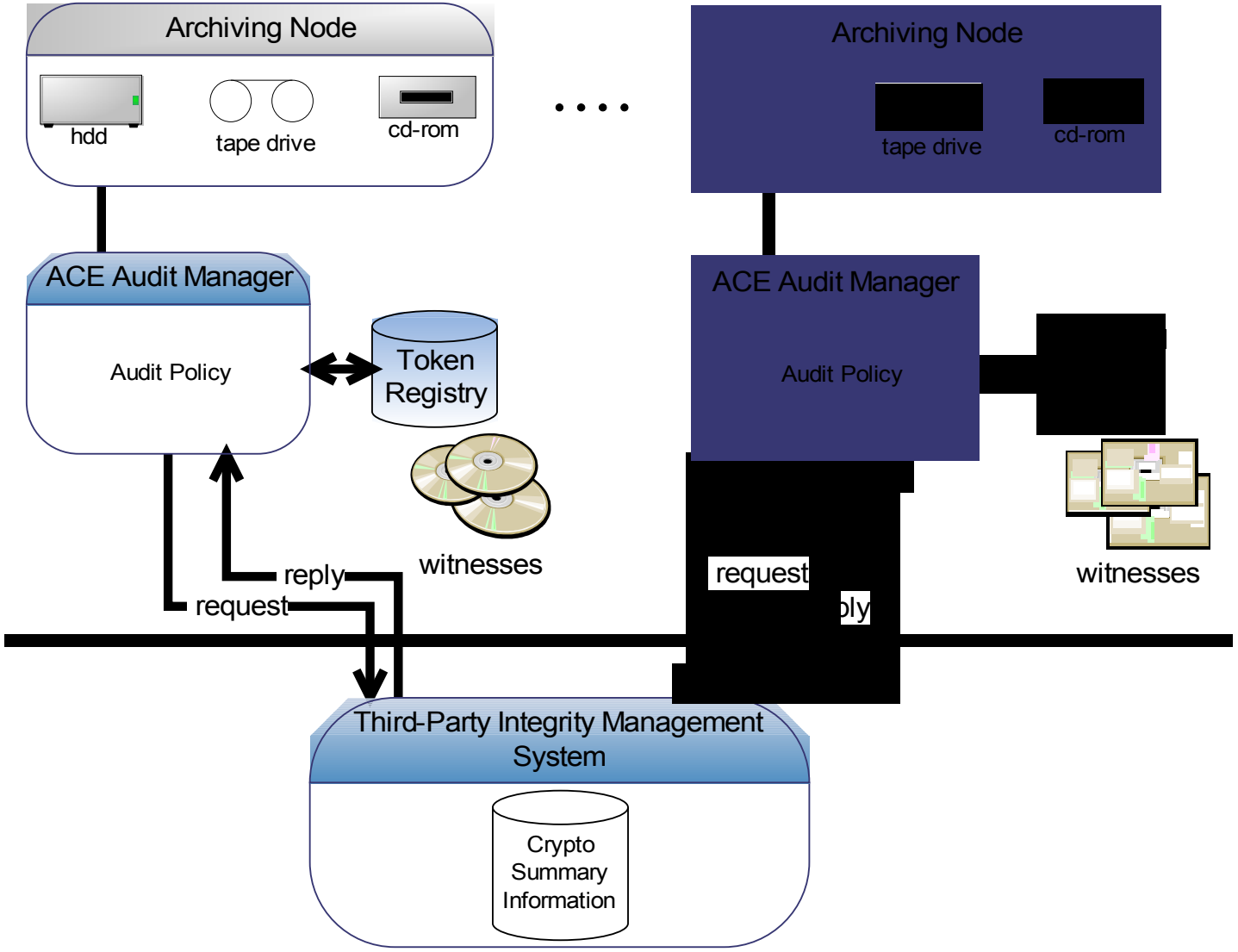
ACE – Auditing Control Environment

- Software to protect the integrity of digital assets in the long term
  - Hardware/media degradation
  - Security breaches, malicious alterations
  - Infrequent access to most data
  - Evolution of cryptographic schemes
- Underpinnings are based on rigorous cryptographic techniques.
- Scalable, cost-effective, and can interoperate with any archiving architecture.

ACE – Basic Methodology

- Builds on cryptographic hashing by introducing additional layers of trust.
  - Layers of cryptographic summary information
- Is not confined to the local processes of the archive, and assumes a third-party, which is not fully trusted.
- An independent party can assert the correctness of any object in the future based on the archive's information and publically available information.

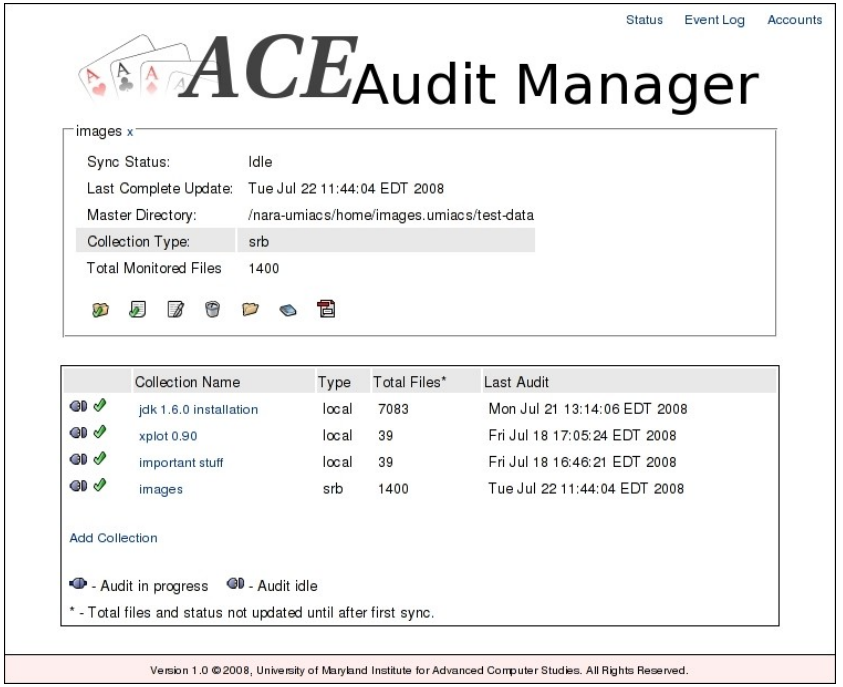
ACE – System Architecture



Components

- IMS – issues tokens for hashes that are to be monitored.
  - WSDL available
  - Java API for bulk operations (uses WSDL)
- Audit Manager(s) – Local, per-archive installations. Monitor bitstreams locally
  - May be independent or part of larger software

Audit Managers



ACE Audit

- **Audit Local Files**: Audit Manager periodically scans all files and compares stored digests with computed digests.

- **Audit Local Manager**: Manager computes round summary for each digest using that digest and its token. This is compared to value stored on the IMS.

- **IMS Audit**: Round summaries are used to compute witness values. These are compared with offsite witness values.

ACE Summary

- TPAP
  - Audit 1.1Tb of images
  - 1.5+ million small files (1.2Tb)
  - Single portal for collections on disk, SRB, iRODS
- Chronopolis
  - 3 Collections
  - 5+ million files, 12.2Tb total
- High performance, Scalable
- Version 1.0 publically available
  - http://adapt.umiacs.umd.edu/ace

Tracking and Replication Monitoring

- Portal that provides overview of a collection status over different zones.
- Ensures that new objects are replicated to relevant sites.
- Tracks files at master locations and periodically copy new files to replica sites.
- Log actions on a collection and errors during replication

Web Archiving: Compact Storage and Fast Retrieval

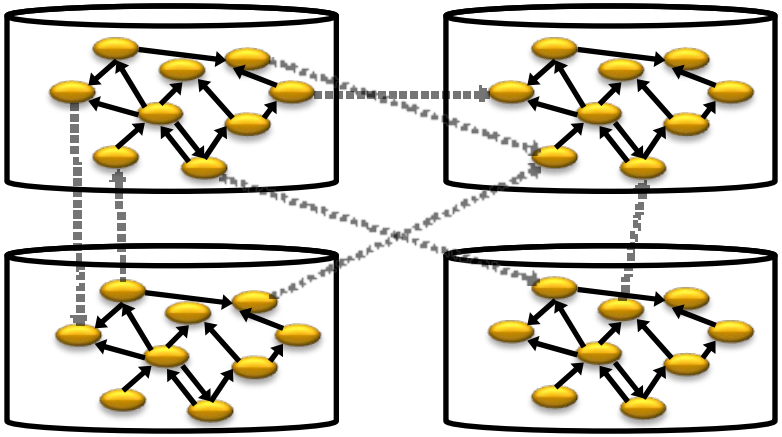
- New technology for storing and indexing web archives.
- Uses standard web containers (WARC) and stores unique contents – detect duplicates before storage.
- Indexing structure based on advanced multiversion B-trees.
- Significantly improved storage and performance over existing technologies.

Scalable Technology for Information Discovery of Web Archives

- Allows discovery through a combination of words and time spans.
- Efficient for handling temporal queries rather than “search and then filter”:
  - “Retrieve documents containing September 11 which were written before 2001”
- Returned web links are ranked according to an appropriate scoring function.
- Allows the possibility of coalescing similar versions of a web page.

Organization of Archived Web Contents

- Efficient browsing of archived web contents based on web graph analysis and graph partitioning techniques.
- Archived web contents are organized into web containers using standard WARC formats.



Other Technologies

- PAWN – Related:
  - APIs for different packaging technologies (METS and XFDU).
  - ICDL Book Builder – Interface to enable bulk ingestion of digital objects already managed by a database.
- FOCUS (FOrmat CUration Service): a scalable, and secure registry for persistent information and services applied to formats.

Conclusion

- Initial effort started through an ERA project, which has grown substantially over the last few years.
- Focus has been on platform and architecture – independent tools and services that are scalable and cost effective.
- Empirical testing and evaluation using a wide variety of NARA and NDIIPP collections and different infrastructures.
- Partnerships have played a crucial role.
- http://adapt.umiacs.umd.edu