

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

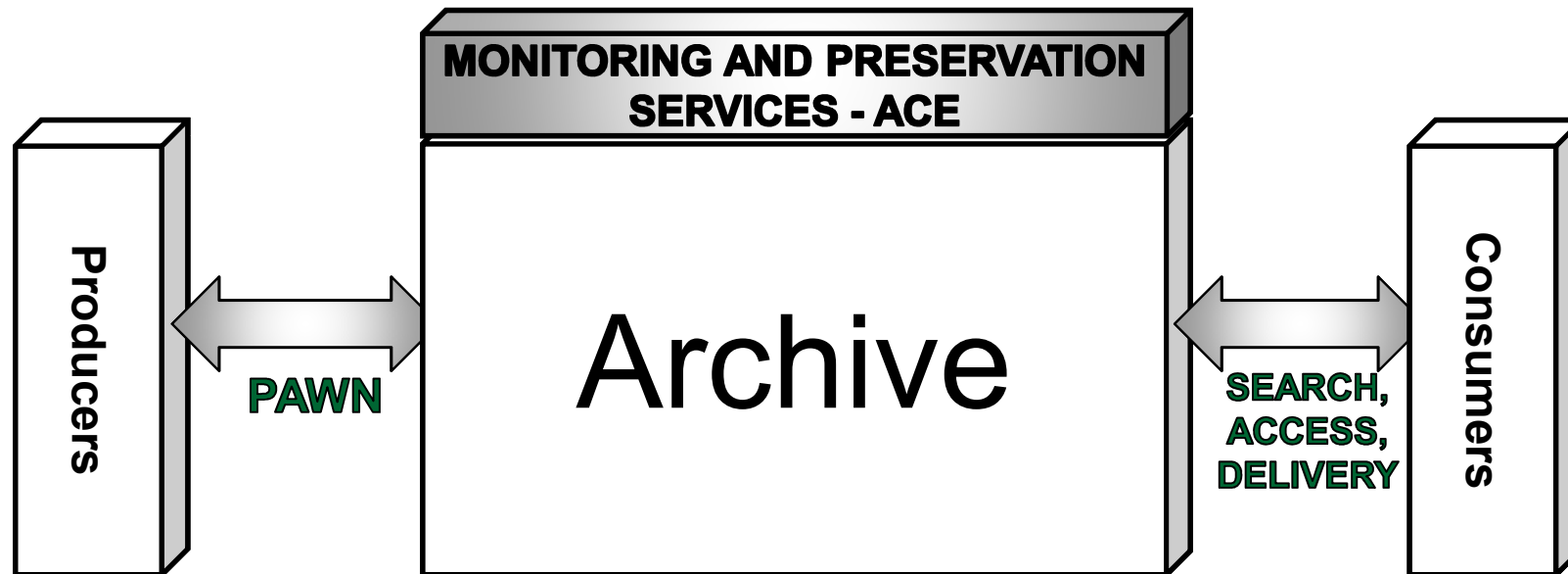
Joseph JaJa, Mike Smorul, and Sangchul Song
Institute for Advanced Computer Studies
Department of Electrical and Computer Engineering
University of Maryland, College Park

Research Goals

- Development of flexible, platform-independent modular tools and technologies, specifically targeting long term preservation and access.
- Evaluation and demonstration on different architectures (centralized and distributed) and large scale heterogeneous collections (government records, scientific data, web data, images, etc.)

Flexible, Layered, OAIS-Compliant Architecture

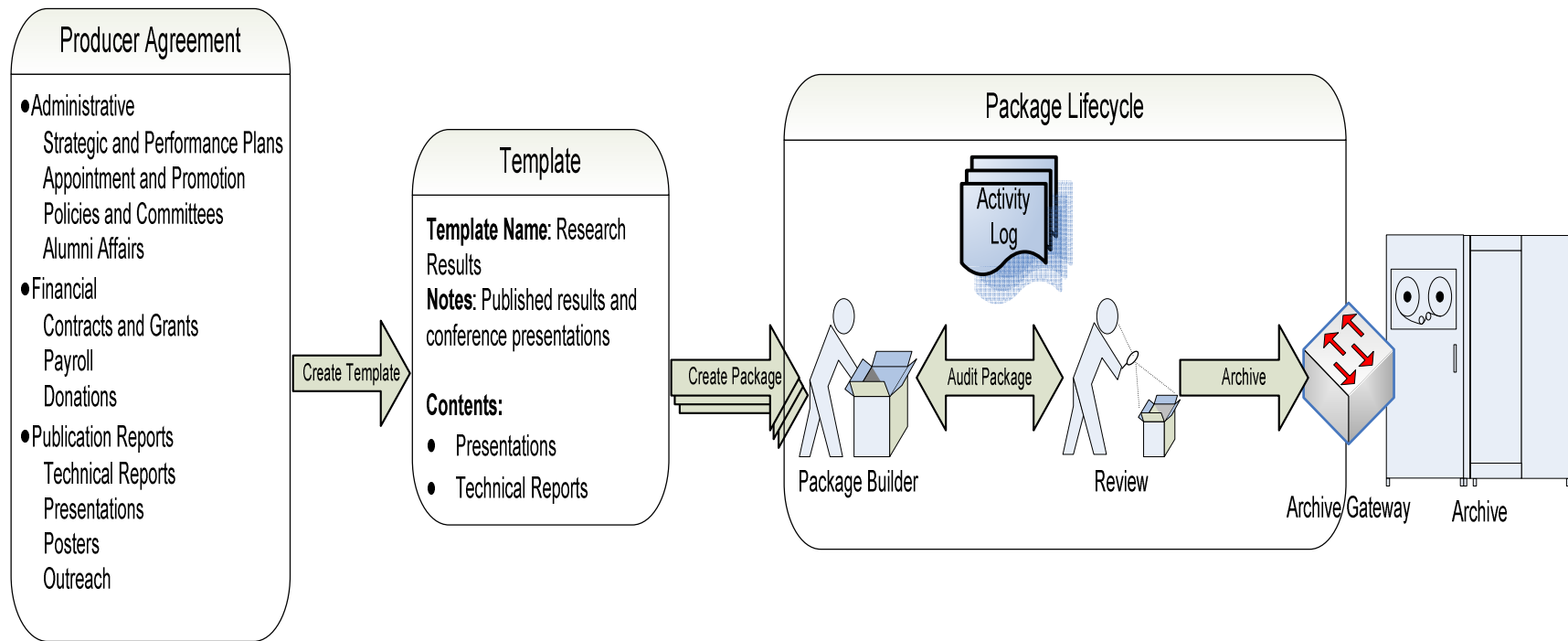
Open Standards, Platform-Independent Components



PAWN – Producer Archive Workflow Network

- Software that provides a flexible and customizable distributed ingestion framework
- Automate the process of submitting or pulling data into an archive in multiple contexts.
- Handles the process in a reliable and secure fashion:
 - From package assembly
 - To archival storage
- Simple interface for end-users and archive managers

Package Workflow Overview



Sample End to End workflow

- On ingest, filtering processes run to perform basic validation.
 - Ensure file is valid and virus free
 - EXIF metadata extraction
 - All approved files are automatically pushed to archival storage
- Archivist may login and handle rejected or non-archived items
- Manual process is invoked to force files to archival storage

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 Audit Manager


[Status](#) [Event Log](#) [Accounts](#)









ACE Audit Manager


ICPSR - UMIACS x

Audit Status: Idle
Last Complete Update: Sun Feb 08 10:12:39 EST 2009
Directory: /chron-umiacs/home/srbChron-umiacs.umiacs/icpsr
Collection Type: srb
Total Monitored Files 4830625

 more...

	Collection Name	Type	Total Files*	Last Audit
 	CDL - UMIACS	srb	46762	Wed Feb 11 21:29:06 EST 2009
 	sio-gdc - UMIACS	srb	197718	Mon Feb 09 16:55:12 EST 2009
 	NC State	srb	608424	Wed Feb 11 14:33:48 EST 2009
 	ICPSR - UMIACS	srb	4830625	Sun Feb 08 10:12:39 EST 2009

[Add Collection](#)

 - Audit in progress  - Audit idle

* - Total files and status not updated until after first sync.

Version 1.2 © 2009, University of Maryland Institute for Advanced Computer Studies. All Rights Reserved. [ACE Website](#)

ACE Summary

- Software to track the availability and integrity of the archive's data holdings.
 - Auditing – local service to periodically verify integrity of files
 - Hash integrity – remote, auditable service to secure hash
 - Extensively tested – main bottleneck is network and I/O bandwidth.
- Chronopolis Environment (NDIIPP Project)
 - Multiple Heterogeneous Collections
 - 5+ million files, 13TB
- High performance, Scalable
 - A single manager can audit over 6 million files a day
- Version 1.2 available























































Tracking and Replication Monitoring

- Portal that provides overview of the status of all the collections in the archive.
- Enforces policies regarding availability and replication.
- Tracks files at master locations and periodically copy new files to replica sites.
- Log actions on a collection and errors during any processing.
- Currently, incorporated with ACE.



Replication Monitor

Tracking and Replication Monitor

Status Event Log Accounts

	Collection Name	Total Files*	
 	cdl		   
 	sio-gdc		   
 	icpsr-AIP		   
 	icpsr-Archive0		   
 	icpsr-GUIDE		   
 	icpsr-TIGER		   
 	icpsr-database		   
 	icpsr - dlib		   
 	NC State		   

[Add Collection](#)

 - Synchronization in progress  - Synchronization idle

* - Total files and status not updated until after first sync.

Version 2.4 (8/08) © 2008, University of Maryland Institute for Advanced Computer Studies. All Rights Reserved.

Access Technologies for Long Term Archives

- Search and information discovery within a temporal context.
- Content exploration to enable knowledge discovery.
- Test and validation on significant scale web archives (in collaboration with the Library of Congress and the Internet Archive.)

Scalable Technology for Information Discovery of Web Archives

- Allows discovery through a combination of words and time spans. Web objects are ranked within a temporal context.

[September 11 attacks - Wikipedia, the free encyclopedia](#)

The **September 11** attacks (often referred to as nine-eleven, written ...
en.wikipedia.org/wiki/September_11,_2001_attacks

[September 11 Digital Archive](#)

to collect, preserve, and present the history of the **September 11**, ...
911digitalarchive.org/

... and 4 million other pages pertaining to the 9/11 Attack ...



[Ethiopian calendar - Wikipedia, the free encyclopedia](#)

Thus the first day of the Ethiopian year, 1 Mäskäräm, for years ...
en.wikipedia.org/wiki/Ethiopian_calendar - 43k

... and only 560 other pages that are irrelevant to the 9/11 Attack

More in Archiving'2009

Conclusion

- Focus on platform and architecture – independent tools and services that are specifically tailored to handle core issues in long term archiving.
- Empirical testing and evaluation using a wide variety of collections and different infrastructures.
- Released tools to manage distributed ingestion, monitoring, and integrity of archived objects.