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

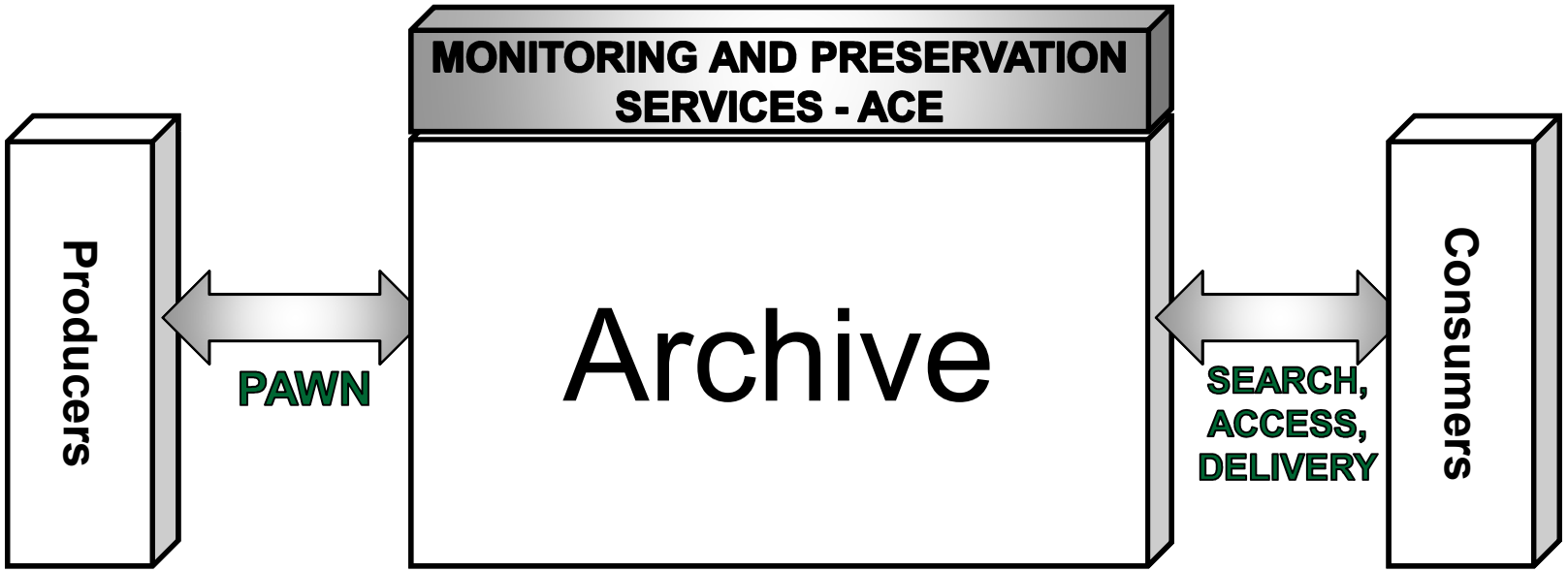
Joseph JaJa, Mike Smorul, and Sangchul Song
Institute for Advanced Computer Studies
Department of Electrical and Computer Engineering
University of Maryland, College Park

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

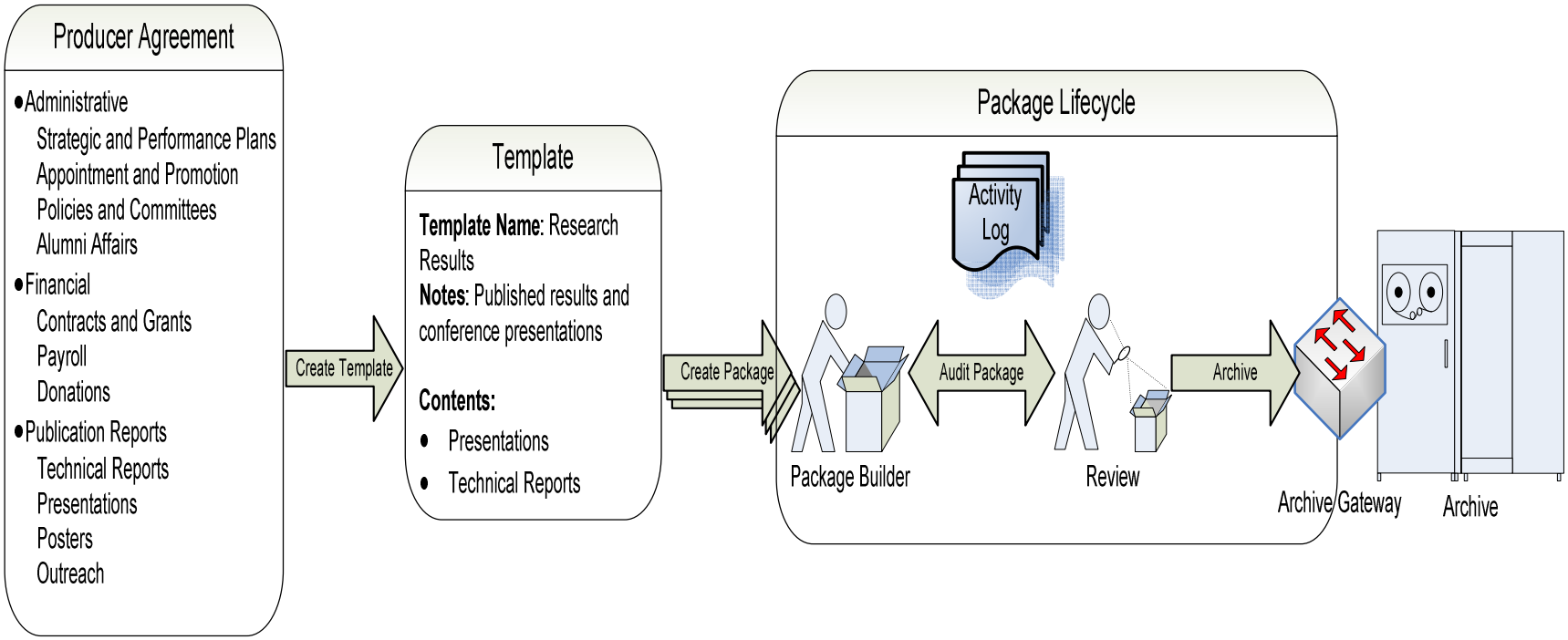
Open Standards, Platform-Independent Components



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



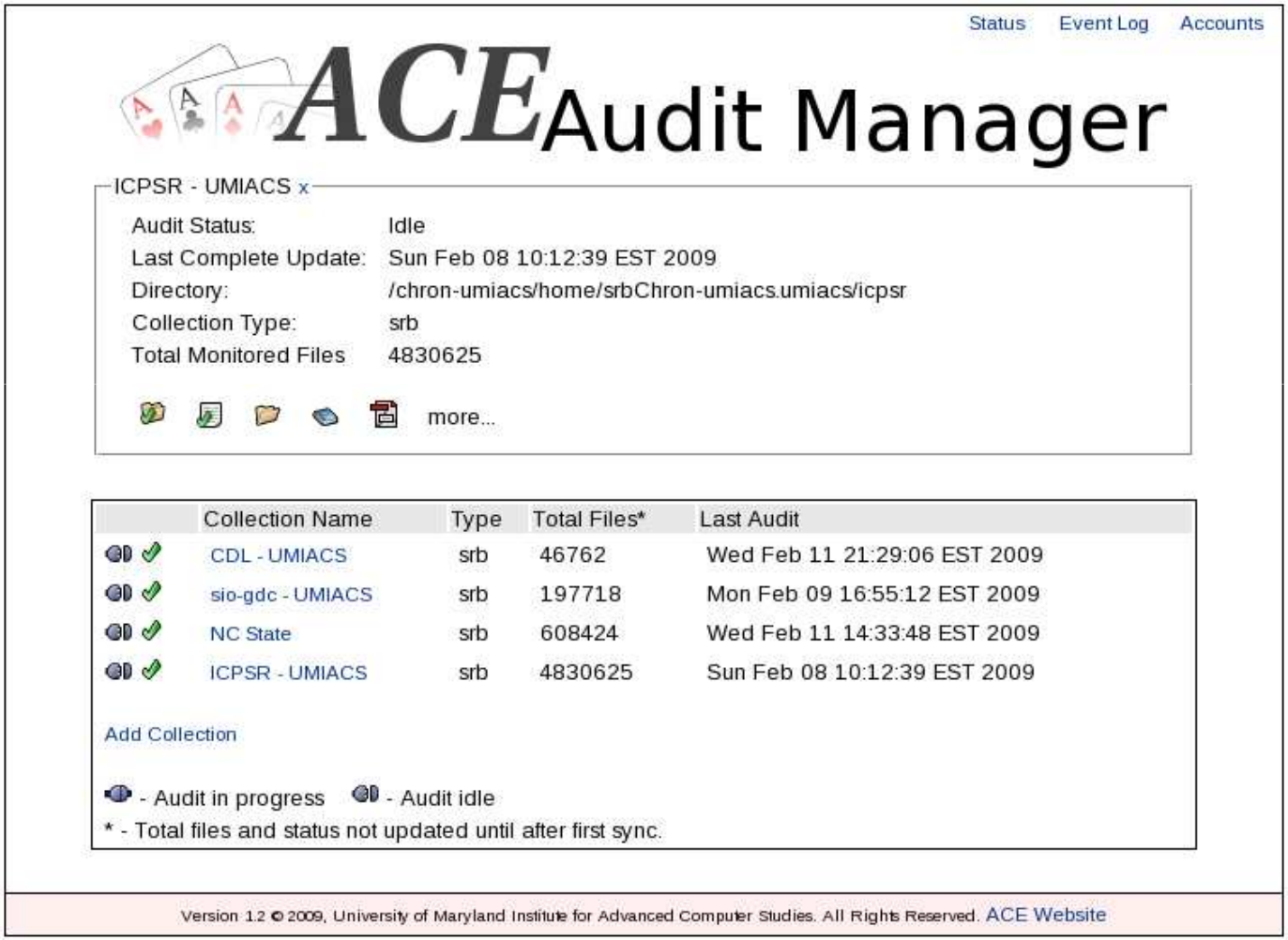
Sample End to End workflow

- 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



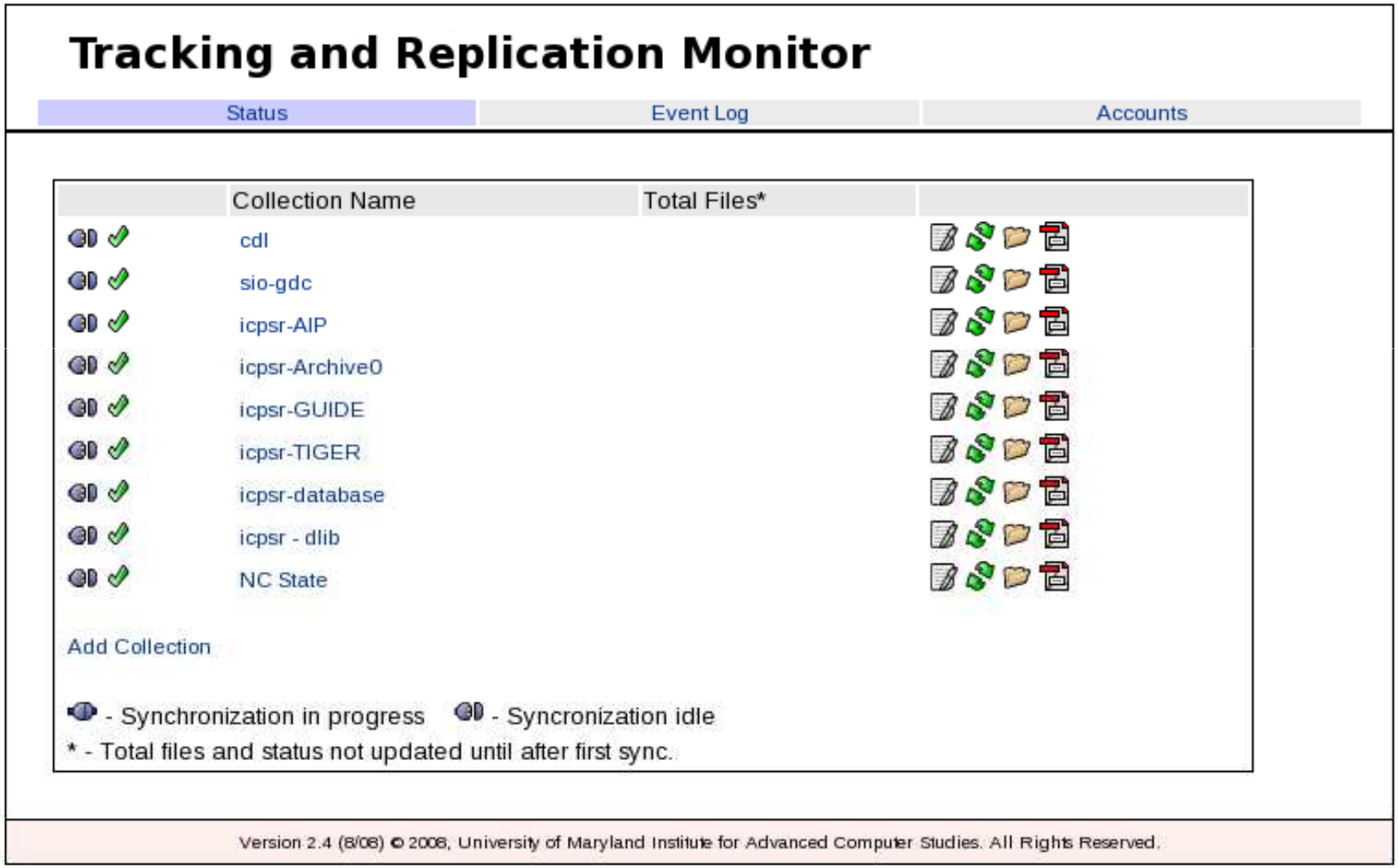
ACE Summary

- 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

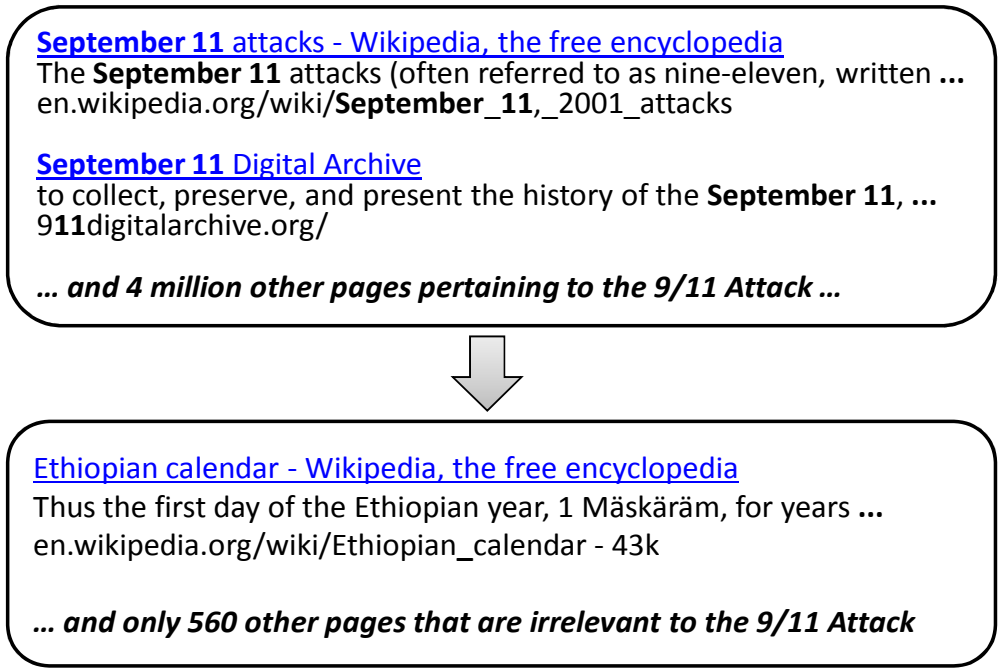


Access Technologies for Long Term Archives

- 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.



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.