

ADAPT

An Approach to Digital Archiving and Preservation Technology

Principal Investigator: Joseph JaJa

Lead Programmers: Mike Smorul and Mike McGann

Graduate Students: Sang Song and Muluwork Geremew

Institute for Advanced Computer Studies

University of Maryland, College Park

Research Objectives

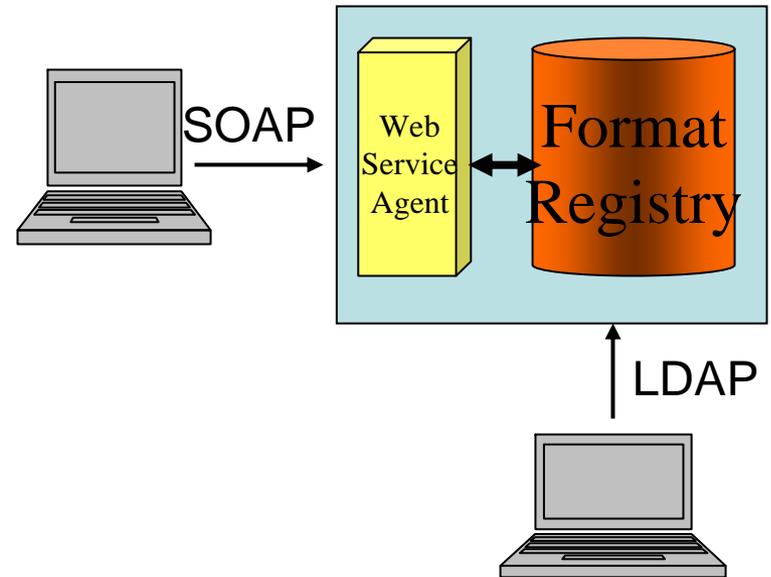
- Development of tools and technologies for:
 - Automated Distributed Ingestion – flexible platform for Producer-Archive Interactions
 - Management of Preservation Processes
 - Monitoring, Integrity Auditing, and Preservation Services.
- Evaluation and demonstration of tools on widely different collections.

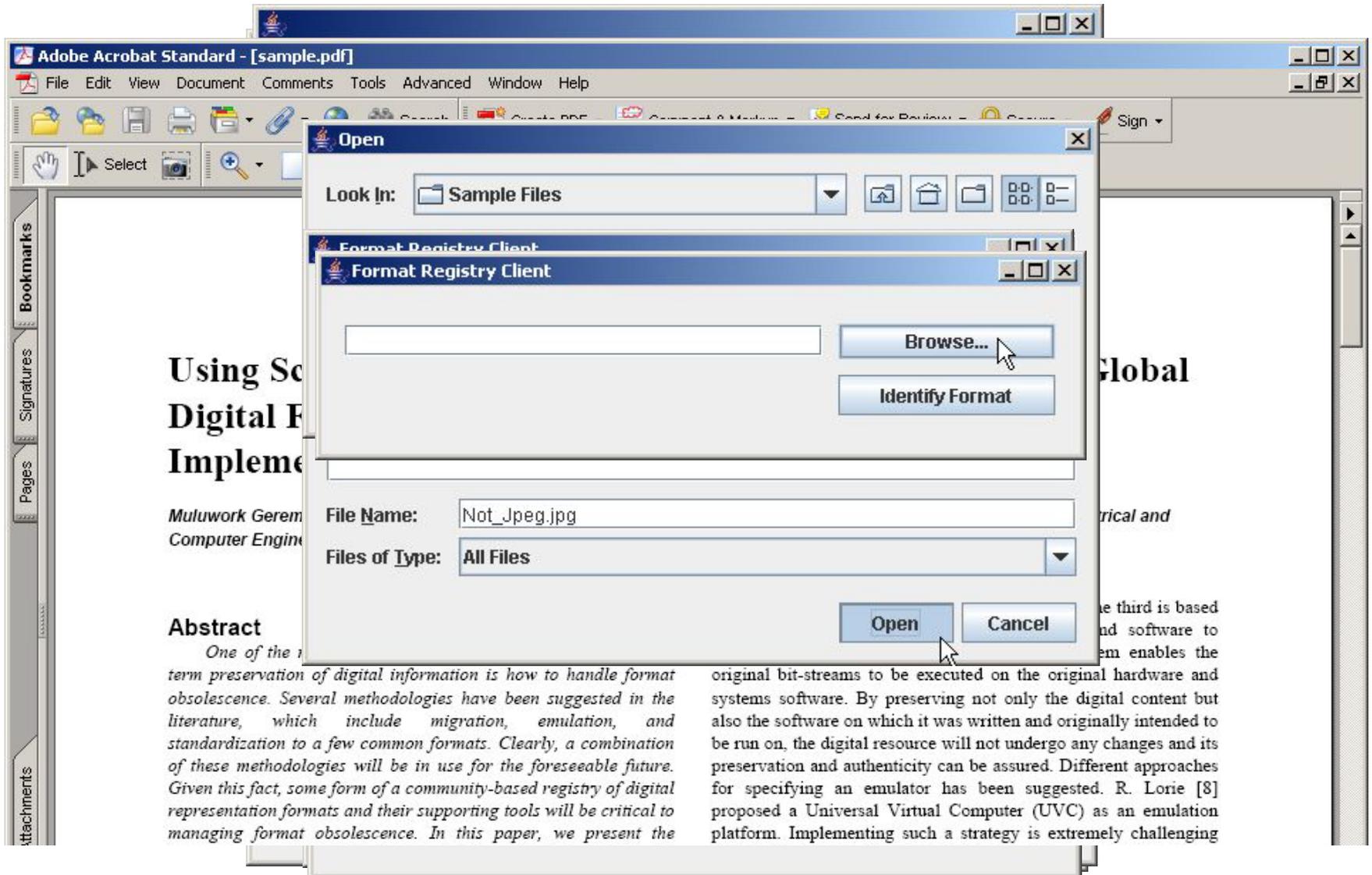
Recent Major Accomplishments

- **FOCUS** – a scalable, and secure registry for persistent information and services applied to formats.
- **ACE (Auditing Control Environment)** - a policy-driven software environment to continually verify the integrity of an archive's holdings.
- **PAWN** – Producer-Archive Workflow Network software platform for data ingestion.
- **SRB Replication Monitor** – 3rd party replication in a data grid environment

FOrmat CUration Service

- Maintains persistent information on digital formats and applications to access and manipulate them.
- Accessible either
 - Directly through LDAP
 - Or indirectly through SOAP (Web Services)





Using So Digital F Impleme

Muluwork Gerem
Computer Engine

Abstract

One of the most important aspects of digital information preservation is how to handle format obsolescence. Several methodologies have been suggested in the literature, which include migration, emulation, and standardization to a few common formats. Clearly, a combination of these methodologies will be in use for the foreseeable future. Given this fact, some form of a community-based registry of digital representation formats and their supporting tools will be critical to managing format obsolescence. In this paper, we present the

Global

ritical and

the third is based
and software to
em enables the

original bit-streams to be executed on the original hardware and systems software. By preserving not only the digital content but also the software on which it was written and originally intended to be run on, the digital resource will not undergo any changes and its preservation and authenticity can be assured. Different approaches for specifying an emulator has been suggested. R. Lorie [8] proposed a Universal Virtual Computer (UVC) as an emulation platform. Implementing such a strategy is extremely challenging

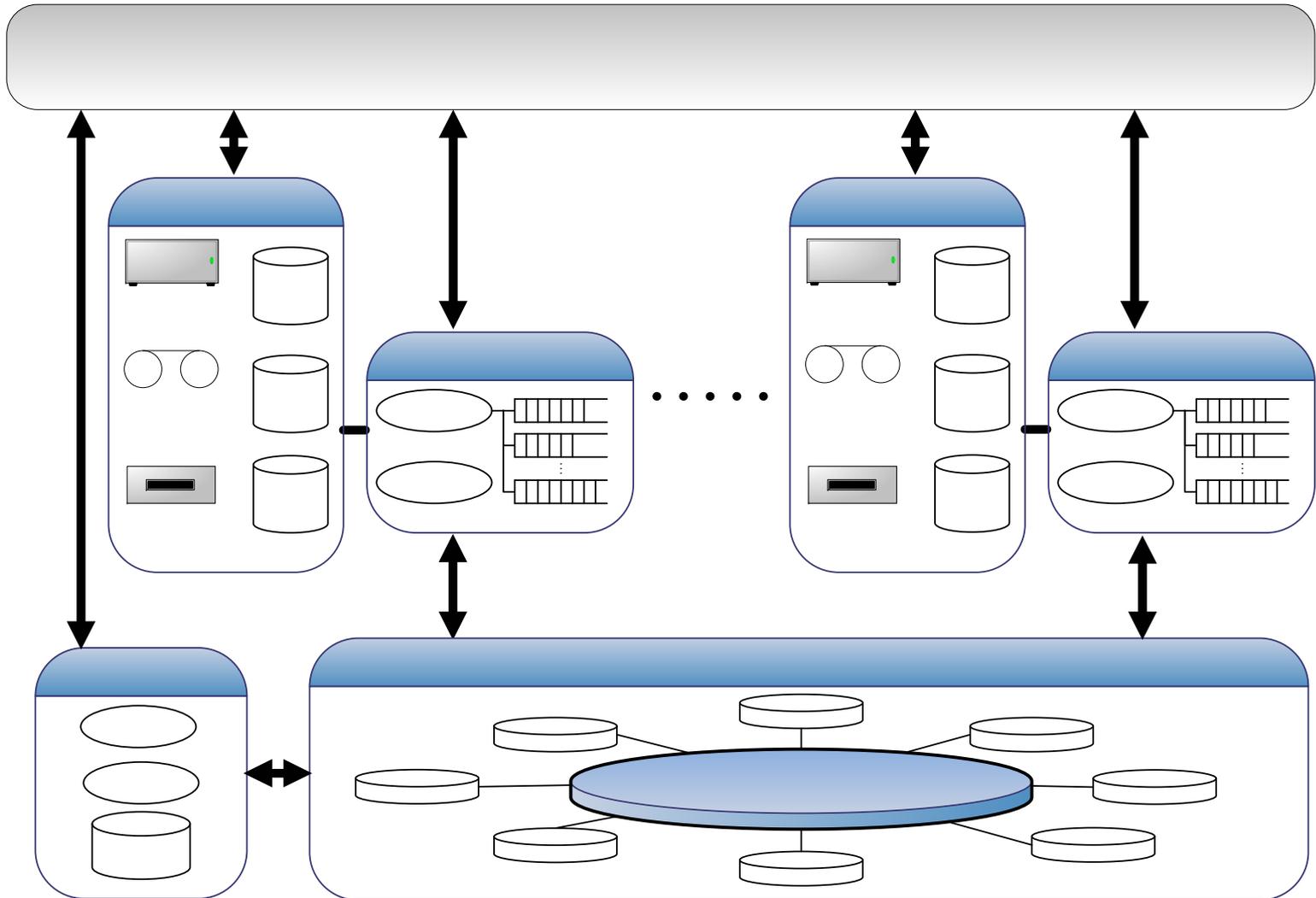
Integrity Auditing Service

- Many types of errors:
 - Media or hardware degradation
 - Technology evolution/upgrades
 - Operational errors
 - Malicious alterations
 - Hardware/software malfunctions
 -
- Digital objects are subject to transformations and changing standards/protocols.

Basic Ideas

- Auditing service is managed and run independently of the archiving system.
- Active and user-triggered auditing.
- Time-stamped certificates that enable the verification of the integrity of the object throughout its lifetime – auditable record of every transformation.
- Highly available and secure service with the ability to detect and correct errors.

Overall Structure



Software Components

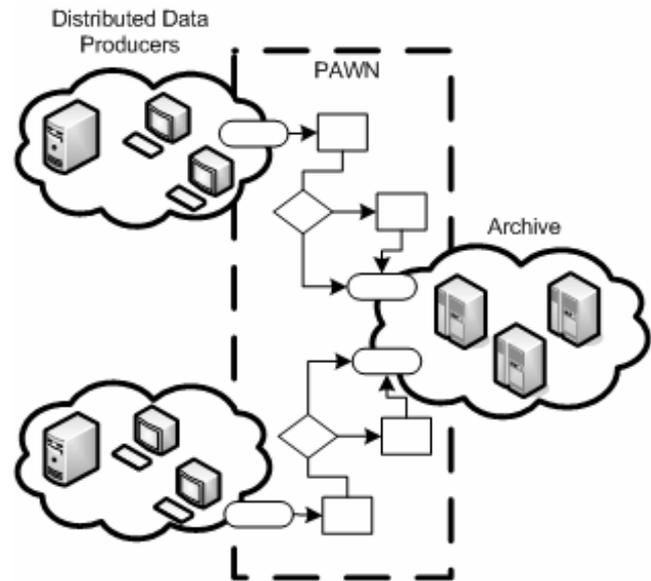
- **Audit Manager:** registers objects to be audited, and performs auditing either actively or as triggered by user/archive.
- **Certificate Management System:** An independent, highly available, and highly secure environment for preserving and ensuring the integrity of the certificates.
- **Replica Monitor:** Verifies the availability of the data in the archive using the object ids in the CMS.

PAWN

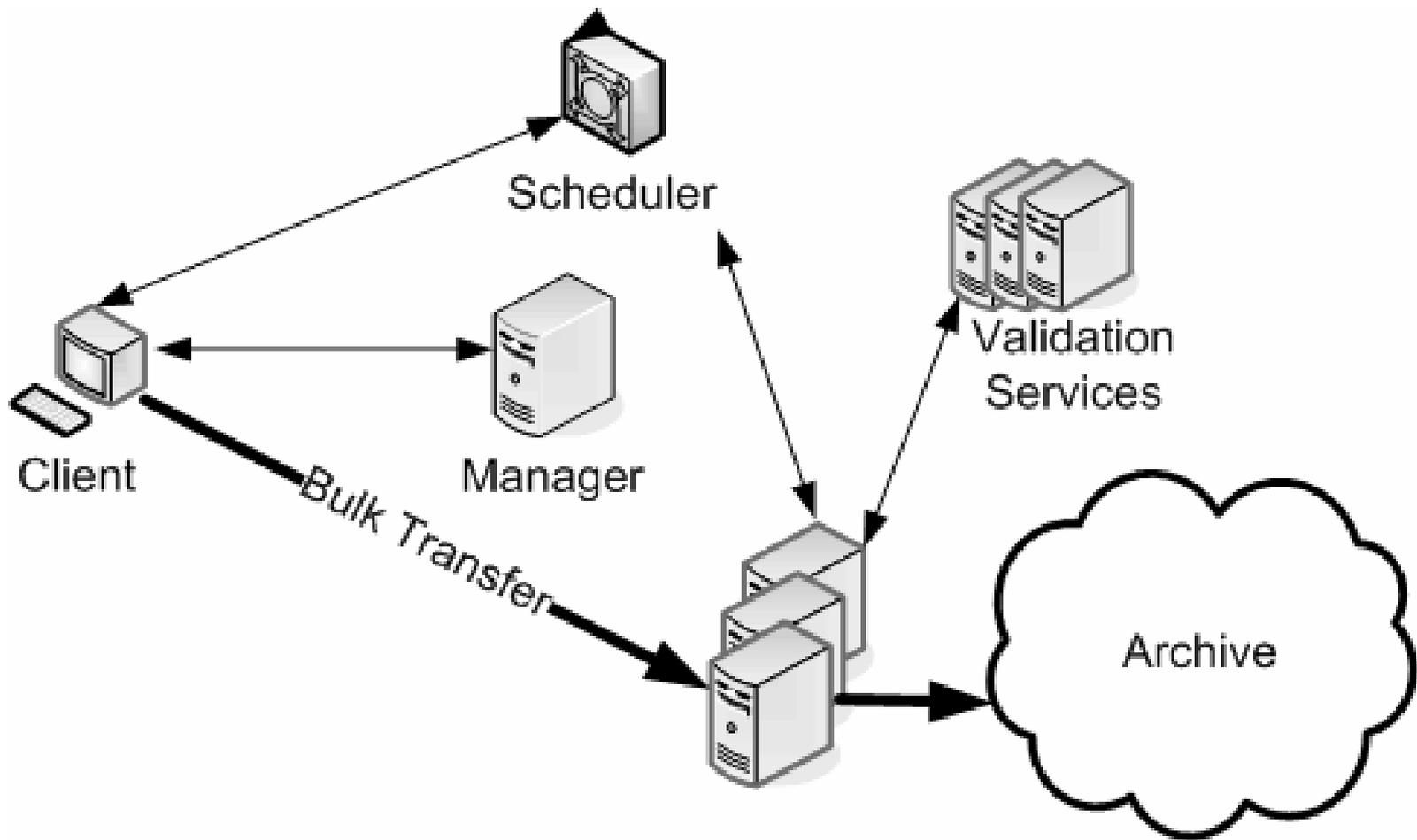
- Flexible platform for creating custom package ingest workflows.
- Handle complex interactions while providing simple end-user ingestion.
- Accountability of transfer and guarantee of data integrity.
- Scalable infrastructure.

Distributed Ingestion with PAWN

- Multiple producing sites with different requirements.
- Separation of administrative responsibility.
- Customizable roles for various parties.



Components

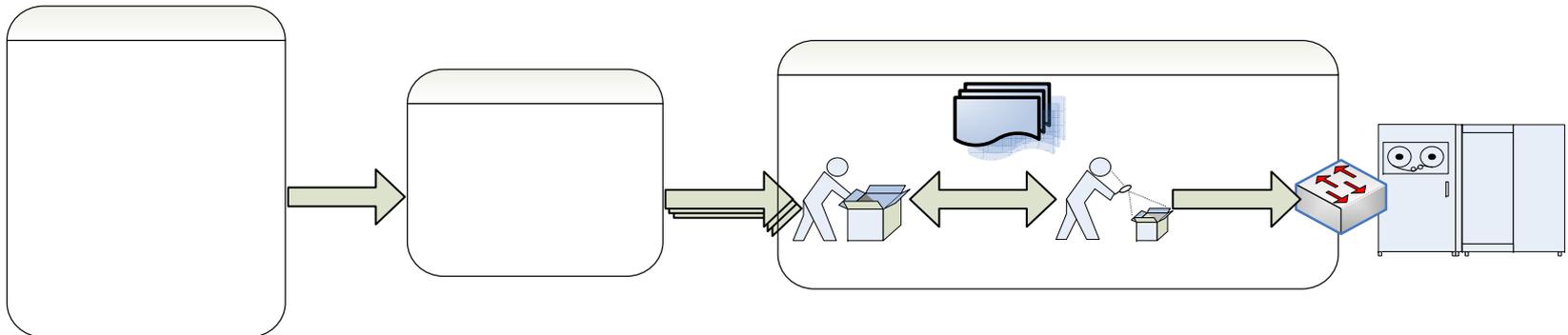


Software Components

- **Management Servers** – Track administrative functionality and high level package details for a set of domains.
- **Scheduler** – Allocate resources from receiving servers for client packages
- **Receiving Server** – Holding pool for packages in pawn, handles 3rd party package operations.
- **Client** – Creates packages and submits to receiving server.

Package Workflow Overview

1. Create Producer-Archive Agreement
2. Client package template.
3. Create package based on template
4. Once approved, packages can be archived
5. Rejected packages can be held until rectified or deleted for resubmission.

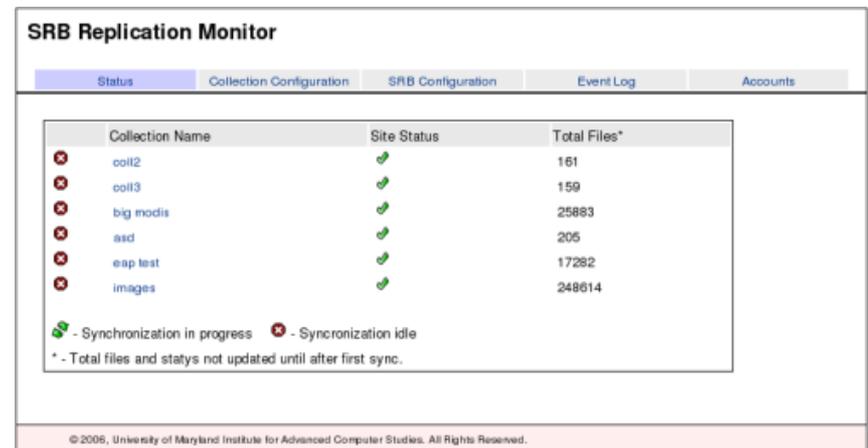


Extensible Platform

- Customizable roles for ingestion.
 - Arbitrary grouping of actions within PAWN.
- API for creating custom clients.
 - Hierarchical package building.
 - PAWN handles transport and tracking.
- Pluggable modules for communicating with various archive resources

Replication Monitoring

- Automatically synchronize collections between master and mirror sites.
- Log any actions or anomalies.
- Support multiple collections.



The screenshot displays the SRB Replication Monitor interface. It features a navigation bar with tabs for Status, Collection Configuration, SRB Configuration, Event Log, and Accounts. The main content area contains a table with the following data:

Collection Name	Site Status	Total Files*
coll2	✓	161
coll3	✓	159
big modis	✓	25883
asd	✓	205
eap test	✓	17282
images	✓	248614

Legend:  - Synchronization in progress  - Synchronization idle
* - Total files and statys not updated until after first sync.

© 2006, University of Maryland Institute for Advanced Computer Studies. All Rights Reserved.

Replica Monitor Demonstrations

- Transcontinental Persistent Archive Prototype
 - 5.5million files between UMD, Archives I and Archives II
 - 1.2Tb image collection between UMD and SDSC
 - Select collections replicated to US Navy, Defense Logistics Agency
- Chronopolis testbed
 - >5Tb replicated monitored between SDSC, UMD, NCAR

Conclusion

- Research program focusing on tools and environments for ingestion, management of preservation processes, and in the near future access for long term digital archives.
- Software prototyping and testing on a wide variety of collections that are available locally.
- Tools to be used by the Chronopolis Consortium, NARA, and NDIIPP partners.