

## RESEARCH BRIEF

# PAWN: PRODUCER-ARCHIVE WORKFLOW NETWORK

### The challenge

Federal agencies churn out uncounted millions of documents, all of which must be cataloged, stored, made available for sharing, converted to electronic format in most cases, destroyed or preserved in perpetuity, as regulations may require.

The National Archives and Records Administration's (NARA) challenge is to identify stable formats and advise agencies on best practices and standards to ensure that important records are accessible to agencies for as long as needed and to researchers, if determined to be permanently valuable and worthy of preservation in the National Archives.

Eventually, all federal records will be tied into the Electronic Records Archive (ERA)—a system for preserving records and making them easily available. ERA will be deployed in September 2007 when four agencies—the Patent and Trademark Office, Bureau of Labor Statistics, Navy Oceanographic Office and National Nuclear Security Administration—will test its effectiveness.

But those goals are complicated by the same advances in technology that are meant to enable them. Ever-changing computer systems and formats frustrate the government's efforts to create, archive and document content. The sheer volume of records—both existing records and those being created every minute—is also part of the challenge, as are the numerous formats in which records are created and the resultant lack of compatibility.

### The research

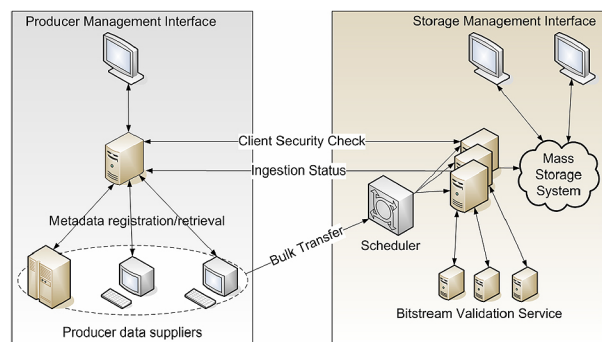
ISR-affiliated Professor Joseph Jájá (UMIACS/ECE) is at the forefront of helping NARA preserve federal records. His team's Producer-Archive Workflow Network (PAWN) software will enable agencies to submit their records remotely, securely, easily and with the assurance that it is transmitted correctly. PAWN is a component of Jájá's larger ADAPT project, an approach to digital archiving and preservation technology, and was funded by NARA.

PAWN takes advantage of emerging technologies in high speed network attached storage (NAS), hierar-

chical storage management systems, and networked systems that virtualize interconnected storage over IP and fiber-channel networks. These advances promise to consolidate distributed data stores onto large-scale professionally managed enterprise storage environments.

PAWN was originally developed to capture core elements required for long term preservation of digital objects as identified by researchers in the digital library and archiving communities. PAWN also can be extended—as is the case with the ERA system—to enable multiple clients at a number of distributed sites to prepare, organize, and bulk transfer large scale data onto clusters of servers that securely verify the integrity of the data, register metadata, and store the data in an enterprise storage environment.

PAWN allows detailed description, auditing, and organization of the data, and soon will allow for efficient management, access and disaster recovery. The basic software components are based on open standards and web technologies, and are platform independent.



Software components in PAWN.

### Inside PAWN

PAWN allows efficient staging, arrangement, and assembly of data at various sites, followed by parallel bulk transfers into receiving servers whose loads are managed by a scheduler. The receiving servers validate data and metadata, organize the data into collections, and register the metadata into a unified metadata database before storing it in the enterprise storage management environment.

PAWN's software components include a management server and clients at each producer, as well as verification services and receiving servers connected to mass storage systems.

Because a number of people at each producer are engaged in preparing and transferring data to receiving servers, the management server acts as a central point for the initial organization of the data, and for tracking bitstreams and metadata functionality. The server:

- Provides the necessary security infrastructure to allow secure transfer of bitstreams between the producer and the receiving site.
- Assigns a unique identifier for each bitstream to be transferred.
- Provides an interface for bitstream organization and metadata editing.
- Accepts checksums/digital signature, system metadata and other client supplied metadata.
- Tracks which bitstreams have been transferred to the receiving servers.

A client runs on each machine to automatically register metadata information and bulk transfer data to the data center. The client:

- Bulk registers bitstreams, checksums and system metadata;
- Assembles a valid data stream;
- Transmits the data stream to the center either directly or through a third party proxy server.

The center's cluster of servers receives data from the producers, managed through a scheduler, accepting data and initiating verification/validation processes on the bitstream. Each receiving server:

- Securely accepts a data stream from clients as assigned by the load balancer;
- Processes data streams and initiates verification/validation processes;
- Coordinates authentication with the management server at the corresponding producer site;
- Verifies with the management server that all data streams have arrived intact; and
- Provides temporary storage for incoming data streams until they can be validated and then transferred to mass storage systems.

## Research team

Joseph JáJá  
Principal Investigator; ISR-affiliated Professor

Mike Smorul, Lead Programmer and Manager, Persistent Archives Project

Mike McGann and Fritz McCall, Programmers

Qingmin Shi, Postdoctoral Associate

## Collaborators

San Diego Supercomputer Center (SDSC) and the National Archives and Records Administration (NARA): Three node servers at SDSC, University of Maryland, and NARA are linked through the SDSC Storage Request Broker (SRB) data grid, managing several terabytes of significant NARA-selected collections.

## Funding

This research is sponsored by the National Archives and Records Administration and the National Science Foundation under the Partnerships for Advanced Computational Infrastructure (PACI) program.

## Contact

**Joseph JáJá**  
ISR-affiliated Professor  
Electrical and Computer Engineering  
and the Institute for Advanced Computer Studies  
3433 A.V. Williams Bldg.  
University of Maryland  
College Park, MD 20742

Phone: 301.405.1925 Fax: 301.405.6707

Email: [joseph@umiacs.umd.edu](mailto:joseph@umiacs.umd.edu)

## For more information

Dr. JáJá's home page: [www.isr.umd.edu/faculty/gateways/jaja.htm](http://www.isr.umd.edu/faculty/gateways/jaja.htm).

ADAPT web site: [umiacs.umd.edu/research/adapt/](http://umiacs.umd.edu/research/adapt/).

PowerPoint presentation on PAWN: [umiacs.umd.edu/research/adapt/papers/erpanet.ppt](http://umiacs.umd.edu/research/adapt/papers/erpanet.ppt).

Story about PAWN at *Federal Times* website: [federaltimes.com/index.php?S=2841121](http://federaltimes.com/index.php?S=2841121)