

Building a Reference Implementation of a Preservation Environment

REAGAN W. MOORE and RICHARD MARCIANO

Abstract: The Transcontinental Persistent Archive Prototype project (TPAP) is sponsored by the National Archives and Records Administration, and is researching the development of a reference implementation of a preservation environment. The approach is based on the use of the integrated Rule-Oriented Data System (iRODS) data grid to enforce and validate trustworthiness assessment criteria. We will present an analysis of the required components, discuss the current state of implementation, and identify the areas where substantial research is still needed. This project integrates technology from the San Diego Supercomputer Center (iRODS), the Producer-Archive Workflow Network from the University of Maryland, and the SHAMAN project (Sustaining Heritage through Multivalent Archiving) micro-services for parsing documents. The capabilities sustained within the preservation environment are based on the NARA Electronic Records Archives capabilities list, and the assessment criteria are based on the Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist. For each identified capability, the required operations are encapsulated in micro-services that are executed at the storage location, under the control of rules that implement the management policies needed to enforce TRAC criteria. Rules are also defined that periodically query the system to verify compliance, and automate recovery procedures when problems are found. The reference implementation then consists of the record management environment, the preservation management rules, the management processes that implement preservation services, and the rules that verify compliance with assessment criteria.

About the authors:

Reagan W. Moore is Visiting Research Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He was formerly Associate Director for Data Intensive Computing at the San Diego Supercomputer Center. He coordinates research efforts in development of massive data analysis systems, scientific data publication systems, and persistent archives. An ongoing research interest is support for information based data-intensive computing. Moore is an active participant in NSF workshops on digital libraries and Knowledge Networks. Recent publications include a chapter on data-intensive computing in the book "The Grid: Blueprint for a New Computing Infrastructure." Moore has been at SDSC since its inception, initially being responsible for operating system development. Prior to that he worked as a computational plasma physicist at General Atomics on equilibrium and stability of toroidal fusion devices. He has a Ph.D. in plasma physics from the University of California, San Diego, and a B.S. in physics from the California Institute of Technology.

Richard Marciano is Visiting Research Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He was formerly Director of the Sustainable Archives and Library Technologies (SALT) Laboratory and

Lead Scientist in the Data and Knowledge Systems (DAKS) Group at the San Diego Supercomputer Center (SDSC), at the University of California San Diego (UCSD). He is also an affiliate professor in the Urban Studies and Planning Program in the Division of Social Sciences and founding member of the Regional Workbench Consortium (RWBC) at UCSD. The SALT Lab is an interdisciplinary unit focused on developing information technology strategies and conducting research in the area of digital materials and records collection and preservation. Dr. Marciano's interests are in data management, digital archiving, and long-term preservation. Current research projects include InterPARES 2, Persistent Archives Testbed (PAT), Preservation of Electronic Records in an RMA (PERM), and Incorporating Change Management in Archival Processes (ICAP). He holds degrees in Avionics and Electrical Engineering (National School of Civil Aviation, Toulouse, France) and an M.S. and Ph.D. in Computer Science from the University of Iowa; he also worked as a postdoc in computational geography.