

Reagan Moore



San Diego Supercomputing Centre, US

moore@sdsc.edu

Reagan Moore is Director of Data and Knowledge Systems at the San Diego Supercomputer Center (SDSC). He coordinates research efforts in development of data grids, digital libraries and preservation environments. Developed software systems include the Storage Resource Broker data grid and the integrated Rule-Oriented Data System. Supported projects include the National Archives and Records Administration Transcontinental Persistent Archive Prototype, the National Science Foundation National Science Digital Library persistent archive, the California Digital Library Digital Preservation Repository and the Worldwide Universities Network data grid. An ongoing research interest involves the use of data grid technology to automate execution of management policies and validate trustworthiness of repositories.

Moore has been at SDSC since its inception in 1986, 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, (1978) and a B.S. in physics from the California Institute of Technology (1967).

GRL2020 Position Paper

Scientific research projects now rely upon the analysis of large data collections, assembled from simulation output, observational data and experimental data. The research digital holdings require the same attention to provenance, representation information, authenticity and integrity that are needed within digital libraries. To facilitate use by the scientific community, the data needs to be turned into digital reference collections. Assertions about completeness, consistency and authoritative source need to be proven for each collection to establish trust in the data. A major challenge for the digital library community is the ability to curate, manage and preserve massive digital holdings that are petabytes in size, and that comprise hundreds of millions of files. Managing data at this scale requires the automation of not only curation procedures, but also administrative tasks for data management and validation procedures for verifying assertions about the collection. At the petabyte scale, the management of a digital collection is a dynamic process that requires continual monitoring of the digital holdings. A second challenge is that the management of massive collections requires distribution and replication of the digital holdings to minimize risk of data loss. A convergence between three software technologies is occurring: data grids for organizing distributed data into a shared collection; digital library services for curating and displaying data; and preservation environments for managing technology evolution. Future digital libraries need to provide the capabilities of all three systems. A third challenge is the need to free digital holdings from their creation environment. We need the ability to characterize the structures and information content present within a digital record independently of the original creation application. There are

multiple attempts at doing this ranging from the Data Format Description Language effort of the Open Grid Forum, to the Java-based Multivalent parsing technology for office products.

Within the next three years, technology will be available to automate the enforcement of management policies and minimize the amount of labor needed to manage massive data collections. The approach requires the integration of rule engines with storage systems, the mapping of management policies to rules that are enforced directly at each storage resource and the mapping of management procedures to micro-services that are executed at each storage resource under the control of the rule engine. State information is maintained that tracks the outcomes of applying each micro-service. Assertions about collection properties are then mapped to queries on the state information. An example is the iRODS – integrated Rule Oriented Data System. Such environments are capable of enforcing retention/disposition policies, human subject approval flags for Institutional Research Boards, HIPAA patient confidentiality requirements and Trusted Repository Assessment Criteria. A reasonable goal within the coming three years is the ability to define policy sets (rules and associated micro-services) that implement the management policies required by a specific community, enforce the application of these policies, and then automate assessments about desired collection properties. Through use of data grid technology, internationally shared collections can be assembled that are jointly curated by researchers from multiple institutions.

The creation of reference collections of digital holdings requires the participation of science domain experts. Most disciplines are actively engaged in projects that are attempting to define standards for representation information for their scientific domain. This includes definition of standard semantic terms to describe the physical meaning of the data, standard data formats for structuring the data and standard manipulation services for interacting with the data. Each community also defines the management policies controlling access, distribution, retention and review procedures. The process can be accelerated by enabling: Improved representation information for scientific data. A standard syntax is needed for encoding representation information; Composable manipulation routines (micro-services) for parsing data. The ability to assemble processing pipelines from sets of micro-services will enable broader use of scientific data; Improved policy expression languages that map management policies to execution of management processes. This in turn can be expressed in terms of server-side workflows that can be applied directly at storage systems.