Digital Curation and Preservation: Defining the Research Agenda for the Next Decade

Reagan W. Moore
moore@sdsc.edu
http://www.sdsc.edu/srb

Background

- NARA research prototype persistent archive (SDSC, U Md, NARA)
- NSF National Science Digital Library persistent archive
- NHPRC Persistent Archive Testbed
- NSF/Library of Congress Digital Archive
- California Digital Library - Digital Preservation Repository
- UCSD Libraries image archive
- InterPARES VanMap - GIS preservation

Storage Resource Broker 3.3.1

Application

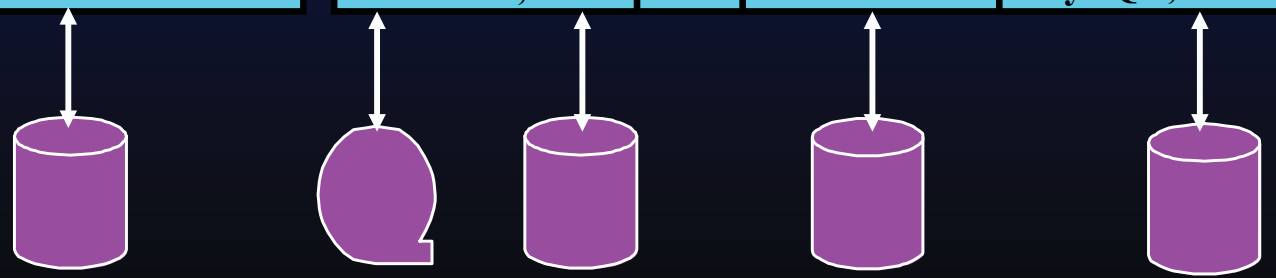
C Library, Java	Unix Shell	Linux I/O C++	NT Browser, Kepler Actors	DLL / Python, Perl, Windows	HTTP, DSpace, OpenDAP, GridFTP	OAI, WSDL, (WSRF)

Federation Management

Consistency & Metadata Management / Authorization, Authentication, Audit

Logical Name Space	Latency Management	Data Transport	Metadata Transport

Database Abstraction	Storage Repository Abstraction			
Databases - DB2, Oracle, Sybase, Postgres, mySQL, Informix	Archives - Tape, Sam-QFS, DMF, HPSS, ADSM, UniTree, ADS	ORB	File Systems Unix, NT, Mac OSX	Databases - DB2, Oracle, Sybase, Postgres, mySQL, Informix



UCSD

SDSC

Preservation

- **Extract record from its creation environment**
  - Authenticity
    - Assertions by record creator
    - Static provenance information
- **Import record into the preservation environment**
  - Integrity
    - Assertions by archivist
    - Dynamic information context, mapping from preservation environment to external changing world
  - Infrastructure independence
    - Assertion that the preservation environment has no dependencies upon a particular proprietary product, format, or protocol

State of the Art

- **Grid - workflow virtualization**
  - Support execution of jobs (processes) across multiple compute servers
- **Data grid - data virtualization**
  - Manage properties of a shared collection that is distributed across multiple storage servers
  - Trust virtualization - manage authentication, authorization
- **Semantic grid - information virtualization**
  - Reason across inferred attributes from multiple collections.

Preservation Environment Reference Model

Preservation management virtualization

Preservation trust virtualization

Preservation workflow virtualization

Preservation data virtualization

Preservation knowledge virtualization

UCSD

SDSC

Management Virtualization

- **Characterization of management policies independently of the implementation**
  - Validation policies
  - Lifetime policies
  - Access policies
  - Federation policies
  - Presentation policies
  - Consistency policies

Trust Virtualization

- **Management of ownership of records independently of storage systems**
  - Collection owned data
  - At each remote storage system, an account ID is created under which the preservation environment stores files
  - Management of roles for permitted operations
  - Management of authentication of users
  - Management of authorization

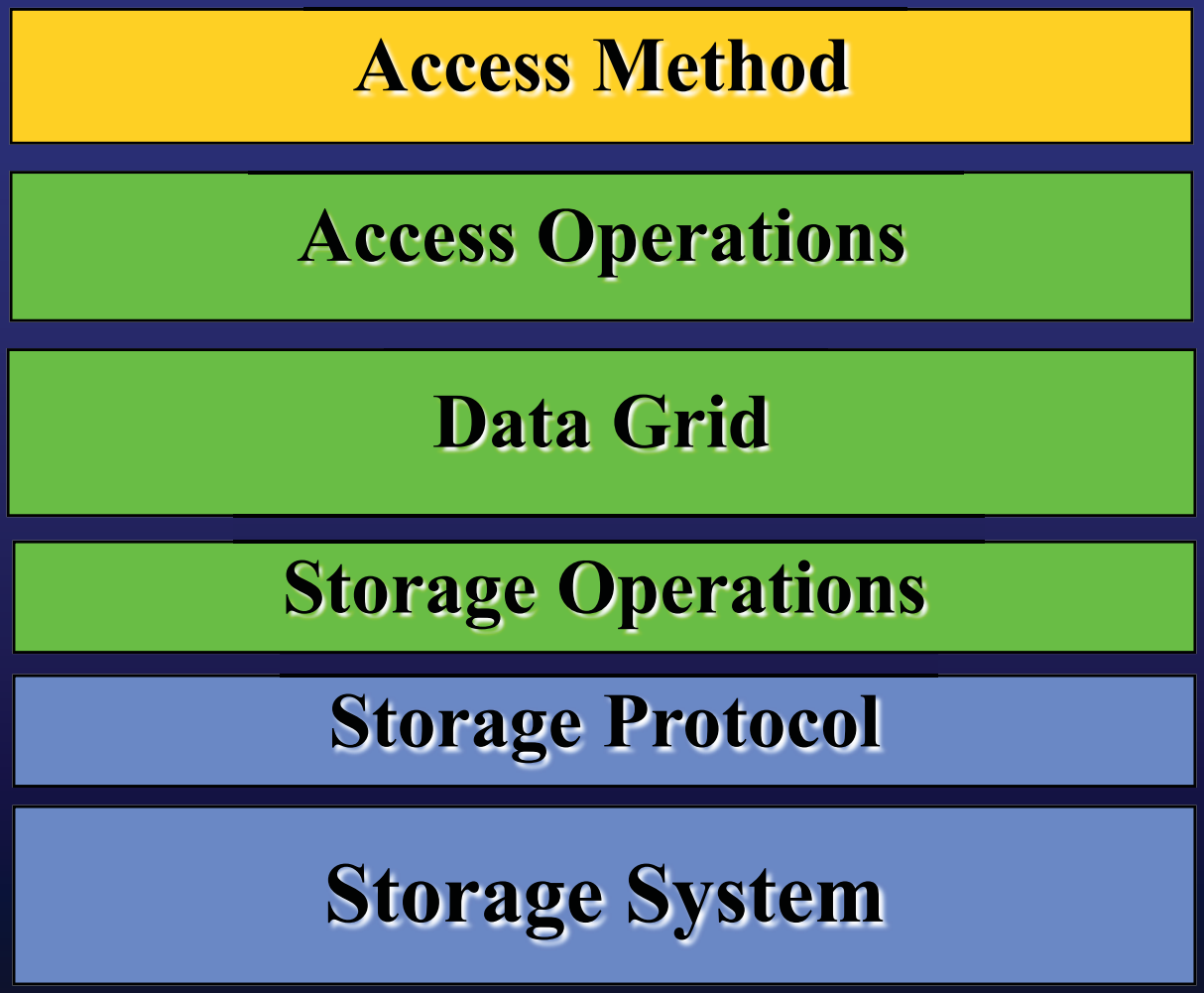
Workflow Virtualization

- **Management of execution of preservation processes across distributed resources**
  - Management of execution state
  - Management of relationships between jobs
  - Management of interactions with remote schedulers

UCSD

SDSC

Data Virtualization



Map from the operations used by the access method to a standard set of operations used to interact with the storage system

Data Virtualization

Data Access Methods (C library, Unix, Web Browser)

Data Collection

Storage Repository

- Storage location
- User name
- File name
- File context (creation date,…)
- Access constraints

Data Grid

- Logical resource name space
- Logical user name space
- Logical file name space
- Logical context (metadata)
- Control/consistency constraints

Data is organized as a shared collection

UCSD

SDSC

Federation Between Data Grids

Data Access Methods (Web Browser, DSpace, OAI-PMH)

Data Collection A

Data Collection B

Data Grid

- Logical resource name space
- Logical user name space
- Logical file name space
- Logical context (metadata)
- Control/consistency constraints

Data Grid

- Logical resource name space
- Logical user name space
- Logical file name space
- Logical context (metadata)
- Control/consistency constraints

Access controls and consistency constraints
on cross registration of digital entities

UCSD

SDSC

Knowledge Virtualization

- **Management of relationships between record components or between records independently of the storage systems**
  - Characterization of hierarchical metadata
  - Logical relationships - OWL
  - Spatial/structural relationships
  - Temporal/procedural relationships
  - Functional relationships

Research

- **Persistent objects**
  - Characterize the structure of digital entities
  - Migrate the characterization forward in time
  - Parse the digital entity based on the characterization
    - DFDL / OpenOffice / Multivalent Browser
  - Separate behaviors from parsing
- **Preservation workflow systems**
  - Management virtualization
  - Automate application of preservation policies
  - Automate migration between preservation environments

Research

- **Preservation environment validation**
  - Consistency checking of semantics and syntax of authenticity and integrity metadata
  - Validation of assertions made about distributed environment

Reference Models

- **OAIS - record handling**
- **Preservation environment infrastructure**
- **Preservation management policies**
- **Preservation utilization properties**
  - Curation
  - Implied knowledge

Evaluation Metric

- **Migrate records into an alternate preservation environment while maintaining authenticity and integrity**
- **Generic infrastructure**
  - Community standards for semantics
  - Community standards for formats
  - Community standards for services

Synergy in Data Management

- **Common requirements across:**
  - Data grids
  - Digital libraries
  - Persistent archives
  - Collection building
  - Analysis pipelines
  - Real-time sensor streams

For More Information

Reagan W. Moore

San Diego Supercomputer Center

moore@sdsc.edu

http://www.sdsc.edu/srb/

UCSD

SDSC