Data Curation and Preservation

R. Moore (SDSC), Françoise Genova CDS), A. Szalay (JHU)
Approach

• Curation
  ▪ Assemble the information and knowledge needed to support future use of a collection of records

• Then Preservation
  ▪ Management of technology while preserving authenticity and integrity
Discipline Expertise

• Standard vocabulary
  ▪ Uniform Content Descriptors

• Standard data format
  ▪ FITS, VOTable

• Standard services
  ▪ SIAP, VOStore

• Build collection - typically done by a data center
  ▪ Authenticity metadata
  ▪ Integrity metadata

• Standard curation and preservation services
  ▪ Workflow for automating application of services
    ▪ Curation of data and metadata
  ▪ Infrastructure independence
Curation versus Preservation

• Convergence of requirements
  ▪ Digital library metadata to support discovery
  ▪ Archivist metadata to support authenticity and integrity

• Automated processing (digital libraries)
  ▪ Extract metadata
  ▪ Register metadata and files
  ▪ Validation of semantics and syntax
  ▪ Validation of integrity
  ▪ Packaging of data and metadata
  ▪ Workflow templates to control curation and archival processes

• Infrastructure independence (persistent archives)
  ▪ Standard operations for interacting with new technology
Curation and Preservation Communities

- IVOA Preservation Interest Group
  - IAU Commission 5 - Data
  - GGF Preservation Environments Research Group
- Astronomy preservation bodies
  - ADS - Bibliographies, references, cross-reference to data
  - CDS - Bibliographies, references, cross-reference to data
  - Project specific archives
  - Agency specific archives
- United States
  - DSpace digital library - MIT
  - Fedora digital library - Cornell University
  - NHPRC InterPARES - International Research on Permanent Authentic Records in Electronic Systems
  - NSF Chronopolis - Federated Digital Preservation across Space and Time
  - Library of Congress NDIIPP - National Information Infrastructure Preservation Program
  - NARA Electronic Records Archive
- European
  - ERPAnet - Electronic Resource Preservation and Access Network
  - UK Digital Curation Centre
  - DELOS - Network of Excellence on Digital Libraries
  - DILIGENT - Digital Library Infrastructure on Grid Enabled Technology
Curation

- Set of processes to assemble the information needed to access, manipulate, and manage material

- Appraisal - selection of what to preserve
- Accession - controlled import of the data
- Arrangement - how to structure the material
- Description - how to provide the authenticity, descriptive, integrity metadata
- Preservation - creation of the archival form, and storage
- Access - discovery and manipulation
IAU Commission 5

- **Working Group Astronomical Data**, Chair: Ray Norris
  - Promoting open access to observatory archives
  - Promoting the concept of the Virtual Observatory
  - Promoting common formats for astronomical data to facilitate sharing and entry into public-domain databases.

- **Working Group Designations**, Chair: Marion Schmitz
  - IAU Recommendations for Nomenclature

  - LISA (Library and Information Services in Astronomy)

- **Working Group FITS**, Chair: William Pence, vice-chair: François Ochsenbein
  - Standard data interchange and archiving format

- **Working Group Virtual Observatories**, constitution in progress, contact person: Françoise Genova
  - International Virtual Observatory Alliance

- **Task Force Preservation and Digitization of Photographic Plates**, Chair: Elizabeth Griffin
  - Pisgah Astronomical Research Institute archive
Preservation Environments Working Group

• Co-chairs
  ▪ Bruce Barkstrom (NASA Langley)
  ▪ Reagan Moore (SDSC)

• Preservation principles
  ▪ Authenticity - ability to assert that the information needed to describe a digital record remains linked to the record
  ▪ Integrity - ability to assert that the record remains uncorrupted and that the management of the record can be tracked
  ▪ Infrastructure independence - ability to incorporate new technology within a preservation environment

• GGF Requirements document
  ▪ Document being produced for GGF14 - June 26.

• Synergy with IVOA
  ▪ Requirements for the preservation of collections
    ▪ Standard vocabulary - Uniform Content Descriptors
    ▪ Standard encoding format - FITS
    ▪ Standard services for manipulating the encoding format
Digital Library Technology

- **DSpace** - digital library that provides standard data curation services
  - Import files, add metadata, validate, create METS profile, assemble collection, archive
- **DSpace - SRB integration**
  - Support distribution, replication of data
  - Support federation of DSpace instances
  - Based on a Java interface (JARGON) to the SRB
  - Integration done by David Little, UCSD Libraries
- **DSpace digital library services remained the same**
  - SRB provides a distributed data management system to support DSpace files
  - Goals are to support access of files at a remote site, distribution of files between sites, replication to a remote site,
Integration of Knowledge Management with Data Grids

- **Fedora - NSF NSDL project**
  - Digital Library technology that supports relationships on records (annotation, logical organization, display constraints, …)
  - Uses RDF to describe relationships
  - Plan to integrate with SRB in 2nd quarter FY05 for access to NSDL persistent archive

- **NSF ITR project on Constraint-Based Knowledge Systems (SDSC)**
  - Characterize constraints as relationships that may be imposed dynamically, reified (characterize application of constraint and save result), and updated (characterize state information about reification)
  - Classes of constraints for data administration, access control, organization, presentation, and transformation
Standards

- Open Archival Information System - OAIS
  - Archival Information Package - AIP
- Metadata Encoding and Transmission Standard - METS
  - Profiles for structuring metadata
- Open Archives Initiative Protocol for Metadata Harvesting - OAI-PMH
  - Simple API for retrieving metadata
- Hierarchical Data Format version 5 - HDFv5
  - Standard characterization of groups of files in a container
- Data grid containers
  - Physical mapping of files into containers
- ISO Producer-Archive Interface Methodology
  - ISO DIS 14721 - development of producer archive submission pipelines
Projects

- **InterPARES project**
  - Preservation of Canadian MOST collection

- **TeraGrid Data Intensive Analysis**
  - Replication of sky surveys onto Teragrid for data intensive analyses
  - Providing 50 TBs of on-line disk storage
  - Hyperatlas - Roy Williams

- **NSF Chronopolis**
  - Replication of sky survey image collections to mitigate risk of data loss
  - Integration of digital library technology with data grid technology to build preservation environments

- **Library of Congress National Digital Information Integration and Preservation Program - NDIIPP**
  - Migrate collections between preservation systems