



NPACI

Knowledge-Based Persistent Archives

Reagan W. Moore

San Diego Supercomputer Center

9500 Gilman Drive, La Jolla, CA 92093-0505

Phone: 858 534-5073 FAX: 858 534-5152

E-mail: moore@sdsc.edu

**Presented at the THIC Meeting at the Bahia Hotel
998 West Mission Bay Dr, San Diego CA 92109
on January 16, 2001**



THIC Inc.

The Premier Advanced Recording Technology Forum

NPACI



Data Intensive Computing Environment

Staff

- **Reagan Moore**
- **Chaitan Baru**
- **Sheau Yen Chen**
- **Charles Cowart**
- **Amarnath Gupta**
- **George Kremenek**
- **Bertram Ludäscher**
- **Richard Marciano**
- **Arcot Rajasekar**
- **Abe Singer**
- **Michael Wan**
- **Ilya Zaslavsky**
- **Bing Zhu**

Students - GSRA

- **Martin Kuhl**
- **Liyang Sui**
- **Yang Yu**
- **Valter Crescenzi**

Students - Undergrad Interns

- **Peter Shin**
- **Roman Olshanowsky**
- **Shabbar Tambawala**
- **Pratik Mukhopadhyay**
- **+/- NN**

Topics

- Persistent archive functionality
- Characterization of
 - Data / Information / Knowledge
- Integration of Digital Library, Grid environments, and Persistent Archives

Persistent Archive

- Manage digital objects for the “life of the republic”
- Maintain ability to discover and access digital objects while supporting hardware and software systems evolve

Fundamental Concept for a Persistent Archive

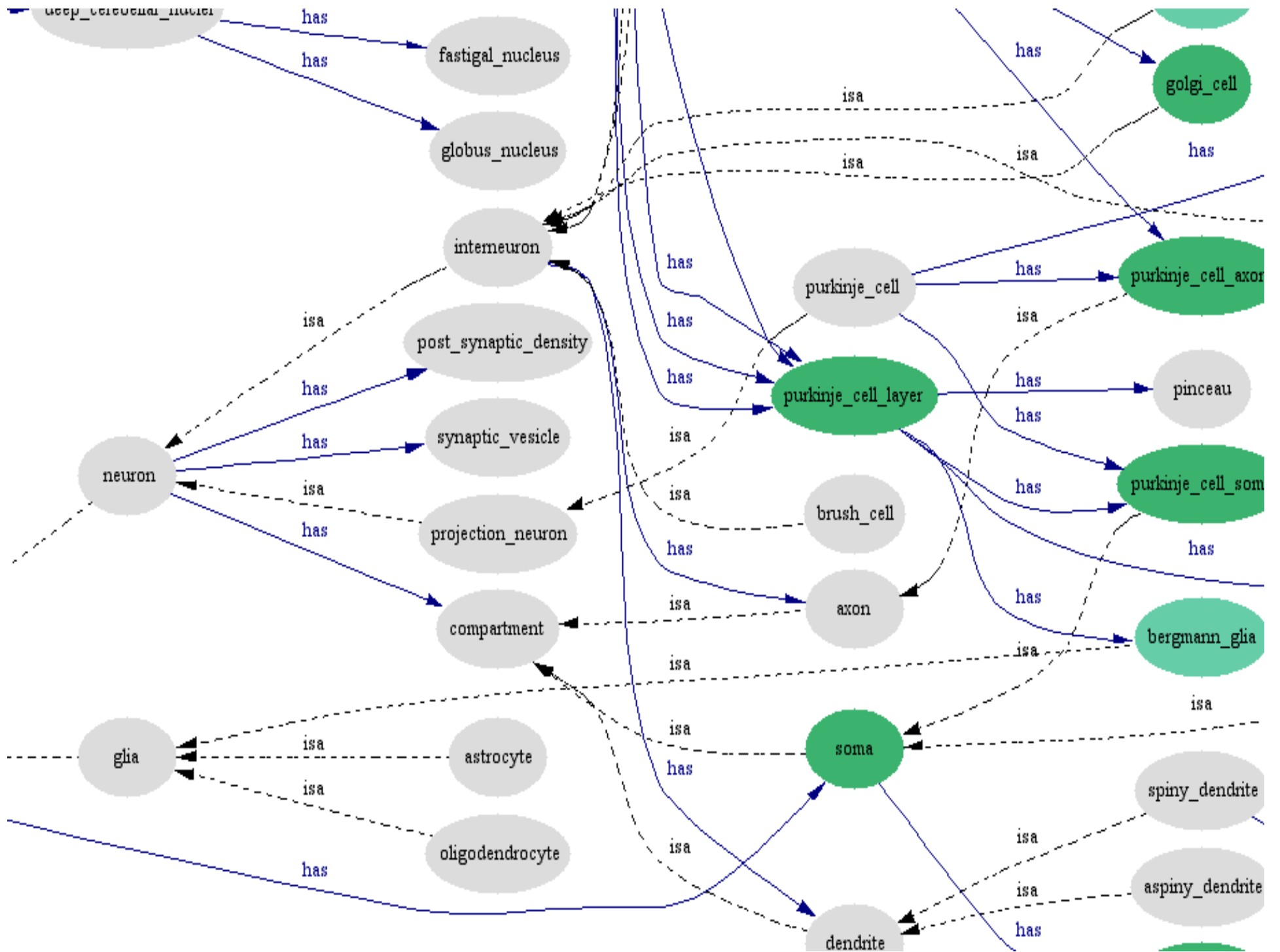
- Persistence requires migration over time onto new technology
- While the migration occurs, a persistent archive must be able to interoperate with both the old technology and the new technology.
- A persistent archive is an interoperability system.

What Types of Interoperability are Needed?

- Data management (digital objects)
 - Ability to work with multiple types of storage systems, across separate administration domains
- Information management (attributes)
 - Ability to define a collection independent of database choice
 - Ability to migrate collection onto new databases
- Knowledge management (relationships)
 - Ability to manage relationships
 - Ability to map domain concepts to collection attributes

Simplest Definitions

- Data
 - Digital object
 - Objects are streams of bits
- Information
 - Any tagged data, which is treated as an attribute.
 - Attributes may be tagged data within the digital object, or tagged data that is associated with the digital object
- Knowledge
 - Relationships between attributes
 - Relationships can be procedural/temporal, structural/spatial, logical/semantic, functional



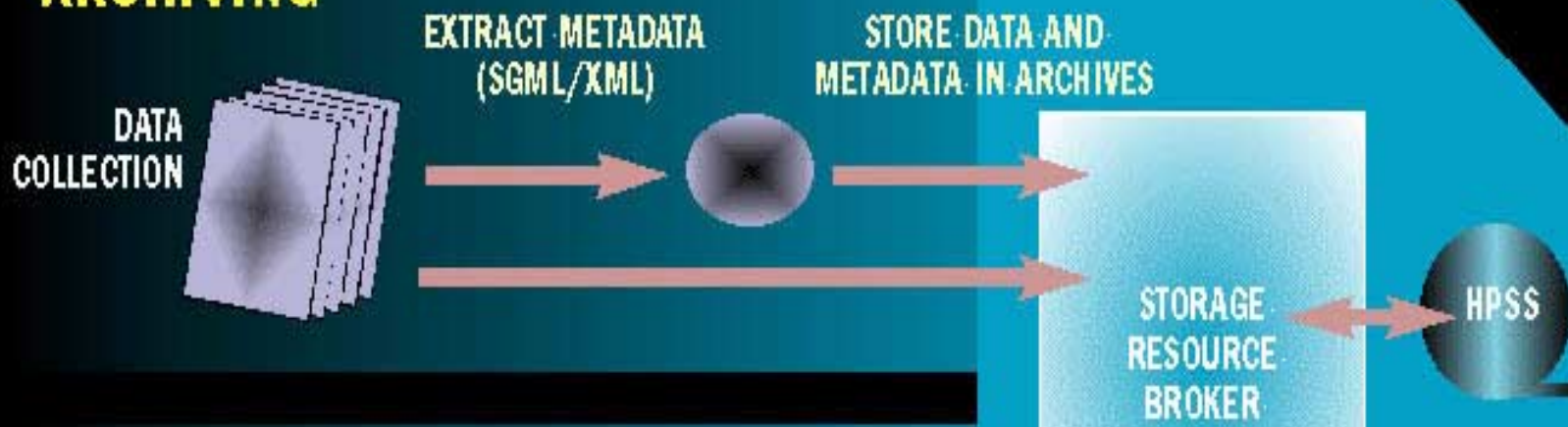
Implicit Concepts for Persistent Archive

- Infrastructure independence
 - Data set access
 - Authentication
 - Collection management
 - Presentation
 - Non-proprietary formatting
- Information models
 - XML - Information markup language
 - GML - Graphics markup language
- Support for ingestion, management, access
 - Accessioning workbench, archive, access workbench

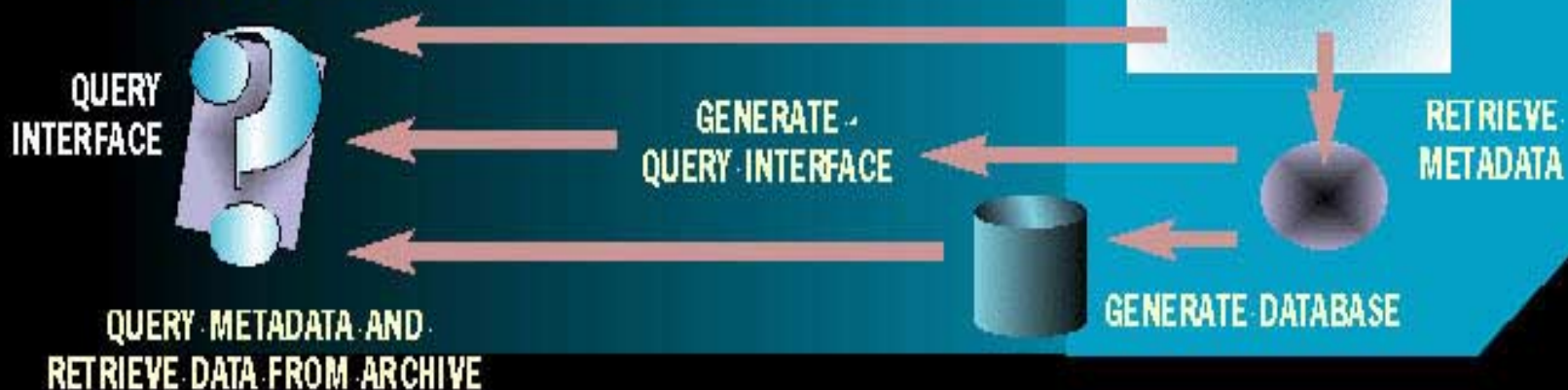
Standard Information Markup Language

- XML representation of metadata attributes
 - Standardization of DTDs - MOA II DTD for text
 - Standardization of markup language
- XML based representation of collection structure
 - Attributes defining the physical layout of a schema into relational tables (foreign keys, attribute data types, ...)
- XML databases & XML organized data collections
 - Commercial systems: Excelon, TAMINO, Oracle8i,
- XML based Topic Maps
 - Represent relationships between collection domain concepts, collection attributes

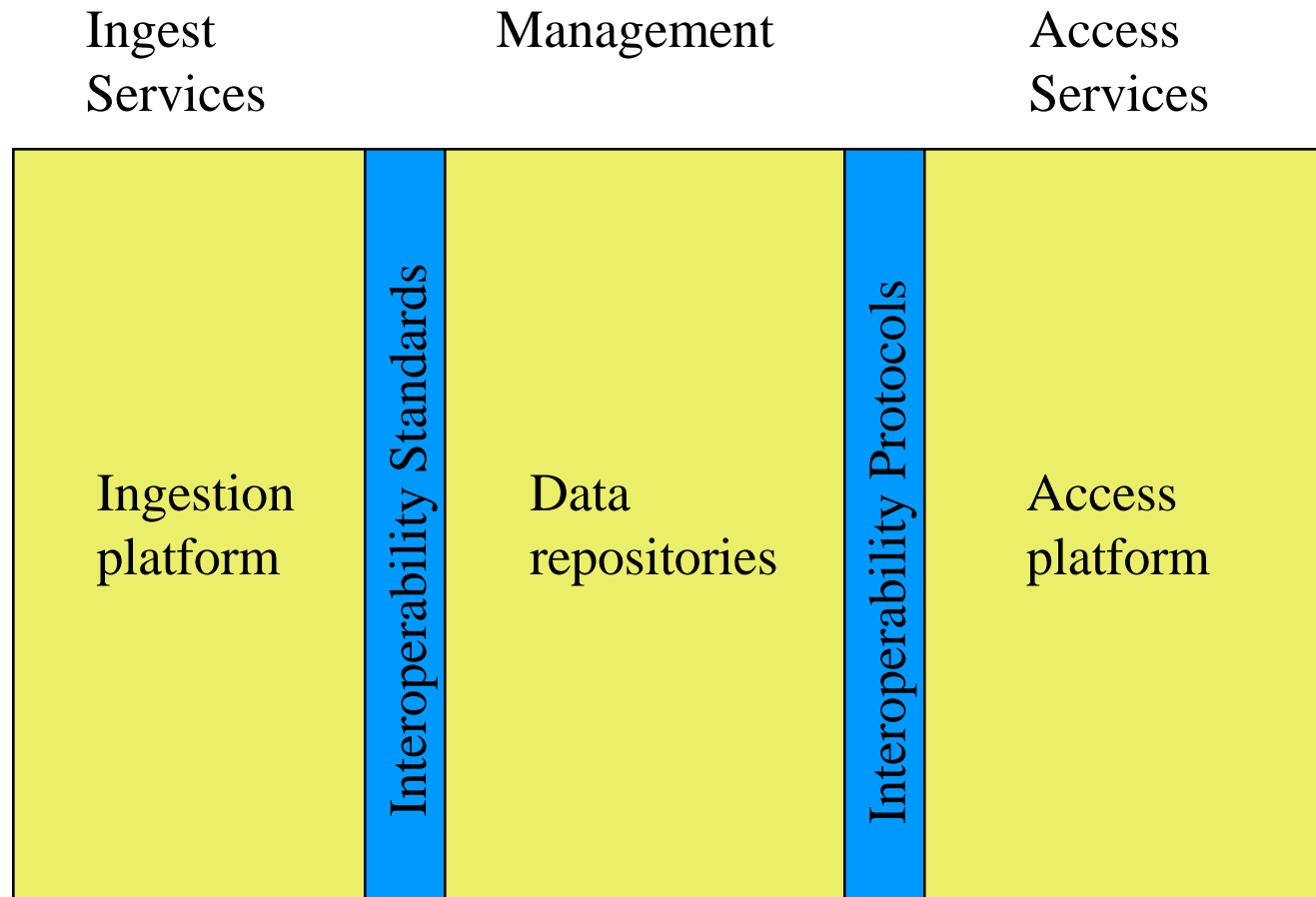
ARCHIVING



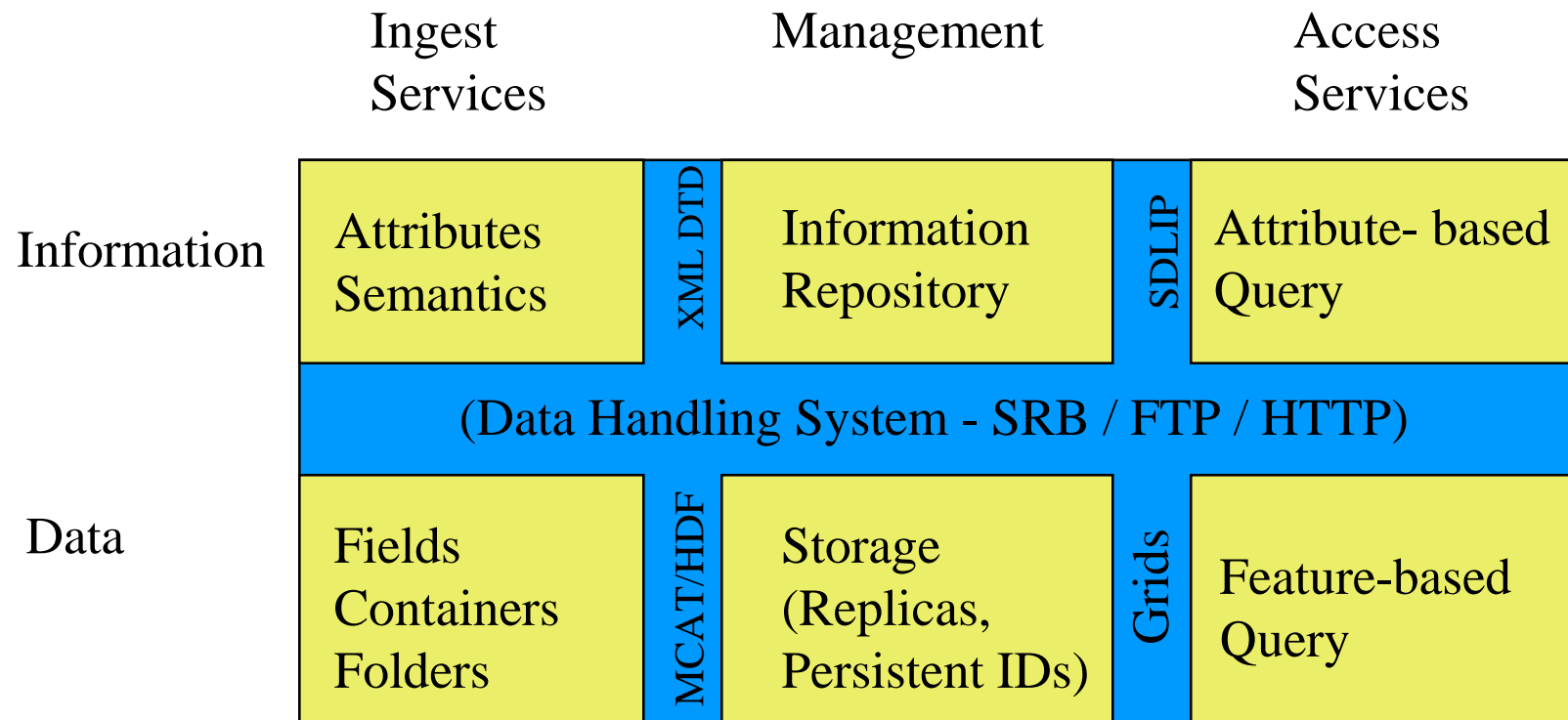
RETRIEVAL



Data Archive



Collection Based Persistent Archive



Knowledge Based Persistent Archive

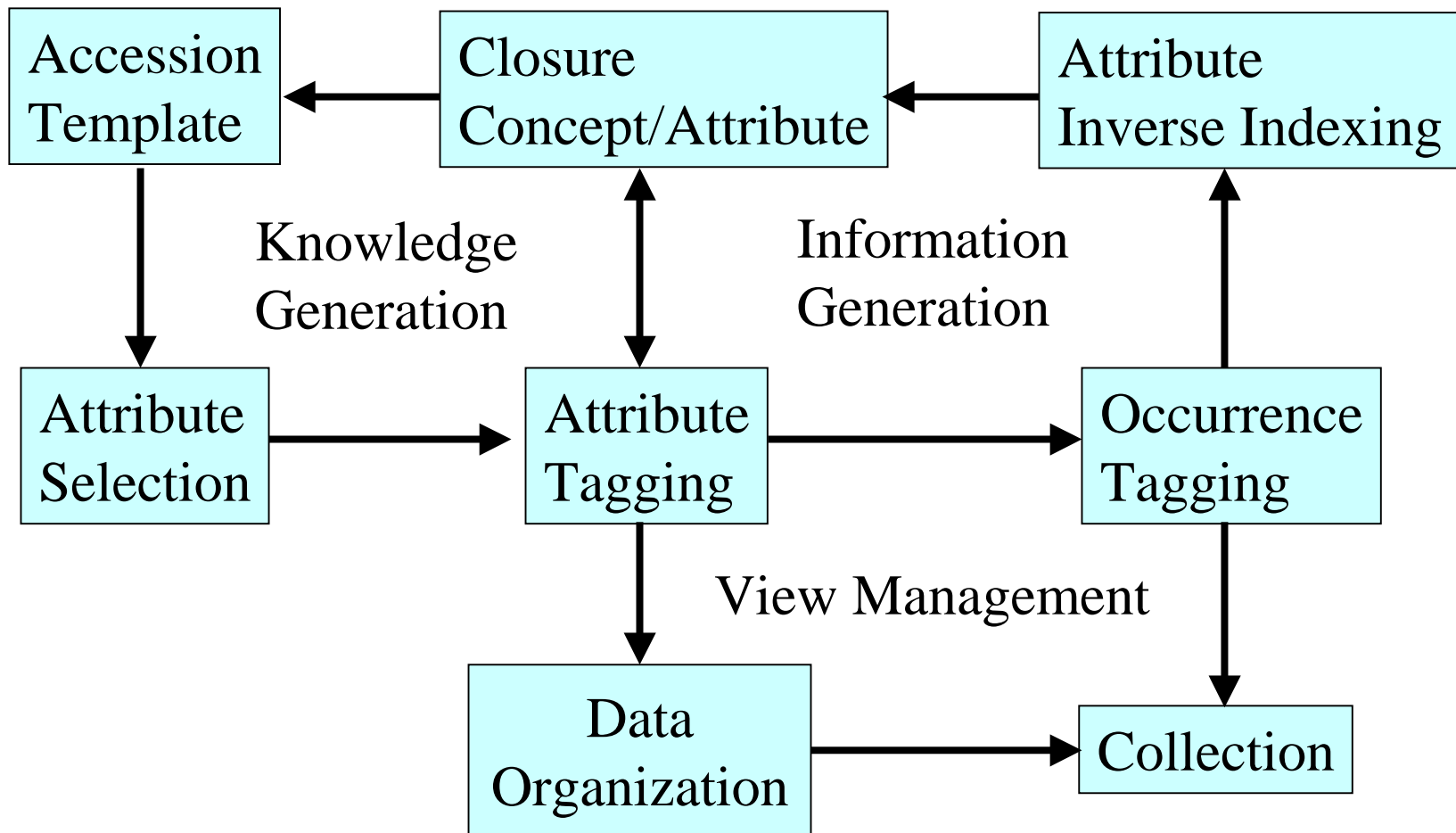
Ingest
Services

Management

Access
Services

Knowledge	Relationships Between Concepts	XTM DTD	Knowledge Repository for Rules	Rules - KQL	Knowledge or Topic-Based Query / Browse
	(Topic Maps / Buckets / Model-based Access)				
Information	Attributes Semantics	XML DTD	Information Repository	SDLIP	Attribute- based Query
	(Data Handling System - SRB / FTP / HTTP)				
Data	Fields Containers Folders	MCAT/HDF	Storage (Replicas, Persistent IDs)	Grids	Feature-based Query

Ingestion Processes for Collection Creation



Information Management Projects

- Digital Libraries
 - NSF Digital Library Initiative, Phase II - UCSB, Stanford
 - Digital Embryo digital library - GMU
 - NPACI Digital Sky - Caltech 2MASS sky survey
 - CDL - AMICO
 - NSF NSDL - UCAR / DLESE
- Grid Environments
 - NASA Information Power Grid - NASA Ames
 - DOE Data Visualization Corridor - LLNL
 - DOE Particle Physics Data Grid - Stanford, Caltech
 - NSF Grid Physics Network - U Fl
- Persistent Archives
 - NARA Persistent Archive
 - NHPRC - Scalable archives



Communities Providing Technology

- Archival storage - HPSS, ADSM, SANs
- Data handling - Storage Resource Broker
- Databases - XML, Object relational
- Digital libraries - services, information discovery
- Data grids - collection federation, finding aids
- Computational grids - remote execution
- Library - catalogs, DTDs, finding aids
- Archivist - archival procedures

Digital Library Data Management (CDL)

- Persistent identifiers
 - Ability to move a data set without the name changing
- Data set replicas
 - Management of multiple copies of a data set
- Archival backup of data sets
 - Integration of disk data caches with archival storage
- Persistent archives
 - Management of a collection through multiple cycles of technology evolution

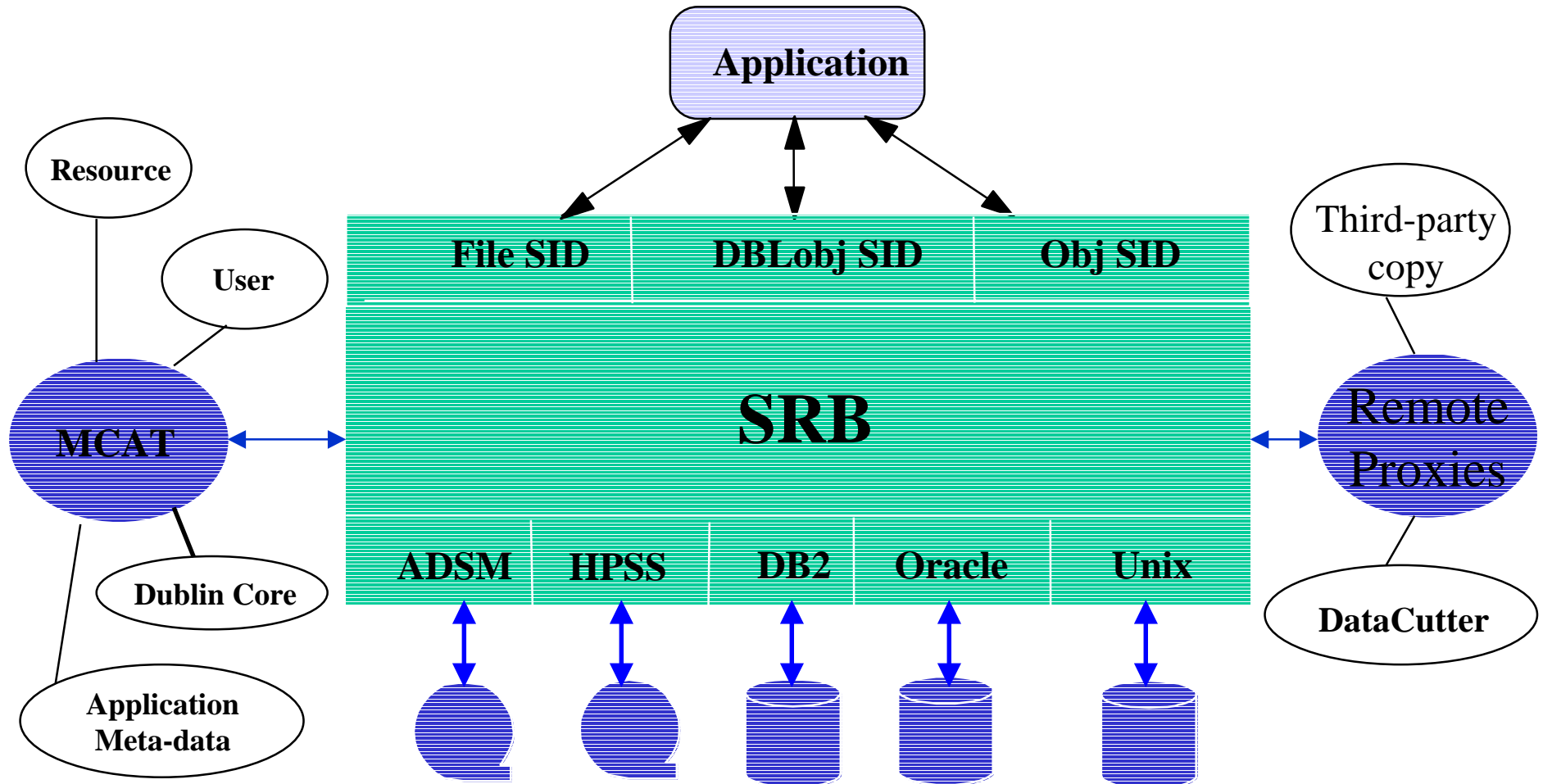
Storage Transparencies

- Location transparency
 - Distribution of data collection across multiple physical resources
- Name transparency
 - Attributed based access to data
- Protocol transparency
 - Common API for access to remote data resources
- Time transparency
 - Minimization of data access latency

Common Data Management Hierarchy

- Persistent Archives
 - Storage of information model, data model, along with data
- Data Grid
 - Access to data in a different administration domain
- Digital Library - services
 - Interlib - ADEPT, UC Berkeley Digital Library
- Data Collection
 - Extensible Meta-data catalog - EMCAT
- Data handling
 - SDSC Storage Resource Broker - SRB
- Archival Storage
 - High performance storage system - HPSS

SDSC Storage Resource Broker & Meta-data Catalog



Applications

- Support for distributed data collections
- Federation of data collections to form digital library
- Integration of digital libraries with archives
- Finding aids for federation of digital libraries through mediation of information
- Data grids for data access
- Persistent archives

Further Information

<http://www.npaci.edu/DICE>

