

Digital Libraries, Data Grids, and Persistent Archives

Reagan W Moore

San Diego Supercomputer Center

10100 John Jay Hopkins Dr, La Jolla CA 92093

Phone: +1-858-534-5073

E-mail: moore@sdsc.edu

Presented at the THIC Meeting at the Hilton San Diego/Del Mar

Del Mar CA 92014-1901

on January 22, 2002

Digital Libraries, Data Grids, and Persistent Archives

Reagan W. Moore
San Diego Supercomputer Center
moore@sdsc.edu
<http://www.npaci.edu/DICE/>

Data and Knowledge Systems Group

Staff

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

Graduate Students

- **A. Bagchi**
- **S. Bansal**
- **A. Behere**
- **R. Bharath**
- **S. Bharath**
- **M. Kulrul**
- **L. Sui**

Undergraduate Interns

- **N. Cotofana**
- **M. Shumaker**
- **J. Trang**
- **L. Yin**
- **+/- NN**

Accessing Data

- How do you access storage systems at remote sites in someone else's administration domain?
- How do you organize distributed data into a cohesive collection with global, persistent identifiers?

Information Management Projects

- **Digital Libraries**

- CDL - AMICO
- DARPA/USPTO - patent digital library
- NLM Visible Embryo digital library - GMU
- NSF Digital Library Initiative, Phase II - UCSB, Stanford
- NSF NPACI Digital Sky - Caltech 2MASS sky survey
- NSF NSDL - UCAR / Columbia / Cornell / UCSB

- **Data Grid Environments**

- DOE Data Visualization Corridor - LLNL
- DOE Particle Physics Data Grid - Stanford, Caltech
- NASA Information Power Grid - NASA Ames
- NIH Biomedical Informatics Research Network
- NSF Grid Physics Network - U Florida
- NSF National Virtual Observatory - Johns Hopkins University / Caltech
- NSF Southern California Earthquake Center - ISI

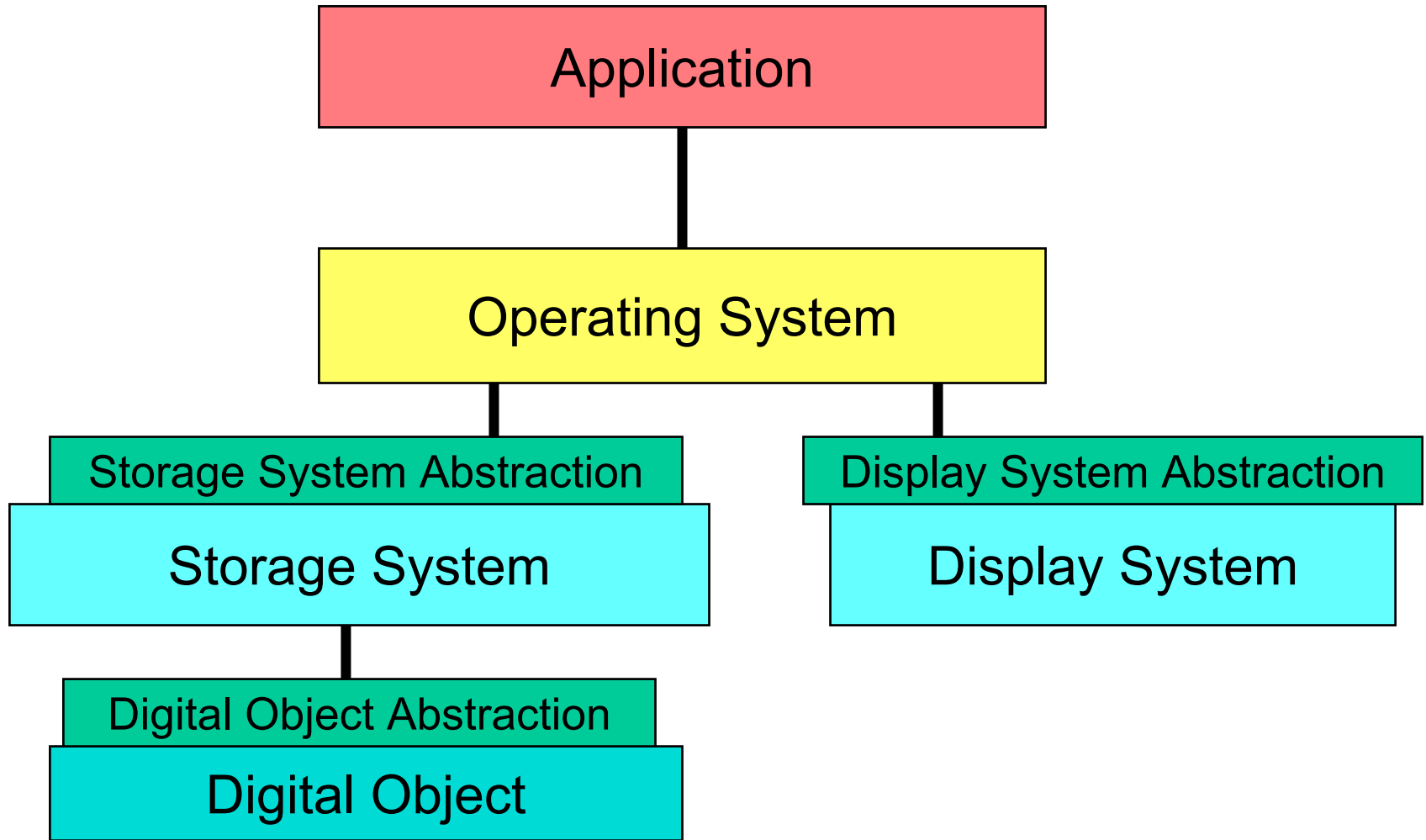
- **Persistent Archives**

- NARA Persistent Archive
- NHPRC - Archivist workbench

Specifying levels of Abstraction

- Technology management becomes simpler if the persistent archive infrastructure operates on abstractions, rather than an explicit physical implementation of a resource
- Can we abstract
 - Digital objects
 - Storage

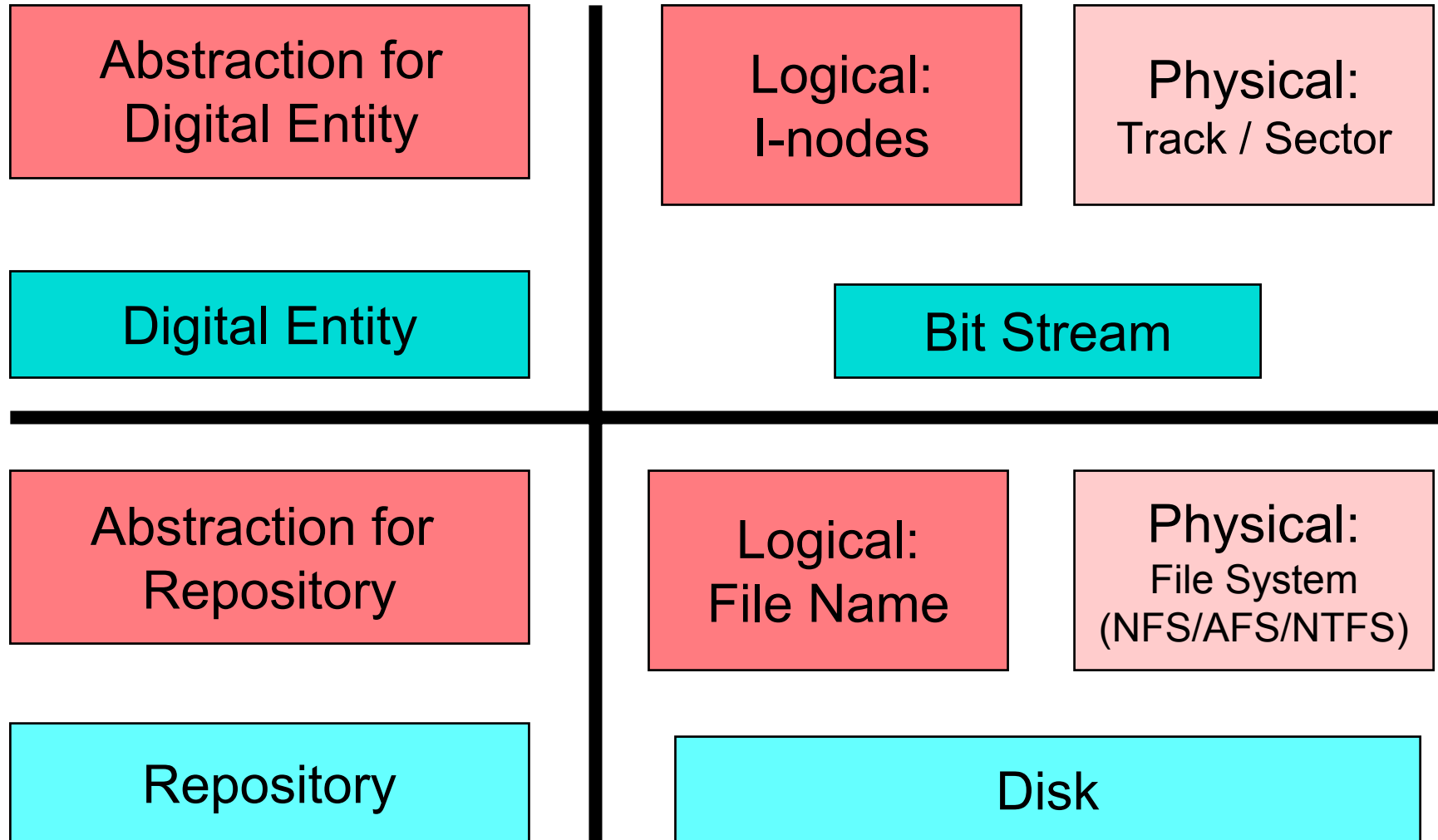
Technology Management



Types of Digital Entity Abstractions

- Logical representation
 - What does the digital entity represent?
 - What is the associated meaning?
- Physical representation
 - What is the physical structure of the digital entity?

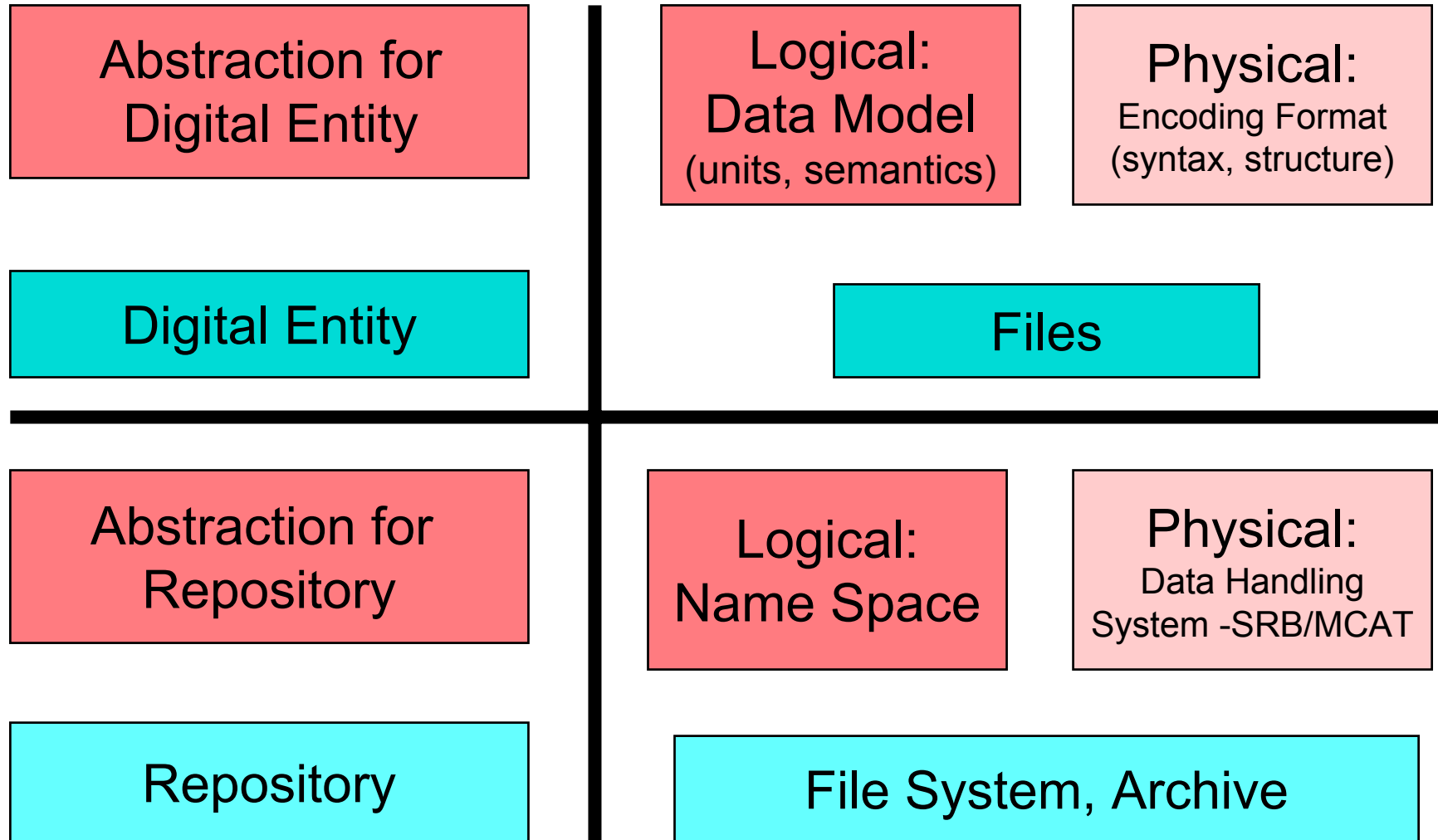
Levels of Abstraction for Bits



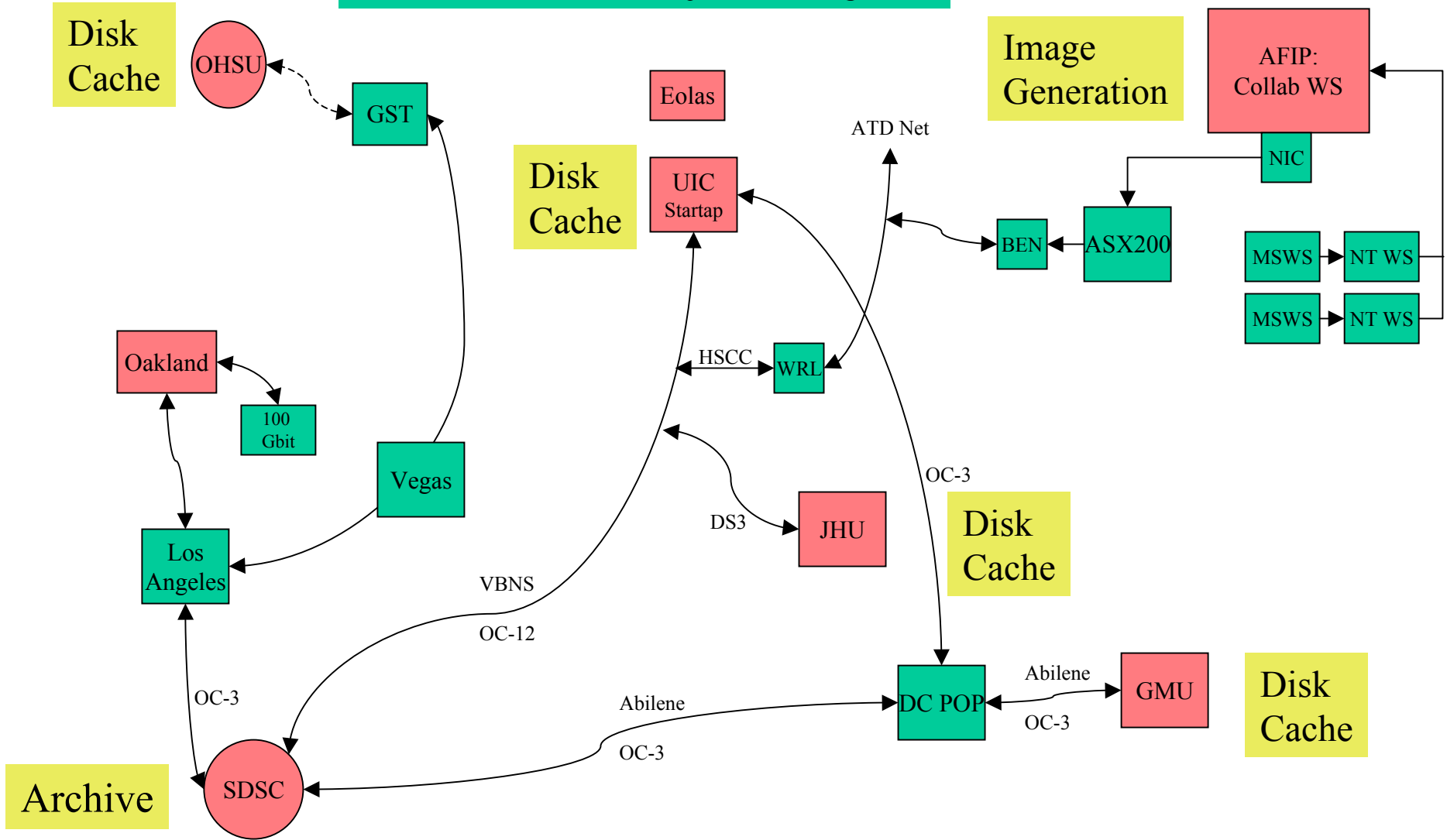
Managing Distributed Storage

- Separate the organization of digital objects from their physical storage
 - Logical Name Space to manage attributes about the digital objects
 - Data handling system to manage interactions with remote storage systems
- Create storage abstraction layer
- Storage Resource Broker (SRB) provides data management system

Levels of Abstraction for Data



Visible Embryo Project



Disaster Response

- Support replicas - provide multiple copies of a data set stored at multiple sites, but accessed by the same logical file name
- On access, map from logical file name to the physical file name. If the file is not accessible, automatically fail over to a replica.

SDSC Storage Resource Broker & Meta-data Catalog

Application

C, C++, Libraries	Linux I/O	Unix Shell	Java, NT Browsers	DLL / Python	Prolog Predicate	Web
--------------------------	------------------	-------------------	--------------------------	---------------------	-------------------------	------------

Clients

Consistency Management / Authorization-Authentication

Logical Name Space	Latency Management	Data Transport	Metadata Transport
---------------------------	---------------------------	-----------------------	---------------------------

Prime Server

Catalog Abstraction

Storage Abstraction

Databases DB2, Oracle, Sybase	Archives HPSS, ADSM, UniTree, DMF	HRM	File Systems Unix, NT, Mac OSX	Databases DB2, Oracle, Postgres
---	---	------------	--	---

Servers



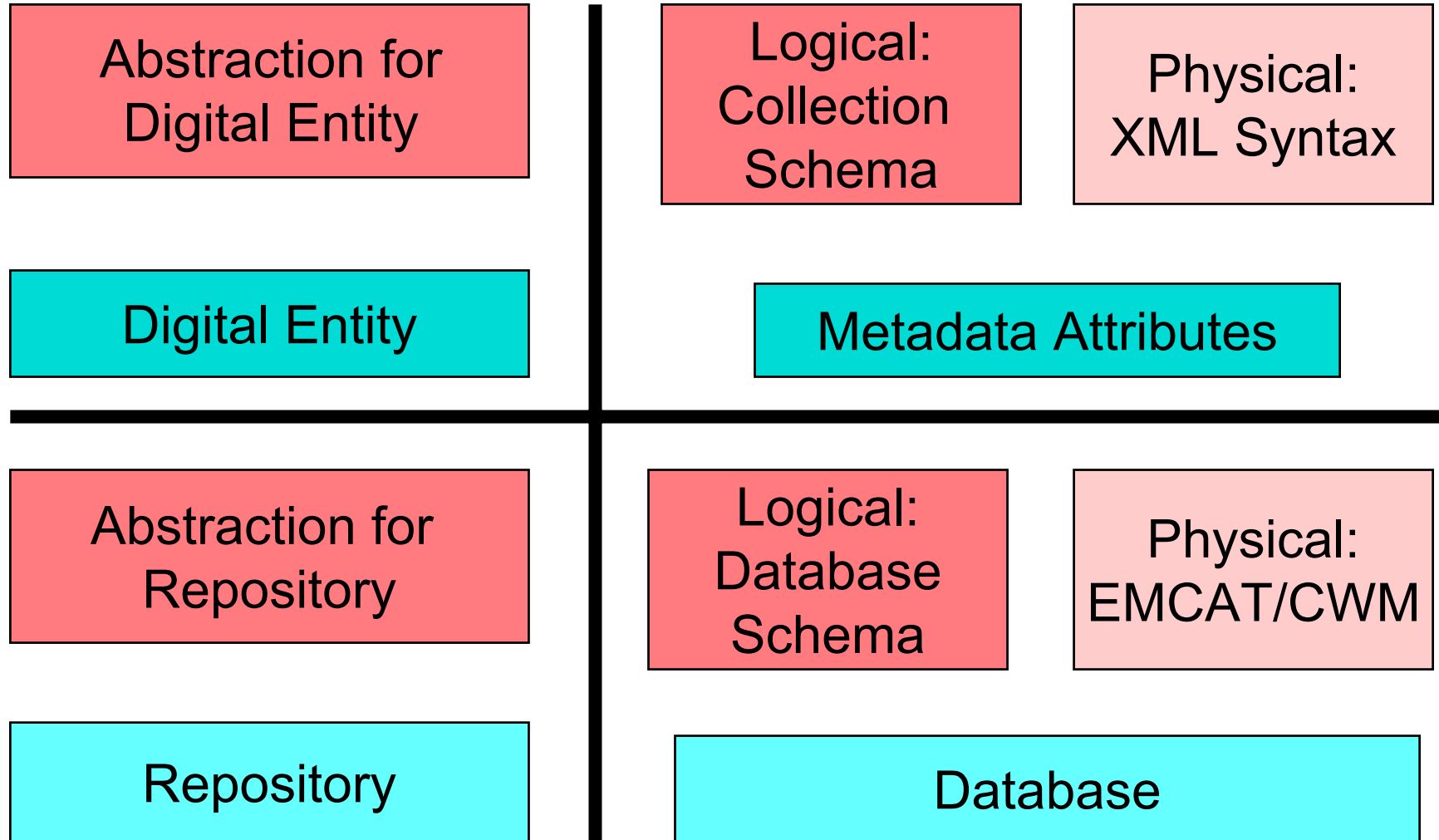
Information Management- Logical Name Space

- Set of attributes to describe digital entities that are registered into the logical name space
 - SRB metadata - Unix file system semantics
 - Provenance metadata - Dublin Core
 - Resource metadata - User access control lists
 - Discipline metadata - User defined attributes
- Each digital entity may have unique attributes

Information Management

- Abstraction layer for interacting with information repositories
 - Manage the schema and physical table structures of a database
 - Extensible schema
 - User defined attributes
- Extensible Metadata CATalog (EMCAT) manages collections
- mySRB.html interface supports dynamic collection creation

Levels of Abstraction for Information



National Virtual Observatory Data Grid

1. Portals and Workbenches

2. Knowledge
& Resource
Management

3. Metadata
View

Data
View

Catalog
Analysis

Bulk Data
Analysis

Concept space

Standard APIs and Protocols

4. Grid
Security
Caching
Replication
Backup
Scheduling

5. Information
Discovery

Metadata
delivery

Data
Discovery

Data
Delivery

Standard Metadata format, Data model, Wire format

6. Catalog Mediator

Data mediator

Catalog/Image Specific Access

7. Compute Resources

Derived Collections

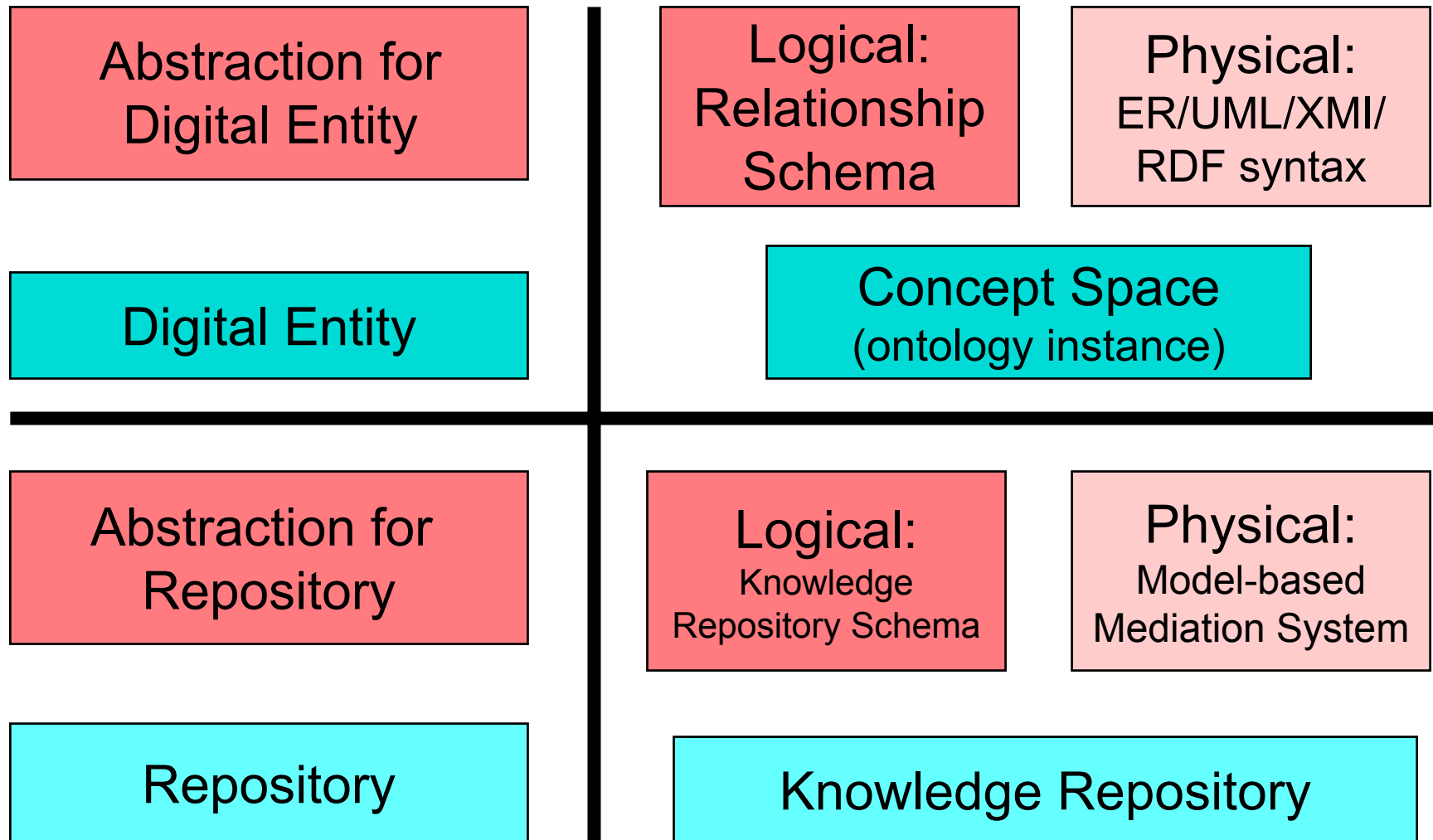
Catalogs

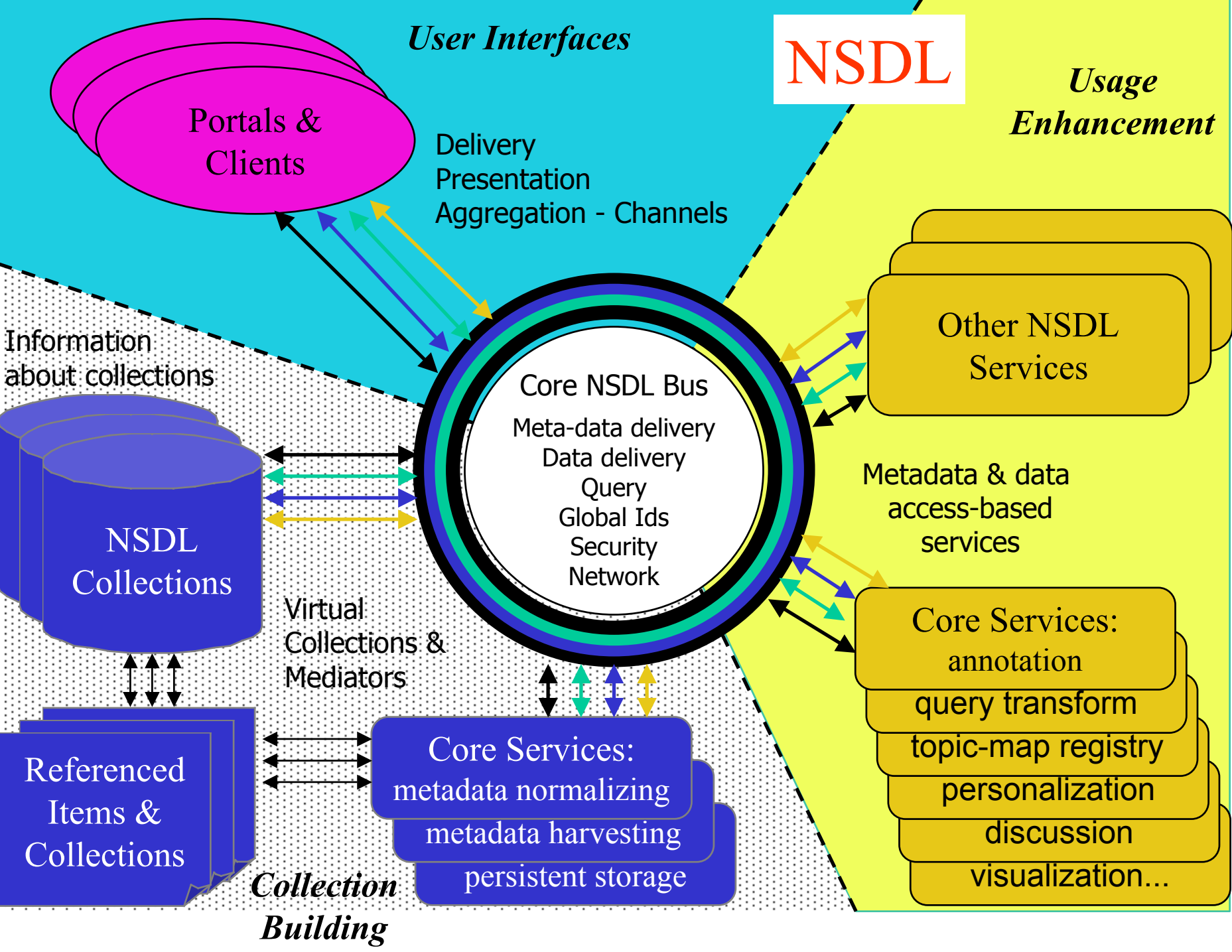
Data Archives

Knowledge Management - Discovery across Collections

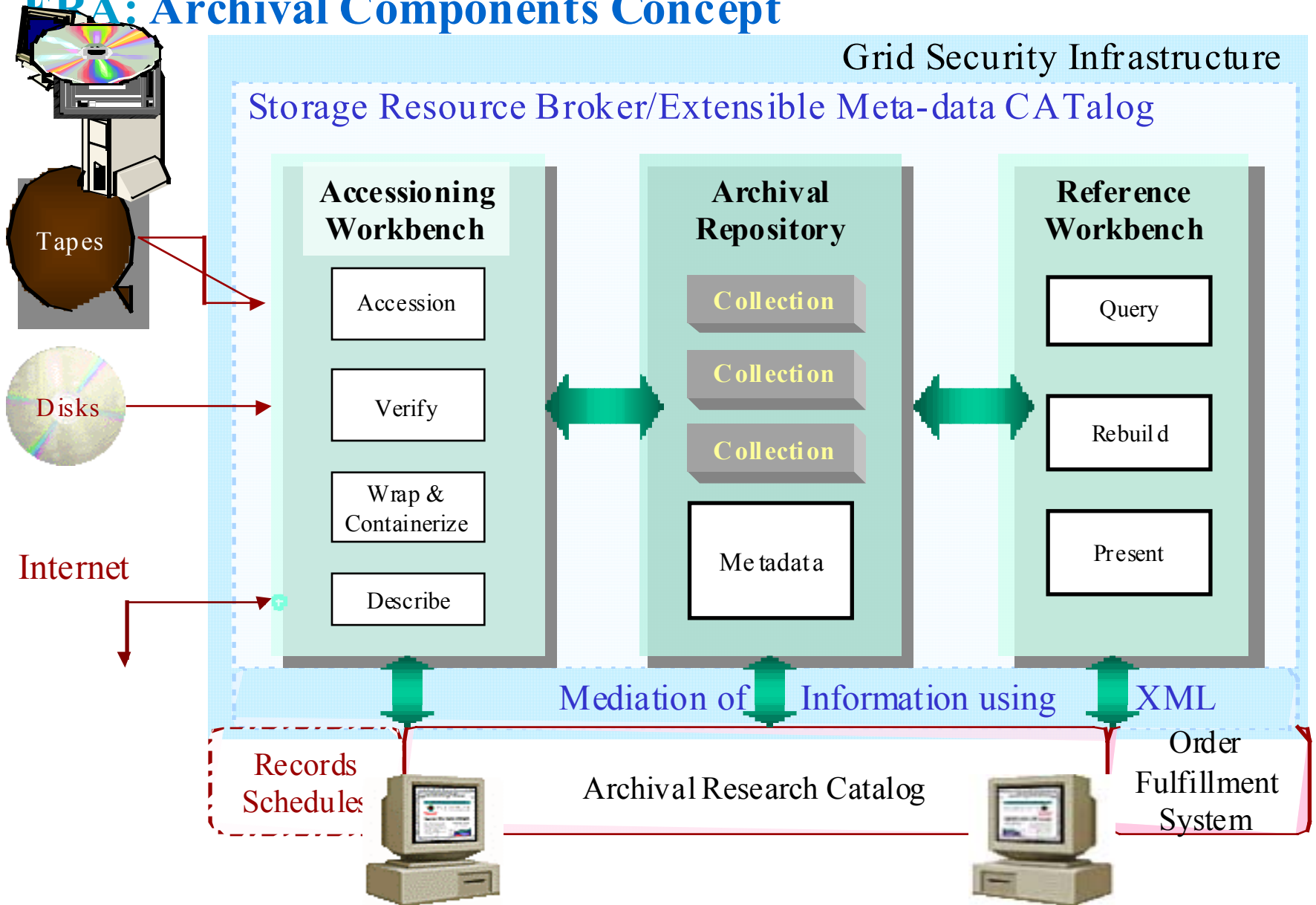
- Characterization of relationships between attributes
 - Semantic / logical - cross-walks
 - Procedural / temporal - records management
 - Structural / spatial - GIS
- Abstraction layer for knowledge repositories
- Mapping from collection attributes to discipline concepts
- Model-based Mediation supports mapping from knowledge relationships to rule-based inference engines

Levels of Abstraction for Knowledge





FPA: Archival Components Concept



Further Information

- Academic

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

- Commercial - Storage Resource Broker

constantin.scheder@gat.com