

# Building a Persistent Archive

## Technology Transfer to Federal Agencies

**Reagan W. Moore**

**San Diego Supercomputer Center**

**moore@sdsc.edu**

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

# Data and Knowledge Systems Group

## Staff

- **Reagan Moore**
- **Chaitan Baru**
- **Ilkai Altintas**
- **Sheau Yen Chen**
- **Charles Cowart**
- **Amarnath Gupta**
- **George Kremenek**
- **M. Kulrul**
- **Bertram Ludäscher**
- **Richard Marciano**
- **A. Memon**
- **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**
- **L. Sui**

## Undergraduate Interns

- **N. Cotofana**
- **D. Le**
- **J. Trang**
- **L. Yin**
- **+/- NN**

# Topics

- Management of digital entities
- Accelerating progress through common technology for digital libraries, data grids, and persistent archives
- Application of NSF developed technology to NARA, NASA, NIH, NLM, DOD, DOE, DOJ, LC

# Data Management Systems

- Data sharing environments
  - Share data across administration domains
- Digital Libraries
  - Provide information discovery, presentation services
- Data Grids
  - Federate multiple digital libraries
- Persistent Archives
  - Manage technology evolution

# Common Approach (Digital Library, Data Grid, Persistent Archive)

- Logical name space used to organize digital entities, and associate attributes
- Separation of information management from data storage management
- Definition of abstraction mechanisms for dealing with repositories
- Emergence of need for knowledge management

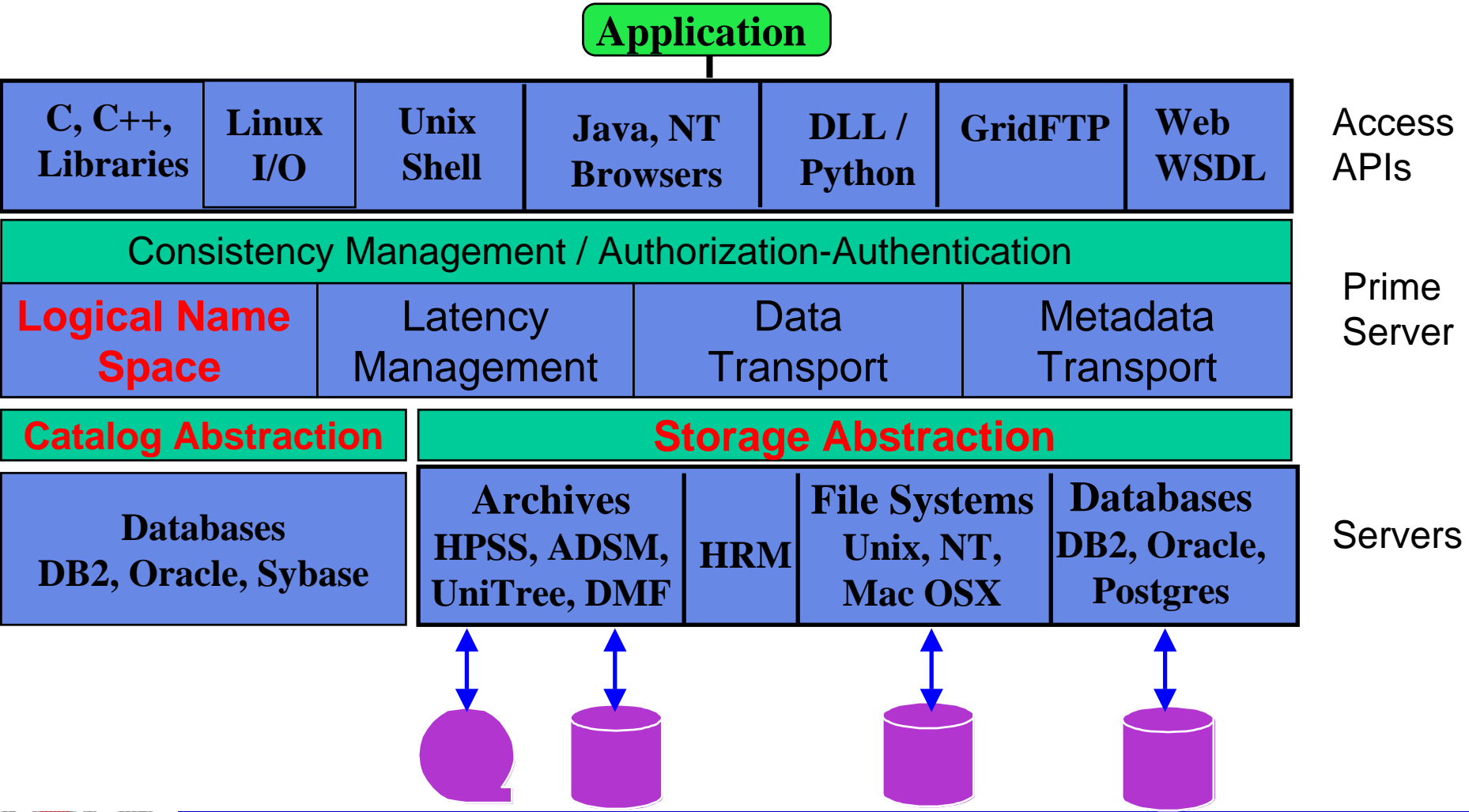
# Persistent Archives

(Similar requirements to a data grid)

- Name transparency
  - Find a file by attributes (map from attributes to global name)
- Location transparency
  - Access a file by a global identifier (map from global to local file name)
- Access transparency
  - Use same API to access data in archive or file cache
- Authenticity
  - Disaster recovery, replicate data across storage systems
  - Audit and process management

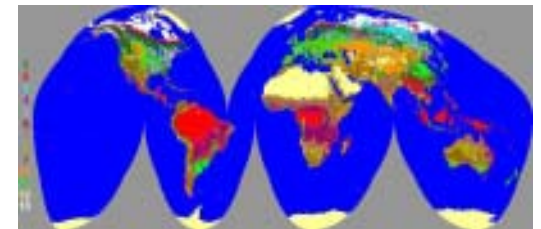
# SDSC Storage Resource Broker & Meta-data Catalog

## Levels of Abstraction



# SRB Collections

- **Astronomy**
  - CACR Computing Resource
  - National Virtual Observatory
  - 2 Micron All Sky Survey
  - DPOSS Collection
  - Hayden Planetarium
- **Ecology and Environmental Sciences**
  - CEED
  - Bionome
  - HyperLTER
  - Land Data Assimilation System
  - Knowledge Networks for BioComplexity
- **Medical Sciences**
  - Digital Embryo
- **Molecular Sciences**
  - JCSG, Synchrotron Data Repository
  - AFCS, Alliance for Cell Signaling
- **NeuroSciences**
  - Biomedical Information Research Network
  - TeleScience Portal
  - Brain Databases
  - Brain Data Archiving
- **Computer Science**
  - DataStreaming
  - AppLeS, DataCutter



# SRB Collections

## •Physics and Chemistry

- PPDG, Particle Physics Data Grid
- GriPhyN
- BaBar
- GAMESS

## •Digital Libraries and Archives

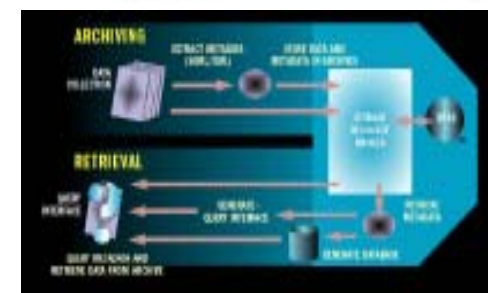
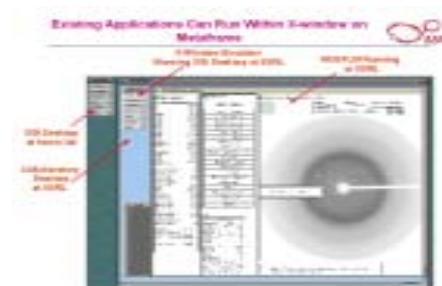
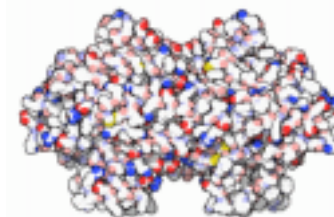
- SIO Digital Libraries
- California Digital Library
- ADEPT
- Stanford Digital Library Project
- National Archives and Records Administration

## •Data Grids

- ROADNet, Real-time Observatories Application and Data management
- E-Science at CLRC, UK Grid Starter Kit
- DOE ASCI Data Visualization Corridor
- NASA Information Power Grid
- DOE SciDAC - Portal Web Services
- NPACI Portal Projects

## •Education

- Transana
- Digital Insight
- NSDL National STEM Education Digital Library



# Digital Entities

- Digital entities are “images of reality”, made of
  - **Data**, the bits (zeros and ones) put on a storage system
  - **Information**, the attributes used to assign semantic meaning to the data
  - **Knowledge**, the structural relationships described by a data model or semantic relationships between attributes
- Every digital entity requires information and knowledge to correctly interpret and display

# Data, Information, and Knowledge

## Content of Digital Entities

- **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

# Knowledge Creation Roadmap

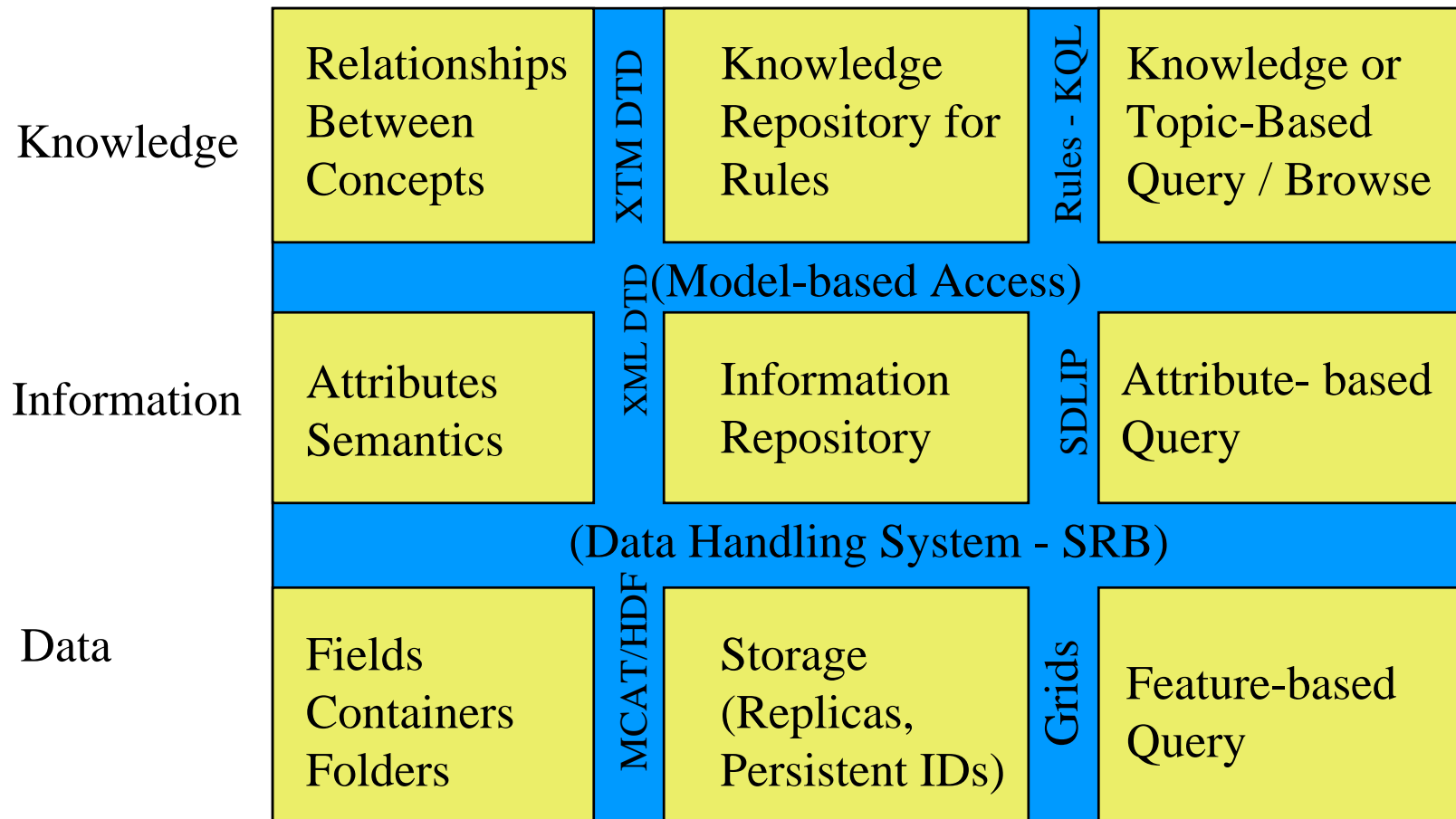
- Knowledge syntax (consensus)
  - RDF, XMI, Topic Map
- Knowledge management (recursive operations)
  - Oracle parallel database
- Knowledge manipulation (spatial/procedural rules)
  - Generation of inference rules and mapping to data models
- Knowledge generation (scalable inference engine)
  - Application of inference rules in inference engine

# Knowledge Based Data Grids

Ingest  
Services

Management

Access  
Services



# Further Information

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