Knowledge-based Grids

Reagan Moore
San Diego Supercomputer Center

(http://www.npaci.edu/DICE/)

Data Intensive Computing Environment

Chaitan Baru	Mediation of information
Walter Crescenzi	Web site wrapping
Amarnath Gupta	Rule-based mediation
Bertram Ludaescher	Self-instantiating archives
Richard Marciano	Knowledge management
Xufei Qian	Knowledge mining
Arcot Rajasekar	Collection management
Michael Wan	Data handling
Ilya Zaslavsky	GIS systems

Charlie Cowart	Browsers
Sheau Yen Chen	Digital Embryo project
George Kremenek	Information Power Grid / 2MASS
Bing Zhu	Particle Physics Data Grid

Technologies for Managing Storage in the Web

- Grids
- Data Grids
- Digital Libraries
- Persistent Archives
- Knowledge-based Grids

Storage Management

- **Logical representations for storage systems**
  - Store bits of data
- **Logical representations for information repositories**
  - Logical representations for collections (Information about digital objects)
  - Store attributes about data
- **Logical representations for knowledge repositories**
  - Store relationships between attributes

Grid Services

- **Grids provide access to distributed resources: computing, storage, sensors, display devices,…**

- **Middleware services**
  - Remote job execution
  - Remote file access
  - Authentication across administration domains
  - “Single sign-on”

  - Examples - Globus, Legion

Globus Layered Grid Architecture (By Analogy to Internet Architecture)

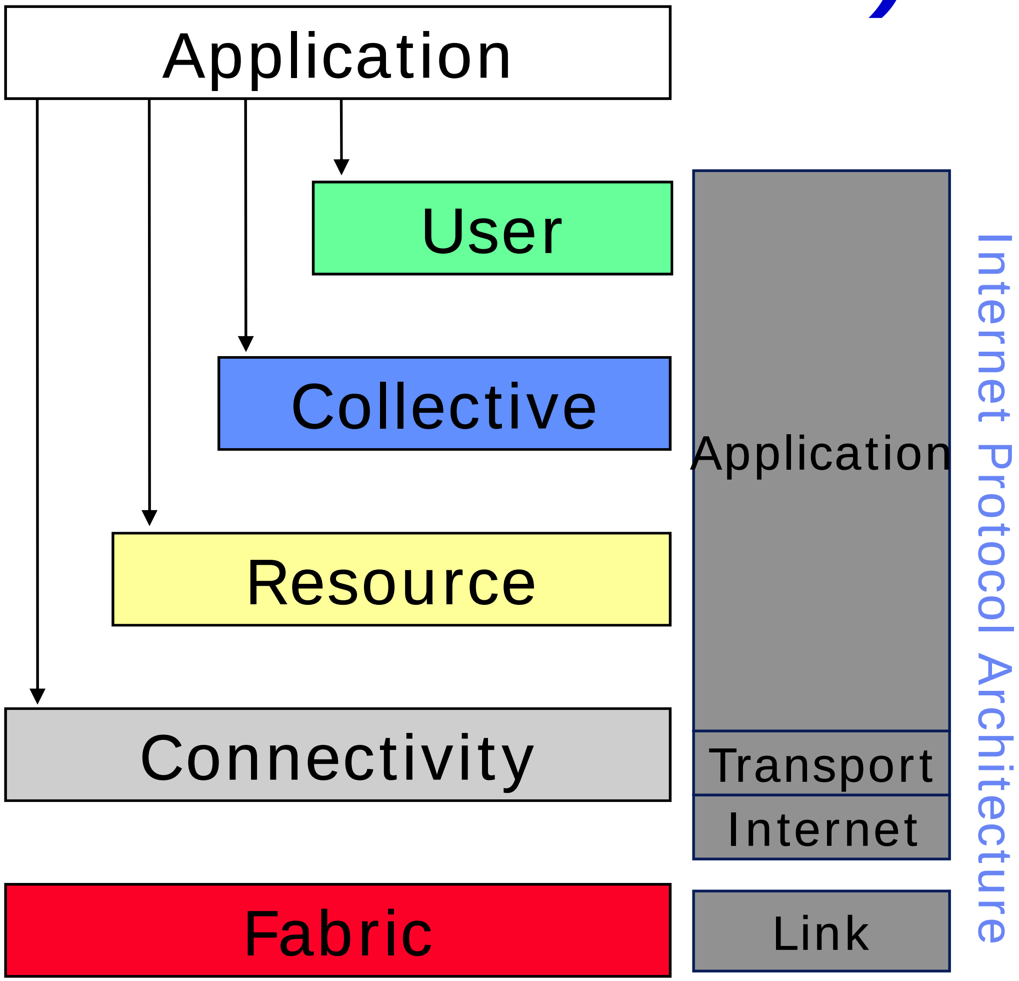
“Specialized services”: user- or appln-specific distributed services

“Managing multiple resources”: ubiquitous infrastructure services

“Sharing single resources”: negotiating access, controlling use

“Talking to things”: communication (Internet protocols) & security

“Controlling things locally”: Access to, & control of, resources



Application

User

Collective

Resource

Connectivity

Fabric

Application

Transport

Internet

Link

Internet Protocol Architecture

Data Grid

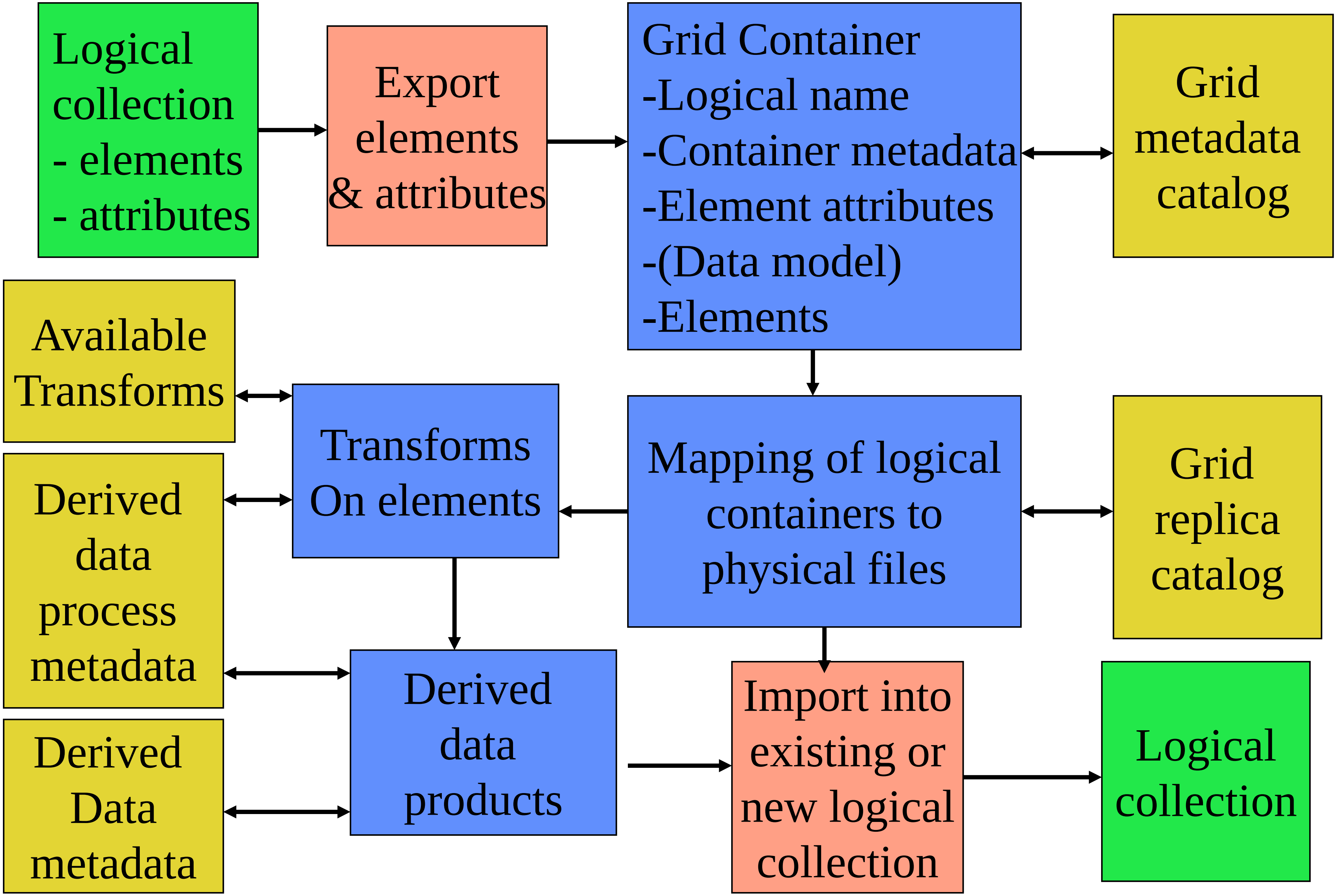
- **Supports management of data objects across a distributed set of storage resources**

- **Extends Grids to include data management. Challenges are:**
  - Object discovery
  - Managing context for objects (organization into collections)
  - Managing relationships between objects (concept spaces)
  - Integration of collections into data grids

Collection-based Storage

- **Access millions to billions of data objects within a collection**
  - Astronomy sky surveys - 2-Micron All Sky Survey
  - 5 million images, 10 TBs of data
- **Access requirements**
  - Replicate between two HPSS archives
  - Provide access to individual images
  - Provide access to thousands of images

Linking Collections with Data Grids



Logical collection
- elements
- attributes

Export elements & attributes

Grid Container
-Logical name
-Container metadata
-Element attributes
-(Data model)
-Elements

Grid metadata catalog

Available Transforms

Transforms On elements

Mapping of logical containers to physical files

Grid replica catalog

Derived data process metadata

Derived Data metadata

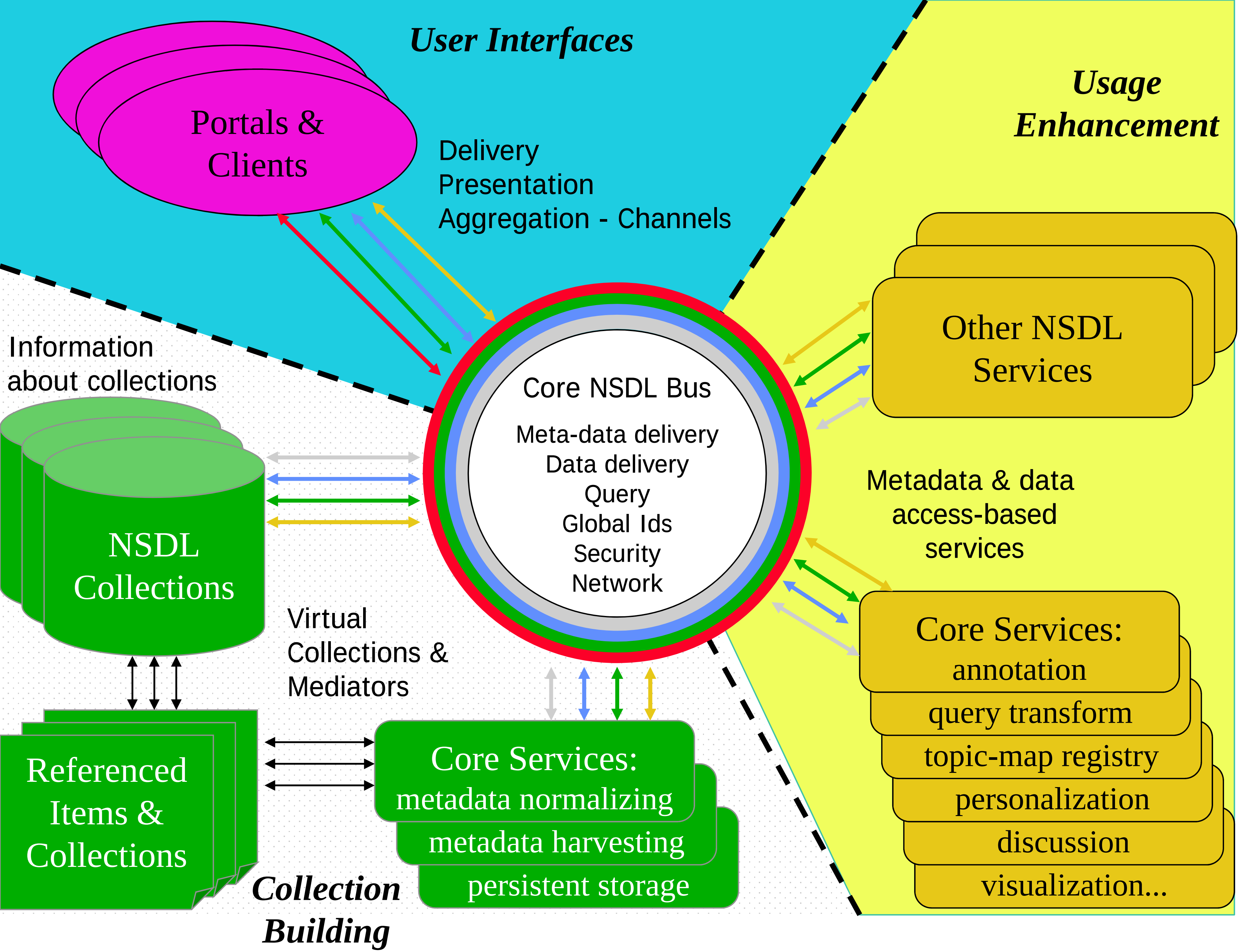
Derived data products

Import into existing or new logical collection

Logical collection

Digital Libraries

- **Provide services to discover, access, manipulate information organized in collections**
- **Discover**
  - Digital library standards for provenance metadata - Dublin Core
  - Information catalog characterization - MCAT
  - Schema composition for extensible discipline attributes - EMCAT
  - Discovery mechanisms based on XML syntax - XQuery
- **Access**
  - Metadata delivery mechanisms for information content using XML - SDLIP
- **Manipulate**
  - Extensions to XQuery - for manipulation of scientific data



User Interfaces

Portals & Clients

Delivery
Presentation
Aggregation - Channels

Usage Enhancement

Information about collections

Core NSDL Bus

Meta-data delivery
Data delivery
Query
Global Ids
Security
Network

Other NSDL Services

Metadata & data access-based services

NSDL Collections

Virtual Collections & Mediators

Core Services:
annotation
query transform
topic-map registry
personalization
discussion
visualization...

Referenced Items & Collections

Core Services:
metadata normalizing
metadata harvesting
persistent storage

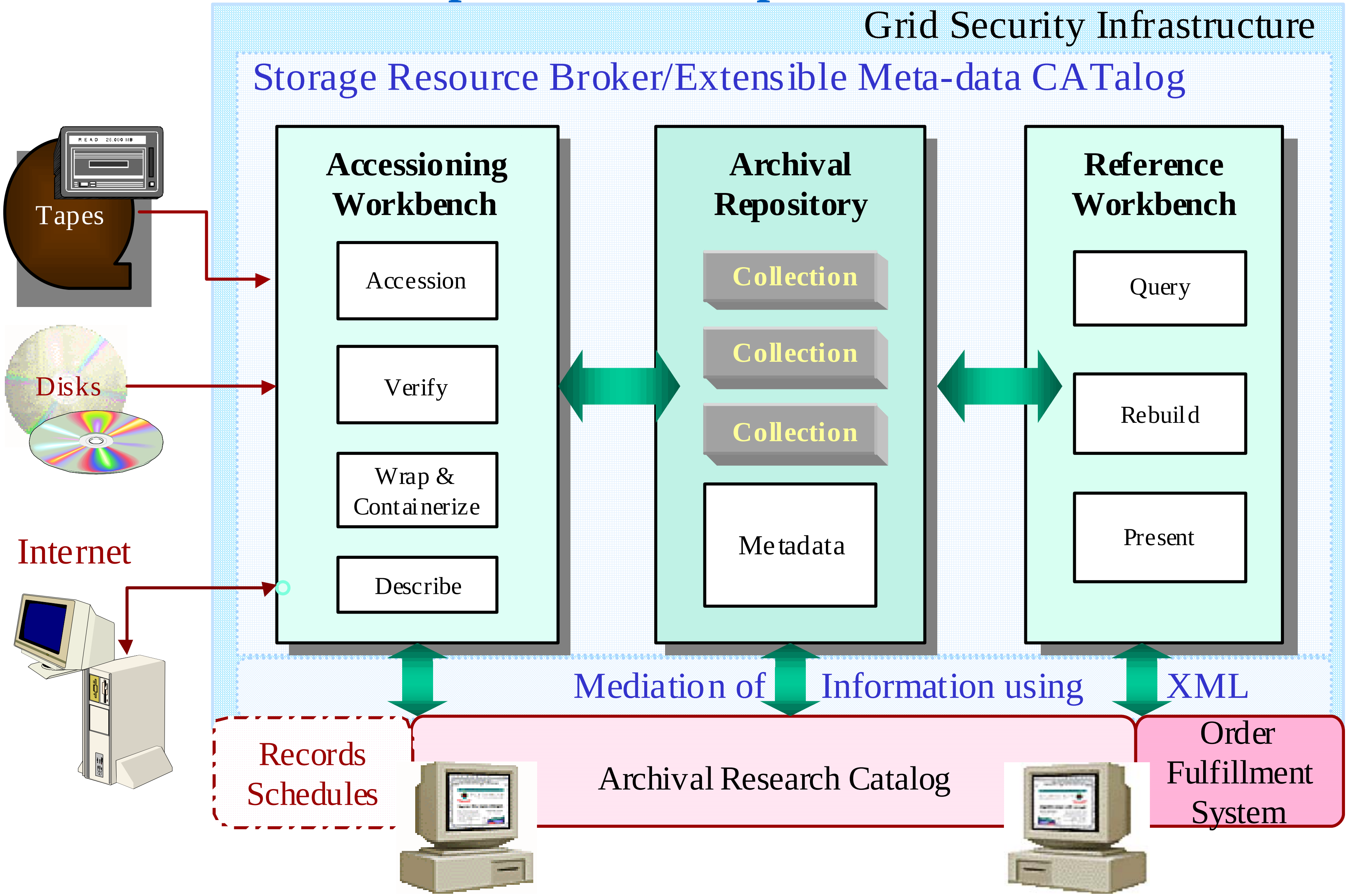
Collection Building

Persistent Archives

- **Provide interoperability mechanisms to migrate collections from old technologies to new technologies**

- **Requires ability to migrate across:**
  - Media
  - Storage systems
  - Collections
  - Information markup language standards

ERA: Archival Components Concept

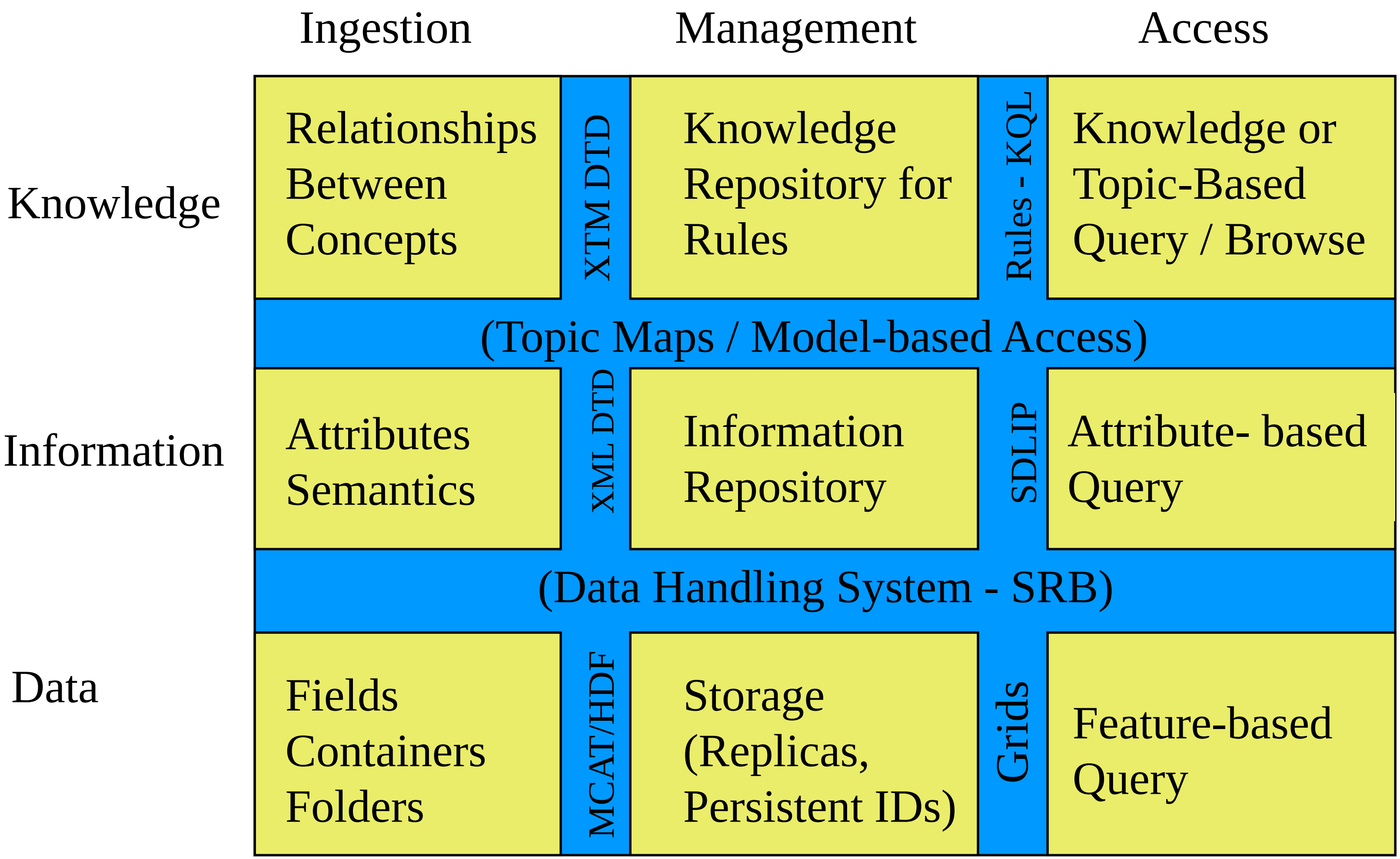


Evolution of Grids

- **File-based access**
  - Digital objects identified by path name
- **Collection-based access**
  - Digital objects identified by collection attributes
- **Knowledge-based access**
  - Digital objects identified by domain concepts

Map from concepts used by a discipline to collection attributes to local file name

Knowledge Based Grid

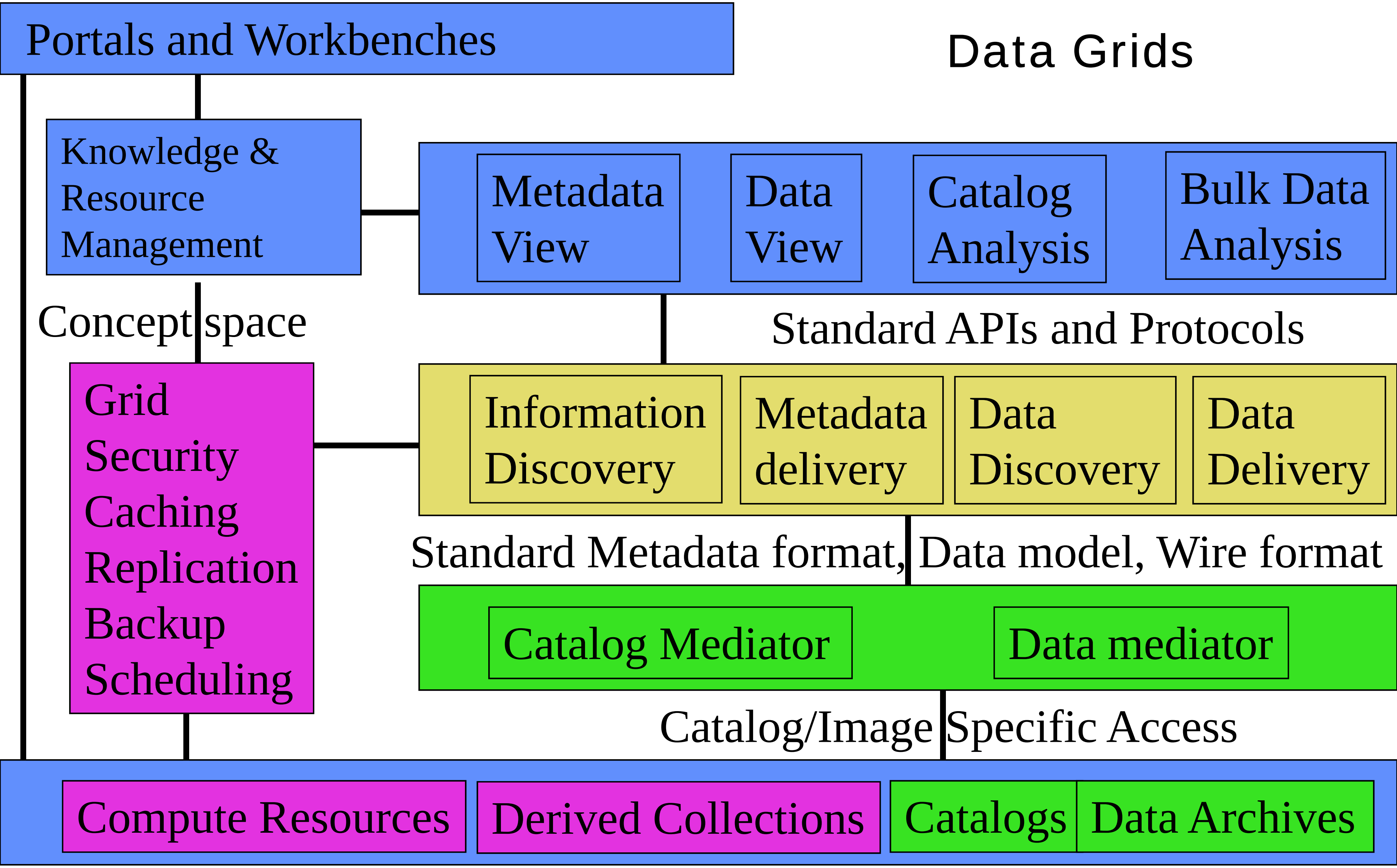


	Ingestion		Management		Access
Knowledge	Relationships Between Concepts	XTM DTD	Knowledge Repository for Rules	Rules - KQL	Knowledge or Topic-Based Query / Browse
	(Topic Maps / Model-based Access)				
Information	Attributes Semantics	XML DTD	Information Repository	SDLIP	Attribute- based Query
	(Data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

Common Web Storage Management Hierarchy

- **Knowledge-based Grids**
  - Concept based access
- **Data Grid**
  - Access to data across administration domains
- **Digital Library**
  - Services applied to information
- **Data Collection**
  - Manage information
- **Data handling**
  - Manage access to storage systems
- **Persistent Archives**
  - Manage evolution of software and hardware storage systems

Data Grids


Portals and Workbenches
Knowledge & Resource Management
Metadata View
Data View
Catalog Analysis
Bulk Data Analysis
Concept space
Standard APIs and Protocols
Grid
Security
Caching
Replication
Backup
Scheduling
Information Discovery
Metadata delivery
Data Discovery
Data Delivery
Standard Metadata format, Data model, Wire format
Catalog Mediator
Data mediator
Catalog/Image Specific Access
Compute Resources
Derived Collections
Catalogs
Data Archives

Collection Interactions

- Provide a logical representation for a collection (schema, table structure)
- Register a collection as an object within a grid
- Dynamically generate SQL commands for attribute-value based discovery
- Export data elements form collection into containers
- Manipulate containers (replicate, cache, transport)

SDSC Storage Resource Broker & Meta-data Catalog

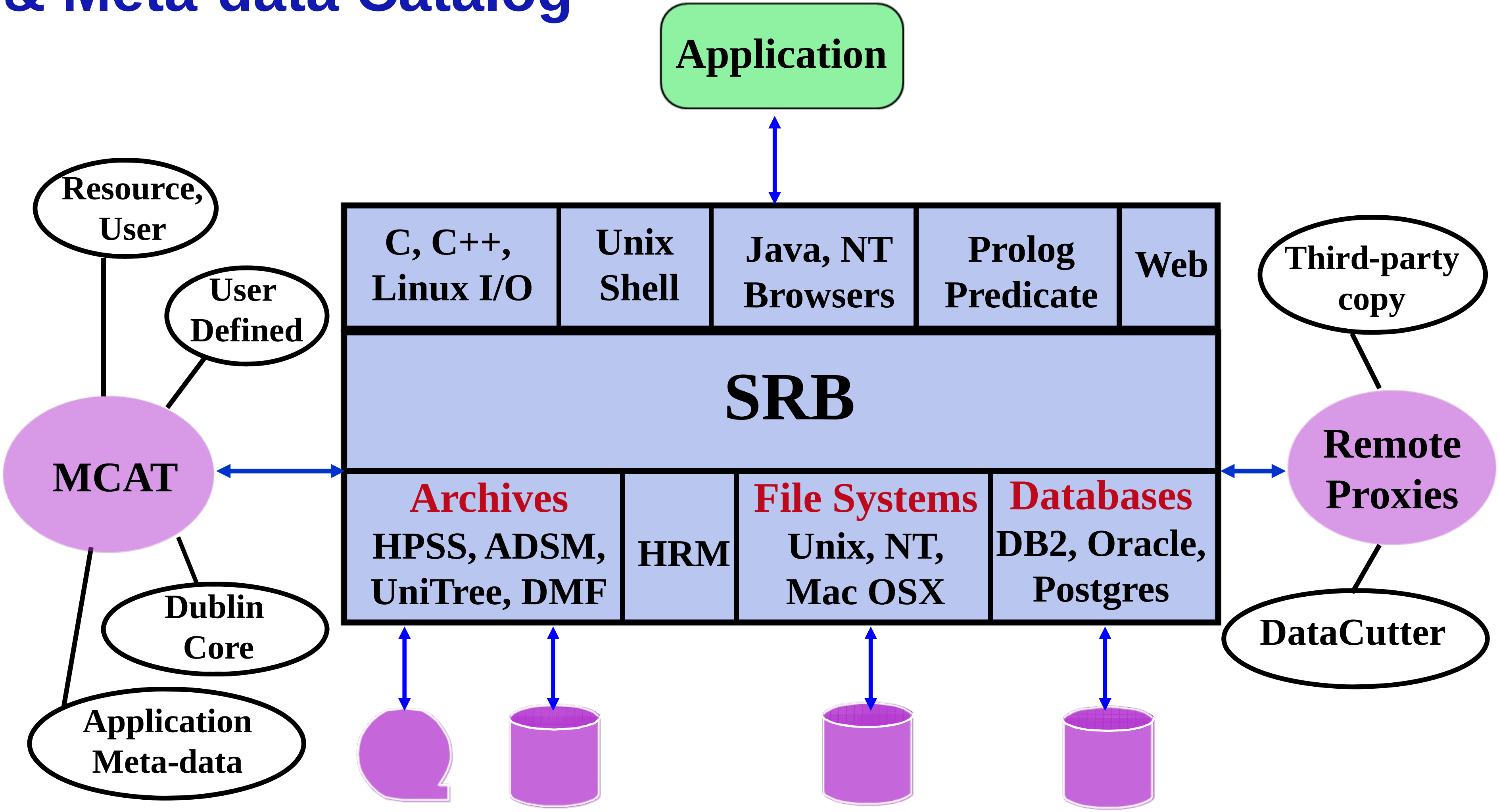


Table Access Interface

- **Facility to access tabular data using SRB API**
- **View SQL queries as Locators (Path Names or URI)**
- **Apply open, close, read, write operations**
- **Provide for very general queries to specific queries**
  - any query on a database to soft queries to hard-coded queries
- **Access Result Table as a Stream**
- **Provide Server-side operations to present results**
  - Forms, HTML, XML, …
  - Data Wetting, Charting, Visualization
- **Multi-modal Ingestion**
  - SQL ingestion
  - Packed Ingestion - useful in data movement and replication
  - Directly ingest data marked by HTML, XML, ...

Shadow Objects

- **A feature for registering partial physical locations**
  - Partial path in a file system allows one to access files under a directory
  - Partial SQL query allows for modification at access time.
- **Registering a null query allows for any query to be allowed**

Server-side Presentation

- **Markup data before sending to client**
- **Generic markup - HTML, XML**
- **Specific markup - Template**
- **Template Language**
  - Allows data element variables
  - Control structure - if-then-else, for , nested
  - Object-in-object
- **User specifies mark up at query time**
- **Can be used for other data streams also!**

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

Further Information

http://www.npaci.edu/DICE