On Building a Reference Implementation of a Preservation Environment

Reagan W. Moore

Richard Marciano

Arcot Rajasekar

Mike Wan

{moore,marciano,rajasekar,mwan}@diceresearch.org

http://irods.diceresearch.org

http://srb.diceresearch.org

UNC

Preservation Environment Reference Implementation

- **Starter kit for assembling a preservation environment**
- **Provides initial**
  - Assessment criteria
  - Preservation policies
  - Preservation procedures
  - Preservation clients
  - Preservation framework

Concepts Driving Architecture

- **Infrastructure Independence**
  - Virtualization mechanisms needed to manage distributed data
  - Virtualization mechanisms needed to establish trust
  - Virtualization mechanisms needed to enforce management policies
  - Virtualization mechanisms needed to validate assessment criteria
- **Scalability**
  - Manage collections with 100 files or 100 million files
  - Manage collections with 10 Gigabytes or 10 Petabytes
- **Federation**
  - No single preservation environment is sufficient
  - Must be able to migrate records between environments

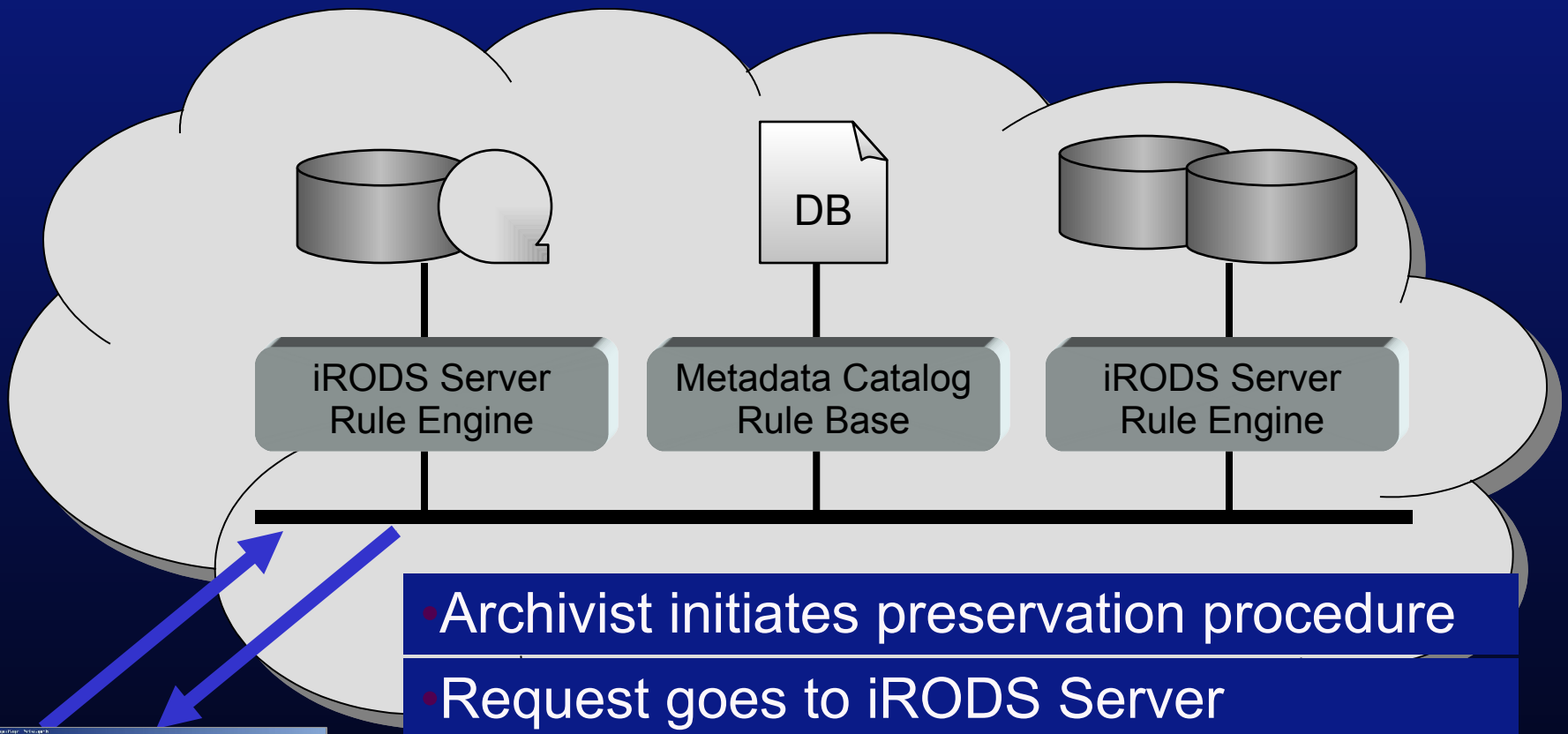
iRODS - integrated Rule-Oriented Data System

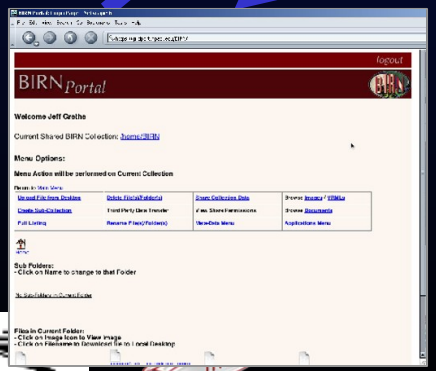
Data Management Environment	Conserved Properties	Control Mechanisms	Remote Operations
Management Functions	Assessment Criteria	Policies	Procedures
	Data grid – Management virtualization		
Data Management Infrastructure	Persistent State	Rules	Micro-services
	Data grid – Data and trust virtualization		
Physical Infrastructure	Database	Rule Engine	Storage System

Preservation Assertions

- **Authenticity**
  - Maintain links between representation information, provenance information, descriptive information and record
- **Integrity**
  - Ensure original bits have been preserved through replicas, checksums, synchronization
- **Chain of Custody**
  - Ensure preservation procedures controlled by an identified archivist
- **Original Arrangement**
  - Maintain the relationships between records
- **Trustworthiness**
  - Verify assessment criteria
  - Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist

Distributed Preservation Framework

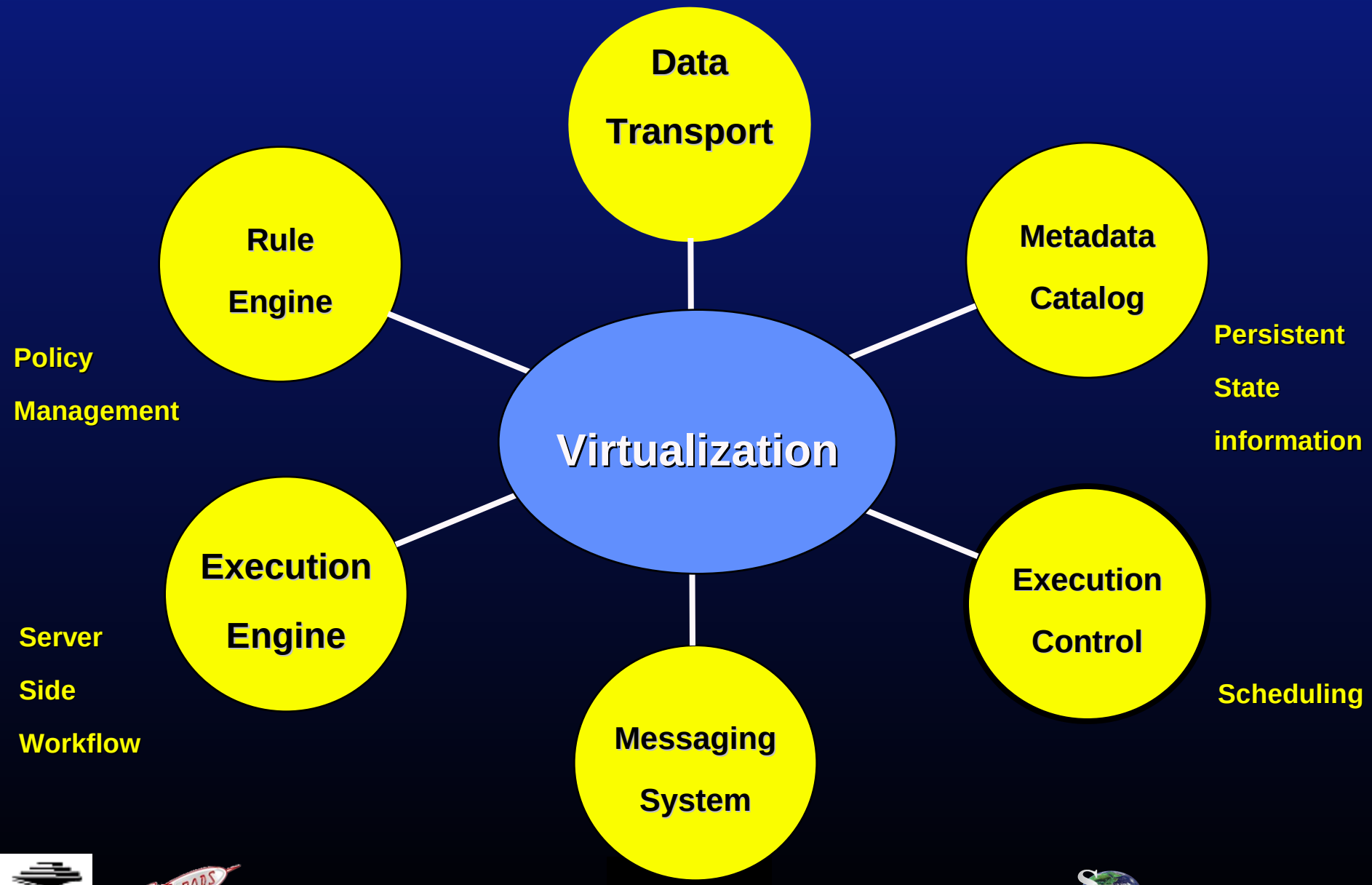




- Archivist initiates preservation procedure
- Request goes to iRODS Server
- Server looks up information in catalog
- Catalog tells which iRODS server has data
- 1st server asks 2nd to perform procedure
- The 2nd iRODS server applies policies

UCSD

Distributed Management Framework



Preservation Procedures

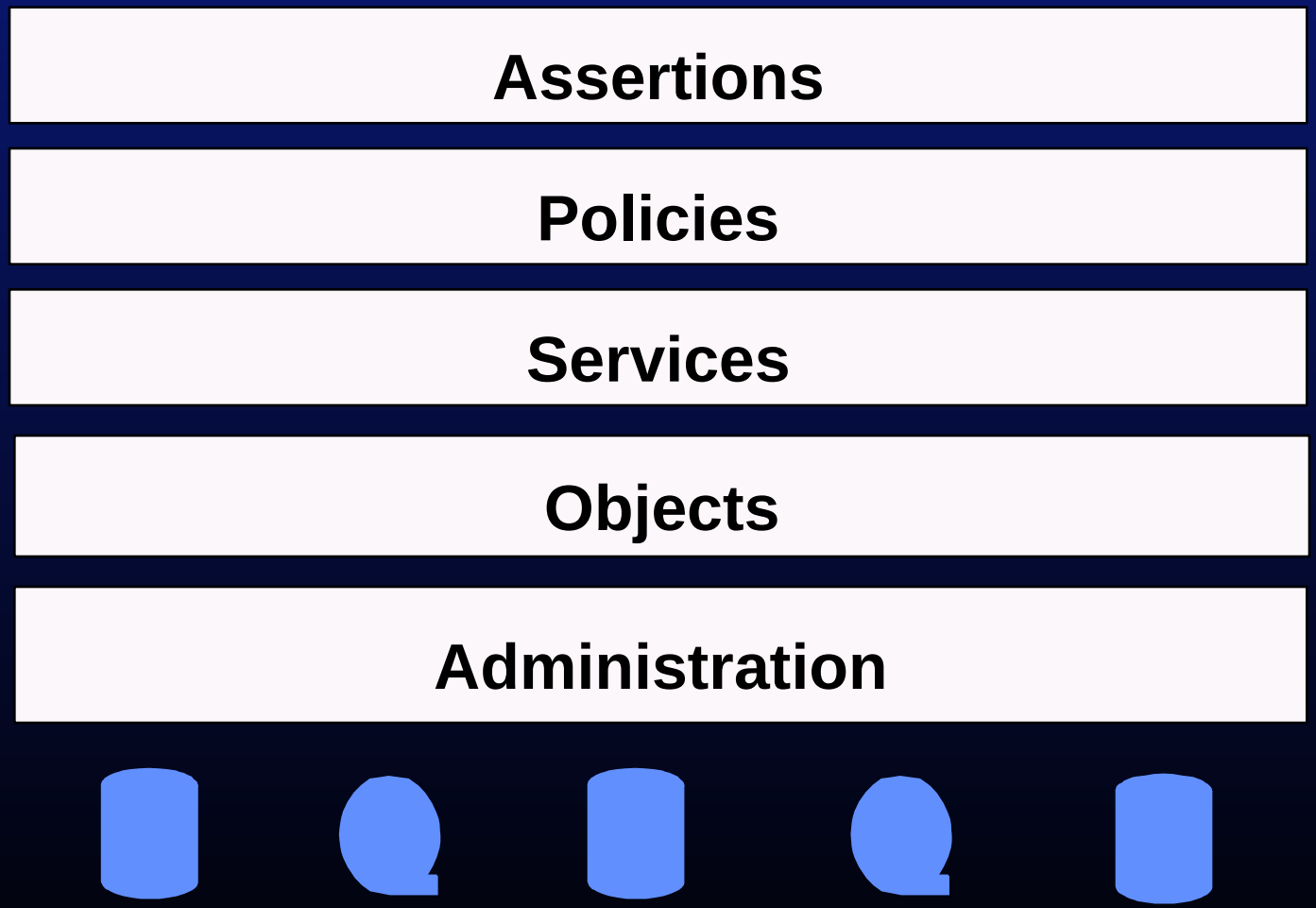
- Appraisal
- Accession
- Arrangement
- Description
- Preservation
- Access

- Examples of 200 executable procedures definable from the Electronic Records Archive capabilities list

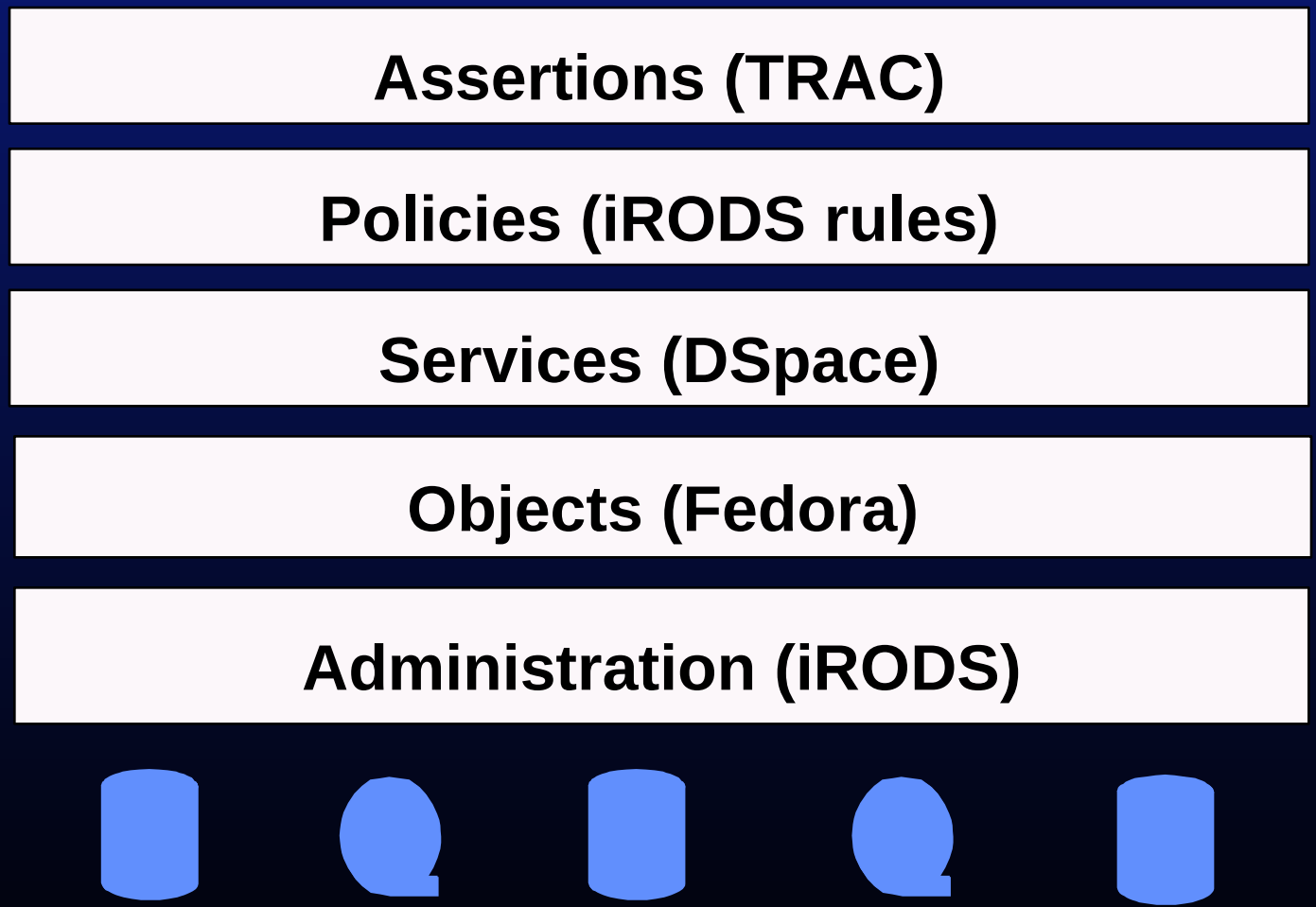
Multiple Policy Control Levels

- **Access level**
  - DSpace - accession procedures
- **Digital library middleware level**
  - Fedora - relationships between record
- **Data grid**
  - iRODS - administrative policies
- **Storage system**
  - Physical control

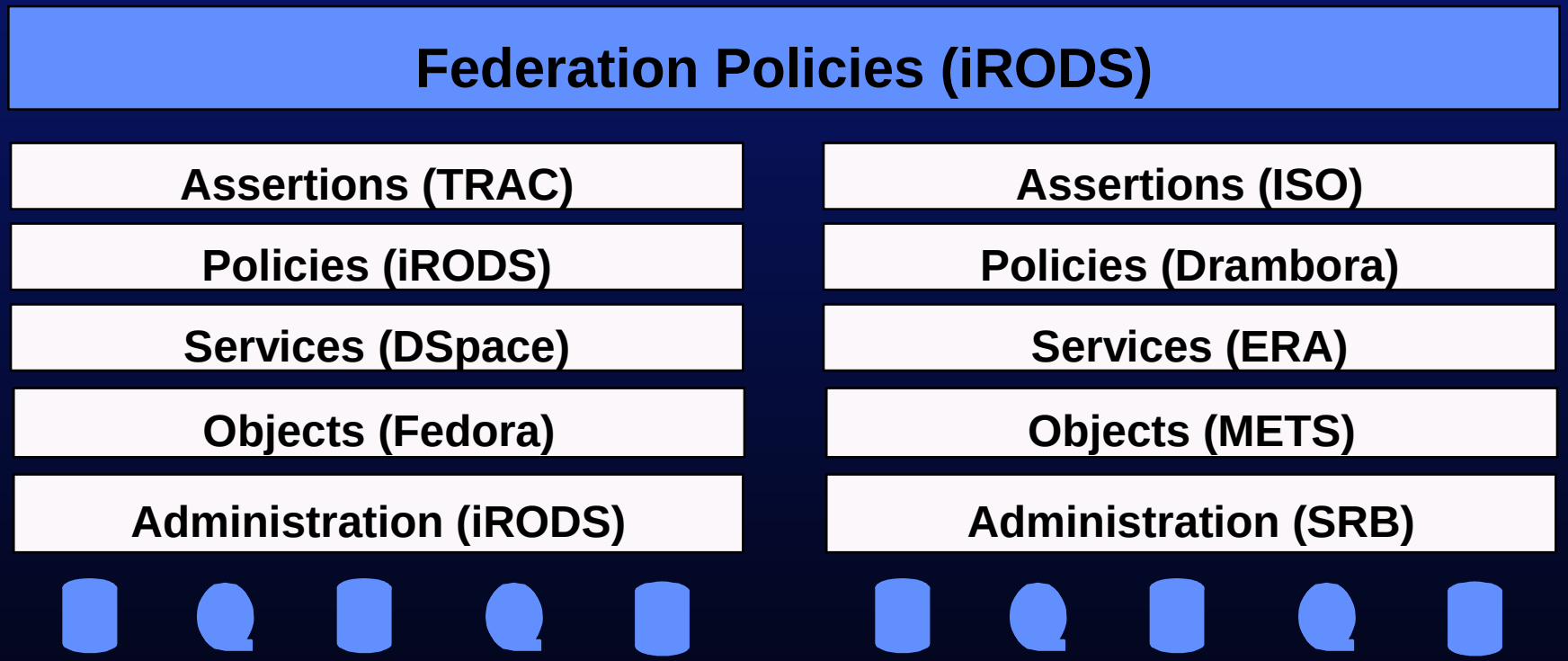
Preservation Environment Reference Implementation



Example Starter Kit



Federation of Preservation Environments



Assertion

- **Data management applications apply many of the same procedures**
  - Data format parsing
  - Metadata manipulation
  - Data administration tasks
- **Each application applies different management policies**
  - Migrate records between data life cycle stages by changing the management policies
  - Can create generic infrastructure that can be used to implement collections, digital libraries, persistent archives

Reference Implementations

- **Data grids**
  - **Share data** - organize distributed data as a collection
- **Digital libraries**
  - **Publish data** - support browsing and discovery
- **Persistent archives**
  - **Preserve data** - manage technology evolution
- **Real-time sensor systems**
  - **Federate sensor data** - integrate across sensor streams
- **Workflow systems**
  - **Analyze data** - integrate client- & server-side workflows

For More Information

Reagan W. Moore
University of North Carolina, Chapel Hill
rwmoore@renci.org

http://irods.diceresearch.org