



Policy-based Data Management

integrated Rule Oriented Data System (**SDCI**)

Reagan W. Moore (DICE-UNC)

Arcot Rajasekar (DICE-UNC)

<http://irods.diceresearch.org>

<http://datafed.org>



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





iRODS

❑ Integrated Rule Oriented Data System

- DICE – Reagan Moore
- Concepts – Arcot Rajasekar
- Architect – Mike Wan
- Security / metadata / production – Wayne Schroeder
- Rule engine – Hao Xu
- User interface (Java) – Mike Conway
- Applications – Antoine de Torcy
- Administration – Sheau-Yen Chen

❑ E-iRODS (enterprise version developed by RENCi)

- Management – Charles Schmitt
- Production version – Jason Coposky
- Test environment – Terrell Russell
- Administrator interface – Lisa Stillwell
- Tutorials – Leesa Brieger



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL



renci



Examples of “National” Infrastructure

□ Data Grids

(data sharing)

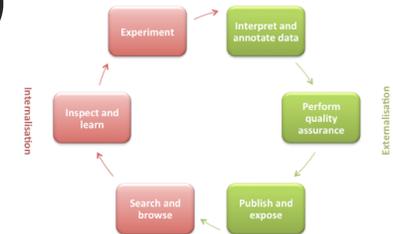
- National Optical Astronomy Observatory
- Ocean Observatories Initiative
- The iPlant Collaborative
- Babar High Energy Physics
- Broad Institute genomics data grid
- WellCome Trust Sanger Institute genomics data grid



□ Digital Libraries

(data publication)

- French National Library
- Texas Digital Library
- UNC-CH SILS LifeTime Library



□ Repositories / Archives

(data preservation)

- NASA Center for Climate Simulation
- Carolina Digital Repository



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Approaches

- ❑ **National data grid**
 - Single system supporting data sharing across institutions
 - Australian Research Collaboration Service
 - Top down approach
- ❑ **Federation environment**
 - Establish trust mechanisms to enable data access between systems
- ❑ **Collaboration Environment**
 - Support data sharing across community resources
 - Requires interoperability mechanisms to enable access to remote repositories
 - Register data into a logical collection
 - Bottom-up federation of existing data management systems

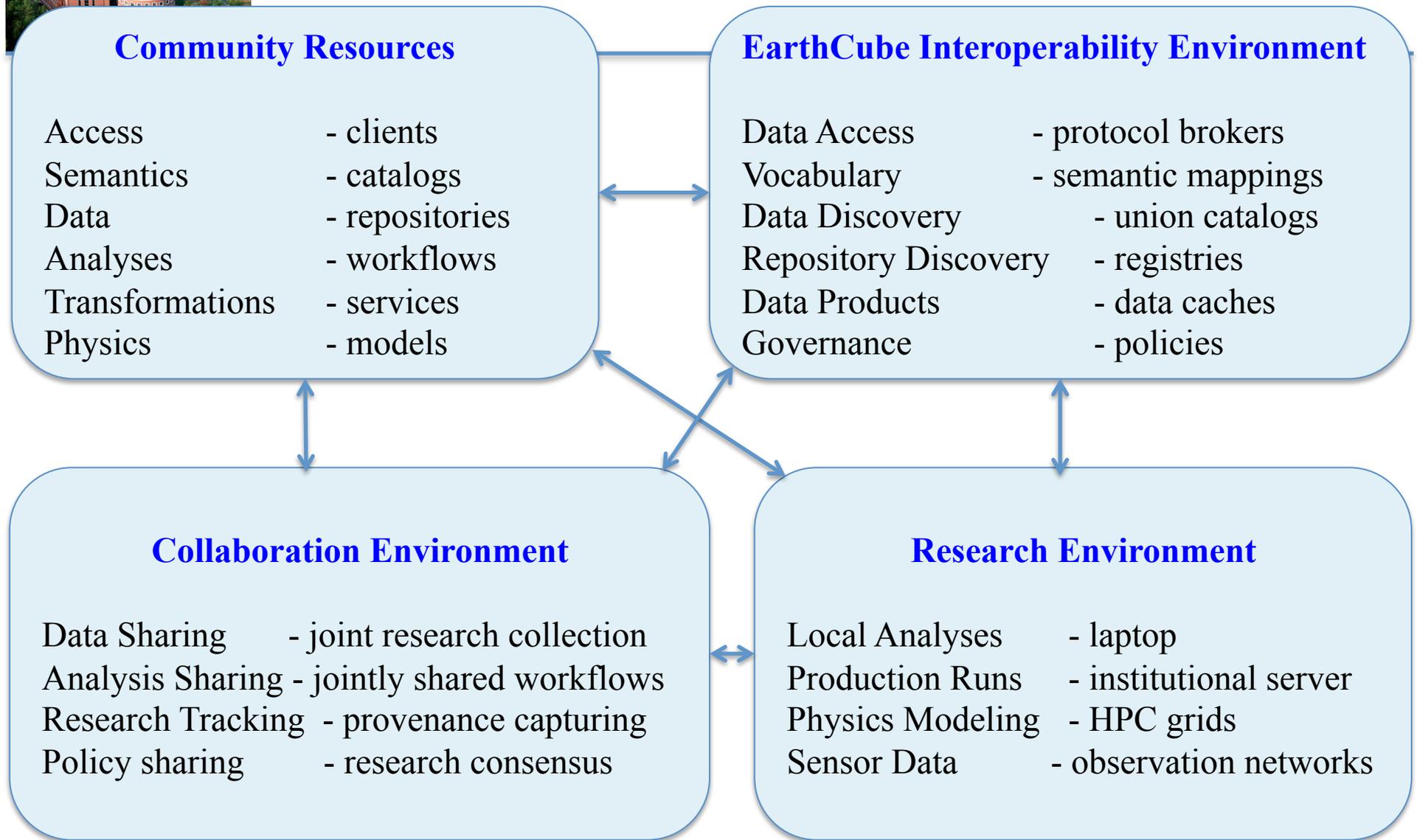


THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





EarthCube Infrastructure Components



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Policy-Based Data Environments

- ❑ **Purpose**
 - Reason a collection is assembled
- ❑ **Properties**
 - Attributes needed to ensure the **purpose**
- ❑ **Policies**
 - Controls for enforcing desired **properties**,
 - mapped to computer actionable rules
- ❑ **Procedures**
 - Functions that implement the **policies**
 - Mapped to computer executable workflows
- ❑ **Persistent state information**
 - Results of applying the **procedures**
 - mapped to system metadata
- ❑ **Property verification**
 - Validation that **state information** conforms to the desired **purpose**
 - mapped to periodically executed policies



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Goals and Impact

- ❑ **Collaborative research**
 - Sharable collections
 - Sharable workflows
- ❑ **Reproducible science**
 - Automate data retrieval, transformation
 - Re-execution and provenance of workflows
- ❑ **Reference collections**
 - Community knowledge resources (catalogs, repositories)
 - Manage data life cycle through evolution of policies as user community broadens
- ❑ **Student participation in research**
 - Policy controlled research analyses



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





DFC Vision - Data Driven Science

□ Enable reproducible science through collaborative research on shared workflows and data collections

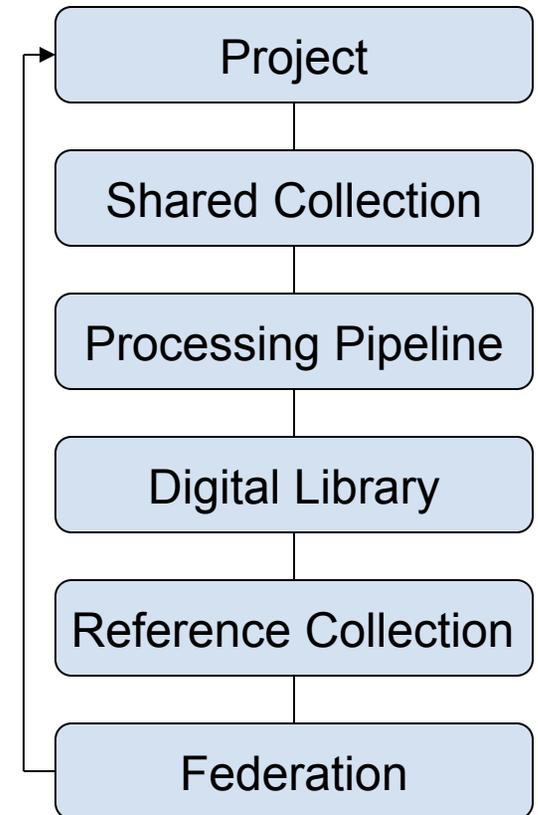
- Researcher management of workflows and data
- Policy-based management of entire scientific data life cycle from data analysis pipelines to long-term sustainability of reference collections

□ Implement NSF national scale data cyber-infrastructure

- Federation of exemplar data management technologies from national research initiatives
- Provision of interoperability mechanisms
- Proven technology implemented in extant data grids

□ Integrate “live” research data collections into education initiatives

- Student digital libraries accessing national data sets



**Community-based
Collection Life Cycle**



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Community-based Collection Life Cycle

The driving purpose changes at each stage of the data life cycle

Project Collection	Data Grid	Data Processing Pipeline	Digital Library	Reference Collection	Federation
Private	Shared	Analyzed	Published	Preserved	Sustained
Local Policy	Distribution Policy	Service Policy	Description Policy	Representation Policy	Re-purposing Policy

Stages correspond to addition of new policies for a broader community
Virtualize the stages of the collection life cycle through policy evolution



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Building Community Resources

- ❑ **Digital libraries use collections to define context**
 - Provenance information
 - Descriptive information
 - Administrative information
- ❑ **Policy-based data management use procedures to encapsulate domain knowledge**
 - Workflows for generation of data
 - Workflows for administration of data
 - Workflows for enforcement of management policies
 - Workflows for verifying collection properties



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Computer Actionable Knowledge

<input type="checkbox"/> Data	objects	bits
<input type="checkbox"/> Information	names	metadata
<input type="checkbox"/> Knowledge	relationships between names	procedures
<input type="checkbox"/> Wisdom	relationships between relationships	policy points
<input type="checkbox"/> Data	bits	Posix I/O
<input type="checkbox"/> Information	metadata	Relational database
<input type="checkbox"/> Knowledge	procedures	Workflows
<input type="checkbox"/> Wisdom	policy points	Rule engine

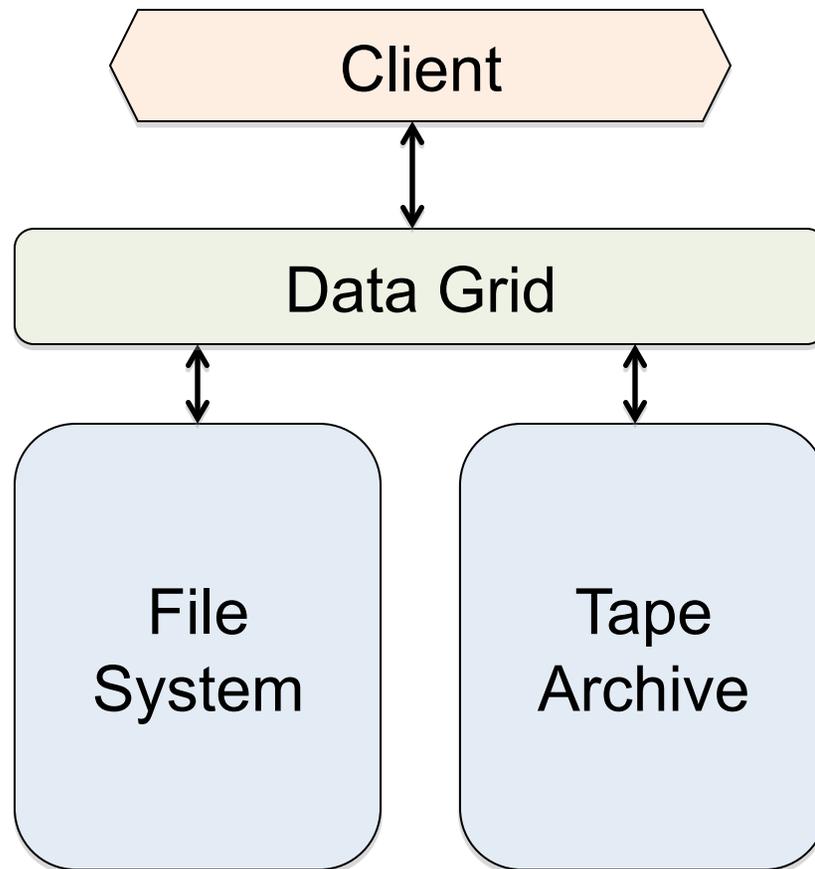


THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Shared Collections – Data Grid



50 clients: web browser, unix shell command, ...

Data grid middleware provides global name, single sign-on, policy enforcement, metadata, replication

Multiple types of systems can be used to store data

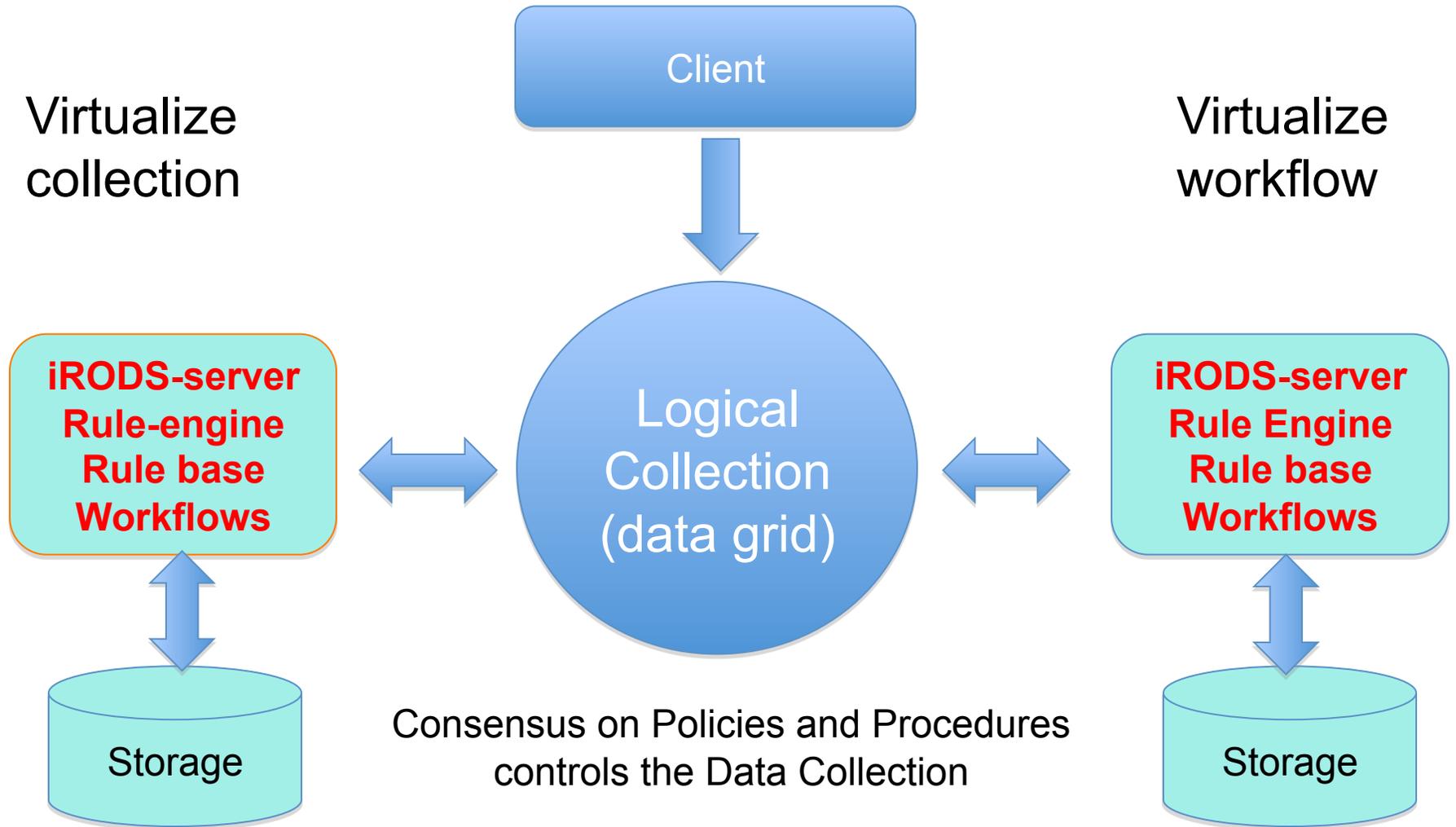


THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Policy-based Data Management

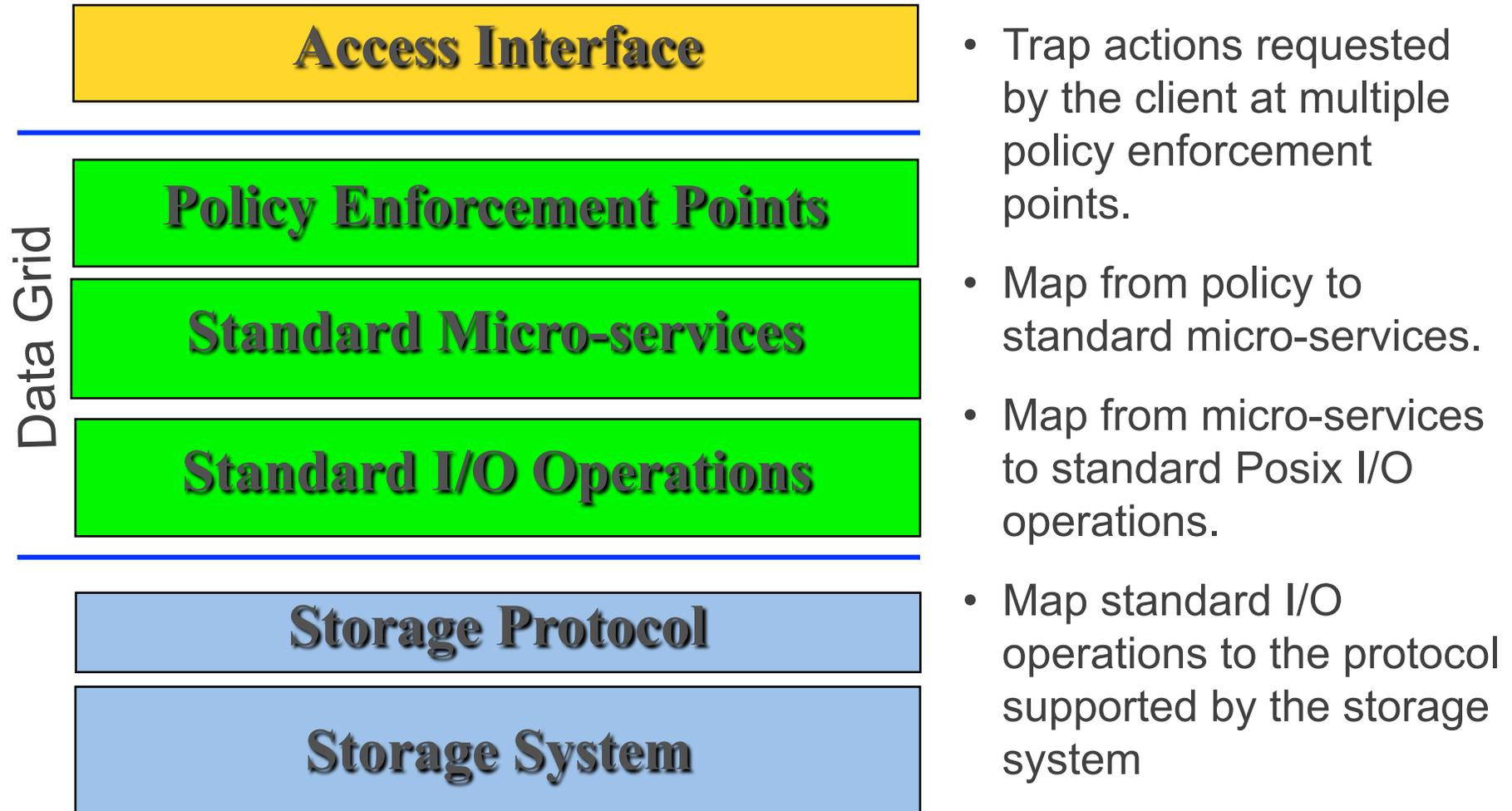


THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Data Workflow Virtualization



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Virtualization Mechanisms

Name Space	Operations	Virtualization interface
Users	Authentication, authorization, groups	GSSAPI / PAM
Objects	Partial I/O, move, copy, replicate, share	Posix I/O & staging
Collections	Organization, Browsing	System metadata
State information	Add, update, delete, query	Catalog interface to DBMS
Resources	Load leveling, fault tolerance, grouping	Storage drivers
Policies	Management, administrative, verification	Policy language
Procedures	Basic functions on each name space	Workflows

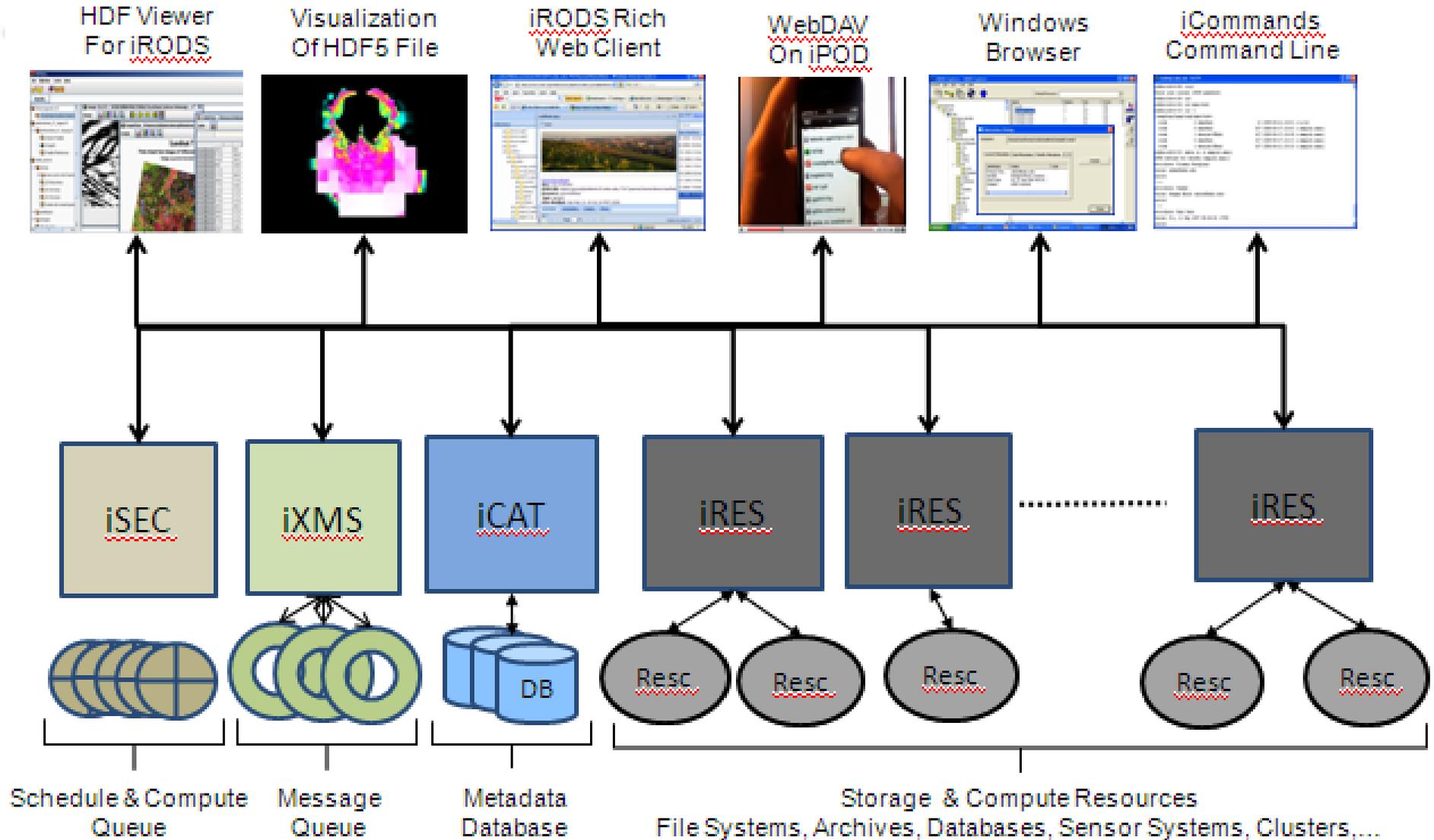


THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





iRODS Distributed Data Management



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Rule to count metadata values

```
myTestRule {
#Input parameters are:
# String with conditional query
#Output parameter is:
# Result string
  msiExecStrCondQuery(*Select,*QOut);
  foreach(*QOut) {
    msiPrintKeyValPair("stdout",*QOut)
  }
}
INPUT *Select=$"SELECT count(META_DATA_ATTR_VALUE),
order(META_DATA_ATTR_NAME), META_DATA_ATTR_NAME where
COLL_NAME like '/lifelibZone/home/rwmoore%%'"
OUTPUT ruleExecOut
```



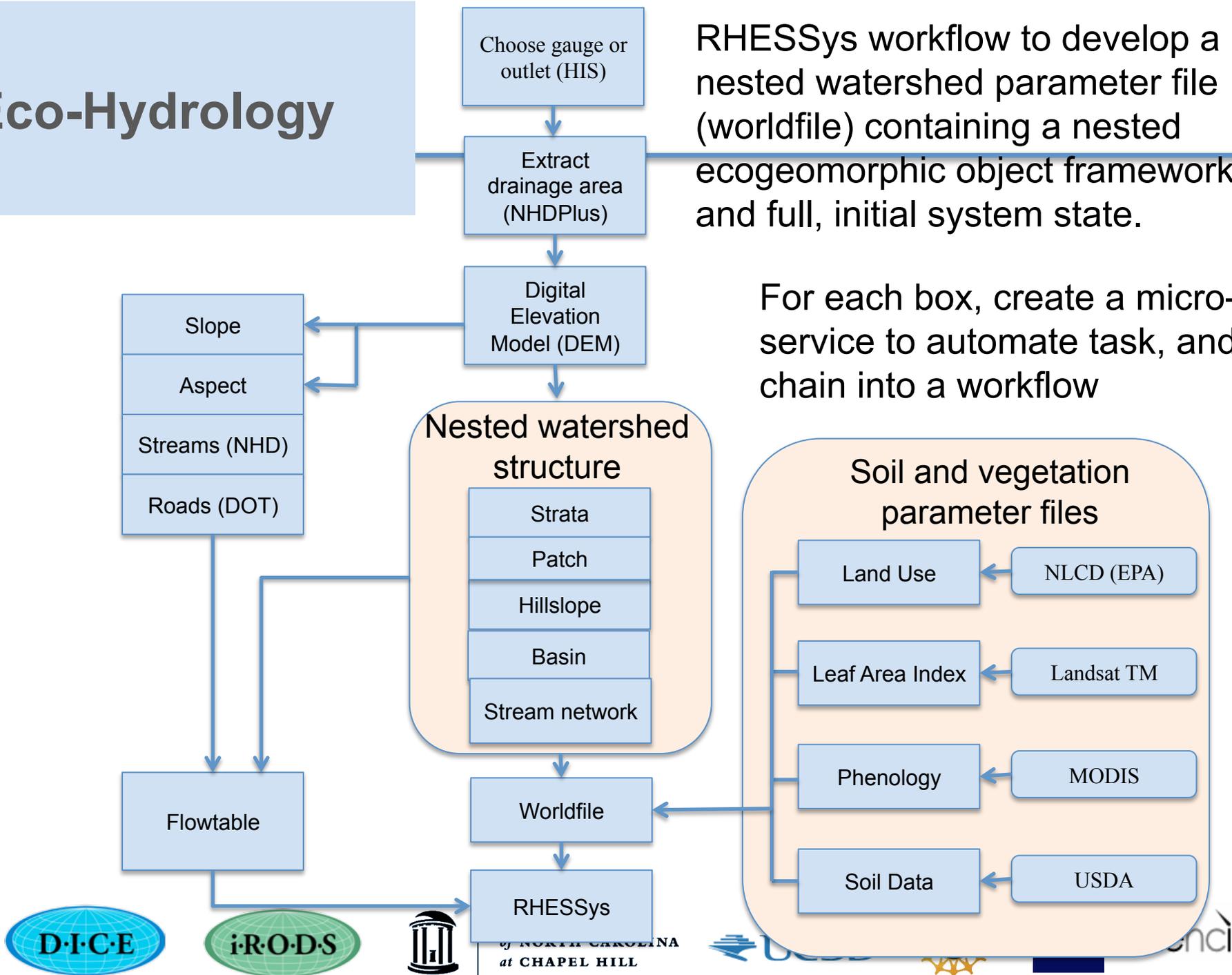
THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL



Eco-Hydrology

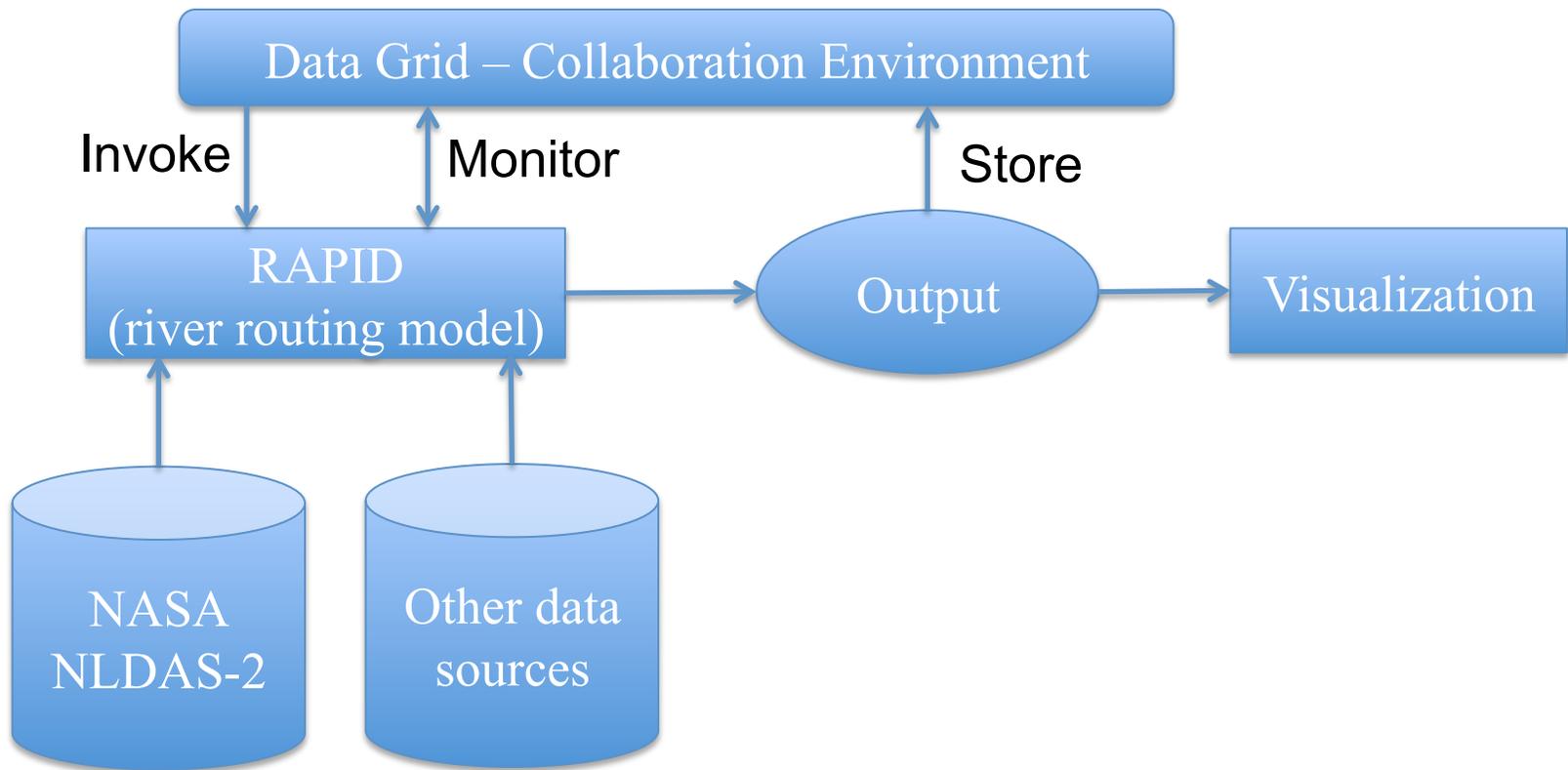
RHESSys workflow to develop a nested watershed parameter file (worldfile) containing a nested ecogeomorphic object framework, and full, initial system state.

For each box, create a micro-service to automate task, and chain into a workflow





Event-Driven Real-Time Drought Analysis/Prediction Workflow



<http://rapid.ncsa.illinois.edu:8080/rapid/>



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





SILS LifeTime Library

□ Student digital libraries

- Enable students to build collections of
 - ✓ Photographs
 - ✓ MP3 audio files
 - ✓ Class documents
 - ✓ Video
 - ✓ Web site archive

□ Resources provided by School of Information and Library Science at UNC-CH

- Student collections range from 2 GBytes to 150 Gbytes
- Number of files from 2000 to 12,000



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





SILS LifeTime Library Policies

□ Library management

- Replication
- Checksums
- Versioning
- Strict access controls
- Quotas
- Metadata catalog replication
- Installation environment archiving

□ Ingestion

- Automated synchronization of student directory with LifeTime Library
- Automated loading of MP3 metadata



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Student Collections

# files	# metadata	Size	#metadata/ file	Collection	Metadata type	Metadata load
2111	8684	16.0 GB	4.1	iTunes	AVUs	XML load
2734	4500	4.3 GB	1.6	Photo	Tags	Hand
1109	8174	1.2 GB	7.4	Photo, Music	Tags, AVUs	Hand
5697	15472	47.0 GB	2.7	iTunes	AVUs	ASCI load
1692	8098	0.1 GB	4.8	Photo	AVUs	Hand
125	1100	0.8 GB	8.8	iTunes	AVUs	XML load

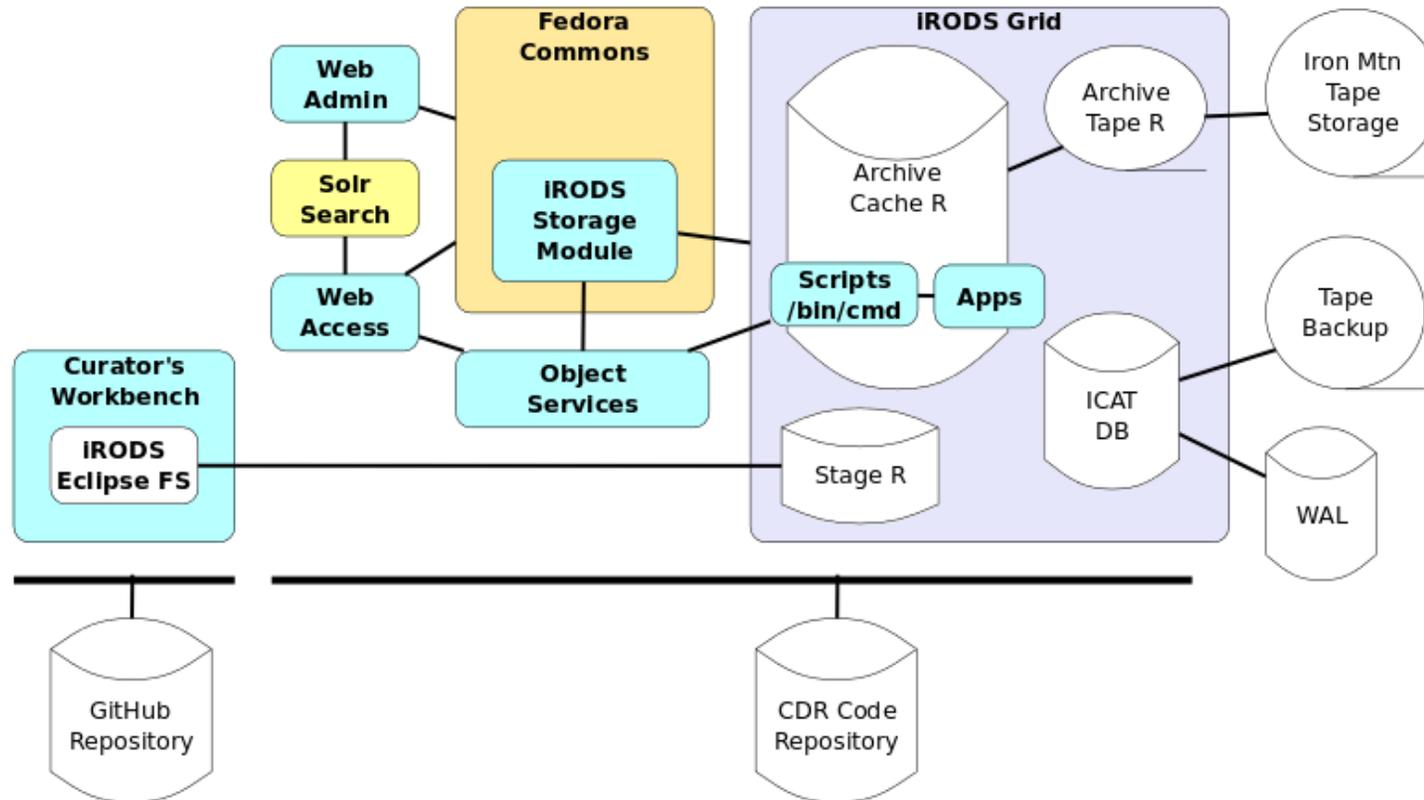


THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Carolina Digital Repository



Policy-Driven Repository Infrastructure project
funded by the Institute for Museum and Library Services

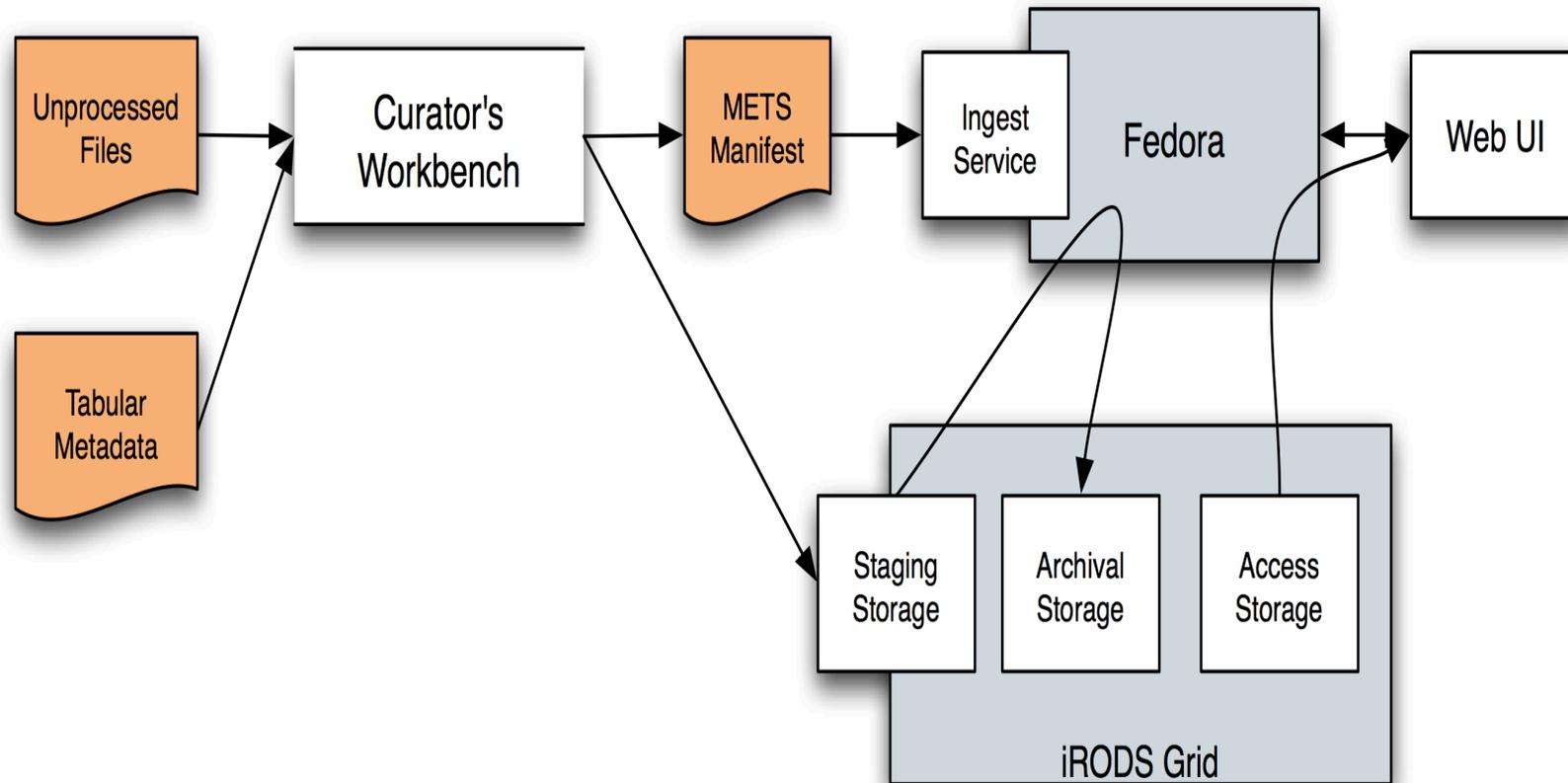


THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Carolina Digital Repository Ingest Workflow

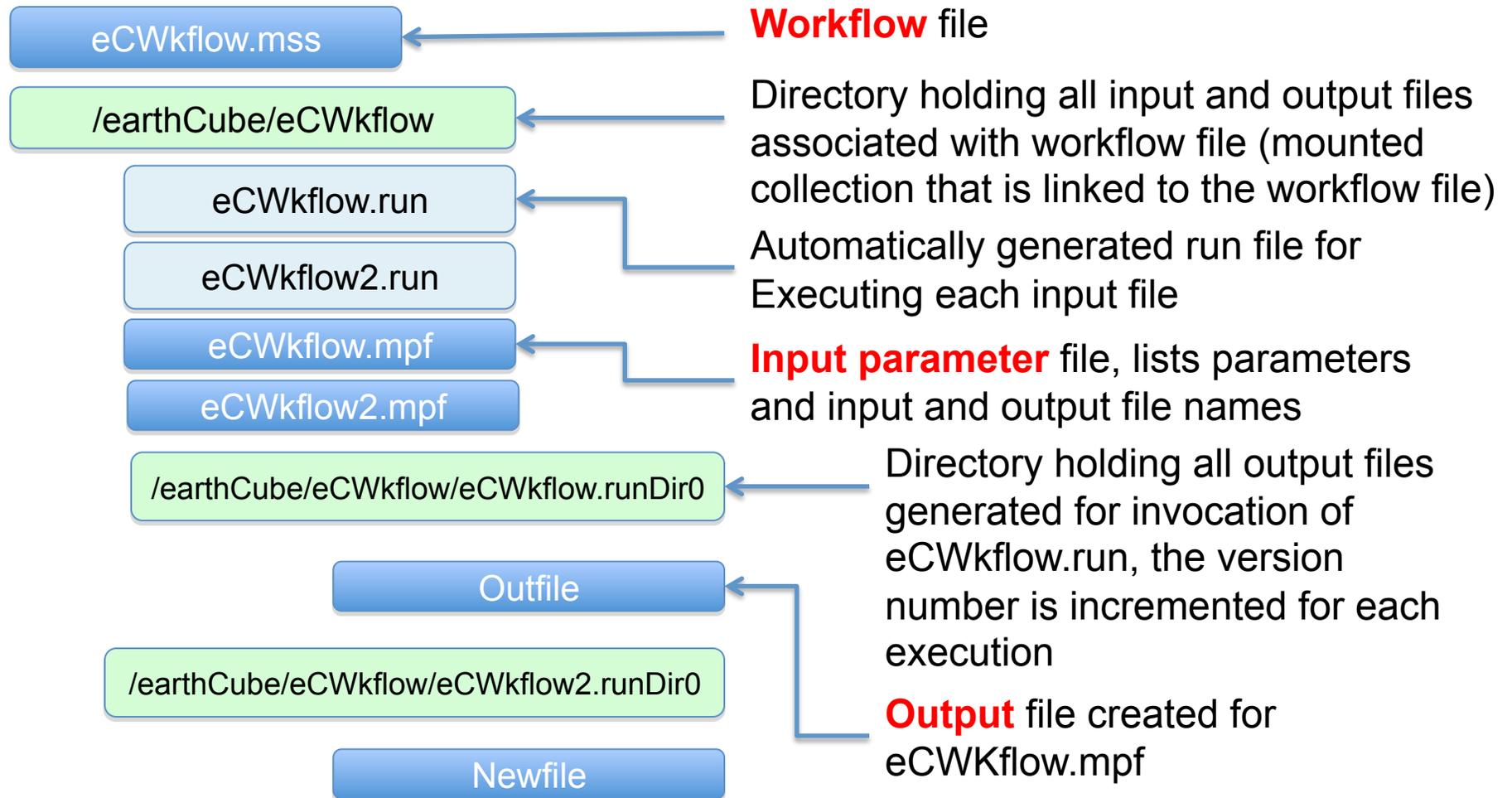


THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Capturing Workflow Provenance



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





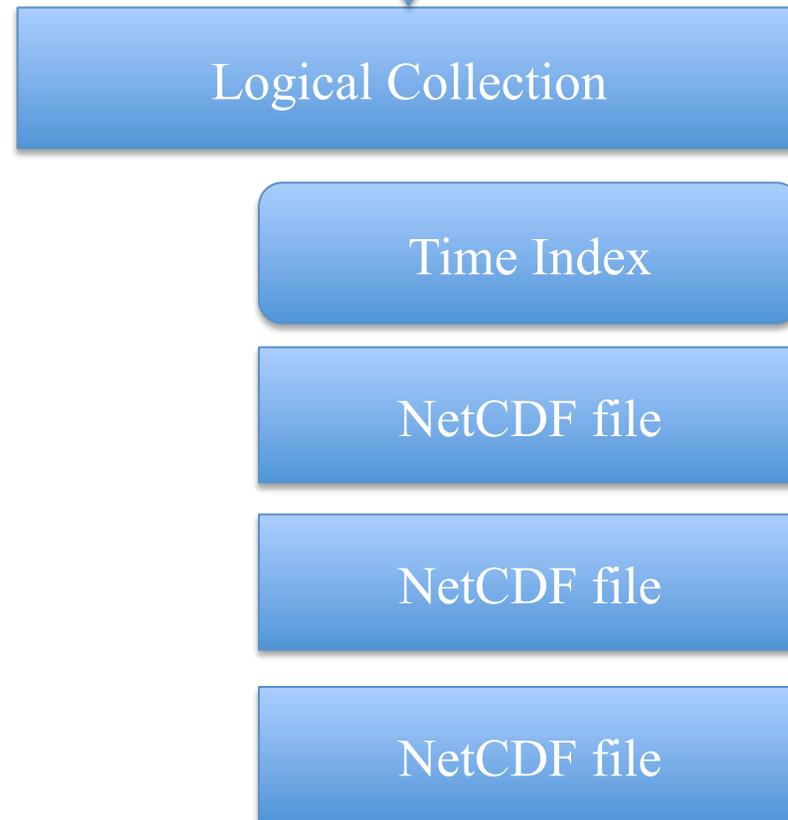
Automating Time Series Data Access

Client
Requests time period

Data grid automatically generates a time index into all files deposited into the collection.

Each access defines the desired time period, and the data grid retrieves data from the relevant files.

Being developed for iRODS 3.3 for use by OOI



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Publications

- ❑ Rajasekar, R., M. Wan, R. Moore, W. Schroeder, S.-Y. Chen, L. Gilbert, C.-Y. Hou, C. Lee, R. Marciano, P. Tooby, A. de Torcy, B. Zhu, “iRODS Primer: Integrated Rule-Oriented Data System”, Morgan & Claypool, 2010.
- ❑ Ward, R., M. Wan, W. Schroeder, A. Rajasekar, A. de Torcy, T. Russell, H. Xu, R. Moore, “The integrated Rule-Oriented Data System (iRODS 3.0) Micro-service Workbook”, DICE Foundation, November 2011, ISBN: 9781466469129, Amazon.com



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





iRODS - Open Source Software

- ❑ <http://irods.diceresearch.org>
- ❑ **Distributed under BSD license**
 - Current version is iRODS 3.2
 - Typically have three releases per year
 - Scale of capabilities:
 - 338 system attributes (users, files, collections, resources, rules)
 - 272 basic functions (micro-services)
 - 80 policy enforcement points
 - Downloads
 - 39 countries
 - 62 US academic institutions



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





Future Directions

- ❑ **Integration with storage controllers**
 - DDN SFA12KE – provides virtual machine environments within the storage controller
 - Run an iRODS data grid in the storage controller
 - Automate the application of policies for feature detection, indexing, metadata extraction
- ❑ **Integration with Software Defined Network Overlays for Future Internet Architecture**
 - Integrate virtual networks with virtual collections
 - Use network policies to support addressing by file name, access controls in the network, data caching in the network, data distribution



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL





iRODS - Open Source Software

Reagan W. Moore

rwmoore@renci.org

<http://irods.diceresearch.org>

NSF OCI-0940841 “DataNet Federation Consortium”

NSF OCI-1032732 “Improvement of iRODS for Multi-Disciplinary Applications”

NSF OCI-0848296 “NARA Transcontinental Persistent Archives Prototype”

NSF SDCI-0721400 “Data Grids for Community Driven Applications”



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL



renci