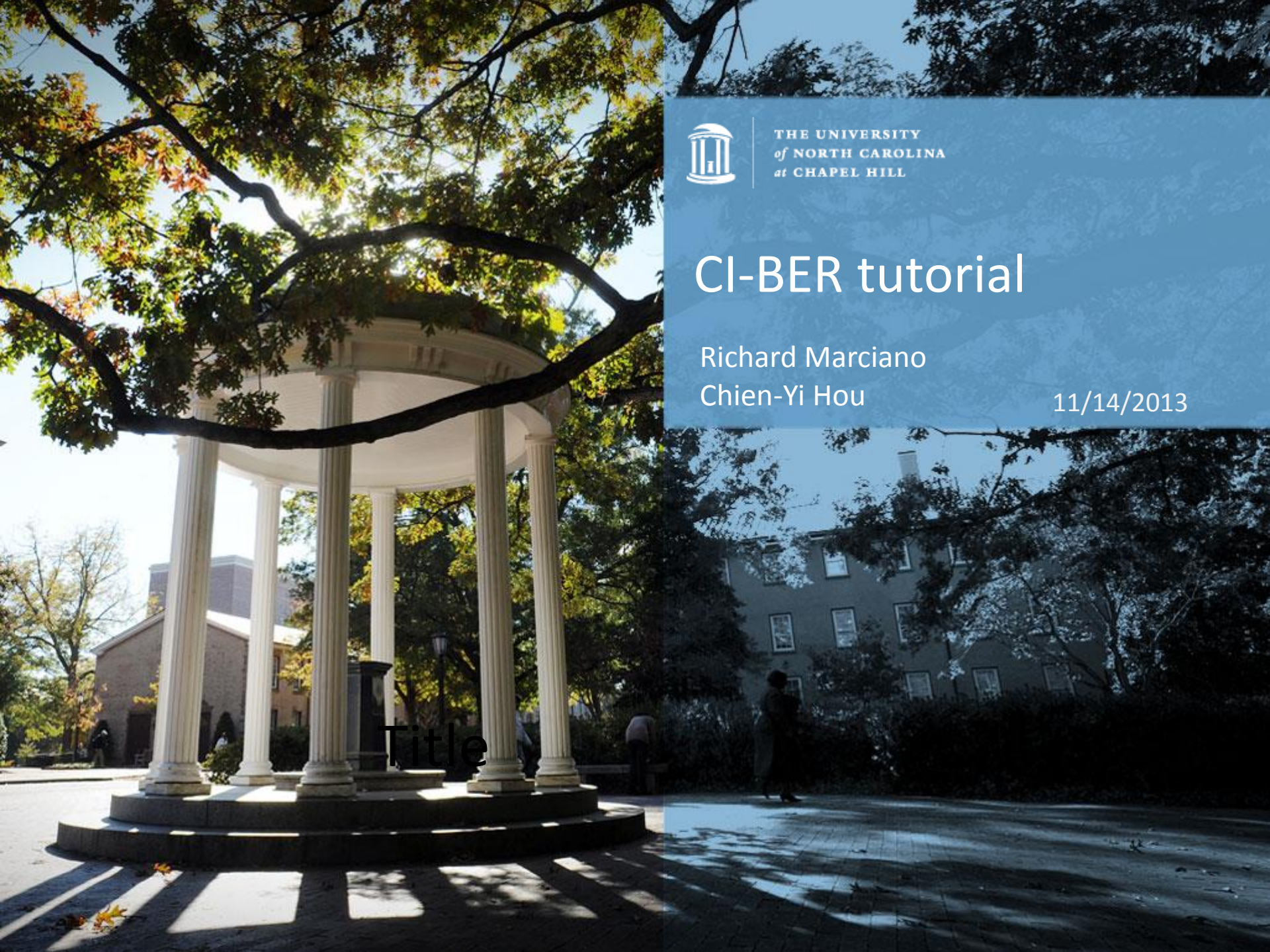

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL
CI-BER tutorial
Richard Marciano
Chien-Yi Hou
11/14/2013
Title

Our goal today…

- Learn how to access CI-BER data collection

Outline

- Part 1
  - Introduction to iRODS
  - What data do we have in CI-BER?
- Part 2
  - Hands on with icommands

What is iRODS?

- is software middleware that manages a highly controlled collection of **distributed** digital objects, while enforcing **user-defined Management Policies** across the multiple storage locations.
- open source, BSD license

How iRODS is used?

- A **data grid** for sharing data across collaborations
- a **digital library** for publishing data
- a **preservation environment** for long-term data retention
- a **data workflow** for data processing
- a **system** for federating real-time sensor data streams

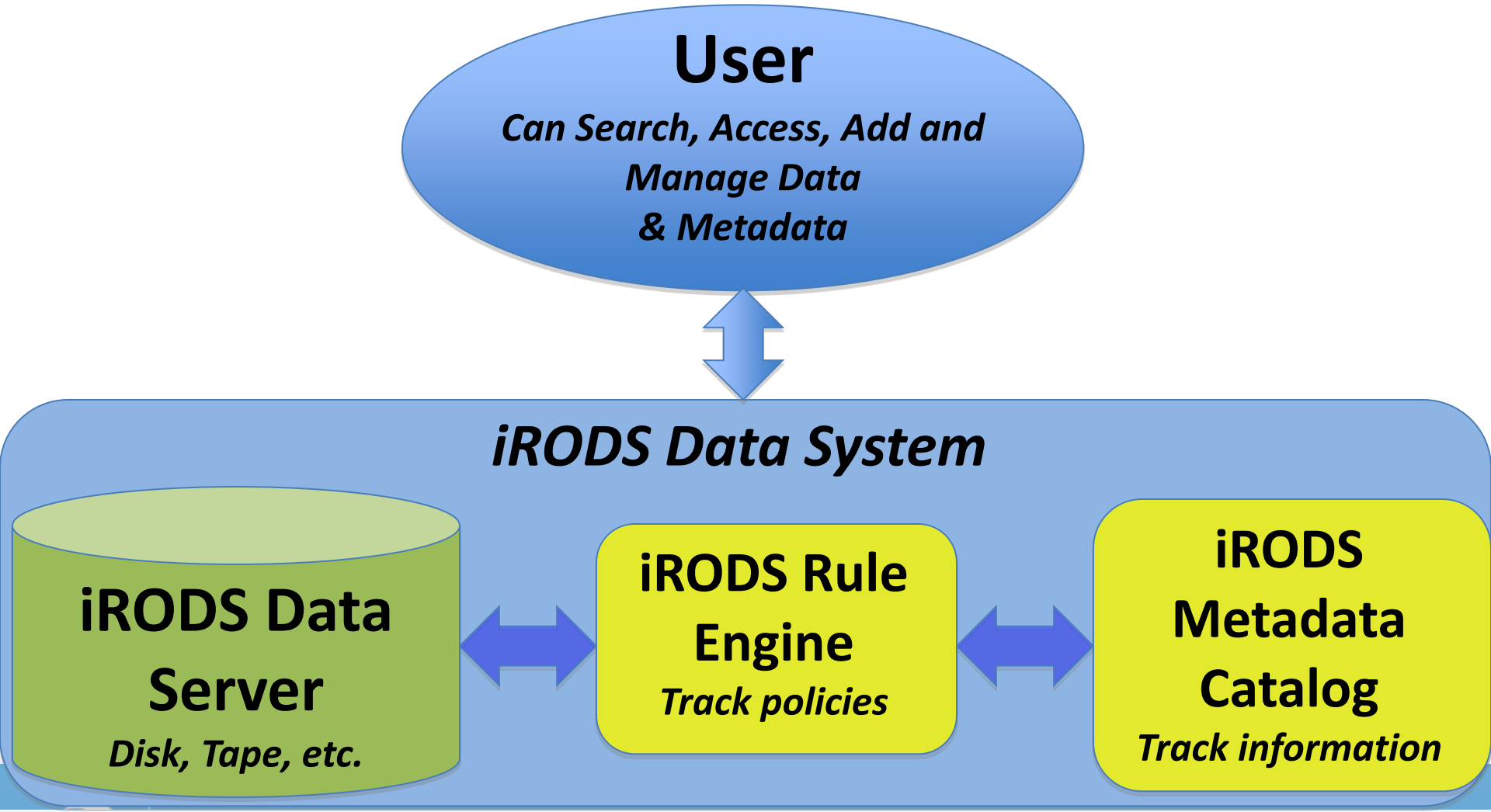
Organize Distributed Data into a Sharable Collection

- Project repository
  - MotifNet - manage collection of analysis products
- Institutional repository
  - Carolina Digital Repository for UNC collections
- Regional collaboration
  - RENCI Data Grid linking resources across North Carolina
- National collaboration
  - NSF Temporal Dynamics of Learning Center
  - Australian Research Collaboration Service
- National Library
  - French National Library
- National Archive
  - NARA Transcontinental Persistent Archive Prototype, Taiwan
- International collaboration
  - BaBar High Energy Physics (SLAC-IN2P3)
  - National Optical Astronomy Observatory (Chile-US)

iRODS Glossaries

- Object
- Metadata
- Resource
- Micro-service
- Rule

Overview of iRODS Architecture

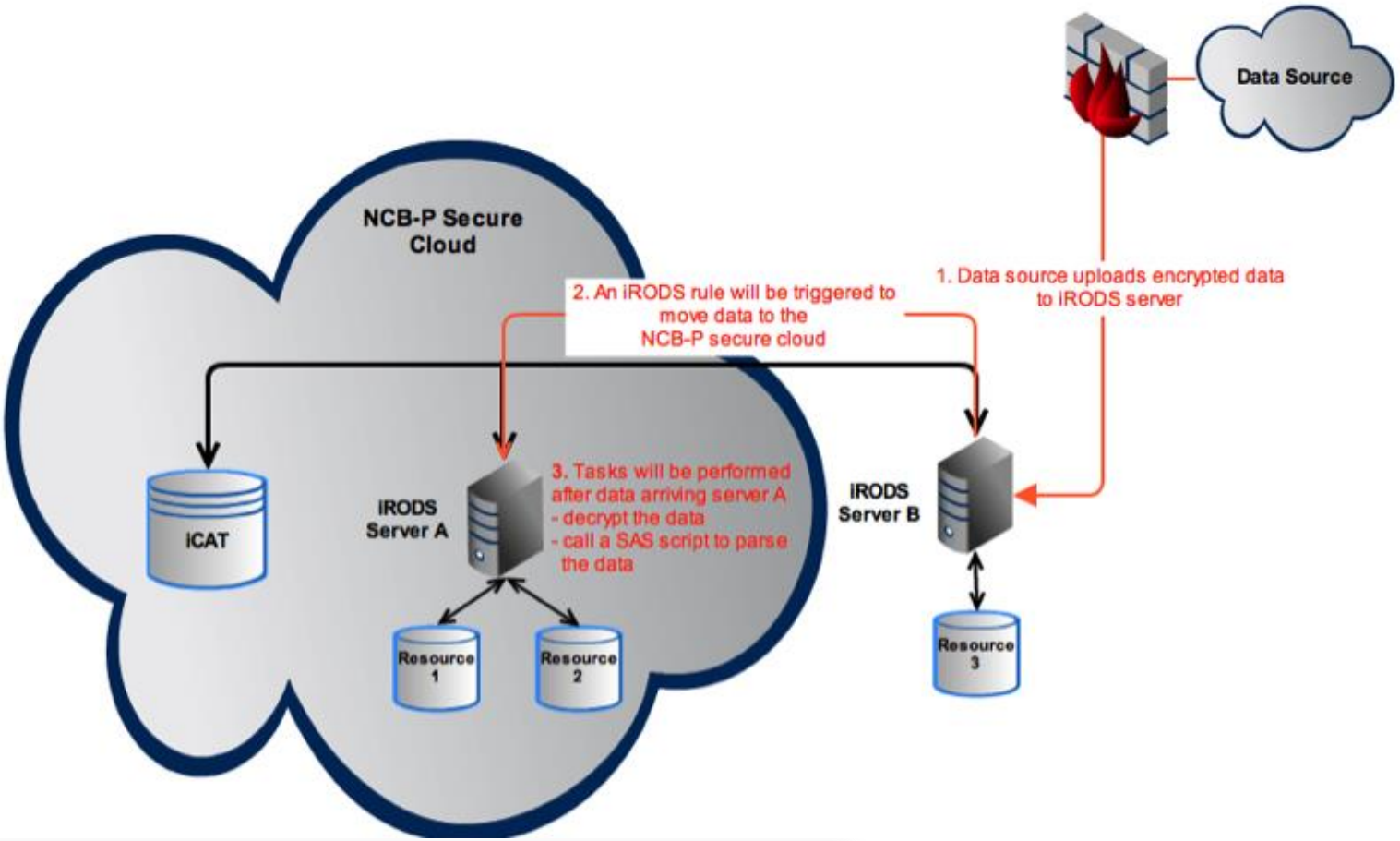


*Access data with Web-based Browser or iRODS GUI or Command Line clients.

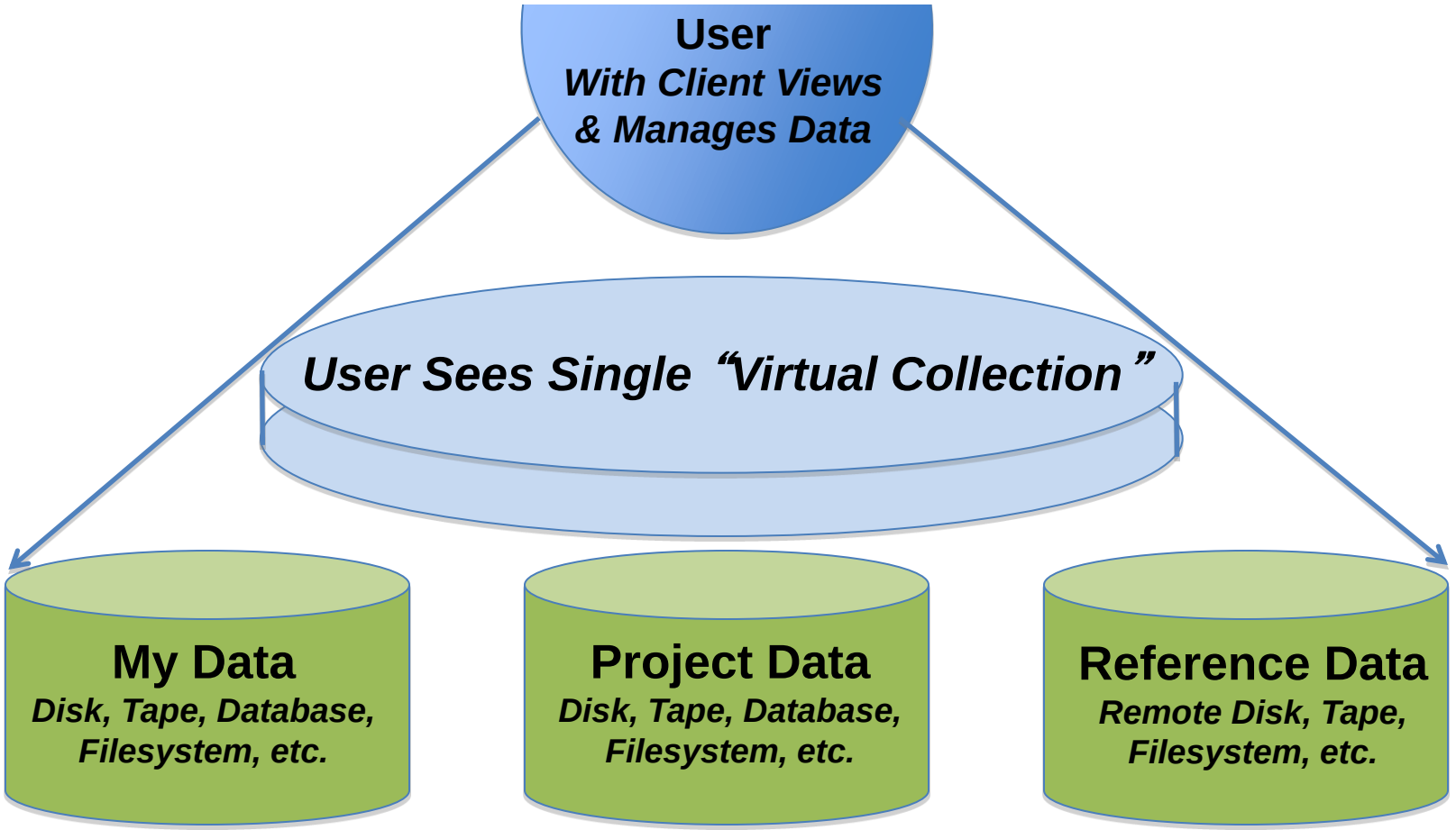
iRODS Rule Example

```
acPostProcForPut{
  ON($rescName == “resource-uiuc”){
    delay("<PLUSET>1m</PLUSET>"){
      msiSysReplDataObj(“resource-unc”,”null”);
    }
  }
```

iRODS Use Case Example – NCB-P

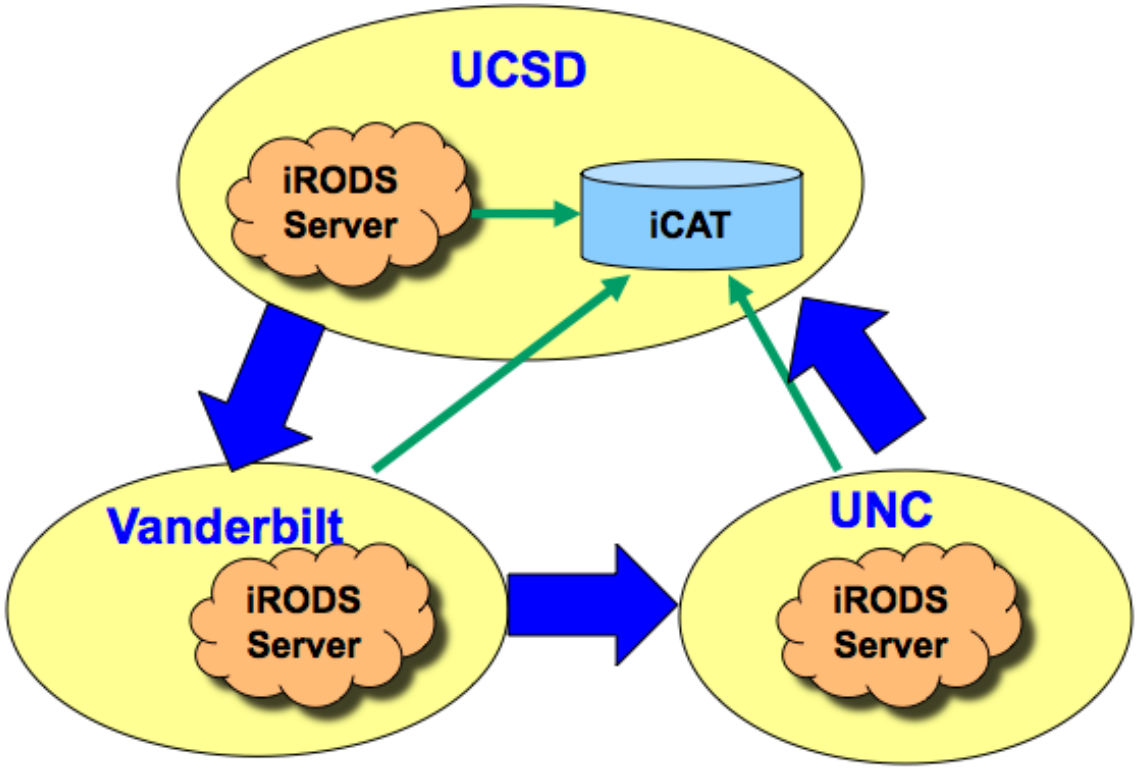


iRODS Shows Unified "Virtual Collection"



The iRODS Data System can install in a "layer" over existing or new data, letting you view, manage, and share part or all of diverse data in a unified Collection.

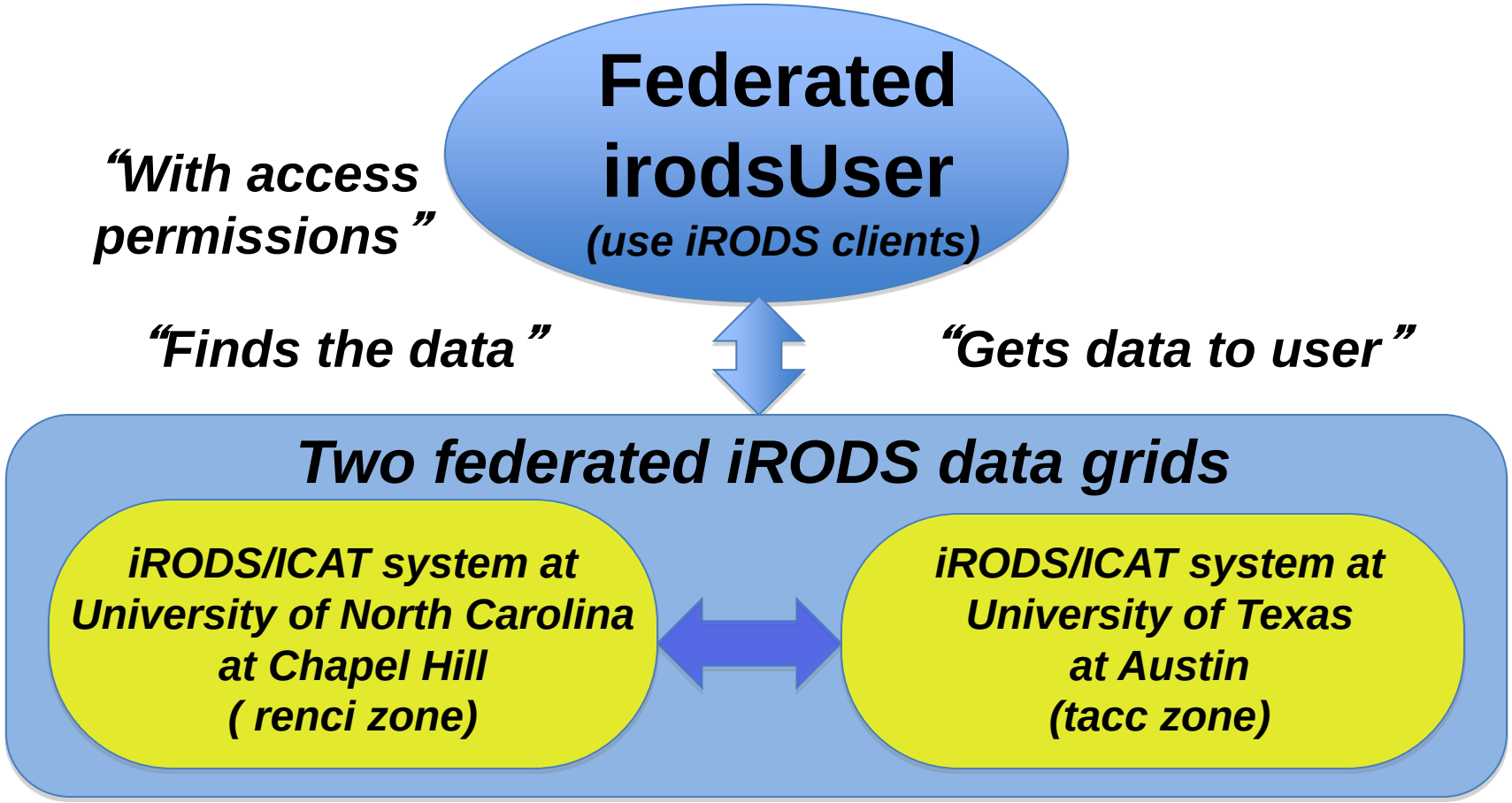
iRODS Use Case Example - TDLC



Data Grid Federation

- Motivation
  - Improve performance, scalability, and independence
- To initiate the federation, each Data Grid administrator establishes trust and creates a remote user
  - iadmin mkzone B remote Host:Port
  - iadmin mkuser rods#B rodsuser
- Use cases
  - Chained data grids - NOAO
  - Master-slave data grids - NIH BIRN
  - Central archive - UK e-Science
  - Deep archive - NARA TPAP
  - Replication - NSF Teragrid

Accessing Data in Federated iRODS

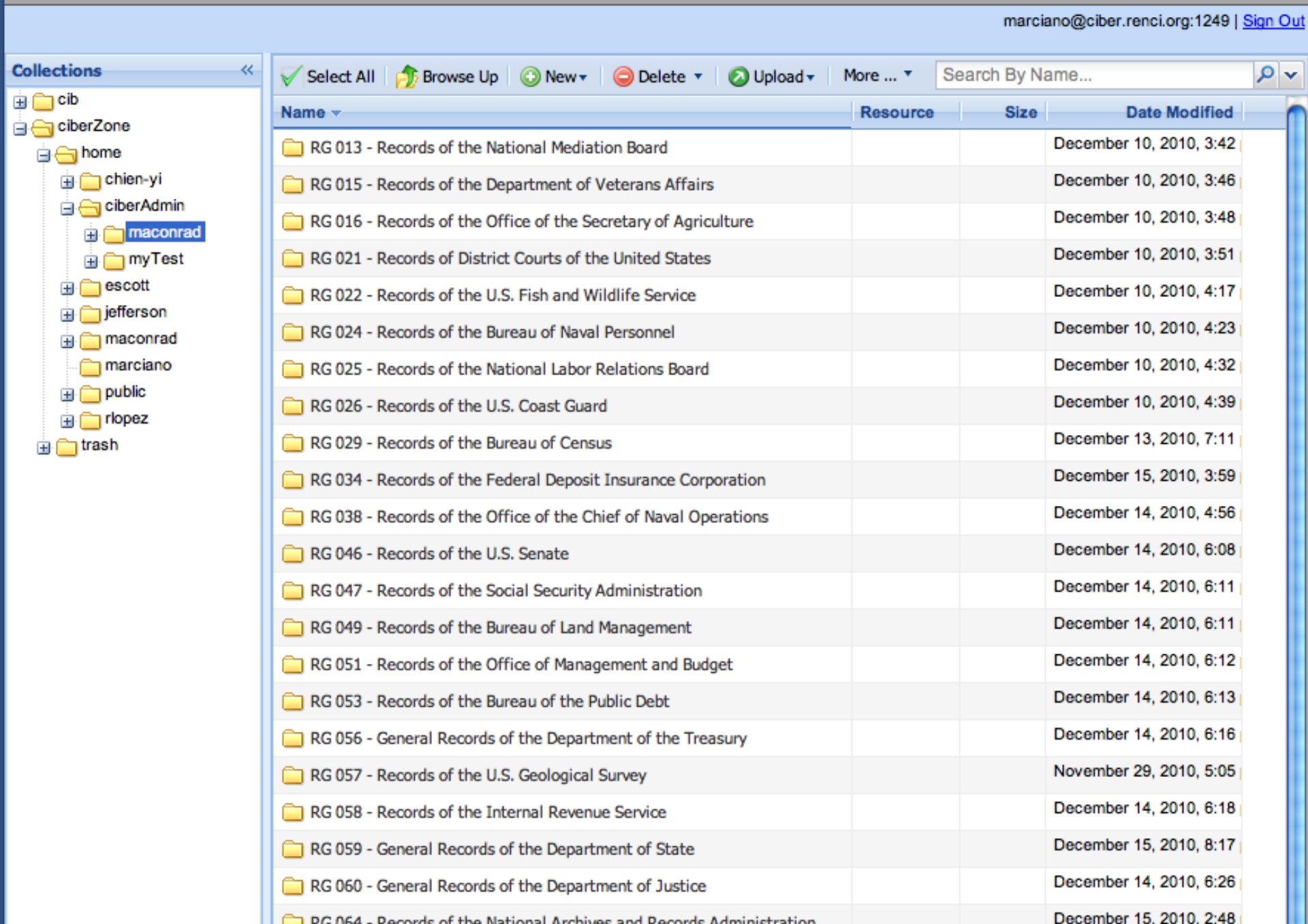


Federated irodsUsers can upload, download, replicate, copy, share data, manage & track access, in either zone.

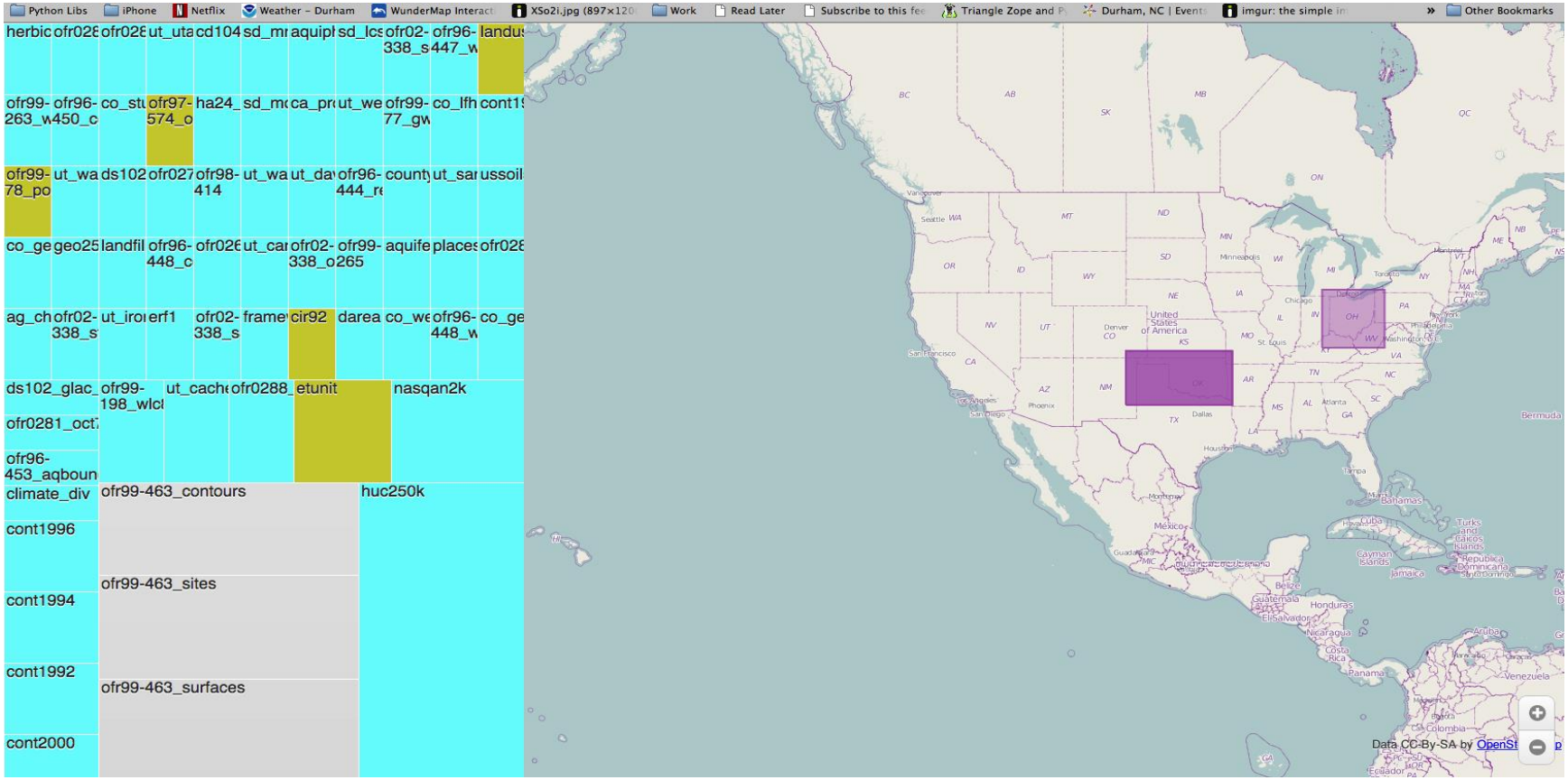
The CI-BER project

- **CI-BER** (CyberInfrastructure for Billions of Electronic Records):
  - Funded by NARA / NSF (2010-2013)
  - See: http://ci-ber.blogspot.com/
  - **Big data management** project based on the integration of heterogeneous datasets:
    1. **Testbed collection** of 100M files and 50TB of data with content from over 100 federal agencies.
    2. Comprises nearly 6,000 file types ranging from a few files to tens of millions of files each, and including diverse file types (text, desktop publishing, databases, audio, video, GIS, XML, etc.), and historical, cultural, social science, and scientific content.
    3. The CI-BER testbed has primarily been used to support the development of scalable record visualization of e-records (geo-analytics) and test the development of national federated infrastructure.

1. Testbed collection



2. Geo-analytics prototype



- **CI-BER directly contributes to one of the objectives of the White House Big Data Initiative:**
  - **Advanced technologies that support Big Data and data analytics:**
    - Setting up of a big data test-bed for managing billion-record scale distributed collections.
      - Exploring computational infrastructure (Hadoop, noSQL)
      - Conducting scalable data analysis. The CI-BER team have built open-source interfaces to big data test-beds based on interactive map interfaces, where users can interact with records at scale from an iPad, cell phone, or web browser. Software at:
        - » github: https://github.com/jeffheard/geoanalytics
      - Relevant papers include:
        - » Cyberinfrastructure tools for Big Geographic Data", OGRS (Open Source Geospatial Research & Education Symposium). Oct. 24-26, 2012 Yverdon-les-Bains, Switzerland. (http://youtu.be/5XCpED0qrJ4)
        - » IEEE LDAV 2011 Symposium on Large-scale Data Analysis and Visualization: "A System for Scalable Visualization of Geographic Archival Records", Oct. 2011, Providence, RI. (http://www.slideshare.net/richardjmarciano/a-system-for-scalable-visualization-of-geographic-archival-records)

- **CI-BER directly contributes to two of the objectives of the White House Big Data Initiative:**
  - **Educate and expand the Big Data workforce:**
    - Scholars:
      - IEEE BigData 2013 workshop on Big Humanities Data, workshop on Oct. 8, 2013 in Santa Clara, http://www.ischool.drexel.edu/bigdata/bigdata2013/topics.htm
    - Professionals:
      - Workshops on CI-BER conducted at national professional meetings attended by archivists, records managers, and IT staff: NAGARA 2013 and SAA 2013.
    - Graduate Students and post-docs:
      - NSF Pan-American Advanced Studies Institutes Program (PASI), workshop taught on Jul. 24, 2013, http://artcaonline.org/news/new-collaborations-at-the-center-of-pasi-2013/
      - Virtual School on Computational Science and Engineering (VSCSE), Data Intensive Summer School, workshop taught on Jul. 8, 2013, http://www.vscse.org/summerschool/2013/bigdata.html
    - Undergraduate Students:
      - Duke University Bass Connections initiative for undergraduate scholars. Workshops in the fall of 2013 and spring of 2014, http://today.duke.edu/2013/05/bassconnx
  - **Improve key outcomes in public awareness and education through demonstrations of Big Data applications:**
    - Citizens:
      - CI-BER has launched a citizen-led initiative with the city of Asheville, N.C., including citizen groups, non-profits, city organizations, and universities, exploring the potential of crowdsourcing around the integration of urban renewal data, and census, economic, and planning content. The partnership is with the Southside Community Advisory Board and seeks to remap a historically African-American neighborhood in Asheville, adversely impacted in the 70s and 80s during urban renewal.
      - A public workshop was held on Aug. 3, 2013, http://www.citizen-times.com/article/20130804/LIVING/308040026/

Break for 5 minutes

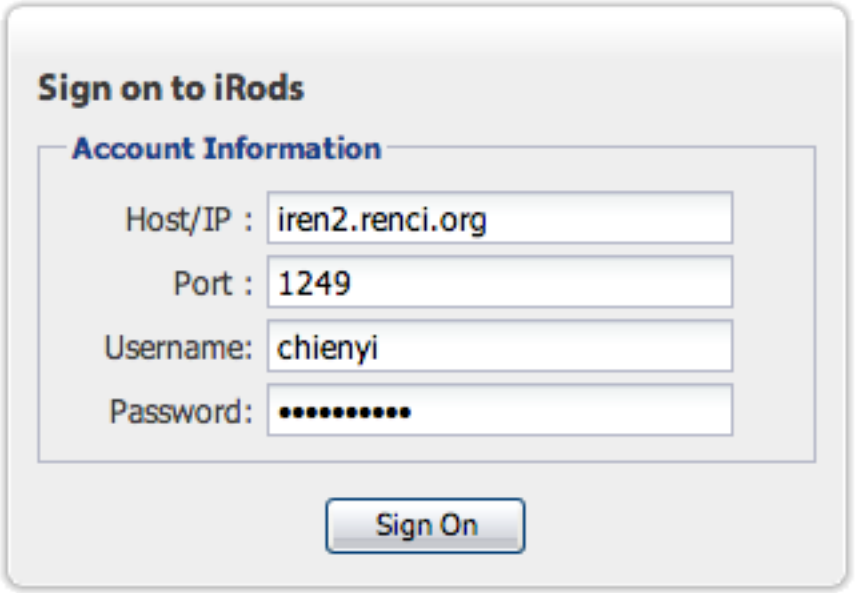
iRODS Clients

Clients

- icommands - unix shell commands
- C I/O library
- Jargon - Java I/O library
- PRODS - PHP iRODS client API
- iDrop – desktop and web GUI
- Python load library
- Fedora / DSpace digital library
- Taverna / Kepler workflows
- Web / Windows browser
- WebDAV - Davis / FUSE
- Parrot / SAGA
- GridFTP
- REST-ful API – based on Jargon Core (planning…)

iRODS Web Browser

- https://irods.org/web/

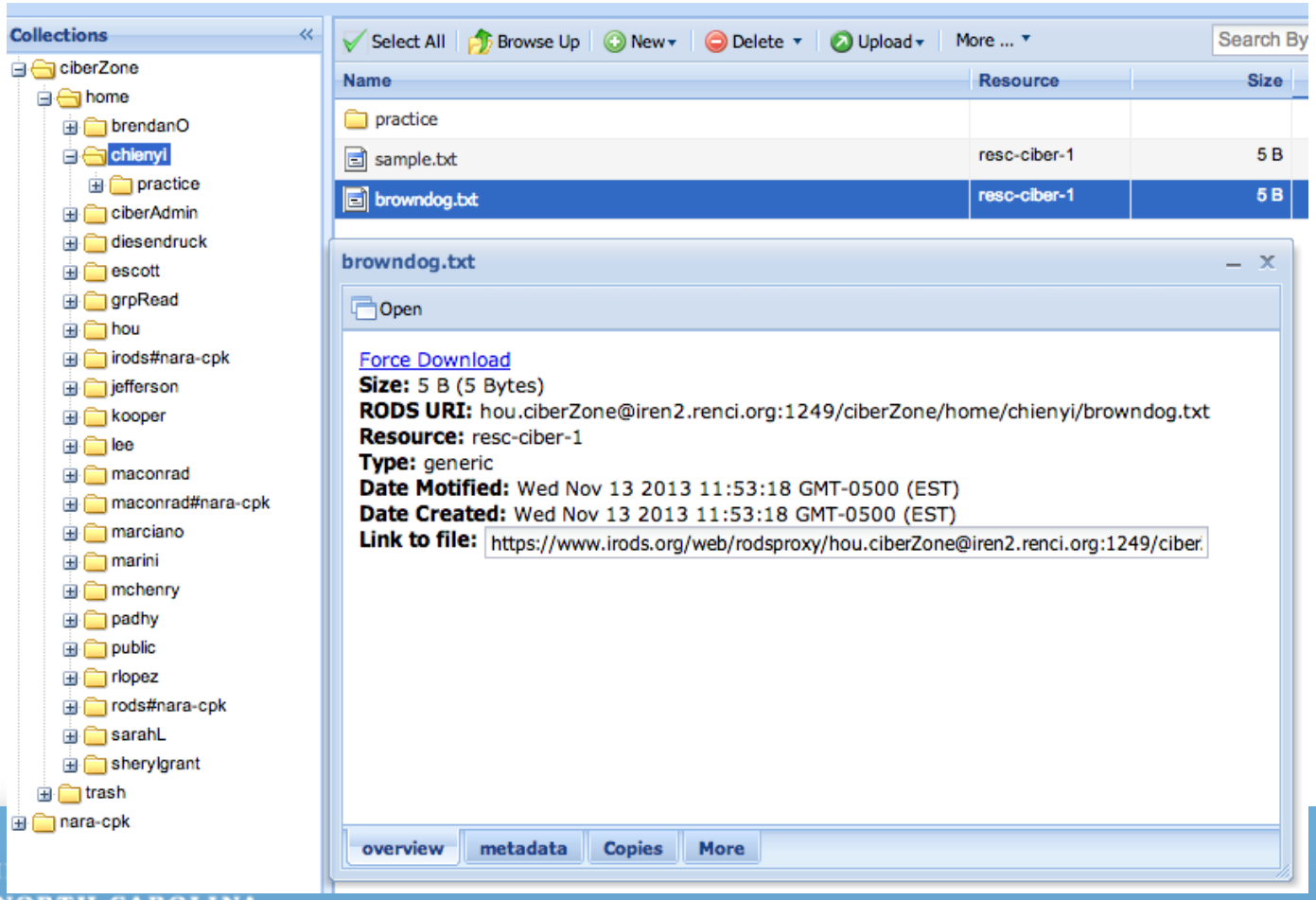


Web browser

- Directory style listing of collections
- List metadata
- List resources
- Load and retrieve files
- Replicate files
- Invoke interactive rule execution

iRODS Web Browser Practice

- Get browndog.txt from ‘chienyi’



Installation of icommands

iRODS installation

- Download appropriate installation manual from iRODS Wiki http://irods.org
- Installation procedure will take
  - Up to 30 minutes for server/catalog/clients
  - Up to 10 minutes for server/clients
  - About 3 minutes for clients
- We will do a client install

iRODS - Unix/Linux/Mac Installation

- https://www.irods.org/download3.2.html
  - Tar file
    - Installation script (Linux, Solaris, Mac OS X)
    - Automated download of PostgreSQL, ODBC
    - Installation of PostgreSQL, ODBC, iRODS
    - Initiation of iRODS collection

iRODS Installation- Unix

- Unpack the release tar file
  - gzip -d irods.tgz
  - tar xf irods.tar
- cd into the top directory and execute
  - ./irodssetup
- It will prompt for a few parameters

iRODS Source Distribution

- INSTALL.txt
- LICENSE.txt
- Makefile
- README.txt
- Configure
- Vault

- irodsctl
- irodssetup

- COPYRIGHT
- CVS
- bin
- clients
- config
- doc
- install
- installLogs
- lib
- modules
- nt
- scripts
- server

irodssetup

- Set up iRODS
- ----------
- iRODS is a flexible data archive management system that supports many different site configurations. This script will ask you a few questions, then automatically build and configure iRODS.

- There are four main components to iRODS:
- 1. An iRODS server that manages stored data.
- 2. An iCAT catalog that manages metadata about the data.
- 3. A database used by the catalog.
- 4. A set of 'i-commands' for command-line access to your data.

- You can build some, or all of these, in a few standard configurations. For new users, we recommend that you build everything.

iRODS Client Installation

- iRODS configuration setup
- ---
- This script prompts you for key iRODS configuration options.
- Default values (if any) are shown in square brackets [ ] at each
- prompt. Press return to use the default, or enter a new value.

- For flexibility, iRODS has a lot of configuration options. Often
- the standard settings are sufficient, but if you need more control
- enter yes and additional questions will be asked.

- Include additional prompts for advanced settings [no]?

iRODS Client Installation

- iRODS configuration (advanced)
- ----------
- iRODS consists of clients (e.g. i-commands) with at least one iRODS
- server. One server must include the iRODS metadata catalog (iCAT).

- For the initial installation, you would normally build the server with
- the iCAT (an iCAT-Enabled Server, IES), along with the i-commands.

- After that, you might want to build another Server to support another
- storage resource on another computer (where you are running this now).
- You would then build the iRODS server non-ICAT, and configure it with
- the IES host name (the servers connect to the IES for ICAT operations).

- If you already have iRODS installed (an IES), you may skip building
- the iRODS server and iCAT, and just build the command-line tools.

- Build an iRODS server [yes]? no

iRODS Client Installation

- iRODS can make use of the Grid Security Infrastructure (GSI)
- authentication system in addition to the iRODS secure
- password system (challenge/response, no plain-text).
- In most cases, the iRODS password system is sufficient but
- if you are using GSI for other applications, you might want
- to include GSI in iRODS. Both the clients and servers need
- to be built with GSI and then users can select it by setting
- irodsAuthScheme=GSI in their .irodsEnv files (or still use
- the iRODS password system if they want).

- Include GSI [no]? no

iRODS Client Installation

- Confirmation
- ----------
- Please confirm your choices.
- ----------
- GSI not selected

- Build iRODS command-line tools
- ----------
- Save configuration (irods.config) [yes]?
- Saved.

- Start iRODS build [yes]?

iRODS Client Installation

- Build and configure
- ------------------
- Preparing...
- Configuring iRODS...
- Step 1 of 4: Enabling modules...
- properties
- Step 2 of 4: Verifying configuration...
- No database configured.
- Step 3 of 4: Checking host system...
- Host OS is Mac OS X.
- Perl: /usr/bin/perl
- C compiler: /usr/bin/gcc (gcc)
- Flags: none
- Loader: /usr/bin/gcc
- Flags: none
- Archiver: /usr/bin/ar
- Ranlib: /usr/bin/ranlib
- 64-bit addressing not supported and automatically disabled.

iRODS Client Installation

- Step 4 of 4: Updating configuration files...
- Updating config.mk...
- Created /iRODS/config/config.mk
- Updating platform.mk...
- Created /iRODS/config/platform.mk
- Updating irods.config...
- Updating irodsctl...

- Compiling iRODS...

- Step 1 of 2: Compiling library and i-commands...

- Step 2 of 2: Compiling tests...

- Done!

iRODS Client Installation

- -----

- To use the iRODS command-line tools, update your PATH:
- For csh users:
- set path=(/iRODS/clients/icommands/bin $path)
- For sh or bash users:
- PATH=/iRODS/clients/icommands/bin:$PATH

- Please see the iRODS documentation for additional notes on how
- to manage the servers and adjust the configuration.

- Change the path name to your installation path

Environment Variables

- In home directory
  - cd ~/.irods
  - vi .irodsEnv
- Default values to describe settings for interacting with your data grid

Environment File

```
# iRODS personal configuration file.
#
# iRODS server host name:
irodsHost 'iren2.renci.org'
# iRODS server port number:
irodsPort 1249

# Default storage resource name:
irodsDefResource 'resc-ciber-1'
# Home directory in iRODS:
irodsHome '/ciberZone/home/chienyi'
# Current directory in iRODS:
irodsCwd '/ciberZone/home/chienyi'
# Account name:
irodsUserName 'chienyi'
# Zone:
irodsZone 'ciberZone'
```

User Configuration

- To use the iRODS 'i-commands', update your PATH:
- For csh users:
  - set path=(/storage-site/iRODS/clients/icommands/bin $path)
- For sh or bash users:
  - PATH=/storage-site/iRODS/clients/icommands/bin:$PATH

icommands I

- iinit: login
- UNIX like commands:
  - ipasswd: change password
  - ils: list files/directories
  - icd: change directory
  - imkdir: create directory
  - icp: copy files/direcotries
  - imv: change name of file/directory
  - ichmod: modify permission of file/directory

icommands II

- iget: download files/directories
- iput: upload files/directories
- imeta: show/edit metadata
- ireg: register files/directories to iCAT
- irepl: replicate files to other resources
- irule: execute a rule file

Check out ihelp or for more icommands

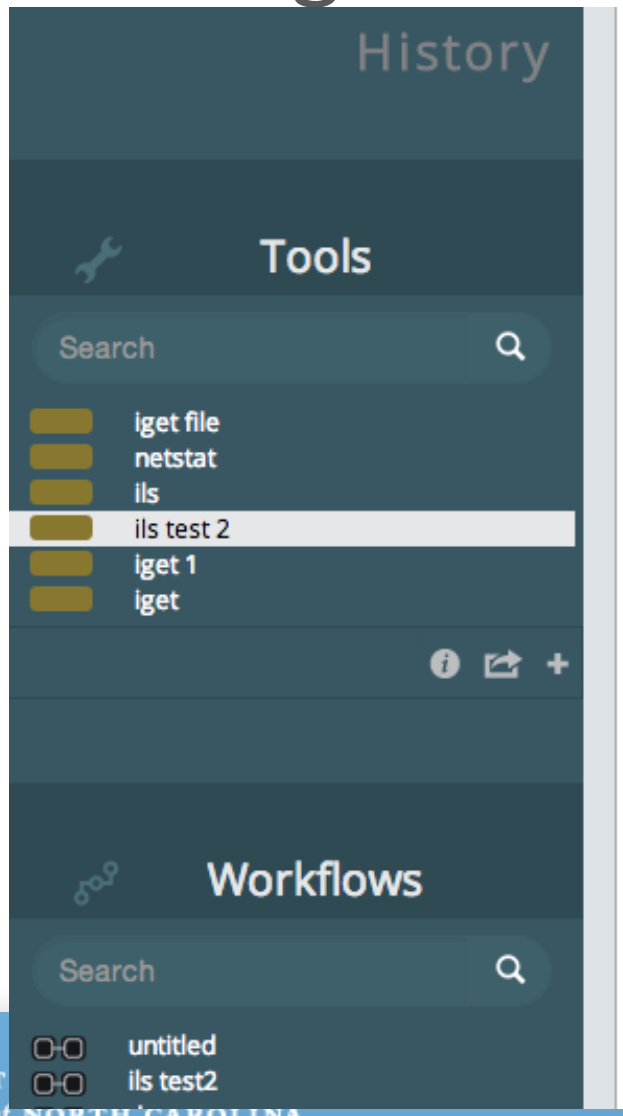
icommands practice

1. Change your password
2. Download sample1.txt and sample2.txt from /ciberZone/home/chienyi
3. Upload a file to your home directory
4. Add “author” metadata to the file
5. Change access permission to allow others to get your file

icommand Practice - Answers

- ipasswd
- iget /ciberZone/home/chienyi/sample1.txt
- iput sample3.txt .
- imeta add -d sample2.txt author hou
- ichmod read grpRead sample2.txt

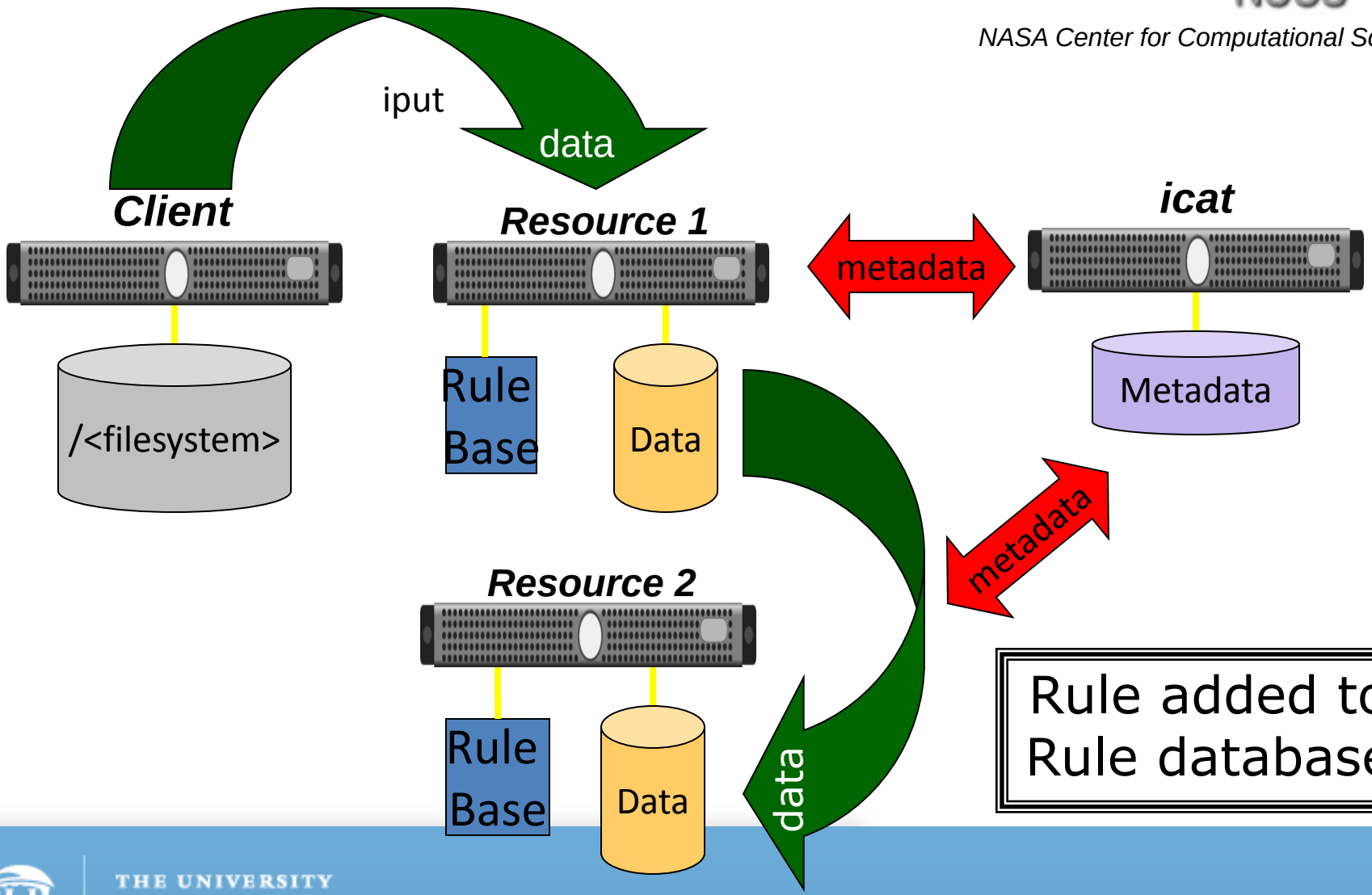
Cyberintegrator



iput With Replication

NCCS

NASA Center for Computational Sciences



iRODS Distributed Data Management