DAITF and the DataNet Federation Consortium

Reagan W. Moore

University of North Carolina at Chapel Hill

rwmoore@renci.org

Topics

- **DAITF and WebDataForum**
  – Working groups
- **DataNet Federation Consortium**
  – Interoperability between systems
  – Federation of data management systems

Research Data Alliance

- **International effort to understand data sharing and interoperability**
  - 16 working groups have been proposed
- **Policy-based interoperability working group**
  - Collections are governed by policies
  - Share policies and associated procedures to understand what can be done with the shared data
  - Makes explicit the relationship between domain knowledge and the procedures that encapsulate the knowledge into workflows
  - Expect to share the basic functions (micro-services) that are chained to create workflows

Workflows as Domain Knowledge

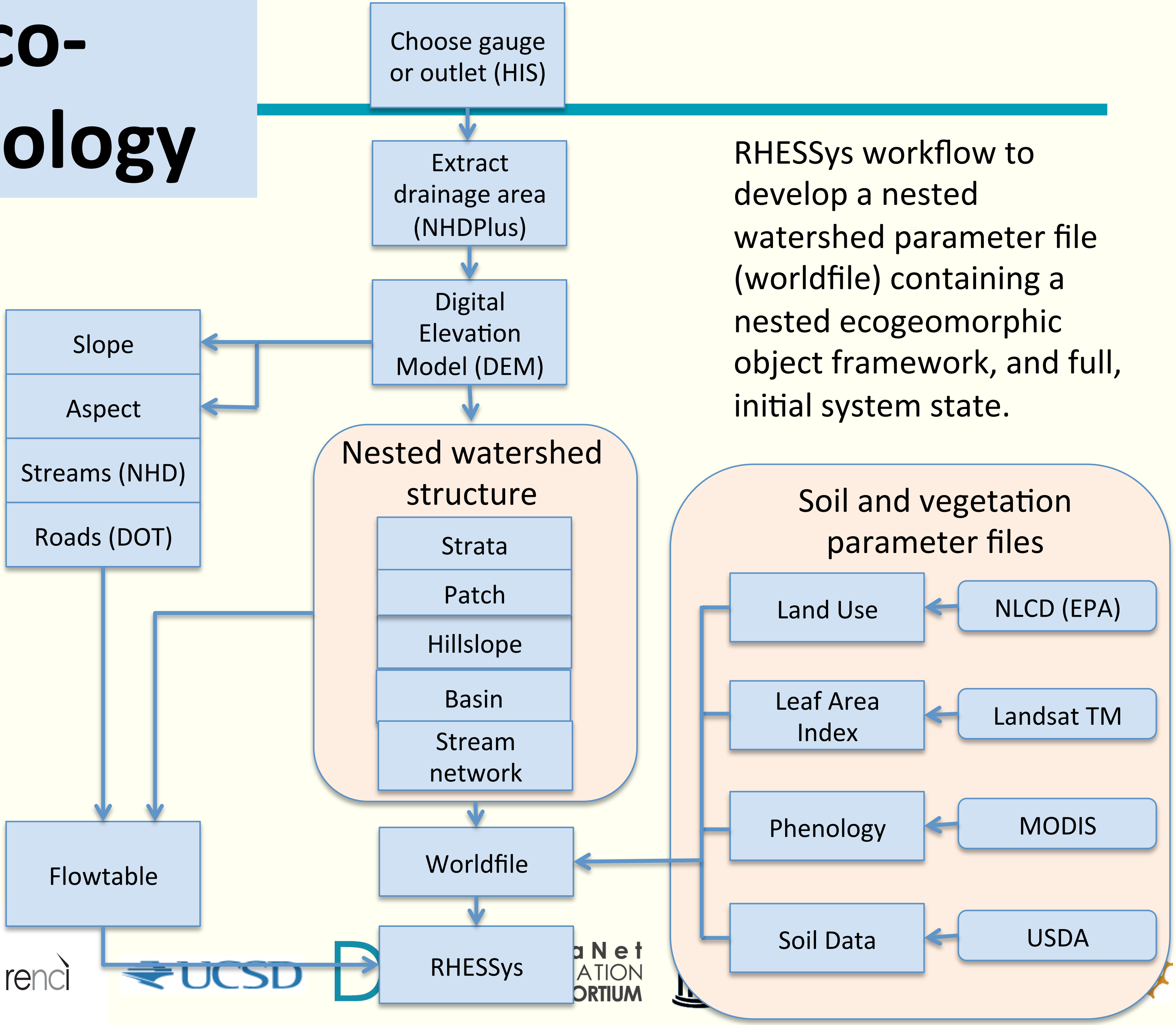
- **An example is the hydrology community**
  - Access multiple federal repositories to acquire digital elevation maps, precipitation, soil, land cover data
  - Process each data set to transform to the required physical variables and coordinate system
  - Each process step is encapsulated into a separate basic function (micro-service)
  - Chain the micro-services together to implement an analysis
- **VIC watershed analysis**
  - 3085 files
  - Requires about 3 hours to run as a workflow, versus about 3 months to do the data aggregation and transformation by hand

Use Cases

- **Demonstrate reproducible science.** A use case could include the registration, storage, sharing, and re-execution of a workflow. The hypoxia use case from the Cross-Domain and Brokering Concept groups could be used as an example.
- **Automate data retrieval.** A use case could demonstrate remote access to a data collection, retrieval of desired data sets, transformation, and use in an analysis workflow. An eco-hydrology example that automates access to digital elevation maps and land use coverage is being built.
- **Integrate community resources with collaboration environments**. An example would be use of the DAB protocol to identify and cache local copies of relevant data sets for local analysis.
- **Integrate multiple community resources.** A use case could be demonstration of invocation of multiple workflow systems within the same analysis. An example is the integration of Cyberintegrator workflow with collaboration environments to support drought prediction.

Eco-Hydrology

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



Choose gauge or outlet (HIS)

Extract drainage area (NHDPlus)

Digital Elevation Model (DEM)

Slope

Aspect

Streams (NHD)

Roads (DOT)

Nested watershed structure

Strata

Patch

Hillslope

Basin

Stream network

Soil and vegetation parameter files

Land Use

NLCD (EPA)

Leaf Area Index

Landsat TM

Phenology

MODIS

Soil Data

USDA

Flowtable

Worldfile

RHESSys

iRODS Rule for RHESSys

Modular workflow composed by chaining basic transformation

- Define input variables
- Call functions to apply each transformation step
- Store results in shared collection

```
main {
 getExtentForGageReachcode(*gageReachcode, *extentInNHD_Vect_Coords);

convertExtentToNHD_DEM(*extentInNHD_Vect_Coords,
*extentInNHD_DEM_Coords);

 extractTileFromNHD_DEM(trimr(*extentInNHD_DEM_Coords, "\n"));

 importDEMTileIntoNewGRASSLocationAsUTM(*extentInNHD_Vect_Coords,
*newLocPhysPath, *newLocObjPath);

 delineateWatershedForNHDGage(*nhdStreamGageID, *newLocPhysPath,
*newLocObjPath);
}
```

D·I·C·E renci UCSD DFC DataNet FEDERATION CONSORTIUM

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

THE NATIONAL ARCHIVES ARCHIVES.GOV

NSF

Science Requirements Working Group

- **Focus on the knowledge needed to understand a scientific data set**
- **Common properties**
  - Data format (e.g. HDF5, NetCDF, FITS, …)
  - Coordinate system (spatial and temporal locations)
  - Geometry (rectilinear, spherical, flux-based, …)
  - Physical variables (density, temperature, pressure)
  - Physical units (cgs, mks, …)
  - Accuracy (number of significant digits)
  - Provenance (data generation steps, calibration steps)
- **Domain specific and project specific properties**
  - Physical approximations (incompressible flow, adiabatic, equation of state, …)
  - Semantics (domain knowledge for term relationships)
  - Domain Semantics (domain knowledge& term relationships)
  - Extended Semantics (project-specific properties)

DataNet Federation Consortium

Data Driven Science

- **Implement national data grid**
  - Federate existing discipline-specific data management systems to enable national research collaborations
- **Enable collaborative research on shared data collections**
  - Manage collection life cycle as the user community broadens
- **Integrate "live" research data into education initiatives**
  - Enable student research participation through control policies

Project

Shared Collection

Processing Pipeline

Digital Library

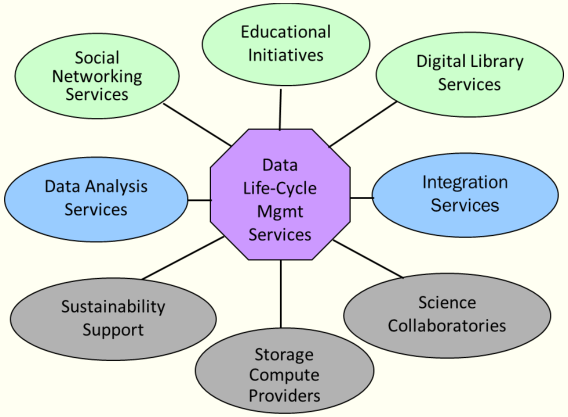
Reference Collection

Federation

Collection Life Cycle

Cyber-infrastructure Partners:

Univ. of North Carolina, Chapel Hill
Univ. of California, San Diego
Arizona State University
Drexel University
Duke University
University of Arizona
University of South Carolina



Science and Engineering Initiatives:

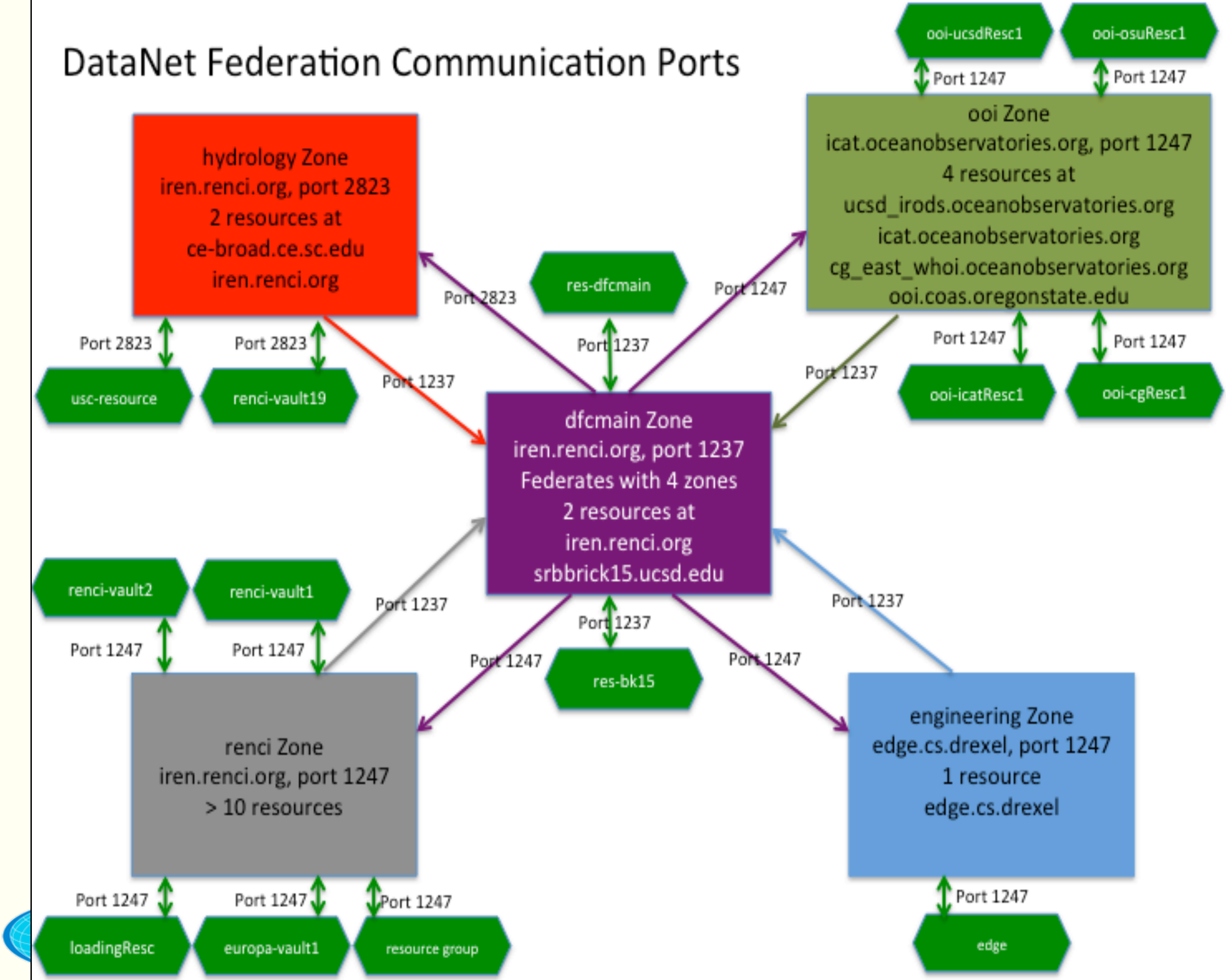
Ocean Observatories Initiative
the iPlant Collaborative
CUAHSI
CIBER-U
Odum Social Science Institute
Temporal Dynamics of Learning Center

National Science Foundation Cooperative Agreement: OCI-0940841

i·R·O·D·S Policy-based data management

D·I·C·E renci UCSD DFC DataNet FEDERATION CONSORTIUM THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL THE NATIONAL ARCHIVES ARCHIVES.GOV NSF

DataNet Federation Communication Ports



hydrology Zone
iren.renci.org, port 2823
2 resources at
ce-broad.ce.sc.edu
iren.renci.org

ooi Zone
icat.oceanobservatories.org, port 1247
4 resources at
ucsd_irods.oceanobservatories.org
icat.oceanobservatories.org
cg_east_whoi.oceanobservatories.org
ooi.coas.oregonstate.edu

dfcmain Zone
iren.renci.org, port 1237
Federates with 4 zones
2 resources at
iren.renci.org
srbbrick15.ucsd.edu

renci Zone
iren.renci.org, port 1247
> 10 resources

engineering Zone
edge.cs.drexel, port 1247
1 resource
edge.cs.drexel

DataNet Federation Consortium

- **Uses extensibility mechanisms provided in iRODS**
  - Storage resource drivers

    Issue protocol of remote repository (Posix I/O)
  - Micro-Service Structured Object

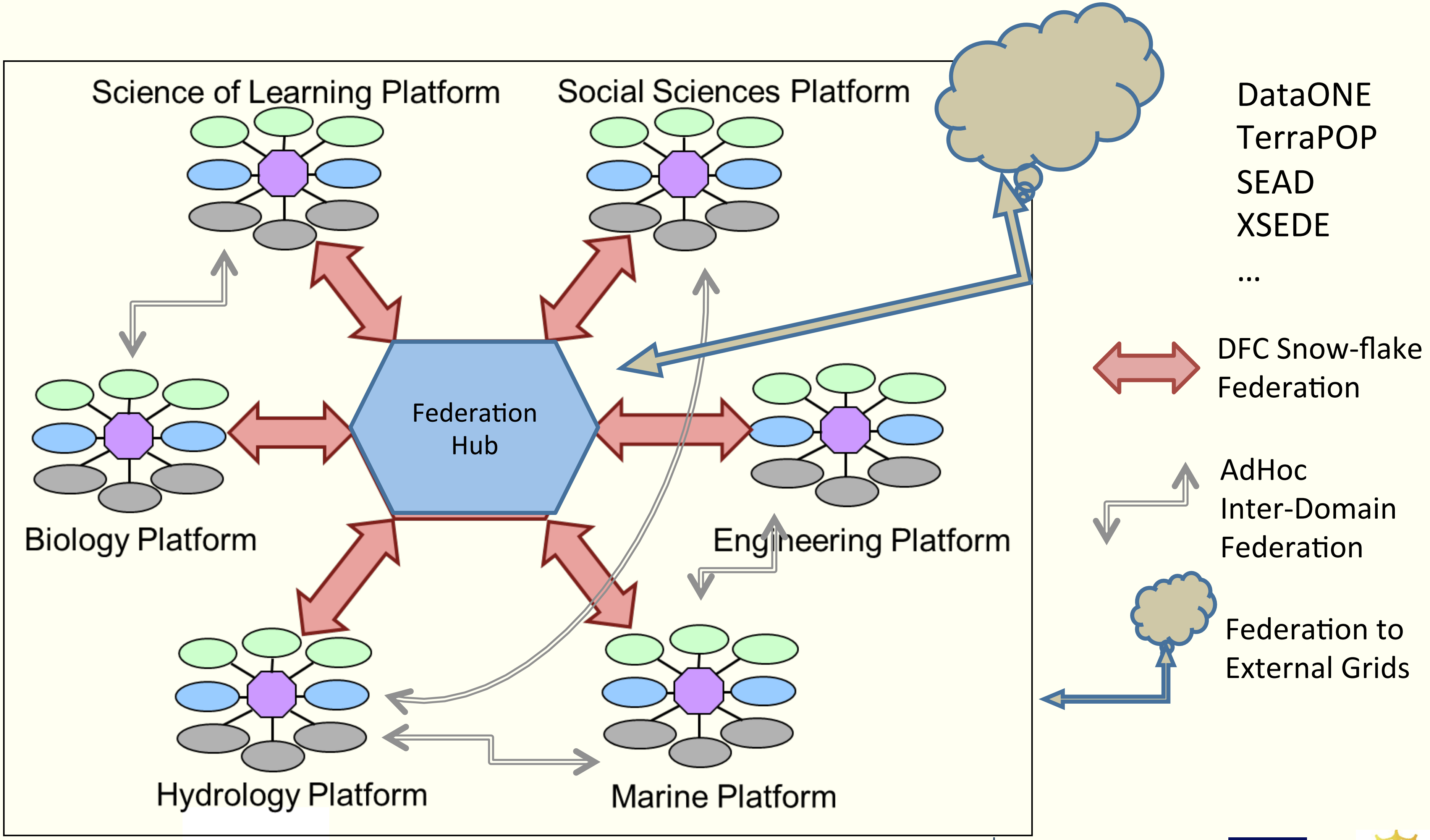
    Issue protocol of remote repository for get and put

    Used to create soft links
  - Federation policies

    Control interactions between data management systems

DFC Federation

(Federation of DFC Platforms)



DataONE
TerraPOP
SEAD
XSEDE
…

DFC Snow-flake Federation

AdHoc Inter-Domain Federation

Federation to External Grids

Three Areas of Tech Dev

- **Authentication**
  - iRODS supported Secure Password, Kerberos and GSI
  - iRODS extended to support PAM/LDAP
  - Pluggable Authentication Module
  - PAM is part of the X/Open Single Sign-on (XSSO) standard
- **Workflow Virtualization**
  - SRB virtualized users, resources, data and collections
  - iRODS virtualized policies and micro-services
  - iRODS extended to virtualize workflows
- **Data Book GUI**
  - A Face Book for Data
  - Under design and development

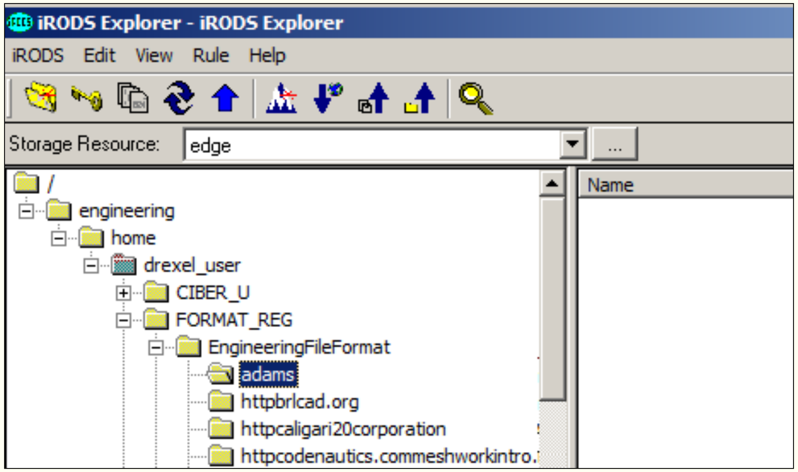
Engineering Demo

Isaac Simmons, William Regli

Drexel University

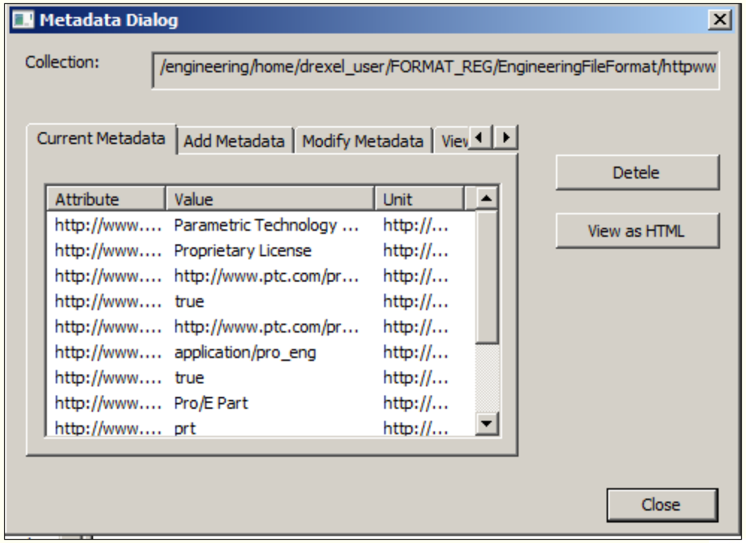
Format Registry

- **Centralized store for file format knowledge**
- **Configured by OWL schema**
- **Stored in iRODS**
  - AVU metadata
  - Sample files
- **Engineering file formats (63)**



Storage Resource: edge

/
engineering
home
drexel_user
CIBER_U
FORMAT_REG
EngineeringFileFormat
adams
httpbrlcad.org
httpcaligari20corporation
httpcodenautics.commeshworkintro.



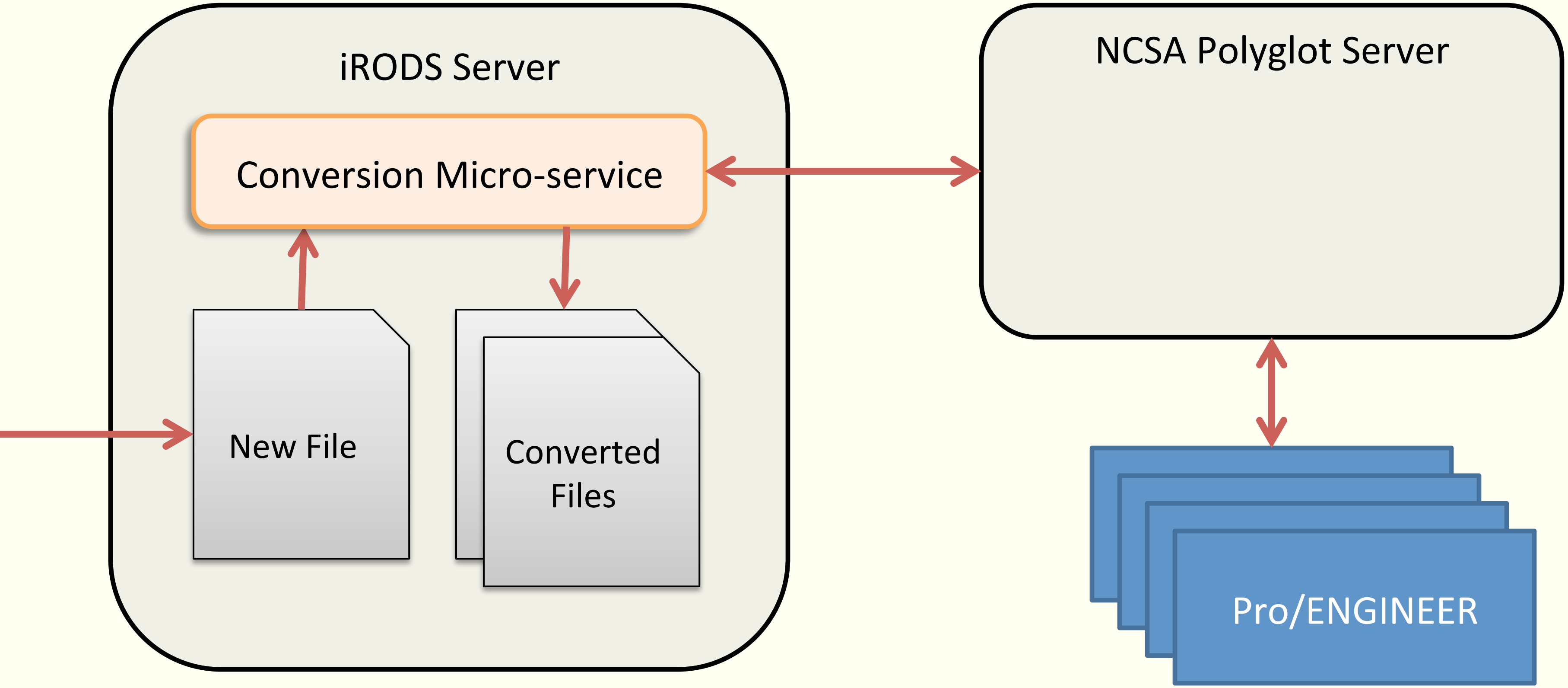
Collection: /engineering/home/drexel_user/FORMAT_REG/EngineeringFileFormat/httpww

Attribute	Value	Unit
http://www....	Parametric Technology ...	http://...
http://www....	Proprietary License	http://...
http://www....	http://www.ptc.com/pr...	http://...
http://www....	true	http://...
http://www....	http://www.ptc.com/pr...	http://...
http://www....	application/pro_eng	http://...
http://www....	true	http://...
http://www....	Pro/E Part	http://...
http://www....	prt	http://...

File Conversion

- **NCSA Polyglot server**
  - Code reuse server for file conversions
- **iRODS micro-service**
  - Upload from iRODS to Polyglot HTTP server
  - Download converted files from Polyglot to iRODS
- **Convert proprietary 3D CAD formats into open formats appropriate for archival and consumption**

File Conversion



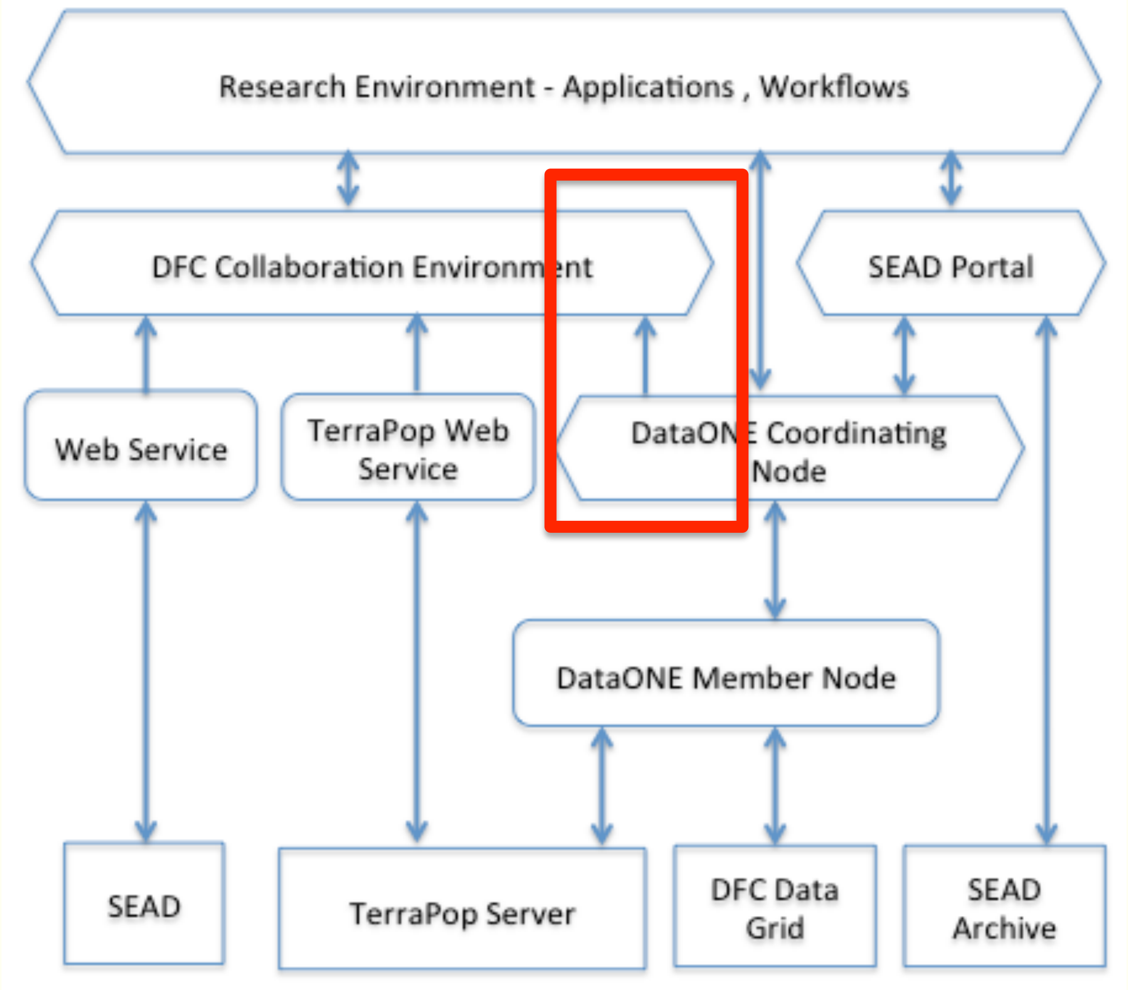
MediaWiki Integration

- CIBER-U is a MediaWiki installation used for engineering design education
  - 41 courses, 9 universities, 32 faculty
- MediaWiki offers limited file curation capabilities – Augment with iRODS file storage
- Bring iRODS capabilities to the processes users are already using

DFC + DataONE Interoperability

- Goal: support interoperability between a DFC data grid and DataONE
- Task: Retrieve a file from DataONE, load into a DFC collaboration environment and add metadata

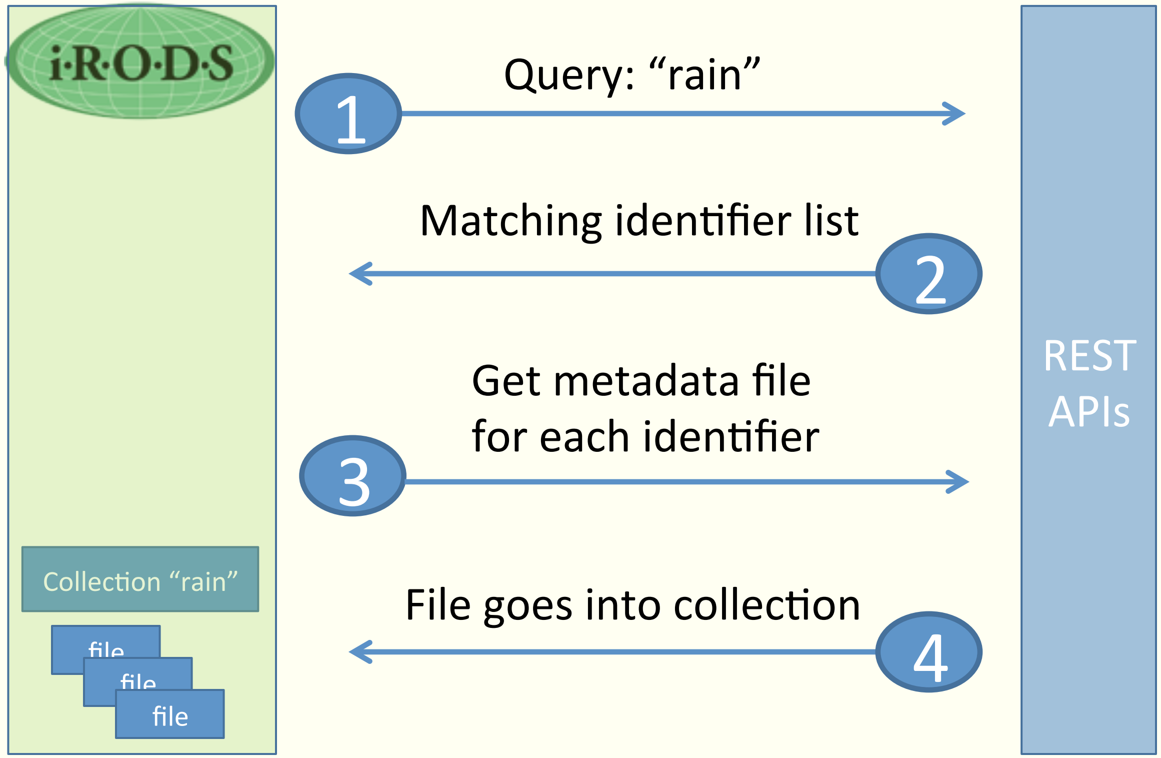
Interoperability Approach

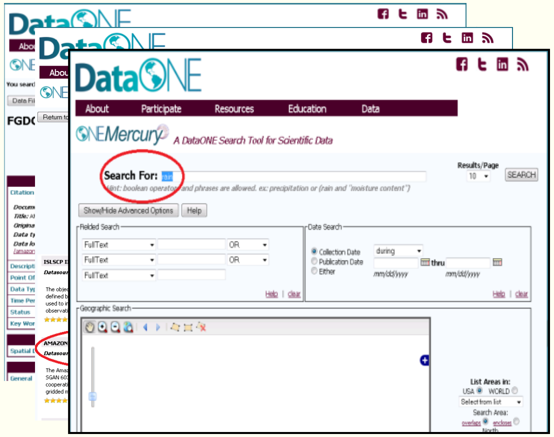


How It Works

1. Query DataONE Coordinating Nodes with SOLR query
2. Create iRODS collection with same name as query
3. Get list of identifiers for metadata files from search
4. Download the metadata file for each identifier
5. Store the metadata file in DFC data grid

What the Demo Shows

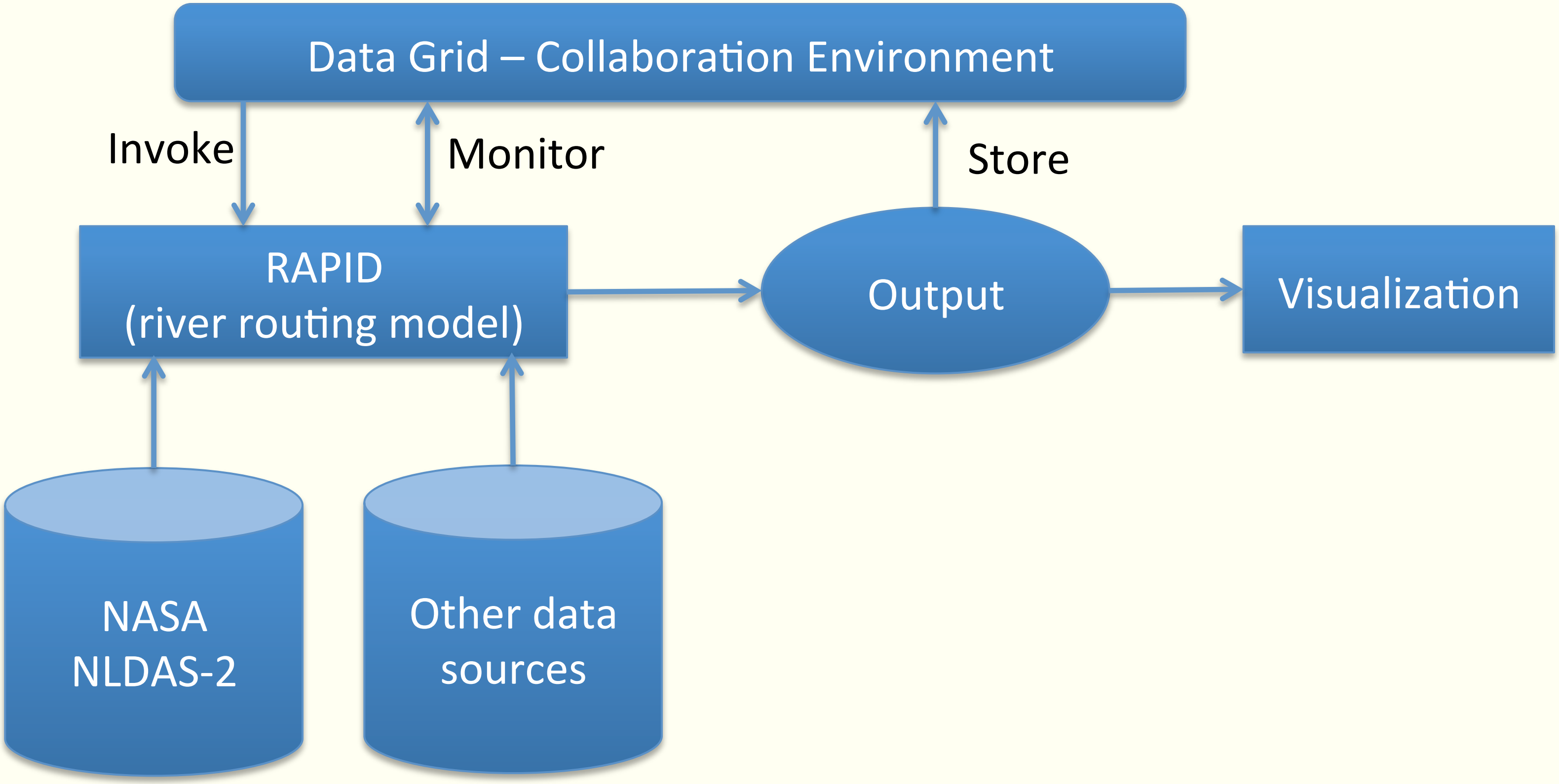




Mercury Web portal

EarthCube Demonstrations

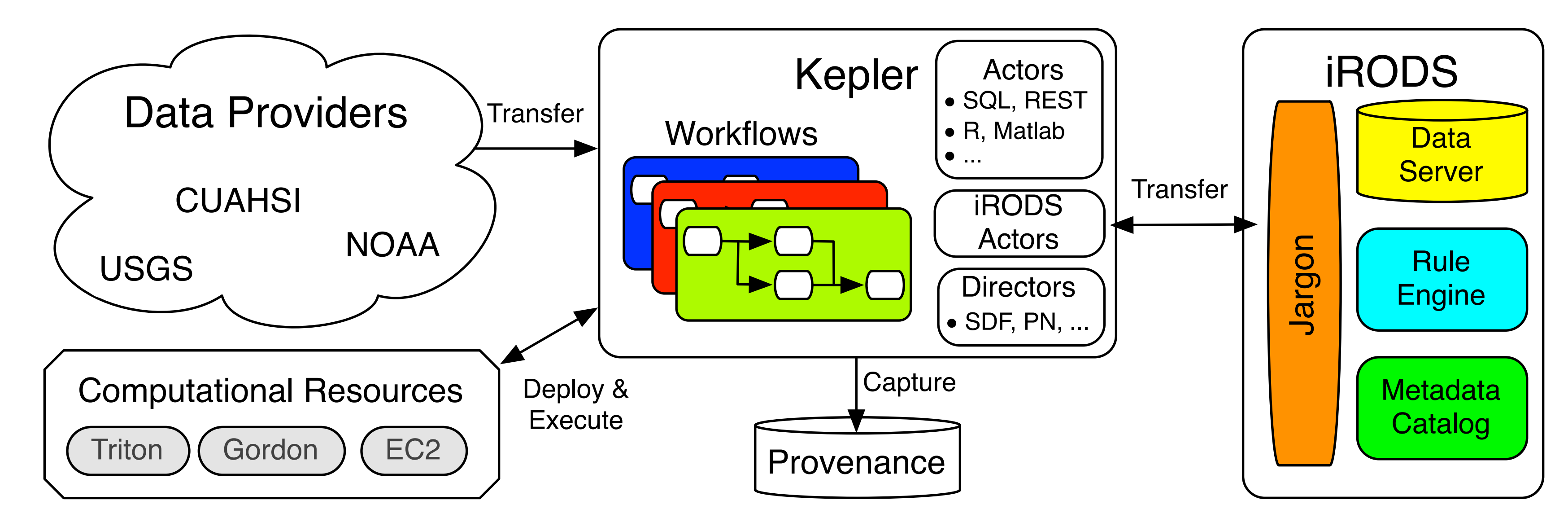
Event-Driven Real-Time Drought Analysis/ Prediction Workflow



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

Hypoxia in the Gulf of Mexico

- Dissolved oxygen plots stored in iRODS repository
- Plots can be annotated on web portal
  - annotations stored as iRODS metadata



D·I·C·E renci UCSD DFC DataNet FEDERATION CONSORTIUM THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL THE NATIONAL ARCHIVES ARCHIVES.GOV NSF

DataNet Federation Consortium

http://www.datafed.org