DataNet Federation Consortium

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

RENCI

Years 1-2:
✓ 6 FTEs
✓ 8 Staff
✓ 5 Faculty
✓ 3 Students

Years 1-2:
✓ 4 Universities
✓ 3 Consortia
✓ 2 Companies

Please credit the DataNet Federation Consortium when referencing this information.
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 cyberinfrastructure
  - 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
DFC Vision

• Collaboration environment enables research:
  – Registration of remote data sets into local workspace
  – Registration of analysis workflows
  – Tracking of provenance of workflows
    ▪ Input data sets and input parameters
    ▪ Correlation of workflows with research data products

• Reproducible science
  – Sharing of workflows and data
  – Re-execution of workflows by collaborators
The Challenge: 
*To Enable Data-Driven Science*

Deliver the capability to manage, mine, and analyze data in near real time and build sustainable collections.

The Future: an Explosion of Data
Use Case: National Water Model

Hydrologic scientist have expressed a “grand research challenge” of building a National Water Model for flood and drought applications.

Terrain in the Neuse River Basin, NC constructed from 390 million LiDAR measurements

Flooding in the Mississippi River Basin, August 1993 observed from satellite

Achieving this goal will require a system like DFC to handle the massive data requirements.

Source: terrain.cs.duke.edu

Source: nasa.gov

4/5/2012
Data and Model Integration Needed to Support Hydrologic Science

- **DFC**
  - Physical Data
  - Weather and Climate Models

- **Observations**
  - Monitoring Station 92374

- **Hydrologic Models**

- **Socioeconomic Data**

- **CUAHSI HIS**
Ocean Sensing Systems Paradigm Shift

Platform-centric Sensing Systems

Net-centric, Distributed Autonomous Sensing Systems

Federation of real-time sensor data
Use Case: Ocean Observatories

- **Use data grid technology within OOI**
  - Manage real-time sensor data stored as files
- **Collaborate on installation of data grids at NOAA**
  - Ingest and access to NOAA data records
- **Federate with NOAA data archives**
  - Automate deposition into NOAA archive
  - Generate climate data records
State of Engineering Design Education

- Is America Falling Off the Flat Earth? (2007)
DFC Supports NAE “Grand Challenges” Efforts

- CUASHI
- Drexel “Engineering Cities” Initiative
- TDLC
- Drexel’s ACIN Initiative
- iRODS and DFC tools are the tools for data-driven scientific discovery!
### Community-based Collection Life Cycle

Each collection life cycle stage re-purposes the original collection

<table>
<thead>
<tr>
<th>Project Collection</th>
<th>Data Grid</th>
<th>Data Processing Pipeline</th>
<th>Digital Library</th>
<th>Reference Collection</th>
<th>Federation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private</td>
<td>Shared</td>
<td>Analyzed</td>
<td>Published</td>
<td>Preserved</td>
<td>Sustained</td>
</tr>
</tbody>
</table>

- The stages correspond to addition of new policies for a broader community.
- We virtualize the stages of the collection life cycle through policy evolution.

4/5/2012
Goals and Challenges

Opportunity and Capabilities

Scientific Goals

• Data-driven science
• National data cyberinfrastructure deployment
• Enable collaborative research
• Improve research productivity

Challenges

(capturing metadata)
(massive collections)
(controlled sharing)
(workflow registration)

New Opportunities and Capabilities

• Time-dependent data support
• Federation of repositories
• Reference collections for educational use
• Virtualization of stages of data life cycle

(new data modalities)
(integration across systems)
(active collections)
(sustainability)
Addressing the Challenges

Highly extensible, scalable, modular policy-based framework

- Policy-based data management
- Interoperability across clients and storage resources
- Automation of administrative tasks
- Validation of assessment criteria
- Workflow management
- Federation of repositories

Sustainability of infrastructure, collections, policies

- Repurposing of collections
- Federation of repurposed collections
- International development collaborations (EUDAT)
- Enterprise-iRODS version (RENCI)
Communities of Practice

Policies for automating administrative tasks

1. Science & Engineering Domains
2. Education And Outreach
3. Technology And Research
4. External Advisory Board
5. Sustainability And Institutions
6. International Projects

Policies and Standards

Standards Groups

Facilities And Operations

Policies for federation and re-purposing

Policies for metadata, formats, APIs

NSF data management policy examples

Policies for international collaboration

Policies for IPR and access

Policies for federation and re-purposing

Policies for domains

4/5/2012
Science Domain Partners

Years 1-2

GeoScience - Hydrology
Jon Goodall – USC

Ocean Observatories Initiative (OOI)
John Orcutt – UCSD

Cyberinfrastructure-Based Engineering Repositories for Undergraduates (CIBER-U)
William Regli – Drexel

Years 3-5

Temporal Dynamics of Learning Center (TDLC)
Andrea Chiba – UCSD

iPlant Collaborative
Sudha Ram – University of Arizona

Odum Institute for Research in Social Science
Tom Carsey, Jon Crabtree – UNC-CH
Extended Collaboration Communities

- RENCI
- EUDAT
- DDN
- iPLANT
- SDCI
- Other Open Source Contributors
- Product Realization Integrated Digital Enterprise (PRIDE), DoE/NNSA NA-122
- DFC Technology COP
- DFC Science Collaborators OOI, CIBER-U, Hydrology
- Open Source iRODS
- Distributed Bio
- IN2P3
- EarthCube Testbed

EarthCube Testbed

4/5/2012
Transition to Production Infrastructure

Executive Committee
Members are:
- The PI
- The Project Manager
- Community of Practice and Science Domain representatives (rotating membership)
  - Manages overall vision
  - Manages external relationships
  - Meets quarterly

Management Team
Members are:
- The PI
- The Project Manager
- Community of Practice and Science Domain representatives
  - Manages internal activities and implementation
  - Meets weekly

Project Manager
- Manages
  - finances
  - personnel
  - risk tracking and mitigation
  - project information systems
- Liaises with university & sponsoring agencies
- Performs other duties as assigned

Strategic (vision)  →  Tactical (operations)
Strategic Framework for Organizing Work Plan

- Real-time Data
- Repository Infrastructure
- Operational Evaluation
- Test & Measurement
- Reference Collections
- Enabling Technologies
- Systems/Applications
- Technology Integration
- Requirements
- Barriers/Risks
- Products & Outcomes
- OOI Hydrology CYBER-U iP C TDLC Odum

Technology Base
- National Scale Cyber-Infrastructure
- Repository Federation
- Reference Collections

Technology Elements
- User Interfaces & Interoperability (Cobalt)
- Autonomic Applications
- Distributed Trust Mechanisms
- Semantically Self-Describing Data & Sensors
- Digital Library / Workflow Systems (Fedora, Taverna)
- Policy-based Data Management (iRODS)

Fundamental Insights
- Data Life Cycle Policy Sets
- Verification of Rule Correctness
- Rule Generation
- Consistency of Rule Bases
- Expressive Policy Language
- Evaluating Collection Worth

Barriers/Risks
- Automation
- Validation Mechanisms
- Monitoring & Control
- Policy Consensus

Research
- Policy-based Data Management, Sustainability

4/5/2012
Evolution of Domain Policies

Key Management Elements

Policies and Standards

Science and Engineering

Identify / Create New Policies

Science and Engineering

Data Cyber-infrastructure Technology and Research

Facilities and Operations

Analyze

Iterate with rapid prototypes

Prioritize requests

Implement and test

Install and administer

4/5/2012
Policy-Based Roadmap

• **Purpose** - reason a collection is assembled
• **Properties** - attributes needed to ensure the purpose
• **Policies** - enforce and maintain properties
• **Procedures** - functions that implement the policies
• **State information** - results of applying procedures
• **Assessment criteria** - validate that state information conforms to the desired purpose
• **Federation** - controlled sharing of logical name spaces to enable repurposing of community collections for long-term sustainability
DFC Deliverables

**Year 1**
- Implement a National Data Management Infrastructure
- Requirements Analysis and Middleware Installation
- Within first 3 months: Develop Program Execution Plan

**Year 2**
- Develop Reference Implementations
- Develop Interoperability Mechanisms
- Research: Education Policies

**Year 3**
- Develop Services for Structured Data
- Develop Social Network Services for Data Use in Education
- Research: Policy Validation

**Year 4**
- Automate Validation of Assessment Criteria
- Implement Rules to Validate Assertions about Collections
- Research: Collection Value

**Year 5**
- Build Consensus for Standards
- Validate Social Networking Models
- Develop Sustainability Models
- Research: Policy Creation

4/5/2012
Our Approach - Federation

Social consensus for the sharing of data, policies, methods, practice

• Each community controls their own collection policies
  – Policies can be dynamically changed, versioned, and audited

• Explicit computer actionable rules control the type of federation interactions
  – Peer-to-peer, central archive, master-slave data distribution, chained data grids, deep archives

Interoperability mechanisms to support technology integration

• Community specific clients, policies and procedures
• Cross registration of data
• Structured information resource drivers
Questions?