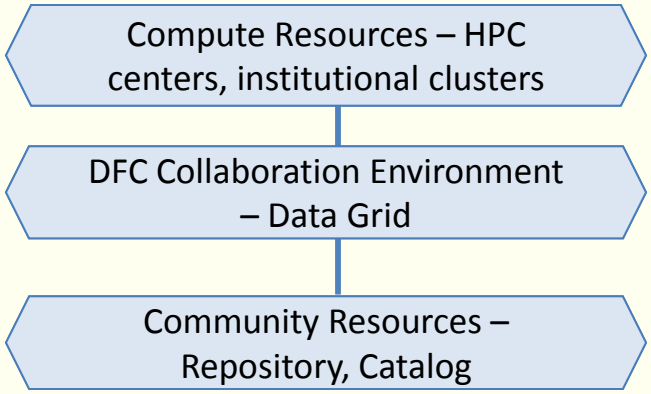
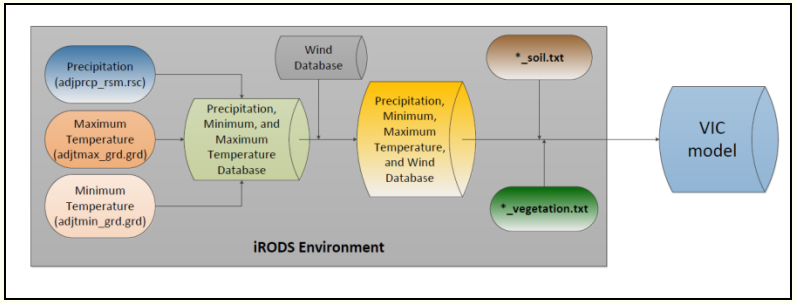
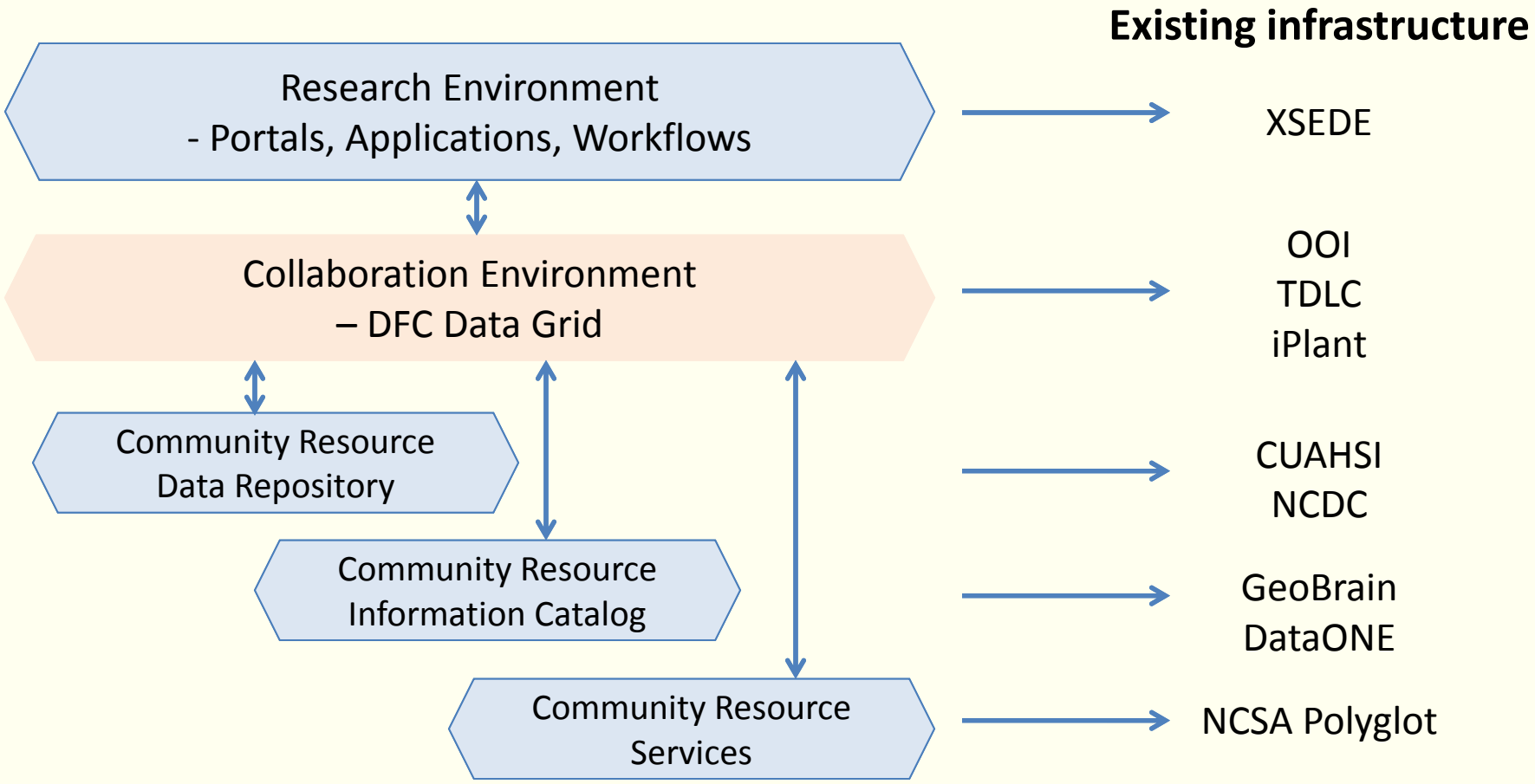
NSF OCI: #0940841 DataNet Federation Consortium

- Enable collaborative research
  - Sharing of data, information, and knowledge
- Build national data cyberinfrastructure
  - Federation of existing data management systems
- Support reproducible data-driven research
  - Encapsulate knowledge in shared workflows
- Enable student participation in research
  - Policy-controlled access to “live” data





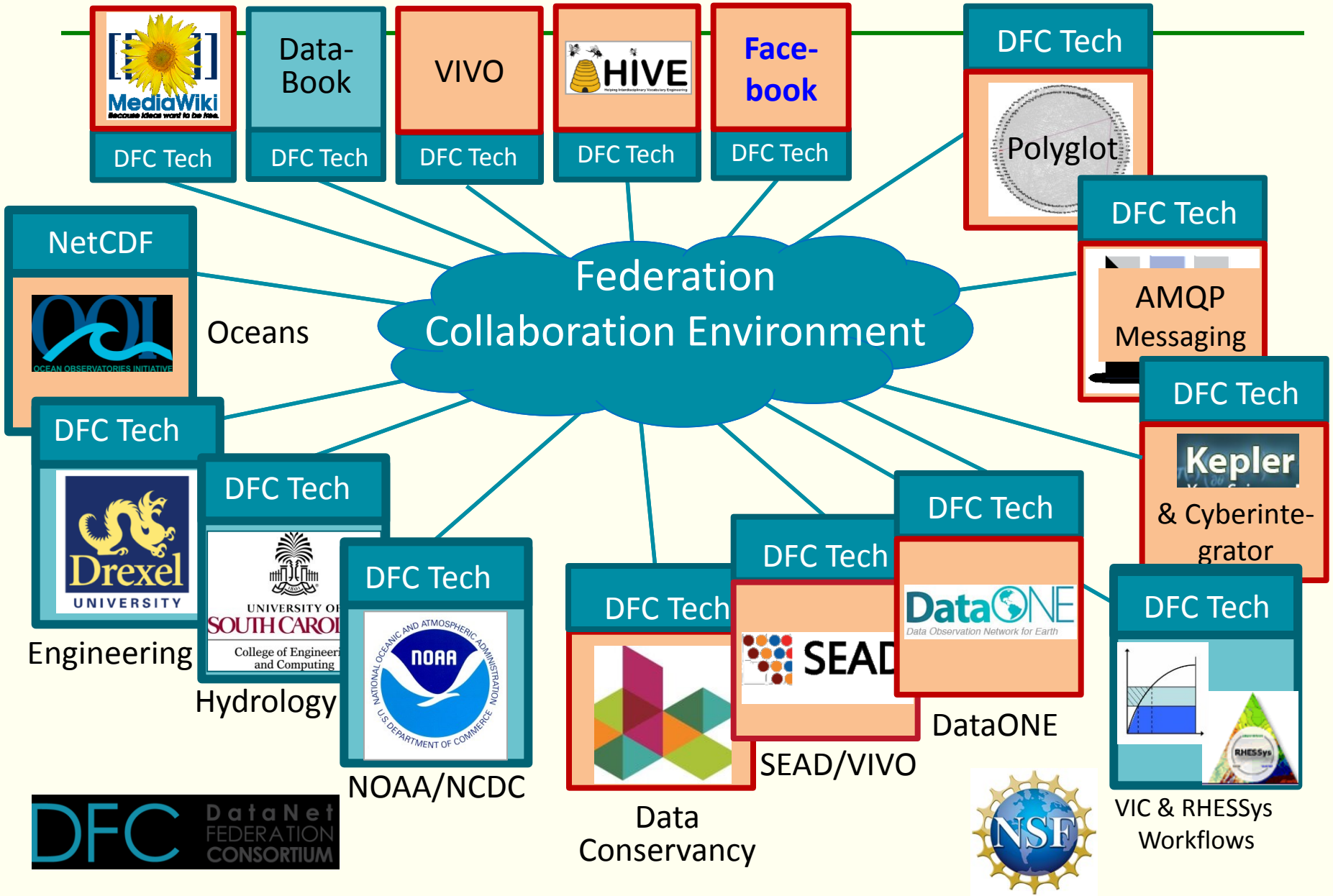
National Infrastructure



NSF DataNet Federation Consortium

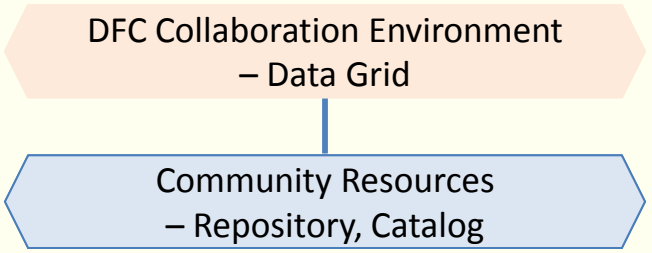
Enabling Collaboration through Interoperability

DFC iRODS-based middleware enables interoperability between heterogeneous clients, data, and service resources

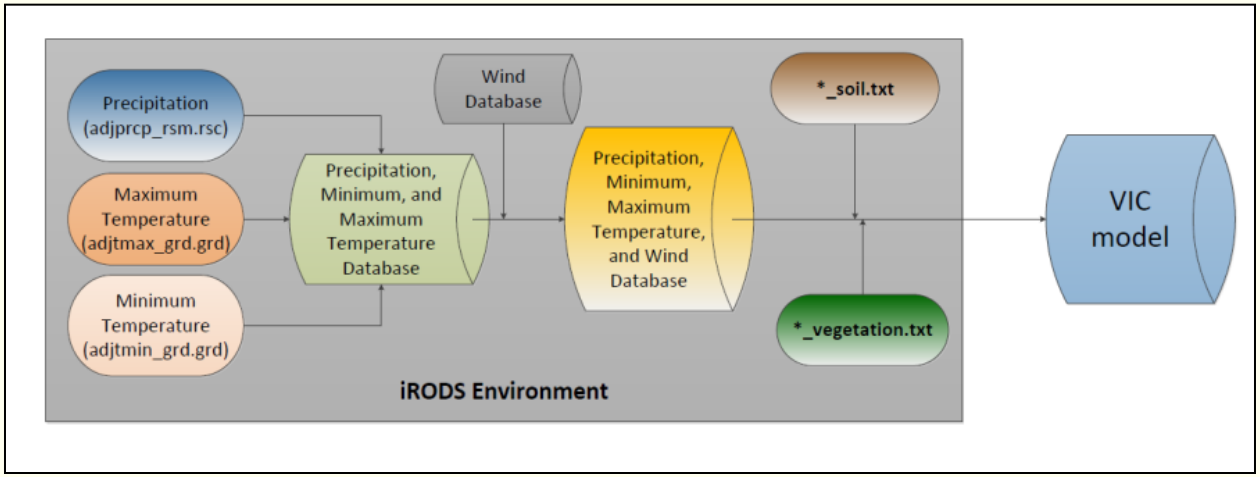


Practitioners' Perspective

- Build community resource
  - Address explicit purpose for formation of a collaboration
  - Build community consensus on provenance, descriptive, system metadata
  - Capture domain knowledge (procedures for interoperability, research analyses, management)
  - Share data, procedures, workflows



- Enable reproducible data-driven research through workflows



Challenges

- DFC uses iRODS policy-based data grid to handle:
  - Acquisition of all relevant data for research
    - Develop micro-services that can access external repositories
  - Distribution of data management effort
    - Use data grid to automate replication of data between agencies
  - Automation of the application of domain knowledge
    - Share workflows used in research analyses
  - Management of policies for data control
    - Enforce policies at each storage location

1. Metadata virtualization (manage properties of metadata – creation time, storage location, access controls, schema)
2. Knowledge virtualization (manage processes that generate metadata – provenance, descriptive, administrative)

iRODS Policy-Based Data Management

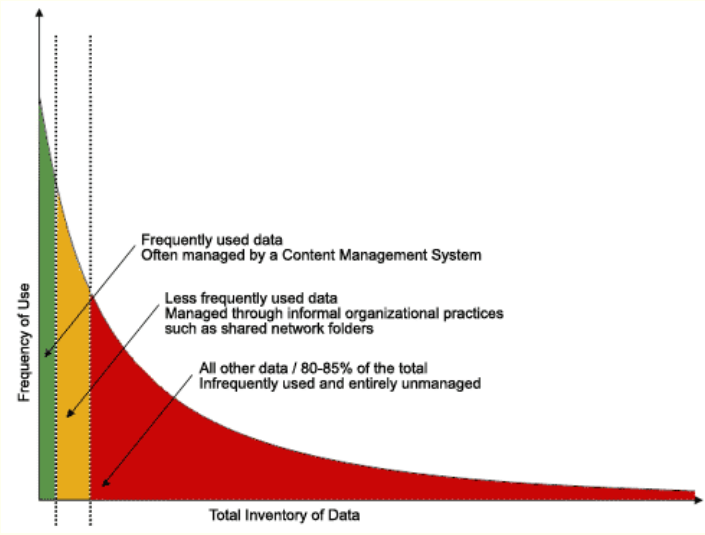
- ***Purpose*** - reason a collection is assembled
- ***Properties*** - attributes needed to ensure the **purpose**
- ***Policies*** - rules to enforce and maintain collection **properties**
- ***Procedures*** - functions that implement the **policies**
- ***Persistent state information*** – metadata from applying the **procedures**
- ***Property assessment criteria*** – validation that **state information** conforms to the desired **purpose**
- ***Federation*** - controlled sharing of **logical name spaces**

- We capture domain knowledge in policies and procedures, and evolve policies to implement data life cycle stages
- Broadening of impact corresponds to evolution of policies to represent consensus of a new larger community

NSF Data Bridge: Solving the First & Last Mile Problems in Big Data

First Mile: Bring the Long-tail of Science Data into Mainstream

Last Mile: Automate Linking, Clustering, and Discovery of Interesting Relationships in Heterogeneous Data



Data Bridge: NSF-funded Big Data Project

–Apply Socio-metric Network Analysis (SNA) to data

–Explore Relationships between Data, Users, Resources, Methods, Workflows, …

–Link through Multi-dimensional vectors

- Similar to, but for data:

LinkedIn YouTube facebook twitter Google+ TAGGED

–Incentives:

- Enable participation in a larger collaboration
- Raise awareness of local data and bring low value per byte data into shared collections

More Information

- DataNet Federation Consortium
  - http://datafed.org
  - UNC-CH, UCSD, Drexel, USC
- Integrated Rule Oriented Data System (iRODS)
  - http://irods.diceresearch.org
  - Application of data grids include
    - NOAA National Climatic Data Center
    - NASA Center for Climate Simulations
    - French National Library
    - Broad Institute genomics data grid
    - International Neuroinformatics Coordinating Facility