iRODS: Data Sharing Technology integrating Communities of Practice

Reagan W. Moore, Arcot Rajasekar
Data Intensive Cyber Environments Center
University of North Carolina at Chapel Hill
Chapel Hill, NC USA 27599-3360
{moore,sekar}@diceresearch.org

Introduction: Cyber infrastructure for sharing digital content is at the cusp of an explosion. There is a convergence of sophisticated social networking, digital library tools and persistent digital archive frameworks with data grid technology. This enables the sharing of digital content from the small to the very large for time frames that can span decades. For national scale projects like the Ocean Observations Initiative (OOI), Temporal Dynamic Learning Center (TDLC), Large Scale Synoptic Survey (LSST), iPlant Collaborative (iPC), and Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI), petabytes of data will be collected and stored across distributed heterogeneous resources under multiple administrative organizations. The diversity of applications and services needed by participating disciplines, and the heterogeneity of data resources in these disciplines, pose challenges in providing a seamless integrated system or infrastructure that integrates policy management, administrative organization and sustainability models. To provide scalable solutions for these and other projects, a service-oriented federation framework is needed that can help not only federate data and services across disciplines and societal communities but also during technology evolution and for long-term fiscal sustainability for current and future usage. The service-oriented framework requires multiple collaborating *communities of practice* to communicate and interact at multiple levels of management and establish policies for data sharing and organization. We call this paradigm the **F**ederation of Cyber Environments through **C**ommunities of **P**ractice (FCP).

Federation Through Communities of Practice: In the traditional science and engineering data life cycle, research data are initially organized as shared collections within projects, then published in digital libraries [such as the protein data bank-PDB] and finally preserved as reference collections for use within education and future research initiatives. Currently, the interactions between and among these communities are on an ad-hoc basis using stove-pipe models that do not extend beyond narrow and sometimes one-off applications. FCP departs from traditional data life-cycle management in providing a uniform platform within which the communities can integrate their services, enforce their internal policies and interact with other communities using well-defined policy commitments. The policy-based interactions that FCP advocates are akin to the standardization of protocols used for communication between pairs of electronic devices. FCP provides an integration paradigm for *communities of practice* to establish management policies that enable access to scientific data at all stages of the data life cycle.

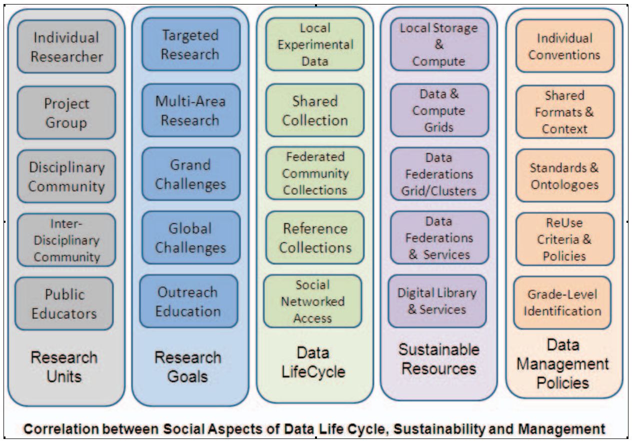


Figure 1. Data Life Cycle Relationship to Social Networks

Each collaborating science and engineering discipline has an objective or purpose that drives their initial choice of data to assemble in a collection. During the life cycle of their data, the purpose may evolve from managing data for a specific researcher, to formation of a shared collection to foster collaboration in a project, to publication in a digital library for use in a wider research domain, to formation of reference collections as the authoritative resource for evaluating future research, to federations with data collections from other disciplines to explore new research initiatives. Each change in the data life cycle corresponds to use of the collection by a broader community, and an associated evolution of management policies to meet the requirements of the expanded set of users. The context that entails the provenance, description, authenticity, integrity, and use of the data must encompass a broader range of knowledge to ensure that the larger community of users will be able to correctly apply and interpret the data. Standards to define semantics, data formats, and analysis tools require a consensus by each new community. Long-term data management requires evolution of management policies to address requirements of an expanding user community.

There is a strong correlation between data life cycle, broadening support for collection use by multiple communities, evolution of data management policies, and sustainability. Figure 1 lists these correlations, which drive the requirements for national scale infrastructure. The horizontal rows represent the stages of the data life cycle. The vertical columns denote the research units that generate the data collections, the motivating research goals, the storage resources used, and types of management policies that may control properties of the shared collection. Building generic infrastructure requires fundamental computer science research into the social network principles that underlie formation of research collections, their associated management policies, assessment criteria, and organizing principles.

The FCP addresses these research challenges through interactions of six communities of practice: 1) Science and Engineering; 2) Facilities & Operations Center; 3) Data Cyberinfrastructure Technology and Research; 4) Policy and Standards; 5) Institutions and Sustainability; and 6) Outreach and Education. These communities of practice are social networks that provide coordination points for seeking input from external groups, for promoting the findings of the FCP, and for extending collaborations to new communities. The communities of practice will lead the formation of a social consensus on the policies and procedures for managing the data life cycle. The FCP will build social networks that integrate the findings of each community of practice, resolve these findings into a consistent set of data grid policies for managing the data life cycle, and demonstrate long-term sustainability through federation across institutional support commitments and storage facility providers.

The FCP provides a virtual platform that integrates existing and emerging social networking services and virtual world technologies for supporting next generation data-intensive collaborative research. By combining policy-based data life-cycle management to provide a solid infrastructural foundation and social interaction services to provide innovative usage models, the FCP will enable long-term preservation of scientific and engineering data as well as incorporate revolutionary data sharing methodologies that facilitate the use of science and engineering data in research and education.

Technology for FCP:

The FCP is based on a service-oriented architecture that implements the multiple layers of federation that are needed for effective data sharing. Moreover, to ensure the services are applicable to diverse types of users from individual researchers to regional collaboratories to national projects and institutions, we propose a standard platform – called the FCP platform that can be used for extensible and scalable data sharing. The aim and activities of FCP are geared towards building a solid common data management platform across multiple disciplines and collections, enabling new discoveries through interdisciplinary questions that previously could not be asked. This is characterized as the socialization of data sharing through the generation of a social consensus on data sharing policies.

The architecture of FCP is a federation of multiple cyber-infrastructure nodes, or *FCP platform,* that consists of a set of distributed services coordinated by a data grid system. FCP Platforms federate to form the FCP Cyber-Infrastructure for the communities of practice as a whole and allows communities to share data and services. The FCP Platform (Figure 2) is a service-oriented architecture providing a broad range of services for technological sustainability in a scalable and extensible solution. The services include back-end services for access to science data collaboratories, compute and storage services, data analysis workflows, and data integration software. FCP Platform also includes front-end services such as digital library interfaces, and social networking interfaces. The FCP Platform also provides cross-cutting services for education, outreach and sustainability support. All of these services are tied together by a data life-cycle management service that migrates data between data life cycle stages.

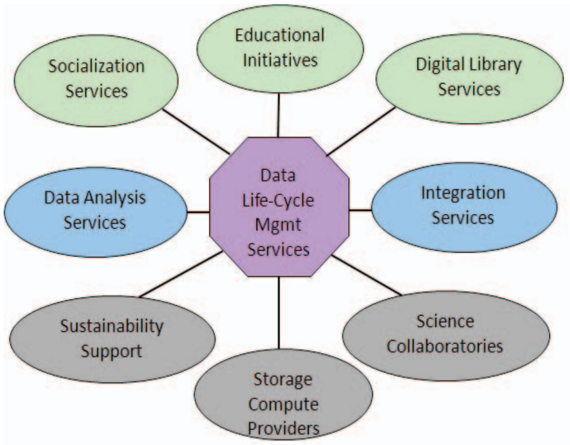
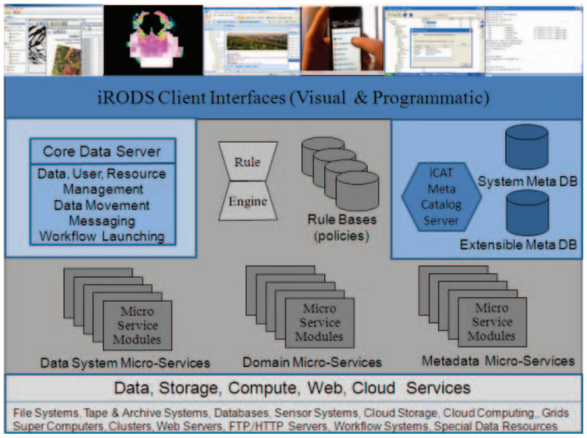


Figure 2. FCP Platform

The Data Life-cycle Management Service is based on the integrated Rule Oriented Data System [1,2]. iRODS was developed by the Data Intensive Cyber Environments Center (DICE) at the University of North Carolina and the University of California, San Diego. At its core, iRODS (Figure 3) is a data server that supports multiple driver interfaces for accessing data from storage systems such as file systems, tape archives, databases, cloud storage, etc. The architecture is extensible and more data resources can be plugged in with very little development. iRODS federates distributed and heterogeneous data resources into a single logical file system (called the collection hierarchy) and provides a modular interface to integrate new client-side applications. For data management in a wide-area network level, iRODS provides services for user authentication, access authorization and usage auditing, optimized data movement protocols and rich support for metadata at multiple levels of data collections. It also provides facilities for data placement, caching, replicating, copying, moving, and versioning, and supports the concepts of retention, disposition, integrity checking and validation. To help keep track of persistent data such as access control lists, user authentication, replica locations, etc, iRODS also has an integrated metadata system called iCAT. The iCAT is built upon a relational database such as Postgres and keeps track of over 60 attributes necessary for data life-cycle management.


iRODS Client Interfaces (Visual & Programmatic)
Core Data Server
Data, User, Resource Management
Data Movement
Messaging
Workflow Launching
Rule
Engine
Rule Bases (policies)
iCAT Meta Catalog Server
System Meta DB
Extensible Meta DB
Micro Service Modules
Micro Service Modules
Micro Service Modules
Data System Micro-Services
Domain Micro-Services
Metadata Micro-Services
Data, Storage, Compute, Web, Cloud Services
File Systems, Tape & Archive Systems, Databases, Sensor Systems, Cloud Storage, Cloud Computing,, Grids
Super Computers, Clusters, Web Servers, FTP/HTTP Servers, Workflow Systems, Special Data Resources

iRODS also has a built-in distributed rule-engine. Administrators and collection owners can encode policies as rules for managing their data collections. The iRODS system provides comprehensive policy-driven data management for all data life-cycle stages including data ingestion/acquisition, access/dissemination, metadata-based categorization and discovery, data protection, security and privacy, long-term preservation and curation, and integration with a wide variety of tools and user interfaces. Services for data creation, ingestion, organization into collections, association of metadata and annotations, and publication and disposition of data all require application of different management policies at each phase of the data life cycle. Also, for each collection and discipline the policies will differ leading to the requirement of an extensible and flexible management system. IRODS provides this capability by mapping policies to computer actionable rules. Rules can be defined for each life-cycle stage. For example, rules for making the desired number of replicas, assigning access control, and computing checksums can be defined and applied on a per collection basis. Similarly, one can define rules for auditing, accounting, redaction or post-processing and vetting of confirmed delivery to be applied on a collection and/or user basis. Management polices are developed for deposition, acquisition, access control, integrity, trustworthiness and privacy (including constraints such as HIPAA), replication, transformation, retention, curation, discovery, access, disposition, data interoperability, and for standards and institutional policy enforcement . These policies govern the usage model for the collections, define the collection assessment criteria, and define the expectations of the originating community.

Data Sharing based on Communities of Practice: The Science and Engineering community both generates and consumes data. The policies they define for scientific data sharing constitute a community of practice. The policies range from the identification of descriptive and provenance metadata that need to be associated with each data set, the verification and validation procedures that are needed to ensure data quality, and policies for data sharing (including Institutional Review Board policies, access control needs and what usage models are allowed). The Facilities & Operations Center provides storage and computational facilities for storing, analyzing and visualizing data. Output from simulations is compared with observational data to evaluate research results. The facilities community requires policies for control of storage and computational resources, including management of protocols for using physical resources, integrating networks, and auditing accounting services. An Institutions and Sustainability community provides the institutional support and long-term framework for sustaining the whole data sharing enterprise. Policies for this community include institutional commitments, intellectual property requirements and cost-based models for business management. The Data Cyber-infrastructure Technology and Research community develops data life-cycle management services. Associated policies include controlling access to data, launching and running workflows for the other services, and keeping track of the virtualization needed to make the system technology independent of choice of vendor. Policies for this community are also discussed in the previous section as part of the iRODS description. A Policies and Standards community interacts with the other communities to standardize and propagate policies across multiple user communities. They will need meta-policies about how to define core properties, control creation of new properties and version property sets. Finally, an Education and Outreach community provides the policies for engaging outside communities in use of the data collections for applications beyond the scientific domain including decision making, education and multi-disciplinary applications. From this perspective, there are multiple communities of practice that develop the policies needed to implement production data management systems. All of these communities need to interact to develop a comprehensive set of policies that can be used to automate administrative tasks, enforce management functions, and validate assessment criteria.

Acknowledgement: The research in this paper is funded by the NARA supplement to NSF SCI 0438741, (2005-2008) and the NSF OCI-0848296 grant (2008-2012). Funding was also provided by NSF ITR 0427196 (2004-2007) and NSF SDCI 0721400 (2007-2010).

References:

[1] A. Rajasekar, M. Wan, M. Moore, and W. Schroeder, “A Prototype Rule-based Distributed Data Management System,” HPDC workshop on Next Generation Distributed Data Management, Paris, France, 2006.

[2] IRODS: integrated Rule Oriented Data System, https://www.irods.org/index.php. (This reference has links to several appers and tutorials on iRODS).