INTEGRATED RULE ORIENTED DATA SYSTEM (IRODS)

Todd BenDor
Associate Professor
Dept. of City and Regional Planning
UNC-Chapel Hill
bendor@unc.edu

http://irods.org/
Important note:

- Most of this is blatantly ripped off from iRODS official documentation
  - iRODS executive summary/introduction
  - Reagan Moore, UNC SILS
  - Adil Hasan, Liverpool
Model integration and flexibility needs sophisticated tools for collecting, curating, and updating databases.

- What happens when our model integration dreams come true?

- Maintaining data provenance and facilitating loose/tight coupling is key.
Overview

- What is iRODS?
- Who Uses iRODS?
- What Can iRODS Do?
What is iRODS?

iRODS is integrated Rule Oriented Data Management System.

iRODS is open source data grid middleware that implements...

- Data Virtualization
- Automation of Data Operations
- A Robust Metadata Catalog
- Data Management Policy Enforcement and Compliance Verification
Based on considerable experience from Storage Resource Broker (SRB) developed by Data Intensive Cyber Environment (DICE) group at UNC, UCSD, SD Super Computer Center.

- Found many groups used SRB to store large quantities of data.
- A lot of server-side post-processing of the data is required (e.g. replicate files, convert to different format, checksum etc).
- Almost all management is policy driven.
Emergence of iRODS

- SRB experience motivated requirements for a new data management system
  - Contained all SRB functionality
  - Add work-flow to manage server-side post-processing
    - Model coupling potential grows
  - Configurable – only include the 'services' you need
  - Open-source – SRB license imposed sever restrictions on the academic community
Claims to fame...

- An infinitely configurable data janitor
  - iRODS is the kind of technology you need to host everyone’s unstructured data.

- A powerful data migration tool

- A data preservation technology

- A tool for providing fine-grained privacy and security controls
  - Can be seen as a basis for a Digital Repository/Archive
  - Digital Repository/Archive is a Policy Driven System
Extensible: iRODS has command-line clients, APIs for numerous programming languages, and web clients
- Data is accessed using familiar APIs
- Supports new plug-ins for storage resources, authentication mechanisms, microservices, and network protection
- iRODS lets system administrators roll out an extensible data grid without changing their infrastructure.

OPEN SOURCE
How will models interface with large amounts of heterogeneous data?

- Databases
- Network Services
- Authentication Mechanisms
- Storage Resources
- Coordinating Resources
- Databases
- Microservices

Familiar APIs

- Command Line Client
- Integrated Custom Client
- Web Client
iRODS is Middleware

Middleware

noun software that acts as a bridge between an operating system or database and applications, especially on a network

- abstracts out the low-level I/O
- provides a uniform interface to heterogeneous storage systems

Applications/Users

iRODS Data Grid

Operating System

Filesystem

(Heterogeneous) Storage Systems
Data Virtualization across Grids

- Administrators control how the grid is presented to users
  - Implement replication, load-distribution, and archiving policies that are completely transparent to the user.

- Independent grids can be federated with one another to allow controlled access to remote grids or grids operated by separate workgroups.
With iRODS, any agent can initiate any action upon any trigger.

This powerful capability allows administrators to automate policies such as:

- Validating checksums every time a new file is placed in a folder.
- Backing up a set of files every second Thursday.
- Archiving data that hasn’t been accessed in over 1 month.
- Logging each time a file is replicated or destroyed.
- Permitting a file to be accessed by multiple independently defined user groups.

These operations can be distributed to the storage resource or client.
Policy Enforcement and Compliance

Verifications

- Metadata + automation = infrastructure to enforce mandated data management policies
  - Records retention and privacy protection requirements
- Audit trails generated by iRODS can be used to verify compliance with policy.
Who Uses iRODS?

- iPlant: 15,000 users on an iRODS data grid with 100 million files
  - http://www.iplantcollaborative.org/
- IN2P3 (French Nat. Inst. Nuclear and Particle Physics): over 6 PB of data managed by iRODS
- Sanger Institute (Genome Research): 20+ PB of iRODS data
- NASA Center for Climate Simulations: 300 million metadata attributes
- CineGRID (exchanging digital media): sites distributed across Japan-US-Europe
What Can iRODS Do?

iRODS simplifies data grid management

Data on different storage devices at different locations can be centrally managed.

In situ migration to new hardware can be managed by replicating the legacy resource before repurposing or decommissioning it.

Backup and archiving are transparent and highly configurable using the automation and metadata capabilities in iRODS.

BnF (2008)
What Can iRODS Do?

iRODS simplifies data discovery, data validation, and data processing.

User-defined and intrinsic metadata make stored data searchable.

Validation and analytical tools can be automated to process incoming data.

The results and process steps can be stored in the iCAT metadata catalog.
What Can iRODS Do?

- **Data at scale**
- **Large number of users**
- **Complex management tasks**
- **Critical policy enforcement**
Where iRODS fits?

Client interacts with digital repository to access data

Access

Digital Repository

Storage

IRODS Spans Digital Repository and Storage Domains
Rules

- Policies are implemented in iRODS as rules.
- Rule is a series of logically connected steps.
- Each step realised as a micro-service.
- IRODS rules fully featured:
  - Contain loops and branches.
  - Can have rules contained within rules.
- IRODS rules read from a rule-file (called core.irb by default).
Questions/ideas?

Todd BenDor
Department of City and Regional Planning
bendor@unc.edu
http://todd.bendor.org