iRODS: integrated Rule-Oriented Data System
http://www.irods.org

The iRODS software supports the formation of shared collections from distributed data, while automating the execution of management policies. The software has been developed with funding from the National Science Foundation and the National Archives and Records Administration. An international collaboration is pursuing the development of additional capabilities. We present the development history that tracks the emergence of rules required for the internal consistency of the data management environment, administrative rules that minimize the labor required to manage the system, assessment rules that validate assertions about the collection properties, and application rules that control execution of desired processing steps.

Rules control the execution of remote micro-services that are simple, well-defined functions that implement specific tasks. A distributed rule engine executes the rules at each storage location where data resides. Programmatic constructions such as recursion, iteration, and conditional tests are supported. The use of server-side workflows minimizes the amount of data sent over networks. The server-side workflows can be executed periodically, minimizing the amount of labor required to maintain large shared collections. The enforcement of management policies is effectively decoupled from the client interface used to access the system. Rules implement transparent policies for distributed data management.

The iRODS system provides standard operations for manipulating data (Posix I/O) and for interacting with structured information resources. The persistent information needed to manipulate data is managed within the iCAT metadata catalog. The information needed to access structured information is managed within the structured information resource. Queries are made on the structured information resource to retrieve the information required to process the data. This interface makes it possible to build and manage a collection that is distributed across arbitrary autonomous data management implementations.

SDSC-led Projects
  o 4/1/06 – design of micro-service architecture to support server-side workflows
  o 12/20/06 – version 0.5 released, containing micro-service framework, rule engine, metadata catalog, file manipulation operations, and logical name spaces for users, files, resources, rules, micro-services, persistent state information, and installation script
  o 6/1/07 – version 0.9 released, supporting data movement, replication, migration, integrity checking, a trash system, system metadata, user defined metadata, metadata queries, administrative functions, and error reporting and recovery. In addition to these, iRODS includes a restart capability (restart file) built into many operations (put, get, replicate, etc), socket reconnection to deal with unreliable networks, and improved database performance and security via the use of ‘bind variables’.
  o Semantic verification of the rule system initiated
  o Java client (JARGON) initial version
  o Active web client initial version
• NARA supplement to NSF SCI 0438741, “Cyberinfrastructure; From Vision to Reality” - Transcontinental Persistent Archive Prototype (TPAP) (2005-2008)
  o Creation of an iRODS testbed for preservation
  o Development of rules needed to migrate records from the TPAP SRB testbed to the iRODS testbed
  o Implementation of a subset of the rules defined in the Electronic Records Archive capabilities list and the Trusted Repository Audit Checklist
  o Development of micro-services to control processing of result sets, support transformative migration of document formats (images), and migration of structured metadata
  o Improved installation design to handle ancillary application packages
  o Initial audit trail implementation, and micro-services for parsing audit trails to track compliance with assessment criteria
  o Development of a mounted collection interface to support standard operations on structured information resources (aggregations of files, remote file systems)
  o Development of structured information interface for creating and parsing AIPS
• NARA supplement to NSF SCI 0438741, “Cyberinfrastructure; From Vision to Reality” - Developing Scalable Data Management Infrastructure in a Data Grid-Enabled Digital Library System (2005-2006)
  o Characterization of Trusted Repository Assessment Criteria as iRODS assessment rules
• NARA supplement to NSF SCI 0438741, “Cyberinfrastructure; From Vision to Reality” – Research Prototype Persistent Archive Extension (2006-2007)
  o Distribution of rules between iRODS and DSpace
  o Specification of administrative rules to minimize labor support requirements
  o Assessment of integrity metrics and availability criteria, and mapping to iRODS rules
  o 3/2008 - Planned implementation of the administrative rules for minimizing labor requirements
  o 5/2008 – Testing and documentation of administrative rules for robustness and performance
• NSF SDCI 0721400, "SDCI Data Improvement: Data Grids for Community Driven Applications" (2007-2010)
  o 12/2007 – first production release which supports basic data grid functionality: structured information interface, xMSG notification service, 64-bit architectures, micro-services for executing web service calls
  o 6/2008 – support for partitioning result sets by location in structured information, location on a tape, and location at a storage location
Collaborative Development Activities

- NSF Ocean Observatory Infrastructure
  - Rules for controlling real-time data streams
  - Integration with message bus technology
- NSF supplement to ITR 0427196, “Constraint-based Knowledge Systems for Grids, Digital Libraries, and Persistent Archives” – Temporal Dynamics of Learning data grid
  - 2/2008 - Rules for managing human subject approval processes
- Monash University
  - 6/2008 – Support for alternate authentication mechanisms (Shibboleth)
- SHAMAN project
  - 6/2008 – Productization of iRODS release, including documentation, load testing, deployment tools, automated updates, and monitoring system
  - 12/2008 - Tools to add or delete micro-services
  - 12/2008 - Tools to support rule creation
  - Integration of Multivalent Document media adaptors (parsers for office products) with iRODS micro-services
  - Integration of Multivalent Document behaviors with micro-services
  - Integration of Data Format Description Language format parsing technology
- James Cook University
  - 6/2008 – Perl, Python, and PHP load libraries
- UK e-Science data grid
  - 6/2008 – HPSS driver
  - 6/2008 – mounted collection interface for tar files
  - User community meeting in Edinburgh
- IN2P3 (Lyon, France)

- 6/2008 – support for initial deadline scheduling for long-running processes (tracking progress against remaining workload)
- 12/2008 – support for initial data grid federation (minimal exchange of rules and micro-services)
- 12/2008 – support for additional storage systems
- GSI support - possibly in collaboration with ZIH, Germany
- Support for multiple types of mounted collections
- Support for extensible iCAT metadata
- Extended semantic verification of the rule system
- Intelligent guided rule generation interface
- Extensions to the xMSG notification service
- Support for extensible metadata
- Extensions to the rule engine to support parallel operation, and reification of rules to improve performance
- Driver for HPSS, NCAR mass store, DCache
- Driver for DB2, mySQL, SQLServer
- Windows port and inQ client port
- Rule creation GUI
- MCAT to ICAT migration tools
- Support for forwarding rules for execution in other workflows (Kepler, Taverna, Chimera)
- Migration tools for importing data from SRB data grids
- Maintenance of SRB data grid technology

Papers and Tutorials: