School of Information and Library Science
University of North Carolina, Chapel Hill
INLS 624– Policy-based Data Management
[Last Updated: 2012-11-11]

Spring 2013
Meeting Time: Friday 9:00 – 11:45
Location: 214 Manning
Credits: 3
Instructor: Reagan Moore
Office: 216 Manning (Moore)
Phone: 919-962-9548 (Moore)
E-Mail: rwmoore [at] renci [dot] org
Office Hours: 2:00 PM on Monday, 2:00 PM on Friday, or by appointment
Course Web Site: https://sakai.unc.edu/portal/site/

COURSE DESCRIPTION

This class will prepare students to develop and implement policies for validating digital repositories and management of digital collections. This includes formulation of policies that can be enacted through computer actionable rules, adapting existing rules and developing new rules. The rules will be applied in the LifeTime Library. The underlying technology is the integrated Rule-Oriented Data System (iRODS), which organizes distributed data into a sharable collection. Rules are used to automate collection administration, or enforce policies, or validate assessment criteria. Students will work in groups to define policies and identify rules for particular digital management situations.

Topics covered will include: policy-based data management systems; policies for data sharing, digital libraries and preservation environments; and trustworthiness assessment criteria. An overview of the iRODS data management system will be provided.

Students will receive accounts in the LifeTime Library, and will be able to develop rules to manage their personal digital library. Previous experience with programming will be very helpful but is not required. Knowledge of the material in INLS-461 "Information Tools" will be very helpful.

Institutions applying policy-based data management systems include:
- UNC-CH Carolina Digital Repository / SILS LifeTime Library
- Data Direct Networks SFA10KE disk storage system
- NASA Center for Climate Simulation
- French National Library
- Texas Digital Libraries

COURSE OBJECTIVES

Upon completion of this course, you should be able to:
- Articulate requirements for trustworthy and sustainable repositories

- Express data management policies that can be supported by computer-executable rules that control operations performed within the iRODS data grid
- Build a representative set of rules appropriate to a specific data management context (e.g. shared collection, digital library, preservation environment, reference collection)
- Install and configure an iRODS data grid
- Test and apply specific rules within iRODS on a set of test files
- Identify ways to verify whether a rule set correctly enforces desired collection properties

PREREQUISITE KNOWLEDGE REQUIRED

Knowledge of the C programming language or a scripting language is desirable, but it is not required. Sufficient information will be provided for students to generate rule sets on their own.

HARDWARE AND SOFTWARE REQUIREMENTS

Students should have access to a laptop (Mac, Windows, or Unix operating system) and will install on their laptops– with help from the instructors – the iRODS software, which is available as an open source download from http://diceresearch.org. If you foresee any problems with this laptop requirement, you should let the instructors know as soon as possible. Students will build and modify rules to control their own personal data grid. Students will receive an account within the SILS LifeTime Library.

COURSE EXPECTATIONS

- Complete readings BEFORE CLASS each week. Manage your time accordingly.
- Come to class on time.
- Participate in discussions – counts as 15% of your total grade for the course.
- Demonstrate concerted effort to successfully complete all lab exercises, and submit products from the exercises that reflect this effort.
- Practice "respectful and informed ignorance." Will Rogers said, "Everybody is ignorant, only on different subjects." This class will be most effective if everyone feels comfortable asking questions, so respect the questions of others. Bring to class your own informed questions about the week's materials (i.e. be able to convey how you've tried to understand the issues and what still remains unclear to you).

Special Needs: If you feel that you may need an accommodation for a disability or have any other special need, please make an appointment to discuss this with the instructor. We will best be able to address special circumstances if we know about them early in the semester.

COURSE REQUIREMENTS

1. Adequate preparation – read required materials each week
2. Participation in class discussions – active engagement with course material, raising questions, contributing to classroom discussions
3. Class participation

4. Completion of in-class lab exercises - Instructions will be provided each week as part of the lab.
5. Group policies and rule set - In a group of students, you will develop a document that includes a set of policies associated with a specific digital management function and a set of rules to support those policies. Each group is expected to write 10 policies during the semester. Example policies will be provided in each class.
6. Final exam - This will be an in-class exam. In order to prepare, you will be provided a list of questions near the end of the semester, and the actual exam questions will be a subset of the list you have already received.

EVALUATION

- Class participation: 15%
- Completion of in-class lab exercises: 30%
- Group policies and rule set: 25%
- Group presentation: 5%
- Final exam: 25%

Based on UNC Registrar Policy for graduate-level courses (http://regweb.unc.edu/resources/rpm24.php), both assignment and semester grades will be H, P, L or F. Few students will obtain an "H," which signifies an exceptionally high level of performance (higher than an "A" in an A-F systems). The following is a more detailed breakdown:

H = Superior work: complete command of subject, unusual depth, great creativity or originality

P = Satisfactory performance that meets course requirements (expected to be the median grade of all students in the course).

L = Unacceptable graduate performance: substandard in significant ways

F = Performance that is seriously deficient and unworthy of graduate credit

COURSE READINGS

Required Text:

The main texts for this course will be provided as pdf files:

Rajasekar, Arcot, Michael Wan, Reagan Moore, Wayne Schroeder, Sheau-Yen Chen, Lucas Gilbert, Chien-Yi Hou, Richard Marciano, Paul Tooby, Antoine de Torcy, and Bing Zhu. *iRODS Primer integrated Rule-Oriented Data System*, ISBN 978-1-60845-333-7, Morgan & Claypool. (available through UNC-CH library at no cost)

Ward, Jewel, Michael Wan, Wayne Schroeder, Arcot Rajasekar, Antoine de Torcy, Terrell Russell, Hao Xu, and Reagan Moore. *The integrated Rule-Oriented Data System (iRODS) Micro-service Workbook*, ISBN 978-1-46646-912-9, Amazon.com. (pdf file available through the class Sakai web site)

Access to Other Readings:

Most other readings for this class are available at specified URLs. In some cases, the reading will be available through Sakai. NOTE: Accessing licensed online materials can

require you either to use a computer with a UNC IP address (generally, a SILS or UNC Library computer) or visit the associated sites through a UNC proxy server. See: <http://proxy.lib.unc.edu/setupinfo.html>http://proxy.lib.unc.edu/setupinfo.html

COURSE SCHEDULE AND TOPICS

NOTE: Most weeks of this course will follow a structure of lecture and discussion. The first half of the Friday session will be devoted to new topics. The second half will be devoted to hands-on application of policies.

Week 1 (January 11) - Course Introduction and Preparation

We will discuss policy-based data management systems, the capabilities they provide, and how they are used to support a variety of data management applications.

Read:

- Moore, Reagan W. "Building Preservation Environments with Data Grid Technology." *American Archivist* 69, no. 1 (2006): 139-58. http://www.metapress.com.libproxy.lib.unc.edu/content/176p51l2w5278567/fulltext.pdf

Friday, January 11 - Introduction to Policy-based Data management

- Policy Based Data Management overview
- Applications in data grids, digital libraries, persistent archives
- Properties / policies / procedures / assessment criteria
- ISO MOIMS-rac trustworthiness assessment criteria

LAB: iRODS Software Installation and Configuration

- Installation of personal iRODS data grid
- LifeTime Library accounts
- Formation of project groups for class project

Week 2 (January 18) – Overview of iRODS architecture

We will discuss the integrated Rule Oriented Data System, which is widely used in academia and the federal government to support large data collections.

Read:

- *iRODS Primer* - Chapter 1 (Introduction), Chapter 2 (iRODS), Appendix A (iRODS Shell Commands)

Friday, January 18 – Rule-based data management architecture

- Data grid virtualization mechanisms
- Policy / procedure implementation
- Policy enforcement
- Federation of data grids

Lab: LifeTime Library accounts

- Set up access to LifeTime Library
- Use web interface

Week 3 (January 25) – Policies to Assess Trustworthiness
Multiple standards efforts have attempted to define repository trustworthiness assessment criteria. We will examine how validation policies can be turned into computer actionable rules.

Read:

- *iRODS Micro-service Workbook*, Chapter 1
- "Metrics for Digital Repository Audit and Certification." Consultative Committee for Space Data Systems. White Book 3. January 2009. http://wiki.digitalrepositoryauditandcertification.org/pub/Main/WebHome/MetricsForDigitalRepositoryAuditAndCertificationWBv03a.doc

Friday, January 25 - Definition of "Policy" and Strategies for Writing Policies

- Trustworthiness assessment policies
- Examples of policies from existing projects and repositories
- How to write policies that can be implemented with specific rules
- Evolution of policy definitions

LAB: Developing Examples of Policies

- Groups select policies they will develop and apply
- Define types of rules that will enforce policies

Week 4 (February 1) – Types of Policies
We will examine how policies can be automatically invoked and used to enforce desired collection properties. Policies can also be used to automate administrative tasks and validate collection properties.

Read:

- *iRODS Primer* - Chapter 3 (iRODS Architecture)
- Green, Ann, Stuart Macdonald, and Robin Rice. "Policy-Making for Research Data in Repositories: A Guide." Edinburgh, UK: EDINA and University Data Library, University of Edinburgh, 2009. <http://www.disc-uk.org/docs/guide.pdf>http://www.disc-uk.org/docs/guide.pdf

Friday, February 1 - System Enforced Policies

- Policy enforcement points
- Policies invoked on common actions
- Management policies, automation of tasks, validation policies

LAB: Policy rule language

- Rule syntax for writing policies
- iRule command for interactive rule execution

Week 5 (February 8) – LifeTime Library Clients
Data management applications can be accessed through a variety of clients, ranging from digital library interfaces such as Fedora to web browsers, web services, workflow systems, file system interfaces, I/O libraries (C, C++, Fortran), load libraries (Java, Python), Unix tools, grid tools, synchronization interfaces, portals, and domain specific libraries.

Read:

- NASA Evaluation of iRODS Dan Duffy, NASA Center for Computational Sciences, March 2009.
- Moore, Reagan. "Towards a Theory of Digital Preservation." *International Journal of Digital Curation* 1, No. 3 (2008). http://www.ijdc.net/index.php/ijdc/article/viewFile/63/42

Friday, February 8 - Synchronization Interfaces, Web Browsers

- iDrop client interface
- iDrop-web browser
- iCommands – unix shell commands
- Windows browser

LAB: Using Interfaces to iRODS

- First policy is due
- Load an initial collection using a preferred interface

Week 6 (February 15) - Introduction to Rule Language
Policy-based systems enable each person to implement policies that control their collections. A rule language provides a way to chain basic functions into a workflow that the data management system can execute. These workflows can be quite sophisticated, enabling the properties of the entire collection to be validated.

Read:

- *iRODS Primer* - Chapter 4 (Rule-Oriented Programming), 5.2 (Rules), 5.3 (Rule Grammar), 6 (iRODS Micro-Services)
- *iRODS Micro-service Workbook*, Chapter 2

Friday, February 15 - Policy, Action, Rule and Micro-Service:

- Definitions, Roles and Relationships
- Core-re rule base, automated rule enforcement
- irule command for interactive execution
- Workflow variables

LAB: Rule Writing Session

- Rules to manipulate workflow variables

Week 7 (February 22) – Workflow composition
Workflow languages provide standard operators for manipulating structured information. An example is the processing of information returned after a query on the metadata catalog, and the use of the information to drive execution of desired tasks such as replication, or metadata extraction, or assignment of retention periods.

Read:

- *iRODS Primer* - Chapter 5 (The iRODS Rule System) up to 5.6 [Note: You already read 5.2 and 5.3, which were assigned last week)
- *iRODS Micro-service Workbook*, 4.67-4.85
- Owens, Evan. "Automated Workflow for the Ingest and Preservation of Electronic Journals." In *Archiving 2006: Final Program and Proceedings, May 23-26, 2006, Ottawa, Canada*, edited by Stephen Chapman and Scott A. Stovall, 109-12. Springfield, VA: Society for Imaging Science and Technology, 2006. http://www.portico.org/news/Archiving2006-Owens.pdf

Friday, February 22 - Managing and Invoking Rules

- Workflow operators for conditional tests
- Workflow operators for looping over micro-services
- Session variables

LAB: Manipulating session variables

- Write rules to manipulate session variables

Week 8 (February 29) – Metadata Management
Each digital collection has associated metadata to track the provenance and description of digital files. In addition, system-level metadata tracks the results of all operations performed within the data management system. A major goal of any data management application is the self-consistent management and preservation of all metadata.

Read:

- *iRODS Primer*- 5.6 (Default iRODS Rules), 5.7 (Session Variables available for each rule)
- *iRODS Micro-service Workbook*, 4.125-4.150
- PREMIS Editorial Committee. *PREMIS Data Dictionary for Preservation Metadata, Version 2.0*. March 2008. [Read p.1-21 in detail, then browse through the rest of the document] http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

Friday, February 29 - Manage Descriptive and Provenance Metadata

- Persistent state variables
- Descriptive and provenance metadata loading
- Querying the metadata catalog

LAB: Retrieve User-Defined Metadata

- Write rules to query catalog
- Write rules to load metadata

Week 9 (March 8) - Conditions for Rules
A digital library may contain multiple collections, each with their own controlling policies. It is possible to manage multiple collections within the same data management system through use of conditions on policies. Policies can be checked for whether they should be applied to a given collections.

Read:
- *iRODS Micro-service Workbook*, 4.95, 4.101

Friday, March 8 - Examples and types of conditions to include in rules
- Conditions on session variables
- Conditions on input variables
- Conditions on workflow processing

LAB: Tuning rule application
- Write rule to restrict operations to a specific collection
- Write rule to restrict operations to a user group
- Two policies are due

Week 10 (March 15) - NO CLASS (Spring Recess, March 7-17)

Week 11 (March 22) - Timing of Rules
Within data management systems, policies can be applied atomically on every client interaction through policy enforcement points, or may be deferred for execution at a later time, or may be executed periodically. Typically, assessment criteria are implemented as periodic policies, while the enforcement of management policies is done atomically on each client access.

DUE: MARCH 22 BY START OF CLASS - DRAFT of Final Group policy and rule set for feedback from instructors

Friday, March 22 - Atomic, Periodic and Deferred Rules - how and why to use each
- Deferred rule operation
- Periodic rule operation

LAB: Creating three different rules for a given operation: atomic, periodic and deferred
- Write rule to defer validation of a checksum
- Write rule to periodically validate checksums
- Four policies are due

Week 12 (March 29) - Holiday

Week 13 (April 5) - Scheduling Execution of Policies
Very long running tasks may be needed to validate assessment criteria that check every file in the collection. Through use of the workflow language, it is possible to monitor the execution rate, and implement policies that are designed to finish by a specified date. The assessment criteria are run at a predictable I/O rate that minimizes impact on the digital library.

Read:

- Just-in-time rule, Sakai class web site

Friday, April 5 - Just-in-time policy execution

- Management of policy execution
- Examples of policy scheduling

LAB: Application of scheduling

- Apply just-in-time scheduler to control processing of a collection
- Six policies are due

Week 14 (April 12 – Validating policies
Every action that is performed within a digital library can be validated. This is typically done by checking the state information that is generated by a management policy, or by parsing an audit trail to track compliance over time, or by applying a procedure on each file within the system. An example of the latter is the validation of checksums to ensure data have not been corrupted.

Read:

- *iRODS Micro-service Workbook*, 4.188-4.192 – audit trails
- *iRODS Micro-service Workbook*, 4.238-4.242 – integrity checks

Friday, April 12 - Verification of policies

- Query information from metadata catalog
- Parse audit trails
- Re-compute desired property

LAB: Policy validation

- Implement a validation policy for one of your management policies
- Eight policies are due

Week 15 (April 19) – Linking external data sources
A digital library may contain physical files and pointers to files that reside in other data management systems. We will examine how to build a logical collection that links to material at external web sites.

Read:

- *iRODS Micro-service Workbook*, 4.216-4.229

DUE: APRIL 19 BY START OF CLASS - Final Group policy and rule set

Friday, April 19 - Soft links

- Management of links to external data management systems
- Active objects, Policy encoded objects

LAB: Accessing remote web sites

- Select web site to add to your collection through a soft link
- Write a rule to periodically harvest web site
- All 10 policies are due

Week 16 (April 26) - Group Presentations and Discussion
The policies developed within the course will be applied to the LifeTime Library. The goal is to understand how to automate management of the library, automate ingestion of material into the library, and automate validation of the library properties.

Friday, April 26 - Group presentations and class discussion of policies

- 15 minute talks on policy development by each group

Week 17 (May 3) – Final Exam
The exam will consist of essay questions about types of policies that can be used, and applications of policy-based data management systems.

Friday, May 3 - Final exam preparation and course evaluations

- Answer questions