Policy-based Data Management

INLS 624-01W

Instructor: Reagan Moore

**COURSE DESCRIPTION**

This class will prepare students to develop and implement policies for 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 developed for 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 define policies and implement rules for one of their own collaborations that is building a data collection. This includes identifying the collection properties that are required by a social consensus across the collaboration, selecting policies that will be applied to control the data collection, and implementing example policies using the iRODS data grid.

Topics covered will include: iRODS data grid architecture, design templates for policies, and appraisal of existing policy sets for collections, digital libraries, data centers, preservation environments, protected data, and National Science Foundation data management plans.

Students will receive accounts in the LifeTime Library, and will be able to develop rules to manage a proposed collaboration. 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.

Projects applying policy-based data management systems include:

* Collections: UNC Lineberger Cancer Institute, UNC Genomics data grid
* Collaborations: the DataNet Federation Consortium, the iPlant Collaborative
* Libraries: SILS LifeTime Library, French National Library
* Archives: Carolina Digital Repository, NOAA Climatic Data Center
* Processing pipeline: HydroShare

The material for the course is available at

<https://dfcweb.datafed.org/idrop-web2/home/link?irodsURI=irods%3A%2F%2Firen2.renci.org%3A1237%2Fdfcmain%2Fhome%2FDFC-public%2FPolicy-course>

This includes

* Policy-course-workbook.  This contains a transcription of the videos and the slides
* Slides.  A separate slide set is made for each of 120 topics, 30 use cases, 120 exercises, and 15 homework assignments.
* Videos.  A separate video is available for each of the 30 classes.  Each class covers 4 topics and 1 use case.
* Rules.  This provides 65 rule examples that will be examined in the course.
* Papers.  The papers referenced in the course are also available here.

**COURSE OBJECTIVES**

Upon completion of this course, you should be able to:

* Articulate requirements for automating management of a data collection
* 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 context (e.g. shared collection, digital library, preservation environment, reference collection)
* 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). 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 collection. Students will receive an account within the SILS LifeTime Library.

**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 STRUCTURE**

1. The class consists of 30 sessions (twice per week for 15 weeks).
2. Each session will contain 4 topics, 4 related exercises, and an application example.
3. Each week a homework assignment will be given – typically a 1-page essay.
4. A final will be given requiring a 10 page essay.

**EVALUATION**

A “fill-the-pipe” grading scheme will be used

* Each exercise is worth 1 point
* Each homework assignment is worth 10 points
* Final is worth 75 points

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: >314 points

P = Satisfactory performance (expected to be the median grade): 285-314 points

L = Unacceptable graduate performance: 255-284 points

F = Performance that is seriously deficient: <255 points

**COURSE READINGS**

Five books will be used. PDF files for each book are available in Sakai.

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 on Sakai, INLS624.001,SP16)

Chen, Sheau-Yen, Mike Conway, Jon Crabtree, Cal Lee, Sunitha Misra, Reagan W. Moore, Arcot Rajasekar, Terrell Russell, Isaac Simmons, Lisa Stillwell, Helen Tibbo, Hao Xu, *Policy Template Workbook – iRODS 4.2*, (pdf file available on Sakai, INLS624.001.SP16)

Chen, Sheau-Yen, Mike Conway, Jon Crabtree, Cal Lee, Sunitha Misra, Reagan W. Moore, Arcot Rajasekar, Terrell Russell, Isaac Simmons, Lisa Stillwell, Helen Tibbo, Hao Xu, *Policy Examples Workbook – iRODS 4.2*, (pdf file available on Sakai, INLS624.001.SP16)

Moore, Reagan, *Policy-based Data Management*, (pdf file available on Sakai, INLS624.001Sp16).

**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 TOPICS**

Week 1 - Jan 11

Homework 1 - Identify 10 essential properties for a digital library that should be

conserved

1 Present requirements for course, syllabus

**Exercise** - Get an account on the LifeTime Library. Calculate number of

exercises, and homework assignments you need for a desired

grade. Identify which areas are most important to you

2 Present LifeTime Library

**Exercise** – Load description of course expectations into shared collection for

class materials on the LifeTime Library.

3 Present policy concept graph

**Exercise** - Trace the concept of integrity through the policy concept

graph - purpose, policy, procedure, persistent state, assessment

4 Describe types of purposes and properties

**Exercise** - For your collaboration, which properties will be important

Application 1 – LifeTime Library

5 Describe types of policies (PEP, delayed, interactive)

**Exercise** – Design a research collaboration, what type of policies will be

needed

6 Describe application of policies (management, administration, assessment

versus users, collection, technology vs meta-policies)

**Exercise** - Across the policy categories, which control the users of a

collection versus the properties of the collection

7 Explain virtualization concept - manage properties independently of choice

of technology

**Exercise** - Which technologies will pose challenges for management (tape

storage, authentication, identifiers)

8 Explain virtualization required by research collaborations (collection,

workflows, data flows)

**Exercise** - For your collaboration, define at least one collection, one

workflow, one data flow that will be needed

Application 2 - Present DFC federation hub

Week 2 - Jan 18

Homework 2 - identify 10 requirements for a digital library that can be automated

9 Explain virtualization of collections - types of properties

**Exercise** - Identify which collection properties will be most important

10 Explain virtualization of workflows - types of properties

**Exercise** - Identify which workflow properties will be most important

11 Explain virtualization of data flows - types of properties

**Exercise** - Identify which data flow properties will be most important

12 Present iRODS data grid

**Exercise** - Which components of the data grid are needed for integration

with hardware technologies, software technologies

Application 3 - Present survey of 25 disciplines using policy-based data management

13 Explain collection life cycle and evolution of context

**Exercise** - Identify the stage of a collection life cycle that will represent your

collaboration

14 Explain LifeTime Library access and use

**Exercise** - Log onto LifeTime library, load 5 files into a collection

15 Demonstrate Cloud browser client for accessing repositories

**Exercise** - Try setting metadata, access controls

16 Demonstrate Cyberduck client for accessing data

**Exercise** - Install, Try accessing and loading 5 files

Application 4 - Dataverse use case

Week 3 - Jan 25

Homework 3 - Identify persistent state attributes that are needed for your 10 requirements

17 Explain persistent state information

**Exercise** - Which persistent state information attributes are needed to

describe data properties

18 Explain irods rule language syntax

**Exercise** - Write a simple rule that says Hello World

19 Explain rule language variables

**Exercise** - Identify important session variables for identifying person

connecting to the data grid

20 Explain identifiers, operations on identifiers

**Exercise** - Identify which identifier can provide location, access control,

arrangement, descriptive metadata

Application 5 - Google search on iRODS for user communities

21 Explain micro-services

**Exercise** - Identify micro-services that can be used to manipulate data

22 Explain tickets

**Exercise** - Create a ticket for sharing data (iDrop-web).

23 Demonstrate WebDAV client for accessing data

**Exercise** - Try accessing and loading 5 files using WebDAV

24 Demonstrate iCommands for accessing data

**Exercise** - Install and try loading 5 files

Application 6 - Carolina Digital Repository

Week 4 - Feb 1

Homework 4 - Create a policy template for 5 NSF DMP policies

25 Explain policy template Create policy template for access control

**Exercise** - Create a policy template for NSF DMP policy

26 Explain policy-enforcement points - for users, collection management,

technologies

**Exercise** - Which type of PEP is most important for your collaboration

27 Explain types of enforcement, PEP vs delayed vs interactive policies

**Exercise** - Which is more important, PEP versus delayed rule execution

28 Explain relationship between data, information, knowledge, wisdom and

realization in procedures

**Exercise** - Identify wisdom needed to control collaboration (when and

where controls should be applied)

Application 7 -iPlant use case

29 Explain data fabrics (discovery, collaboration, federation)

**Exercise** - Identify policy differences between types of data fabrics

30 Explain core-re - basic policies needed by everyone

**Exercise** - Which basic policies are related to controlling technology versus

controlling users

31 Explain policy requirements for NSF DMP

**Exercise** - Pick the policy requirements for your collaboration from NSF DMP policies

32 Explain policy requirements for protected data

**Exercis**e - For your data, identify types of protected information

Application 8 - French National Library use case

Week 5 - Feb 8

Homework 5 - Create a policy template for a preservation policy

33 Explain policy requirements for preservation

**Exercise** - Consider what you will need 20 years from now to understand a

context for your data

34 Explain policy requirements for RDA data centers

**Exercise** - How do the data center policies differ from preservation policies?

35 Explain policy requirements for LifeTime Library

**Exercise** - Identify 3 additional policies for a digital library (consider DMP)

36 Explain policy requirements for data sharing

**Exercise** - Identify 3 additional policies for data sharing (consider DMP)

Application 9 - LifeTime Library rules

37 Explain idea behind auditing operations - provability

**Exercise** - Pick an event to audit that would prove the data management

system is working

38 Explain pluggable rule engine - iRODS, python, javascript

**Exercise** - Which policy-enforcement points are needed to trap the events

you proposed to audit

39 Explain data types - string, integer, double, boolean

**Exercise** - Write a policy that adds two number, and writes out the result.

Concatenate two strings and write.

40 Demonstrate iDrop-web client for accessing data

**Exercise** - Try running a rule from the client

Application 10 - IN2P3 use case

Week 6 - Feb 15

Homework 6 - Create a policy template for a protected data policy

41 Explain micro-service input and output W

**Exercise** - write a rule that creates a directory - msiCollCreate

42 Explain list creation, extension, modification

**Exercise** - Write a rule that creates a list. Then modify the list by replacing

one element.

43 Explain rule language workflow operators - loops

**Exercise** - Write a rule that adds the numbers from 1 to 10.

44 Explain rule language workflow operators - conditional tests

**Exercise** - Write a rule that writes out the value for the 5th iteration of a loop

from 1-10

Application 11 - DPN Digital Preservation Network

45 Explain queries on persistent state

**Exercise** - Write a query on your USER\_NAME to find your USER\_ID

46 Explain dot operator

**Exercise** - Write a rule that conts the number of files your have in your

collection.

47 Explain schema indirection for metadata

**Exercise** – Create some metadata attributes on two files. Run a rule to

discover the files based on the attribute values.

48 Explain rule for listing metadata

**Exercise** - Set an attributes on a file, then list the attributes

Application 12 - DDN use case

Week 7 - Feb 22

Homework 7 - Modify a log rule to concatenate information into a manifest

49 Explain policy function library

**Exercise** - Write a rule that uses a policy function for initializing a list

50 Explain log files

**Exercise** - Write a rule that creates a log file

51 Explain checking input parameters

**Exercise** - Write a rule that checks a collection input parameter

52 Explain unix time and date manipulation

**Exercise** - Write a rule that gets the systems time, converts to human format,

and prints

Application 13 - EMC client

53 Explain delayed execution, remote execution

**Exercise** - Write a rule to do a delayed execution of a write to a log file

54 Explain reproducible research

**Exercise** - Create a manifest file for a collection

55 Explain workflow structured objects

**Exercise** - Generate a workflow that writes to the manifest file

56 Explain workflow provenance

**Exercise** - Re-execute rule and check entries to the manifest file

Application 14 - Hydroshare use case

Week 8 - Feb 29

Homework 8 - Modify a metadata rule to identify files with similar metadata names

57 Explain metadata parsing - XML

**Exercise** - Create XML file for loading metadata on 2 files

58 Explain metadata parsing - pipe

**Exercise** - Create pipe-delimited file for loading metadata on 2 files

59 Explain tags - units field

**Exercise** - Create two tags, then list the tag AVUs

60 Explain HIVE ontology

**Exercise** - Select a preferred metadata schema for your collection

Application 15 UNC-CH genomics data grid

61 Explain rule for verifying presence of required metadata

**Exercise** - Set metadata on a collection, then check that the files have the

metadata

62 Explain rule for verifying compliance with HIVE ontology

**Exercise** - Compare metadata on a file with the selected HIVE metadata

schema

63 Explain rule for exporting metadata as XML

**Exercise** - Create XML file for the metadata on the files in your collection

64 Explain rule for exporting metadata as JSON

**Exercise** - Create JSON file for the metadata on the files in your collection

Application 16 – Genomics use cases, Bayer and Genedata

Week 9 - Mar 7

Homework 9 - Identify 10 assessment criteria that should be periodically evaluated

65 Explain assessment criteria

**Exercise** - Decide which state information should be verified (integrity,

metadata, replicas)

66 Explain report generation - internal file, external file

**Exercise** - Write a rule to manage externally generated reports

67 Explain rule for summarizing collection size

**Exercise** - Run a rule to summarize your collection size and number of files

68 Explain storage quotas

**Exercise** - Retrieve the amount of storage used on each resource

Application 17 - TDLC use case

69 Explain rule for checking quota

**Exercise** - Check how much space you have left

70 Explain rule for creating list of top 10 users

**Exercise** – Run rule to check # files, # metadata attributes, and storage

71 Explain rule for checking storage cost

**Exercise** - Check the cost for storing your data

72 Explain rule for identifying storage systems being used by the data grid

**Exercise** - Run rule to see track information about resources

Application 18 - Google search on iRODS - what changed

Week 10 - Mar 21. Note spring break Mar 14

Homework 10 - Show how groups can be used to manage restricted access

73 Explain groups

**Exercise** - List the members of the Class624 group

74 Explain rule for adding users

**Exercise** - Decide what properties should be associated with each new user

(trash directory, home directory, quota)

75 Explain rule for identifying file extensions

**Exercise** – Set file data types in your collection

76 Explain rule for checking format type

**Exercise** - Run a rule to identify data types in your collection

Application 19 -ADCIRC use case

77 Explain examples of PII, PCI, PHI

**Exercise** - Create a file with a birth date in the text and store in the data grid

78 Explain BitCurator for detecting personal information

**Exercise** - Run bitcurator to find the file with a birth date

79 Explain IPR constraints

**Exercise** - Identify which files should be proprietary

80 Explain public vs anonymous vs account access

**Exercise** - Set access for other members of the class on a file. Set public

access on a second file

Application 20 - neutrino experiment use case

Week 11 - Mar 28

Homework 11 - Identify 5 types of disposition policies

81 Explain rule for setting access approval flag or use agreement flag

**Exercise** - Add metadata to a file with an approval access flag

82 Explain rule for enforcing access controls - inheritance

**Exercise** – Set inheritance on a collection

83 Explain rule for checking all access controls on a collection

**Exercise** - Identify all the persons including group members who have access

to a sub-collection

84 Explain rule for time-dependent access control

**Exercise** – Check a time-dependent access control

Application 21 - NOAA NCDC use case

85 Explain rule to check for viruses on ingestion

**Exercise** - Run clamscan on a file to detect virus

86 Explain retention policies

**Exercise** - Define retention policies for your files

87 Explain disposition policies

**Exercise** – Define disposition policies for your files

88 Set disposition flag

**Exercise** - Set up a disposition policy to process the disposition flag

Application 22 - NASA NCCS use case

Week 12 - April 4

Homework 12 - Modify a staging rule to create a file version if the file already exists

89 Explain encryption

**Exercise** - Choose the type of encryption you want for your proprietary file

90 Explain rule for encryption

**Exercise** - Run the encryption rule on the file, reset the size

91 Explain rule for creating versions

**Exercise** – Run a versioning rule. Choose between time stamps or

incremental version number

92 Explain rules for backup

**Exercise** - Run a backup rule

Application 23 - EUDAT

93 Explain rule for staging

**Exercise** - Create a staging directory, load files from stage area into an

archive

94 Explain rule for summarizing collection properties

**Exercise** – Review your decision on the properties that your research

collection should enforce

95 Explain integration of a summarization report with a manifest file

**Exercise** - Run rule to sumarize your collection

96 Explain rule for manifest creation for a bacup directory

**Exercise** – Define how you would creae a manifest file for every time you

create a backup

Application 24 - French data grid

Week 13 - April 11

Homework 13 - Identify 5 assertions that a preservation environment should enforce

97 Explain preservation assertions - authenticity, integrity, chain of custody,

original arrangement

**Exercise** - Which of the preservation assertions should be applied to a digital

library

98 Explain RLG-NARA req and trustworthiness

**Exercise** – Define 5 requirements for a trustworthy repository.

99 Explain TRAC req

**Exercise** - Pick a requirement and choose a rule (see Odum policies)

100 Explain ISO 16363 req

**Exercise** - Pick a requirement and choose a rule (see DMP policies)

Application 25 - Syndicate

101 Explain ISO 16363 policy set

**Exercise** - Pick a requirement and define a policy template (see HIPAA

policies)

102 Explain integrity mechanisms- checksums, replication

**Exercise** – How often should you verify checksums?

103 Explain rule for BagIT

**Exercise** - Run bagit to generate a tar file

104 Explain notification rule

**Exercise** - Decide which types of events deserve a notification email

Application 26 - BestGrid

Week 14 - April 18

Homework 14 - Identify 5 assessment criteria that should be periodically validated.

How are these related to the assertions

105 Explain integrity assessment

**Exercise** - Decide on the number of replicas you need. Where are they

stored?

106 Explain rule for verifying checksums

**Exercise** - Run a checksum verification rule on your collection

107 Explain rule for verifying replicas and distribution

**Exercise** - Verify replicas on your collection

108 Explain auditing meta-rules

**Exercise** - Identify the events that you want to audit (access, changes,

accounts)

Application 27 - DataONE

109 Explain rule for parsing audit events

**Exercise** - Run a rule to summarize auditing events

110 Explain rule for verifying compliance with access controls

**Exercise** - Run rule for checking access controls

111 Explain data sharing rule set - 11 rules

**Exercise** - Identify collaboration requirements for quota, strict access

controls, replication

112 Explain federation - tightly coupled.

**Exercise** - Access the dfcmain federated data grid. Try loading 5 files

Application 28 - VISR

Week 15 - April 25

Homework 15 - Summarize the most important concepts learned in the course

113 Explain rule for staging data across a federation

**Exercise** - Run rule to copy 5 files to the federated data grid

114 Explain federation - loosely coupled

**Exercise** – Identify remote services that would be useful to your

collaboration

115 Explain federation - asynchronous

**Exercise** – Which asynchronous federation mechanism is most useful

116 Explain information exchange models

**Exercise** – Compare storing information in a file versus metadata

Application 29 - DataBook

117 Explain production capable policies - requirements

**Exercise** - Examine the production policies and select the set your

collaboration should enforce

118 Explain production capable policies - management

**Exercise** - Select two production policies to run and execute them

119 Explain rule for policy management and distribution

**Exercise** - Outline the steps needed to manage policies across distributed

servers

120 Discuss quality control - assertions about collection

**Exercise** - Identify the set of quality metrics needed for your research

collaboration

Application 30 – DataBridge

Final Paper – May 2

Write a 10-page paper that describes how you would apply the concepts learned in the course to a research collaboration or digital library. The major source for this paper can be the homework assignments submitted during the semester. Please turn the paper in by the evening of May 2, 2016.