Automation of Preservation Functions

Integrated Preservation Infrastructure Prototype

Data Intensive Cyber Environments Center
University of North Carolina at Chapel Hill

- http://dice.unc.edu

Institute for Neural Computation
University of California, San Diego

- http://diceresearch.org
- http://irods.org

Sustaining Heritage Access through Multivalent ArchiviNg
University of Liverpool

- http://shaman-ip.eu/shaman/

Renaissance Computing Institute

- http://www.renci.org

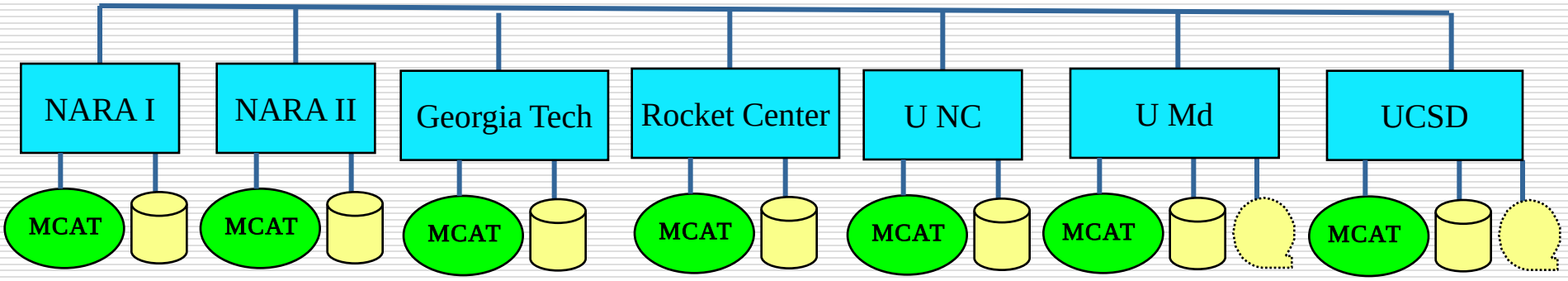
Shaman / iRODS / EnginFrame

- Sustaining Heritage Access through Multivalent ArchiviNg
- Cheshire3 / Multivalent
  - European Union funded grant
  - Rob Sanderson
  - Paul Watry (PI - University of Liverpool)
  - Ken Arnold
  - Jerome Fuselier
  - John Harrison
  - Fabio Corubolo
  - Ann Gledson
  - Adil Hasan
- Integrated Rule Oriented Data System
  - Reagan Moore
  - Wayne Schroeder
  - Mike Wan
  - Arcot Rajasekar
  - Antoine de Torcy
  - Chien-Yi Hou
  - Richard Marciano
- RENCI - EnginFrame
  - Leesa Brieger
  - Mike Conway
- DCAPE - Policies

National Archives and Records Administration
Transcontinental Persistent Archive Prototype

Demo-14

Federation of Seven Independent Data Grids



Extensible Environment, can federate with additional research and education sites. Each data grid uses different vendor products.

Automation through Policies

- Policies controlling ingestion of records
  - Cheshire3
- Policies controlling indexing and arrangement
  - Cheshire3 and Multivalent
- Policies controlling preservation
  - iRODS
- Policies controlling assessment criteria validation
  - iRODS and EnginFrame
- Policies controlling presentation
  - Multivalent and Cheshire3

SHAMAN Approach

SHAMAN
Sustaining Heritage Access through Multivalent ArchiviNg

- Perpetual Data Access
  - Keep data in original format
  - Media Engines read and display original data file
- Evolvable
  - Pluggable framework for support of new media
  - Framework and Engines adapt to new platforms
  - Preserve the preservation environment
  - New display mechanisms can be applied to legacy formats
- Scalable
  - Do not have to migrate the entire collection to new formats
  - Parsing is done on display
- Living Documents
  - Format-independent annotations

iRODS Approach

- Infrastructure independence
  - Manage properties of the records independently of the choice of storage system
  - Manage properties of the preservation environment
    - Policies
    - Procedures
    - State information
- Provide bulk operation support
  - Parallel I/O
  - Containers - tar files
  - Metadata load
  - Remote filtering
- Enforce management policies at the remote storage location

EnginFrame Approach

- Provide presentation layer for records
  - List records and descriptive metadata
- Provide presentation layer for preservation environment
  - List users
  - Parse and filter audit trails
- Support interactive invocation of iRODS rules
  - List management rules
  - List management procedures (micro-services)

DCAPE: Distributed Custodial Archival Preservation Environments

Purpose:

Build a distributed production preservation environment that meets the needs of archival repositories for trusted archival preservation services

Distributed partnership of 11 institutions: 33 people

*** STATES:**

- California
- Michigan
- North Carolina
- Kansas
- Kentucky
- New York

*** UNIVERSITIES:**

- Tufts University
- West Virginia University
- UNC (SILS/RENCI/DICE Center)

*** CULTURAL ENTITIES:**

- Getty Research Institute

*** INTERNATIONAL PARTNERS:**

- Carleton University (Geomatics and Cartographic Research Centre)

D·I·C·E Data Intensive Cyber Environment | i·R·O·D·S | THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL | UCSD | UNIVERSITY OF LIVERPOOL | SHAMAN Sustaining Heritage Access through Multivalent ArchiviNg | NSF | THE NATIONAL ARCHIVES ARCHIVES.GOV

DCAPE Preservation services

- Ingestion (SIP validation, packaging)
- Staging
- Archival storage
- Administration
- Preservation planning
- Access
- Common services
- Management

Examples... Starter list of 25 services

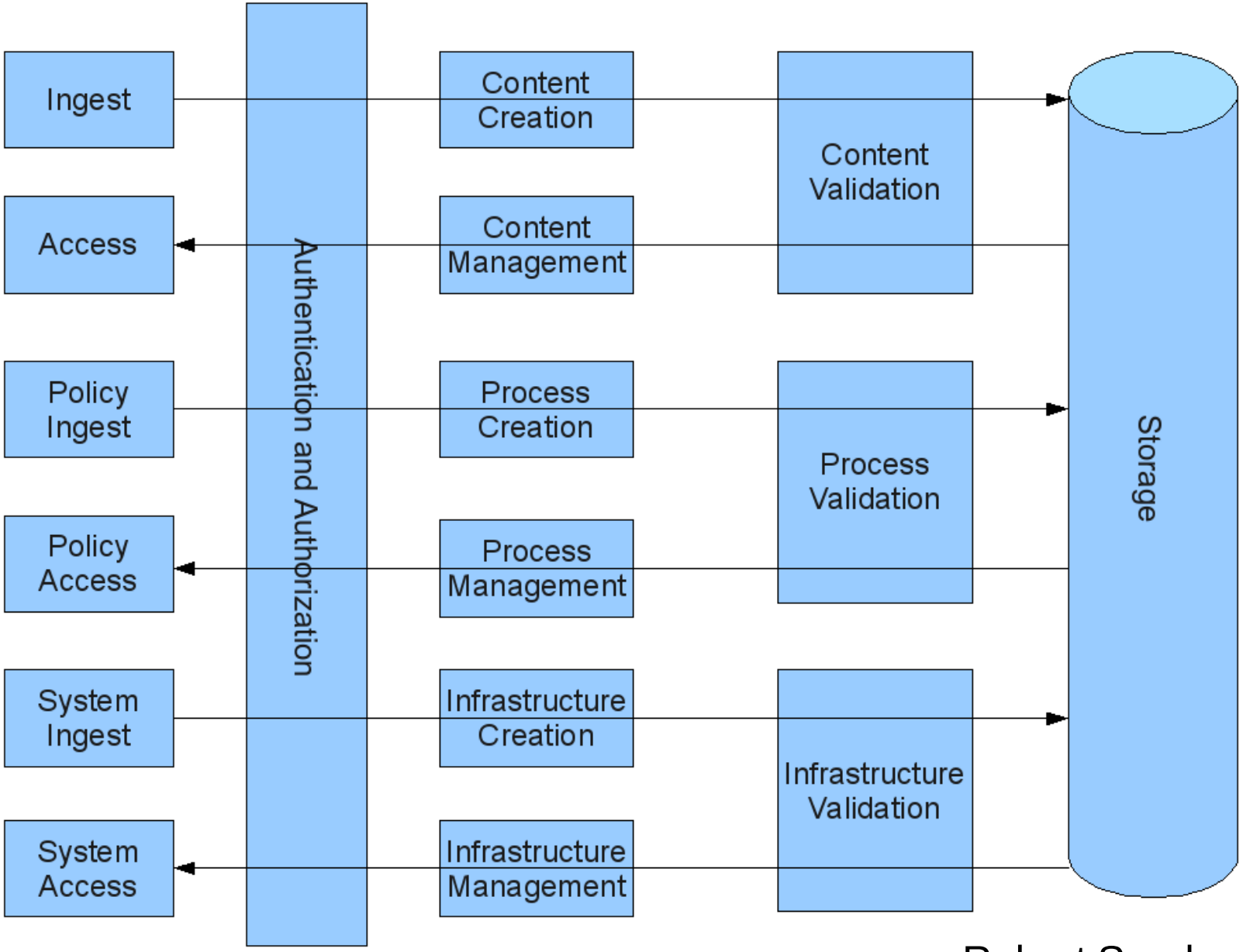
1. Authentication of submitter
2. Upon acceptance load into Virtual Loading Dock
3. Metadata submission template
4. SIP metadata creation
5. Virus checks
6. Authenticate content
7. Identify encryption, compression, other access issues
8. Document chain of custody
9. Document open and restricted records and apply security controls
10. Accept or reject
11. Automatic metadata extraction
12. Run hash verfication
13. Submission and migration metadata management

Examples (cont.)…

14. Verify and confirm archival storage of AIP post ingestion
15. Replicate AIP
16. Run error checks and monitor error logs
17. Run fixity checks
18. Maintain an activity log
19. Monitor file formats requiring migration
20. Document migration process
21. Confirm and apply current policies
22. Identify SLA associated with content
23. Notification of change in access status
24. Create authentic DIP, with the ability to certify the DIP
25. Export DIP to specified format

Implementation: Policies

- Policies controlling administration are stored in iRODS
  - Policies are preserved, and can be ingested in exactly the same way as content
    - We expect most policies to be related to authorization and validation
- All policies controlling presentation are stored in Cheshire database as RDF / XML documents which then get searched by an iRODS rule
  - Policies consist of links to Cheshire workflows which implement the preservation process
- All of the iRODS hooks can be used as triggers for policies
  - Allows policies on content, infrastructure, users



Robert Sanderson

Ingest (Implementation)

- SWORD (Simple Web Service Offering Repository Deposit)
  - Standard for ingesting data into archive
- Web based interface
  - Under development
- iRODS iCommands
  - Support bulk operations

Discovery (Implementation)

- Uses Cheshire digital library system integrated with iRODS
  - Cheshire processing workflows
  - Scalable system (can run on cluster)
  - Cheshire indexes and software both archived
- Discover documents by content and metadata
  - Automating resource discovery across domains and formats
- Interfaces generated from Z39.92 (Zeerex description of search service)

Presentation (Implementation)

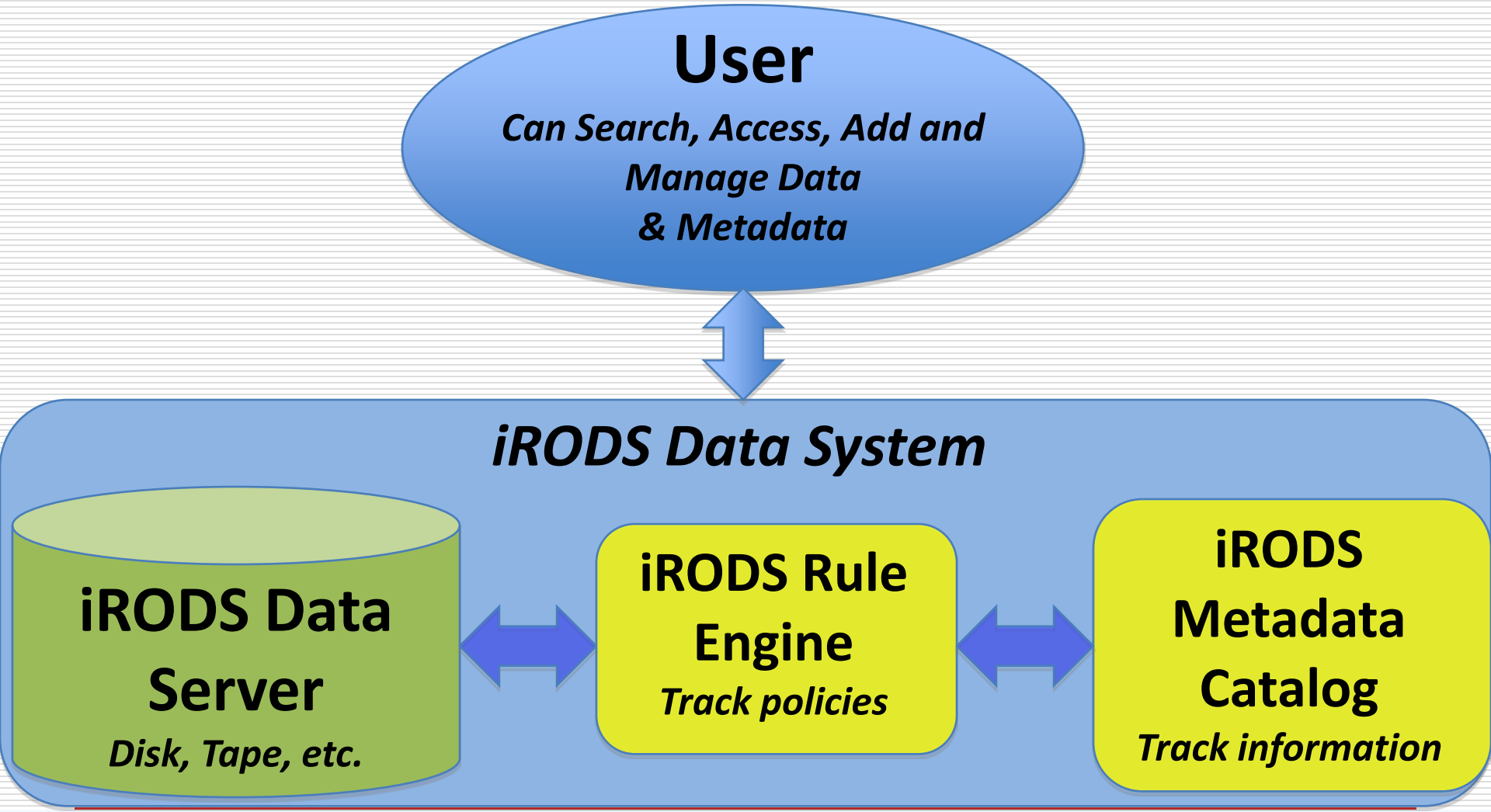
- Multivalent browser technology
  - Supports format parsing in a media adaptor
  - Supports behaviors for manipulating the parsed data
- Extensible
  - Perfect fidelity as opposed to conversion / migration
  - Access to all parts of a document (vs. emulation)
  - Generalizes to any media
  - Cross-format, distributed annotations
- Sustainable
  - Independent of original applications
  - Independent of operating systems and processor and machine class

iRODS Concepts

- Preservation is the extraction of records from the creation environment and ingest into the preservation environment
- Preservation is communication with future systems that use new protocols and standards
- Preservation is management of communication from the past and validation of actions by prior archivists
- Preservation is based on specified policies and procedures, which drive requirements for descriptive and system metadata

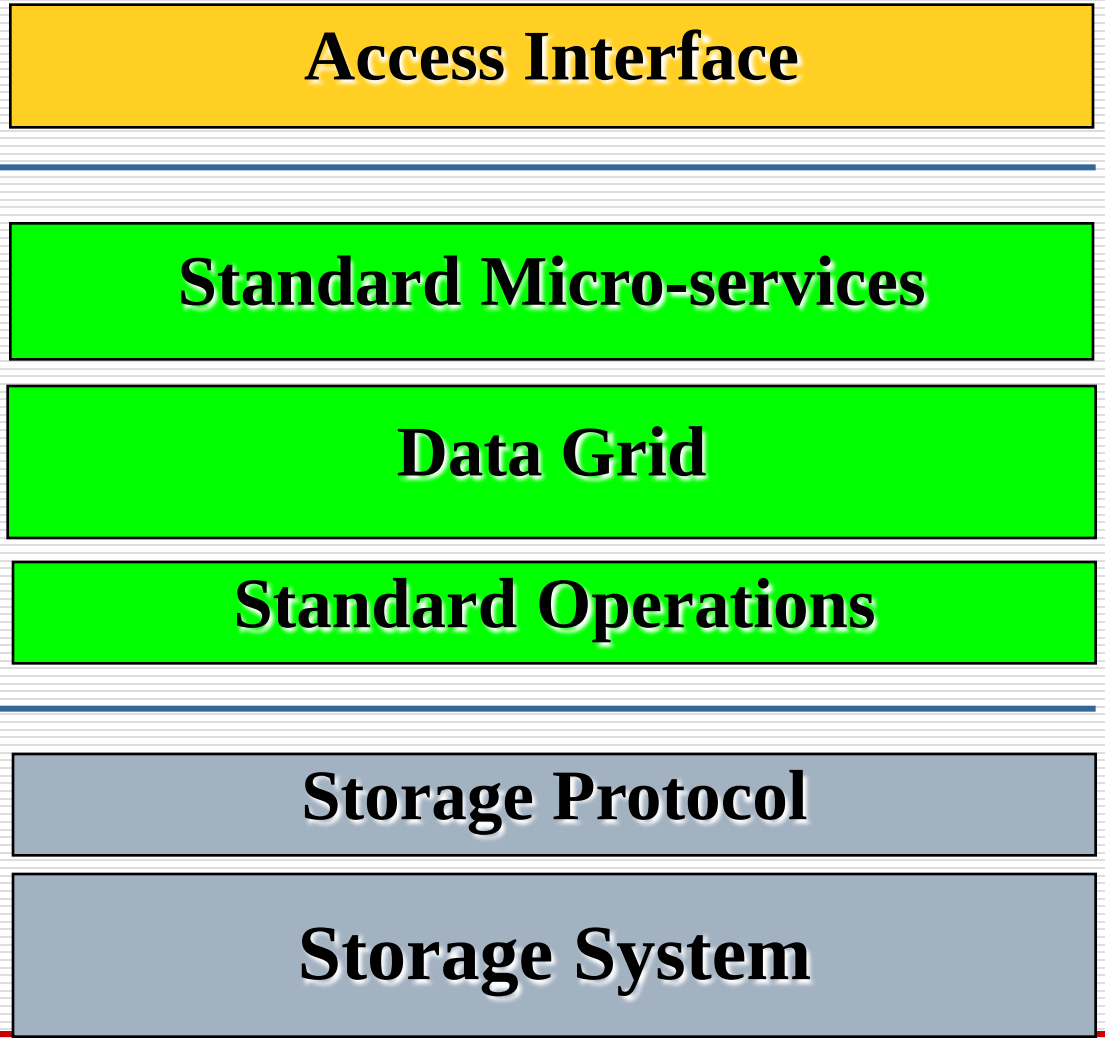
- The preservation environment itself will evolve

Overview of iRODS Data System



*Access data with Web-based Browser or iRODS GUI or Command Line clients.

Data Virtualization



Map from the actions requested by the access method to a standard set of micro-services.

Map the standard micro-services to standard operations.

Map the operations to the protocol supported by the operating system.

iRODS Rules

- Policies implemented as computer actionable rules
  - Rules control the execution of procedures
  - Rule types - Atomic (immediate), Deferred, Periodic
- Procedures implemented as remotely executable workflows
  - Workflows implemented by chaining micro-services together, providing recovery procedures to handle failures
- Each workflow defined by:
  - Event, Condition, Action chains (micro-services, other Rules), Recovery chains

Micro-services

Demo-8

- Function snippets – perform a small, well-defined operation/semantics, e.g.
  - computeChecksum
  - replicateFile
  - integrityCheckGivenCollection
  - zoomImage
  - getSDSSImageCutOut
  - searchPubMed
- Chained to implement iRODS Rules (workflows)
  - Invoked by the distributed iRODS Rule Engine
  - Currently C functions, Python scripts; Java in development
  - Able to execute remote Web-services
- Functional decomposition of client actions

State Information

Demo-9

- The execution of each micro-service generates state information that is stored in the iCAT metadata catalog
  - Example - the mapping from logical file name to physical file location
  - Example - the value of a checksum
- The state information can be queried.
  - Can verify value of state information against expected value as an assessment criterion

ISO MOIMS
repository assessment criteria

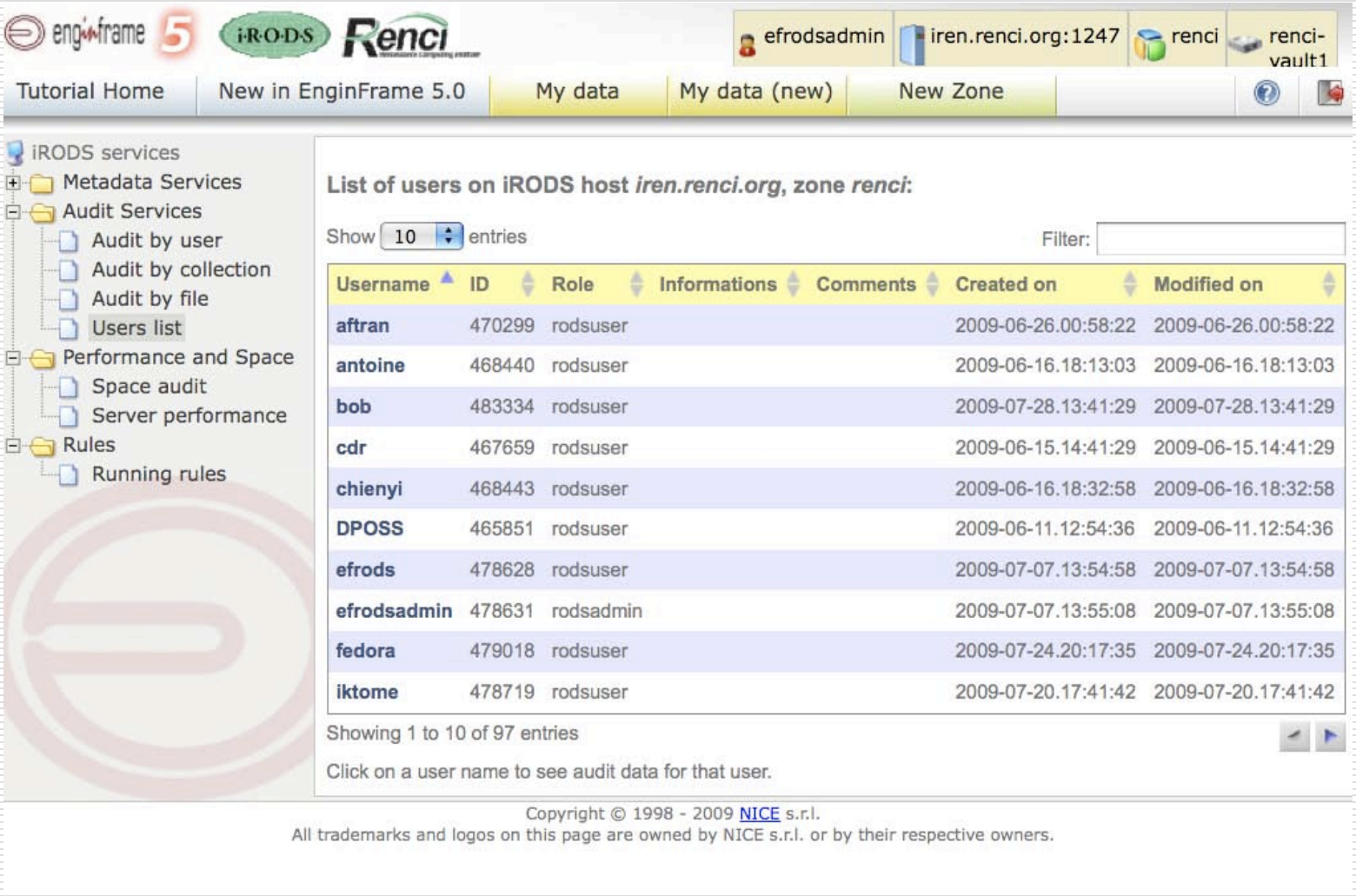
- Are developing 106 rules that implement the ISO assessment criteria

90	Repository has a documented history of the changes to its operations, procedures, software, and hardware
91	***Verify descriptive metadata against semantic term list***
92	***Verify status of metadata catalog backup (create a snapshot of metadata catalog)***
93	***Verify consistency of preservation metadata after hardware change or error***

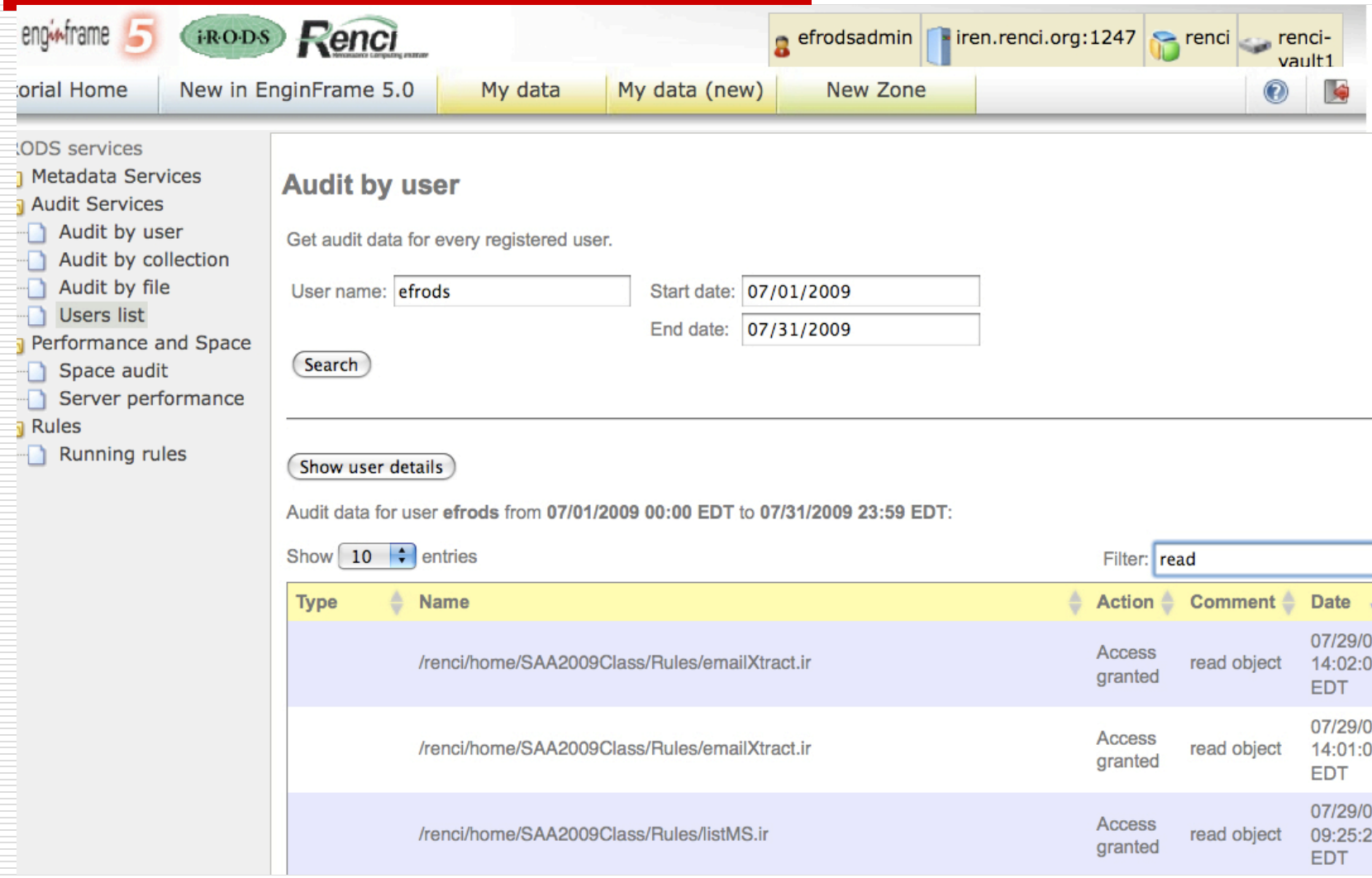
EnginFrame Portal

- Web interface to capabilities provided by iRODS data grid
  - Display files and collections
  - Interactive invocation of iRODS rules
- Parsing of audit trails
  - Audit trails record operations performed within the iRODS data grid
    - Data manipulation
    - Access control changes
    - Attempted access
    - Audit user activities
    - Audit file manipulation
    - Audit collection manipulation
    - Audit actions
- Expression of iRODS actions as web services

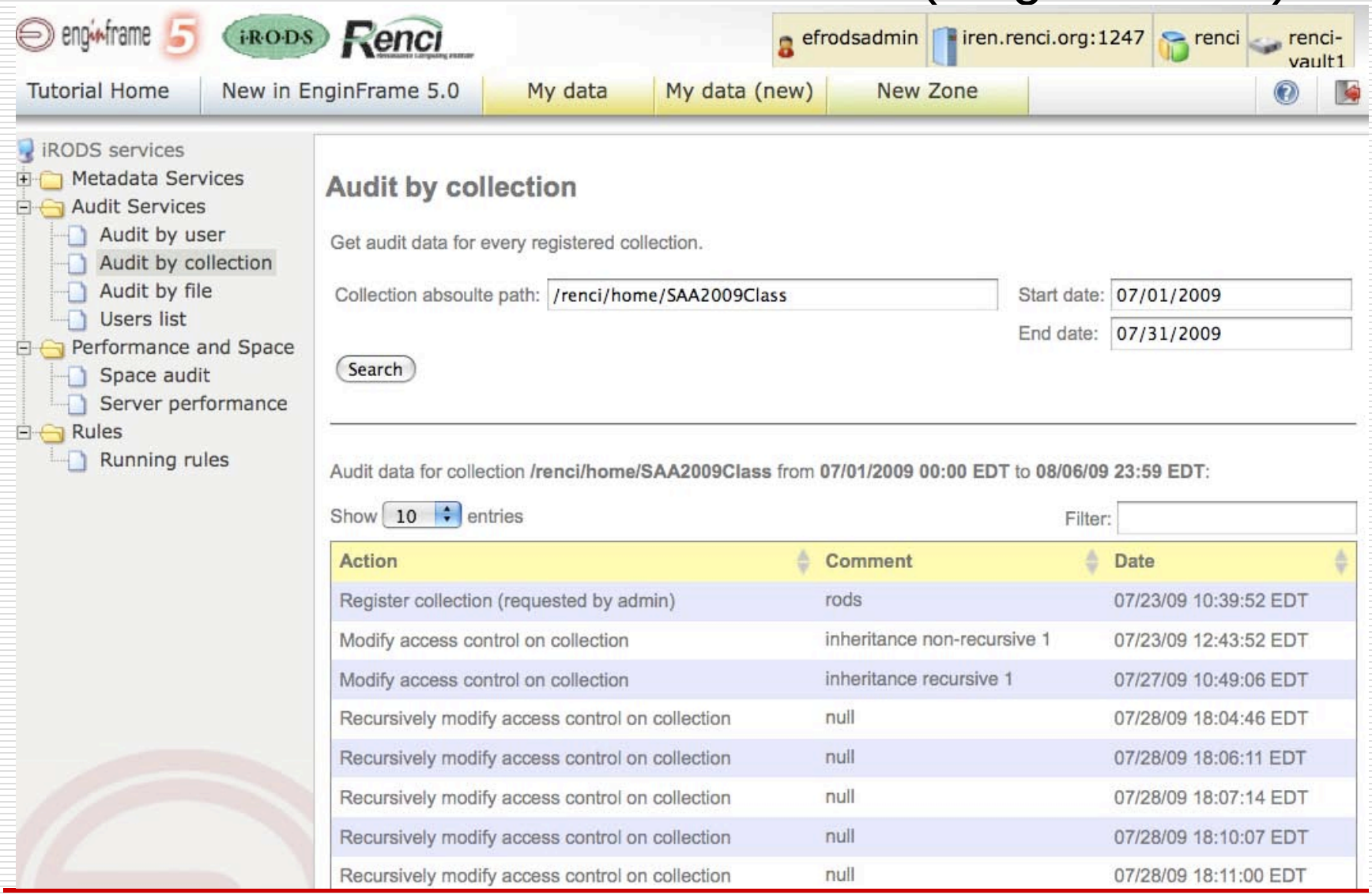
User Listing (EnginFrame)



User-level Audit (EnginFrame)



Collection-level Audit (EnginFrame)



Community Driven Development
iRODS Version 2.1

- Released on July 10, 2009 under BSD open source license
  - Added support for Kerberos authentication (DoD)
  - Added support for mySQL database (SLAC)
  - Created iRODS standard C I/O library (NASA)
  - Preservation micro-services (NARA)
  - 64 policy enforcement points within framework (SHAMAN)
  - Added monitoring system (IN2P3)
  - Web-DAV interface (ARCS)
  - Added compound resource (disk cache/tape archive) (UK)
  - File aggregation (tar file manipulation) (UK)
  - Fedora bulk load interface (CDR/UNC)
  - Virtual Computing Laboratory integration (NCSU/RENCI)
  - 32 bug fixes

Next set of planned extensions

- EnginFrame interface (grid portal) (EU)
- Shibboleth authentication (CDR)
- Query on integer attribute values (EPA)
- Cloud storage interface (OOI)
- Message bus interface for control (OOI)
- Preservation policies (NARA)
- Quotas (IN2P3)
- HPSS archive parallel I/O interface (Teragrid)
- Recursion protection (SHAMAN)
- NetCDF remote filtering (OOI, EPA)

For More Information

Reagan W. Moore
rwmoore@renci.org
http://irods.diceresearch.org

NSF OCI-0848296 "NARA Transcontinental Persistent Archives Prototype"
NSF SDCI-0721400 "Data Grids for Community Driven Applications"