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

🗢 UCSD 👹

UNIVERSITY OF LIVERPOOL

SHAMAN

<u>http://shaman-ip.eu/shaman/</u>

Renaissance Computing Institute

http://www.renci.org

D·I·C·E

i-R-O-D-S

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

i-R-O-D-S

D·I·C·E

- 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

Federation of Seven Independent Data Grids

Demo-14



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

🗢 UCSD 👹

LIVERSITY OF

SHAMAN



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

🗢 UCSD 👹

UNIVERSITY OF LIVERPOOL

- iRODS and EnginFrame
- Policies controlling presentation
 - Multivalent and Cheshire3



SHAMAN Approach



- 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

LIVERSITY

Scalable

D·I·C·E

- Do not have to migrate the entire collection to new formats
- Parsing is done on display
- Living Documents

i-R-O-D-S

- 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

D·I·C·E

- 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

i-R-O-D-S

D·I·C·E

- Parse and filter audit trails
- Support interactive invocation of iRODS rules
 - List management rules
 - List management procedures (micro-services)

UCSD UNIVERSITY OF

DCAPE: Distributed Custodial Archival Preservation Environments

Purpose:

D·I·C·E

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

* UNIVERSITIES:

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

* CULTURAL ENTITIES:

i-R-O-D-S

- Getty Research Institute

* INTERNATIONAL PARTNERS:

- Carleton University (Geomatics and Cartographic Research Centre)

KansasKentucky

- New York



DCAPE Preservation services

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

D·I·C·E

- 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

i-R-O-D-S

D-I-C-E Data Intensive Cyber Environments

- 11. Automatic metadata extraction
- 12. Run hash verfication
- 13. Submission and migration metadata management

LIVERSITY CAROLINA CONCINATION OF LIVERSITY OF

SHAMA

Examples (cont.)...

- 14. Verify and confirm archival storage of AIP post ingestion
- 15. Replicate AIP

i-R-O-D-S

D-I-C-E

- 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

🗢 UCSD 🦉

LIVERSITY OF

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





Sustaining Heritage Access through Multivalen



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

D·I·C·E

i-R-O-D-S

- 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

UNIVERSITY OF LIVERPOOL

¥

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

i-R-O-D-S

D·I·C·E

- Independent of original applications
- Independent of operating systems and processor and machine class

LIVERSITY

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

User

Can Search, Access, Add and Manage Data & Metadata

iRODS Data System



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

Data Virtualization

	Access Interface Standard Micro-services	Map from the actions requested by the access method to a standard set of micro-services.				
	Data Grid	Map the standard micro-services to standard operations.				
	Standard Operations Storage Protocol	Map the operations to the protocol supported				
	Storage System	by the operating system.				
D-L- Data Intensive Cyber	D-I-C-E i-R-O-D-S II THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL CHAPEL HILL					

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:

i-R-O-D-S

D·I·C·E

 Event, Condition, Action chains (micro-services, other Rules), Recovery chains

CSD UNIVERSITY OF

Micro-services

- Demo-8
- Function snippets perform a small, well-defined operation/semantics, e.g.
 - computeChecksum
 - replicateFile
 - integrityCheckGivenCollection
 - zoomImage

i-R-O-D-S

D·I·C·E

- getSDSSImageCutOut
- searchPubMed
- Chained to implement iRODS Rules (workflows)
 - Invoked by the distributed iRODS Rule Engine
 - Currently C functions, Python scripts; Java in development

UNIVERSITY OF LIVERPOOL

- Able to execute remote Web-services
- Functional decomposition of client actions

State Information

- 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

CSD UNIVERSITY OF



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

AROLINA CONCOL UNIVERSITY OF LIVERPOOL SHAMAN



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

i-R-O-D-S

D-I-C-E

Expression of iRODS actions as web services

of NORTH CAROLINA CONSCIENT OF UNIVERSITY OF UNIVERSITY

SHAMAN

User Listing (EnginFrame)

Tutorial Home New in E	nginFrame 5.0		My data	My data (ne	ew) New	Zone	0
IRODS services Handbard Services Handbard Services Handbard Services Handbard Services	List of users on iRODS host <i>iren.renci.org</i> , zone <i>renci</i> : Show 10 • entries Filter:						
Audit by collection	Username 🔺	ID 🝦	Role 🍦	Informations 🝦	Comments 🍦	Created on	Modified on
Users list	aftran	470299	rodsuser			2009-06-26.00:58:	22 2009-06-26.00:58:
Performance and Space	antoine	468440	rodsuser			2009-06-16.18:13:	:03 2009-06-16.18:13:
Space audit	bob	483334	rodsuser			2009-07-28.13:41:	29 2009-07-28.13:41:
	cdr	467659	rodsuser			2009-06-15.14:41:	:29 2009-06-15.14:41:
Running rules	chienyi	468443	rodsuser			2009-06-16.18:32:	:58 2009-06-16.18:32:
	DPOSS	465851	rodsuser			2009-06-11.12:54:	36 2009-06-11.12:54:
	efrods	478628	rodsuser			2009-07-07.13:54:	:58 2009-07-07.13:54:
	efrodsadmin	478631	rodsadmin			2009-07-07.13:55:	:08 2009-07-07.13:55:
	fedora	479018	rodsuser			2009-07-24.20:17:	:35 2009-07-24.20:17:
	iktome	478719	rodsuser			2009-07-20.17:41:	:42 2009-07-20.17:41:

Copyright © 1998 - 2009 NICE s.r.l.

ATIONA

All trademarks and logos on this page are owned by NICE s.r.l. or by their respective owners.



User-level Audit (EnginFrame)

enginframe 5 (iroos	Renci gefrodsadmin giren.re	enci.org:1247 🎧 renci- vault1
orial Home New in Er	nginFrame 5.0 My data My data (new) New Zone	 Image: Second sec
 ODS services Metadata Services Audit Services Audit by user Audit by collection Audit by file Users list Performance and Space Space audit Server performance Rules Running rules 	Audit by user Get audit data for every registered user. User name: efrods Start date: 07/01/2009 End date: 07/31/2009 Search Show user details Audit data for user efrods from 07/01/2009 00:00 EDT to 07/31/2009 23:59 EDT:	
	Show 10 + entries	Filter: read
	Type 🝦 Name	🔷 Action 🔷 Comment 🌲 Date 💡
	/renci/home/SAA2009Class/Rules/emailXtract.ir	Access read object 07/29/09 granted EDT
	/renci/home/SAA2009Class/Rules/emailXtract.ir	Access read object 07/29/09 granted EDT
	/renci/home/SAA2009Class/Rules/listMS.ir	Access read object 07/29/09 granted EDT

FL HILL

CAROLINA COUCSD UNIVERSITY OF UNIVERSITY OF UNIVERSITY OF USHAMAN

ATIONA



	Renci	8 efro	odsadmin iren.renci.	org:1	247 👸 renci 🥪
Tutorial Home New in E	EnginFrame 5.0 My data My data	(new)	New Zone		C
 iRODS services Metadata Services Audit Services Audit by user Audit by collection Audit by file Users list Performance and Space Space audit Server performance Rules 	Audit by collection Get audit data for every registered collection. Collection absoulte path: /renci/home/SAA2009C	lass	Start End o	date: date:	07/01/2009 07/31/2009
Kunning fules	Audit data for collection /renci/home/SAA2009Class	ss from 07	7/01/2009 00:00 EDT to 08/	/06/09 Filter:	23:59 EDT:
	Action	\$ C	Comment	\$	Date
	Register collection (requested by admin)	r	ods		07/23/09 10:39:52 ED
	Modify access control on collection	ir	nheritance non-recursive 1		07/23/09 12:43:52 ED
	Modify access control on collection	ir	nheritance recursive 1		07/27/09 10:49:06 ED
	Recursively modify access control on collection	n	ull		07/28/09 18:04:46 ED
	Recursively modify access control on collection	n	null		07/28/09 18:06:11 ED
	Recursively modify access control on collection	n	ull		07/28/09 18:07:14 ED
	Recursively modify access control on collection	n	ull		07/28/09 18:10:07 ED

the UNIVERSITY of North Carolina at CHAPEL HILL ARCHIVES.00V

Community Driven Development iRODS Version 2.1

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/RENC

NA CSD UNIVERSITY OF USHAMAN

Released on July 10, 2009 under BSD open source

• 32 bug fixes

D-I-C-E Data Intensive Cyber Environments i-R-O-D-S

Next set of planned extensions

- EnginFrame interface (grid portal)
- Shibboleth authentication
- Query on integer attribute values
- Cloud storage interface
- Message bus interface for control
- Preservation policies
- Quotas

D·I·C·E

- HPSS archive parallel I/O interface
- Recursion protection

i-R-O-D-S

NetCDF remote filtering

(EU) (CDR) (EPA) (OOI)(OOI)(NARA) (IN2P3) (Teragrid) (SHAMAN) (OOI, EPA)

UNIVERSITY OF LIVERPOOL

For More Information

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

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

