Distributed Custodial Frameworks for Archival Preservation

Amy Rudersdorf

Director, Digital Information Management Program

State Library of North Carolina

Two Questions:

- Why don’t *real* preservation solutions exist after all the time & resources that have been spent by a lot of really smart people and powerful institutions?

- Why don’t some (any?) of us have established, trusted digital preservation programs (mature policies, institutional support, and stable repositories)?

Answers (sort of)

- Because the challenges are huge and the resources are small
  - Most institutions will never have the resources needed to build and manage in-house digital preservation programs
  - The objects we are trying to preserve are constantly changing and increasing in complexity
  - Every repository cannot hire experts in preservation technologies
  - The solutions we develop today will not work forever

Solution [?]

- A trusted, sustainable preservation service that repositories of all types and sizes could employ to support their digital preservation activities and responsibilities.
- Ideally, it would:
  - be distributed
  - be custodial
  - include preservation actions that answer the needs of both libraries and archives

Introducing . . .

- **Distributed Custodial Archival Preservation Environments (DCAPE)**
  - *Main Goal: to build a distributed production preservation environment that meets the needs of digital repositories for trusted archival preservation services.*
  - Grant funded by NHPRC in 2008   (RE10010-08)
  - Started in December 2008, will run for 2 ½ years
  - Over 30 individuals at 10 institutions, including 4 staff at UNC
  - http://dcape.org

DCAPE Partners

- Cultural Entity: Getty Research Institute
- Cyberinfrastructure: West Virginia University, Carleton University (Canada)
- State Archives: California, Kansas, Michigan, Kentucky, North Carolina, New York
- State Library: North Carolina
- University Archives: Tufts
- UNC: Renaissance Computing Institute (RENCI) and School of Information and Library Science (SILS)

Grant PI & Development Team

- The Center for **Data Intensive Cyber Environments** at the University of North Carolina at Chapel Hill
  - Richard Marciano, Reagan Moore, et al.
  - Develops and manages iRODS
  - “Advanced open source technologies for complete life cycle managing, sharing, and preserving of digital data”

Let’s break down the main goal:

- **Distributed**: Physical custody of collections is hosted outside of the originating repository by a trusted preservation service
- **Custodial**: Originating repository retains legal custody
- **Archival Preservation**:
  - Originating repository remains responsible for archival functions, including preservation and access activities
  - Access to collections is controlled by the originating repository
  - Trusted preservation service provides originating repository with a complete audit trail for all items in hosted collections

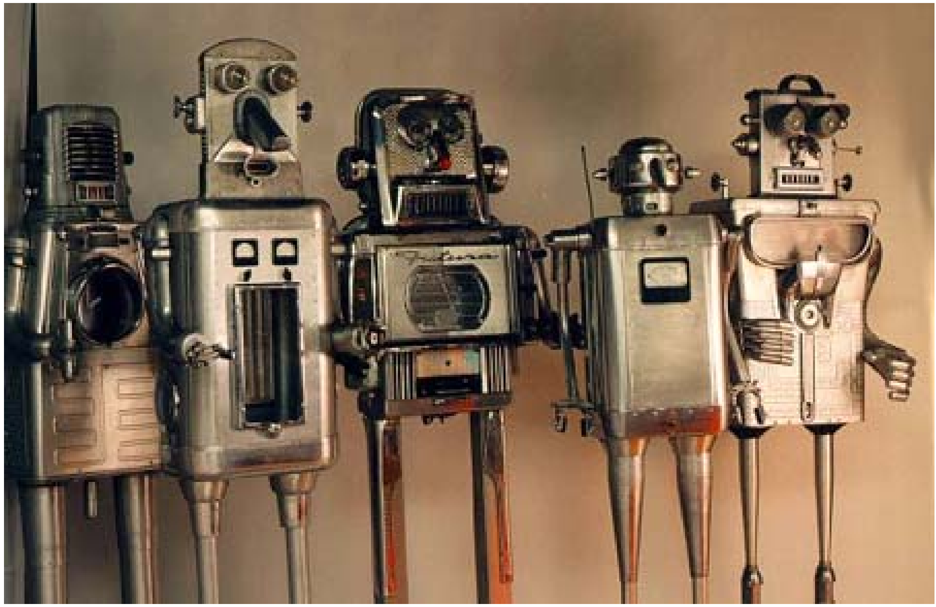
Let’s break down the main goal:

- “Trusted” = TRAC & OAIS compliant
- Services are based on policies (“rules”) defined by the user
    - A series of rules might “look” like this:

      “When my files are ingested, replicate them in three different locations and run a checksum on each file. Bit-check files every month until I say otherwise. Alert me to any changes.”

Other project goals #1/3

- The software infrastructure will automate many of the administrative tasks associated with the management of digital repositories.
  - Examples of automated tasks:
    - Authentication, replication, migration, obsolete file management, preservation metadata management



Other project goals #2/3

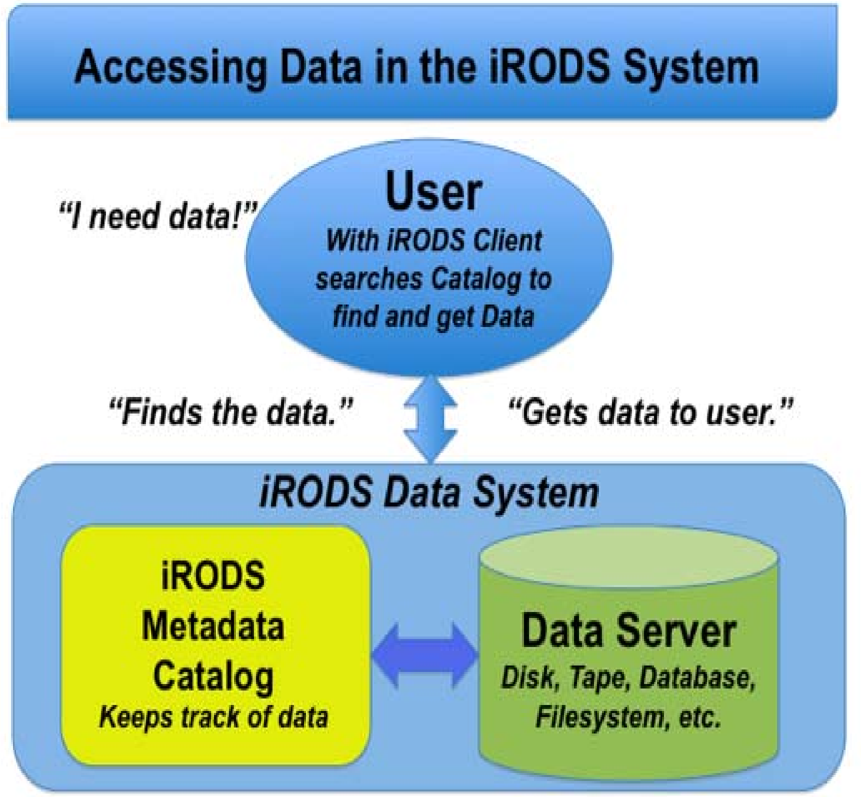
- The preservation service will reduce the need for repositories to build their own digital preservation systems in-house.
  - This is especially appealing to small institutions or institutions with little IT or administrative support

Other project goals #3/3

- A business model will be developed to sustain the preservation service over time.
  - At some point, down the road this will be established, so. . .*something* may cost *something*. . .but none of that has been defined yet.

iRODS introduction

- “i Rule Oriented Data Systems”
- Preservation environment that provides rules-based automation of archival and preservation functions (basically, repeatable policy-based services)
- Standard and optional services will be available


Accessing Data in the iRODS System
“I need data!”
User
With iRODS Client searches Catalog to find and get Data
“Finds the data.”
“Gets data to user.”
iRODS Data System
iRODS Metadata Catalog
Keeps track of data
Data Server
Disk, Tape, Database, Filesystem, etc.

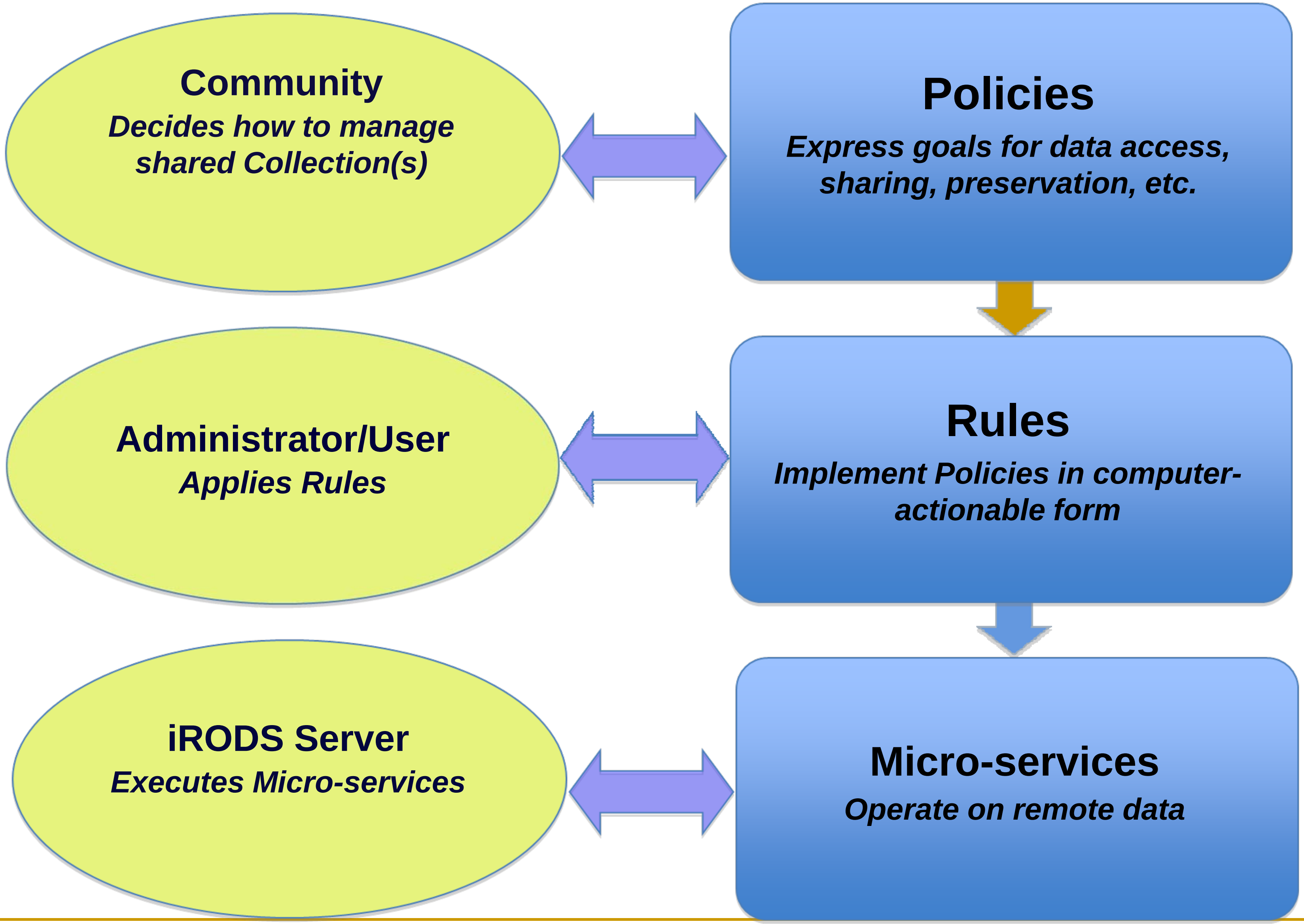
Users can search for, access, add/extract metadata, annotate, analyze & process, replicate, copy, share data, manage & track access, subscribe, and more.

iRODS introduction

- Associate the rule-based (policies-driven) data management system to combine:
  - Data Objects
  - Collections
  - User Groups
  - Storage Systems

  - For Example: *A particular group might ingest a particular collection, and another group might access a subset of that collection from another location.*

"Layers" in iRODS: From Users to Storage



Community

Decides how to manage shared Collection(s)

Policies

Express goals for data access, sharing, preservation, etc.

Administrator/User

Applies Rules

Rules

Implement Policies in computer-actionable form

iRODS Server

Executes Micro-services

Micro-services

Operate on remote data

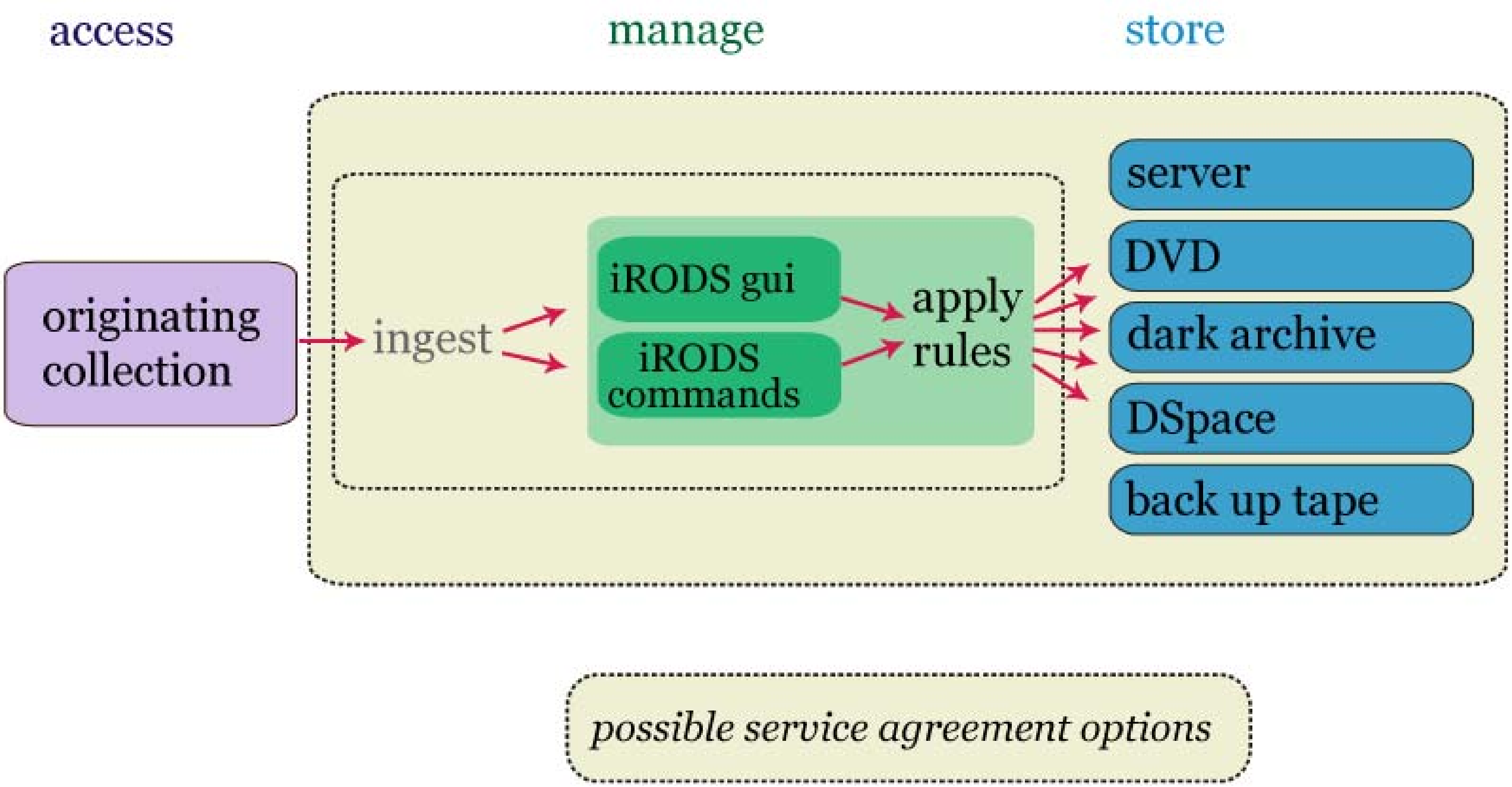
iRODS intro: Policies in Action!

- Originating Institution specifies policies
  - e.g., *“Make X Copies of Accessioned Records”*
- Break Policies Down into Rules
  - e.g., *“Put one copy at Rocket Center”* [and] *“Put one copy at UCSD”* [and] *“Verify Copies are Identical”*
- Break Rules Down into Micro-Services
  - e.g., *“Put one copy at Rocket Center.”*
    - Read File --> Copy File --> Create Checksum --> Copy Checksum --> etc.
- Micro-Services Can Be Combined into Complex Workflows
  - Execute: periodically, on-demand, delayed start, anywhere on the network

iRODS introduction

- Shared service should reduce costs for individual repositories compared to the cost of building and maintaining in-house preservation capabilities
- Provides hooks to existing CMSs, DAMs, and repositories
- Acts as “middleware” or as a back-end system
- https://www.irods.org

DCAPE & iRODS: What a team!



DCAPE Tasks (Underway)

- Execute service agreements between UNC and existing partners to govern use of test collections.
- Define and implement rules (defined by partners) and services (based on OAIS framework) for iRODS to perform on test collections.
- Ingest test collections into iRODS and validate rules and services.
- Develop business model (including costs) for sustaining a repository service based on iRODS.
- Develop model service agreements that define standard and optional services of the repository.

DCAPE Tasks (Future)

SHAMAN
Sustaining Heritage Access through Multivalent ArchiviNg

- DCAPE/DICE team involved in SHAMAN project grant
  - Enable systems to render back files without interacting with the original environment.
  - Driver driven
  - Basically, emulation without the hard and software.
  - Will be added to iRODS (and thus DCAPE) when it is “stable.”

DCAPE is “More”

- More than a storage service or environment . . .
- More than a reference tool . . .
- DCAPE will provide the capability for all sorts of digital repositories to fulfill their responsibility to preserve . . .

DCAPE

The Obligatory "Questions?" Slide