iRODS - Second Generation Data Management Software or DataGrids 2.0

Mark Conrad, Archives Specialist, Electronic Records Archives, NARA

Richard Marciano, Director, Sustainable Archives & Library Technologies, DICE group, UCSD

What are Data Grids?

Data Grids are “middleware services”

- Software that sits between applications and data sources

So, What?

What are Data Grids Good For?

Data Grids allow you to access data:

- In any format
  - Files, databases, streams, web, programs,…
  - Documents, images, data, sensor packets, tables,…
- Stored in any type of storage system
  - File Systems, tape silos, object ring buffers, sensor streams,…
- Stored anywhere over a wide area network
  - Across organizational, administrative and security boundaries
- ***Without having to know the system addresses, paths, protocols, commands, etc. needed to retrieve it!***

What are Data Grids Good For?

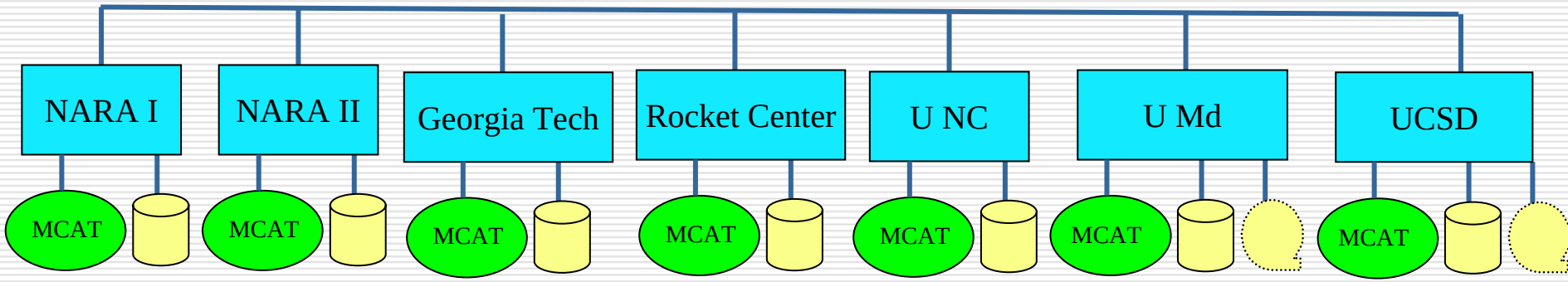
- Scalability
  - Millions of files
  - Petabytes of Data
- Evolvability
  - Infrastructure Independence
  - Across Generations of Software
- Extensibility
  - Deal with Technologies not yet Dreamed of

What are Data Grids Good For?

- Collections Managed by the DICE Group:
  - 1+PetaBytes, 170+ Million files
  - Multi-disciplinary Scientific Data
    - Astronomy, Cosmology
    - Neuro Science, Cell-Signalling & other Bio-medical Informatics
    - Environmental & Ecological Data
    - Educational (web) & Research Data (Chem, Phys,…)
    - Earthquake Data, Seismic Simulations
    - Real-time Sensor Data
  - Growing at 1TB a day
  - Supporting large projects: TPAP, TeraGrid, NVO, SCEC, SEEK/Kepler, GEON, ROADNet, JCSG, AfCS, SIO Explorer, SALK, PAT …

What are Data Grids Good For?

- TPAP - NARA Transcontinental Persistent Archive Prototype
  - Federation of Seven Independent Data Grids



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

Ten Years of Data Grid 1.0 - What's Missing

- Automatic Policy Execution
  - Increasing Scale
  - Managing System Administration
  - Visualization
  - Virtualization
  - Customization

Data Grids 2.0 – Policies in Action!

Specify Policies

- "Make X Copies of Accessioned Records"

Break Policies Down into Rules

- "Put one copy at Rocket Center"
- "Put one copy at UCSD"
- "Verify Copies are Identical"

Break Rules Down into Micro-Services

- "Put one copy at Rocket Center."
  - Read File
  - Copy File
  - Create Checksum
  - Copy Checksum
  - Etc.

Micro-Services Can Be Combined into Complex Workflows

- Execute them: Periodically, On-demand, Delayed Start, Anywhere on the network

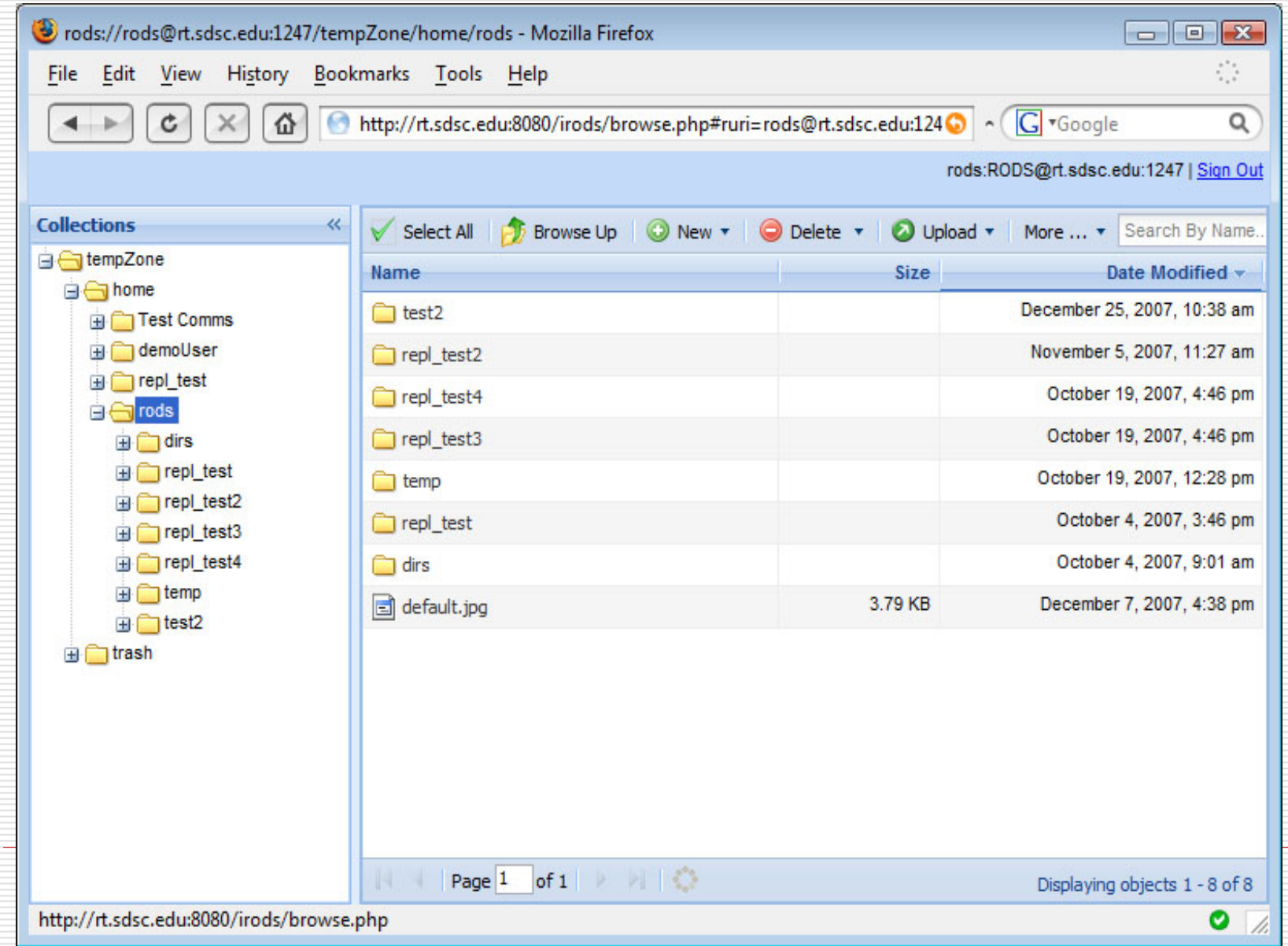
Rule-based Data Management

- Associate Rules with Combinations of:
  - Data Objects
  - Collections
  - User Groups
  - Storage Systems
- For Example:
  - Particular User Groups when Accessing a Particular Collection

Demo Time!

- If the Internet Pixies Cooperate…
- Otherwise We Have More Slides!

iRODS Browser



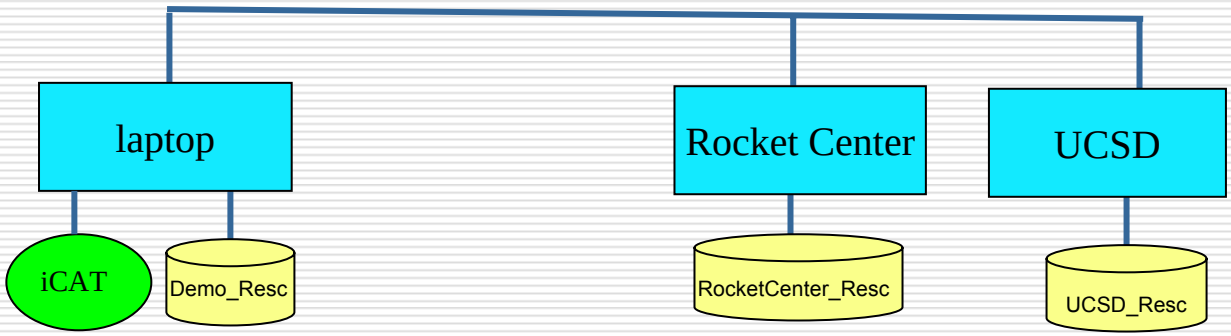
Environment Used for the Demos

Resources:

local laptop: demo_Resc

Rocket Center: RocketCenter_Resc

UCSD: UCSD_Resc



Collections: under /tempZone/home/National_Archives

RG064-NARA	**NARA:**	*Records_of_the_National_Archives_and_Records_Administration*
RG266-SEC	**SEC:**	*Records_of_the_Securities_and_Exchange_Commission*
RG255-NASA	**NASA:**	*Records_of_the_National_Aeronautics_and_Space_Administration*
RG059-DS	**DS:**	*General_Records_of_the_Department_of_State*

Preview of Demos

- Rules w. Collections:
  - Make X Copies of Each Record on Accession and Distribute them Geographically
  - Make X Copies After a Specified Delay
- Rules w. Data Objects:
  - Verify that the Copies are Identical
  - Verify that the File has not been Corrupted
  - Extract Metadata
- Rules w. Collections:
  - Automated Reference

Rules w. Collections

- Make X Copies of Each Record on Accession and Distribute them Geographically
  - Upon a “put” operation into collection:

    “RG064-NARA” on the Rocket Center Resource

    …automatically put a copy into collection:

    “RG064-NARA” on the UCSD Resource

$ ls

emailrule.ir email.tag sample.email test.txt

$ ipwd

/tempZone/home/National_Archives:

$ ils

/tempZone/home/National_Archives:

C- /tempZone/home/National_Archives/RG064-NARA
C- /tempZone/home/National_Archives/RG266-SEC
C- /tempZone/home/National_Archives/RG255-NASA
C- /tempZone/home/National_Archives/RG059-DS

$ ils RG064-NARA

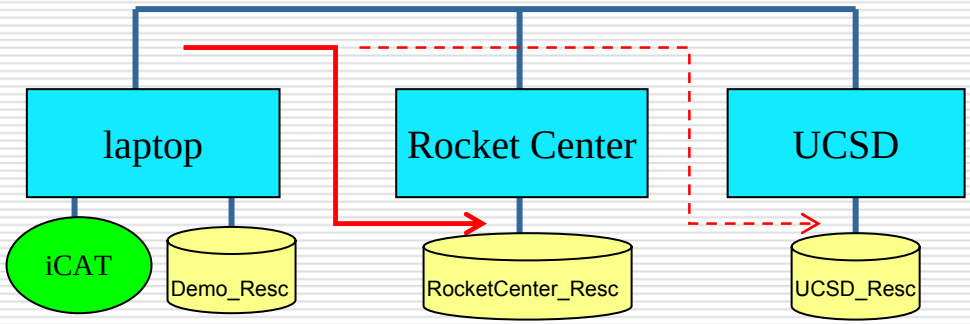
/tempZone/home/National_Archives/RG064-NARA:

$ iput –R RocketCenter_Resc test.txt RG064-NARA

$ ils RG064-NARA

/tempZone/home/National_Archives/RG064-NARA:

test.txt



$ ils -l RG064-NARA

/tempZone/home/National_Archives/RG064-NARA:

rods 0 RocketCenter_Resc 30 2008-06-04.22:48 & test.txt
rods 1 UCSD_Resc 30 2008-06-04.22:48 & test.txt

Rules w. Collections & Delayed Start

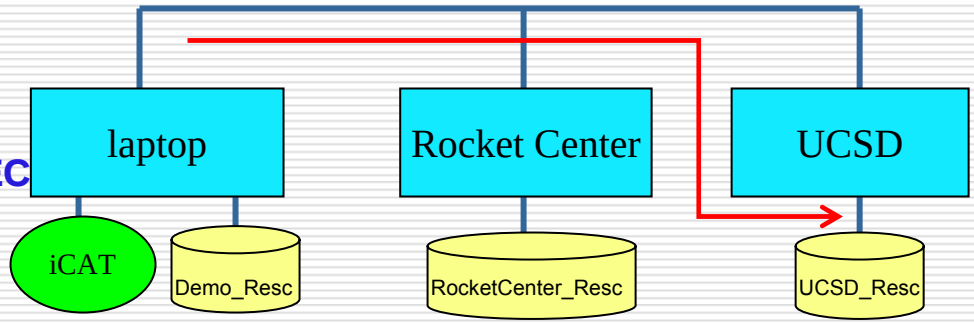
- Make X Copies After a Specified Delay
  - Upon a “put” operation into collection:

    “RG266-SEC” on the UCSD Resource

    ...automatically put a copy into the *“RG266-SEC” collection* on the Rocket Center Resource

  - Delay this replication by 1 minute

```
$ ils
/tempZone/home/National_Archives:
        C- /tempZone/home/National_Archives/RG064-NARA
        C- /tempZone/home/National_Archives/RG266-SEC
        C- /tempZone/home/National_Archives/RG255-NASA
        C- /tempZone/home/National_Archives/RG059-DS
$ ils RG266-SEC
/tempZone/home/National_Archives/RG266-SEC:
$ date
Wed Jun  4 22:49:09 PDT 2008
$ iput –R UCSD_Resc  test.txt   RG266-SEC
$ ils RG266-SEC
/tempZone/home/National_Archives/RG266-SEC:
  test.txt
$ ils -l RG266-SEC
/tempZone/home/National_Archives/RG266-SEC:
  rods      0 UCSD_Resc              30 2008-06-04.22:49 & test.txt
$ iqstat
id     name
10159 msiSysReplDataObj(RocketCenter_Resc,null)
$ date
Wed Jun  4 22:51:38 PDT 2008
```



$ iqstat

No delayed rules pending for user rods

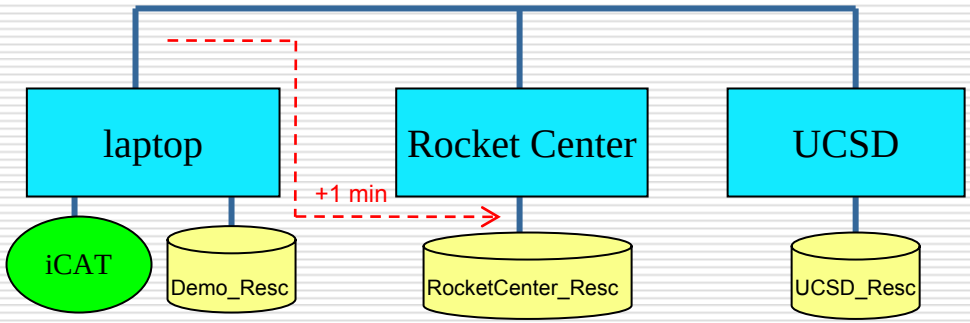
$ date

Wed Jun 4 22:51:44 PDT 2008

$ ils RG266-SEC

/tempZone/home/National_Archives/RG266-SEC:

  test.txt



$ ils -l RG266-SEC

/tempZone/home/National_Archives/RG266-SEC:

rods	0 UCSD_Resc	30 2008-06-04.22:49 & test.txt
rods	1 RocketCenter_Resc	30 2008-06-04.22:51 & test.txt

Rules w. Data Objects

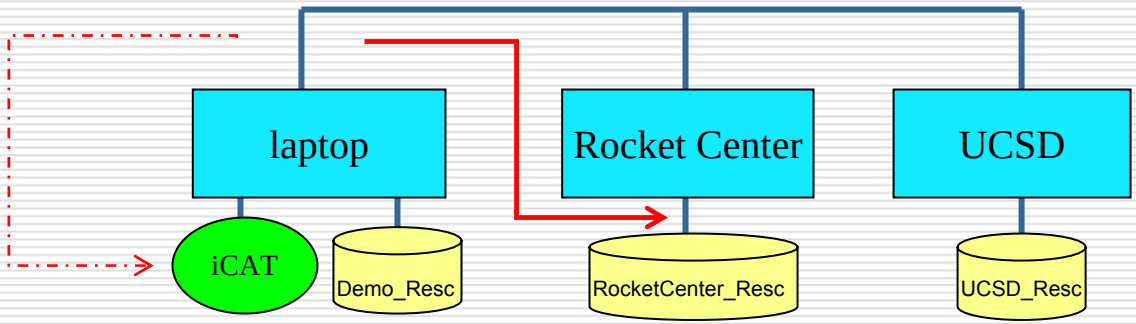
- Verify that the Copies are Identical
  - create a checksum on the client side
  - transfer the data onto the server
  - read the server copy and compute the checksum
  - verify that both checksums match
  - register the checksum in the iCAT
- Verify that the file has not been corrupted at regular points in the lifetime of the file

$ iput -vK –R RocketCenter_Resc test.txt RG255-NASA

test.txt 0.000 MB | 0.012 sec | 0 thr | 0.002 MB/s

$ ichksum -K RG255-NASA/test.txt

test.txt 5c90eb3be2957f48a43cf9e4db8ad600



Rules w. Data Objects

- Extracting Metadata
  - Apply a tag onto a data file in collection:

    RG255-NASA

  - Extract header information from the data file

  - Register the header information as metadata in iCAT associated with the data file

$ ils RG255-NASA

/tempZone/home/National_Archives/RG255-NASA:

$ ls

emailrule.ir email.tag sample.email test.txt

$ cat sample.email

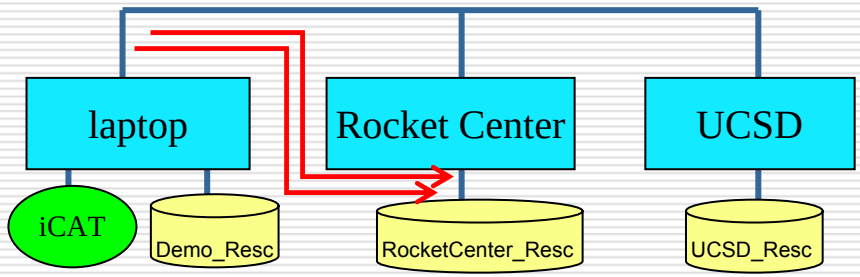
Date: Wed, 4 Jun 2008 07:37 AM
From: Charlotte Walters <cwalters@unm.edu>
To: mark.conrad@nara.gov, marciano@sdsc.edu
Subject: DigIn conference
Mark & Richard,
Just got a request for your bios...Please send them ASAP OR I will make up something very juicy!
See you soon!
Thanks,
Charlotte
Charlotte A. Walters
Political Archives
CSWR/Special Collections
University Libraries
MSC05 3020
University of New Mexico
Albuquerque, NM 87131-0001
505.277.3279

$ cat email.tag

<PRETAG>X-Mailer: </PRETAG>Mailer User<POSTTAG>
</POSTTAG>
<PRETAG>Date: </PRETAG>Sent Date<POSTTAG>
</POSTTAG>
<PRETAG>From: </PRETAG>Sender<POSTTAG>
</POSTTAG>
<PRETAG>To: </PRETAG>Primary Recipient<POSTTAG>
</POSTTAG>
<PRETAG>Cc: </PRETAG>Other Recipient<POSTTAG>
</POSTTAG>
<PRETAG>Subject: </PRETAG>Subject<POSTTAG>
</POSTTAG>
<PRETAG>Content-Type: </PRETAG>Content Type<POSTTAG>
</POSTTAG>

$ iput –R RocketCenter_Resc sample.email RG255-NASA
$ iput –R RocketCenter_Resc email.tag RG255-NASA

$ ils RG255-NASA
/tempZone/home/National_Archives/RG255-NASA:
  email.tag
  sample.email

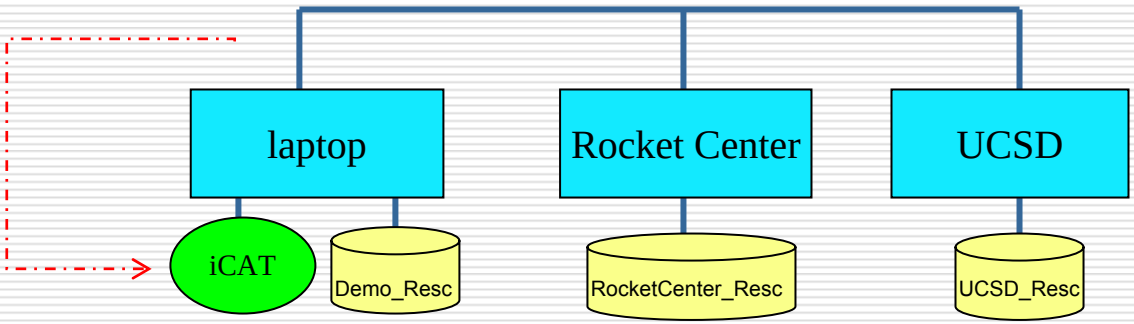


$ imeta ls -d RG255-NASA/sample.email

AVUs defined for dataObj RG255-NASA/sample.email:

None

$ irule -F emailrule.ir



$ imeta ls -d RG255-NASA/sample.email

AVUs defined for dataObj RG255-NASA/sample.email:
---
attribute: **Primary Recipient**
value: mark.conrad@nara.gov, marciano@sdsc.edu
units:
---
attribute: **Sender**
value: Charlotte Walters <cwalters@unm.edu>
units:
---
attribute: **Sent Date**
value: Wed, 4 Jun 2008 07:37 AM
units:
---
attribute: **Subject**
value: DigIn conference
units:

Rules w. Collections & Notification

- Automated Reference

  Upon a “put” operation into collection:

  ”RG059-DS”

  ...send mail to:

  “*marciano@diceresearch.org*”

$ ils

/tempZone/home/National_Archives:

C- /tempZone/home/National_Archives/RG064-NARA
C- /tempZone/home/National_Archives/RG266-SEC
C- /tempZone/home/National_Archives/RG255-NASA
C- /tempZone/home/National_Archives/RG059-DS

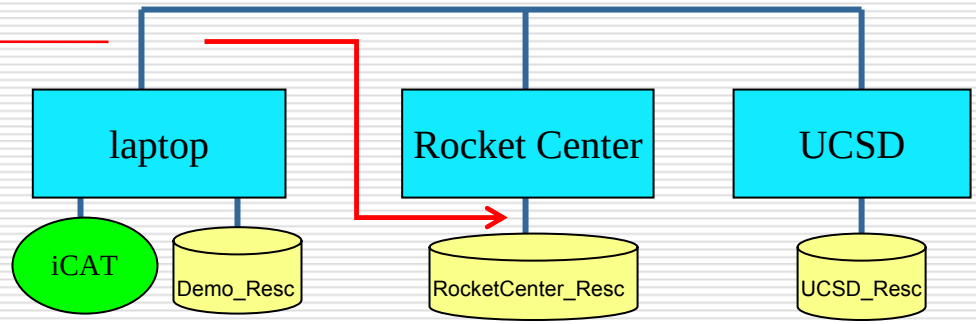
$ iput –R RocketCenter_Resc test.txt RG059-DS

$ ils RG059-DS

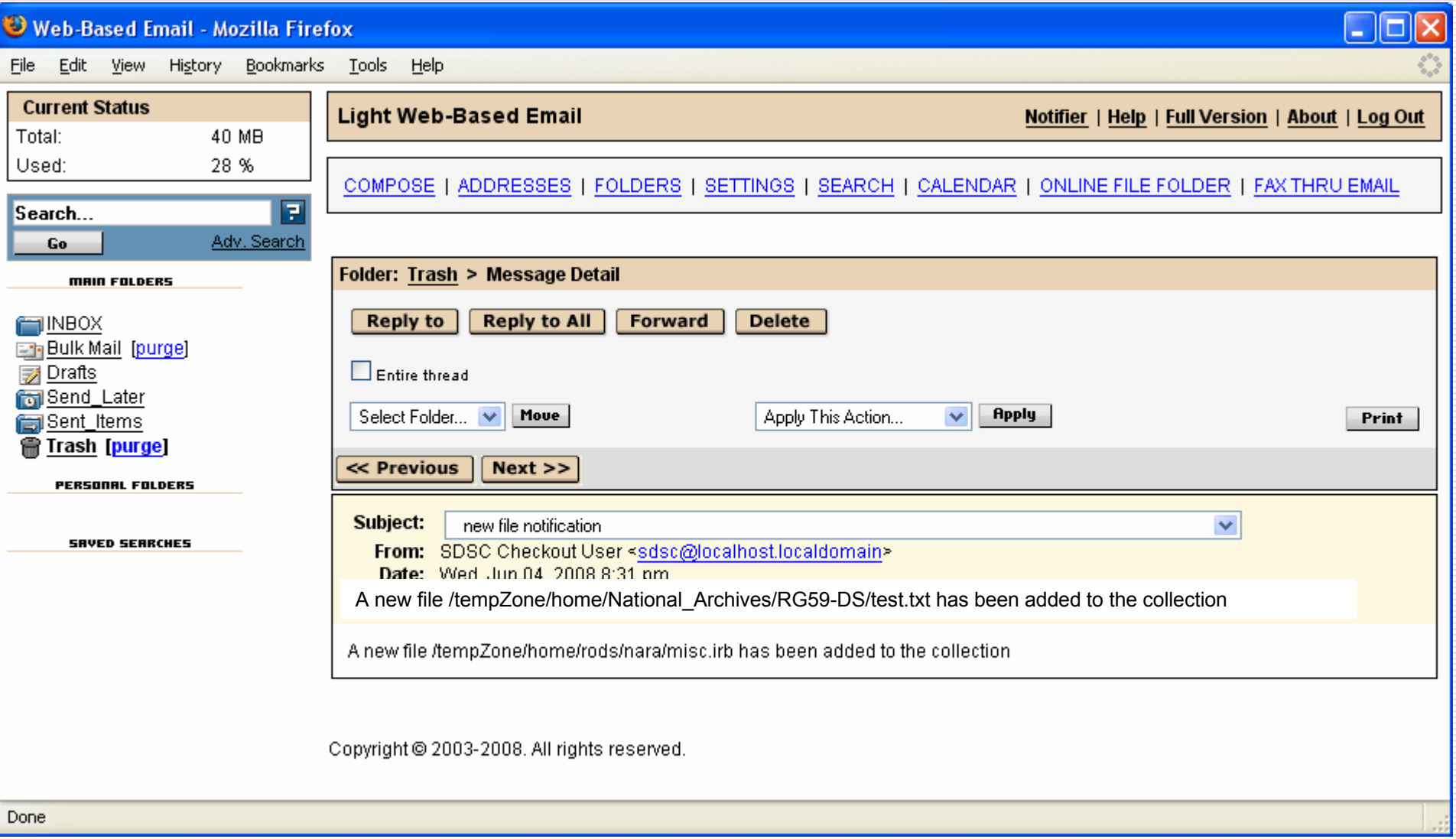
/tempZone/home/National_Archives/RG059-DS:

test.txt

e-mail
marciano@diceresearch.org



Notification of Collection Accretion delivered…



For More Information

Mark Conrad
ERA Research
mark.conrad@nara.gov
http://www.archives.gov/era/

Richard Marciano
DICE Group
marciano@diceresearch.org
http://www.DiceResearch.org