The iRODS Research Data Archive of the JGU: Introduction to its usage

Alvaro Frank
Researchers want

- Minimize risk of data loss
- Facilitate
  - Reproducibility
  - Teamwork
  - Follow-up projects
  - Storage
- Increase reputation > reuse of data > citations

Safe & long term data archival: iRODS
Benefits of the JGU iRODS

- Two copies at different locations (Tape+SSD/HDD)
- Encrypted Storage
- Granular Access control (permissions)
- Granular metadata
- Publication possibility
- Command line Linux tools
- WebUI*
- 10 years +

- Works just like a remote virtual filesystem
iRODS and FAIR principles

- **Findable**
  - Use Metadata standards: RADAR, Data Cite, Schema.org, ...
  - Author, Title, Keywords, ...

- **Accessible:**
  - Persistent open repositories: iRODS, Libraries, DRYAD*, figshare*, ...
  - Cloud: Dropbox, Drive, Azure, etc.

- **Interoperable:**
  - Unencrypted, open documented standards and formats.

- **Reusable:**
  - Documented, reproducible, accurate and relevant metadata, licensed.
Publishing JGU iRODS data

OpenStreetMap Traces

Data set description

The data set consists of complete virtual machine images (Virtual Box) collected over a period of more than 2 years. The images are based on Debian Linux and contain a small Linux base, a Mate display manager, standard multimedia applications, and software for navigation including an OpenStreetMap client from Germany. All data was updated on an almost weekly basis. During the observation time the image size increased from about 5 GB to about 5.5 GB without installing any additional software packages from user side.

Citation

Please cite the following paper in case you are using our traces for your research:

2018

Author/Publisher URL

Download

The metadata of the individual VMs including their iRODS tickets are available in the DedupVMBackups Collection.

The metadata of each individual file can be downloaded using the ticket numbers from the overall metadata, e.g., using https://irods-web.zdv.uni-mainz.de/irods-rest/rest/dataObject/zdv/project/zdvresearch/suesst/DedupVMBackups/VMs/backup-20160419.tar/metadata?ticket=KKaeFL1EEYXXtLH.

Individual virtual machines can finally be downloaded again using the ticket number received from the metadata, e.g., https://irods-web.zdv.uni-mainz.de/irods-rest/rest/fileContents/zdv/project/zdvresearch/suesst/DedupVMBackups/VMs/backup-20160419.tar?ticket=KKaeFL1EEYXXtLH.

https://irods-web.zdv.uni-mainz.de/irods-rest/rest/dataObject/zdv/project/zdvresearch/suesst/DedupVMBackups/VMs/backup-20160419.tar/metadata?ticket=KKaeFL1EEYXXtLH

https://irods-web.zdv.uni-mainz.de/irods-rest/rest/fileContents/zdv/project/zdvresearch/suesst/DedupVMBackups/VMs/backup-20160419.tar?ticket=KKaeFL1EEYXXtLH

https://research.zdv.uni-mainz.de/research-data/multimedia-traces/
iRODS workflow (Command Line*)

- generate data
- document data
- prepare metadata
- move data to ZDV*
- choose remote /path/location
- archive meta+data to iRods

- download data: zdv internal
- publish iRODS tickets
- done

- create DOI*
- download data: web
Using iRODS at the JGU

Access a ZDV Linux system
- Office PC
- ZDV servers

Data must be uploaded from a ZDV Linux system
- If not there already, it needs to be uploaded

Username + password

```
linux3:~ # ssh afrankra@linux.zdv.uni-mainz.de
```

SSH key

```
linux3:~ # ssh -J hpcgate.zdv.uni-mainz.de afrankra@miil01.zdv.uni-mainz.de
```

```
linux3:~/demo # ls mydata/
file001.bin  file003.bin  file005.bin  file007.bin  file009.bin
file002.bin  file004.bin  file006.bin  file008.bin  file010.bin
```

```
linux3:~/demo # scp -r mydata/ afrankra@linux.zdv.uni-mainz.de:/temp-data/
```
**iRODS commands**

<table>
<thead>
<tr>
<th>Linux</th>
<th>iRODS</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>pwd</code></td>
<td><code>ipwd</code></td>
</tr>
<tr>
<td><code>ls</code></td>
<td><code>ils</code></td>
</tr>
<tr>
<td><code>cd</code></td>
<td><code>icd</code></td>
</tr>
<tr>
<td><code>mkdir</code></td>
<td><code>imkdir</code></td>
</tr>
<tr>
<td><code>rm</code></td>
<td><code>irm</code></td>
</tr>
</tbody>
</table>

**Used for normal Linux files and directories**

**Used between virtual iRODS files and directories (objects and collections)**

- `download` ➡️ `iget`
- `iput` ➡️ `upload`
- `imeta`, `iticket`

[https://researchdata.uni-mainz.de/wiki-archiving-research-data/](https://researchdata.uni-mainz.de/wiki-archiving-research-data/)
Initialize account

Test connection to irods:

```
linux3:~/demo # ils
/zdv/home/afrankra:
  C- /zdv/home/afrankra/.irods
```

Sometimes needed if ils does not work:

```
linux3:~/demo # kinit -p afrankra
Password for afrankra@UNI-MAINZ.DE: 
linux3:~/demo # iinit
```

https://researchdata.uni-mainz.de/wiki-archiving-research-data/
Create a remote virtual directory

```
linux3:~/demo # ils
/zdv/home/afrankra:
    C- /zdv/home/afrankra/.irods
linux3:~/demo # mkdir /zdv/home/afrankra/vdemo
linux3:~/demo # ils
/zdv/home/afrankra:
    C- /zdv/home/afrankra/.irods
    C- /zdv/home/afrankra/vdemo
linux3:~/demo # icd /zdv/home/afrankra/vdemo
linux3:~/demo # ipwd
/zdv/home/afrankra/vdemo
linux3:~/demo #
```

```
linux3:~/demo # pwd
/uni-mainz.de/homes/afrankra/demo
```

Always use FULL paths!

different!

https://researchdata.uni-mainz.de/wiki-archiving-research-data/
iPut data into the remote virtual directory

```
linux3:~/demo # ls mydata/
file001.bin  file003.bin  file005.bin  file007.bin  file009.bin
file002.bin  file004.bin  file006.bin  file008.bin  file010.bin

linux3:~/demo # iput -v -r mydata/ /zdv/home/afrankra/vdemo
Running recursive pre-scan... pre-scan complete... transferring data...
```

```
<table>
<thead>
<tr>
<th>file</th>
<th>size</th>
<th>time</th>
<th>thread</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>file001.bin</td>
<td>0.000 MB</td>
<td>9.440 sec</td>
<td>0 thr</td>
<td>0.000 MB/s</td>
</tr>
<tr>
<td>file002.bin</td>
<td>0.000 MB</td>
<td>9.271 sec</td>
<td>0 thr</td>
<td>0.000 MB/s</td>
</tr>
<tr>
<td>file003.bin</td>
<td>0.000 MB</td>
<td>8.077 sec</td>
<td>0 thr</td>
<td>0.000 MB/s</td>
</tr>
<tr>
<td>file004.bin</td>
<td>0.000 MB</td>
<td>7.800 sec</td>
<td>0 thr</td>
<td>0.000 MB/s</td>
</tr>
<tr>
<td>file005.bin</td>
<td>0.000 MB</td>
<td>5.822 sec</td>
<td>0 thr</td>
<td>0.000 MB/s</td>
</tr>
<tr>
<td>file006.bin</td>
<td>0.000 MB</td>
<td>7.684 sec</td>
<td>0 thr</td>
<td>0.000 MB/s</td>
</tr>
<tr>
<td>file007.bin</td>
<td>0.000 MB</td>
<td>8.204 sec</td>
<td>0 thr</td>
<td>0.000 MB/s</td>
</tr>
<tr>
<td>file008.bin</td>
<td>0.000 MB</td>
<td>10.044 sec</td>
<td>0 thr</td>
<td>0.000 MB/s</td>
</tr>
<tr>
<td>file009.bin</td>
<td>0.000 MB</td>
<td>11.286 sec</td>
<td>0 thr</td>
<td>0.000 MB/s</td>
</tr>
<tr>
<td>file010.bin</td>
<td>0.000 MB</td>
<td>9.083 sec</td>
<td>0 thr</td>
<td>0.000 MB/s</td>
</tr>
</tbody>
</table>
```

[https://researchdata.uni-mainz.de/wiki-archiving-research-data/](https://researchdata.uni-mainz.de/wiki-archiving-research-data/)
iPut data into iRods

```
linux3:~/demo # iput -v -r mydata/ /zdv/home/afrankra/vdemo
Running recursive pre-scan... pre-scan complete... transferring data...
C - /zdv/home/afrankra/vdemo/mydata:
  file001.bin  0.000 MB   9.440 sec   0 thr   0.000 MB/s
  file002.bin  0.000 MB   9.271 sec   0 thr   0.000 MB/s
  file003.bin  0.000 MB   8.077 sec   0 thr   0.000 MB/s
  file004.bin  0.000 MB   7.800 sec   0 thr   0.000 MB/s
  file005.bin  0.000 MB   5.822 sec   0 thr   0.000 MB/s
  file006.bin  0.000 MB   7.684 sec   0 thr   0.000 MB/s
  file007.bin  0.000 MB   8.204 sec   0 thr   0.000 MB/s
  file008.bin  0.000 MB  10.044 sec   0 thr   0.000 MB/s
  file009.bin  0.000 MB  11.286 sec   0 thr   0.000 MB/s
  file010.bin  0.000 MB   9.083 sec   0 thr   0.000 MB/s
```

```
linux3:~/demo # lls /zdv/home/afrankra/vdemo/mydata
/zdv/home/afrankra/vdemo/mydata:
  file001.bin
  file002.bin
  file003.bin
  file004.bin
  file005.bin
  file006.bin
  file007.bin
  file008.bin
  file009.bin
  file010.bin
```

https://researchdata.uni-mainz.de/wiki-archiving-research-data/
Adding metadata with imeta

```
linux3:~/demo # imeta ls -C /zdv/home/afrankra/vdemo/mydata
AVUs defined for collection /zdv/home/afrankra/vdemo/mydata:
None

linux3:~/demo # imeta set -C /zdv/home/afrankra/vdemo/mydata Description "dummy data"
linux3:~/demo # imeta set -C /zdv/home/afrankra/vdemo/mydata License "GPL 2.0"
linux3:~/demo # imeta ls -C /zdv/home/afrankra/vdemo/mydata
AVUs defined for collection /zdv/home/afrankra/vdemo/mydata:
attribute: Description
value: dummy data
units: 
---
attribute: License
value: GPL 2.0
units:
```

https://researchdata.uni-mainz.de/wiki-archiving-research-data/
iGet data from iRods

```bash
linux3:~/demo # mkdir downloaded
linux3:~/demo # cd downloaded/
linux3:~/demo/downloaded # iget -r /zdv/home/afrankra/vdemo/mydata
linux3:~/demo/downloaded # ls ./mydata
file001.bin  file003.bin  file005.bin  file007.bin  file009.bin
file002.bin  file004.bin  file006.bin  file008.bin  file010.bin
```

```bash
linux3:~/demo # mkdir downloaded
linux3:~/demo # iget -r /zdv/home/afrankra/vdemo/mydata ./downloaded/
linux3:~/demo # ls downloaded/mydata/
file001.bin  file003.bin  file005.bin  file007.bin  file009.bin
file002.bin  file004.bin  file006.bin  file008.bin  file010.bin
```
Making data public for download (read only)

```
linux3:~/demo/downloaded # iticket create read /zdv/home/afranka/vdemo/mydata/file001.bin
        ticket:SAu5B1g3P8jHfz8

linux3:~/demo/downloaded # wget -O file001.bin -v https://irods-web.zdv.uni-mainz.de/irods-rest/rest/fileContents/zdv/home/afranka/vdemo/mydata/file001.bin?ticket=SAu5B1g3P8jHfz8
--2021-10-18 20:21:51-- https://irods-web.zdv.uni-mainz.de/irods-rest/rest/fileContents/zdv/home/afranka/vdemo/mydata/file001.bin?ticket=SAu5B1g3P8jHfz8
Resolving irods-web.zdv.uni-mainz.de (irods-web.zdv.uni-mainz.de)... 134.93.178.33
Connecting to irods-web.zdv.uni-mainz.de (irods-web.zdv.uni-mainz.de)|134.93.178.33|:443... connected.
HTTP request sent, awaiting response... 200 200
Length: 81 [application/octet-stream]
Saving to: 'file001.bin'

file001.bin 100%[========================================] 81.00 KiB/s in 0s

2021-10-18 20:21:51 (38.6 MB/s) - 'file001.bin' saved [81/81]
```

https://researchdata.uni-mainz.de/wiki-archiving-research-data/
Simplified archival

```
linux3:~/demo # source /usr/local/bin/i_init.sh
Setting up iRODS Login Profile

List of existing iRODS directories from afrankra ( ils for details ):
/zdv/home/afrankra private directory
/zdv/project/<PROJECT NAME> project directory
/zdv/home/public every registered user can read/write/delete
/zdv/trash/home/afrankra private trash bin

Current iRODS working directory ( ipwd )
/zdv/home/afrankra

linux working directory ( pwd )
/uni-mainz.de/homes/afrankra/demo
Notice the difference! Archives go into iRODS

iRODS commands are similar to unix commands with a preceeding 'i':
ils, ipwd, mkdir, iput, icp,.imv, irm, imeta
```
Simplified archival

These wrappers starting with 'i_<cmd>' make working with iRODS easier:
i_init, i_archive, i_metaupdate, i_publish, i_downlinkget, i_ticketget, i_exit

Your iRODS current directory will be displayed as:
iRODS:/zdv/home/afrankra
linux3:~/demo # icd vdemo
iRODS:/zdv/home/afrankra/vdemo
linux3:~/demo # ls
downloaded meta.json mydata
iRODS:/zdv/home/afrankra/vdemo
linux3:~/demo # cat meta.json
{
  "Title":"Mogon II sensor data.",
  "Description":"Sensor data from the Mogon II HPC @ JGU",
  "Subject":"Computer Science, Machine Learning",
  "ResourceType":"Binary Data",
  "PublicationDOI": "http://doi.org/10.1109/hipc.2019.00047",
  "URL":"https://research.zdv.uni-mainz.de/",
  "Contributors":"AAAAAAAA BBBBB, CCCCCC DDDDDD",
  "Project":"ZXY DFG"
}
Simplified archival

```
irods:/zdv/home/afranka/vdemo

linux3:~/demo # i_archive mydata meta.json

Upload a local directory or file to a remote iRODS directory (current default) using Metadata from a flat JSON file.

Contributors: AAAAAAAA BBBBB, CCCCCC DDDDDDD; Description: Sensor data from the Mogon II HPC @ JGU; Project: ZX Y DFG; PublicationDOI: http://doi.org/10.1109/hpc.2019.00047; ResourceType: Binary Data; Subject: Computer Science, Machine Learning; Title: Mogon II sensor data.; URL: https://research.zdv.uni-mainz.de/;

Uploading using this command:

input -T -K -r mydata --metadata="Contributors: AAAAAAAA BBBBB, CCCCCC DDDDDDD; Description: Sensor data from the Mogon II HPC @ JGU; Project: ZX Y DFG; PublicationDOI: http://doi.org/10.1109/hpc.2019.00047; ResourceType: Binary Data; Subject: Computer Science, Machine Learning; Title: Mogon II sensor data.; URL: https://research.zdv.uni-mainz.de/;"

Running recursive pre-scan... pre-scan complete... transferring data...

Data uploaded:

ils -r mydata
/zdv/home/afranka/vdemo/mydata:

  file001.bin
  file002.bin
  file003.bin
  file004.bin
  file005.bin
  file006.bin
  file007.bin
  file008.bin
  file009.bin
  file010.bin
```

https://researchdata.uni-mainz.de/wiki-archiving-research-data/
Simplified archival

iRODS:/zdv/home/afrankra/vdemo
linux3:/demo # i_publish /zdv/home/afrankra/vdemo/mydata
Creates a ticket for a remote iRODS file or directory and provides the public web download links to all files.
	
	iticket create read /zdv/home/afrankra/vdemo/mydata
		
		ticket:0nHg4uXerLHEmmm
This ticket can be used to retrieve the data
	
	OnHg4uXerLHEmmm

First Web link is to download the file, second link to download Metadata:
wget https://irods-web.zdv.uni-mainz.de/irods-rest/rest/fileContents/zdv/home/afrankra/vdemo/mydata/file001.bin?ticket=OnHg4uXerLHEmmm -x --content-disposition
curl https://irods-web.zdv.uni-mainz.de/irods-rest/rest/dataObject/zdv/home/afrankra/vdemo/mydata/file001.bin/metadata?ticket=OnHg4uXerLHEmmm

wget https://irods-web.zdv.uni-mainz.de/irods-rest/rest/fileContents/zdv/home/afrankra/vdemo/mydata/file002.bin?ticket=OnHg4uXerLHEmmm -x --content-disposition
curl https://irods-web.zdv.uni-mainz.de/irods-rest/rest/dataObject/zdv/home/afrankra/vdemo/mydata/file002.bin/metadata?ticket=OnHg4uXerLHEmmm

https://researchdata.uni-mainz.de/wiki-archiving-research-data/
Recommendations

- Always use full paths for iCommands
- iRODS API is installed in ZDV systems (linux/mogonii)
- Document data: metadata + technical metadata
- Making data open for web download? > use tar.gz!
- Files over 1TB? > split tar.gz!
- Check the documentation for more functionality

https://researchdata.uni-mainz.de/wiki-archiving-research-data/
https://docs.irods.org/master/icommands