



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

Alvaro Frank



Zentrum für Datenverarbeitung

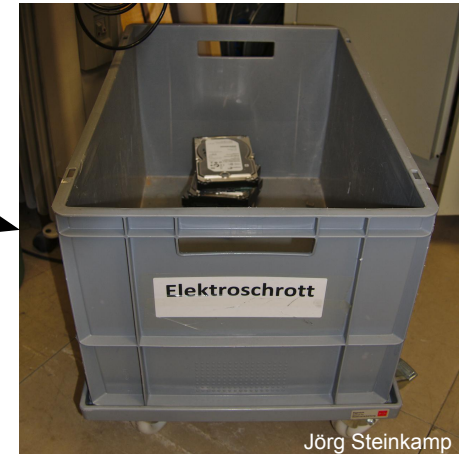
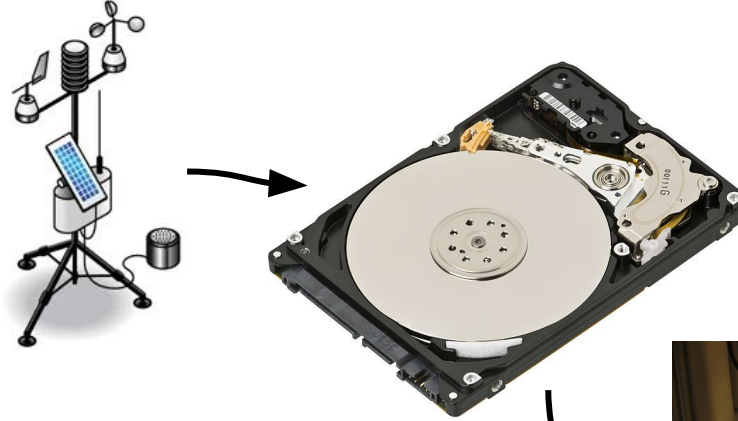
JOHANNES GUTENBERG
UNIVERSITÄT MAINZ



Why do Research Data Management (with iRODS)

Researchers want

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



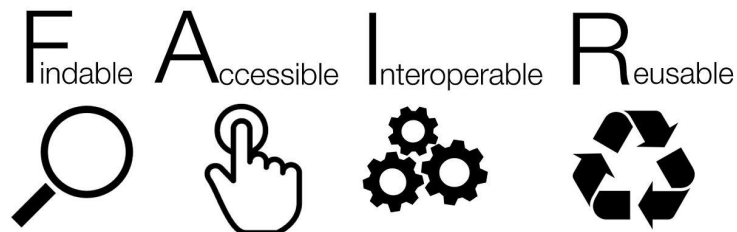
→ Safe & long term data archival: iRods

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



- **F**indable
 - Use Metadata standards: RADAR, Data Cite, Schema.org, ...
 - Author, Title, Keywords, ...
- **A**ccessible:
 - Persistent open repositories: iRODS, Libraries, DRYAD*, figshare*,...
 - ~~Cloud: Dropbox, Drive, Azure, etc.~~
- **I**nteroperable:
 - Unencrypted, open documented standards and formats.
- **R**eusable:
 - 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 periode 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 from Germany. All data was update on an almost weekly base. During the observation time the image size increased from about 5 GiB to about 5.5 GiB 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

- Tim Süß, Tunahan Kaya, Markus Mäsker, and André Brinkmann. 2018. Deduplication Analyses of Multimedia System Images. In *USENIX Workshop on Hot Topics in Edge Computing (HotEdge)*. Author/Publisher URL 

Download

The metadata of the inividual 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>.

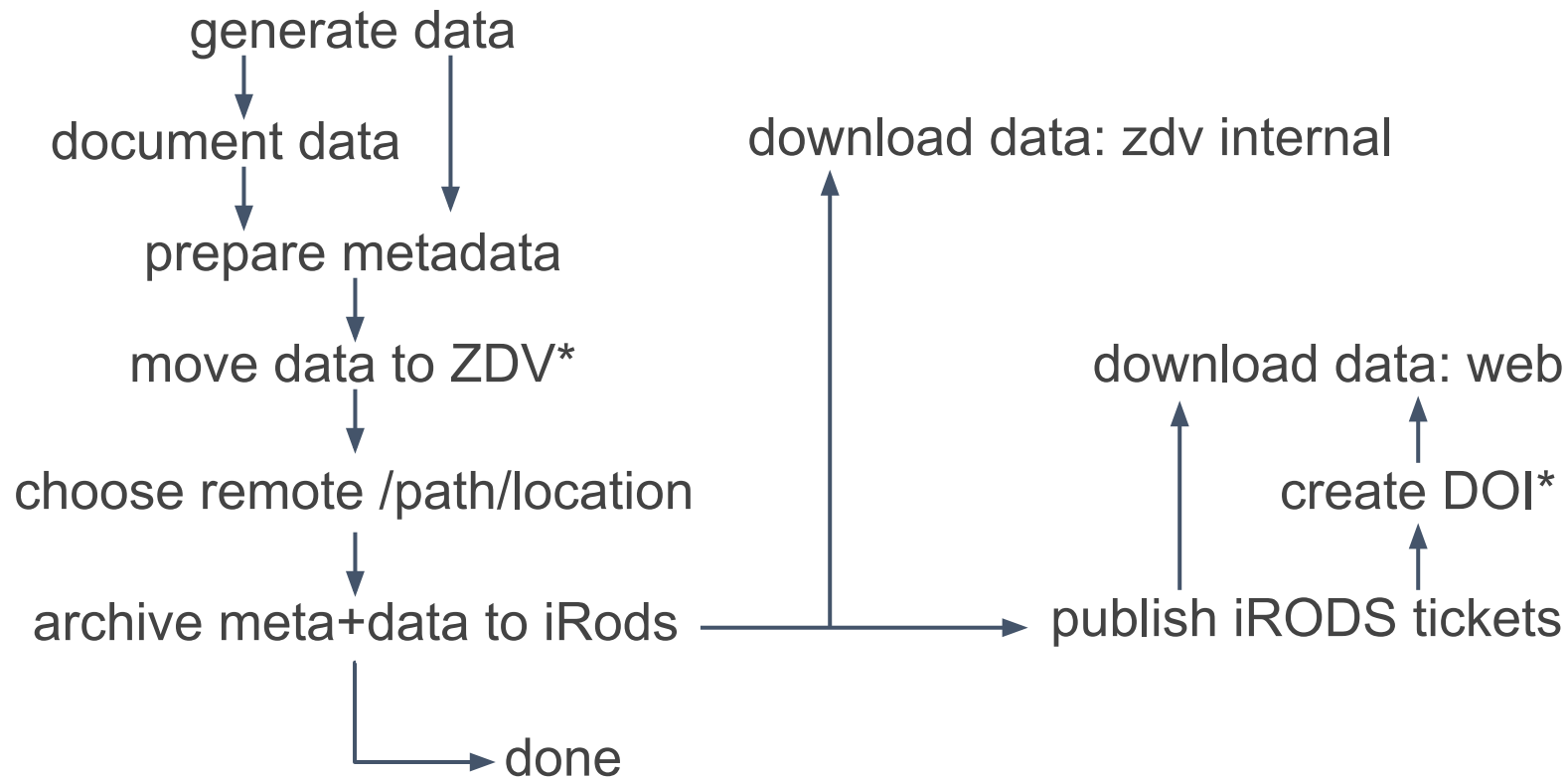
Individual virtual machines can finally be downloaded again using the ticket numbert received from the metadata, e.g., <https://irods-web.zdv.uni-mainz.de/irods-rest/rest/fileContents/zdv/project/zdvresearch/suesst/DedupVMBackups/VMs/backup-20160419.tar?ticket=KKaeFL1EEYXXtLH>.

<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*)



Using iRODS at the JGU

Access a ZDV Linux system

- Office PC
- ZDV servers

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
```

Data must be uploaded from a ZDV Linux system

- If not there already, it needs to be uploaded

```
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

Used for normal linux
files and directories

Linux	iRODS
pwd	ipwd
ls	ils
cd	icd
mkdir	imkdir
rm	irm

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

download ← iget
 iput → upload
 imeta, iticket

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
```

Create a remote virtual directory

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

Always use FULL paths!

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

different!

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...
```

```
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

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 # ils /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
linux3:~/demo #
```

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:
```

iGet data from iRods

```
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
```

```
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/afrankra/vdemo/mydata/file001.bin
ticket:SAu5BIg3P8jHfz8
linux3:~/demo/downloaded # wget -O file001.bin -v https://irods-web.zdv.uni-mainz.de/irods-rest/rest/fileContents/zdv/home/afrankra/vdemo/mydata/file001.bin?ticket=SAu5BIg3P8jHfz8
--2021-10-18 20:21:51-- https://irods-web.zdv.uni-mainz.de/irods-rest/rest/fileContents/zdv/home/afrankra/vdemo/mydata/file001.bin?ticket=SAu5BIg3P8jHfz8
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  --.-KB/s    in 0s

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

linux3:~/demo/downloaded # ls
file001.bin  mydata
```

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, imkdir, 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":"AAAAAAA BBBBBB, CCCCCC DDDDDD",  
  "Project":"ZXY DFG"  
}
```

Simplified archival

```
iRODS:/zdv/home/afrankra/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;AAAAAAA BBBB, CCCCC DDDDD;;Description;Sensor data from the Mogon II HPC @ JGU;;Project;ZX
Y DFG;;PublicationDOI;http://doi.org/10.1109/hipc.2019.00047;;ResourceType;Binary Data;;Subject;Computer S
cience, Machine Learning;;Title;Mogon II sensor data.;;URL;https://research.zdv.uni-mainz.de/;
Uploading using this command:
iput -T -K -r mydata --metadata="Contributors;AAAAAAA BBBB, CCCCC DDDDD;;Description;Sensor data from t
he Mogon II HPC @ JGU;;Project;ZXY DFG;;PublicationDOI;http://doi.org/10.1109/hipc.2019.00047;;ResourceTyp
e;Binary Data;;Subject;Computer Science, Machine Learning;;Title;Mogon II sensor data.;;URL;https://resear
ch.zdv.uni-mainz.de/;"
Running recursive pre-scan... pre-scan complete... transferring data...
Data uploaded:
ils -r 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
```

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:OnHg4uXerLHEmmm
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
```

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>