Data Management in the HPC
Tools and Workflows

February 24, 2022
osf.io/98bzd/

Fernando Rios
Research Data Management Specialist, Library – frios@arizona.edu
data.library.arizona.edu

Chris Reidy
Research Facilitation Manager, Research Computing– chrisreidy@arizona.edu
What you will learn

Part 1: learned about policies and basic tools/approaches for data management

Part 2: Learn about additional tools for transferring, organizing, and archiving HPC data.

Why: Setting up workflows means better collaboration, reproducibility => better research

Assume

• Familiarity with the Unix shell
Sample Transparent and Reproducible Research Pipeline

Basic workflow – less reproducible

Get data

Compute

Export

Data, Figures

Advanced workflow – More Reproducible

Get

Manage

docker

S

Get

Manage

Open Science Framework

Links

Collab w/ Co-authors (manuscripts, data)

Publish & archive

GitHub

Manage

ReDATA
arizona.figshare.com
Get Data
## UA Storage Refresher

<table>
<thead>
<tr>
<th>Tier</th>
<th>Uses</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tier 1 – HPC storage</td>
<td>High performance. Active research. <strong>Limited permanent storage.</strong></td>
<td>Home directory PI allocations.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tier 2 – General storage</td>
<td>Less frequent access Copy subsets as needed to tier 1. <strong>Backups</strong></td>
<td>Google Drive Box Your computer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tier 3 - Archival storage</td>
<td>Store data after project completion <strong>Publish data that supports publications</strong></td>
<td>UA ReDATA Amazon Glacier</td>
</tr>
</tbody>
</table>
HPC Storage Refresher

Tier 1

File transfer node

Bastion host

Login node

Shared data storage

Computes nodes
Getting data into the HPC

Many ways to do it...

Small, infrequent transfers

OOD.hpc.arizona.edu

General purpose

$ wget

FileZilla

WinSCP

rsync

Large transfer >100GB, scheduled transfers, transfer outside UA

globus
Rclone

• Mature software for working with cloud storage
• Mirroring, syncing, encryption, union and more
• Supports 40+ providers
  • Gdrive
  • Box
  • Amazon S3
  • OneDrive
  • SFTP
  • ownCloud
  • ...
Tier 2 to Tier 1 transfer

Get data

Google Drive

RCLONE

Get code

GitHub

docker

S

Compute

Export

data, figures

Manage

Mendeley

box

Dropbox

Collab w/ Co-authors (manuscripts, data)

Manage

Open Science Framework

Publish & archive (w/ DOI)

ReDATA

arizona.figshare.com
Basic Rclone workflow

The sample data is also available on the OSF osf.io/7rbpd

• Log in to HPC
  • ssh <netid>@filexfer.hpc.arizona.edu

• rclone config
  • Create the remote UA-gdrive – see rclone documentation

• rclone lsf UA-gdrive:'/OSF/HPC Demo'
  • Test the connection

• rclone copy UA-gdrive:'/OSF/HPC Demo' .
Advanced Rclone workflow

Mounting: Access Google as if it were a local folder (e.g., can use standard commands like ls, cat, cp).

- Configure the remote (if not configured already)
- Request an interactive session on a cluster
  - Mounting doesn’t work on filexfer or login nodes
  - `rclone mount UA-gdrive:'/OSF/HPC Demo' ~/Desktop/mount/ &`
  - Mount into a folder (e.g., mount)

- `cd ~/Desktop/mount ; ls`
- Test the connection

- `fusermount -uz ~/Desktop/temp/directory`
- Note: lots of caveats!! See rclone docs
Now what?
File Management

Need a strategy before doing anything else
Data Management Best Practices

In Part 1 we covered versioning and file/folder organization.

**Better**

```
Study001_Raw
  └ BiopsyData
    ├ 20161101_Study001_Biopsy_visit1.xls
    ├ 20161101_Study001_Biopsy_visit1_v2.xls
    ├ 20161101_Study001_Biopsy_visit1_v3.xls
```

**Add version and/or date (ISO 8601)**

```
GW_model
  ├ elevation.mat
  ├ depth_wt.csv
  ├ well_loc.csv
  ├ flow_model.m
  ├ flow_model2.m
  ├ flow_model_final.m
  ├ flowlines1.png
  ├ flowlines2.png
  ├ contours.png
```

**Descriptive folder names**

```
GroundwaterModel
  └ Code
    ├ 20170402_FlowModel_v1.m
    ├ 20170410_FlowModel_v2.m
    ├ 20170511_FlowModel_v3.m
  └ Inputs
    ├ TerrainElevation.m
    ├ DepthToWaterTable.csv
    └ WellLocations.csv
  └ Outputs
    ├ 20170402_Flowlines_FlowModelv1.png
    ├ 20170402_Contours_FlowModelv1.png
    └ 20170415_Flowlines_FlowModelv2.png
```

**Document the naming scheme**

**Bad**

```
GW_model
  └ elevation.mat
  └ depth_wt.csv
  └ well_loc.csv
  └ flow_model.m
  └ flow_model2.m
  └ flow_model_final.m
  └ flowlines1.png
  └ flowlines2.png
  └ contours.png
```

**“FINAL.doc”**

```
  └ FINAL.doc!
  └ FINAL_rev.2.doc
     └ FINAL_rev.18.comments7.corrections9.MORE.30.doc
     └ FINAL_rev.22.comments49.corrections10.#@%WHYDIDICOMETOGRADEschool?????doc
```


www.phdcomics.com
Project Setup

- Do you have a more complex project?

![Image of project structure]

**Basic Metadata**

- `AUTHORS.md`
- `LICENSE`
- `README.md`

**Each stage of data in its own folder**

- `bin`
- `config`
- `data` - Data from third party sources.
- `interim` - Intermediate data that has been transformed.
- `processed` - The final, canonical data sets for modeling.
- `raw` - The original, immutable data dump.
- `docs` - Documentation, e.g., doxygen or scientific papers (not tracked by git)
- `notebooks` - Ipython or R notebooks
- `reports` - For example, LaTeX, Markdown, etc., or any project reports
- `figures` - Figures for project reports
- `src` - Source code to process data

**Software in its own folder**

- `data`
- `external` - Any external source code, e.g., pull other git projects, or external libraries
- `models` - Source code for your own model
- `tools` - Any helper scripts go here
- `visualization` - Scripts for visualisation of your results, e.g., matplotlib, ggplot2 related.

[GitHub link: https://github.com/mkrapp/cookiecutter-reproducible-science]
Cookiecutter

- Install: See instructions on OSF (osf.io/9ceqd).
- cookiecutter gh:mkrapp/cookiecutter-reproducible-science

(puma) frios@r2u06n1 ~$ cookiecutter gh:mkrapp/cookiecutter-reproducible-science
You've downloaded /home/u17/frios/.cookiecutters/cookiecutter-reproducible-science
ownload it? [yes]: yes
full_name [Mario Krapp]: Fernando Rios
email [mariokrapp@gmail.com]: frios@arizona.edu
github_username [mkrapp]: zoidy
project_name [Name of your science project]: HPC Demo 2021
project_slug [hpc-demo-2021]:
project_short_description [A short description of your project]: Hello world!
release_date [2021-10-04]:
version [0.1.0]: 1.0
(puma) frios@r2u06n1 ~$ |

Other templates
https://github.com/luke-gregor/cookiecutter-science-pub
https://github.com/mnarayan/cookiecutter-data-science
File & Space Management Tools

- Checking your space & file limit: uquota

```
frios@login2 ~> uquota
frios home & PBS /extra/frios
   used         soft limit      hard limit      files/limit
 46.84M           14G            15G          1044
 56.72G          200G           200G          2/120000
```

- Checking folder usage and count: NCDU

HPC installation instructions and pre-compiled binaries: https://osf.io/98bzd/
File & Space Management Tools

- Keeping file names tidy with renameutils

Pre-compiled version for HPC https://osf.io/98bzd/

Credit: ostechnix.com
Data Mgmt Best Practice: Storage & Backup

“I decide what data is important while I am working on it and typically save it in a single location”

Do

• 3-2-1: If possible, 3 copies, 2 different storage types, 1 copy offsite
• Keep offline backups if possible. Sync clients could propagate changes unintentionally

Avoid:

• Storing sensitive data on an unencrypted laptop or flash drive or insecure servers
• Relying on cloud storage for the only copy!


For HIPAA compliance, use UA Box Health account.
Backup and Restore to Google Drive (Tier 1 to Tier 2)

- Rclone to transfer directly to/from Google Drive, Dropbox, S3, Box... many more
- TIP: Transferring lots of little files is slow. Put everything in a tar archive
Finish Project and Export

- Export to Google Drive via Globus
Why Globus?

- Designed for large data
- Reliable – supports resuming
- Can initiate remotely, no babysitting transfers
  - Email on completion

https://www.globus.org/data-transfer
Globus Demo

• Go to https://www.globus.org/ and log in

• In the File Manager tab,
  • Click the left hand Search box and search for the collection “UA HPC Filesystems”
  • In the box on the right, search for “UA Google Drive”

• You may have to authorize a bunch of requests
Scheduled/recurring transfers using Globus

- Beta Globus Timer
  https://pypi.org/project/globus-timer-cli/
- pip3 install --user globus-timer-cli
- globus-timer session login

```
(elgato) frios@cpu1 ~> globus-timer session login
Please log into Globus here:
-------------------------------
https://auth.globus.org/v2/oauth2/authorize?client_id=bc7th.globus.org%2Fv2%2Fweb%2Fauth-code&scope=profile+email+wgYS81mcVAwZSrzsfgMhZEQH3qmOd5V-d-LM&code_challenge_method=Line+Interface+on+cpu1.elgato.hpc.arizona.edu
-------------------------------
```

Enter the resulting Authorization Code here: |
Scheduled/recurring transfers using Globus

- Set up the transfer. Transfer some files from xdisk to Google Drive (may need to authorize up to 3x)

```bash
globus-timer job transfer \
--name my-job \
--label "Timer Transfer Job" \
--interval 120 \
--start '2021-10-04T12:57:00' \
--source-endpoint 7c4462b2-7ca4-4f44-820a-xxxxxxxxxxxxx \
--dest-endpoint 26b96369-5f03-4742-9ab8-xxxxxxxxxxxx \
--verify-checksum \
--sync-level 2 \
--item '/home/u17/frios/xdisk/coca/Shared Files/' '/xdisk/coca/Shared Files/folder1' true
```

Don’t have to be logged in to the HPC. Get email notification + status in Globus web

Source and dest endpoint UUIDs from Globus web globus.org

```
(puma) frios@junonia ~$ globus-timer job status 1ced207e-29d9-4d98-a4aa-e6368f3f4369
Name: my-job
Job ID: 1ced207e-29d9-4d98-a4aa-e6368f3f4369
Status: loaded
Start: 2021-10-04T19:57:00+00:00
Interval: 0:02:00
Next Run At: 2021-10-04T20:03:00+00:00
Last Run Result: RUN COMPLETE
```
Scheduled/recurring transfers using Globus

- Job runs automatically
- Check status
  - globus-timer job status <job_id>

Globus Notification <no-reply@globus.org>

[EXT]SUCCEEDED - Timer Transfer Job

To Rios, Fernando - (frios)

External Email

TASK DETAILS
Task ID: e7a36dde-2871-11ec-95d4-853490a236f9
Task Type: TRANSFER
Status: SUCCEEDED
Source: UA HPC Filesystems (7c4462b2-7ca4-4f44-820a-b3ae9f7865fd)
Destination: HPC UA GDrive (26b96369-5f03-4742-9ab8-d4e9de3dcb8b)
Label: Timer Transfer Job
https://app.globus.org/activity/e7a36dde-2871-11ec-95d4-853490a236f9/overview
Project is Done – publish an article

Work done, data analyzed, paper ready to submit

Now what?

- Archive the “final” data
- Cite it in the paper => get credit
- Publish it in a data repository
UA Research Data Repository (ReDATA)

- Long-term archival repository for “final” data
- Get a DOI
- Comply with funder, journal policies, UA retention policy
- Get help improving the data for reuse
- You don’t have to worry about keeping data around, even if you leave UA
Go to arizona.figshare.com, log in
Takeaways

• Setting up data management workflows increase efficiency and support doing good research

• By linking together both UA-provided and 3\textsuperscript{rd} party tools and resources, you can build a solid workflow at no cost

• Refer to documentation and guides
  • public.confluence.arizona.edu/display/UAHPC
  • data.library.arizona.edu/osf
  • data.library.arizona.edu/redata
The data cooperative is a group of library-based data services providers
https://data.library.arizona.edu

Data Management
Consulting, data management plans, support for DMPTool, OSF, data archiving via ReDATA

Data Science & Visualization
Data analysis & data visualization support through consulting and instruction

Geospatial Support
Data management consulting, data management plans, data archiving via ReDATA

data.library.arizona.edu/data-management/events-schedule-current
Pipelines: Data news, events for UA
https://redata.tiny.us/dm-news