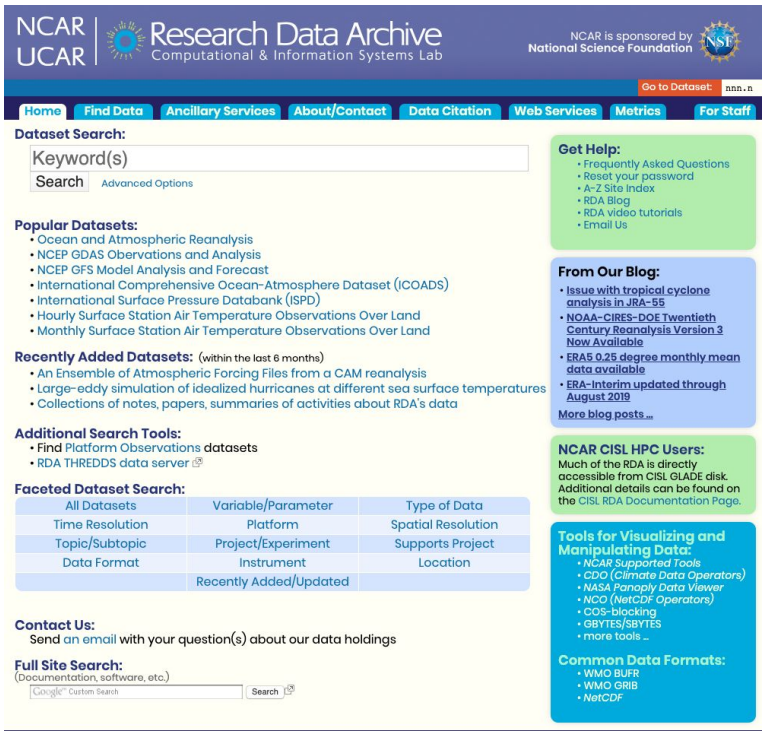
Muntaha Pasha
03/02/2020
Rotation 3 - DECS

DECS Rotation (NCAR SARP)

I want to start off this report by explaining the purpose of all this documentation. I'm writing this report to recount my experience here in the DECS team of CISL and what I learned, as well as what I am taking away from this rotation! In this rotation, I worked with Bob Dattore. He was my mentor, and showed me a lot of what data archiving looks like and how it works.

1. Introduction to DECS

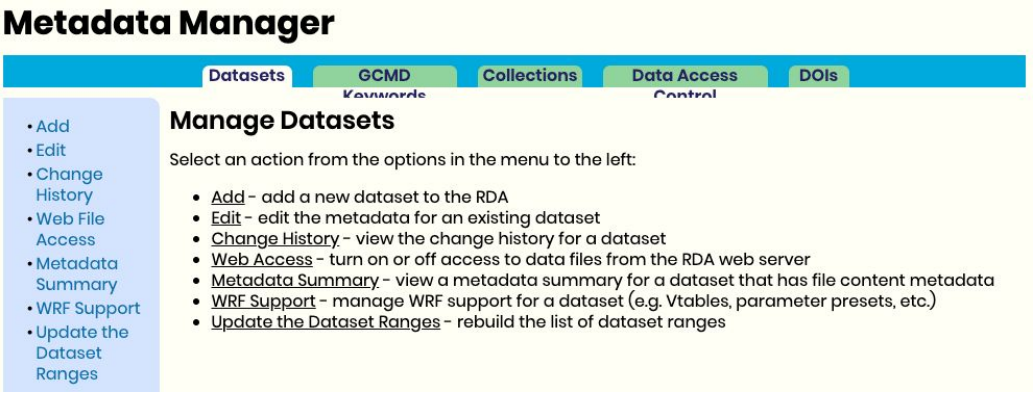
The first thing I did in this rotation was learn what the DECS team does, how they curate and archive data, and it was really eye opening for me because I was never exposed to this particular side of data engineering and management at school. I always figured data engineering just encompassed writing queries against large databases to pull and analyze information… but it was so much more than that! Data needs to be curated, it needs to be taken in from external sources, and it needs to be manipulated and put out in several different forms centering around user readability, and simplicity while still maintaining the data to be free of errors.



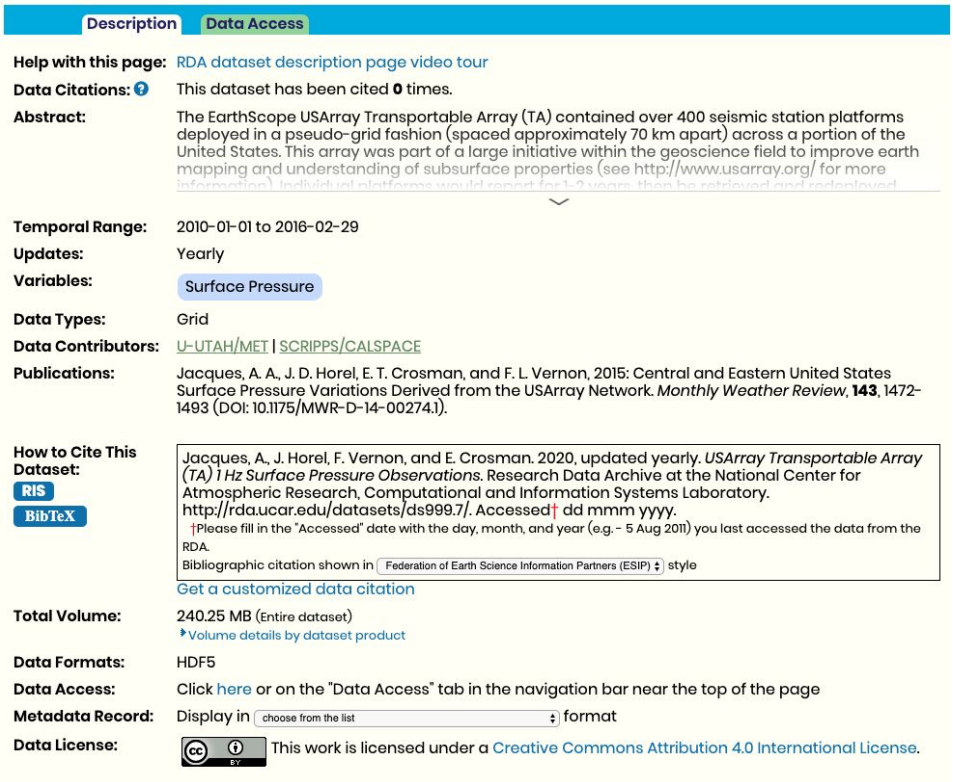
I was initially acquainted with the rda-data website that DECS maintains and manages. It includes several hundreds of databases, most of them being accessible widely to the public and researchers all over the globe. I was able to explore multiple aspects of this website, including learning what a faceted search was. Faceted searches are a lot more user centered, as you can pick certain aspects of the data you want to see and specifically find those. For instance, if you want to see only data for a specific geographical region, or if you're looking for a particular type of data like ice cover or air particulates, you can look for those in the faceted search. It implements a lot of unique and interesting filters that you can use to narrow down what you want.

2. Learning the Technology

The most important part of joining any team is to really learn all about the utilities they use and interact with on a daily basis! One of these was the metadata manager.



DECS has this Metadata Manager feature on their website, where all the metadata about a certain database can be manually entered or scanned in! Usually databases have headers which you can parse to get all of this information, but I decided to play around with it and manually enter some data!



Here's an example showing some of the things I entered as a trial run in the metadata manager. It covers things like data formats, and authors, including everything from different variables to summaries about the data. It's a great way of knowing exactly what your data covers, what kinds of measurements its' taking, and more about the contributors/authors of the data! Usually DECS scans all of this

information in from the database itself, but I wanted to manually enter some information and see what it looked like! From this, Bob and I also talked about the different kinds of data, since there are a lot of particular formats they can come in. Often times DECS deals with GRIB and NETCDF files, but there's plenty of other data formats that they see a lot of as well.

3. Accessing the Data

In this rotation, I learned primarily how to archive data and put it into the metadata manager by scanning it in through. I tried my hand at writing bash scripts to do all of that for me, and had a really great time learning about how DECS does all the archiving of their data so quickly! The first thing I had to do each morning was log into the cheyenne supercomputer. This is where all the data was stored, and the source of the archiving. I had my own work directory where I was able to access some data and learn how to archive it properly.

```
cisl-green:~ mpasha$ ssh cheyenne.ucar.edu
Warning: Permanently added the ECDSA host key for IP address '128.117.181.200' to the list of known hosts.
Token_Response:
*******************************************************************************
*                Welcome to Cheyenne - February 13, 2020
*******************************************************************************
                Today in the Daily Bulletin (dailyb.cisl.ucar.edu)

       - Experts available to collaborate on visualizing data
       - Reminder: Extended Cheyenne downtime scheduled for February 25-29
       - Tutorial February 20: Using Globus to transfer and share data
       - University request for large-scale allocations are due March 23

Quick Start:          www2.cisl.ucar.edu/resources/cheyenne/quick-start-cheyenne
User environment:     www2.cisl.ucar.edu/resources/cheyenne/user-environment
Key module commands:  module list, module avail, module spider, module help
CISL Help:            support.ucar.edu -- 303-497-2400
-------------------------------------------------------------------------------


mpasha@cheyenne6:~> echo $PATH
/ncar/opt/slurm/latest//bin:/opt/clmgr/sbin:/opt/clmgr/bin:/opt/sgi/sbin:/opt/sgi/bin:/glade/u/apps/ch/opt/netcdf/4.7.3/intel/18.0.5/bin:/glade/u/apps/ch/
opt/ncarcompilers/0.5.0/intel/18.0.5/mpi:/glade/u/apps/ch/opt/mpt/2.19/bin:/glade/u/apps/ch/opt/ncarcompilers/0.5.0/intel/18.0.5:/glade/u/apps/opt/intel/2
018u4/vtune_amplifier/bin64:/glade/u/apps/opt/intel/2018u4/inspector/bin64:/glade/u/apps/opt/intel/2018u4/advisor/bin64:/glade/u/apps/opt/intel/2018u4/com
pilers_and_libraries/linux/bin/intel64:/glade/u/apps/opt/globus-utils:/glade/u/apps/ch/opt/usr/bin:/glade/u/apps/ch/opt/lmod/8.1.7/bin:/usr/local/bin:/usr
/bin:/bin:/usr/bin/X11:/usr/games:/opt/ibutils/bin:/usr/lib/mit/bin:/glade/u/apps/opt/hpss/hpss-client.5.0.2.p5/bin:/glade/u/apps/ch/os/usr/bin:/opt/pbs/b
in:/sbin:/bin:/gpfs/u/home/rdadata/bin:/ncar/rda/setuid/bin
mpasha@cheyenne6:~> dsarch -sd -ds ds999.7 -ud P
Set dataset info of ds999.7 ...
No change of dataset record for ds999.7!
```

```
mpasha@cheyenne6:~> cd /glade/collections/rda/transfer/download/pasha
mpasha@cheyenne6:/glade/collections/rda/transfer/download/pasha> ld
ld: no input files
mpasha@cheyenne6:/glade/collections/rda/transfer/download/pasha> ls
station_A30A_2010_1Hz_pressure_data.h5  station_I37A_2012_1Hz_pressure_data.h5  station_Q34A_2011_1Hz_pressure_data.h5
station_A30A_2011_1Hz_pressure_data.h5  station_I37B_2014_1Hz_pressure_data.h5  station_Q34A_2012_1Hz_pressure_data.h5
station_A31A_2010_1Hz_pressure_data.h5  station_I37B_2015_1Hz_pressure_data.h5  station_Q35A_2010_1Hz_pressure_data.h5
station_A31A_2011_1Hz_pressure_data.h5  station_I37B_2016_1Hz_pressure_data.h5  station_Q35A_2011_1Hz_pressure_data.h5
station_A31A_2012_1Hz_pressure_data.h5  station_I37B_2017_1Hz_pressure_data.h5  station_Q35A_2012_1Hz_pressure_data.h5
station_A32A_2010_1Hz_pressure_data.h5  station_I38A_2010_1Hz_pressure_data.h5  station_Q36A_2010_1Hz_pressure_data.h5
station_A32A_2011_1Hz_pressure_data.h5  station_I38A_2011_1Hz_pressure_data.h5  station_Q36A_2011_1Hz_pressure_data.h5
station_A32A_2012_1Hz_pressure_data.h5  station_I38A_2012_1Hz_pressure_data.h5  station_Q36A_2012_1Hz_pressure_data.h5
station_A33A_2010_1Hz_pressure_data.h5  station_I39A_2011_1Hz_pressure_data.h5  station_Q37A_2010_1Hz_pressure_data.h5
station_A33A_2011_1Hz_pressure_data.h5  station_I39A_2012_1Hz_pressure_data.h5  station_Q37A_2011_1Hz_pressure_data.h5
station_A33A_2012_1Hz_pressure_data.h5  station_I39A_2013_1Hz_pressure_data.h5  station_Q37A_2012_1Hz_pressure_data.h5
station_A36M_2013_1Hz_pressure_data.h5  station_J30A_2010_1Hz_pressure_data.h5  station_Q38A_2010_1Hz_pressure_data.h5
station_A36M_2014_1Hz_pressure_data.h5  station_J30A_2011_1Hz_pressure_data.h5  station_Q38A_2011_1Hz_pressure_data.h5
station_A36M_2015_1Hz_pressure_data.h5  station_J30M_2017_1Hz_pressure_data.h5  station_Q38A_2012_1Hz_pressure_data.h5
station_A36M_2016_1Hz_pressure_data.h5  station_J31A_2010_1Hz_pressure_data.h5  station_Q39A_2010_1Hz_pressure_data.h5
station_A36M_2017_1Hz_pressure_data.h5  station_J31A_2011_1Hz_pressure_data.h5  station_Q39A_2011_1Hz_pressure_data.h5
station_B30A_2010_1Hz_pressure_data.h5  station_J31A_2012_1Hz_pressure_data.h5  station_Q39A_2012_1Hz_pressure_data.h5
station_B30A_2011_1Hz_pressure_data.h5  station_J32A_2010_1Hz_pressure_data.h5  station_R30A_2010_1Hz_pressure_data.h5
station_B31A_2010_1Hz_pressure_data.h5  station_J32A_2011_1Hz_pressure_data.h5  station_R30A_2011_1Hz_pressure_data.h5
station_B31A_2011_1Hz_pressure_data.h5  station_J32A_2012_1Hz_pressure_data.h5  station_R31K_2016_1Hz_pressure_data.h5
station_B31A_2012_1Hz_pressure_data.h5  station_J33A_2010_1Hz_pressure_data.h5  station_R31K_2017_1Hz_pressure_data.h5
station_B32A_2010_1Hz_pressure_data.h5  station_J33A_2011_1Hz_pressure_data.h5  station_R32A_2010_1Hz_pressure_data.h5
station_B32A_2011_1Hz_pressure_data.h5  station_J33A_2012_1Hz_pressure_data.h5  station_R32A_2011_1Hz_pressure_data.h5
station_B32A_2012_1Hz_pressure_data.h5  station_J34A_2010_1Hz_pressure_data.h5  station_R32B_2014_1Hz_pressure_data.h5
station_B33A_2010_1Hz_pressure_data.h5  station_J34A_2011_1Hz_pressure_data.h5  station_R32B_2015_1Hz_pressure_data.h5
station_B33A_2011_1Hz_pressure_data.h5  station_J34A_2012_1Hz_pressure_data.h5  station_R32B_2016_1Hz_pressure_data.h5
station_B33A_2012_1Hz_pressure_data.h5  station_J35A_2010_1Hz_pressure_data.h5  station_R32B_2017_1Hz_pressure_data.h5
station_B34A_2010_1Hz_pressure_data.h5  station_J35A_2011_1Hz_pressure_data.h5  station_R32K_2016_1Hz_pressure_data.h5
station_B34A_2011_1Hz_pressure_data.h5  station_J35A_2012_1Hz_pressure_data.h5  station_R32K_2017_1Hz_pressure_data.h5
station_B34A_2012_1Hz_pressure_data.h5  station_J36A_2010_1Hz_pressure_data.h5  station_R33A_2010_1Hz_pressure_data.h5
station_B35A_2010_1Hz_pressure_data.h5  station_J36A_2011_1Hz_pressure_data.h5  station_R33A_2011_1Hz_pressure_data.h5
station_B35A_2011_1Hz_pressure_data.h5  station_J36A_2012_1Hz_pressure_data.h5  station_R33M_2016_1Hz_pressure_data.h5
station_B35A_2012_1Hz_pressure_data.h5  station_J37A_2010_1Hz_pressure_data.h5  station_R33M_2017_1Hz_pressure_data.h5
station_B35B_2014_1Hz_pressure_data.h5  station_J37A_2011_1Hz_pressure_data.h5  station_R34A_2010_1Hz_pressure_data.h5
station_B35B_2015_1Hz_pressure_data.h5  station_J37A_2012_1Hz_pressure_data.h5  station_R34A_2011_1Hz_pressure_data.h5
station_B35B_2016_1Hz_pressure_data.h5  station_J38A_2010_1Hz_pressure_data.h5  station_R34A_2012_1Hz_pressure_data.h5
station_B35B_2017_1Hz_pressure_data.h5  station_J38A_2011_1Hz_pressure_data.h5  station_R35A_2010_1Hz_pressure_data.h5
station_C30A_2010_1Hz_pressure_data.h5  station_J38A_2012_1Hz_pressure_data.h5  station_R35A_2011_1Hz_pressure_data.h5
station_C30A_2011_1Hz_pressure_data.h5  station_J39A_2011_1Hz_pressure_data.h5  station_R35A_2012_1Hz_pressure_data.h5
station_C31A_2010_1Hz_pressure_data.h5  station_J39A_2012_1Hz_pressure_data.h5  station_R36A_2010_1Hz_pressure_data.h5
station_C31A_2011_1Hz_pressure_data.h5  station_J39A_2013_1Hz_pressure_data.h5  station_R36A_2011_1Hz_pressure_data.h5
station_C31A_2012_1Hz_pressure_data.h5  station_K30A_2010_1Hz_pressure_data.h5  station_R36A_2012_1Hz_pressure_data.h5
station_C32A_2010_1Hz_pressure_data.h5  station_K30A_2011_1Hz_pressure_data.h5  station_R37A_2010_1Hz_pressure_data.h5
station_C32A_2011_1Hz_pressure_data.h5  station_K30B_2017_1Hz_pressure_data.h5  station_R37A_2011_1Hz_pressure_data.h5
station_C32A_2012_1Hz_pressure_data.h5  station_K31A_2010_1Hz_pressure_data.h5  station_R37A_2012_1Hz_pressure_data.h5
station_C33A_2010_1Hz_pressure_data.h5  station_K31A_2011_1Hz_pressure_data.h5  station_R38A_2010_1Hz_pressure_data.h5
```

Above, there is a sample of some data files I was archiving. They deal with pressure measurements. To group all these files accordingly, (since its a lot to look at!) Bob and I went through

mpasha — ssh cheyenne.ucar.edu — 154×67

```
#! /bin/bash

echo "hello"
cd /glade/collections/rda/transfer/download/pasha/
for year in 2010 2011 2012 2013 2014 2015 2016 2017
do
echo $year
for file in $(ls *$year*)
do
echo $file
#f=`echo $file|sed "s/\/glade\/collections\/rda\/transfer\/download\/pasha\///"`
#echo $f
dsarch -ab -ds ds999.7 -lf $file -mf /FS/DECS/DS999.7/$year/$file -wf $year/$file -df HDF5 -gi $year
done
done
cd
```

how to group it by something like the year, which would make it easier for the users to find the files they want by the group ID.

This is what our script ended up looking like. I was really surprised bash scripting offered its own format of looping, it was something I was unaware of and didn't really know I could do! This was a great way to be able to execute a single command on one file onto hundreds of other files at the same time. Replacing each file name in the command every time was super time consuming so this really narrowed down the tedious work load and even taught me a new way to script.

[HPSS holdings]

Csh Download Script | Python Download Script | Jupyter notebook download Script

Group Summary

Group ID	Data Description	FILE COUNT
2010 Files	Data Files for Year 2010	200
2011 Files	Data Files for Year 2011	220
2012 Files	Data Files for Year 2012	175
2013 Files	Data Files for Year 2013	18
2014 Files	Data Files for Year 2014	22
2015 Files	Data Files for Year 2015	26
2016 Files	Data Files for Year 2016	45
2017 Files	Data Files for Year 2017	50
TOTAL	8 Groups	756

Then after we had run the script, we were able to see our results in the metadata manager. Here the users can see a neat and organized cluster of files instead of a gigantic list of unorganized files. These databases are immensely large, including up to thousands of files, so it's vital that the information be put out to users in a very organized and readable format.

4. **My Archiving Process**

Alas, it was my turn to put what I had learned with Bob to the test! I was tasked to archive some annual data recently released by the HadISD, "*a global sub-daily dataset based on the ISD dataset from*

NOAA's NCDC. As well as station selection criteria, a suite of quality control tests has been run on the major climatological variables."

You can check out their website and data here: https://www.metoffice.gov.uk/hadobs/hadisd/

The first thing I did was to take the 2019 files, download them onto my repo on the Cheyenne supercomputer, and then start to unzip them all to convert them into their bare bones file type. To download them, I had to use commands on my terminal.

```
mpasha@cheyenne3:/glade/work/mpasha> ls
WMO_030000-049999.tar.gz  WMO_100000-149999.tar.gz  WMO_250000-299999.tar.gz  WMO_450000-499999.tar.gz
WMO_050000-079999.tar.gz  WMO_150000-199999.tar.gz  WMO_350000-399999.tar.gz  WMO_500000-549999.tar.gz
WMO_080000-099999.tar.gz  WMO_200000-249999.tar.gz  WMO_400000-449999.tar.gz
mpasha@cheyenne3:/glade/work/mpasha> wget -O WMO_000000-029999.tar.gz "https://www.metoffice.gov.uk/hadobs/hadisd/v311_202001p/data/WMO
_000000-029999.tar.gz"
--2020-02-20 10:05:13--  https://www.metoffice.gov.uk/hadobs/hadisd/v311_202001p/data/WMO_000000-029999.tar.gz
Resolving www.metoffice.gov.uk (www.metoffice.gov.uk)... 23.77.91.112
Connecting to www.metoffice.gov.uk (www.metoffice.gov.uk)|23.77.91.112|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 879046954 (838M) [application/x-gzip]
Saving to: 'WMO_000000-029999.tar.gz'

100%[=======================================================================================>] 879,046,954 2.87MB/s   in 4m 52s

2020-02-20 10:10:05 (2.87 MB/s) - 'WMO_000000-029999.tar.gz' saved [879046954/879046954]
```

This is an example of me pulling a tar file from their data and downloading it into my personal repo. After I downloaded all of the 2019 files and had them all ready to go in my folder, it was time to take all the file extensions off. A lot of these data files are heavily compressed together, so even unzipping and untarring them takes a while! After a few quick commands were written to strip the files of the compressions, it was time to start archiving the data. For this, I had to make sure one archiving command worked on the data, and if it did, then I could put that command into a script and run it for all the other files.

```
#!/bin/bash

for file in *.nc;
do
dsarch -ab -ds ds463.4 -lf $file -mf /FS/DECS/DS463.4/$file -wf $file -df CFNetCDF -gx -md
done
~
```

This was the script I ended up writing! It was small, but it took each file ending in .nc and archived that into the metadata manager. I then ran it, and after debugging the errors, I finally got it running! It was really amazing to see it working and doing the archiving correctly.

5. Concluding Thoughts

Coming into this rotation, I had no idea what to expect! I had such a narrow scope of what data engineering entailed, and came into it expecting everyone was writing SQL queries against some kind of database and pulling information from it. However, I quickly came to learn that data engineering encompassed so, so much more than just querying a database. To me, it was almost like an art of sorts, learning how to properly display data to users in a readable way, how to download such huge amounts of data, how to group data accordingly, how to write scripts to make the archiving process faster and more efficient… it really opened my eyes to the immersive opportunities within this field, and even though my time spent working on the project was small, I definitely feel as if I learned a lot from all of this.

Muntaha Pasha
03/02/2020
Rotation 3 - DECS

A huge thank you to Bob for being my mentor in this, and patiently guiding me through all the little bumps in the road in regards to permissions and computer errors! It's truly been an experience I'll remember.