DDWG meeting Minutes: May 31 2022 12:45 UTC

Past year summary

Kevin Krieger gave a quick summary of the past year's activities. For the presentation, see pre-recorded videos on the workshop website. One question came up regarding 'slices' and how files are produced for Borealis style radars. Each slice in an experiment will produce its own data file. The first slice of an experiment will produce a file with 'a' as the identifier, the second slice will produce a file with 'b' as the identifier, and so on. Please see this link which provides extensive documentation and examples for how this works.

Overview of the new NSSC data website

An introduction and examples were given for the new NSSC website used for data distribution. Features include batch downloading of files, requesting files from PI groups using a web page form, and data inventory and metadata views. Please see: https://superdarn.nssdc.ac.cn/

There was some discussion about the PI permission requests seen here under 'Request Data':

The Request data form will only email the PIs whose radar data is within the request, and it seems like a useful common place for all approvals.

File inclusion standards discussion

There currently exist differences between USASK, BAS and NSSC mirrors due to different file inclusion standards.

- Currently, USASK is utilizing backscatter to check each and every DMAP formatted files before placement on the mirror.
- NSSC is utilizing pyDARNio to check DMAP formatted files before placement on the mirror.
- BAS currently has no DMAP file format checking. The IT group at BAS has not upgraded the operating systems on their virtual machines, which is a prerequisite to installing python so that either backscatter or pydarnio can be run.

Simon: Has built a C program to do DMAP file checking after encountering a set of JME files that broke a server due to analysis software requesting much more RAM than was necessary or available. Perhaps it is possible for BAS to install this software?

Various points were made regarding when, where and what types of file checks should be done:

- Simon: A policy/procedure should be created for both file checking as well as for files that are changed on mirrors so the changes propogate to everyone
- Mikko: What about only staging good quality data files? Should we require radar PIs to implement checks?
- Paul: Mirrors still need to intervene manually, but if PI groups were to do checks then this would reduce the frequency of issues that come up that mirror operators need to deal with.
- Simon: Can checks be automated in RST?
- Evan: Worries about coordination of software between all PI groups, as it is already difficult enough between USASK, BAS, NSSC
- Mikko: As a radar operator: would like to know if there are issues with radar data ASAP, so will be implementing checks regardless of what the mirrors do.
- Evan: There's another example of files that have issues that are not caught by these checks: DAT files with timestamps that are out of order
- Simon: Falls under this working group to analyze the files for issues like this, shouldn't be knowingly putting files with issues into distribution
- Kevin K: Believes the DDWG is responsible for checking file integrity, not the content/data in the files - this should fall to the PI group supplying the data
- Kevin S: Do we really know what 'good values' are for file data? What about the data standards working group - checking file data/content seems like it should fall under that group
- Mikko: As radar operators, we should be responsible for providing data that is good.
- Simon: There may be too much onus on the PIs to check their data, too much effort.
- Paul: One issue is that many processes are automated and might not stop issues in time so files would end up in the distribution before anything can be done. We should strive for not putting known bad data on the mirrors.
- Jianjun: For ZHO: due to the network limitations, data will be delayed 1 month - we check the quality before transfer as to not waste bandwidth on files that have issues.

A summary of the discussion

Where should files be checked?

Files could be checked at PI institutions, and/or at the mirrors. There are pros and cons to both. If files are checked at PI institutions - then faulty files would not end up at the mirrors, where they are currently placed in a parallel directory to the main distribution - this would preclude anyone from attempting to work

with the faulty files, and potentially fixing them (sometimes there is simply one record that causes the whole file to be flagged as faulty). However, reducing the amount of faulty files that end up at the mirror institutions could reduce the amount of work required to manage the mirrors, as well as allowing the PI institutions to be aware of any potential data issues from their radars more quickly.

What checks should be done?

Currently, the only checks available to be done are to determine if the files are readable, not corrupt, and that the DMAP file structure is self-consistent. No checks are done on the actual data itself. A majority of opinions heard agreed that this was appropriate, but it would also be possible - given direction and consensus, to do various checks on the data itself - one example brought up was to check for timestamps that are out of order. The premise being that no known bad data should be distributed.