DDWG meeting minutes: May 17 2023 12 UTC

Day	Date	Time	Location	Next Meeting
Wednesday	May 17 2023	12:00 - 13:15 UTC	Zoom	TBD

Attendees

Name	Institution	Shorthand
Kevin Krieger [Organizer]	USASK	KKr
Simon Shepherd	Dartmouth	SS
Evan Thomas	Dartmouth	ET
Akira Sessai Yukimatu	NIPR	ASY
Paul Breen	BAS	PB
Tim Barnes	BAS	TB
Jianjun Liu	PRIC	JL
Judy Stephenson	UKZN	JS
Nozomu Nishitani	ISEE	NN
Tim Yeoman	Leicester	TY

Summary

The meeting was held from ~12:00UTC until 13:15 UTC May 17th 2023. After some slides and updates from USASK, BAS, and UNIS, we discussed data file differences (and more specifically, how failed files are being handled) amongst the mirrors - this was the bulk of the discussion. After that, we discussed several other outstanding tasks such as the updated bistatic Dartmouth files, wallops file transfers, the outdated DDWG data flow figure, the inventory tool on the BAS website, the automated email notifications from BAS, and finally a note for future discussion on the sounding files from recent normalsound modes.

Action items from the discussions:

1. KKr will send out/make available the lists of file differences that are generated monthly between the USASK and BAS/NSSC mirrors.
2. DDWG needs to do some research on the data checking with pydarnio/backscatter
3. DDWG should collate and make available information about how each mirror checks files
4. The automated emails from BAS should be implemented for all data managers
5. The Wallops data flow needs to be set up to BAS
6. Bistatic files from CVE/CVW need to be looked at, determine if they need to be updated on the mirrors.
7. Existing data flow diagram needs updating
8. Discuss possibility of adding other data products to distribution, like sounding files that have no other home.

General Announcements

KKr went through several slides (available here: https://github.com/SuperDARN/DDWG/blob/main/reports/ddwg_2023_I

1. Data amount overview
2. Data gaps emails sent and some files retrieved from that
3. FRDR - Federated Research Data Repository
4. Files on github updated, if you have comments or need clarification let me know
5. Storage of 100TB received by USASK for the next year
6. Monthly automated comparison tool developed along with visual tool to see differences between mirrors

7. One unscheduled downtime at SFU impaced USASK mirror
8. One scheduled downtime in March of USASK mirror
9. Updates from Paul Breen at BAS
    - JME files now being downloaded from mirror
    - BPK updated, TIG,UNW removed
    - ICE, ICW being transferred
    - CVE, CVW re-enabled
    - FIR - some gaps from OCT and DEC 2022
10. Update for LYR:
    - Recommissioning starting very recently
    - Airport interlock system needs to be in place
    - Data flow to BAS
    - Data checking should be implemented (bzip2 compression, size, dmap checking using pydarnio)
    - Expects to be normally operated in August this year
11. Update from Fuli [APOLOGIES - I found these in my email junk directory]
    - From 22-10-01 to 2023-04-1, NSSC Mirror has synchronized a total of 1451.32GB of data from 10 radar stations from Bas, and a total of 1265.73GB of data from 15 radar stations from GLOBUS. 91.57GB of data was acquired and uploaded to the mirror from the JME radar station.
    - 756 files (from 2023 Feb. 9th to Mar. 31st.) were not synchronized on time from JME radar due to the transfer software unexpected exit. This problem was solved in April.
    - Upgrade Globus Connect Personal endpoints(s) to version v3.2.0 on December 12, 2022
    - On April 4, 2023, the synchronization of ICE and ICW radar station data from bas mirror station was added to NSSC mirror.
    - Update the strategy of data synchronization from BAS and USASK to synchronize their respective radar stations. (The original strategy was to synchronize the data of all radar stations from both sides, which could duplicate updates and possible updates to old versions of data.)
12. KKr went through slides to show new visual tool comparing mirrors
13. Went through visual tool that shows last two weeks of data from each radar

DDWG Discussion topics

Visual tool - differences between mirrors

Can we expose the file lists that generate the visual tool? Yes we can. KKr will do this at some point in the future. They can be hosted on the website superdarn.ca, or via the globus server.

Failed files? What are the reasons?

Question brought up - why are some files labeled as 'failed', but RST or dmapdump doesn't fail when making fitacf? The process was automated to generate the failed files. They would need to be handled on a case-by-case basis to determine what exactly is wrong. There may be differences between pydarnio and RST for example.

Need to get to the point to where we all agree on what file checks are being done. The failed files list is available on the globus server (all_failed.txt) which contains an error message from the tool used to check the file.

RST has had some bug fixes in the last few years, so some of the files may now not be problematic.

Can the failed file error message be put in the tool tip on the visual tool?

BAS still can't run python, but governance would be useful to implement a policy. There may be risk of moving reasonable data out of the distribution.

A fine line between removing too much data and not having enough quality control.

All files that are marked as failed are still available on the globus mirror, in a parallel directory to the mirror. Up to a user to do their due diligence to find

Everyone has to agree on the policies to use to exclude/include data. Need to have consensus. Should the PIs have this call, and the DDWG should implement?

Should be putting together some prep-work to prepare information before sending up a question to the PIs. Worries that if it is kicked up to the PIs, then nothing will happen. Need to have something in hand before asking the PIs. Sounds like it needs a larger discussion beyond the DDWG, but should start the discussion here.

Need someone to do the work of finding out why the errors occur. Need to look deeply into pydarnio. KKr sends out list of files that are failed to each of the data managers during the gaps emails.

One thing we can do is collate the list of which mirrors implement which checks. Need to find good place to hold this information for the BAS/USASK/NSSC mirrors.

Are NSSC and BAS the same? NSSC also is using pydarnio to check files, so it is likely closer to USASK than BAS right now.

Main points so far on discussion:

1. Need to do more research on pydarnio and what it is doing
2. Collate a list of failed files and error messages and send out
3. Ask PIs for some volunteers to look through failed files in a systematic way
4. Documentation about what is happening
5. DDWG should start discussion first, do some research before posting something to PIs.

Pydarnio doesn't go into the parameters - it doesn't look into the content of the files but rather the structure of the DMAP format.

Bistatic updated files from CVE/CVW

2019 through 2020? Files that were from receive only mode needed updating - so Simon/Evan updated these files, they still need to be updated on the mirrors which will be a manual process. List sent out in late 2022. Paul has had issues with receiving emails from the mailing list.

Can be sorted out by looking at logs to determine if the files existing on the BAS mirror are old versions or not.

WAL data flow

APL data transfer server/flow has changed. Preference to start sending to BAS now. List of outstanding data flows had only WAL on it left to transfer to BAS. They prefer to push to BAS, Paul waiting on that. WAL was offline for several years, but up as of late 2022.

BAS automated data flow emails

Paul had implemented automated emails to data managers to let them know when their data was not flowing Only first cohort of data managers had been set up, second cohort had not been set up yet, on the todo list.

The threshold for when an email is sent is configurable and should be long enough to not be 'spammy', but short enough to know about the issue within a reasonable amount of time. Set at 30 days.

Dartmouth data flow

Simon currently taking care of a backlog manually, eventually wants to get this to an automated system. Putting up all data at one time would be too much data, would take a long time, so working on chunks.

It's good to know when large amounts of data are coming down the pipeline from any radar/server.

DDWG data flow figure

The existing data flow figure needs to be updated. An updated version will be useful to visually show how system works.

Data inventory tool on BAS website

Due to PYK/STO being decommissioned, BAS inventory tool removes the radars, but it will still be useful to show the historical data for all radars.

Sounding files

Something to think about and discuss in the future - sounding files (example: from Evan's recent normalsound experiments) to be placed on the mirror? Otherwise they may be lost to time.