**MWA Project Meeting – ‘Adapting to a post-limitless archive’ Discussion**

*The large MWA archive is an excellent resource, and we want to keep MWA data for as long as possible. However, the archive is near full and we need to make room for new observations. Are there options (other than deleting), to create space in the archive?*

**Questions to think about…**

* What are some options?

1. Fringe stopping
2. Processing the archive & replacing with averaged data?
   * MWA ASVO/Birli (corrections, averaging, MS or uvfits)
   * This is a major shift in philosophy & policy
   * Science impacts? Re-use impacts?
   * VCS, EoR, Transients?
   * Logistics/Resources
3. Compression (lossy/lossless)
4. Selective tile/coarse channel deletions
5. Build a culture of review & reporting data with low utility for delete
6. Something else???? (Please share!)

* For your science area, what are the pros and cons of having e.g. averaged data instead of raw data?
* We have limited compute and operations resources to implement new features. What features should we prioritise? E.g. imaging over averaging?
* How old is your oldest data? Are you concerned about data being deleted? Does the new data retention policy alleviate those concerns?

**Discussion notes/ideas**

* Keep time resolution for transient searches.
* Better fringe stopping through correlator itself
* If researchers could put data back into archive, could make reprocessing simpler
* EoR preference to keep spectral resolution – different use cases per science area
* Options for corrections (delays, moving sources etc) available when specifying observations. Dipole delays would also need to be corrected for in correlator if also using other instruments
* QU: Why wouldn’t we prioritise new data over old data? Focus on reducing size of new data coming in.
* Setonix is another compute resource that could speed up re-processing times
* Tape speed is bigger limiting factor. So we shouldn’t delay decisions on this front.
* What corrections need to happen pre-archive? (metadata) Those details are v important for EoR
* QU: Is any form of averaging better than deletion? (‘Infinitely!’)
* SKA Regional Centre activities highly related. Distributed archiving as a possibility? Having multiple accessible copies. Rucio as an option has been explored previously.
* MWA playing precursor role in Big Data era
* Re: fringe stopping. Variable delay rates great for Space Domain Awareness. Already done in code – not tested
* Average some sets and keep others as raw data depending on science case
* QU: Does new Policy cover calibration data? Very useful resource, support could be added explicitly.
* Ratio of archived data per observing mode
* QU: What happens in Data Retention Policy when PI is no longer part of the project? All MWA members can access all data >18mths, and can request retention, even if they weren’t the PI of the dataset
* Could publication policy be changed to capture obsID information?
* Will there be another ARDC round for ringfenced data collections? We’re definitely looking for all retention opportunities.
* *However*, onus for data retention is on researchers to comply with their institutes’ policies
* Recording voltage data is non-traditional for pulsar research, not sustainable. Could there be an option in proposal form where PIs state a timeline for keeping voltage data only on a short-term basis? For non-survey projects where interest is in limited number of targets. Interim strategy. -> We are then dependent on users to action things on a limited timeline, which can be difficult.
* Real-time beamforming is another data reduction option for VCS applications.
* Real-time calibration solutions for pulsar research
* Can MWA ops provide cost of Pawsey storage resources (from ARDC form) on wiki – yes!
* Scoping the problem – data quality and data access metrics? Using FAIR principles as guidelines for MWA researchers in data storage
* Current issue - no established feedback mechanism for researchers to indicate which files/observations were useful from the archived dataset they pulled
* Suggestion for flagging ‘good’ data as well, via ranking system
* Downloaded data > data that has never been touched? Difficult to determine pre-2018
* Ionosphere quality tables exist from EoR RTS – could these be expanded on by other science groups?