ATNF Data Archives and OPAL

For ATUC

Minh Huynh and Andrew O’Brien | Mar 2023

Australia’s National Science Agency

Recent CASDA Developments

- Containerisation of deposit modules, for move to Setonix (data mover nodes)
- Speed up of validation and release tasks, and ability to delete unreleased observations
- Completed CARTA proof-of-concept installation, but a solution with better I/O is needed (probably Setonix data mover node)
- UI/UX usability study project completed.

CASDA Work: Current Big Ticket Items

- Migrate deposit modules to new servers (*Setonix data mover node*)
- Installation and integration with CARTA *(Setonix data mover node)*
- ATOA analysis and migration to CASDA
- UI-UX refresh *(implement recommendations from study)*

CASDA Work: Other Requests

1. Improve validation workflow and allow validation updates
2. Include beam tables in CASDA cutouts
3. Improve download speeds
4. UI search: show integration time and ability to filter by Stokes/pol
5. Custom DOIs
6. Derived data/L7: Ability to deposit FITS tables and other file types

Parkes Pulsar DAP

- UX/UI usability study of Pulsar domain complete, IM&T have report
- Move to S3+tape storage planned
    - Better access to data across collections
    - Increase DAP ingest rates
        - Better accommodate large UWL+CryoPAF data rates. But future DAP ingest rate unclear.

- What to archive with CryoPAF? (max data rates of up to ~9TB/hour in extreme search mode)
- Already unable to archive all UWL datasets (users must specify an endpoint for projects > 10 TB/sem)
- Working on data storage and access policy required for CryoPAF

ATOA Migration to CASDA

- BIGCAT and CryoPAF will increase data rates in 2023+
  - ATCA BIGCAT up to 0.5 PB per year, nominal max rate 100MB/sec
  - Parkes UWL and CryoPAF continuum/spectral data 100 to 200 TB(?) per year
    - UWL is currently ~20 TB per semester to ATOA
- High Level Requirements:
  - Meet the future Parkes and ATCA data rates
  - Have same download and search capability as current ATOA
  - Data must be served in a manner meeting ATNF Data Policy (permissions)

ATOA Staged Migration to CASDA

Stage 1: take incoming (raw) new data, i.e. allow for switchover to CASDA

- ATCA (BIGCAT)
- Parkes spectral line and continuum (UWL and CryoPAF)
- VLBI (correlated data, FITS-IDI)
- **Planned completion: end of 2023**

Stage 2: migrate current archived data

- Migrate ATCA, Parkes and VLBI (raw) data currently in ATOA
- **Planned Completion: mid 2024**

Stage 3: migrate remaining data

- Science data from Mopra (e.g. MALT90), HI Surveys?, RFI data, others?
- Need to do an assessment of best place for this data, may not be CASDA
- **Planned Completion: mid 2024**

Questions from ATUC

Will the new ATOA in CASDA serve calibrated and processed datasets?

- Not initially. Potential for calibrated and processed data sets in future if there are resources and developed pipelines.
- Compute at Pawsey could be requested to support ATOA users.

Questions from ATUC

ATOA data is initially accessible via an internal machine (e.g. Carina/Kaputar) where the raw dataset could be partially processed and reduced in size by the project teams, before data transfer **(i.e. bypassing ATOA!).** Will such an interim processing capability be available in future?

- ATOA data will still pass through Marsfield (at least initially), so expect data will be still available via Kaputar/Carina.
- But is compute+disk at Marsfield sufficient?

Questions from ATUC

What are the direct download limits? Limitations on data transfer to users?

- CASDA has a default 0.5 TB web download limit, but individual users can be set to unlimited
  - Practical issue that 1 TB can take a day or more to transfer now
- Flagged data transfers with both CASDA and Pawsey teams
  - Need to do some benchmarking and testing
- GLOBUS could be an option
- **Should reduce data transfers as much as possible by using Pawsey/Marsfield compute**

Questions for ATUC

1. How would you like to search for data in ATOA/CASDA? Will you mainly be downloading data from an individual telescope/project, or data from all telescopes for given params (e.g. position), or a combination of both?
2. Large data rates will take a long time to transfer, as well as ingest into ATOA/CASDA. What is an acceptable rate? E.g. continuum observation in ATOA/CASDA within 24 hours? (~10 GB/hour observation available within 24 hours)

OPAL Updates

OPAL updates since Oct 2022

- Security and bug fixes
- Improvements to development processes
  - Easier testing, streamlined deployments
- Coversheet changes
  - Simple version warnings
  - Added “Observing Experts”
  - Parkes: added fields for expected data rate, volume, and end-point
- TAC interface changes
  - Proposal investigator anonymisation
  - Hide normalised scores
  - Other minor fixes

Planned OPAL updates

- Backlog of 53 issues
- Major updates planned for 2023OCTS
  - ASKAP guest science proposals
  - Unicode support
  - Updates to scientific keywords
- Other planned major updates
  - Long-term project proposals
  - Observation table upgrade
  - Expand internal API

Thank you

OPAL feedback welcome at **atnf-datsup@csiro.au**

Australia's National Science Agency

Thank you

CSIRO Astronomy and Space Science
Minh Huynh
Senior Data Scientist and Astronomer, ATNF Science Group Leader

+61 8 6436 8696
Minh.Huynh@csiro.au

Australia’s National Science Agency