Data Access and Archive Developments

For ATUC

Minh Huynh | Apr 2024

Australia’s National Science Agency

ATNF Archives: Current State

- Parkes Pulsar Archive (Data Access Portal, CSIRO Canberra Data Center)
  - Parkes Pulsar data
- CASDA (DAP, Pawsey Supercomputing Research Center)
  - ASKAP SDP (ASKAPsoft) outputs, and ASKAP SST valued-added (L7) data
- Australia Telescope Online Archive (ATOA, Marsfield)
  - ATCA, Parkes spectral line and continuum data
  - LBA data (not ingested since late 2022, FITS-IDI parsing issue)
- Classic/normal Data Access Portal (CSIRO Canberra Data Center)
  - CSIRO's enterprise-wide system
  - User derived data, e.g. pulsar timing products, ATCA science data products, ASKAP L7 data (SPICE RACS)

ATNF Archives: Future State

- Parkes Pulsar Archive (Data Access Portal, CSIRO Canberra Data Center)
  - Parkes Pulsar data

- CASDA (DAP, Pawsey Supercomputing Research Center)
  - ASKAP SDP (ASKAPsoft) outputs, and ASKAP SST valued-added (L7) data
  - Australia Telescope Online Archive
    - ATCA, Parkes spectral line and continuum data, LBA

- Classic/normal Data Access Portal (CSIRO Canberra Data Center)
  - User derived data, e.g. pulsar timing products, ATCA science data products, ASKAP L7 data (SPICE RACS)
  - ATOA science-ready datasets, e.g. Mopra surveys such as MALT90

Recent updates and Development Priorities

Parkes Pulsar Archive: recent updates

- DAP developments:
  - Recent move to S3 storage
  - GLOBUS integration
  - Ingest rate has increased, but publishing rate still an issue
- For projects > 10 TB, still taking weeks to get data to overseas institution, days to get to a local/Aust one
- Trialing AWS to get data to users

Parkes Pulsar Data Access Options

CryoPAF pulsar (search) data rates are a challenge (potentially ~100 TB per month)

Options:

- Push to user-defined endpoint (being done now)
  - Amazon AWS cloud S3 (ingress cheap, egress ~$100/TB, user pays?)
- CSIRO proto-platform (AWS, EASI) being explored
- Set of default pipelines at Parkes, pre-process data for users?
- Need to ensure pulsar projects can be completed – carefully consider CryoPAF pulsar search time allocation

CASDA Recent Work

- Integration of CARTA (Oct 2023 release)
- ATOA migration
  - MVP to support LBA, UWL/CryoPAF and ATCA BIGCAT
  - Work on deposit and data access sides
- Acacia storage failure mitigation in Feb
- Various bug fixes

CASDA-ASKAP Development Priorities

- Data retention resilience: options being developed
  - E.g. Copy off site all primary images/cubes, catalogues, L7 data (20%, > 1 PB/yr)
- Software maintenance and performance on Pawsey infrastructure
- CARTA enhancements (e.g. use of fits2idia)
- Download enhancements: GLOBUS, direct S3 bucket access
- VO and TAP upgrades
- FITS table support for catalogues
- Improved L7 user-derived data support for incremental deposits (collection updates)
- Global catalogues assessment, replace with RACS/EMU
- Custom DOIs

ATOA Staged Migration to CASDA

Stage 1: take incoming (raw) BIGCAT, UWL/CryoPAF spectral-line+continuum, and LBA data

- MVP development of CASDA deposit module and data access/UI complete
- LBA data Dec-2021 in process of being deposited into prod, ETA very soon/imminent
- Tests on Parkes and BIGCAT data ongoing, to be ready for commissioning later this year

Stage 2: migrate existing ATOA (~ 500 TB)

- Parsers need to be written, substantial amount of work
- Expect migration to start in a few months, finish end of 2024
- **Planned Completion: mid/late 2024**

Stage 3: migrate science data (Mopra surveys etc)

- Need to do an assessment of best place for this data, probably normal/classic DAP
- **Planned Completion: late 2024**

CASDA-ATOA Development Priorities

- Ongoing support:
  - Issues and bugs will only become clear once BIGCAT and CryoPAF are online

Other Priorities:

- Data resilience, copy off-site of ATOA data
- Ability to download portions of a dataset (e.g. channels or beams of interest)
- Full integration into CASDA and use of existing functionality (e.g. VO and TAP services, python astroquery)

Open questions / other considerations

- Pawsey infrastructure and performance
  - Need to meet increased deposit and data-access loads with ASKAP+ATOA
  - Contention with other workflows at Pawsey

- CRACO
  - What to archive in long term?
  - What do the space weather community need in terms of functionality and data types?

- Do the development priorities align with ATUC's thinking?

Summary

- Major changes to ATNF archives underway to support new instrumentation
  - Consolidation into DAP/CASDA

- Data access and archiving is a crucial element of the National Facility

- A significant body of work has been identified, but resourcing going forward is unclear. ATUC support and strong statements will help.

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