The Future of Archiving

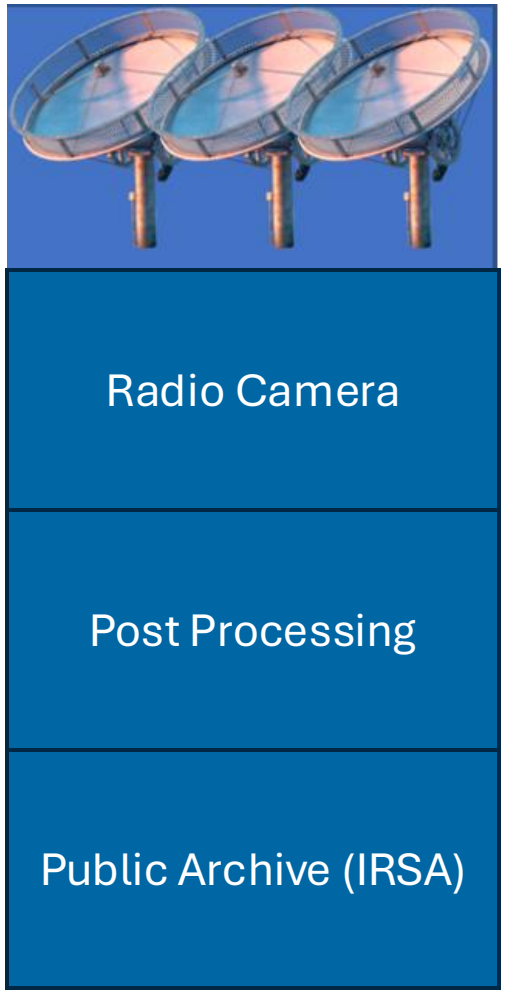
Rachel Akeson, Caltech/IPAC

The Future of Archiving

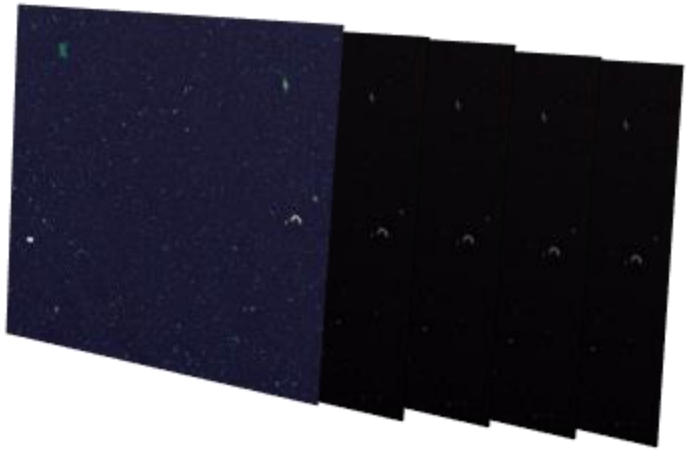
Rachel Akeson, Caltech/IPAC

DSA-2000, ngVLA and the era of surveys and big data

DSA-2000 Data Flow



- Release of processed data products every ~4 months
  - **95 TB** all-sky continuum images
  - **1 TB** of extracted pol cubes
  - **21 TB** of extragalactic HI cubes
  - **100 TB** of Galactic HI cubes
  - **100 TB** of pulsar folded profiles
  - **10 TB** FRB positions, spectra
- IPAC will produce a source catalog with ~1 billion sources



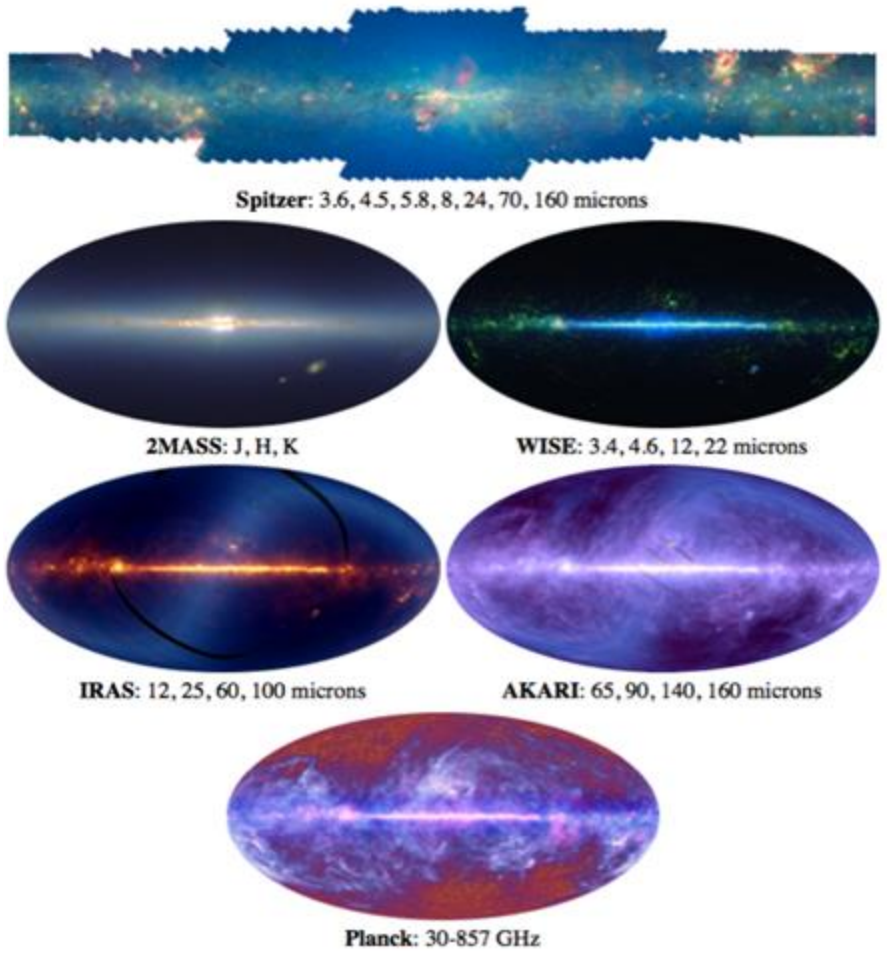
Schmidt Sciences

DSA-2000

Caltech

DSA-2000 data at IRSA

- DSA-2000 public data will be hosted at the NASA/IPAC Infrared Science Archive (IRSA)
  - Current IRSA holdings include all-sky coverage in 29 bands from IRSA, WISE, 2MASS, Planck and other missions
  - Also time domain images and catalogs from ZTF (Zwicky Transient Facility)
  - https://irsa.ipac.caltech.edu
- Standard IRSA capabilities
  - Includes search, visualization and download capabilities for images and catalogs
  - Web-based and application programming interfaces (APIs)
  - Data products (images, spectra, catalogs) will follow IVOA data access and metadata standards



Plus SPHEREx and Euclid in 2025

ngVLA "High Level Data Product" (HLDP) & Archiving Concepts

- Provide HLDP pipelines (=processing workflows) for "standard mode" observations (>80% of all observations)
  - Archive Raw visibilities, calibration/flagging results & HLDPs (e.g., images, cubes)
  - Calibrated visibilities available and generate upon request
  - *Exactly what HLDPs will be produced for each mode is still being defined (see Wilner et al. 2024*, ngVLA Memo #125)*
- "Resource limited operations": Cannot afford the compute/storage to produce full FOV images & cubes for all cases
  - Instead, collect information from PIs in proposal that informs what HLDPs to create (e.g. limited FOV, spectral grasp)
  - Make some standard products from most PI observations
- Enable sub-setting (spatial cutouts, binning, averaging, spectral subsets) to reduce download volume
- Provide capable "reimaging" facilities for PIs & Archival researchers:
  - Produce and archive modified or new HLDPs from archived visibilities (e.g., re-weighting, unimaged fields, etc.)
  - Resources will be limited; some requests may require separate reprocessing proposal
- Access will be through NRAO Archive: Will include ability for cross-facility queries & data access
- Average Data Rate: 7.6 GB/s → 240 PB/yr

** Wilner et al. 2024 "Key Science Goals for the ngVLA: Update from the ngVLA Science Advisory Council (2024)" https://library.nrao.edu/public/memos/ngvla/NGVLA_125.pdf, https://www.arxiv.org/abs/2408.14497*

Key Requirements (ROP)

Source: ngVLA Computing Memo 11, R. Hiriart

Visibility Data Rate	Imaging Data Rate
Average Data Rate: 1.93 [GVis/s]	Average High Level Data Product Data Rate: 0.66 [GBytes/sec]
Average Data Rate: 15.48 [GBytes/s]	Average Image Size (1 Plane): 57.07 [Mbytes]
Average Data Rate: 40.11 [PBytes/month]	Max Image Size (1 Plane): 307.37 [Mbytes]
Peak Data Rate: 33.46 [GVis/s]	Average Cube Size: 2.38 [TBytes]
Peak Data Rate: 267.69 [GBytes/s]	Max Image Cube Size: 37.70 [TBytes]
90% Quantile Data Rate: 2.01 [GVis/sec]	Average Number of Pixels: 595.09 [Gpixels]

Key Requirements (ROP)

Source: ngVLA Computing Memo 11, R. Hiriart

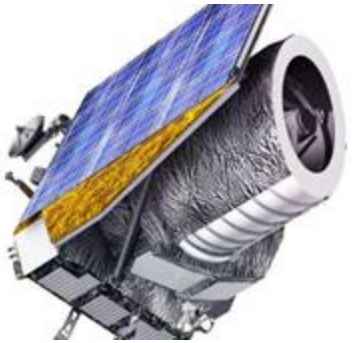
Visibility Data Rate	Imaging Data Rate
Average Data Rate: 1.93 [GVis/s]	Av [G
Average Data Rate: 15.48 [GBytes/s]	Average Image Size (1 Plane): 57.07 [Mbytes]
Average Data Rate: 40.11 [PBytes/month]	Max Image Size (1 Plane): 307.37 [Mbytes]
Peak Data Rate: 33.46 [GVis/s]	Average Cube Size: 2.38 [TBytes]
Peak Data Rate: 267.69 [GBytes/s]	Max Image Cube Size: 37.70 [TBytes]
90% Quantile Data Rate: 2.01 [GVis/sec]	Average Number of Pixels: 595.09 [Gpixels]

4 hr observation ~109 TB.

Requires ~1000 cores to process in a few days

DSA-2000 and ngVLA will operate in an era of surveys and big data

Euclid: 2023



Rubin: 2025

Sphere-X: 2025

Roman: 2026

UVEX: 2029

And many more!

- Many science cases will require observations from multiple wavelengths and/or facilities
  - It is not realistic for most users to have full copies of all these data sets

Interoperability

- Many archives, including IRSA and NRAO, use International Virtual Observatory Alliance (IVOA) standards to enhance interoperability
  - For example, Table Access Protocol (TAP) queries can be used to search observation metadata across archives
- Data products (images, spectra, catalogs) which use IVOA data access and metadata standards are accessible with many tools:
  - IVOA-aware GUI tools (e.g. TOPCAT, Aladin, IRSA Viewer, MAST Portal, Xamin, Rubin Portal).
  - Python libraries pyVO and astroquery
- Users can search/retrieve from multiple archives using same query mechanisms

The future of interoperability

- **Big Data** challenges current interoperable workflows. The community is responding with:
  - cloud storage and computing
  - cloud-friendly access standards/tools, e.g. S3 pointers, cutouts without download
  - cloud-friendly format standards, e.g. Parquet for catalogs, with associated tools
  - (increasingly cloud-based) server-side analysis (e.g. Fornax, SciServer, Astro Data Lab, CANFAR, TIKE)
- **Big Time Domain data** presents additional specific interoperability challenges that are being addressed by the IVOA's Time Domain Interest Group: https://wiki.ivoa.net/twiki/bin/view/IVOA/IvoaVOEvent
- **Big Radio data** is being discussed in the IVOA's Radio Interest Group: https://wiki.ivoa.net/twiki/bin/view/IVOA/IvoaRadio
- Participation by the community is essential to make interoperability a reality!