VERY LARGE ARRAY SKY SURVEY

Observing, data products, data management, resource requirements

Steven T. Myers (NRAO Socorro)

Galactic Center (Survey) Multiwavelength Image
Credit: X-ray: NASA/UMass/D.Wang et al., Radio: N
RAO/AUI/NSF/NRL/N.Kassim, Mid-Infrared: MSX

Outline

- Observing
- VLASS Data Products
- Data Storage and Archive Server
- Processing pipelines
- Resource Requirements and Allocations

Observing Schedule

- 7+ years spanning 6 cycles (88 mos. 2016-23)
  - Impact: 1480 hrs/cycle

Cycle	Config.	ALL-SKY	COSMOS	ECDFS	Elais-N1	Total
1	B	906				906
1	A		60	325	188.5	573.5
2	B	906				906
2	A		60	335	188.5	573.5
3	B	906				906
3	A		60	325	188.5	573.5
4	B	906				906
4	A		60	325	188.5	573.5
5	B	906				906
5	A		60	325	188.5	573.5
6	B	906				906
6	A		0	335	188.5	523.5
Total		5436	300	1960	1131	8827

Scheduling

- Scheduling considerations
  - where practical use “fixed LST blocks” as long as possible
    - ECDFS is essentially a fixed block (narrow LST window)
    - COSMOS & Elais-N dynamic (modest LST window)
    - ALL-SKY most easily scheduled as an ordered series of granular (~4 hr?) blocks, each with narrow LST start range, or longer 6-hour blocks
  - self-contained calibration (factored into overheads)
    - for ALL-SKY primary calibration every ~8-12 hours (2-3 blocks)
- Schedule creation
  - will use Python script generation (e.g. as in CNSS Stripe-82)
  - these text files can be read into OPT and submitted as usual
    - NOTE: if you don’t know about this option ask us!
- Observing
  - blocks interruptible (for TOO alert responses, overrides, etc.)
    - modularity of ALL-SKY blocks planned

Calibrators

- Fundamental
  - observe standard flux density & polarization angle calibrators
    - 3C286 is fundamental standard
  - polarization leakage using low-pol or tracked through P.A.
    - can do low-pol in blocks, but leakage expected to change only slowly
  - nearby gain calibrators
    - coarse grid of astrometric (VLBI) calibrators
    - link to finer grid of secondary bright (>100mJy) calibrators
    - need a good list, ideally do a search before scheduling (T&DP)
    - after observing starts, ALL-SKY will generate a really good list!
  - self-calibration
    - done after observing, but we expect to hit a useable source (>6 mJy) near beam center every 16 seconds to 1 minute!
    - this “global sky model” will be tremendously useful

Observing Management

- A day in the life of the VLASS
  - carried out by “Observer of the Day” and Data Analysts
  - schedules pre-generated
  - schedules queued (previous day)
  - appropriate schedules observed (in proper order if necessary)
  - QL calibration run, QA assessment
    - to look for big issues, possibly to tune parameters for imaging
  - QL imaging run, QA assessment
    - to look for errors that would prevent fast staging in to QL archive
    - might require more careful hands-on processing
  - when competed & passed, stage cal data and images to archive
    - requirement is 48 hrs after observing
    - to give leeway for re-processing would like to be able to complete within 24 hours on average

VLASS Basic Data Products

- Deliverables by NRAO (with SSG collaboration where possible):

Product	Timescale	Notes
Raw Data	immediate	no proprietary period
Calibrated Data	1 week	same, served from archive
Quick-Look Images	48 hrs.	continuum only, simple QA
Quick-Look Catalog	w/QLI	only basic image object finding
Single-Epoch Images	6 mos. (T1) 12 mos. (T2,pol)	better quality assurance
Single-Epoch Catalog	w/SEI	more object parameters
Cumulative Images	12 mos. (T1) 16 mos. (T2,pol)	produced after each epoch after first
Cumulative Catalog	w/CI	more detailed

VLASS Basic Data Products

- Deliverables by NRAO (with SSG collaboration where possible):

Product	Contains
Raw Data	full dataset
Calibrated Data	calibrated dataset (or raw+tables)
Quick-Look Images	I images (continuum only)
Quick-Look Catalog	flux density,position,size
Single-Epoch Images	IQU (continuum,spectral cubes*)
Single-Epoch Catalog	flux density,position,size,spectrum
Cumulative Images	IQU (continuum,spectral cubes*)
Cumulative Catalog	flux density,position,size,spectrum

* spectral cubes may be limited/compress (see TIP)

Raw and Calibrated Data

- Raw visibility data
  - ALL-SKY: 5436 hours @ 25MB/s = 489 TB
    - this high sustained data rate necessary for accurate OTFM
  - DEEP: 3391 hours @ 25MB/s = 305 TB
    - this assumes observed in same mode as ALL-SKY (0.45 sec)
    - beneficial for uniformity of data for survey
    - is possible to reduce rate, e.g. use 1-2 sec integrations
- Calibrated data
  - we do not plan to store a copy of the data with the calibration applied (this would require almost twice the data storage)
  - we plan to provide the flagging & calibration tables and prescription (script) for application to the raw data (both QL and final calib.)
    - could be applied by the archive server or user
    - similar to process used by ALMA

Image Sizes

- ALL-SKY (~2.5") at 0.6" (36Mpix/deg^2) :
  - 34000 deg^2 : 1.22Tpix = 4.9TB per "image"
  - Continuum : 9 images = 44TB (7 images, 3 epochs = 100TB)
  - Spectral Cubes : (1024ch, 5 images = 25000TB = 25PB)
    - 5.5 Ppix! This would be a lot of image pixels to sift through!
    - TOO MUCH FOR ARCHIVE AS PLANNED! must compress <1%
      - "postage stamp" cutouts, channel averaging, processing on-demand (POD)
- DEEP (~0.8") at 0.2" (324Mpix/deg^2):
  - 10 deg^2 : 3.24 Gpix = 13GB per "image"
  - 10 deg^2 in 2390 passes x deg^2 : 774 Gpix = 3TB
  - Final Continuum : 9 images = 116 GB
  - Spectral Cubes: (1024ch, 5 images = 66TB)

Continuum Images

- Images at relevant resolution (0.2” or 0.6”)
  - Stokes I,Q,U
    - no Stokes V planned for BDP, squint correction and alterations to calibration would be more costly
  - MFS Taylor-term images
    - spectral index α ($dI/d\ln\nu$) and curvature “β” ($d^2I/d^2\ln\nu$)
    - possibly stored as native clean output tt0 ($I*\alpha$) and tt1 ($I*\beta$) to avoid having to mask the divide by zero regions
  - Uncertainty maps (including mosaic weighting) for all images
    - plan to save single thermal “uncertainty” map for QU (it represents complex polarization map P=Q+iU), if needed store separate Q,U noise maps based on data
    - these would nominally be computed as part of imaging based on input data, but could also be (additionally?) computed empirically by processing the actual images
  - Up to 9 total continuum images

Spectral Cubes

- Full spectral resolution (1024 channels, 2 MHz resolution)
  - Stokes IQU
  - Uncertainty maps for I and P=Q+iU
  - up to 5 total maps
- Coarse resolution cubes (16 spectral windows, 128 MHz)
  - Stokes IQU
  - Uncertainty maps for I and P=Q+iU
  - up to 5 total maps
- Compression options for full (or high) resolution
  - “postage stamp” cutouts around bright sources
  - frequency averaged (e.g. in λ^2 for RM)
  - other compression options?
  - or, process on demand…

10M obj.
20” cutouts
40TB
~1% of total

Object Catalogs

- Basic Data Product catalogs
  - will evaluate source finding software/algorithms (e.g. AIPS SAD, Aegean, etc.)
  - See Hancock et al. 2012, Mooley et al. 2013 for more info
- Basic Fit Quantities (example):
  - Position, and uncertainty (likely centroid of I emission)
  - Peak Flux Density (continuum) in IQU, and uncertainty
  - Spectral Index at Peak (Stokes I) and uncertainty
  - Integrated Flux Density (continuum) in IQU, and uncertainty
  - Integrated Spectral Index (Stokes I) and uncertainty
  - Basic Shape information IQU, for example
    - ellipsoidal (Gaussian) bmaj, bmin, bpa or similar
    - concentration index or similar

Quick Look (QL) products

- To enable fast response transient science and enable quick turn-around for Quality Assurance and user science,
- Stokes I continuum images (and uncertainty maps), basic object catalogs (at higher flux cutoff, >10σ e.g.)
- Delivery: Calibration and Imaging completed, images available in archive within 48 hours following completion of Schedule Block observing
- Implementation: streamlined version of CASA JVLA calibration pipeline and a (new) QL imaging pipeline
- Fallback: Caltech-developed AIPS-Lite + CASA pipeline used for CNSS (Stripe-82) and COSMOS
- Q: provide for citizen science?

Single-Epoch (SE) products

- Calibrated data
  - our 1-week delivery of the calibrated data may be refined as we learn more (e.g. better RFI excision)
- Continuum images (Stokes IQU, α) and spectral cubes (IQU) and uncertainty maps (7 cont. + 5 cubes)
  - if Q,U measured noise separate (8 cont. + 6 cubes)
- Delivery: within 6 months (ALL-SKY: I, α) or 12 months (ALL-SKY: QU, all DEEP) of end of cycle observations
  - we anticipate having pretty good preliminary versions earlier (e.g. within a month) but budget 6 months as a conservative guarantee
  - Q: Would it be valuable and preferable to make these available as products earlier (with appropriate cautions) or wait for exhaustive tests?
  - Q: provide for citizen science ("RG zoo")?

Cumulative Final (CF) products

- Continuum images (Stokes IQU,α,β) and spectral cubes (IQU) and uncertainty maps (9 cont. + 5 cubes)
  - if Q,U measured noise separate (10 cont. + 6 cubes)
- Delivery: within 12 months (ALL-SKY: I,α,β) or 16 months (ALL-SKY: QU, all DEEP) of end of cycle observations after the first (keep only the most recent)
  - we anticipate having pretty good preliminary versions earlier (e.g. within a month) but budget 12(16) months as a conservative guarantee as with SE products
  - we would anticipate providing images from this suite as primary citizen science product

Archive & Data Storage

- In long-term planning for the JVLA, we project to have storage for 3300TB in 2017, and 5700 in 2020
  - Would like to use <10% for VLASS image storage in 2020
- In the TIP, we estimate VLASS usage as 1280 TB
  - Raw data = 794 TB
  - Image products = 486 TB
- This is 22% (8.5% for images) of the total archive in 2020
  - Can accomodate in projected NRAO DM budget plans
- But, this storage budget requires significant reduction or compression for the image spectral cubes
  - strongly impacts polarization (RM) studies
  - impacts line searches

Types of Images & Image Cubes

- pixel sizes set by properly sampling PSF
  - conservative: 0.2” (A) & 0.6” (B) ~ x4 oversampling of a good mid-band PSF (x3 at upper end)
- continuum images
  - single planes IQU (plus α and β from MFS)
  - uncertainty (noise) maps for each (Q+iU or Q,U)
- coarse spectral cubes
  - per 128MHz spectral window or similar
  - TIP assumed 14 spw useable, may actually get only 11 in many areas)
- full spectral cubes
  - assume 896 of the 1024 2MHz chans (14 spw)

ALL-SKY images

- 33885 deg^2 at 0.6” pixel size → 1.2Tpix (4.8TB)
- QL images
  - 2 continuum images (I + unc.) = 2.4Tpix (9.6TB)
- SE images (3 epochs)
  - 3 x 7 continuum images (IQUα+unc.) = 25.2Tpix (100.8TB)
  - 3 x 5 coarse (14 planes, per good spw) cubes = 252Tpix (1008TB)
    - too much! compress/cutout/drop spw ~ 10:1 to 100.8TB
  - 3 x 5 full cubes (896ch) would take 16.1Ppix (64.5PB) = NO WAY!
- CF images (best combined images)
  - 9 continuum images (IQU$\alpha\beta$+unc.) = 10.8Tpix (43.2TB)
  - 5 coarse cubes (14spw) = 84Tpix (336TB) – compress 10:1 to 33.6TB
  - 5 fine cubes (896ch) = 5.4Ppix (21.5PB)
    - compress 200:1 to 107.5TB

DEEP Images

- 10 deg^2 at 0.2” = 324Mpix/deg^2 → 3.24Gpix (13GB)
- QL images : 2390 deg^2 x passes = 774Gpix (3.1TB)
  - 2 continuum images = 1.55Tpix (6.2TB)
- SE images : 58 deg^2 x epochs = 18.8Gpix (75.2 GB)
  - 7 continuum images = 131.6Gpix (526.4 GB)
  - 5 coarse cubes x 14 spw = 1.3Tpix (5.3TB)
  - 5 full cubes x 896 ch = 84Tpix (337TB) – DO NOT INCLUDE
    - could consider highly compressed versions
- CF images : 10 deg^2 = 3.24GPix (13GB)
  - 9 continuum images = 29.2 Gpix (116.6GB)
  - 5 coarse cubes x 14 spw = 226.8Gpix (907GB)
  - 5 full cubes x 896 ch = 14.5Tpix (58TB)

Storage Requirements

Data Product	All-sky (TB)	Deep (TB)	Total (TB)	Comments
Raw data	489	305	794	at 25MB/s
Quick-look continuum images	29	6	35	I only
Single epoch continuum images	101	0.5	102	IQU
Single epoch coarse spectral image cubes	101	5.3	106	128 MHz resolution
Cumulative final continuum images	43	0.1	43	IQU
Cumulative final coarse spectral image cubes	34	0.9	35	128 MHz resolution
Cumulative final fine spectral image cubes	107	58	165	2 MHz resolution compressed
Total	**904**	**376**	**1280**	

Processing On-Demand (POD)

- This is a very attractive alternative to making and storing every possible image in the archive
- POD: use a mini-pipeline to create the requested images and image cubes from the calibrated when demanded (e.g. requested from the archive)
  - used for the old CLASS (8GHz VLA grav. lens survey of 10^4 sources total), we could make images faster than it took to observe them!
  - for VLASS will require significant (super)computing to make the turnaround fast enough (1 day?) to be practical
    - NRAO (J. Robnett) investigating NSF XSEDE to do this (using CNSS)
  - maintain smaller buffer of recent/frequent images
  - we think this is the way to go for VLASS & future
    - TIP estimate 2.1 FTE years for this

BDP Pipelines

- Quick Look (QL) pipeline
  - calibration/flagging (drastic editing for speed)
  - imaging (continuum only for BDP)
  - object finding & analysis
  - QA at each stage
- Standard Reduction (SR) pipeline
  - all the above, with more careful processing (e.g. optimized RFI)
  - ALL-SKY: add polarization, spectral index (& curvature), coarse spectral cubes (compressed), fine cubes (compressed)
  - DEEP: same, plus coarse & fine spectral cubes (uncompressed)
  - probably a more refined “guided” mode to handle high dynamic range and complex regions
- Make these pipelines available to users!

Calibration

- Data import, (Hanning smoothing), flagging, (antenna position), Tsys, initial phase, solve delays, bandpass, final phases & amplitudes, flux scale, (polarization delay, leakage, angle), final flagging, (apply, export to QLI)
- Implemented (w/o polarization) in AIPS-Lite & CASA custom script (Stripe-82) and in CASA production pipeline of NRAO
- CASA Benchmarks (in TIP using single CNSS 3h block):
  - custom CASA script = 10.8 hours (single node, single process)
  - CASA pipeline = 66.25 hours (mostly plotting, other bottlenecks)
  - expect to process in real time 3h multi-proc single node
  - parallelization of calibration important
  - AIPS-Lite much faster (more procs/ node)

Imaging

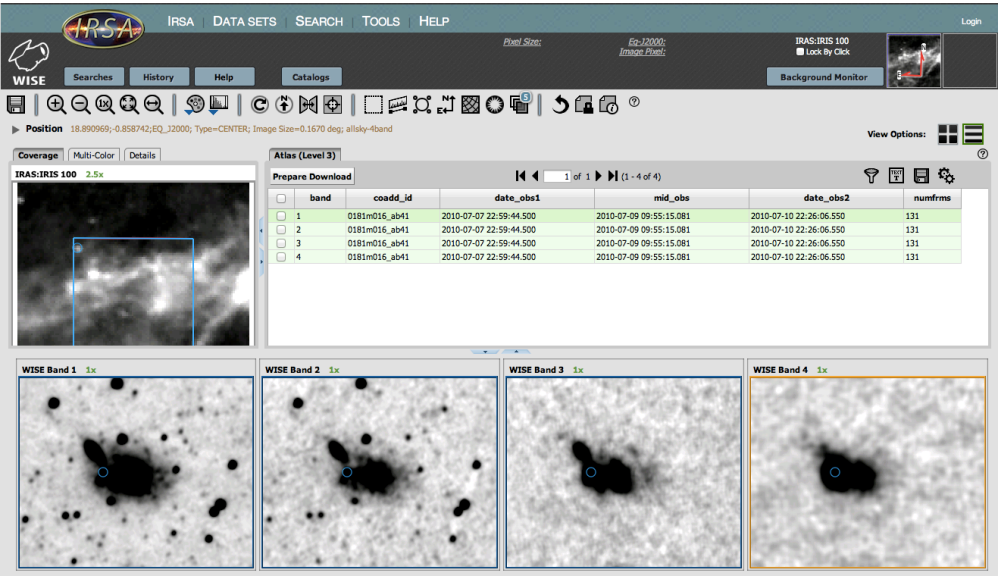
- image joint sub-mosaics, linear mosaic to full mosaics
  - each submosaic separate → easy parallelization
- MFS for continuum, cubes, polarization
- incorporate self-cal (at start w/prev. model)
- ionosphere correction (TECOR)?
- pointing self cal (source density might allow this sometimes)
- CASA Benchmarks:
  - custom CASA script CASA took 82 hours (single proc, single node)
  - could be done in 3h using 7 nodes x 4 procs (or similar)
  - VLASS processing more complex
  - need improvements in CASA performance
    - these are being worked on now

Analysis

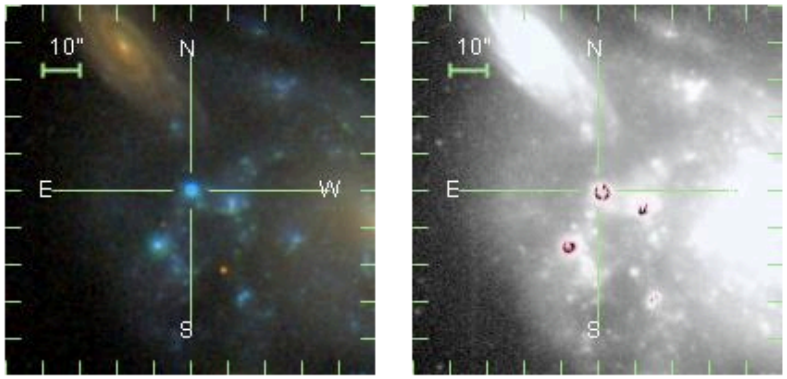
- Object (transient & durable) finding
  - use (modified?) source finders, e.g. SAD, Aegean
- Visualization
  - look at raw/calibrated data, calibration solutions, etc.
  - look at Tpix of images (!), use new viewer (ALMA dev)
- Multi-messenger
  - tools to quickly match VLASS to other views of the sky
  - example: the Caltech NRAO Stripe-82 Survey Marshal
    - developed by Kunal Mooley, used to identify transients from numerous transient candidates
    - something like this would be more generally useful
    - possibly for citizen science

The CNSS Marshal

- Example from Kunal Mooley's CNSS marshal:
  - vs. SDSS, FIRST, Hodge et al., PTF, WISE, Vizier, NED

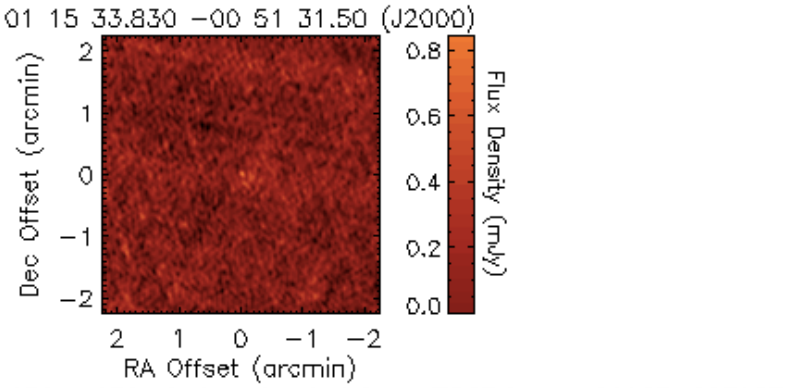


SDSS



Loading...

FIRST



Test Case – Stripe-82

1” cell
1 term MFS
I only
W-projection (128)
84 μJy rms

- Observing block
  - 3 hours, ~250GB, 2025 OTFM target phase centers (“fields”)
  - SB 13B-370.sb28581653.eb28626177.56669.781848645835
- Calibration (CASA JVLA pipeline)
  - 66.25 hours (22x block duration), mostly spent plotting
- Calibration (CASA custom script)
  - 10.5 hours (3.6x block duration), single process on cluster node
- Imaging (CASA custom script)
  - 81.56 hours (27x block duration), single process on cluster node
- Total (CASA custom scripts) – current code
  - 92 hours (31x block duration)
  - 4 procs (imaging) = 31 hours (<48 hours QL!)
  - 120 procs (30 nodes) plausible for VLASS

How will we run this?

- Data staging, QL Pipeline triggering, basic QA by Data Analysts (guided by Observer of the Day)
- Alerts to community of new data available
- SE/CF pipeline operations under supervision of technical group with data analyst support (&OoD) running separately from QLP, longer term
- Temporary (scratch and medium term) storage
  - on Lustre system accessible to processing cluster(s)
  - TIP estimate: 25TB “live” for calibration, 216TB for imaging
    - this is long-term once survey gets going
    - careful staging of data on/off needs to be coordinated
  - past experience is that we like to keep lots of data on disk
    - survey start-up will probably need 50-100TB

Algorithm & Software Development

- Improvements in algorithms (particularly imaging) and software (mostly CASA) needed or beneficial
- The big win: performance improvements in CASA
  - memory footprint for calibration, imaging should be optimized
  - removal of bottlenecks: applycal (CalLibrary use)
- Mosaicking
  - New Clean: AW-projection MSMFS for joint (uv-plane) deconvolution (under testing), get ultimate sensitivity
  - Current: Single “field MSMFS” or per-spw joint deconvolution
  - Also: integrated auto-boxing beneficial (not required)
- RFI auto-flagging: always need better!!!
- Analysis: RM synthesis, other?

Resourcing

- Computing:
  - will need to use significant Lustre storage (250TB)
  - will need to use large amount of cluster nodes
    - 36 nodes (fewer with performance improvements)
  - use of POD (e.g. with XSEDE) could be a game-changer
- Archive
  - 800TB data, 500TB images (can be accomodated)
  - need better archive interface (NRAO and/or EDS)
- People
  - Test & Development (1 year) – 3.1 FTE yr
  - Operations, observing, BDP production – 4.3 FTE/yr for 7 years
  - POD development – 2.1 FTE yrs
  - Comm/Edu/Outreach liaison – 0.5 FTE/yr
  - Algorithm Development - ? 1 FTE yr (w/2 FTEs)?
- Coffee – yes and lots of it!
- SSG/Community
  - your time is valuable! what is optimal interface?
  - students & post-doc program would be fantastic
    - separate or integrated into NRAO programs?