Science Ready Data Products and the ngVLA

Jeffrey S. Kern
National Radio Astronomy Observatory
Charlottesville VA 22903

Abstract

With angular resolution comparable to ALMA and future ELTs, the ngVLA will be the centimeter wavelength instrument contributing to multi-wavelength astronomy throughout the next decades. To maximize the impact of the ngVLA it is important that it be a facility available to all astronomers, not only to those who consider themselves radio astronomers. Building on the successes of the ALMA and VLA pipelines NRAO has undertaken a multi-year project to provide science ready products to the users of our telescopes and archive. The SRDP initiative addresses one of the technical risk areas of the ngVLA, and demonstrates the Observatory's ability to deliver this critical functionality.

The SRDP project is currently refining and reviewing the technical and scientific requirements for the existing NRAO telescopes. Capabilities will be made available to the community as they are developed and mature. As the project progresses, a rich archive of science quality radio images from both ALMA and the VLA will be provided to the astronomical community. The current status, concepts, and plans for the SRDP project are presented, as are the implications for the ngVLA facility.

ngVLA Data Processing

The next generation VLA will enable science at unprecedented resolution and sensitivity throughout the centimeter to millimeter regime. Although the science is revolutionary, the technology is mostly an update to *current* state of the art electronics and digital processing capabilities. Table 1 (below) summarizes several of the key science use case requirements and the implied post-correlator data rates.

In developing the data management plan for the ngVLA there are two key questions that must be answered:
"*Are the raw visibilities from the correlator archived*?" and "*Where is the processing to be done*?"

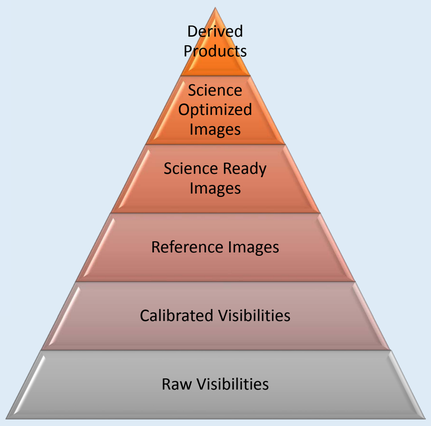
Based on the current key science drivers, and assuming that other projects have similar data rates, the raw visibilities can be archived, provided that historic trends in the cost of storage continue. It is worth noting that it would be cost prohibitive to store the visibilities at the current epoch.

While the ngVLA does not have the same exaflop computing problem that the SKA project does, it will require a modest cluster to process most observations. Moving the observed data from the archive to the processing center however is another problem. At the beginning of early science, in 2035, a reasonable assumption is that 10-gigabit networks are commonplace. For traffic from the archive to an arbitrary node assuming 25% to 50% efficiency is reasonable, thus for a project producing 10 TB/hour of data, transferring that data will be 5 to 10 times longer than the observation. Of course for dedicated point to point transfer significantly higher data rates are possible. Based on this analysis, we conclude that most processing will be performed at dedicated processing centers rather than at PI institutions as is done now. The SRDP project is preparing NRAO for this transition using the data from our current telescopes.

Center Frequency [GHz]	Bandwidth [GHz]	Channel Width [kHz]	Field Of View [arcmin]	Target Resolution [mas]	Raw Data Rate [TB/hour]	Raw Data Rate [PB / Month]
100	20.0	2500	0.08	5	14	10
27.3	13.5	576	0.08	5	40	29
40.5	4	13	1.0	50	50	37
7.9	8.8	131	317	500	11	8
16.4	8.2	273	4.7	100	5	4

Table 1: Key parameters and the associated data rates for selected Key Science Goals. Although these data rates are cost prohibitive in the current epoch $\mathcal{O}$($100K) per PB historical trends in storage predict that the cost per petabyte will be $\mathcal{O}$($5K) in 2035, making storage of the raw data expensive $\mathcal{O}$($1M) per year but tractable.

The NRAO SRDP Project



Tiers of product levels as designated by by the SRDP Project for delivery to users.

The Science Ready Data Products project at NRAO is a multi-year effort to maximize the scientific impact of NRAO's telescopes by:

- Allowing our users to spend more time focused on science and less time on data processing.
- Decreasing the barriers to using NRAO facilities for the broader astronomical community.
- Curating a rich archive of images to enable archival studies.

The SRDP project is in the initiation phase and is developing the overall requirements that will guide the project. Science ready data products will address use cases for both telescope and archive users.

At left is a representations of the tiers of products as considered by the SRDP project. An important concept is the three tiers of images. Reference images, which are informative to the user, but require additional effort before being ready for science analysis. Science ready images which are any image for which some science can be done, this means understanding the uncertainties in the flux scale and sensitivity of the image. And science optimized images, which are created for a specific scientific intent, through iterative interaction with the researcher.

In defining the requirements for the SRDP project we are guided by the following set of use cases. These are still under development and may change as we refine and review the requirements. The table at bottom right summarized current status of the primary use cases by telescope.

Standard Calibration the automated generation of a science quality assured calibration for supported observational modes.
Standard Imaging the automated generation of science quality assured images for supported observational modes. Objective is a homogeneous set of images in the archive to support telescope and archive users.
Optimized Imaging targeted generation of science quality assured images as requested by a science user. Images will be quality assured (in conjunction with the user) and delivered to both the requestor and to the archive.
Combined Imaging takes data taken from multiple configurations (including the ALMA Total Power Array) to produce quality assured images with better flux recovery.
Time Critical Observation's modifies the standard processes to support triggered and target of opportunity observations.
Large Projects represent a large investment both on the part of the observatory, and on the part of the user. This use case focuses on maximizing the return to all stakeholders, and leveraging the special relationship between these projects and the observatory.
Curation and Reproducibility describes how the SRDP project will assist the NRAO user community in documenting reproducible data reduction processes.

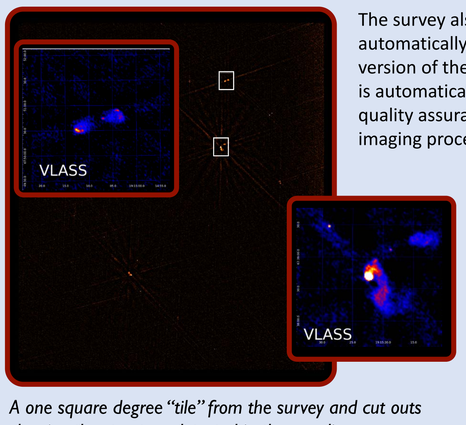
Restore one or more measurement sets to the calibrated and flagged state they were at the end of the standard calibration process. The calibrated measurement sets might be delivered to the PI directly or serve as the initial state for other processes.
Recalibration is the process of redoing a calibration, either with a different version of the supporting software tools, or with additional inputs from the user. Quality assured recalibration products are stored in the archive.

Use Case	ALMA	VLA	VLASS
Standard Calibration	Science Ready	Reference	Science Ready
Restore / Recalibration	None	None	N/A
Standard Imaging	Reference	None	Quick Look: Science Ready Single Epoch: Preliminary
Optimized Imaging	None	None	None
Large Project	None	None	None

The SRDP Project is currently seeking a full time Project Scientist to lead the heuristics development and validation team. Details and application instructions available at: http://jobs.jobvite.com/nrao/job/oAoy6fwm

SRDP Precursors at NRAO

The VLA Sky Survey is a three-epoch survey of the sky visible from the VLA. Totaling nearly 40,000 square degrees this S-Band (2-4 GHz) survey will achieve 120 μJy RMS in each epoch, and a cumulative sensitivity of 69 μJy. When complete this high resolution (2.5 arcsec) survey is expected to have detected over 5 million radio sources.



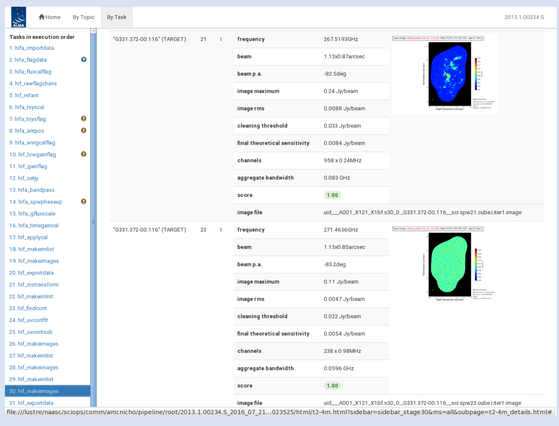
A one square degree "tile" from the survey and cut outs showing the structure detected in these radio sources.

The survey also represents the first science ready data products automatically generated for the Jansky VLA. Using a slightly modified version of the standard VLA calibration pipeline, each VLASS observation is automatically calibrated upon completion of observation. After a quality assurance step, the calibration products are stored and the imaging processing is begun.

Over the lifetime of the survey it will produce and quality assure:

- 1350 Calibrations
- 108K Quick Look Images
- 108K Single Epoch Images
- 216K Cumulative Images
- 3.7M Coarse Spectral Images
- 170K Fine Spectral Images

Lessons learned from this experience are already, and will continue to be, applied to the SRPD Project

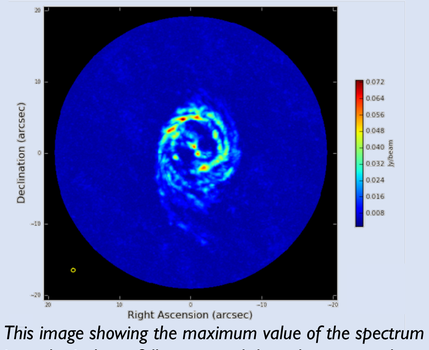


The Pipeline Weblog, initially developed for the ALMA pipeline, and now used by the VLA pipeline as well, is a hierarchical representation of the pipeline processing that provides both a summary view for the novice, and a detailed view for the expert. The SRDP project will expand the use of the weblog in the archive.

The ALMA pipeline has been used throughout Cycle 4 for the automated calibration and imaging of ALMA data. Over 95% of Cycle 4 standard mode observations were successfully calibrated by the pipeline. Approximately two thirds with no human intervention. NRAO and the ALMA project continue to refine the heuristics in order to decrease the fraction of observations that require human intervention.

The imaging pipeline has been used throughout Cycle 4 to produce reference images, suitable for establishing the quality of the observations.

Improving the quality of the images, and providing interfaces to allow user specified re-imaging are high priority deliverables from the NRAO SRDP Project.



This image showing the maximum value of the spectrum in each pixel was fully generated through automated processes.

The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc.