Title: SRDP Revolution	***Author***: Kern, Glendenning, Robnett	***Date***: 2019-07-10
NRAO Doc. #: 530-SRDP-043-MGMT		***Version***: 1.0

The Science Ready Data Products Revolution at the NRAO

PREPARED BY	CONTACT	ORGANIZATION	DATE
J. Kern	jkern@nrao.edu	NRAO	2019-07-10
B. Glendenning	bglenden@nrao.edu	NRAO	2019-07-10
J. Robnett	jrobnett@nrao.edu	NRAO	2019-07-10

Change Record

VERSION	DATE	REASON
1.0	2019-07-10	First version for submission to Decadal Survey

TABLE OF CONTENTS

1 INTRODUCTION

In the era of multi-messenger and multi-wavelength astronomy, the radio and sub-millimeter domain provides unique tools to probe a wide variety of astrophysical phenomena that are relevant to nearly all fields in astrophysics. Until recently, these powerful diagnostics have been utilized predominately by experts in radio interferometry techniques. The specialized techniques required to transform the Observatory's raw data into useful astronomical images have acted as a barrier to more widespread use of these powerful tools. The Science Ready Data Products (SRDP) initiative will increase the accessibility of the National Radio Astronomy Observatory's (NRAO) instruments to non-radio astronomers and increase the scientific output of the astronomical community.

For future radio telescopes, such as the next generation Very Large Array (ngVLA), the science ready model is made essential by the large data volumes and associated processing costs. It is no longer viable to have individual astronomers maintain the computing resources and expertise to process radio data at their home institution. The shift of computing load from the researcher to the Observatory has profound effects on data management operations at radio observatories.

This paper describes the background, goals, status, and plans of the SRDP project, concentrating on the software, data management, and computing aspects.

2 BACKGROUND

In 2014, the Atacama Large Millimeter/submillimeter Array (ALMA) Observatory began delivering quality-assured image products to observers by removing this barrier, a first for a major general-purpose observatory at these wavelengths. The ALMA telescope is highly oversubscribed, and has a broad and diverse scientific user base. Although the level to which this success is predicated on the generation of science ready products is not known, it undoubtedly contributes to the broad appeal of the ALMA telescope.

Building on the success of the ALMA pipeline, in 2017 the NRAO initiated the SRDP project. The objective of this project is to maximize the scientific impact of the radio telescopes operated by the NRAO on behalf of the National Science Foundation (NSF). By shifting the responsibility for creation of scientific quality images from the researcher to the Observatory, the SRDP project significantly lowers the barriers to novice users of the NRAO facilities. For existing experts, it permits astronomers to focus on the science rather than data reduction.

Data management provides the vital link between the data produced by facilities and the astronomers and astrophysicists that transform the data into scientific results. This vital role will evolve across the field in the coming decade as the next generation of telescopes produce ever larger volumes of data. The transition is particularly acute in the field of radio astronomy where the next generation of instruments (ngVLA, ALMA 2030, and SKA) are unsupportable by current operational models. Pioneered at radio and sub-millimeter wavelengths by the ALMA telescope, the Observatory's development of SRDPs will increase the use and impact of radio telescopes, place new operational constraints on data management operations at facilities, and drive technological and algorithmic changes within the supporting software.

3 GOALS

For our users and the Observatory, the SRDP Initiative will enable:

- Making ALMA, the VLA, and the Very Long Baseline Array (VLBA) accessible to non-radio astronomers (both as PIs and from archival data re-use by others). Increasing the user base of NRAO telescopes can only help in terms of assuring the best possible science from the telescopes, and in making the case for continued NSF support.
- Even for radio astronomers proficient in data calibration, flagging, and imaging, it will make them more efficient users of telescopes (analysis time per paper), effectively shifting effort from manual data reduction activities to scientific investigations.

In the next decade, the SRDP project will deliver science-ready calibrations and images for the Jansky VLA, the VLBA, and early science for the ngVLA. The benefit of these products extends beyond the radio observing community—the high-quality images produced by the SRDP project constitute a rich and accessible resource for archival researchers, increasing the scientific utility of the underlying investment of telescope time. For ALMA, the SRDP project will provide more processing and data access options, for example, the ability to re-execute imaging pipelines with non-default parameters.

4 STATUS AND PLAN

The approach to SRDP is supported by development in six critical areas: enabling external data processing, capturing scientific intent in proposals, improving the processing pipelines, modernizing the Archive, development of advanced algorithms, and modernizing CASA.

4.1 External (to NRAO) Processing

Individual researchers reducing data are primarily focused on minimizing the latency in the processing workflow to achieve high quality products as quickly as possible. The high-performance computing (HPC) model attempts to deliver this through broad parallelization, accepting the inevitable inefficiencies in order to decrease the latency of each job. Observatory-based processing, on the other hand, seeks to most efficiently utilize the investment in hardware, allowing individual processing requests to proceed more slowly in order to maximize the total number completed. High throughput computing (HTC) models are better suited to Observatory-based processing.

The VLA Sky Survey (VLASS), a synoptic microwave survey covering all of the sky visible from the VLA in three separate epochs, exemplifies this issue. While individual high-quality VLASS single-epoch (SE) images do present a computational challenge, the main hurdle is the sheer volume of calibration and imaging pipeline workflow executions. At a minimum, there will be over a half a million discrete workflow executions covering 34,000 square degrees. This challenge of completing a large volume of workflows within a fixed window is a prototypical HTC case.

Tests were performed on a variety of remote facilities: within the North American ALMA Science Center, within XSEDE facilities like the San Diego Super Computing Center, and within Amazon Web Services (AWS). While each of these approaches were successfully demonstrated, the entire suite of external options lacked coherency and would require bespoke software to enable operational computing at scale and to address automated data transfers and state tracking. A collaboration in development with the U. Wisconsin Center for High Throughput Computing (CHTC) and several other partners will provide access to more appropriately designed HTC workflow management systems and enable access to the extensive publicly available computing resources within the Open Science Grid (OSG).

4.2 Proposal and Observation Preparation Tools

To help increase the accessibility of NRAO and GBO instruments to the astronomical community, NRAO is redesigning the Proposal Submission Tool (PST). The PST is the first point of contact for most users and represents a significant barrier for novice users. Strongly recommended by the NRAO Users Committee, the new PST will focus on the scientific parameters (sky regions, sensitivity, resolution, frequency), minimizing the need to specify instrumental details (sub-band placement, baseline board pairs, sidebands, local oscillators) and data reduction parameters. This science-centered approach delegates the technical aspects of observing to a later phase of the process, enabling the proposers to focus first and foremost on their science program.

Novice users will find observing preparation simple, using Observatory-specified calibration strategies, automatically defined schedules, and standard correlator modes. The metadata for these observations will be captured in a database and used to inform the SRDP data processing pipeline. Technical experts maintain access to lower level configuration parameters, enabling opportunities for novel projects or those requiring unusual observing configurations.

NRAO is engaging the global radio community as the Observatory undertakes the re-implementation of the proposal tool. Concurrently, the international ALMA partnership is initiating a redesign of the successful ALMA Observing Tool (OT), and the Square Kilometer Array (SKA) is entering the construction phase and will be designing and implementing their version of the proposal tool. While differences in the telescopes, requirements, and timescales likely preclude a single tool, a similar look and feel and reuse of some modules (such as the spatial and spectral visualization components) is possible. This approach may simplify using more than one of the premier interferometers of the next decade.

4.3 Data Processing Pipelines

NRAO leads the ALMA Pipeline development effort with important contributions from East Asia (NAOJ) and Europe (ESO), and has also been developing pipelines for the VLA. The ALMA pipelines can now generally calibrate and make reference images for the common observing modes with little or no human intervention. The VLA has a calibration pipeline which is executed for most observations and the results are made available to users. In June of 2019, the NRAO began to perform quality assurance of the calibration pipeline results for a growing subset of observations as part of the SRDP pilot project. Even without the quality assurance, the pipeline calibrations are often useful. For the VLA Sky Survey (VLASS), the VLA calibration pipeline produces well-calibrated data for most observations. In addition, there is a VLASS imaging pipeline which automatically produces most of the quick look images and is currently being upgraded to produce the higher quality single epoch and cumulative images.

The NRAO is committed to continued improvement of the Common Astronomy Software Applications (CASA) calibration and imaging pipelines in support of SRDP. For all telescopes, the overall goal is to ensure that the pipelines produce images adequate for scientific use and publication, for all common observing modes. This will be accomplished by allowing users to re-execute (on NRAO-provided computing systems) the pipelines with access to a modest set of parameters, and over time by increasing the sophistication of the pipelines so that these re-executions will be minimized.

4.4 Archive

To maximize the use of NRAO data, those data need to be available to as many users as possible in addition to the original PI. Perhaps the most important aspect of this is the one mentioned above, i.e.

inclusion in the archive of pipeline produced images. This is particularly important for non-PI observers, who may not understand the calibration intent of the various sources in the observation. In addition, the NRAO will:

- Replace the NRAO Archive Access Tool (AAT) with an easy to use, modern, user interface that can search for, and return data from, all NRAO telescopes, including ALMA. (At present, ALMA data is in a separate archive without a common interface).
- The Archive holdings, of both raw and SRDP generated products, will be made available through Virtual Observatory (VO) mechanisms. This will allow better support of multi-messenger/wavelength science and better integrate radio/sub-millimeter data into the global astronomy data ecosystem.
- NRAO will endeavor to fully link its archived data sets with the published literature, e.g., through the dataset identifiers now in use in the Astrophysics Data System. Through such links, scientists can easily find the data underlying research papers and, while searching the archive, find the research papers in which the observations are presented and analyzed.
- NRAO will implement image visualization (using CASA's community developed CARTA image visualizer) and simple analysis tools on its servers, so users will not be required to download images to their own machines in order to do science.

Whenever the flagging, calibration, or imaging algorithms (in CASA) or heuristics (in the pipeline) change significantly, improved products can be made available. This will generally happen upon request by an archive user, but possibly via reprocessing all archive data in cases of very significant algorithmic improvements.

4.5 Algorithms

Imaging and image deconvolution steps constitute the dominant computing and data I/O bottlenecks in an end-to-end processing of data from modern interferometric telescopes.

A suite of imaging and image deconvolution algorithms exists in CASA. Imaging algorithms (to convert irregularly sampled visibility data to raw images) range from standard imaging without direction-dependent corrections to projection algorithms that include corrections for the effects of non-coplanar baselines and antenna far-field power patterns (for wide-field effects, antenna pointing offsets, and instrumental polarization variation across the observing band). Different combinations of imaging and image deconvolution algorithms are possible in CASA and involve trade-offs between computing costs, imaging performance, and algorithmic complexity.

Classical algorithms ignore many low-level effects and are insufficient for full-sensitivity wide-field wide-band imaging with current telescopes such as the VLA and ALMA. The NRAO has developed advanced imaging algorithms that account for these effects that affect the imaging performance of NRAO telescopes. These algorithms are inherently more computationally expensive and increase computing load by 10x to 100x. In addition to investigating and creating reference implementations of these algorithms, to be successful in accessing external (existing) facilities, we will need to ensure that these expensive algorithms fit the paradigm of the facilities. Perhaps most notably, ensuring that the memory required per core is not excessive by structuring algorithms to require smaller portions of the data to be in memory, staging other portions to persistent storage.

Radio Frequency Interference (RFI) corrupts the data, particularly in the VLA frequency range, but possibly also in the future ALMA. Some of the most advanced RFI flagging algorithms have been implemented in CASA. These algorithms are quite mature and have been in use for many years. However due to the complexity of the RFI in the data, configuring these algorithms optimally for the variety of RFI in a particular dataset is challenging. Algorithms R&D has started to evaluate if machine learning (ML) approach could automate configuring these algorithms. The crucial input to such approaches is a diverse set of training data and NRAO data archives are an excellent source for this.

4.6 CASA Refresh

CASA is NRAO's flagship post-processing package for the VLA and ALMA. An active forward-looking development path for CASA is critical for leveraging the investments of NRAO and the community moving forward into the next decade.

CASA extends NRAO's history of analysis software leadership through the coming decade and into the next. CASA is succeeding in the core mission of facilitating science with the world's largest interferometers: ALMA and the VLA. However, the success of the CASA package extends beyond NRAO telescopes, being used for data reduction at several operating telescopes worldwide. In the next decade, CASA will continue to unlock the potential of NRAO's telescopes by delivering scalable functionality from laptops, workstations, and modest clusters housed at NRAO facilities to the nation's supercomputing facilities through Open Science Grid and XSEDE.

The impact of the CASA package extends far beyond the day-to-day science operations of NRAO telescopes. Scientists pushing the frontiers of radio interferometric imaging around the world use the *casacore* libraries, a subset of the CASA package. The existence of the common casacore infrastructure provides a conduit for the rapid exchange and collaboration on implementation and algorithmic developments.

However, the foundations of CASA and casacore are 25+ years old and maintenance difficulties are growing. It is time to modernize it. As just one example, casacore is not thread safe, which makes certain classes of performance optimizations impossible. CASA is being upgrade to Python 3, enabling CASA to more easily interoperate with the astropy ecosystem. Furthermore, CASA is being modularized such that elements of CASA can be imported into Python itself (including Jupyter notebooks), rather than being restricted to the CASA environment. This will allow researchers to more easily access CASA functionality necessary for data analysis. Future developments will include updating the measurement set (the CASA data model) specification to handle modern data (e.g., phased-array feeds), and the underlying technical infrastructure will be replaced with a modern open-source solution with good performance scaling characteristics.

5 DISCUSSION

5.1 Data Management Operations

Historically, the NRAO has followed an HPC model when investigating problematic imaging cases and has a mixed history of engaging with HPC facilities. Much of the effort has been in enabling broader parallelization breadth to reduce imaging time for complex wide-field, wide-band, low-frequency continuum imaging cases. As discussed in Section 3.1 for the Observatory as a whole, reducing the end to end execution time of a single observation is less important than ensuring that the ensemble of all processing can keep up with observing and reprocessing steps, and NRAO is exploring a shift to an HTC model.

5.2 Manual Processing Will Always Be Possible

The NRAO recognizes that there will always be cases where unusual science goals or technical requirements require human-directed data reduction. Some radio astronomy experts will want to verify the quality of the automatic reductions versus what they can do by hand, perhaps in a non-NRAO supported software package. This skepticism is natural, and provides a helpful verification of existing procedures or identification of problems with them to the Observatory, and should not be discouraged. This requires that:

- It will be possible for users to re-execute the pipeline, tweaking parameters, when the standard pipeline is "almost" good enough. (This is implemented in the standard CASA pipeline infrastructure.)
- For users who want to reduce their data with the NRAO pipelines or CASA, NRAO will provide an implementation ported to the national computing infrastructure, as well as in-house cluster facilities. The in-house facilities will have less latency for startup and will be appropriate for smaller datasets; the national HPC infrastructure will be appropriate for larger problems (e.g., peak data rate observing resulting in many TB of raw data size). For smaller datasets and analysis, NRAO will continue to support CASA on desktop and laptop systems.
- For users who want to process data using software packages other than CASA, the Observatory will provide the raw visibility data in SDM/BDF format. Additionally, CASA provides limited support for translation of UVFITS and FITS-IDI formats.

NRAO is committed to supporting the vision of providing sufficient cyber-infrastructure to ensure PIs with limited resources are able to perform their science. This will be achieved by giving PIs secure access to the internal clusters, leveraging the same MyNRAO user account used for proposal submission, helpdesk support, and archive access. An enduring, yet evolving, security program underpins this access.

6 SUMMARY

Radio interferometry has traditionally been a field which requires considerable technical expertise to turn the raw instrumental data (visibilities) into derived products (images, catalogs) suitable for scientific analysis. Even for people with sufficient technical expertise, this work requires a considerable investment in time and resources (both human and computational) that would better be spent on the science analysis. The NRAO's SRDP initiative is aimed at exactly removing this burden for users of NRAO data (both PI groups and archival researchers) for most observations on ALMA, the VLA, and the VLBA. Besides being important for both existing and prospective users of NRAO's current telescopes, this is essential preparatory work for the next generation of radio telescopes. The ngVLA and an anticipated large upgrade of ALMA will have such large computational requirements that traditional user-driven data processing on modest personal or departmental scale computing resources will not be possible.

Although the SRDP initiative will require various organizational changes (QA processes; data processing operations of similar importance to telescope operations), this paper concentrates on the most important software development and computing infrastructure items which are:

- Access to expanded computational resources, both for direct users of SRDP and to fulfill Observatory requirements (VLASS imaging). The emphasis is on HTC rather than single-job parallelization (HPC).
- Updating the VLA/VLBA (and GBT) Proposal and Observing Preparation tools to better capture the scientific intent of the observations to aid in downstream processing (as well as a number of

other important requirements, including modernizing and improving the user interfaces, and supporting new proposal review paradigms).

- Improving the pipelines, both by increasing the breadth of observing modes that can be handled automatically, and by attaching pipeline re-execution with modified parameters to a user interface so users can steer the pipeline executions to be more exactly relevant for their science.
- Modernizing the Archive interface so it can ingest and serve science ready products, provide advanced capabilities including server-side visualization with CARTA, and in general present a more modern and higher-performance interface to users.
- Continuing to press forward with algorithms, filling out the coverage of imaging (wide-band, wide-field, full-polarization, pointing errors) and other algorithms (e.g., RFI) as needed for current and future radio telescopes. Sample implementations will need to take into account characteristics necessary for them to execute efficiently on available (in-house and external) computing infrastructure.
- Updating the CASA infrastructure so it can support modern computing paradigms, be easier to maintain, and scale to the needs of future radio telescopes, most notably the ngVLA.

SRDP operations have already begun in the form of a pilot program, and SRDP capabilities will be rolled out in successive waves over the next five years, starting in late 2019. If you are interested in using SRDP and following its progress, visit https://science.nrao.edu/srdp and the new https://archive-new.nrao.edu/.