NRAO VLA Archive Survey

Jared H. Crossley, Loránt O. Sjouwerman, Edward B. Fomalont, and Nicole M. Radziwill

National Radio Astronomy Observatory,
520 Edgemont Road, Charlottesville, Virginia, USA

ABSTRACT

The Very Large Array (VLA) radio telescope, operated by the National Radio Astronomy Observatory (NRAO), has been collecting interferometric data (visibilities) since the late 1970's. Converting visibility data into images requires careful calibration of the data, fast Fourier transform processing, and deconvolution methods. To make VLA data accessible to the astronomical community, the NRAO has undertaken the NRAO VLA Archive Survey (NVAS). The goal of NVAS is to produce images, calibrated data, and diagnostics from the visibility data archive and make these data products available to all astronomers. Survey results are obtained from a software pipeline, the details of which are described here.

Keywords: Surveys, Data Pipeline, VLA, Synthesis Imaging

1. INTRODUCTION

The Very Large Array* (VLA) has been an exceptionally productive radio telescope. The array consists of 27 25-m diameter antennas arranged in a Y-shaped configuration and located near Socorro, New Mexico, USA. There are four array configuration sizes that allow for a wide variety of angular resolutions; the maximum array size varies between 0.3 km (D-Configuration) to 35 km (A-Configuration). The array remains in one configuration for a period of about four months. The VLA is capable of recording radio frequencies from 0.1 to 43 GHz. The telescope is open to all researchers whose proposals are refereed by an international group that decides on the observing priorities. Approximately 50% of the submitted proposals receive some VLA observing time; the telescope is used over 65% of the time for basic astronomical research.

The data from all VLA antennas are transported to the central control building where a correlator measures the coherence of signals between each of the antenna pairs. This spatial coherence function, called the *visibility*, is stored in an archive for processing. The archived data is then downloaded from http://archive.nrao.edu and analyzed by the project scientists to obtain images using the Astronomical Images Processing System (AIPS) software package, developed and maintained by the National Radio Astronomy Observatory (NRAO). The major processing steps to obtain high-quality images of a selected region of sky and at the required range of frequencies are described below. The responsibility for analyzing the raw visibility data to form images belongs solely to the project scientists. Over 150 papers based on VLA data are published each year.

One year after the observations, the raw data become publicly available for general use, but require significant effort to produce useful images. In order to make these archive data available to the larger part of the astronomical community, the NRAO VLA Archive Survey (NVAS) has been created. NVAS (also called the *VLA Pipeline Survey*) contains images, calibrated visibility data, and diagnostic plots produced from VLA archive data using an automated data reduction system—a data pipeline. After imaging, NVAS data are combined with the VLA raw data archive and collectively provide archive users with access to raw, calibrated, and imaged VLA data accompanied by graphical diagnostics.

NVAS processing of archive data is an ongoing project. Presently, roughly one-quarter of the VLA archive has been imaged for the survey. NVAS contains 72 839 images of 15 590 unique sky positions. Complete processing of the VLA archive is expected by the end of 2010.

Send correspondence to J.H.C. E-mail: jcrossle@nrao.edu

*The Very Large Array is an instrument of the National Radio Astronomy Observatory, a facility of the National Science Foundation, operated under cooperative agreement by Associated Universities Inc.

Table 1. Present State of NVAS

Number of Images	72 839
Unique Sky Positions	15 590
Dates	1991 to 2003
Frequencies	1 to 50 GHz
Observing Modes	Continuum
Array Configurations	D, C, B, and some A

This article describes the NVAS Pipeline System in detail. (A scientific analysis of NVAS will be published by L.O.S. at a later date.) The current status of NVAS is described in § 2. The four main parts of the pipeline system are described in § 3. An analysis of the NVAS pipeline is contained in § 4. Future plans are described in § 5. A summary is given in § 6.

2. CURRENT SURVEY STATE

The present state of NVAS (as of 2008 April) is described in this section and in Table 1. The processed data is currently limited to continuum observations (to simplify the pipeline start-up effort); spectral line data will be added to the survey in the future (see § 5). Frequencies below 1 GHz are omitted because of the increased difficulty of calibration and imaging. With these constraints, we have processed all B-, C-, and D-array configuration data between 1991 and 2003; A-array configuration data in this time range is partially processed. A-array imaging requires more CPU time than imaging data from smaller array configurations. Early in the NVAS effort, A-array data was avoided to save time while testing and debugging. A map of the current NVAS sky coverage is shown in Figure 1. The map is generated in the archiving stage described in § 3.

3. THE NVAS PIPELINE SYSTEM

The NVAS Pipeline System consists of 4 stages that automate the generation of the survey data. These stages are: data acquisition, data processing, data archiving, and data validation.

1. **Data acquisition**

   In this stage, raw visibility data, as limited to the frequencies and arrays listed above, are selected from the VLA archive and downloaded to a local disk. The data are then loaded into the Astronomical Image Processing System (AIPS) and prepared for further processing in Stage (2), below.

   Data acquisition is accomplished through Perl scripts that automate, step-by-step, the process a user would go through to acquire data from the VLA archive via World Wide Web interface. The Perl scripts are started from a command line and controlled by command line parameters. In the simplest and most highly automated case, a range of time is selected by specifying start and stop dates on the command line. The Perl scripts then

   (a) query the VLA raw data archive for data files satisfying criteria of date range, observing frequency, and observing mode; filter out files that should not be included in our survey[†];

   (b) group related raw data files;

   (c) download each group of data files to a local disk;

   (d) load (fill) the data into AIPS; separate the data by frequency, removing frequencies that are below 1 GHz; verify that at least one known flux calibrator is present;

   (e) generate an AIPS *runfile* (batch file) that will be executed in Stage (2).

[†] Filtered files currently include observations for other VLA surveys (such as NVSS and FIRST) and VLA system tests that we know are not useful.

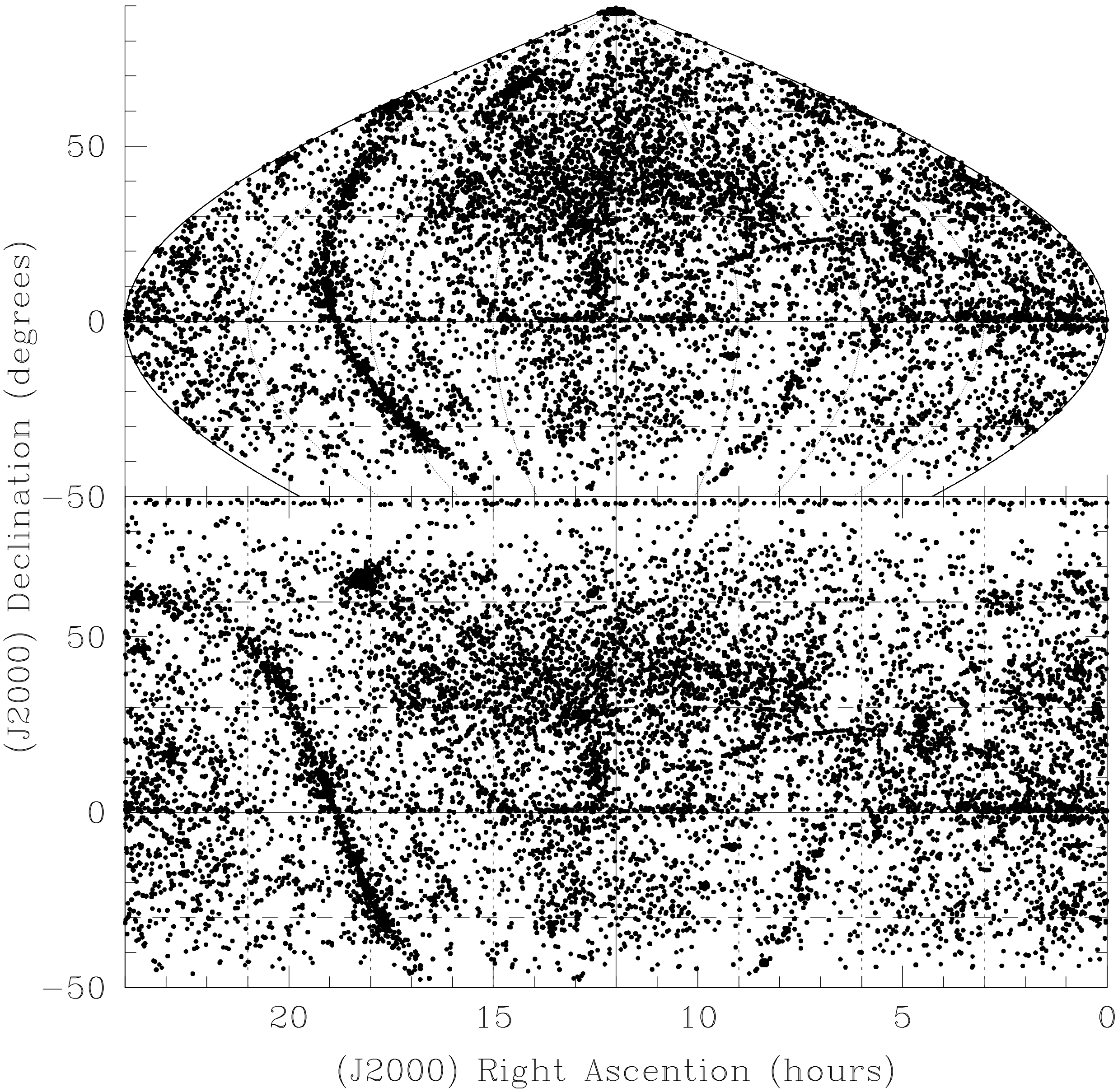
(J2000) Declination (degrees)
50
0
−50
50
0
−50
20
15
10
5
0
(J2000) Right Ascention (hours)

Figure 1. Map of NVAS sky coverage. At this time, the survey contains 15 590 unique sky positions.

Communication with the raw data archive is achieved by HTTP requests. The Perl scripts issue these requests via UNIX utility `curl`. Some of the work in this stage is accomplished by generating and executing shell scripts; those shell scripts are archived in Stage (3), below, for future reference.

Appropriately grouping related raw data files (step [b] above) to obtain the best images is difficult for two reasons. Primarily, calibration of data spanning a large range of time is challenging because wave front distortions (cause by, for example, the weather and ionosphere) vary greatly over long timescales; in addition, the telescope itself varies with time due to maintenance and typical mechanical and electronic imperfections. It is important to calibrate all such effects properly to obtain the highest quality image.

In a larger data set, covering a larger span of time, it is more likely that a subset of poorly calibrated data will negatively affect the whole data set. A secondary issues is that neither the observer's intentions nor the end-user's intentions are known at the time of NVAS processing. Concatenating well-calibrated visibility data of a single source decreases noise (or increases image sensitivity). However, information about temporal variability is lost when data from multiple observing epochs are combined to form one image.

With both of these challenges in mind, we have chosen in our first phase of NVAS processing to limit the time span of visibility data by focusing on 1-day-long file groups; some exceptions have been made when data spanning two days or more (and in the same observing project) is densely packed in time. The exact algorithm used for grouping visibility files is described here; the rules are ordered from highest to lowest precedence.

(a) Only files produced by the same observing project are concatenated.

(b) Visibility files are used only once in the survey; in other words, a raw visibility file only appears in one group. Thus, each image in the survey is derived from unique visibility data.

(c) All files starting on the same day are concatenated.

(d) Files that begin on different days are concatenated if either of the following conditions are true,

   i. the time between the end of the earlier file and the beginning of the later file is < 30 minutes;

   ii. the time between the end of the earlier file and the beginning of the later file is < 6 hours *and* the inclusive span of time between the first file starting after day one and the last file in the concatenated group is < 6 hours.

The time required to complete Stage (1) of the Pipeline is primarily limited by the data rate available for archive download and secondarily by the hard disk read-write speed available for loading data into AIPS.

2. **Data processing (using AIPS)**

Data processing is performed by the Astronomical Image Processing System (AIPS)[‡]. The data processing stage can be subdivided into the following steps.

(a) Editing is performed to remove obviously spurious data; the editing algorithm (task FLAGR in AIPS) finds outlier points which significantly deviate above the noise contribution from the expected slow variation of the visibility amplitude for each baseline.

(b) Instrumental gain and phase changes are determined using observations of calibrator sources, with known intensity and position in the sky; observations of the calibrators are mixed in with observations of the target sources.

(c) The Fourier Transform of the calibrated data produces an image for each source and frequency. Because the VLA samples a limited amount of the Fourier plane, the images have defects at the 10% to 30% level. However, these defects are accurately known, so that a much more accurate image of a source can be obtained by a deconvolution method, CLEAN, that decomposes the source into a large collection of point sources. The collection of point sources are then convolved with a Gaussian shaped beam with the same resolution of the original data, to obtain an accurate representation of the sources, free from the original defects.

(d) The image is prepared for publication by correcting for the primary beam shape (flat-fielding), cropping at a radius from the phase center where image sensitivity decreases below some threshold, and appending a copyright label. Images are then exported to JPEG and FITS[1] files; calibrated data are exported as UVFITS[§] files. Additionally, three diagnostic plots are produced that show (1) the

[‡] AIPS was designed by NRAO in the 1970s to provide VLA observers with a means to reduce their data. AIPS is a comprehensive and well-tested synthesis imaging software package, which, despite a cryptic user interface (by standards of contemporary software) and limited scripting capabilities, makes it a reliable platform for generating a VLA survey.

[§] UVFITS is the common name for the FITS *Random groups structure* format, a format popular in radio interferometry.[1]

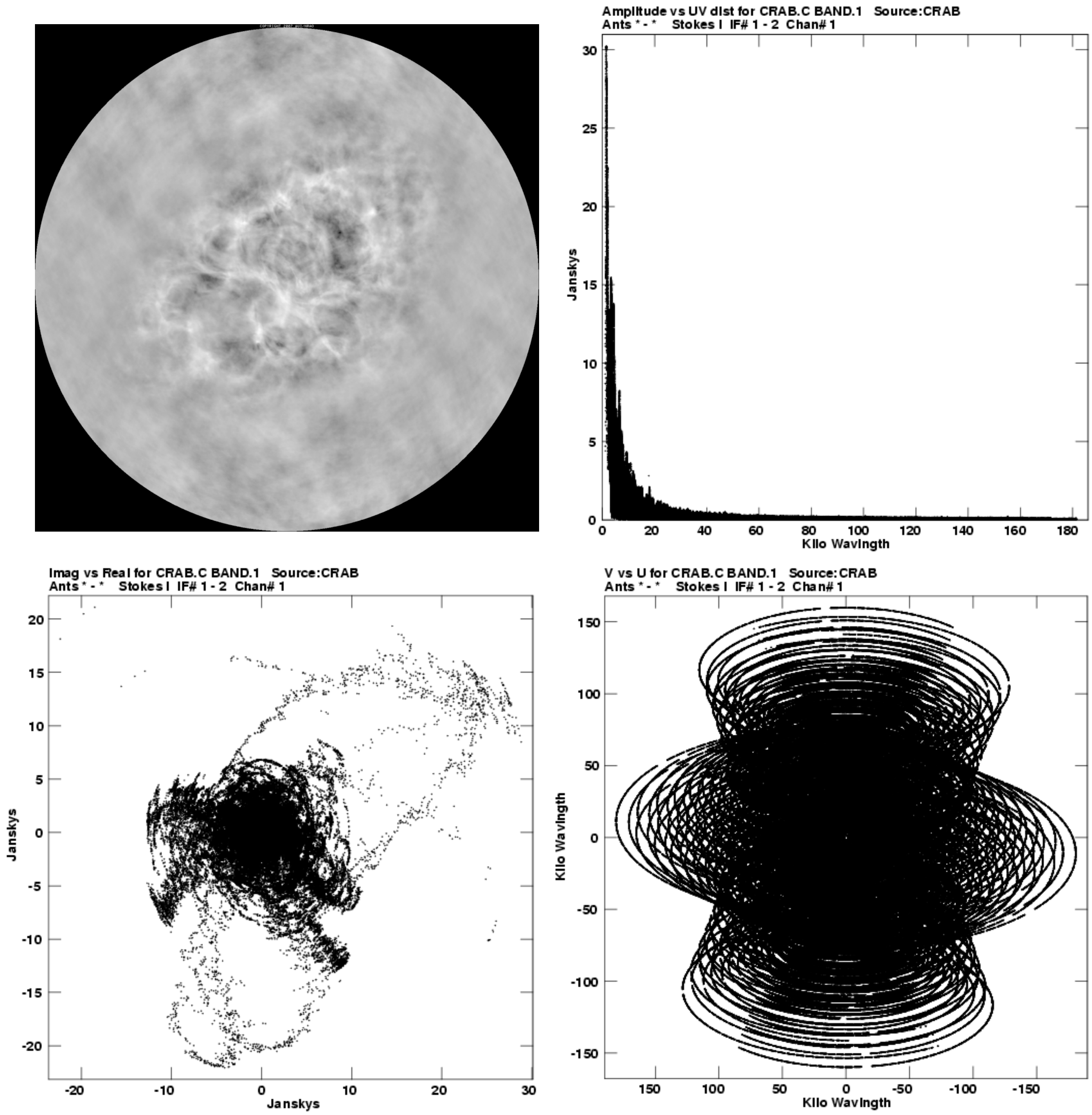

Figure 2. An example NVAS image with diagnostic plots. The image is of the Crab nebula at 4.76 GHz. The top right plot shows visibility amplitude versus baseline length; the bottom left plot shows real versus imaginary visibilities; the bottom right plot shows *uv*-coverage.

uv-coverage, (2) visibilities in the complex plane, and (3) the visibility amplitude as a function of baseline length. See Figure 2 for an example of the diagnostic plots. These diagnostic plots provide important information about the visibility data used to generate the image.

The time required for data processing is limited by CPU speed, disk read/write speed, and algorithm efficiency. The processing time varies widely for continuum data and is determined primarily by the image

Table 2. NVAS Data Products

Single-source files			Multisource files		
File description	#	Format	File description	#	Format
Image data	1	FITS	Visibilities w/ calibration tables	1	UVFITS
Reference images	2	JPEG	Calibration tables (no data)	1	UVFITS
Calibrated visibilities	1	UVFITS	Log files	3+	ASCII
Diagnostic plots	3	GIF	Shell scripts used in Stages 1 & 2	2+	ASCII

resolution or, equivalently, the array configuration size.

3. **Archiving**

   After processing, all data products are moved to the archive server. Data products produced in Stage 2 divide neatly between single-source and multisource files; the former contain information on a single source; the latter contain information on all sources whose visibilities are concatenated in Stage 1. To maximize storage efficiency, files are archived in two directory trees: one for single-source data and one for multisource data. Table 2 lists descriptions of each archived file, the number of files present for each single-source image or multisource project group, and the file formats.

   The data products listed in Table 2 are partially described in Stage 2(d), above. Full-size and thumbnail-size JPEG images of each source are produced for convenient viewing over the Web. Visibility data, with calibration applied, is available for each source; raw visibilities of all sources are also available with calibration tables attached; alternatively, for users that already posses the raw data, calibration tables are provided without attached data. The shell scripts used during Stages 1 and 2 are provided with the data, along with log files generated by the scripts.

   NVAS metadata is stored in directory names and file names. The metadata includes information regarding the sky position, VLA project code, observing frequency, image root-mean-square intensity, and more. This naming scheme allows for easy searching through the archive using filesystem tools. A detailed description of the NVAS naming scheme is provided on the NVAS pilot Web page[¶].

   The archiving stage also generates HTML files that are used to create the image archive Web interface. The HTML files store cross reference information that relates the multisource to single-source files and the single-source to multisource files. The archive Web interface returns information regarding NVAS data (for example, a cone search response) by concatenating HTML files for each applicable source or project code, producing a table of all relevant data.

   Once per day, new data products are copied to the archive server, and the NVAS archive files are updated. The archiving stage does not contribute significantly to the total time required to generate NVAS.

4. **Validation**

   The final pipeline stage is data validation. The data product quality is verified interactively. A private Web interface has been constructed to simplify this task. Currently, less than 5 percent of images generated have been removed in the validation stage.

   If the validating scientist thinks the data product quality is unacceptable, the data files are removed from the archive and saved in a separate location; these files are useful for testing and algorithm development. Removed files are inspected a second time by another scientist, at which point they may be restored to the archive.

   There is not a well defined algorithm for data validation, which is why human interaction is required. Synthesis imaging is a complicated process and errors can arise in many forms. A few of the most common signs of errors are:

[¶] NVAS pilot Web page: http://www.aoc.nrao.edu/~vlbacald

- outlier visibility points (high or low) that were not removed using the flagging algorithm;
- stripes on the image that are caused by outlier points or by sources that are very extended and difficult to clean;
- symmetric patterns in the image which are often caused by very limited observations and the inability to determine the source location;
- an root-mean-square noise on the image which is substantially larger than that expected.

The time required for data validation is limited primarily by the speed of the scientist performing validation.

4. ANALYSIS OF THE NVAS PIPELINE

4.1 Data Quality

To date, NVAS has produced 72 839 images; less than 3 800 images have been removed in the pipeline validation stage (see § 3). Some, if not all, of the survey data will be reprocessed when the editing, calibration, and imaging algorithms improve. All survey images are intended to be of sufficient quality that they are useful for archive browsing and preliminary scientific analysis. It is always advisable for end users to verify image quality for themselves before engaging in extensive analysis or publication. For this reason, calibrated and raw visibility data are provided with the survey images.

4.2 Constraints

The time constraints of each stage of the pipeline system are described in § 3, above. The data processing stage (2) and the validation stage (4) require the most time. Stage (2) is CPU limited, and processing times may decrease as algorithms evolve. Stage (4) is limited by the speed of human interaction.

Not considered in § 3 is the time required to address critical development concerns. For example, as the number of archived data files grow, some common filesystem utilities become incapable of handling the large numbers of files; this requires changes in the Pipeline System scripts. Additionally, as algorithm development continues (especially in data acquisition, Stage [1]), data download, processing, and archiving rates increase, and subsequent stages must be modified to keep up with higher data rates. Development of data processing (Stage [2]) algorithms is also time consuming, but generally does not stop survey processing.

CPU speed, disk input-output speed, and data storage capacity are the primary computer hardware constraints for NVAS. When possible, multiple instances of the processing stage are run on each computer, until disk input-output or the CPU clock speed become limiting factors. Presently, the NVAS archive occupies 1.4 TB of disk space. However, the survey is still growing, and at an increasing rate.

4.3 Personnel

In terms of personnel, NVAS is a small project, with three scientists—one project scientist (L.O.S.), one assistant scientist (J.H.C.) and one scientist providing oversight (E.B.F.)—and one project manager (N.M.R.).

5. FUTURE DEVELOPMENT

The NVAS development plan divides roughly into 3 phases:

1. Image all VLA continuum data (currently in progress). This has been done for most continuum data between 1991 and 2003; approximately 33 percent of the 32 years of VLA archived data[‖]. During this phase, improvements to Stages (1), (2), (3), and (4) have been and will continue to be made. Some of the on-going improvements include:
    - addition of self-calibration to the processing algorithm;
    - improvement of the data editing algorithm;

[‖] It is likely that much of the data prior to the VLA dedication (1980) will not be processable, since fewer antennas were available and special circumstances existed during array construction.

- generation of mosaic images when possible;
- concatenation and imaging of data from observing projects that are greatly extended in time.

2. By the end of Phase (1), the above improvements to the data processing algorithm should be complete. Phase (2) will apply the upgraded algorithm to any continuum data already processed.
3. Extend the survey to include spectral line observations.

We expect to complete most of the above steps by 2010. In the far future, the pipeline system will be used only to process new VLA data. In this sense, NVAS will continue to grow as long as the VLA is operational.

6. SUMMARY

NVAS makes images and calibrated data available, alongside raw visibility data in the VLA archive. Currently, 72 839 images have been produced of 15 590 unique sky positions. NVAS is an on-going project; roughly one-quarter of raw VLA archived data has been processed. The data pipeline used to generate NVAS consists of four stages: data acquisition, processing, archiving, and validation. Once started, the first 3 stages are completely automated. The final stage is human interactive. Less than 5% of images have been removed due to poor data quality. Several improvements to the processing algorithms are currently under development. The majority of the archive survey is expected to be complete by the end of 2010.

REFERENCES

[1] Hanisch, R. J., Farris, A., Greisen, E. W., Pence, W. D., Schlesinger, B. M., Teuben, P. J., Thompson, R. W., and Warnock, III, A., “Definition of the Flexible Image Transport System (FITS),” *Astronomy and Astrophysics* **376**, 359–380 (Sept. 2001).