NRAO Virtual Observatory Plan

Version 0.5, 30 March 2005

Overview

The Virtual Observatory (VO) defines standard mechanisms to discover, access, and analyze data from distributed astronomical archives covering all branches of astronomy. Publishing data and services to the VO is the ideal way to make the results of NRAO's instruments readily available to a wide community. Large projects on NRAO instruments are currently required to make the results publicly available in a timely manner and the VO will simplify this process. In this plan we recommend that most observing data products be made publicly available after a reasonable proprietary period, that the NRAO encourage or require that surveys as well as large PI-based projects publish any appropriate data products to the VO, and that the NRAO provide the necessary facilities to make it as easy as possible to do so.

This plan is composed of several stages with the intent of obtaining the highest short term scientific payoff with the minimal up front cost. Aspects of the plan requiring longer term development and higher costs are deferred until later in the plan.

There is currently a great deal of data available from NRAO instruments that could be published to the VO; indeed, in some cases (primarily the major surveys) data has already been published to the VO by organizations other than the NRAO. The first activity in the NRAO VO plan will be to actively and systematically collect currently available data products and publish these in a uniform fashion via a standard set of services. This includes wide area surveys (e.g. NVSS, FIRST, VLSS) as well as narrower but deeper surveys and pointed observations. Included is a pilot project to mine the VLA archive by producing uniformly generated reference images for a carefully selected subset of the data. In the process of carrying out this first stage we will develop the capabilities required to make it easy to publish NRAO data to the VO.

The next stage of the plan is to change the NRAO policy requiring publication of data products to include publication to the VO. We currently expect users to publish results in professional journals, and publishing the data itself is the next logical step to allow more widespread usage of the results. A recent agreement between ADEC and NVO has created a "dataset identifier" mechanism which can be used in published papers to refer back to the data products upon which a paper is based, allowing direct online access to data products referenced in this way. The NRAO should support this mechanism and allow all raw and processed data products for an NRAO observing project to be associated within the NRAO archive and referenced in publications. In this intermediate case, users would be encouraged or required (where appropriate for a given project) to publish final science data products to the VO, normally by publishing them back to the NRAO archive where they can be integrated with the other data products for an NRAO observing project and made available to the VO via a uniform set of services.

In the longer term, the NRAO "end to end" (E2E) plan is to allow pipeline processing to produce reference data products for most standard observing modes of NRAO instruments. Most of the effort to implement this capability is part of the E2E activity within each telescope project, but integration of the results into the NRAO archive and publication of the data to the VO is part of the overall NRAO VO effort. Publication of data to the VO includes both proprietary and nonproprietary (older) data. VO includes the necessary security mechanisms to safeguard access to proprietary data, allowing the same archive and VO infrastructure to be used for both PI science and general data access and data mining once the proprietary period is over.

A critical element for this plan to work is capturing the observers' intent and making such descriptive information available with the data products, for use to automate observing and post-processing of the data. The E2E plan includes a "project model", a digital description of an entire observing project, which is included in the archive along with the observational data and any processed data products such as images, spectra, and source catalogs, be they from an instrumental pipeline or the direct result of user data processing. The project information includes the proposal and other descriptive project metadata (only parts of which may be externally visible), and for each observation identifies the observing mode and any observing parameters, including the observing script used to obtain the data. Data processing is likewise guided by the observing mode and processing parameters defined for an observation. As data flows through the system the project metadata is updated to keep a full record of all data products associated with the project, and the processing performed to produce them.

Suggested policy changes

We recommend that NRAO require, for projects using NRAO telescopes, which are larger than a certain minimum size, and where the science requires the generation of appropriate data products, the publication of such user derived data products to the VO. Smaller projects should be encouraged to do so as well, both as a matter of policy and by providing the necessary capabilities at NRAO to make it easy for observers to generate and publish their data products. In order to ensure minimum standards of data quality, data products to be accepted for publication will be required to include sufficient valid metadata to identify and characterize the data to the NRAO archive and the VO. Reference to the data in a peer-reviewed paper may also be a reasonable requirement for publication.

VLA and VLBA already have a "big" project category which would be a reasonable place to introduce a policy requiring publication of data products to the VO. GBT does not yet have such a policy but should consider defining one. Any such policy should be flexible enough to be modified on a case-by-case basis, depending upon the nature of the project and whether it makes sense for the project to publish data products to the VO. Many smaller projects with appropriate data products should be strongly encouraged, as part of the telescope time allocation process, to publish data to the VO.

A further suggested policy change is that all project metadata, including those parts of the proposal required to understand and use the data for re-processing or data mining purposes, and any observing and data reduction scripts or other processing information,

be made available with the data products when the data becomes available after a reasonable proprietary period.

How much of the proposal the Observatory requires be public is a matter of policy. For most purposes it may be sufficient to have the cover page information and the formalized project metadata. This metadata should be sufficient to permit automated processing of the data as well as subsequent data mining. There will always be parts of the proposal which should not be public, e.g., the detailed scientific justification, and information from the proposal handling process such as grades or reviewer comments.

These proposed policy changes would take effect in the mid-term part of the plan, once the necessary infrastructure to support publishing of data to the VO by users is in place at NRAO. Near-term work will emphasize getting some useful data online while developing the capability to publish data to the VO. Sufficient non-proprietary data is already available to support near-term development.

Justification for Development at NRAO

There are a number of reasons why the NRAO should support and encourage the publication of the data products from its instruments to the VO; these include:

- Publication of scientifically interpretable data products widens the community of potential "users" of our instruments beyond the traditional radio astronomy community.
- Organized publication of NRAO data and services makes it easier for even radio astronomy experts to access and analyze the data.
- Making it easier to do science with NRAO facilities will increase the amount of science done and reduce the risk associated with a shrinking user base.
- "Reuse" of data products results in more science per unit of investment in the instruments.
- Development and deployment of complex instruments is expensive. Some fraction of the expense is justified to increase the amount of science performed on the resulting data.
- With the growth of online digital archives an increasing fraction of all astronomical data is publicly available data sitting in archives. Data mining of such data is becoming an important part of the astronomical research process. The VO is the means by which this happens.

Implications for NRAO

If the NRAO wants its data products to be available via the VO, then it must provide for the effort and expense of making this happen. The VO is not an archive in the classical sense, but more a collection of standards and tools that help users locate and access data and related resources and analyze data using these resources. Actual long term archiving and curating of the data is done by discipline-specific centers that are familiar with the data, such as the NRAO. For actual online data access the data could be physically made

available either at NRAO or at other locations such as the national supercomputer centers (NCSA, SDSC, etc.,), and the national TeraGrid.

In any case, the NRAO must have expertise in house to collect the data and associated metadata and organize it to be useful for data analysis via the VO. NRAO must bear the cost of operating the primary data archive for data from NRAO telescopes, as well as any pipelines to produce reference data products or calibrations, and the networking infrastructure and software interfaces required to federate this data with other archives. Experience at other astronomical archives indicates that significant effort may be required to curate archive data holdings, and ensure that the metadata is correct and that the data is of acceptable quality. For the very large data sets to be obtained with the EVLA and ALMA, providing substantial computing facilities for remote access and computation may also be required. For the long term, the NRAO must be prepared to curate this data for decades.

System Capabilities

In evaluating what is required for NRAO to be a full participant in the VO, it is difficult to decide what is VO, and what is an essential capability for routine observing, data management, and data processing and analysis. Hence for example the boundaries between NRAO "E2E" and VO are not clear cut. This should not be surprising, as VO is not merely an add-on capability; rather it reflects a fundamental change in the way astronomy will be done in the future. Perhaps a better way to think about VO is as a use-case, a science driver defining an important way NRAO facilities will be used in the future. In the following sections we will focus primarily on VO capabilities, but will also discuss the implications for related development such as for E2E and the observing process.

Overview

VO-related development naturally breaks down into three phases: what we can do in the short term to get something up quickly, what we want to have available when ALMA and EVLA come online toward the end of this decade, and what we need to do in the meanwhile to get there.

Near term (next 12-18 months)

> In the near term, the most productive VO related activity is the collection, organization and publishing on the VO of existing and available datasets. These consist of images or collections of images, and in some cases, source catalogs derived from the images. Included are the major radio surveys and selected deep surveys of smaller fields. At the same time a pilot project will be conducted to produce reference images for a coherent subset of the data in the VLA archive (continuum data for the VLB B configuration from a single semester). Also on this timescale, it should be possible to develop a VO interface to the existing VLA and VLBA data archives to allow VO tools to be used to find and access the extant publicly available raw data.

Intermediate term (next several years)

If the NRAO publication policy is changed to encourage or require users to publish processed science-grade data products to the VO, then a substantial amount of data will become available. On this timescale proposal information should be made available with data products to allow either reprocessing of raw data or understanding of more finished products. The NRAO should also support remote processing of public archive data. This will initially need to be driven interactively but the NRAO should continue to develop a capability for pipeline processing data for more automated processing. Data products generated in this fashion could be published back to the archive when generated. Also on this timescale, GBT data conformant to the new science data model being developed for single dish data will become available in the archive.

Longer term (ALMA/EVLA operational)

By the time ALMA and EVLA are fully operational the E2E system should be functional, providing an automated data management system for each NRAO telescope. All raw and calibrated data will be available in a VO-compliant science data model defined as an external interface so that any software can (in principle) be used to process the raw data. Automated pipeline processing for some subset of the available observing modes will be provided, providing reference images and spectra in the NRAO archive within a few days of observing. VO tools and VO-enabled data analysis software (including but not limited to that provided by NRAO) will be able to access both proprietary and public data. Remote interactive and pipeline processing (for selected observing modes) will be supported, as well as on-demand, on-the-fly generation of image and spectral data from calibrated visibility data.

In what follows we examine in more detail what is required to carry out the program described above. Finally we examine the resources for this effort.

Data Products

The raw and processed NRAO data products useful for data analysis via the VO include the following (in no particular order):

- *Survey images, spectra, and source catalogs*. This includes images, spectral data cubes, spectra, derived source catalogs, and possibly synoptic data, from organized surveys. Both large scale surveys and smaller, deep surveys of designated fields are of interest. Survey data is uniformly and carefully processed and ideal for data mining and analysis via the VO, hence this data is the highest priority for publishing to the VO (most major radio imaging surveys and catalogs are already available via the VO in some form).
- *Raw and calibrated visibility data,* ideally conformant to a formally defined science data model (SDM). While this is the most difficult radio data to use for analysis within the VO, this is the primary data product for NRAO telescopes or any other radio telescopes. While a SDM is in the process of being defined for interferometic data, data in this form will not be routinely available until ALMA and EVLA start to become operational around 2008. In the interim data from the

VLA and VLBA will continue to be produced in its current form. GBT data should transition to a formal SDM sometime in 2006.

- *Pulsar data from surveys*. GBT pulsar data is rarely archived at present due to the large volume of data produced, the lack of any standards for pulsar data, and the difficulty of analysis. Nonetheless, significant GBT time currently goes to pulsar research, and pulsar data is potentially useful for data mining purposes for specialized searches for variable objects or transients. Some fraction of this data (10%) is worth archiving and publishing to the VO, especially for cases where a uniform, large area survey is performed and the data product can be well documented. Particularly important are observations in bands that are RFI free now, but won't be in 10 years. In the longer term, in addition to providing access to the observational data, it might be possible to return simulated synoptic spectra or time series from pulsar data via a standard VO spectral access service.

- *Radio calibrators* and calibration data for specific instruments. These (potentially) include reference images, often synoptic, for radio calibration sources, and calibration database information for specific instrumental data streams. Calibrators are rarely imaged, but generating and publishing reference images, and/or time series fluxes (especially for point sources), for routine calibrator observations, could be valuable for data mining purposes.

- *Reference images and spectra*. Reference data products are uniform data products produced in a mostly automated fashion by routine pipeline processing of the instrumental data stream. In general, such automated processing may only be possible for some subset of the observing modes of an instrument. While reference data products will in general not be as good as what can be produced by interactively processing the data for a specific program, such data is important for use within the VO since it will probably be more uniform than human-processed data and may be better characterized, making the data well-suited for automated analysis including statistical analysis.

- *Final science-grade images, spectra, and time series*. These are the actual final data products produced for a specific science program. Like survey data products they have been carefully produced by a human expert; however these data products represent individual pointed observations. Examples might be a single ultra high resolution image from a VLBA observation, or a time series resulting from a pulsar observation. The quality may vary more than for reference data products depending upon the skill of the project team.

- *Project metadata including the observing proposal*. For subsequent data mining including analysis or reprocessing it is essential to have good metadata describing the observation. This includes at least the formalized content (extracted metadata) of the original proposal, details of any source or calibration observations, processing details, and an accurate characterization and identification of the final data product. All data from a given project should be associated in the archive, including the proposal and other project metadata, the original raw observations, any calibrations, and any processed data products including reference and science-grade images and spectra.

From the perspective of VO the data products from each NRAO telescope may be characterized as follows:

- ALMA. According to the current schedule, ALMA should start producing data of interest to the VO in early 2009. The imaging performance of the telescope should be quite good and reference data products should be of relatively high quality. Scientifically the data will be unsurpassed in the millimeter to sub-millimeter regime. The imaging angular resolution will be excellent, of order 10-100 milliarcseconds, comparable to or better than NGST.

- VLA/EVLA. According to the current schedule the interim correlator starts to function in early 2008, and shared-risk science begins in late 2008. It will probably be 2009 or later before significant data is available for use within the VO. In the interim however, much can be done to interface data from the current VLA to the VO. All VLA raw data is already present in the NRAO archive. Interim processing will use the existing VLA archive format until the new correlator comes online, and observing metadata will initially be limited, although the situation should improve as parts of EVLA E2E gradually come online, such as the proposal and project database.

  Not only are VLA pointed observations scientifically interesting for data mining, they will be useful to the VO to help understand how to characterize interferometric radio data, which may have limited spatial frequency (UV) coverage in the generated images, beam-dependent noise characteristics, RFI, subtle artifacts due to the processing, and so forth. As with photon counting high energy instruments, understanding and uniformly characterizing such data will be essential for reliable automated multiwavelength data analysis with the VO.

- VLBA. All historical and new VLBA data (10 TB or so) should be online in the NRAO archive by the end of 2005. Probably the most valuable VO data product for VLBA will be the fully processed, final science-grade images. We estimate that as many as 10000 such images are potentially available, including 2000 images for the VLBA calibrator survey alone. A proposed VLBA survey of GLAST sources could potentially add another 1000 imaged sources. These images have a resolution of several milliarcseconds and would provide ultra high resolution images in the radio of many scientifically interesting astronomical objects, hence could be invaluable for multiwavelength data analysis with the VO. The issues of image understanding are similar to those for VLA pointed observations.

- GBT. Our first goal for GBT is to capture the raw data in the archive in a form which can be useful for data mining some years in the future. During 2005 an initial science data model for GBT and single dish will be defined which can be used to archive raw and calibrated data. NRAO should be able to start archiving GBT data in this form in 2006. Most GBT data currently consists of 1D aperture spectra. In addition to the raw data we would like to store calibrated reference spectra in a form suitable both for analysis (e.g., processed reference spectra in FITS format), and for preview (e.g., JPEG graphics for display). With the introduction of the Penn Array Camera for commissioning and early science in

Winter 2005/2006, the capabilities of the GBT will expand significantly. Because of the large data volumes from this instrument, it will be particularly important to make advances in pipeline processing to optimize the scientific return from the addition. Pipeline processing to produce reference images is expected for a portion of the observing. GBT is also used for pulsar searches and some fraction of this data is potentially reusable for specialized searches for other types of objects.

Software and Services

The various Virtual Observatory projects (IVOA, NVO, Euro-VO, etc.) are defining standards to facilitate the discovery, access, and distributed analysis of data from astronomical archives. To participate effectively in the VO, NRAO needs to make its data and data access services compliant with the VO. NRAO data processing and data analysis software must also be VO-compliant and VO-enabled.

VO-compliance for data and data access services includes the following:

- Generic dataset metadata (in the SDM) should be patterned after the VO standards for dataset identification, coverage, physical data characterization, and so forth. Since VO actively mediates data at access time this is not strictly required, but doing so will result in higher quality data products and will make construction of the VO services easier.
- All permanent archive data products should be tagged at creation time with an IVO dataset identifier to uniquely identify the data product. If the data is subsequently retrieved from the NRAO archive, or replicated in an external archive, this will allow the origin of the data to be unambiguously determined.
- All data products referenced in the astronomical literature, e.g., science grade images published back to the NRAO archive by an observer, should be tagged with an ADS (ADEC/IVOA) dataset identifier. The ADS operates a service which can be used to automatically find and verify datasets identified in this manner.
- It should be possible to return the results from any archive query operation in VOTable format.
- All catalog data should be made available via either the simple Cone Search protocol, or for major catalogs via the query language based basic SkyNode protocol (NRAO has already implemented cone search services for source catalogs from several surveys including NVSS and FIRST). Implementation of the SkyNode protocol will enable cross-correlation of NRAO source catalogs with those from other sources, e.g., with an uploaded user catalog, or with any of the 5000 or so catalogs in the VizieR service from CDS.
- All image data should be made available via the Simple Image Access Protocol (SIAP), which provides for both data discovery and retrieval. At the most basic level entire archive image files are returned. A more sophisticated version of the

image service would return image cutouts (this is similar to a so-called postage stamp service). In the longer term it is conceivable to implement an image service which computes images on the fly with the requested size, resolution, and sky projection from calibrated visibility data. SIAP can also be used to return preview images in a compressed graphics format.

- Raw visibility data from imaging instruments is probably best made available via a SIAP service (a cone search service would also be possible). In this case the returned metadata describes a virtual reference image which could be generated from the raw visibility data, e.g., by retrieving the data and manually processing it to produce an image. Once an on-demand processing capability becomes available the image generation could be done at the archive.
- Spectral data, e.g., reference spectra from the GBT, will be made available via the Simple Spectral Access Protocol (SSAP). Like SIAP, this service provides both data discovery and retrieval. In the case of image surveys which produce spectral data cubes, and SSA service can be implemented to compute 1D aperture spectra on the fly from any spectral data cube.
- All NRAO data collections and services will be entered into the VO registry. This will permit use of the global VO Registry and standard registry-based tools to discover NRAO data and services. In combination with the data access services, VO applications such as DataScope will be able to find and retrieve data and do multiwavelength data analysis. While at present it does not appear worthwhile for NRAO to locally implement a full-up registry, it may be worthwhile to implement a publishing registry to publish NRAO data collections and services.

VO is not only about data description and access; it is also about distributed multiwavelength data analysis. Another aspect of VO thus concerns the software NRAO gives to users to process and analyze NRAO data. We would like our software to be well integrated with VO, and capable of multiwavelength data analysis combining both radio data from NRAO telescopes, with other data obtained via the VO framework. This does not mean NRAO has to write all the software needed for multiwavelength data analysis, rather our software needs to be well integrated with VO, and the result, combining contributions from multiple branches of astronomy, will be capable of multiwavelength data analysis.

In the past each branch of astronomy (e.g., O/IR, radio, high energy) has tended to have its own data processing and analysis software, specializing primarily in data from a single wavelength regime. The result was that astronomers were forced to learn a different system (sometimes more than one) to process each type of data, with minimal interoperability between software systems from different branches of astronomy. This will probably always be the case to some extent as the requirements for data reduction vary greatly for high energy, O/IR, and radio data. A major goal of VO however is to make multiwavelength, distributed data analysis possible. Since the same system is often used for both data reduction and analysis, it is desirable if processing of different kinds of data can be performed from within the same data processing framework as is used for analysis.

In the ideal world an astronomer could use much the same software system to process and analyze data from any branch of astronomy, merely using a different calibration package to process data from, say, an O/IR instrument or a radio or high energy instrument. This would do a lot to unify observing and integrate radio more into the mainstream of astronomy. NRAO, in developing a new data processing framework to follow-on from AIPS and AIPS++, should work to make to make this software compliant with VO standards and conformant to whatever standards emerge for multiwavelength data processing and analysis.

Ultimately such software should be able to operate equivalently on both local and remote data. One form of remote processing is to run the user interface and scripting layer of the data processing system locally on the user's workstation, with all the heavy processing taking place remotely on a computer which front-ends the archive where the data is stored. This avoids the need to move large amounts of data over the Internet, and makes it relatively easy to exploit cluster computing to speed up the processing. Computational tasks to be used in this way can be exposed as Web services to make them available to any remote software (a major part of VO concerns just this sort of capability). For data analysis the analysis system needs to be able to invoke any remote service. The most common example for VO would probably be the data access services mentioned above (SIA, SSA, etc.), however in principle any functionality from a data analysis system (e.g., a processing task) could be exposed as a Web service and accessed remotely in this fashion.

Timeline

What follows is a more in-depth look at what is required to provide the capabilities outlined earlier. Planning for the mid-term and longer-term phases is only approximate at this stage.

Near-Term

Goals

- Integrate selected processed data products into archive
- Implement VO data access services
- Experiment with pipeline processing to produce reference images
- Develop capability to publish user-submitted data
- Perform R&D for distributed multiwavelength data analysis

Activities

1. VLA archive imaging pilot project (**)
2. Select processed data collections to be ingested into archive
    a. existing large and mid-size surveys, calibrators, etc.
3. Integrate data collections into archive
    a. define metadata, ingest data, verify
4. Implement VO data access services
    a. SIA service for raw visibility data
    b. SIA pointed archive and cutout services for image data
    c. Cone search and SkyNode(*) services for catalog data
    d. *SSA service to extract simulated spectra from data cubes
    e. *VOStore interface for file-level data access

    (*) Advanced capabilities optional for initial development
5. Dataset identifiers

    a. ADS-compliant dataset identifier verification service
6. Science data model development
    a. Interferometry (ALMA, EVLA, VLBA) science data model
    b. GBT / single dish science data model
7. Phase II archive infrastructure development
    a. minimal PDM, userDB, proposalDB, cross-indexing, queries
    b. define metadata for user-submitted data products
    c. develop data publication service
    d. investigate next-generation modular storage technology (e.g. NGAS)
8. Archive replication to NCSA
    a. disk-based currently; SRB possible if network improves
9. R&D for multiwavelength data analysis
    a. execution framework, VO-client, data access protocols

Mid-Term

Goals

- Further development of capability to pipeline VLA, VLBA data
- Deploy capability to publish user-submitted data
- Integrate GBT data into archive
- Develop capability to remotely process archival data
- Develop multiwavelength data analysis capability

Activities

1. Integrate additional processed data products into archive
2. Upgrade VO data access services
3. Add 3D support for access to image data cubes
4. Continue experiments to process and mine VLA, VLBA data
5. Develop tools to enable easy publishing of science grade data
6. Pipeline / DRP support for dataset metadata generation
7. Phase III archive infrastructure development
8. Integrate full project model into archive
9. Integrate proposal and project data with science archive
10. Add capability to store SDM-compliant data
11. Implement data capture for GBT to produce SDM conformant data
12. Implement data capture for EVLA to produce SDM conformant data
13. Integrate GBT data into archive
14. Implement remote processing of archival data
15. Develop distributed multiwavelength data analysis infrastructure

Longer-Term

Goals

- Integrate ALMA and EVLA data into archive
- Dataflow from telescope to archive fully automated
- Routine pipeline processing for some subset of observing modes
- On-the-fly processing of data via VO services
- Deploy distributed multiwavelength data analysis capability

Activities

TBD (premature to define now)

(**) The VLA archive imaging pilot project which would produce images for continuum data for the VLA B configuration at 5 and 8.4 GHz, using data from a single semester (late 1999 to early 2000, consisting of about 300 separate observing programs).

NRAO Archive

The NRAO archive appears as one integrated facility to users, but is distributed in terms of implementation. Part of the archive is centralized; this is what users perceive as "the" NRAO archive, and it is this part which is most closely associated with user data access and VO. Other parts are essential for telescope operations and are part of the operational system of each telescope.

The main NRAO archive (http://archive.nrao.edu) includes data for all NRAO telescopes as well as various radio surveys or other data useful to NRAO users. Currently the archive contains about 18 TB of data, mainly from VLA and VLBA. GBT data will be added once the new single dish SDM has been defined and data is available in this form. Significant ALMA data is not expected until ALMA becomes operational in 2009. Once EVLA becomes operational, observational data will flow into the archive at a rate initially of about 80 TB/yr. The data rate for ALMA is similar. The future data rate for GBT is uncertain, but should be less than for EVLA and ALMA, if most pulsar data obtained for specialized programs is excluded. Growth of the archive will be at a somewhat larger rate due to the need to also store processed data products.

The main NRAO archive provides a common interface to the outside world for all NRAO data. This includes the Web and query interfaces, and all VO data access, data processing, data publication, or other services.

In general, data in the NRAO archive will be replicated to at least two sites, to provide an off-site backup as well as (in some cases) additional bandwidth for data access. The master archive for each telescope is part of the facility operated by the telescope. Currently we are in the process of replicating the entire NRAO archive to NCSA. GBT data will be physically replicated to, and integrated into, the main NRAO archive. The primary archive for ALMA data will be in Chile, and ALMA data for the NRAO community will be replicated to the ALMA regional center (ARC) in Charlottesville (additional ARCs will be located in Europe and Japan). Probably there will be no need for additional replication of ALMA data, hence access will be physically provided directly by the CV ARC, but logically ALMA data will be part of the NRAO archive and will be location transparent in terms of user interface, queries, and data access services.

Currently the main NRAO archive is located at the AOC in Socorro. This will remain the case for at least the next several years while development of the archive infrastructure takes place, since the development teams are currently all centralized in Socorro, as is most of the extant data. Once ALMA data starts to flow in 2009, we could continue to operate the science archive in a distributed fashion, or possibly relocate the operational NRAO archive to CV. In either case the primary archive for EVLA and VLBA data will remain in Socorro.

Operationally, the NRAO archive is involved in almost all aspects of observing and data post-processing. The archive is not a monolith but consists of a number of interlinked subsystems and databases. The centralized databases, which are operated by the main NRAO archive staff, include the following:

- User database
- Project database including proposal

- Science archive with science user interfaces (including VO services)

In addition, a number of databases are required for telescope operations. These are located at each site and are operated by the telescope operations staff as part of each observational telescope system. These telescope operations databases include the following:

- Observatory database
- Telescope observations database (raw and calibrated data)
- Calibrations database
- Monitor data database

In terms of implementation, the intention is for the NRAO archive to be largely common with ALMA, with coordinated development. Hence a common science data model is being developed for all the interferometers (ALMA, EVLA, VLBA). The GBT single dish data model may be optimized for GBT but will be similar in terms of metadata, components, and data representation to the interferometry SDM. Major subsystems such as the user database and the project database including the proposal are common to all telescopes (the content may differ to some extent for each telescope but the model, container, representation, and in many cases the implementation will be the same). Final science data products such as images, spectra, spectral data cubes, etc., are similar or identical (except for telescope-specific metadata) for all telescopes and will conform to VO and FITS standards. In most cases, common VO services for queries and data access will be used for all telescopes.

The current archive uses all disk-based storage, and this is the plan for the future as well. For storage technology ALMA plans to use the next generation archive system (NGAS) from ESO. This is a modular, scalable, disk-based storage technology which uses a PC with a RAID card and 8-16 disks as the storage unit. The same system is used for disk-based data transfer between sites. The NGAS software was recently split off from the proprietary ESO version and open-sourced for use by ALMA, and is available for use by NRAO as well. While no decision has yet been made on the future storage technology for the NRAO archive, NGAS is a strong possibility now that it has been open-sourced.

The various archive subsystems, some of which are distributed to the three main NRAO sites, are interlinked at runtime and in terms of dataflow to the main NRAO archive. The dataflow is discussed in the next section.

Dataflow

Proposal submission is the main entry point to the system and is an NRAO-wide facility. A single proposal may request time on more than one NRAO telescope. An approved proposal becomes a new project in the project database, which is used to describe projects and link all project data products together in the archive. For an observational database, linking project data products is important, even for VO data access, to associate science data products such as images back to the raw observational data for reprocessing, to provide full access to project metadata, and to provide secure access to project data during the proprietary period.

The telescope projects are responsible for observing, including observation preparation, scheduling, and operating the telescope to take data. The telescope projects are also responsible for operation of any automated pipelines used to produce reference data products for the archive. The science archive, including any VO services used to provide access the data, is NRAO-wide.

The project database is part of the central NRAO archive. For observing, the telescope system needs to maintain some project information locally. Raw observational data when obtained is stored locally in the master observational archive for the telescope. Periodically new observations, including updated project metadata, are sent to the main NRAO archive. The new data products ingested and the project metadata is updated.

The dataflow for automated pipeline processing is similar. The telescope project runs the instrumental pipeline to produce new reference data products. These are checked into the NRAO archive (replicating the data in the process), and the project metadata is updated. A telescope archive may or may not permanently store such derived data products. The most important function of the primary telescope archive is to store raw observational data.

Most user access to data is via the main NRAO archive, which integrates all NRAO data and provides common user interface, query, and data access services, including the VO services. The primary emphasis of the main NRAO archive is on science data access and processing. The individual telescope archives may have an independent Web interface and may manage data for specialized programs which never makes it to the main NRAO archive, perhaps because it is too specialized to be of general interest, or because it is too voluminous to be permanently archived (e.g., GBT pulsar data).

Network bandwidth for user access to data is currently an issue. Internet2 provides substantial bandwidth, but the link from the main NRAO archive in Socorro to the Internet2 GigaPop in Albuquerque is currently limited to only 5-10 Mbps (at this rate it takes approximately 1 month to transfer 1 TB of data via the network, assuming the full bandwidth can be sustained). Options are currently being explored to increase this bandwidth. These include tying into the LambdaRail network currently being set up in New Mexico (this could happen as early as summer 2005 and could incur no or minimal cost to NRAO), or buying additional bandwidth from commercial carriers. Replicating bulk data to a site such as NCSA which already has large network bandwidth could increase the overall network bandwidth for user access to data, but (even using physical shipment of hard disks to charge the replica) requires significant network bandwidth to generate and maintain an up to date archive replica.

Support Needed from NRAO

This report represents a first attempt to look at what is required to interface NRAO data to the VO, and integrate VO into NRAO operations. As such, only a preliminary attempt has been made to estimate the resources required for such an effort. Generating such an estimate is complicated by the significant overlap of VO with E2E and science software development. As a guideline we have tried to look mainly at near- and mid-term development of the science archive, including archive infrastructure development, user

interface and VO services, data ingest, data management, and quality control, experiments to mine the VLA archive by producing reference images, and operations. Many aspects of E2E, e.g., project and science data model definition, and development and operation of data processing pipelines by the telescope projects, are excluded here, but would also be required to implement the full program.

Near-Term Activities (ends 2007.0)

Activity	FTE/yr	FTE type(*)	Associated Costs
VLA archive imaging pilot project	0.5	Sci	
Select data collections for ingest	0.1	Sci	
Integrate collections into archive	1.0	SP	2-4 TB storage
Implement VO data access services	1.0	SP+T	
ADS dataset identifiers	0.2	T	
Archive replication to NCSA	0.2	T	better transport?
Phase II archive infrastructure	2.0	SP+T	storage $50K
Database Administrator	0.5	T	
Comp. Div. Support	0.5	T	

(*) Sci – scientist (NRAO scientific staff), SP – scientist/programmer (e.g., J.Benson), T – technical staff, e.g., system administrator, DBA, programmers.

A minimal near-term archive and VO program would require 4-5 FTE (one SP lead, one SP data aide, 1-2 developers, one FTE operations support, not counting oversight, planning, and associated activities by the research staff). Not all of these FTE would necessarily need to be new hires. The current archive staff is about 2 FTE including operations. Related research activities such as the VLA archive imaging pilot project and data analysis framework research are being carried out primarily by the scientific staff and are not included here. Ongoing hardware and software costs are currently about $50K/yr. All computer hardware and storage must be replaced every 3-4 years.

Mid-Term Activities (2007.0 – 2009.0)

Activity	FTE/yr	FTE type(*)	Associated Costs
Integrate new data products	1.0	SP	n TB storage
Upgrade VO services	0.5	SP+T	
Tools to enable easy publishing	0.5	SP+T	
Integrate GBT data into archive	0.5	SP	n TB storage
Phase III archive infrastructure	2.0	SP+T	NGAS storage $200K?
Remote processing of archive data	1.0	SP+T	compute cluster
Database Administrator	0.5	T	
Comp. Div. Support	0.5	T	

The level of effort required in the mid-term is at least comparable to what is required to initiate the program, and may well require additional resources, as much of the development for the operational (2010) archive occurs in the mid-term, and archive

operations can be expected to increase in cost as the archive becomes more functional and is used more.