Ackerman, Mark S., Erik Hofer, and Robert Hanisch. "Collaboratory: The National Virtual Observatory." Gary Olson, Ann Zimmerman, and Nathan Bos (eds.), Science on the Internet, MIT Press, 2007.

# Chapter 7:
# The National Virtual Observatory

Mark S. Ackerman, Erik C. Hofer, and Robert J. Hanisch

Like many scientific communities, the astronomy community faces a coming avalanche of data as instrumentation improves in quality as well as in its ability to integrate with computational and data resources. Unlike scientific fields that are oriented around a small number of major instruments, such as high-energy physics, astronomers use a large number of telescopes located around the world that are designed and calibrated to look at celestial objects in fundamentally different ways. Both space and terrestrial telescopes are designed to observe objects across a narrow part of the energy spectrum, typically focusing on a small part of the spectrum from the infrared to X-ray wavelengths. While each telescope has the potential to reveal and characterize new astronomical objects, even more powerful would be the ability to combine the data produced by each of these instruments to create a unified picture of the observable universe. This data fusion requires federating a large number of data sets, and developing the search and analysis routines that allow investigation across multiple wavelengths.

The National Virtual Observatory (NVO) project is funded by the National Science Foundation (NSF) to provide the cyberinfrastructure necessary to support the federation of a large number of astronomical data sets, allowing search across multiple data sets and the development of simulations that incorporate many types of astronomical data.[1] Through the development of tools and standardized data models, the NVO hopes to enable the combination of multiple pointed-observation telescopes and sky surveys into a large, unified data set that effectively functions as a broadband, worldwide telescope. The NVO is part of a larger effort, known as the International Virtual Observatory Alliance (IVOA), to support data federation and exchange across a number of national and regional virtual observatories.[2]

**The Coming Data Avalanche**

Astronomy is undergoing several revolutions. Like many sciences, new instruments and digital capture provide orders of magnitude more data. One example project, the Sloan Digital Sky Survey, will map more than one hundred million distinct objects.[3] At present (Sloan Digital Sky Survey data release 1) it has mapped only 1.6 percent of the sky, but has already obtained data on fifty-three million objects (see also chapter 1, this volume). Many of these objects, stars, quasars, and galaxies will be mapped multiple times with photometry and spectroscopic measurements. Gaia, a European space-based observatory, will survey one billion objects.[4] Target stars will be monitored in detail about one hundred times over its five-year mission. These surveys will not only map the sky in more detail than ever before; astronomers will also use them to find new objects (such as new brown dwarfs)—and hope to find new classes of objects. They will also find sources for subsequent investigations, such as sources for later X-ray or gamma ray bursts and microlensed (small-scale gravitational lensing) events.

These two surveys are only some of the new observatories coming online. Of the National Aeronautics and Space Administration's four "Great Observatories," Hubble is the most famous.[5] The other two existing Great Observatories are the Spitzer Space Telescope (formerly SIRTF) to observe in the infrared and Chandra to observe in the X-ray. (The Compton gamma ray observatory mission has already ended.) In addition to these four, there are other space- and earth-based observatories, all producing data. Some are current, such as the FUSE (far ultraviolet spectroscopic explorer) mission. Others are planned, such as the James Webb observatory, the successor to Hubble, in process for a launch in 2011. There are also numerous European and Japanese efforts.

These observatories are all exceptional instruments. As an example, the Chandra X-Ray Observatory cost $1.65 billion for development, $350 million for launch costs, and $750 million for operations and data analysis in the first five years. It is able to obtain images at twenty-five to fifty times better resolution than previous X-ray telescopes. The resolution is the equivalent, according to the Chandra science Web site, "to the ability to read a newspaper at a distance of half a mile."[6] In addition to being able to observe black holes and supernovas, this capability provides the ability to do detailed studies of dark matter.

As might be expected from occasional development costs of a billion dollars, these are also extremely complex projects. Four factors contribute to this complexity.

First, these projects can differ in their goals. While all produce huge data streams, the surveys obviously produce the most. Space-based missions tend to target specific objects, based on astronomers' observing-time proposals, but this is shifting toward survey work as well.

Second, different observatories operate in different "wavelength regimes." Spitzer, for example, observes the infrared (between visible and microwave wavelengths at between approximately 0.75 μm and 0.1 mm), but it does not even cover that entire spectrum. As mentioned, Chandra operates in the X-ray, and Compton operated in the gamma ray. Hubble operates in the visible as well as near infrared and ultraviolet (both close to visible light). This mirrors the traditional division of the astronomy community. Each wavelength regime and subcommunity has its own data formats, which are unlikely to change, since the detectors and data can be quite different.

Third, the complexity and costs of the observatories and projects themselves are reflected in the complexity of the institutional arrangements. The Sloan Digital Sky Survey telescopes are at Apache Point Observatory, operated by a consortium of astrophysical institutions. The Sloan Digital Sky Survey itself is a joint effort of thirteen universities and research institutions. While a single institution often controls satellite missions (for example, the Center for Astrophysics at Harvard runs Chandra), the planning is multi-institutional and frequently international.

Finally, as will be discussed below, the data capture and storage are similarly complex, and often idiosyncratic to the institutions and mission involved.

## The Coming Revolution in Astronomical Data Analysis

The revolution of increasing capabilities of the observational instruments and the resulting huge volumes of data are likely to be mirrored in a second revolution. It is thought that a substantial transformation in astronomical and astrophysical work is about to occur. Traditionally, astronomers working across wavelength regimes were rare. It is clear, however, that some research question can be studied best by combining data in different wavelengths. This is particularly true with phenomena that are changing—for example, bursts or supernovas.. Additionally, the large quantities of data and their automatic capture make it possible to watch for dynamic situations—that is, to provide triggers for the automatic detection of changing phenomena. Astronomers to date have lost the precious minutes after the detection of a supernova to bring many instruments to bear on the new supernova.

While the availability (or rather, the potential availability) of data makes new analysis possible, it cannot be overstated that this is a fundamental shift in the nature of the analysis work. Currently astronomers work within a wavelength regime. Like any scientist, their sources of expertise and help are all within their own subcommunities; they understand the instruments and data sets within their own subcommunities, and publish and garner credentials within those subcommunities.

We have also been constantly struck with the concern by astronomers that the next generation will be "armchair astronomers." As stated in study interviews and at conferences, they believe that it will be possible in the near future for astronomers to no longer spend observational time taking data and controlling an instrument, but to merely be able to summon data from these vast data repositories for analysis. This concern, regardless of its merit, clearly reflects an understanding that astronomical analysis—always considered to be at the heart of the profession—is changing and will continue to change in nature.

## Building a New Kind of Observatory

To enable the kind of inquiry that will allow astronomers to rely on a shared data resource rather than a shared instrument resource, a new type of observatory has to be built. This new observatory will not focus on a single instrument as the focal data provider but rather will engage a network of existing and future instruments, enabling the publication and federation of data from that instrument network. As noted in the chapter opening, the NVO is an NSF project to design and build the cyberinfrastructure needed to support this new kind of observatory. The goal of the NVO project is to prototype this new type of observatory. Using information technology to federate the newly available quantities of data, the hope is that astronomers can work in multiple wavelengths, consider temporal phenomena, and perform new forms of statistically based analyses. It is, in fact, an answer to the increasing amount of data, and the astronomers' inability to find, access, or utilize most of it.

A large number of institutions are participating in the NVO project. Members of the collaboration include astronomers, astronomy programmers, and computer scientists from universities and observatories across the United States. The project is co-led by an astronomer and a computer scientist, and is funded by the Information Technology Research program at the NSF (chapter 17, this volume), with oversight from both the Computer and Information Science and Engineering Directorate and the Astronomy and Astrophysics Division. The

NVO's external advisory board also reflects this disciplinary split between computer science and astronomy, with membership of prominent researchers from each field. As with many cyberinfrastructure projects, managing the research interests of both computer and domain scientists can be challenging (see also chapters 17 and 18, this volume). While this problem of competing interests has been problematic for some other collaborations, the technical sophistication of the astronomers working on the project spans many of the gaps that would be expected, causing this partnership between computer science and an application science to work extremely well.

In addition to this internal diversity, the NVO is also the U.S. participant in the IVOA, a larger federation project (see also chapter 1, this volume). The IVOA consists of fifteen similar virtual observatory projects around the world. This coordinating organization serves as a venue for interoperability testing and standardization to ensure that the kind of federation possible on a national scale can also occur on an international scale. The NVO has been critical to the creation of several standards for data formatting and interchange, but must also work to coevolve with other virtual observatories to identify and meet the common needs of IVOA members.

### *Technical Overview*

Technically, the NVO consists of six layers and many components. In this, it is like many other cyberinfrastructure projects. The following is a brief description of the NVO from a technical point of view. (The details are simplified here for clarity; this description will only overview what is required technically.) The NVO includes several efforts that show the requirements for coalition infrastructures. These are being developed by groups of highly geographically distributed people. These efforts include:

- Virtual observatory registries: Registries allow users and applications to search for data. They indicate what observational or derived data are in the archives.

- Query services: These allow astronomer-users to query registries and search for data. As mentioned, this is not simple, as the search occurs within a 3-D space, and on potentially noisy, differently formatted, and incomplete data.

• Portals, analysis procedures, and client tools: These present the NVO capabilities to the astronomer-user. They include, for example, the OpenSkyQuery tool, which can provide data from ten different astronomical surveys.

• A data-access layer: This layer maps the retrieval to actual physical data. An interesting problem is how to provide a common format for data, and there is considerable work being done on standardizing these formats. An international effort has been made to create a standards process for virtual observatory projects.

• Data models: These are detailed data models of various entities important to the NVO. Currently, work is going on to model observational data and a few other simple entities. Additional work is going on to model data tables in archives—how the data are currently stored in archives and what can be simply compared.

• Metadata: Metadata will be automatically attached to queries, and it is hoped that archives will provide more metadata. Currently, there is little standardization within astronomy of this metadata. The initial efforts within the NVO are to standardize coordinate system and provenance metadata.

Figure 7.1, from an NVO report, shows the NVO system architecture.[7] The figure pays more attention to the data services layers. The user layer, which consists of NVO services, applications, and portals, gets short shrift in this figure. Nonetheless, it shows that the NVO architecture is distributed, multilayered, and reasonably complex. New services can be introduced at any layer without affecting the others.

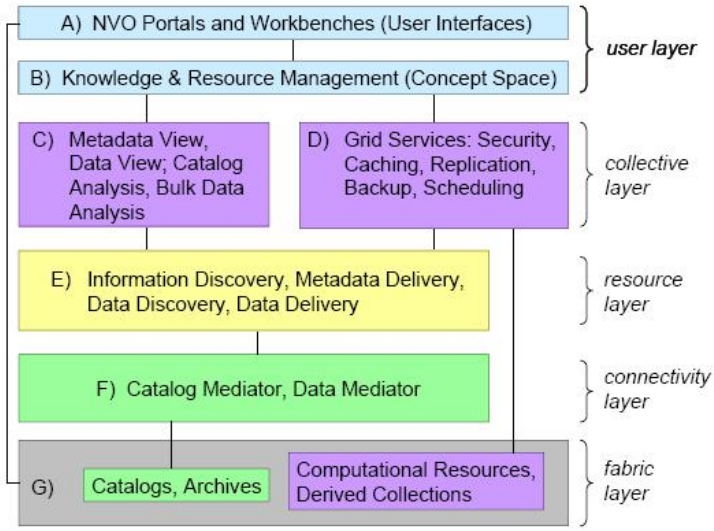


Figure 1: The NVO architecture. Figure is from http://www.us-vo.org/pubs/files/Y2-annual-report1.pdf.

The architecture is perhaps less interesting than what the NVO must do. Figure 7.2 depicts what astronomers want to do.[8] In practice, astronomers would deal with more raw images or data, but figure 7.2 shows that important details are provided in different wavelengths.

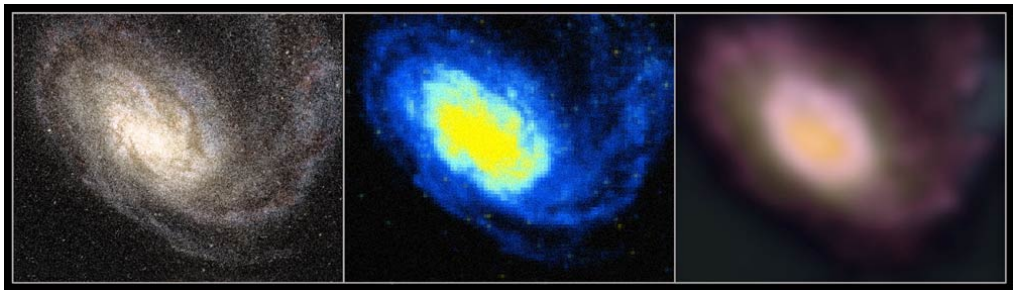

Figure 2: A galaxy in visible, radio, and x-ray. Figure is from http://www.euro-vo.org/avo/gallery/diff_wavelengths_1000.jpg. (Original is in color.)

One should note that this type of data federation has never occurred within astronomy before. While the automation of this process is desired, the NVO team understands that it is more likely that astronomers will not accept a solution that does not allow them to understand the entire retrieval and conversion process. To scientifically interpret any image (or other astronomical data) requires a complete understanding of the observing conditions, the source instrument, and the transformations applied to the collected data. Astronomers require a detailed

understanding of their data; it is an open question how to facilitate that understanding.

Another key area of technical effort has been the standardization of protocols, data formats, and data models. Astronomers have long understood the significance of data standards in data preservation and interoperability. The Flexible Image Transport System (FITS) data format was developed as a standard to solve exactly these problems.[9] The group that created FITS, however, was large and represented many different subcommunities with many different technical goals—the end result being a standard that was extremely general. The huge number of variations due to the different use conventions of FITS has lead to a data standard that is extendable to the point that the standard is virtually unenforceable. Even so, the standard has been crucial to the astronomy community, not just for its intrinsic value as a format, but also as an example of what can happen when standardization efforts do not force difficult decisions to be made.

Largely because of FITS, the NVO group and the larger IVOA community are actively pursuing standards that are meaningful, and balance the need for extendability and adaptability to many uses with the need for some things to remain the same. Presentations, discussions, and arguments about the details of and the need for standards are a major component of the IVOA interoperability meetings that are held twice a year.

### *Challenges and Outreach*

Notably absent from this technical description of project deliverables are the new analysis and comparison tools that will need to be built in order to take advantage of these federated data. A major challenge of the project, which has been funded to provide the basic framework for the NVO cyberinfrastructure, is how to stimulate use of the tools to support the new types of scientific inquiry that the project will enable. Some technology demonstrations, such as producing a merged, multiwavelength image of a single object, are relatively simple given the tools that the NVO has developed for matching and reorienting different data sets so that they are comparable. All that is required is to adapt the existing imaging tools to import data from the NVO. More sophisticated as well as challenging are efforts to build new applications that integrate simulations and analyses with the NVO-federated data, such as a demonstration presented at the 2004 winter American Astronomical Society conference that allowed users to compare theoretical simulations of globular clusters with observed globular clusters in the NVO.

Building these applications faces two challenges, and the NVO is developing solutions for both of them. The first is that astronomers have to be trained how to write programs that leverage the powers of the NVO. This requirement creates a dependence on scientists to do the articulation work (Strauss 1993) (or coordination work) required to connect the NVO capabilities and current scientific practice. The NVO team is only providing the infrastructure and the interfaces required to use that infrastructure; widespread scientific advancement because of the NVO depends on the ability and willingness of others to create the applications that enable discovery. To meet this challenge, the NVO team is targeting junior members of the community by hosting an annual applications "boot camp." This camp allows members of the project team to work with students, postdocs, and junior faculty to help them build applications to introduce them to working with the NVO. This strategy has so far proved successful, with the 2004 summer school resulting in a number of application- and infrastructure-focused student projects.

The second challenge the NVO faces in realizing the scientific vision of the project is the difficulties in training young astronomers to work across wavelength regimes. As noted previously, most astronomers are trained to work within a single wavelength regime. They become experts in the physics of a particular type of observation, and focus their observation at the limits of what is observable. Moving into a different wavelength regime requires mastering an understanding of an entirely new set of observational physics. Furthermore, trying to combine these different observational techniques is even more challenging. In short, the challenge of multiwavelength astronomy is not going to be met just by technical infrastructure. New students must be trained to work across regimes, and think about observation and analysis in fundamentally different ways. The big question now is who will train them—only a handful of astronomers are attempting to work across wavelengths, and many of them are not yet out of graduate school.

Project leaders are hopeful both of these challenges will be overcome. As the possibilities of new discoveries grow, astronomical practice will follow.

In summary, the NVO and the International Virtual Observatory offer a cyberinfrastructure for the next step in astronomy. As more and more space-based observatories as well as earth-based surveys come online, the NVO offers the possibility of doing new types of science, providing astronomers with more data capabilities than they have ever had before. The NVO is likely to change—and revolutionize—astronomy.

**References**

Strauss, A. L. 1993. *Continual permutations of action*. New York: Aldine de Gruyter.

---

**Notes**

1. More information about the NVO is available at <http://www.us-vo.org>.

2. More information about the IVOA is available at <http://www.ivoa.net>.

3. More information about Sloan Digital Sky Survey is available at <http://www.sdss.org>.

4. More information about the Gaia mission is available at <http://sci.esa.int/science-e/www/object/index.cfm?fobjectid=28820>.

5. More information about all of the Great Observatories is available at <http://www.nasa.gov>.

6. Chandra X-ray Observatory, available at <http://chandra.harvard.edu/about/telescope_system3.html>.

7. Figure 7.1 is from <http://www.us-vo.org/pubs/files/Y2-annual-report1.pdf>.

---

8. Figure 7.2 is from <http://www.euro-vo.org/avo/gallery/diff_wavelengths_1000.jpg>.

9. More information about FITS is available at

http://heasarc.gsfc.nasa.gov/docs/heasarc/fits.html.