Enabling Cyber-Discovery in the Next Decades
The Role of the Virtual Observatory

Executive Summary

Cyber-discovery is a science frontier discovery area identified in *New Worlds, New Horizons*. It involves not only the increasing amounts of data being collected by astronomical instruments, but also the increasingly sophisticated analyses that can be conducted on the data and the complementary complex simulations that can be performed. An impending challenge is that the volumes of data sets (both observed and simulated) are at the stage at which movement of the data to the astronomer is becoming impractical. Further, the NSF has recognized, on an agency-wide basis, that cyber-infrastructure—including data preservation, curation, integration, and discovery—is an essential ingredient in fostering transformative scientific research.

The long-term concept for the virtual observatory (VO) is seamless access to data, tools, and the literature: An astronomer at any institution should be able to access the catalogs, databases, and images—regardless of their wavelength range or size—required for his or her science interest; mine these data for relevant sources; and visualize the results. Within the U.S., the Virtual Astronomical Observatory (VAO) is (i) **developing and maintaining standards and protocols** for storing and interchanging data of all types, (ii) **developing cross-cutting tools and services** for accessing and comparing multi-wavelength data, and (iii) **providing for data curation** at all levels. The VAO is implemented via partnerships at multiple levels. Funded jointly by NSF and NASA, the VAO provides international leadership via the International Virtual Observatory Alliance, and it works with individual researchers and projects through science collaborations to develop and expand software tools, services, and libraries.

Information technology has had a profound impact on astronomical research in the past decade. In the next decade, continued IT advances and an increasing reliance on large databases will have a similar effect. Continued support for the VO will be an essential component of conducting the *NWNH* science program.

Astronomical Data and the *New Worlds, New Horizons* Science Program

New Worlds, New Horizons (*NWNH*) discusses the role of astronomical data to be acquired by existing and future facilities both explicitly and implicitly. Discovery is the first science theme in Chapter 2, where it is noted, "New technologies, observing strategies, theories, and computations open vistas on the universe and provide opportunities for transformational comprehension." Cyber-discovery is also described as required in order to give meaning to data collected. It is not enough to collect data—they must be able to be accessed, searched, potentially "cross-[correlated] … at different wavelengths," and even reused for purposes not envisioned by the original investigators for maximal science return.

More generally, the themes of Chapter 2 are so fundamental and so far-reaching that is unlikely that significant progress will be made by observations from a single telescope, or

even in a single portion of the spectrum. Indeed, many of science questions posed make use of multi-wavelength data, and there are numerous examples of comparisons between ground-based TeV, optical, infrared, and radio observations described. Examples include

- Combining optical, millimeter, and radio data to understand the first sources of light in the Universe and their impact on the surrounding intergalactic medium. ("What were the first objects to light up the universe and when did they do it?")
- Combining TeV, optical, and radio data to understand how black holes function. ("How do black holes work and influence their surroundings?")
- Using infrared, millimeter, and radio techniques to detect and study complex organic molecules in interstellar space. ("The Chemistry of the Universe")
- Identifying and classifying transients and variables. ("Time-Domain Astronomy")

Scaling for the Future

Enabled by a variety of technological developments, such as large format CCDs and high-speed digital signal processing chips, the data volume being acquired and processed by astronomers is on an exponential growth path. Astronomy is not unique in this regard; NSF's Office of Cyber Infrastructure describes science as being transformed by "advanced computational facilities (e.g., data systems, computing hardware, high speed networks) and instruments (e.g., telescopes, …) coupled to the development of quantifiable models, algorithms, software and other tools and services to provide unique insights into complex problems in science …." Figure 1 illustrates this trend in astronomy, showing the size of the archives at two NSF Facilities. This acceleration is likely to continue throughout this decade, and into the next. For instance, the Atacama Large Millimeter/Sub-millimeter Array (ALMA) is currently beginning early science operations and its archive is expected to be of order 2 PB after a decade of operation (Lacy et al. 2011), while the LSST archive is expected to exceed 50 PB (Kantor & Axelrod 2010).

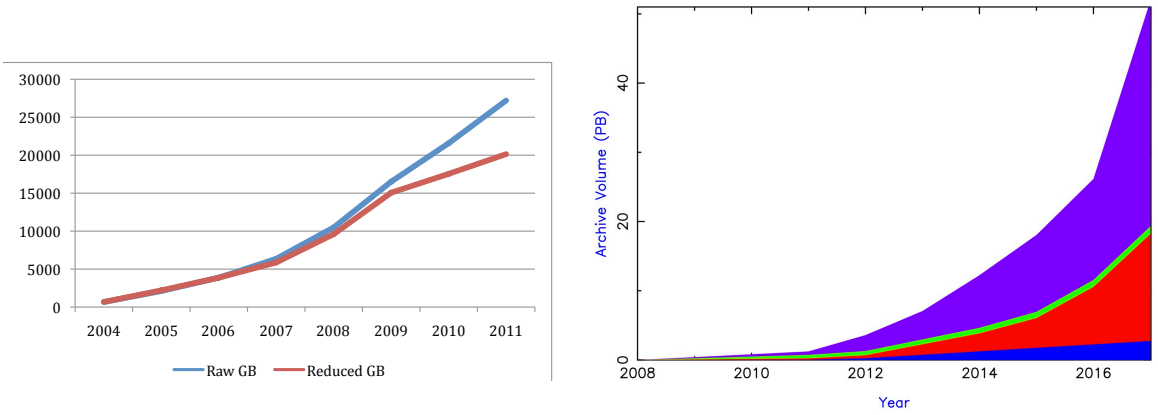

Figure 1. NOAO and NRAO Data Archives as a function of time. (*Left*) The NOAO Data Archive volume, measured in Gigabytes, for the data collected over most of the past decade showing both the raw data volume (blue) and that following a change in file format for the storage (red). The data volume is expected to continue to increase driven by a combination of continued acquisition of new data combined with transferring data currently stored on tape to the current Archive. (*Right*) The NRAO Data Archive volume, measured in *Petabytes*, showing the projected increase as ALMA (blue), the EVLA (red), and the GBT (green and purple) acquire data over the course of this decade.

The increase in the data volume in science archives produces two additional impacts on processing and storage systems. First, as data volumes increases, there is a concomitant increase in demand as more astronomers request larger amounts of data, often using more complex queries, which in turn require additional processing to supply the requested data. Second, combining and analyzing multiple data sets, such as for multi-wavelength analyses, requires both increased processing power and increased sophistication in designing the tools and applications for accessing and manipulating the data.

Needed Capabilities

In order to make maximal use of existing data, optimize the use of observing time (a scarce resource), and enable discoveries by allowing investigations not envisioned when the data were collected, the following set of capabilities are needed, which we list in *priority* order. A challenge associated with obtaining these needed capabilities is the relevant time scales; in contrast to telescopes, which evolve on decade time scales, computational capabilities and data volumes evolve on roughly yearly time scales.

1. **Development and maintenance of standards and protocols** for storing and interchanging data of all types, including images, spectra, time series, catalogs, simulations, and *metadata* about the data. The power and the durability of standards are well illustrated by the Flexible Image Transport System (FITS, Wells et al. 1981; Hanisch et al. 2001; Pence et al. 2010). Developed when CCDs were just coming into wide-spread use and radio interferometers were beginning to produce images routinely, it has been recognized and adopted as a format for storing images at many wavelengths and ultimately has been extended to describe and transmit data of many types.

   As discussed below, the VAO and its predecessor have been and continue to be involved in the development of international standards. As the FITS example illustrates, standards are not static. Rather they must evolve as science questions develop, new data are acquired, and telescopes become more capable. A recent example of this need for continued development is the growing interest in time-domain astronomy.

2. **"Seamless" crosscutting tools and services** for discovering, accessing, and comparing data and the scientific literature. Tools and services that are not tied to a specific observatory—as in the present paradigm—but that are capable of ingesting, analyzing, and displaying data from multiple facilities are required in order to exploit the full range of data being generated and to address the *NWNH* science program.

   The long-term concept for the VO is seamless access to data, regardless of environment: An astronomer at any institution should be able to access the catalogs, databases, and images, regardless of their wavelength range or size, required for his or her science interest; mine these data for relevant sources; and visualize the results. These tools can be developed in a variety of models: (i) By the VAO, in cases for which specialized knowledge is needed, e.g., the development of the capability to cross-match large, multi-wavelength databases (in the spirit of the *NWNH* statement about multi-wavelength cross-correlation); (ii) In a collaboration between the VAO and a project, in which the VAO assists in ensuring that the project develops VO-enabled tools so that other researchers can access and re-use the data (see Partnerships below); or (iii) By individual researchers developing tools based on VO-developed libraries and standards.

3. **Data curation** at all levels, from National facilities to individual investigators. While archiving of (large) data sets used to be considered only the provenance of National facilities, the NSF has introduced a Foundation-wide requirement for investigators to discuss data management in proposals in recognition that even relatively small budget activities are capable of generating large or complex data sets.

   It is not enough to generate and publish data—there must be sufficient and *accurate* metadata so that the data can be re-used for other purposes. The VAO is a key player in maintaining a registry from which data, with accurate metadata, can be identified and retrieved.

Partnerships

Astronomy has been recognized in the broader e-science community as a showcase example of data sharing and virtual scientific organization, and the concept of a Virtual Observatory has been developed via partnerships. The VAO has been implemented under a new model as a limited liability corporation, co-created by AUI and AURA, and structured so as to be maximally responsible to community-wide needs and priorities. The VAO is supported jointly by NSF (~ 2/3) and NASA (~ 1/3), and the National Observatories and NASA data centers are collaborators. Scientific collaborations are being established with research groups and projects: The VAO recently announced a call for proposals to develop tools and services jointly with research groups, and major survey projects (e.g., PanSTARRS, LSST) have begun to collaborate with the VAO.

NWNH (Chapter 3) describes how other nations and regions are mounting increasingly ambitious missions and telescopes. The highest quality science is produced when astronomers can access data regardless of national boundaries. The International Virtual Observatory Alliance (IVOA) is a forum of 20 national projects for coordinating the standards and services used to store, access, and transmit data. U.S. astronomers and software engineers play a significant role in the development of these standards and the IVOA.

- Many of the current VAO team members, as part of the former National Virtual Observatory, were instrumental in forming the IVOA in 2002 and have assumed leadership roles since. The VAO Director, R. Hanisch, served as the first IVOA Chair, and D. De Young (former VAO Project Scientist) also served as IVOA Chair.
- Of the nearly 40 IVOA documents (Recommendations or Proposed Recommendations), U.S. astronomers have been the lead authors on over 40% of them, with U.S. astronomers contributing in total to nearly 30% of the uniquely named co-authors.
- Since the establishment of the VAO (in 2010), U.S. astronomers have served as the chair or vice-chair of an IVOA Working Group for almost 30% of the time.

U.S. engagement in the IVOA has a dual purpose. First, it allows the U.S. to provide leadership in an increasingly multi-polar world. Second, it leverages international investments in standards and services in a manner that allows data from U.S. telescopes to be augmented by data from international telescopes.

State of the Profession and Societal Impact

Cyber-discovery presents both a challenge and an opportunity: The challenge is that the needed skills are ones in which astronomers traditionally have not been trained; the opportunity is that the data volumes and processing involved attract the attention of other communities (e.g., computational sciences, statistics, IT industry). Further, students trained to manage and process of large data volumes have skills relevant to many science fields and even the larger knowledge-based economy (e.g., e-commerce). Finally, VO activities are aligned with the data sharing and interoperability recommendations in "Harnessing the Power of Digital Data for Science and Society" (IWGDD 2009).

Current Status

The Virtual Observatory (VO) was a recommended project in the previous Decadal Survey (*Astronomy for the New Millennium*). Over the past decade, the National Virtual Observatory (NVO) conducted a research program to begin to develop the standards for a range of astronomical data and an underlying software infrastructure. That work has now been transitioned from a research focus (NVO) to an operational focus with the establishment of the Virtual Astronomical Observatory[1] (VAO).

The VAO has begun to roll out the first suite of operational software designed for multi-wavelength and time-domain astronomy in the modern computing environment. These tools include a scalable cross-matching tool designed for rapid multi-wavelength identification of sources in large catalogs, a tool for constructing spectral energy distributions from data stored both locally on the user's machine and stored remotely in VO-aware services, a discovery tool designed to help identify both catalogs and images worldwide relevant to a region of interest, and improvements to the IRAF software to make it aware of and able to use other VO standards and services. The VAO has also pressed forward in the development and maintenance of international standards in order to support emerging research needs and new, complex astronomical data types.

References

Hanisch, R. J., et al. 2001, "Definition of the Flexible Image Transport System (FITS)," A&A, 376, 359

Interagency Working Group on Digital Data (IWGDD) 2009, "Harnessing the Power of Digital Data for Science and Society," Report to the Committee of the National Science & Technology Council

Lacy, M., Halstead, D., & Hatz, M. 2011, "Data Processing and Archiving at the North American ALMA Science Center," in *Astronomical Data Analysis Software and Systems XX*, ASP Conf. Series, Vol. 442, eds. I. N. Evans, A. Accomazzi, D. J. Mink, & A. H. Rots (ASP: San Francisco) p. 57

Kantor, J., & Axelrod, T. 2010, "The Large Synoptic Survey Telescope data management overview," in *Software and Cyberinfrastructure for Astronomy*, Proc. SPIE, Vol. 7740, eds. N. M. Radziwill & A. Bridger, (SPIE) p. 77401N

Pence, W. D., et al. 2010, "Definition of the Flexible Image Transport System (FITS), Version 3.0," A&A, 524, A42

Wells, D. C., Greisen, E. W., & Harten, R. H. 1981, "FITS: A Flexible Image Transport System," A&AS, 44, 363

[1] http://www.usvao.org/