

The Management, Storage and Utilization of Astronomical Data in the 21st Century

A Discussion Paper for the OECD Global Science Forum

Peter Quinn¹, Andy Lawrence², and Bob Hanisch³

¹European Southern Observatory; Director, Astrophysical Virtual Observatory [Europe]; Chairman,
International Virtual Observatory Alliance

²Royal Observatory Edinburgh; Project Leader, AstroGrid [UK]

³Space Telescope Science Institute; Project Manager, National Virtual Observatory [US]

15 March 2004

Summary

The costs and resources associated with the development of forefront astronomical research capabilities often greatly exceed the funding capacities of individual universities, research organizations, and nations (e.g., Atacama Large Millimeter Array [ALMA], The Square Kilometre Array [SKA], and Extremely Large (optical) Telescopes in the 30m-100m class [ELTs]). Collaborative alliances of organizations and nations are therefore being formed to build new, facility-class astronomical observatories across the globe. This expansion and globalization of the astronomical research effort raises a number of major issues that need to be confronted and solved by astronomers, research funding bodies, and governments. Some of these issues are being met by other sciences and some are unique to the research diversity inherent in exploring the Universe through multiple, complementary wavelength windows. In all cases, the challenges of managing, maximally utilizing, and collaboratively sharing the huge volume of digital information flowing from these new observatories is focusing and leading the discussion of critical issues for success. This discussion paper seeks to identify a number of these major issues, to highlight a new vision for a common research infrastructure that will enable some of these issues to be addressed, and further, to identify some of the practical and policy issues associated with long term solutions and the maximal return on global research investments.

Issues for A Global Astronomy

[1] How will global astronomical research projects meet the data volume and computational challenges associated with tackling forefront research problems?

The current and projected rate of data volume growth from new and planned observatories has a doubling time of order six to twelve months. This is faster than the measured rate of increase in performance of computer chips (Moore's Law), which doubles every 18 months. More critically, storage hardware access rates (megabytes/second) are at a relative standstill, presenting a major bottleneck to the transfer of large data sets. The widening gap between the end user and the source of the data (in terms of processing capabilities and download time) has driven a major paradigm shift in the way large research data sets should be processed and accessed. This new paradigm envisages large data sets and computational resources being concentrated at a number of data centres. Through a new software infrastructure, end users will transparently interact with these distributed resources in a similar manner to the transparency

of data access for the existing World Wide Web. However, unlike the Web, data will not be migrated to end users but rather accessed, processed, and explored remotely across the network in a set of distributed data and computational service providers. The explosion in data volume is also driving the development of new software tools and mathematical algorithms that will operate in the distributed resource and service environment.

- How will astronomy benefit from the development of grid middleware and other new technology developments?
- Are astronomical requirements different from those of other disciplines (like High Energy Physics) that are currently leading the requirements definition and prototype phase of the Grid development?
- If data and computational services have to be provided from data centres, who will run them?
- How will they relate to existing astronomical data centres and archives?
- How connected and coherent in their operation do they need to be and how can

diverse global user communities interact with them in an efficient manner?

[2] *How will globally distributed teams of researchers be able to share data from across the electromagnetic spectrum originating from multiple new facilities, work collaboratively with common resources, and finally publish their discoveries and achievements in a manner that can be used by others?*

The multiple spectral windows we have into the origin and evolution of the Universe are provided by specialized telescopes operating on the ground and in space. Each type of observatory (radio, millimetre, infrared, optical, X-ray, gamma ray) has unique detector technologies, data types, data formats, and data volume challenges. These observatories are also operated by different organizations and governments spread across the globe, with different policies and available resources for the capture, processing, and distribution of data, and different tools to support the exploitation of the data. This wavelength diversity is both an opportunity and a challenge for global astronomy, but this is a characteristic that is not shared by the high energy particle physics community where the diversity of data types is relatively small. Deep insights into the fundamental physical processes responsible for the wealth of structures in our Universe are best obtained through a synthesis of information from across the electromagnetic spectrum. These insights are also augmented by theoretical simulations, which produce comparable or even larger data volumes. The challenge is to enable these diverse data sets to *interoperate*. This interoperability of multi-wavelength data must allow physical pictures to be formed in which researchers simultaneously have access to information from radio to X-ray in the same, or readily translatable, physical systems of coordinates and measured properties (e.g. flux, wavelength). Interoperability will be the result of agreed upon international standards for astronomical data across the electromagnetic spectrum.

- How will these new standards be defined and who will maintain them?
- Once the standards are agreed and different data providers adopt them, how will distributed teams of scientists get to know about this?
- How will they be able to discover new data resources that exist in a distributed Grid-like environment?
- Once researchers use data interoperability and produce new physical insights and data

products, how will they be able to share and publish these large volumes of new information?

- Should data be published with the same degree of rigour as scientific papers, including peer review? How will this be achieved technologically?

[3] *How will new observatories with facility-class instruments provide the maximal scientific return on the investment of global public funds that was necessary to create them?*

The set of planned international astronomical facilities represents an amount of public funding that greatly exceeds the traditional commitment of universities, research organizations, and governments to astronomical research. The costs associated with developing and operating ELTs, large radio observatories, and observatories in unique environments (e.g., Antarctica) will be individually of order 1 billion Euros. It is clearly vital that the scientific return on this globally significant investment be maximized. A maximal return will be ensured if the number of scientists who can access data from the facility is maximized. A maximal return therefore hinges on two elements: optimal access and a large user base. Optimal access will require the adoption of a Grid-like approach to the provision of the data and the services that operate on the data. It will also require that the network of data centres and service providers exists, and that they are funded in the long term. Data centres are therefore as important as the facilities themselves. Funding is needed for the physical storage and compute fabric, for expert staff effort in data curation and user support, and in constructing data services.

- How can we ensure that data access (storage, management, and services) is a fundamental part of the funding for these new facilities?

The number of users can be increased in several ways. The most obvious is that data should be captured, stored, and processed in such a way that it can be used by scientists other than those who originally obtained the data. This reuse of astronomical data for new purposes is already being seen as a growing trend at some modern data centres. The process of producing this data heritage and curating it will modify the way observatories are operated and the way instruments are built, and will imply a long term support role for data centres.

- Who will pay for this long term support role?

- What kinds of choices will astronomers have to make in designing new observatories and instruments to ensure data heritage?

Additionally, the number of users will be expanded by providing access free from constraints that originate from institutional policies and national boundaries. While respecting the proprietary rights of investigative teams, data providers must enable open public access to the global community of astronomers if their resources are to be maximally utilized.

- How will such policies be defined and adopted?

Unfortunately, there are technological barriers to the research environment that open data access requires. Many communities of astronomers operate in national IT infrastructures that are not capable of supporting even the Grid-like portals that are the entry points into the network of astronomical data and service providers. This access is as much a research facility as the resources it allows access to.

- How will national planning bodies recognize, measure, and address this need for access?

A Path Forward

Two to three years ago several independent groups of astronomers from around the world began to grapple with some of the difficult questions proposed in the previous section. These groups have now coalesced, and what is now emerging is a collective vision of the path forward. This vision centres on creating a new global astronomical research infrastructure called the *Virtual Observatory*.

The power of the World Wide Web is its *transparency*; it is as if all the documents in the world are inside your PC. The idea of the *Virtual Observatory* (VO) is to achieve the same transparency for astronomical data. In the VO all the world's data is available from your desktop. All archives understand the same query language, can be accessed through a uniform interface, and diverse data can be analysed by the same tools. A central goal is democratisation: the power the scientist has at her fingertips should be *independent of location*. Such an infrastructure will also enable *collaboratories*: informal distributed research teams sharing data, workflows, and analysis results in a transparent virtual storage system.

Transparency is also a goal of *computational Grids*, where a set of distributed computers functions like one supercomputer on your desktop. The VO concept can be seen as a domain-specific example of a *data grid*. However it goes one step further, as what is offered is not just access to the data, but also operations on the data and returned results that are essential for their full exploitation. Today such analysis is done by end-users after downloading data. In the future, the normal mode will be for such calculations (many of which are quite standard) to be *data services* offered by the expert data centres holding the data. These operations then also need to be standardised to be compatible across many archives. The result is a *service grid*. The VO will not be a monolithic system, but, like the Web, will be a set of standards that make all the components of the system *interoperable*: data and metadata standards, agreed protocols and methods, and standardised mix-and-match software components. These standards and software modules constitute the *VO Framework*. To achieve the whole vision, however, data centres, software developers, and facility builders all need to accept the new framework and work within it. Five strands of work are needed:

1. Development of standards and protocols, and their international agreement.
2. Construction of "glue" software components: portal, registry, workflow, user authentication, virtual storage.
3. Uptake by data centres, who need to "publish" to the system, i.e., to write VO compliant data services connected to their holdings.
4. Construction of tools to do science with the data.
5. Establishment and maintenance of resource registries and user support systems.

The VO concept has a high priority in most national astronomy programmes. Many other nations have recognised the issue, but have not yet been able to afford new projects. Large organizations such as ESA, NASA, ESO, and NSF have all recognised the strategic importance of the VO. The community itself, as well as its political leaders, has made its interest clear over the last two years through a number of dedicated VO conferences and workshops and special sessions at large general meetings, for example at the last General Assembly of the International Astronomical Union (Sydney, July 2003). Some of this drive comes from the widespread interest in grid middleware and e-science more generally, but mostly it comes from awareness

of the imminent data flood, and the constantly rising expectations of astronomers concerning the quality and power of web-based tools. The general feeling of most astronomers is that something like this simply has to happen: they are very keen that it happen in an organised and professional fashion.

Making it Happen

In June 2002, at an international VO meeting in Munich (co-sponsored by ESO, ESA, NASA, and NSF), the International Virtual Observatory Alliance (IVOA) was formed (<http://www.ivoa.net>). The alliance of 14 member projects (ESO/ESA, US, UK, Canada, China, Russia, Korea, Hungary, France, Germany, Italy, Australia, Japan, India) seeks to ensure that the essential VO infrastructural technologies and interoperability standards are developed to enable a VO capability on a global scale. IVOA working groups, modelled on the W3C process, and individual projects are actively developing, prototyping, and demonstrating these new capabilities. IVOA is also partaking in the Global Grid Forum, making the VO effort a research group of the GGF to ensure astronomical requirements are discussed and that the VO provides feedback to the GGF community on prototype Grid middleware. The International Astronomical Union is also playing a major role in ensuring the success and viability of the VO vision. A working group has been created within IAU Commission 5 to examine and endorse proposed IVOA standards in much the same way FITS has been endorsed by the IAU as an international astronomical data format over the past 30 years. The IAU is also seeking to enable the maximal return from new facilities. At the most recent General Assembly the IAU endorsed the following resolution¹:

1. *That data obtained at major astronomical facilities should, after a reasonable proprietary period in which they are available only to observers or other designated users of the facility, be placed in an archive where they may be accessed via the internet by all research astronomers. As far as possible, the data should be accompanied by appropriate metadata and other information or tools to make them scientifically valuable.*
2. *That such data should not be subject to intellectual property rights. The form in which data are made available, and the subsequent processing of such data, may be*

appropriately protected by copyright laws, but the fair usage (including educational purposes) of the archive data themselves should not be subject to restrictions.

3. *That funding agencies provide encouragement and support to enable data produced by astronomical research that they fund to be deposited, after some proprietary period as defined above, in recognized data archives which provide unrestricted access to these data.*

The global adoption of this resolution and its support by funding agencies, government bodies, and astronomers is critical to the realization of the VO vision and thereby, the maximal scientific utilization of new astronomical facilities. In the astronomical research environment of the 21st century, the endorsement and financial support of long-term data and data service access cannot be separated from the support of new scientific capabilities.

The work of the IVOA will continue actively until the essential elements of the new astronomical research infrastructure are agreed, developed, and tested. At that time, the community of data centres and data providers will have the enabling technology for them to become VO publishers of data and services. The VO will be this collection of collaborating, interconnected, uniformly accessible data and service providers. Research teams and individual astronomers will be empowered to design tools and run projects within the VO and to return VO-enable scientific results and data products to the broader community. The IVOA's long-term role will be to preserve and guide the development of VO standards and to act as a forum to promote the empowerment of global astronomy through appropriate standards and technologies.

Propositions

- Encourage universal adoption of the above IAU resolution.
- Encourage all data centres and facilities worldwide to follow IVOA standards and recommended IVOA practice.
- Strive to make quality of access to data as independent of location as possible.
- Agencies should invest in the data services infrastructure.
- Access to data services is as important as access to data.
- Every facility or instrument proposal should have a clear plan not just for a data pipeline, but also for data storage, management, and serving of data products.

¹<http://www.iau.org/IAU/Activities/publications/bulletin/pdf/IB94-WEB-12Feb.pdf>, p.35

