

Chapter 1

The World Wide Telescope

Alexander S. Szalay and Jim Gray

Astronomy is a wonderful Grid application because datasets are inherently distributed and yet form a fairly uniform corpus. In particular:

- (1) The astronomy community has a fairly unified taxonomy, vocabulary, and codified definition of metrics and units [24].
- (2) Modern data is carefully peer reviewed and collected with rigorous statistical and scientific standards.
- (3) Data provenance is tracked, and derived data sets are curated fairly carefully.
- (4) Most data is publicly available and will remain available for the foreseeable future.
- (5) Even though old data is much less precise than current data, old data is essential when studying time-varying phenomena.

Each astronomy archive covers part of the electromagnetic spectrum for a period of time and a subset of the celestial sphere. All the archives are from the same sky and the same celestial objects, although different observations are made at different times.

Increasingly, astronomers perform multispectral studies or temporal studies combining data related to the same objects from multiple instruments and archives. Cross-comparison is possible because data are well documented and schematized with a common reference frame, and have clear provenance.

The scale of the data—terabytes now, petabytes soon—means that most data must reside at archives managed by the teams that are gathering and publishing the data. An astronomer wanting to study a particular kind of object or phenomenon cannot download a full copy of each archive for local processing—both because the scientist does not have a spare local petabyte and because it would take too long to do the download. Rather, the scientist must request small (gigabyte-sized now, terabyte-sized in the future) subsets from each archive that represent the few million objects of interest out of the billions at the archives.

1.1 The Virtual Observatory

The Virtual Observatory—sometimes also called the World Wide Telescope—is under

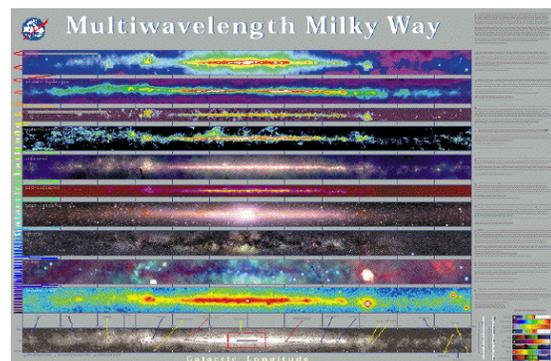


Figure 1: The NASA picture of the Milky Way in 11 spectral bands shows how vastly different the universe looks in different spectral bands. Each band is observed by a different instrument. Together they give a much more complete picture. In addition, there are phenomena at all time scales ranging from milliseconds to billions of years.

construction in many countries [1, 2, 5, 7, 10] It seeks to provide portals, protocols, and standards that unify the world's astronomy archives into a giant database containing all astronomy literature, images, raw data, derived datasets, and simulation data—integrated as a single intelligent telescope [25].

1.1.1 Living in an Exponential World

Astronomical data is growing at an exponential rate, doubling approximately every year as Moore's law improvements in semiconductors provide better computers and detectors. Once a new detector is deployed, data keeps accumulating at a constant rate. The exponential growth in data volumes arises from the continuous construction of new facilities with ever better detectors. New instruments emerge ever more frequently, so the growth of data is a little faster than the Moore's law prediction. Therefore, while every instrument produces a steady data stream, there is an ever more complex worldwide network of facilities with large output data sets.

How can we cope with this exponentially growing data avalanche? The data pipeline processing that analyses the raw detector data and the data storage is linearly proportional to the amount of data. The same technology that creates better detectors also creates the computers to process the data and the disks to save the data. For any individual project, the pipeline-processing task gets easier over time: the rate at which it produces data stays constant, while the cost of the computers required to analyze the data decreases by Moore's law. The first year is the most expensive for pipeline processing. Later the pipeline becomes increasingly trivial as the hardware performance improves with Moore's law and as the software performance bugs are fixed. The data storage costs peak in year two, when the storage demand doubles. Thereafter, the storage demand grows at less than 33% per year, while unit storage costs continue to drop. Thus, the community's total processing, networking, and storage costs are likely to remain stable over time, despite exponential growth in data volumes.

In contrast, the astronomy community's software costs seem to be exploding. Software used in astronomy today has its roots in Fortran, with C, C++, and Java emerging. Components are rarely reused among projects: projects tend to write their own software and use few common libraries. Thus, software costs are claiming a growing share of project budgets. Software costs are typically 25% to 50% of the total project cost. For example, the software investment of the Sloan Digital Sky Survey [9] was about 30%. Much of the work was invested in building the processing pipeline, special data access methods, and Web services. We estimate that more than half of this work is generic. The use of tools such as Condor [22] (Chapter LIVNY), Globus Toolkit™ [15] (Chapter CONCEPTS), Open Grid Services Architecture [16] (Chapter OGSA), virtual data systems such as Chimera [17], SQL databases, and development environments like .Net and Websphere would have made the task much simpler. One challenge the Virtual Observatory faces is to build reusable or prototypical subsystems that subsequent surveys can adapt to their needs.

1.1.2 Making Discoveries

The strongest motivation for building new sky surveys is to make new discoveries. It is important, therefore, to consider when and where new discoveries are made. We believe

that new discoveries are almost always made at the edges or frontiers: Either we need to look much deeper and detect fainter objects, or we have to go to extreme colors, by selecting the edges of a color distribution. We can search for objects of extreme shape (gravitationally lensed arcs) or time-domain behavior (supernovae, micro-lensing).

When the Internet was in its infancy, Bob Metcalfe postulated Metcalfe's law: *The utility of a computer network is proportional to the square of the number of nodes*. It is the number of the different connections one can make that matters. A variant of this law seems to apply here: *The utility of N independent datasets is approximately N^2 in addition to the independent information content of each of the datasets in isolation*. It is the number of connections we can make between fundamental properties that enable us to make new discoveries. A new observation of the sky in a previously unobserved wavelength or a new epoch for time-domain astronomy enables new connections to be made. The utility of a collection of independent observations is proportional to the number of nontrivial connections among them. This non-linear payoff is the motivation behind building multiwavelength sky surveys. By federating datasets from multiple, independent projects, we can make new connections. The early successes of today's sky surveys, Sloan Digital Sky Survey (SDSS) and the Two Micron All Sky Survey (2MASS), prove this point. The number of discoveries made after the first few hundred square degrees of observations (high redshift quasars, brown dwarfs) was far out of proportion to the area of sky. The magnitude of the new results can be explained only when we include the possible number of pairwise comparisons between filters.

1.1.3 Publishing Scientific Data

It is generally believed that scientific data publishing is well understood. There are the *authors*, mostly individuals or small groups, who create the experiments that provide data. Traditionally, authors have written papers that contain the data and explained it. There are the *publishers*, the scientific journals, which print the papers, and nowadays also make them available in an on-line version. There are the *curators*, whose role is filled today by libraries, which organize and store the journals and make them available for consumers. *Consumers* are scientists who want to use and cite the data in their own research.

This model worked well when all the scientific data relevant to the research could easily be included in the publication. The model breaks down, however, with the emergence of large datasets. This breakdown is not unique to astronomy. Particle physics has even larger quantities of data, and a similarly complex picture is emerging in genomic and biology research and in many other disciplines [21].

The author, publisher, and curator roles are clearly present in data-intensive science, but they are performed in different ways. The role of author belongs to collaborations, such as the Sloan Digital Sky Survey, the Human Genome Project, and the Large Hadron Collider at CERN. It takes five to ten years to build the experiment, before the *author* starts producing data. The data volume is so large that it will never be contained in journals—at most small summaries or graphs will be printed. The data is *published* to the collaborations (and the world) through Web-based archives. During the project lifetime, *curation* responsibility rests with the projects themselves. When the collaboration dissolves, the published data is either discarded or moved to a national archive facility for

long-term curation. The *consumers* have to deal with the data from these many sources, often obtaining it from publishers that are not eager to support them. The economic model for long-term curation is difficult because the costs fall to one group and the benefits to others.

1.1.4 Changing Roles

The exponential growth in both the number of data sources and individual data set sizes puts a particular burden on the projects that generate the data: They have the additional roles of data publisher and data curator. It makes sense to spend six years to build an instrument only if one is ready to use the instrument for at least the same amount of time. This means that during the data-production phase of a six-year project, the data grows at a linear rate. Hence, the mean time the data spends in the project archive before moving to the centralized facility is about three years. Turning this around, the national facilities will have all the data that is more than three years old. As the amount of data is doubling every year, in three years the data grows by eightfold. Thus, the archives have only 12% of the total data and less than 25% of the public data (data is typically made public after a year). The vast majority of the data and almost all the “current” data will be decentralized among the data sources —the new *publishers*. This is a direct consequence of the patterns of data-intensive science. These numbers were taken from astronomy; the rates may be different for other areas of science, but the main conclusions remain the same.

Thus, the projects are much more than just *authors*: They are also *publishers* and, to a large extent, *curators*. While scientists understand authorship well, they are less familiar with the responsibilities of the other two roles. These new roles are making many projects spend large software on the software to document, publish, and provide access to the data. Such tasks go far beyond the basic pipeline reductions. Since many projects are experimenting with these roles, much effort is duplicated and much development wasted. We need to identify the common design patterns in the publishing and curation process and to build reusable components and prototypes that others can adopt or adapt.

1.1.5 Metadata and Provenance

As more and more data access is through automated facilities, it is increasingly important to capture the details of how the data was derived and calibrated. This information must be represented in a form that is easy to parse. Even the meaning of data columns can be confusing. One common measure of flux of celestial objects, the so-called Johnson magnitude, has over 150 naming variants, which all connote the same essential concept but with some subtle differences. Unified content descriptors (UCDs) [24] were introduced to address this problem. UCDs are words in a compressed dictionary that was derived by automatically detecting the most commonly used terms in over 150,000 tables in the astronomical literature. Using a UCD designator can be helpful in finding common and comparable attributes in different archives, and serves as a unifying force in data publication.

Archived astronomy data is usually the end product of a complicated processing pipeline, within which the details of the processing (e.g., detection thresholds for objects) are carefully tuned by each project. Currently much of this information is captured in the form of published papers in the literature. There is a slowly emerging trend to describe

the processing pipelines in terms of directed acyclic graphs (DAGs: see Chapter LIVNY) and to create a proper workflow for the data reduction. Once DAGs are widely implemented, they will be the proper way to preserve the data provenance. Custom reprocessing of the data will be then quite easy: One will simply feed different parameters to the workflow. We expect this to be an important part of the Virtual Observatory–Grid interface.

During the loading of the objects into a database, a similar problem arises. One needs to track the heritage of each object, what version of the processing software has created it, and at what date. This requirement leads to yet another workflow system that is closely linked to that of the processing.

Most large astronomy datasets are generated by large collaborations. Typically, these collaborations have a good initial project design document, but as the projects progress, much information exchange is through e-mail exploders. E-mails get archived, but not in a formal sense. Thus, once projects go dormant, these e-mails are deleted. Since most technical decisions during the lifetime of the projects are contained only in the e-mails, these must be carefully archived and indexed; otherwise much of the metadata and provenance information is irretrievably lost.

1.2 Web Services: Using Distributed Data

These problems are not unique to science: Similar issues are emerging in the business world, where companies need to exchange information not only inside their corporate firewalls, but also with others. Exchanging and automatically reading data in various formats has haunted application developers for many years. Finally, a worldwide standard is emerging for data representation: the eXtensible Markup Language (XML). XML is rather complex and was not designed to be human readable. Nevertheless, there are clear grammatical rules for encapsulating complex information in a machine-readable form, and there are style sheets that render XML data to various easily understandable formats.

The most recent XML developments are related to Web services (Chapter OGSA): a standardized way to invoke remote resources on the Web and to exchange complex data. Web services define a distributed object model that lets us build Internet-scale software components and services. The Simple Object Access Protocol (SOAP) specifies how to invoke applications that can talk to one another and exchange complex data. The Web Service Description Language (WSDL) enables an application to find out the precise calling convention of a remote resource and to build a compatible interface. Toolkits, many freely available, link Web services to most modern programming languages and hardware platforms.

1.2.1 Web Services in the Virtual Observatory

Many of the expected tasks in the Virtual Observatory map well to Web services. Astronomers are already accustomed to various analysis packages, such as IRAF, IDL, or AIPS++, that have multilayer APIs [14, 26]. These packages start with a layer of simple image processing tasks and then build a layer of more complex processing steps on top of

that first layer. The packages assume that the data resides in FITS files in the local file system [27], and the processing is done on the workstation itself.

In the Virtual Observatory, most the data will be remote. As a result, data access to remote resources needs to be just as transparent as if it were local. The remote data volume may be huge; therefore, it makes sense to move as much of the data processing as near the data as possible, because in many cases after the first few steps of processing the output volume is dramatically smaller (e.g., extracting object catalogs). In many cases the data not only is remote but does not even exist at the time of the request: It may be extracted from a database with a query. One can carry this situation even further: The requested data may be created by a complex pipeline on the fly, according to the user's specification, like a recalibration and custom object detection run on an image built as a mosaic from its parts. The GriPhyN project [13] calls this concept "virtual data" [17, 18]—data that is created dynamically from its archived components.

1.2.2 Everything for Everybody?

We believe that a multilevel layering of services is the correct Virtual Observatory (VO) architecture. IRAF and AIPS++ are prototypes, but the concept needs to be extended to handle remote and virtual data sources. The core will be set of simple, low-level services that are easy to implement even by small projects. Indeed, we expect that there will be reference implementations of these services that can serve as prototypes for publishing new archives. Thus, the threshold to join the VO will be low. Large data providers will be able to implement more complex, high-speed services as well.

These core services can be combined into more complex portals that talk to several services and create more complex results. Such a design will have modular components, standard interfaces, and access to commercially built toolkits for the lowest-level communication tasks. We need to focus on the astronomy specific issues, and not reinvent the wheel.

We need to carefully define the VO framework and core services. Development resources are scarce and must be used efficiently. It would be a mistake to try to build everything for everybody; it is impossible to make everybody happy. How, then, to choose? A well-known design rule states that 20% of the effort covers 80% of the needs. If the Virtual Observatory builds the right 20% and satisfies 80% of the astrophysics community's needs, it will be counted a success. There is no need to build all possible complex applications—our colleagues have been resourceful in building quite complex applications out of IRAF modules. The VO need only provide clear standards, interfaces, documentation, and reference implementations in order to have most data providers adopt a set of core services. The VO must not lose sight of the ultimate goal: to create a powerful system that is simple to use.

1.3 Hierarchical Architecture

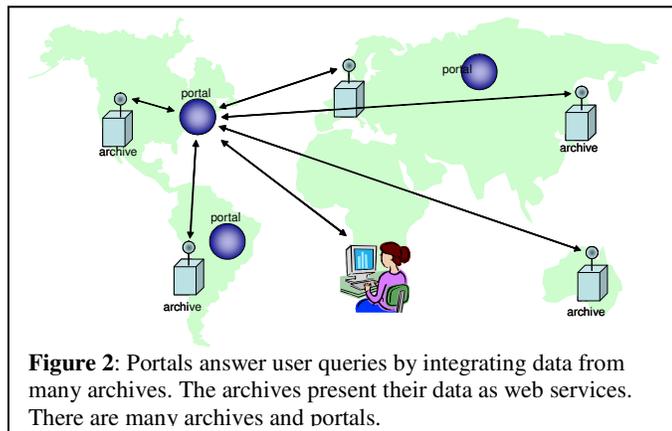
Several independent efforts are under way to build these archives, but all seem to be converging on a common architecture and common core technologies. The emerging common themes are as follows:

Archives: The archives each store text, image, and raw data in blobs or files, and store their schematized data in relational databases. They provide data mining tools that allow easy search and subsetting of the data objects at each archive. They also contain metadata about their contents, describing not only their physical units but also the provenance of the data.

Web Services: The archives provide Web service interfaces for on-demand queries and a file-transfer service for answers that involve substantial computation or data transfer. These Web services all support a common core schema that extends the VOTable data model [11]. The VOTable model specifies a standard coordinate system, standard representations for core astronomical concepts, and standard ways to represent both values and errors. All this is built atop SOAP and XML Schema Definitions (XSD). Some of these Web services are OGSA services to run long-running analysis jobs on the archive's data and to produce derived datasets – but most are interactive tasks that extract data on demand for portals and for interactive client tools.

Registries and Portals: Each archive declares its services with one or more registries. Registries record what kinds of information the archive provides—the sky coverage, temporal coverage, spectral coverage, and resolution. These registries will be widely replicated, and given the overlaps of astronomy with other disciplines, it seems likely that there will be many different kinds of registries, and that a particular service will be registered with many of them. Registries will be used by portals, which serve to answer user queries by integrating data from many archives. We expect that eventually many individuals will build their own custom portals, specializing in solving particular problems.

MAST, GLU, and AstroGrid are sample portals [1, 6, 8.] As one of the first experiments of the international Virtual Observatory effort, more than fifty different archives over the world provide a cone search interface for a Web search [3]. Each of these services, given a point in the sky and a radius (a cone), returns the archives' objects that fall within that cone. The response returns both the metadata (XML Schema) and the data as an XML document as a VOTable. The Cone Service request allows the user to optionally specify the *verbosity* level that controls how much detailed information is returned in the VOTable.



SkyQuery is another example of a portal [23] <http://SkyQuery.Net/>. It integrates five different Web services: SDSS, 2MASS, Faint Images of the Radio Sky at Twenty-cm, and the Isaac Newton Telescope Wide Field Survey, along with an image Web service. These archives are located on two continents at several geographic locations. Additional archives are currently being integrated. The portal accepts queries specifying the desired

object properties. It then decides which archives have relevant data (by querying each of them) and calculates an optimal query plan to answer the question, doing a Bayesian spatial cross-correlation among the objects in each archive. The resulting answer set is delivered to the user in tabular form along with images of the objects. SkyQuery is itself a Web service and can be used as a component of some other portal that renders the information in better ways, or combines these answers with other datasets.

SkyQuery was built using SQL and the .NET tools by four part-time people in less than two months. The use of XML, SOAP, and web services made it possible to easily plug together web services that were running at three different sites. Subsequently, the Isaac Newton Telescope in Cambridge was able to join the federation with a few days' work. The service has operated without problems for a year since then. This experience has made us very enthusiastic about the advantages of web services and good tools to build them.

SkyQuery and its component services are both I/O and CPU intensive. They lend themselves to database and compute farms enabled by web and Grid services. Indeed, some queries cannot be answered interactively; they require reading huge amounts of data or doing extensive calculations. These complex SkyQuery jobs are excellent applications for facilities for submitting, executing, and managing batch computations, and other facilities for bulk data transfer of results.

1.4 The Virtual Observatory and the Grid

Many potential astronomy applications require access at the granularity of objects rather than entire files. Thus, astronomy data on the Grid will generally reside in read-intensive databases that will be accessed by associative query interfaces that subset the data. Comparing multiple observations of the same object taken by different instruments at different times requires accessing and comparing individual records in several different archives. These processes imply a massive use of spatial and other indices, which in turn means a heavy use of databases. The huge data volumes require a parallel search capability with direct random access of certain objects.

Since the same sky is covered by many different surveys, at many different wavelengths, and at many different times, we also need to search and correlate the metadata. Archives and datasets are constructed with great effort and are peer reviewed, so there will be less than a thousand of them. But each archive has fairly extensive metadata. One or more registry services are needed to gather this data together and to provide queries against the metadata. Again, these tasks are best suited to databases.

These are not the only Grid issues that are important to astronomy. Access control must be addressed, but is less important for astronomical data sets than in other sciences. The data has no commercial value, and astronomers are fairly good about sharing—much of the world's astronomy archives are, in fact, public. Resource management, on the other hand, is likely to be important: if remote users can pose complex queries (which presumably will be desirable), then mechanisms will be required to determine who gets to use how much of scarce bandwidth and computing resources.

1.4.1 Data, Networking, and Computation Economics

With the falling costs of computer hardware, projects are able to provide limited public access to their archives at a modest cost. But, where do they draw the line? How much pro bono processing and search should an archive provide to its data?

A simple calculus can be used to decide the least expensive way to structure a distributed archive. First, with progress in disk technology, all data can be kept online, rather than putting it in near-line tape archives. The data and derived products were collected at great expense, so the data should be safely stored at two or more locations. When a user has a new query or computation, several choices exist.

If the query is small, it can just be sent to one of the archive servers. Today “small” means less than ten hours of computation (\$1 of computing), less than a terabyte of disk and LAN bandwidth, and less than 10 GB of network data movement (\$10 of network traffic).

When the query exceeds these limits, some planning is required. With current economics, there is rough parity between (1) one database access, (2) ten bytes of network traffic, (3) 100,000 instructions, (4) 10 bytes of disk storage, and (5) a terabyte of disk or LAN bandwidth. Most astronomy calculations are data intensive. It is rare to find computations of more than 10,000 instructions per byte. This argues strongly for doing the computation near the data. That in turn means that the publishers should provision a Beowulf cluster near the data for large jobs (ones costing more than \$100) and develop some form of chargeback for such services. If that is infeasible, then one alternative is to get a copy of the database either over the Internet or via sneakernet [19]. Paradoxically it can be faster and cheaper to send a terabyte via a storage brick than via the Internet.

1.4.2 Compute-Intensive Tasks

The transformation of raw instrument data into calibrated and cataloged data is a demanding computational task. For example, about 25% of the Sloan Digital Sky Survey budget went into building the software pipelines that process the photographic and spectroscopic data and produce catalogs of stars, galaxies, quasi-stellar objects, spectrograms, spectral lines, and redshift estimates. This work produced over a million lines of code that continuously transform the raw data into the SDSS catalogs. This software is constantly being refined and improved, so all the old data needs to be reprocessed about once a year to produce a new dataset that includes all the data processed in the new and better way. These algorithms consume between 1,000 and 10,000 instructions per byte of input data. Hence, reprocessing the data consumes about 10^{17} instructions—100 exa-instructions operating on about 15 TB of source data. With current processors, that is about ten CPU-years. But we need to do the computation several times per year. The solution is to use computer clusters to process the data in parallel. Indeed, the current Beowulf cluster at Fermilab can reprocess the entire SDSS dataset in about six weeks, and the resulting data can be loaded into databases in a few days.

This pipeline processing is a natural candidate for Grid services and particularly the virtual data toolkit. The SDSS had to build its own batch job scheduler, its own data

lineage tracker, and its own operations interfaces. The European Southern Observatory (ESO) built an even more elaborate and functional system [4]. Today, both these efforts could exploit systems such as the Chimera virtual data toolkit [17] (which in turn builds on the Globus and Condor toolkits) to save a great deal of effort.

1.4.3 Data Mining and Statistics of Terabytes

While much of the pipeline processing of the data into catalogs stored in archives is linear, and has the cost of a few thousand instructions per byte, there are also complex statistical analysis that needs to be performed on the archives subsequently.

In January 2003 both the U.S. and the European community demonstrated several functional research scenarios using the Virtual Observatory. These scenarios covered a wide spectrum: (1) dynamically cross-correlating the SDSS and 2MASS catalogs; (2) scanning hundreds of thousands of files in order to find clusters of galaxies, and testing grid tools in the process [12]; and (3) multiple interfaces rendering the same complex VOTable document created in response to a transient event (gamma-ray burst). These computations demonstrate the enormous diversity of problems the Virtual Observatory needs to address. The SDSS-2MASS cross-correlation found several previously undiscovered rare objects in the process

One major impediment to such spatial and multispectral cross-correlations is that our current correlation algorithms involve the computations of pairwise distances. Brute-force algorithms would have quadratic computational and I/O complexity (compare all pairs of objects). Hypothesis testing over a large feature vector has typically cubic computational and I/O complexity, because of the scaling of matrix inversion. Typical matrix sizes today are in the $10,000^2$ to the $1,000,000^2$ range. Such algorithms are workable for small datasets, but even $M\log N$ algorithms are infeasible for datasets involving billions of objects. We have to move from these exact solutions to approximate and heuristic algorithms that have linear and sublinear computational and I/O costs. Even then, the computations will have to be performed over large numbers of CPUs—for a billion objects even $\log N$ is about 30. The creation of such algorithms stands as a major challenge to the data-mining and statistical community.

1.5 Outreach Using the Virtual Observatory

The World Wide Telescope is also an excellent laboratory to teach Computational Science. It presents a well-documented spatial, temporal, and image database that is geographically distributed. Students can rediscover classical astronomy results using real data and real grid analysis tools, or they can explore new hypotheses and perhaps make new discoveries. The SkyServer project site shows an example of this [20]. Over the past two years the student project part of the site has received more than a million Web hits. We estimate that totals more than 10,000 hours of student instruction in astronomy and computational science.

The Virtual Sky [28] provides image navigation over the whole sky, covering many different surveys with the ease of MapQuest. Through its intuitive interface, one can easily compare images from the different surveys with a large dynamic range of magnifications.

1.6 Summary

The World Wide Telescope is an excellent Grid application. It involves the federation of many distributed terascale data sources. The astronomy community is interested in mining the individual datasets looking for patterns, but there is even greater interest in cross-correlating the datasets to find new phenomena. The construction of each archive involves massive calculations that ingest, analyze, and categorize the instrument data, producing databases and files. Cross-correlation introduces enormous requirements for data movement and computation.

The emergence of these huge datasets changes the roles of authors, publishers, and curators. Science projects become data authors and publishers, and must also do the curatorial work that will allow archives to accept and preserve the data once the project is complete.

Each project will likely publish a database that is externalized as a set of Web services. The astronomy community is in the midst of defining and prototyping these core services. Portals will federate these Web services to provide uniform access to the archives as though they are a single distributed database. Cross-identifications of objects across multiple archives, and searches involving both spatial and time-domain constraints, are typical tasks for the Virtual Observatory. Because of the large sizes of the datasets, their distributed nature, and their rapidly varying content, these tasks need an architecture that is inherently distributed and scalable—the Grid. We view the Virtual Observatory as a small part of the larger effort of defining and applying an Open Grid Services Architecture and creating an international Grid infrastructure.

References

1. AstroGrid Web Site.
2. Astrophysical Virtual Observatory.
3. Cone Service Definition Web Site.
4. European Southern Observatory Web Site.
5. International Virtual Observatory Alliance Web Site.
6. NASA Multimission Archive at Space Telescope Web Site.
7. The National Virtual Observatory Web Site.
8. Simbad Web Site.
9. Sloan Digital Sky Survey.
10. VizieR Web Site.
11. Votable Web Site.
12. Annis, J., Zhao, Y., Voeckler, J., Wilde, M., Kent, S. and Foster, I., Applying Chimera Virtual Data Concepts to Cluster Finding in the Sloan Sky Survey. in *SC'2002*, (2002).
13. Avery, P. and Foster, I. The GriPhyN Project: Towards Petascale Virtual Data Grids, 2001.
14. Croes, G.A. Astronomical Data Analysis Software and Systems II. in Hanisch, R.J., R.Brissenden and Barnes, J. eds. *ASP Conference Series*, ASP, San Francisco, 1993, 156.

15. Foster, I. and Kesselman, C. Globus: A Metacomputing Infrastructure Toolkit. *International Journal of Supercomputer Applications*, 11 (2). 115-129.
16. Foster, I., Kesselman, C., Nick, J. and Tuecke, S. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration, Globus Project, 2002.
17. Foster, I., Voeckler, J., Wilde, M. and Zhao, Y., Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation. in *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, (2002).
18. Foster, I., Voeckler, J., Wilde, M. and Zhao, Y., The Virtual Data Grid: A New Model and Architecture for Data-Intensive Collaboration. in *Conference on Innovative Data Systems Research*, (2003).
19. Gray, J., Chong, W., Barclay, T., Szalay, A.S. and Vandenberg, J. TeraScale SneakerNet: Using Inexpensive Disks for Backup, Archiving, and Data Exchange, Microsoft Technical Report, MSR TR 2002-54, Redmond WA, (2002).
20. Gray, J., Szalay, A.S., Thakar, A.R., Kunszt, P.Z., Malik, T., Raddick, J., Stoughton, C. and vandenBerg, J. The SDSS SkyServer - Public Access to the Sloan Digital Sky Server Data, 2001, 1-11.
21. Gray, J., Szalay, A.S., Thakar, A.R., Stoughton, C. and Vandenberg, J., Online Scientific Data Curation, Publication, and Archiving. in *SPIE Astronomy Telescopes and Instruments*, (Waikoloa, Hawaii, 2002), 103-107.
22. Litzkow, M., Livny, M. and Mutka, M., Condor - A Hunter of Idle Workstations. in *Proceedings of the 8th International Conference of Distributed Computing Systems*, (1988), 104-111.
23. Malik, T., Szalay, A.S., Budavari, T. and Thakar, A., SkyQuery: a Web-Service Approach to Federate Databases. in *Proc. CIDR*, (2003).
24. Mehringer, D.M., Plante, R.L. and Roberts, D.A., ADASS VIII. in *ASP Conf. Ser.*, (1998).
25. Szalay, A. and Gray, J. The World-Wide Telescope. *Science*, 293. 2037-2040.
26. Tody, D. *A Reference Manual for the IRAF Subset Preprocessor Language*, 1983.
27. Wells, D.C., Greisen, E. and Harten, R.H. FITS: A Flexible Image Transport System. *Astron. Astrophys. Suppl.*, 44. 363-370.
28. Williams, R. Virtual Sky Web Site.