The World-Wide Telescope[1]

Alexander Szalay, The Johns Hopkins University

Jim Gray, Microsoft

August 2001

Technical Report

MSR-TR-2001-77

The World-Wide Telescope

Alexander Szalay, The Johns Hopkins University
Jim Gray, Microsoft
August 2001

Abstract *All astronomy data and literature will soon be online and accessible via the Internet. The community is building the* Virtual Observatory*, an organization of this worldwide data into a coherent whole that can be accessed by anyone, in any form, from anywhere. The resulting system will dramatically improve our ability to do multi-spectral and temporal studies that integrate data from multiple instruments. The virtual observatory data also provides a wonderful base for teaching astronomy, scientific discovery, and computational science.*
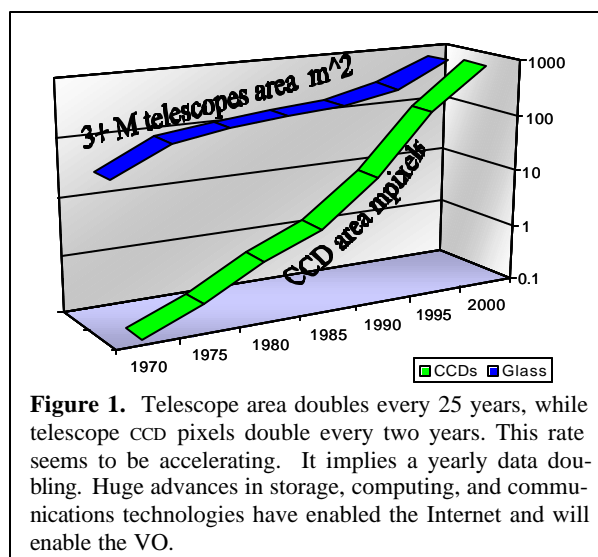
Many fields are now coping with a rapidly mounting problem: how to organize, use, and make sense of the enormous amounts of data generated by today's instruments and experiments. The data should be accessible to scientists and educators so that the gap between cutting edge research and education and public knowledge is minimized and presented in a form that will facilitate integrative research. This problem is becoming particularly acute in many fields, notably genomics, neuroscience, and astrophysics. In turn, the availability of the internet is allowing new ideas and concepts for data sharing and use. Here we describe a plan to develop an internet data resource in astronomy to help address this problem in which, because of the nature of the data and analyses required of them, the data remain widely distributed rather than coalesced in one or a few databases (e.g., Genbank). This approach may have applicability in many other fields. The goal is to make the Internet act as the world's best telescope—a World-Wide Telescope.

The problem

Today, there are many impressive archives painstakingly constructed from observations associated with an instrument. The Hubble Space Telescope [HST], the Chandra X-Ray Observatory [Chandra], the Sloan Digital Sky Survey [SDSS], the Two Micron All Sky Survey [2MASS], and the Digitized Palomar All Sky Survey [DPOSS] are examples of this. Each of these archives is interesting in itself, but temporal and multi-spectral studies require combining data from multiple instruments. Furthermore, yearly advances in electronics enable new instruments that double the data we collect each year (see Figure 1). For example, about a gigapixel is deployed on all telescopes today, and new gigapixel instruments are under construction. A night's observation is a few hundred gigabytes. The processed data for a single spectral band over the whole sky is a few terabytes. It is not even possible for each astronomer to have a private copy of all their data. Many of these new instruments are being used for systematic surveys of our galaxy and of the distant universe. Together they will give us an unprecedented catalog to study the evolving universe—provided that the data can be systematically studied in an integrated fashion.

Already online archives contain raw and derived astronomical observations of billions of objects – both temporal and multi-spectral surveys. Together, they have an order of magnitude more data than any single instrument. In addition, all the astronomy literature is online, and is cross-indexed with the observations [Simbad, NED].

Figure 1. Telescope area doubles every 25 years, while telescope CCD pixels double every two years. This rate seems to be accelerating. It implies a yearly data doubling. Huge advances in storage, computing, and communications technologies have enabled the Internet and will enable the VO.

Why is it necessary to study the sky in such detail? Celestial objects radiate energy over an extremely wide range of wavelengths, from the radio, to infrared, optical, ultraviolet, x-rays and even gamma-rays. Each of these observations carries important information about the nature of the objects. The same physical object can appear to be totally different in different wavebands (See Figure 2). A young spiral galaxy appears as many concentrated 'blobs', the so-called HII regions in the ultraviolet, while it shows smooth spiral arms in the optical. A galaxy cluster can only be seen as an aggregation of galaxies in the optical, while x-ray observations show the hot and diffuse gas between the galaxies.

The physical processes inside these objects can only be understood by combining observations at several wave-

lengths. Today we already have large sky coverage in 10 spectral regions; soon we will have additional data in at least 5 more bands. These will reside in different archives, making their integration all the more complicated.
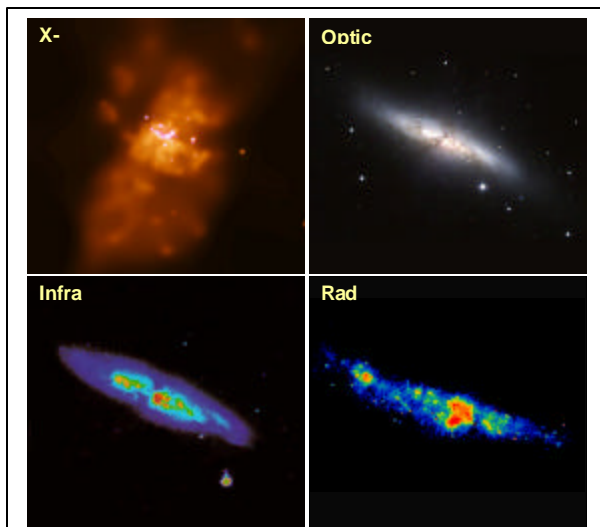

Figure 2: M82, at a distance of 11 million light years, is a rare nearby starburst galaxy where stars are forming and expiring at a rate ten times higher than in our galaxy. A close encounter with the large galaxy M81 in the last 100 million years is thought to be the cause of this activity. The images taken at different wavelengths unveil different aspects of the star forming activity inside the galaxy. Images courtesy of:
(X-ray) NASA/CXC/SAO/PSU/CMU, (Optical) AURA/NOAO/NSF, (Infrared) SAO, (Radio) MERLIN/VLA, compilation from
http://chandra.harvard.edu/photo/0094/index.html

Raw astronomy data is complex. It can be in the form of fluxes measured in finite size pixels on the sky, it can be spectra (flux as a function of wavelength), it can be individual photon events, or even phase information from the interference of radio waves.

In many other disciplines, once data is collected, it can be frozen and distributed to other locations. This is not the case for astronomy. Astronomy data needs to be calibrated for the transmission of the atmosphere, and for the response of the instruments. This requires an exquisite understanding of all the properties of the whole system, which sometimes takes several years. With each new understanding of how corrections should be made, the data are reprocessed and recalibrated. As a result, data in astronomy stays 'live' much longer than in other disciplines – it needs an active 'curation', mostly by the expert group that collected the data.

Consequently, astronomy data reside at many different geographical locations, and things are going to stay that way. There will not be a central Astronomy database. Each group has its own historical reasons to archive the

data one way or another. Any solution that tries to federate the astronomy data sets must start with the premise that this trend is not going to change substantially in the near future; there is no top-down way to simultaneously rebuild all data sources.

The World Wide Telescope

So to solve these problems, the astrophysical community is developing the World-Wide Telescope – often called the *Virtual Observatory* [VO]. In this approach, the data will primarily be accessed via digital archives that are widely distributed. The actual telescopes will either be dedicated to surveys that feed the archives, or telescopes will be scheduled to follow-up 'interesting' phenomena found in the archives. Astronomers will look for patterns in the data, both spectral and temporal, known and unknown, and use these to study various object classes. They will have a variety of tools at their fingertips: a unified search engine, to collect and aggregate data from several large archives simultaneously, and a huge distributed computing resource, to perform the analyses close to the data, in order to avoid moving petabytes of data across the networks.

Other sciences have comparable efforts of putting all their data online and in the public domain – Genbank® in genomics is a good example – but so far these are centralized rather than federated systems.

The Virtual Observatory will give everyone access to data that span the entire spectrum, the entire sky, all historical observations, and all the literature. For publications, data will reside at a few sites maintained by the publishers. These archive sites will support simple searches. More complex analyses will be done with imported data extracts at the user's facility.

And time on the instrument will be available to all. The Virtual Observatory should thus make it easy to conduct such temporal and multi-spectral studies, by automating the discovery and the assembly of the necessary data.

The typical and the rare

One of the main uses of the VO will be to facilitate searches where statistics are critical. We need large samples of galaxies in order to understand the fine details of the expanding universe, and of galaxy formation. These statistical studies require multicolor imaging of millions of galaxies, and measurement of their distances. We need to perform statistical analyses as a function of their observed type, environment, and distance.

Other projects study rare objects, ones that do not fit typical patterns – the needles in the haystack. Again, having multi-spectral observations is an enormous help. Colors of objects reflect their temperature. At the same time, in the expanding Universe, the light emitted by distant objects is redshifted. Searching for extremely red

objects thus finds either extremely cold objects, or extremely distant ones. Data mining studies of extremely red objects discovered distant quasars, the latest at a redshift of 6.28 [QSO]. Mining the 2MASS and SDSS archives found many cold objects, *brown dwarfs,* bigger than a planet, yet smaller than a star. These are good examples of multi-wavelength searches, not possible with a single observation of the sky, done by hand today, automated in the future, we do not know even of data existed—discover on the fly.

The time dimension

Most celestial objects are essentially static; the characteristic timescale for variations in their light output is measured in millions or billions of years. There are time-varying phenomena on much shorter timescales as well. Variations are either transient, like supernovae, or regular, like variable stars. If a dark object in our galaxy passes in front of a star or galaxy, we can measure a sudden brightening of the background object, due to gravitational microlensing. Asteroids can be recognized by their rapid motion. All these variations can happen on a few days' timescale. Stars of the Milky Way Galaxy are all moving in its gravitational field. Although few stars can be seen to move in the matter of days, comparing observations ten years apart measures such motions accurately.

Identifying and following object variability is time-consuming, and adds an additional dimension to the observations. Not only do we need to map the Universe at many different wavelengths, we need to do it often, so that we can find the temporal variations on many timescales. Once this ambitious gathering of possibly petabyte-size datasets is under way, we will need summaries of light curves, and also extremely rapid triggers. For example, in gamma-ray bursts, much of the action happens within seconds after the burst is detected. This puts stringent demands on data archive performance.

Agenda

The architecture must be designed with a 50-year horizon. Things will be different in 50 years – computers will be several orders of magnitude faster, cheaper, and smarter. So the architecture must not make short-term technology compromises. On the other hand, the system must work today on today's technology.

The Virtual Observatory will be a federation of astronomy archives, each with unique assets. Archives will typically be associated with the institutions that gathered the data and with the people who best understand the data. Some archives might contain data derived from others and some might contain synthetic data from simulations. A few archives might specialize in organizing the astrophysical literature and cross-indexing it with the

data, while others might just be indices of the data itself – analogous to Yahoo! for the text-based web.

Astronomers own the data they collect – but the field has a long tradition of making all data public after a year. This gives the astronomer time to analyze data and publish early results, and it also gives other astronomers timely access to the data. Given that data are doubling every year, and given that the data become public within a year, about half the word's astronomy data is available to all. A few astronomers have access to a private data stream from some instrument; so, we estimate everyone has 50% of the data and some people have 55% of the data.

Uniform views of diverse data

The social dynamics of the Virtual Observatory will always have a tension between coherence and creativity – between uniformity and autonomy. It is our hope that the Virtual Observatory will act as a catalyst to homogenize the data. It will constantly struggle with the diversity of the different collections, and the creativity of scientists who want to innovate and who discover new concepts and new ways of looking at things. These two forces need to be balanced.

Each individual archive will be an autonomous unit run by scientists. The challenge is to translate this heterogeneous mix of data sources into a uniform body of knowledge for the scientists and educators who want to use data from multiple sources. Each archive needs to easily present its data in compatible formats and the archives must be able to exchange data.

This uniform view will require agreement on terminology, on units, and on representations – a unified *conceptual schema* (data model) that describes all the data in all the archives in a common terminology. This schema will evolve with time, and there will always be things that are outside the schema, but VO users will see all the archives via this unifying schema, and data interchange will be done in terms of the schema.

We believe that the base representations will likely be done using the emerging standards for XML, Schemas, and SOAP, and web services [W3C], but beyond that there will have to be tools that automatically transform the diverse and heterogeneous archives to this common format. This is beyond the current state of computer science, yet solving this *schema integration problem* will be a key enabler for the Virtual Observatory.

Users will want query the virtual observatory using nice graphical tools, both to pose questions and to analyze and visualize the results. The users will range in skill from professional astronomers to bright grammar school students, so a variety of tools will be needed.

The Virtual Observatory query systems will have to find data among the various versions and replicas stored at the various archives. The user might ask for the correlation between radio, optical, and X-ray sources with certain properties. It will be up to the query system to locate the appropriate datasets, subset them and cross-correlate them, returning the requested attributes from each data source. As with schema integration, there is good technology for automatic indexing, location-transparent, and parallel data search. But the scenario just described is beyond the limits of what computer science researchers can do today.

At its core, the Virtual Observatory is a data manager. There is a well-established discipline of defining and managing data. There are good languages for defining data schemas (structures) and for mapping these data schemas into existing data management systems. But, most of these tools are designed for homogeneous environments with central control – for example managing the assets of an individual corporation, organization, or agency. Tools that integrate heterogeneous databases are still primitive and labor intensive.

The raw astronomy data from a telescope runs through a substantial 'software pipeline' that extracts objects (stars, galaxies, clouds, planets, asteroids) from the data and assigns attributes (luminosity, morphology, classification) to them. These pipelines are constantly improving as we learn new science and as we discover flaws in the pipeline software, so the derived data is constantly evolving into new versions.

Each archive's data will likely be replicated at one or more mirror sites so that a catastrophe at one site will not cause any long-term data loss. Data may also be wholly or partially replicated at other sites so to speed local computations: it is often faster and cheaper to access local data than to fetch it from a remote site.

We assume the Internet will evolve so that copying larger datasets will be feasible and economic (today, datasets in the terabyte range are typically moved by parcel post). Still, it will often be best to move the computation to the petabyte-scale datasets to minimize data movement and speed the computation.

Current data management systems propagate and manage data replicas, but they require administrators to decide when and where the replicas are stored. Administrators must design and manage the replication strategy and must track data versions and data lineage. The computer science challenge is to automate these tasks: to automatically configure and track replicas and to automatically track data lineage and dependencies on different derived data versions as the programs evolve.

Better algorithms

At the micro-level, we expect major advances in computer science algorithms. Hardware performance has improved about a hundred-fold every decade since 1970. There has been a comparable improvement in software algorithms – the simplest example being sorting: sort performance has doubled each year since 1985, half due to improvements in hardware, and half due to improvements in parallel sorting algorithms. We must continue to invest in and investigate new algorithms and data structures for data loading, data cleansing, data search, data organization, and data mining. Statistical methods lie at that heart of most of our data search algorithms, so advances in computational statistics will be essential to the success of the Virtual Observatory.

Current computer architectures are moving towards huge arrays of loosely-coupled computers with local storage. These *Beowulf* clusters harness commodity components and so provide very inexpensive computing. Cellular architectures like IBM's BlueGene project or the Japanese Earth Simulator project carry this to an extreme – millions of computers in one cluster. Computational Grids are clusters of computer clusters that can dynamically allocate resources to the tasks at hand [Grid]. But, not all our algorithms fit this computational structure – some require fine-grain parallelism and shared memory. So, we must either find new algorithms that work well on cluster computer architectures, or we must invest in massive-memory computer machines that can solve these problems.

Education

The Virtual Observatory offers the opportunity to teach science in a participatory way. We can give students direct access to a wonderful scientific instrument. They can use it to make discoveries on their own. Very interesting projects and lectures can be built using the Virtual Observatory tools and data.

Astronomy has a special attraction for all, as demonstrated by planetariums, amateur telescopes, and textbooks. Even very young children can be engaged in many different sciences via astronomy – astronomy has strong ties to physics, chemistry, and mathematics. Astronomy can be used as a vehicle for introducing the basic concepts of all these fields and also used to teach the process of scientific discovery. The two of us, with a lot of help from others, are in the process of creating such a pilot project, using data from the SDSS [SkyServer].

The Virtual Observatory can also be used to teach computational science. Traditionally, science has been either theoretical or empirical. In the last 50 years, computational science emerged as a third approach, first in doing simulations and now increasingly in mining scientific

data. Indeed, most scientific departments now have a strong computational program. The Virtual Observatory is a unique tool to teach these skills

A World-Wide Effort

Like astronomy, the Virtual Observatory is a world-wide effort. Initiatives are underway in many countries with a common goal: join the diverse worldwide astronomical databases into a single federated entity facilitating new research for the worldwide astronomy community. The European national and international astronomy data centers are leading an effort funded by. the European Union. The Astronomical Virtual Observatory (AVO) is led by the European Southern Observatory and also backed by the European Space Agency. The EU also sponsors research in pipeline processing technology to handle the anticipated terabyte data streams from future large survey telescopes. The UK funds the AstroGrid Project to investigate distributed data archives Grid technology. Japan and Australia are setting up their own large archives. In the United States, the National Science Foundation sponsors development of the information technology infrastructure necessary for a National Virtual Observatory (NVO). There is a close cooperation with the particle physics community, through the Grid Physics Network (GriPhyN). NASA supports astronomy mission archives and discipline data centers, while developing a roadmap for their federation.

Impressively, these projects are all cooperating, and are working toward a future Global Virtual Observatory to benefit the international astronomical community and the public alike. There are similar efforts under way in other areas of science as well. The VO has had and will have significant interactions with other science communities, both learning from some and providing a model for others.

References

[Chandra] *Chandra X-Ray Observatory Center*,
http://chandra.harvard.edu/
[DPOSS] *Digitized Palomar Observatory Sky Survey*,
http://www.astro.caltech.edu/~george/dposs/
[Grid] *The Grid: Blueprint for a New Computing Infrastructure*, I.
Foster, C. Kesselman eds., Morgan Kaufmann, 1998.
[HST] *The Hubble Space Telescope*, http://www.stsci.edu/
[NED] *NASA/IPAC Extragalactic Database*,
http://nedwww.ipac.caltech.edu/
[QSO] Fan,X. et al., http://xxx.lanl.gov/abs/astro-ph/0108063
[SDSS] *The Sloan Digital Sky Survey website*, http://www.sdss.org/
[Simbad] *SIMBAD Astronomical Database,* http://simbad.u-strasbg.fr/
[SkyServer] *Public access website to the SDSS data,*
http://skyserver.sdss.org/
[VIzieR] *VizieR Service*, http://vizier.u-strasbg.fr/viz-bin/VizieR
[VO] *Virtual Observatories of the Future*, ASP Conf. Series, V. 225,
R.J. Brunner, S.G. Djorgovski, A.S. Szalay eds., Sept 2000.
[W3C] The Semantic Web http://www.w3.org/2001/sw/ and Web
Services, http://www.w3.org/TR/wsdl
[2MASS] *The Two Micron All Sky Survey*,
http://www.ipac.caltech.edu/2mass/