

AstroGrid : the UK's Virtual Observatory

Andy Lawrence, AstroGrid Project Leader
University of Edinburgh

Paper for the UK e-Science programme All-Hands Meeting Nottingham, September 2003

Abstract

AstroGrid is the UK's contribution to the world-wide drive towards a Virtual Observatory (VO). The VO vision is not one of a software monolith, but rather one of a framework which enables data centres to provide competing and co-operating data services, and which enables software providers to offer a variety of compatible analysis and visualisation tools and user interfaces. AstroGrid is concentrating on the "engine room" work within this vision – the core VO infrastructure – but we are intending nonetheless to produce a working system of daily scientific use to astronomers. Good progress has been made on the necessary international standardisation, and in the last six months we have written real working software for most of the key infrastructure components. We are now setting up a pool of scientist beta testers, and expect to have something of real use by December 2003, and a full working system by December 2004.

(1) Introduction

This paper is in four parts. First I describe the AstroGrid project itself, and its relationship to other endeavours. Next, I summarise in turn the concepts behind the Virtual Observatory and the Grid, and how they relate to each other. I then examine the key technical issues and the progress we have made in solving them. Finally I summarise the overall current status of the project.

(2) The Astrogrid Project

(2.1) AstroGrid

AstroGrid is the UK's contribution to the world-wide drive towards the Virtual Observatory ideal. The AstroGrid consortium comprises (in alphabetical order) the Universities of Belfast, Cambridge, Edinburgh, Leicester, London (MSSL), and Manchester (Jodrell Bank), along with the Rutherford Appleton Lab (RAL) of CLRC. It is primarily a consortium of the main UK astronomical data centres. This is not a coincidence, as our take on the VO concept is one of enhanced and collaborative *services* offered by data centres. However the consortium also contains expertise in astronomical client-side software. The project is largely funded by PPARC, with a budget of 3.7M pounds, but also has support from the EC via the AVO collaboration (see section 3). Our remit is fairly wide, covering solar physics and solar-terrestrial (space plasma) physics as well as optical, radio, and x-ray astronomy. A Phase-A study was undertaken from September 2001 to December 2002, along with the first development of new standards with international partners. Our Phase-B construction phase began in Jan 2003, with a planned end in December 2004. We have already achieved working demonstrations and two open software releases including working grid services. We are a well integrated part of the UK e-science programme, in particular acting as an early adopter of OGSA-DAI software.

(2.2) The international context

There are three large funded VO programmes across the world – AstroGrid, the US–NVO programme, and the European AVO project (Astrophysical Virtual Observatory). AstroGrid has good working relations with US–NVO, and is actually a formal partner in the EC funded European AVO project. The work programmes are not identical but intersect. Formally AVO funds three staff positions which we see as part of the AstroGrid team, and an equivalent number of PPARC funded FTEs (full time equivalents) spread across a larger number of individuals are at the disposal of AVO. We have carefully aligned workpackages to make this feasible. The main distinction between AstroGrid and AVO is timing. AstroGrid is committed to early implementation, even if this requires limited functionality, whereas AVO is an RDprogramme aimed at a follow–on project to construct a full scale working facility. To AVO then, AstroGrid is both a technology development programme, and a very large pilot programme.

Other smaller VO projects have now sprung up all around the world, and we have formed an **International Virtual Observatory Alliance (IVOA)**, which is increasing in importance. It began as a senior management discussion forum, developing a joint roadmap, but soon evolved a set of technical issue working groups, and a series of workshop meetings aimed at producing agreed international standards, for data, metadata, and software module inter–operability. We now have an agreed process for developing standards based on the IETF model.

(2.4) Goals of AstroGrid

The **scientific aims** of AstroGrid are very general and can be summed up as follows :

- to improve the quality, efficiency, ease, and speed of on–line astronomical research
- to make comparison and integration of data from diverse sources seamless and transparent
- to remove data analysis barriers to interdisciplinary research
- to make science involving manipulation of large datasets as easy and as powerful as possible

To achieve this, these are our **practical goals** (the eightfold path) :

1. to develop, with our IVOA partners, standards for data, metadata, data exchange, and provenance
2. to develop a software infrastructure for data services
3. to establish a physical grid of resources shared by AstroGrid and key data centres
4. to construct and maintain an AstroGrid Service and Resource Registry
5. to implement a working Virtual Observatory system based around key UK databases
6. to provide a user interface to that system
7. to provide by construction or by adaptation a set of science user tools to work with that VO system
8. to establish a leading position for the UK in VO work

Informally, we are often guided by four *slogans* :

- The archive is the sky
- Everybody can be a power user
- Shift the results not the data
- A supercomputer on your desk

The explanation for these slogans will arise naturally through the text.

(3) The Virtual Observatory

(3.1) The VO vision

The VO vision can be summed up as the desire to make all archives speak the same language – all searchable and analysable by the same tools, all data sources accessible through a common interface, all data held in distributed databases that appear as one. Much astronomical research will be done by "observing" this virtual sky – *the archive becomes the sky*. The VO then is a system that allows users to interrogate multiple data centres in a seamless and transparent way, which provides new powerful analysis and visualisation tools within that system, and which gives data centres a standard framework for publishing and delivering services using their data. This is made possible by standardisation of data and metadata, by standardisation of data exchange methods, and by the use of a Registry which lists available services and what can be done with them. The long term vision is not one of a fixed specific software package, but rather one of a *framework* which enables **data centres** to provide competing and co-operating *data services*, and which enables **software providers** to offer a variety of compatible *analysis and visualisation tools* and *user interfaces*. The first priority of AstroGrid is to develop the standardised framework which will allow such creative diversity, but we also intend to build a working implementation, with an example interface and a set of tools, and to work with UK data centres to populate the system with real datasets and data services.

(3.2) General Science Drivers

There is no single "killer application", but rather a very general drive towards improvement in three main areas. (i) More ambitious kinds of on-line analysis – not just data download, but complex queries, and the ability to analyse data *in situ*, such as panning across large images, making N-D parameter plots and rotating them, fitting curves etc. This is part of a trend towards on-line working and standardised tools, but remote analysis is also required by the developing *data explosion* in astronomy. (ii) Multi-archive science – collecting all the multiwavelength data on a particular object, cross-matching optical and IR catalogues, or finding the ground-based radar measurements that correspond to the delayed effect of a coronal mass ejection observed with a solar telescope. (iii) Data intensive science – searching for rare objects, calculating correlation functions, or finding clumps in multiparameter space, for billions of objects. All the above can be done now – but slowly and awkwardly. The VO should make such techniques easier, faster, and more transparent – *everybody can be a power user*.

(3.3) Collectivisation and Democratisation

The VO can be seen as part of an inexorable long term trend in astronomy towards communal organisation. The first step was the development of "common-user" or "facility class" instruments. The instrument is already built, documented, and robust, liberating the astronomer to think about collecting data. Next came communally developed data reduction packages, such as Starlink, IRAF, or Midas – one no longer had to hack one's own Fortran. Then came the first on-line archives. Next came a development that has transformed our daily working lives – the availability of push-button on-line information resources, such as Simbad, ADS, and astro-ph. Right now we seem to be in the middle of a trend towards collectivising the collection of data, with large consortium survey projects such as SDSS, 2dFGRS, and UKIDSS. What are the obvious next steps in this process? The first is *interoperability of archives* – making them all speak the same language so one can make joint queries. The second is *automation of resource discovery* – not needing to know the URLs of dozens of different web pages. The third is *facility class analysis tools*. Right now we wouldn't dream of coding our own data reduction routines, but do expect to write our own code for correlation functions, principal component analysis, etc. Such things should become modular and standard – and available on-line attached to the data. Just as with all the previous stages of communal development, although

the evolution seems at first to be in a collectivist direction, the actual effect is to *empower the individual*. With all these tools, you don't need to be in Cambridge or Caltech – one's resources are nearly as good in Sheffield or Florida. This *democratisation* of science is a key driving force in nearly all e–science projects.

(3.4) The Grid

How does the VO concept relate to that of "the Grid"? The Grid concept originally referred to *computational grids*, i.e. distributed sets of diverse computers co–operating on a calculation. The term "Grid" is an analogy with the electrical power grid – vast CPU power is available just by plugging in at home, without needing to know where it comes from – hence the next slogan: *a supercomputer on your desktop*. The relevance of strict computational grids to astronomical data analysis is debateable, but the term "Grid" has expanded to refer to a general sense of transparent access to distributed resources, and a sense of collaboration and sharing. The resources which are shared could be storage, documents, software, CPU cycles, data, expertise, etc. The sharing concerned is usually taken to mean not just an attitude of "help yourself", but a commitment to organised management of resources by a community and/or putting in place mechanisms to ease that sharing. The history of computing can be seen as a gradual evolution towards the grid concept. First came the physical networks, and the protocol stacks, to enable us to pass messages between computers. Next came the World Wide Web, providing transparent sharing of documents. Then came computational grids enabling shared CPU. A new concept is that of a *datagrid*, making possible transparent access to databases. This is close to the Virtual Observatory concept, but to truly reach this ideal, we believe what we need is a *service grid*. This involves not just open access to data sources, but also standardised formats and standardised services, i.e. operations on the data.

(3.4) Geometry of the VO

Two technical issues dictate the structure of our framework. The first is the **I/O bottleneck**. Some problems are limited by CPU–disk bandwidth, which has grown much more slowly than Moore's law, and some are limited by seek time, which has hardly changed at all. This means that searches and analyses of large databases take extremely long unless high throughput parallel facilities are used, along with innovative and efficient algorithms. The second issue is the **network bottleneck**. Networks are improving but are in practice limited by end–point CPUs and firewalls rather than fibre rental, and are not expected to be nearly good enough to routinely move around the new large databases. Given that users can't realistically download large databases, or have room to store them, or have the search and analysis engines required, we are driven to a situation where the data stay put, but the science has to be done next to the data. In other words, data centres have to provide search and analysis services – the slogan is *shift the results not the data*. They will also need to provide file storage and manipulation services so that intermediate results can be worked on before any final download to an end–user's machine.

The above conclusion, together with the fact that the human expertise on any new and exciting dataset will usually live next to the data, dictates the **geometry of the VO**. We do not want a super centralised archive. Neither will we have a truly democratic peer–to–peer network like Napster, or a hierarchical system like the LHC Grid. Rather, what we have is a moderate number of competing specialist *data service centres* and a large number of *end users*. The purist model of independent data centres is in practice likely to be complicated by collaborations between those services. Astronomers will want to cross–match sources in different catalogues on the fly, which seems to require either shifting data across the net, or a single location data warehouse. In fact we expect that collaborating data centres, as opposed to end–users, will be grid–connected by dedicated fat pipes, and an intelligent approach to cross–matching can minimise traffic.

(4) Progress on the key technical issues

(4.1) Standards, Standards, Standards

Our prime targets for progress are as much sociological as technological. We have to evolve, through the IVOA, agreed standard formats for data, metadata, provenance, and semantics. Astronomy has actually been in the vanguard of **data standardisation**, with the FITS format, bibcodes, and so on, but we now must go further. **Provenance** refers to recording the history of where data has come from, who has touched it, which programs have transformed it and so on. This is already normal in good astronomical pipelines, but not standard in archives. As results are extracted from data analysis servers, and passed on to other services and so on, recording this history will become crucial, and we need to agree standard formats for recording such data. **Semantics** refers to recording the *meaning* of columns in a database. A familiar problem is receiving a table with a column labelled "R-mag" and not knowing whether it refers to a Johnson, Gunn, or Sloan R, let alone whether the normalisation is as a Vega magnitude or an AB magnitude. Ideally we want not just to agree terminology for specific quantities, but to specify their relationships in order to allow *software inference*. Finally we also need **software interoperability standards** so that our services can speak to one another. This is relevant to all the areas below.

A key step forward was the development in 2002 of VOTable, an XML-based format for table data, which is now in daily use worldwide. An agreed format for "images" (i.e N-D binary arrays in general) is in progress. VOTable has a method to link to bulk data but it is not satisfactory. NVO have developed a draft standard (Simple Image Access (SIA)) but we are also assessing developing external standards from both the commercial and the academic world (DIME and BinX). In the *semantics* area, CDS Strasbourg have already made an excellent start with their huge tree-structured list of Universal Column Descriptors (UCDs). This has drawbacks, for example not easily allowing parameterised values, or non-tree-like structures, where for example one daughter can have multiple parents. It is also strictly a content descriptor, not a unique column name, so that "Right Ascension" may appear in a database table more than once – as a property of the object and its nearest neighbour for example. For all these reasons, several rival replacement systems are under debate, but they will not reach agreement soon enough for AstroGrid. We are therefore building using an updated UCD system, in the knowledge that re-engineering will be needed in later generations of the VO infrastructure. Likewise, after some initial research, we are ignoring developments in ontology (DAML, OIL, OWL etc), but expect to incorporate these in later VO generations.

(4.2) Internet Technology

To construct a VO, we need to take advantage of several developments in internet and grid technology. The first requirement is **protocols for exchanging and publishing data**. The idea of *web services* has almost solved this problem, with XML data formats, SOAP message wrappers, and Web Service Description Language (WSDL). (Of course, a particular community has to design its own particular XML formats, as in the VOTable example above). The problems are that standard web services are one-to-one, stateless, and verbose, so we need to add methods for linking to bulk binary data, for composing multiple services with lifetime management, and for defining and controlling workflow. These issues are potentially solved by the concept of *grid services* within the standard OGSA framework. AstroGrid's basic plan is to build in 2003 using web services, and then to re-engineer in 2004 with OGSA grid services where appropriate. This applies to production software – in parallel we are already writing experimental software using grid services, and have used them in demonstrations. In the meantime, we have written dozens of web services. Almost everything is a service. Rather than being just a way to deliver content from data providers to users, we have conceived our whole architecture as a set of interlocking services.

(4.3) Registries

Before some portal software can connect a user to data services, it needs to know of their existence, which requires their **publication in a Registry**. There is a developing commercial registry called UDDI, but its structures match poorly onto Astronomy, so we have been constructing a specialised *AstroGrid Registry*. As well as simply advertising service availability, the Registry collates coverage information and other metadata from available datasets, so that many queries, and the first stage of all queries, can be answered directly from the Registry before going to the remote service. We have developed a prototype AstroGrid Registry which can be queried either manually, or automatically by web services. It is defined by an XML schema, and currently instantiated as an XML file and queried via XQUERY. (It may be converted to a DB2 database soon, but actually not all the required information fits neatly into a table structure..) The resource metadata contained complies with a provisionally agreed IVOA standard. The registry has been populated by hand with real resource metadata for a limited set of key data resources. In the latter half of 2003 we will be expanding the resource list and also implementing methods for automatically harvesting updated metadata from published services. It has also become clear to us that the Resource Registry is only one of several we need, with others needed for MySpace and AstroPass. The US–NVO project has also developed a prototype registry based on the provisional standard. There is some disagreement about how to proceed to the next stage, with issues on flat versus deep structures, on fine grained versus coarse grained registries, and on whether there are multiple registries for different purposes or one super–registry. We expect to reach agreement and have a new standard shortly.

(4.4) Single sign–on and collaborative spaces

A key requirement is a method of transmitting **identity, authorisation, and authentication** to achieve the goal of single–sign–on use. One doesn't want a trawl round the world's databases to stop thirteen times and come back and ask you for another password. There are several commercial solutions to this problem but for a variety of reasons they are not appropriate for astronomy. We have chosen to follow the *Community Authorisation Server (CAS)* model from Globus, using X.509 certificates and standardised *distinguished names*, although we are developing a simplified and modified version we are calling *AstroPass*. Users will also want to store intermediate results wherever they came from and analyse them, without having to download data to their own machine. This needs a kind of virtual personal filestore, which we refer to as "MySpace". (The software infrastructure will also need to use such a transparent store). The requirement to have a certificate before you can use AstroGrid will be the most controversial development from the point of view of the typical end–user, but we believe they will soon see the advantage, especially when put together with the MySpace system. It will be possible for example to create arbitrary lists of collaborators who can all see the same protected data.

AstroPass is still at the design stage, but will be implemented during the latter half of 2003. We have made significant progress with MySpace. The virtual filestore is defined through an XML Registry and controlled by "MySpace Manager" which is composed of a set of web services invoking java classes to perform all the functions required. (A good example of everything being a service). As well as being used by the AstroGrid infrastructure extensively, the system is viewable and manipulable by the end–user through an interface called "MySpace Explorer". It is becoming clear that each data centre will want to install a local version of MySpace Manager and that this is a key element of being part of the VO. Currently communication is over HTTP, but shortly collaborating MySpace components will communicate by GridFTP. The MySpace concept and its implementation is currently unique amongst VO projects, but we believe it is crucial to the success of the VO.

(4.5) Data Access and Workflow

In the future, with the VO infrastructure in place, any data centre should be able to publish data services compliant with the framework. Writing and publishing data services costs effort of course. E-science raises the game for data centres just as the arrival of the Web did in the 1990s, and we hope that funders recognise this need. We will be writing a "cookbook" to aid this process. AstroGrid is also however a consortium of data centres and so in the first instance actually constructs constituent services. To date we have integrated a handful of databases in the infrastructure, with more to be added over the coming months. This involves transforming data from a variety of DBMS via JDBC to the VOTable standard, and making output available to the Registry interface and the MySpace system. A problem that became apparent is that standard commercial DBMS expect to work synchronously, whereas output has to be passed around our components asynchronously, requiring some kind of *workflow system*.

"Workflow" often refers to two separate things in e-science projects. The first is the ability of an end-user to compose and control complex jobs, preferably with both a drag-and-drop interface and a scripting mode. Several models of how to do this are emerging from other e-science projects, and we expect to adopt one of these shortly. The second use of "workflow" refers to job control, resource scheduling, query routing and so on. For multi-site distributed operation, these are key issues in the developing world of Grid Technology, from which components will be selected and deployed as necessary. However, even within a single site, synchronising communications between our various web services and databases is a significant issue, and we therefore designed and constructed a message queue system, which we have called AstroMQ.

(4.6) User Interface and Tools

The VO framework will only produce science once there are tools that use it. We can think of these in three categories. (i) Data intensive exploration, visualisation, and analysis services – cross matches, source detection algorithms, cluster analysis, Fourier transforms, etc. These will be CPU and or I/O hungry operations on the data, offered by data centres and/or the Grid as a remote service. (ii) Client tools – simple image display, plotting of returned data, statistical analysis of small tables, etc. We expect that large numbers of these will be written to use the framework by third parties and offered as Java applets or installed as stand-alone applications. (iii) Lightweight tools as above, but offered by data centres as services (usually Java servlets).

AstroGrid is concentrating on the basic VO infrastructure and has little effort to spare to construct tools from scratch. Instead we are selecting a standard set of tools to adapt, and or working with our European and US partners to develop new ones. Three good examples are the Aladin image viewer from Strasbourg, the new SED tool from ESO, and VOPlot developed by the Indian VO programme. However we are writing a specialised AstroGrid user interface. The working assumption is that the user will not need to install any new software to use AstroGrid, but only to start from the AstroGrid web page, linking to pages for querying the registry, examining the MySpace system, setting up certificates, etc. All the working software is at the various servers. The key technology here is Apache Cocoon. Actions begun at the web page invoke Java classes which return html and/or applets which are fed standard data formats.

(5) Project Status

AstroGrid has now completed a Phase A study and two out of eight quarterly iterations in its Phase-B construction phase. It involves a team of 26 people expending 23.4 FTEs of effort, and has spent approximately 40% of its budget. We have written approximately 10,000 lines of production code, with nearly all of this being in the last few months. We have issued two open software releases (on

schedule) and produced three separate demonstration suites.

AstroGrid follows rigorous software engineering standards. The initial staff complement was dominated by scientists and scientist-programmer hybrids, but nearly all the new recruits have been professional developers from a commercial background. We have developed fairly strict development procedures for UML design, version control, unit testing, and so on. Overall we follow a modified version of the "Unified Process", being architecture driven, use-case centric, and iterative, with new workplans each quarter. Part of the rigour is releasing completed software via a public web page at the end of each quarterly cycle.

We have pioneered an open-project collaborative style mediated by new Web technologies. We run AstroGrid News and AstroGrid Forum websites which are readable by anybody, and postable to by registered users. We encourage a wide membership of the Forum well outside the project. It currently has 115 members, including for example Ian Foster. For document development, including use cases, UML designs, and schemas, we set up the AstroGrid Wiki. This uses the Twiki technology which allows registered users to directly edit pages and create new ones. The Twiki was primarily meant for internal project collaboration, but in fact is regularly perused by others, and we have set up "oversight pages" which enable outsiders (especially PPARC committee members !) to track our progress. The Twiki method has now been emulated by other VO projects and the IVOA, and has also stimulated interest more widely across the e-science scene.

We are concentrating on the basic "engine room" work, and have constructed working versions of many of the key components of our architecture – the Registry, the MySpace system, Job Control, Message Queue, Data Access and the Portal – along with experimental grid services, and an example of a heavy duty remote service – the Astronomical Catalogue Extractor (ACE), written together with AVO. The key components still to start are AstroPass, user workflow, and registry harvesting, and wrapping of a variety of tools. The existing software has been demonstrated in several public arenas, but under controlled circumstances. The next key stage is releasing the software for experimentation and feedback by a pool of beta testers. We have set up an AstroGrid Science Advisory Group (AGSAG) composed of astronomers from outside the project. As well as meeting regularly and providing advice, they will act as the pool of beta testers. This will start in October 2003.