PPARC

Astro Grid

The AstroGrid Phase A Report

IoA University of Cambridge

Queen's University Belfast

UCL MSSL

University of Leicester

Rutherford Appleton Laboratory

Institute for Astronomy

Table of Contents

(0) Executive Summary....1
(0.1) Introduction....1
(0.2) Project Progress....1
(0.3) Financial Overview....2
(0.4) Phase B Proposal....2
(0.5) Conclusion....3

(1) Project Vision....5
(1.1) The AstroGrid Vision....5
(1.2) Development of the Project....5
(1.3) General Science Drivers....6
(1.4) Specific Science Requirements....6
(1.5) The Virtual Observatory Concept....7
(1.6) The Grid Concept....7
(1.7) The technical route forward....7
(1.8) World context....9
(1.9) Project Methodology....9
(1.10) Building the VO....9
(1.11) Goals of the AstroGrid Project....10

(2) The Science Analysis Summary....13
(2.1) Summary....13
(2.2) Introduction....13
(2.3) Capturing the Science Requirements....13
(2.4) Shaping the AstroGrid Drivers from the Virtual Observatory Science Cases....14
(2.5) 'The AstroGrid Ten' Science Drivers....15
(2.6) Providing Functionality as Indicated by the AstroGrid Ten....21
(2.7) The AstroGrid Science Input into Partner Virtual Observatory Projects....24
(2.8) The Evolutionary Path....24
(2.9) Closing Remarks....25
(2.10) References....25

(3) Architecture Overview....29
(3.1) Introduction....29
(3.2) Approach....29
(3.3) Use Cases....30
(3.4) Conceptual Model....30
(3.5) Services Model....31
(3.6) Technology Demonstrations....36
(3.7) Technology Choices....36
(3.9) References....37

(4) Virtual Observatory Prototypes....39
(4.0) Introduction....39
(4.1) Current Services, Sites, and Software....39
4.2 Testing CDS and NED with use cases....45
4.3 Discussion....46

(5) Report on Grid Technology....49
(5.1) Introduction....49
(5.2) The natural architectures....49
(5.3) Astrogrid's technology needs....50
(5.4) Technology choices....51
(5.5) Technology−choice matrix....53
(5.6) Experiments....54
(5.7) References....56

Table of Contents

(6) Interoperability Report 57
(6.1) Interoperability 57
(6.2) Computational Infrastructure Interoperability 57
(6.3) Astronomy−specific Interoperability 57
(6.4) Ontology 60
(6.5) References 61

(7) Database Technology and Data Mining 63
(7.1) Introduction 63
(7.2) Use−cases 63
(7.3) Data Exploration and Data Mining 66
(7.4) DBMS Evaluations 67

(8) Pilot Programme Report 71
(8.1) The AstroGrid Pilot Programme 71
(8.2) Pilot Selection 71
(8.3) Highlights from the Pilot Programme 73
(8.4) Summary and Conclusions 85

(9) Financial and Management Report 87
(9.1) Introduction 87
(9.2) Management 87
(9.3) Personnel 89
(9.4) Finances 90
(9.6) References 92

(10) Phase B Plan 93
(10.1) Introduction 93
(10.2) General Approach 93
(10.3) Management 94
(10.4) Project Work Plan 95
(10.5) Personnel Plan 97
(10.6) Financial Plan 98
(10.7) Milestones 99
(10.8) References 101
(10.9) Detailed Effort Estimates 101

(0) Executive Summary

(0.1) Introduction

AstroGrid is one of three major world−wide projects (along with European AVO and US−VO projects) which aim to create an astronomical ***Virtual Observatory (VO)***. The VO will be a set of co−operating and interoperable software systems that:

- allow users to interrogate multiple data centres in a seamless and transparent way;
- provide powerful new analysis and visualisation tools;
- give data centres a standard framework for publishing and delivering services using their data.

The long term vision of the Virtual Observatory is not one of a single software package, but rather of a framework which enables data centres to provide competing and co−operating data services, and software providers to offer compatible analysis and visualisation tools. The first priority of AstroGrid, along with the other VO projects worldwide, is to develop this standardised framework to allow such creative diversity.

However, our intentions are to go beyond this framework. We will develop a working implementation of immediate use to astronomers. As a consortium of data centres and software providers, we will pool resources, including key UK databases, storage, and compute facilities. As a UK e−Science project, our architecture will be firmly based on a data−grid approach: we will make use of grid components produced by other projects and will make our own components freely available to the e−Science community.

This document is the formal report of the AstroGrid Phase A study to PPARC's Grid Steering Committee (GSC). It includes a summary of the project vision, an overview of our intended architecture, a discussion of science requirements, a series of study reports, a summary report on project management and finances, and a description of our Phase B plan.

Following a formal proposal to PPARC in April 2001, AstroGrid began in September 2001 with a one−year Phase A study, with final project funding to be reviewed at the end of Phase A. Following this review, we expect to begin the software construction phase (Phase B) at the beginning of 2003. Although this document is in the first instance intended for this formal review, we intend in due course to circulate it widely. It is available as a single pdf file or as a connected set of pages on the AstroGrid *Wiki* web site.

(0.2) Project Progress

Project progress has been significant, and the greater part of this Phase A Report documents the project's achievements. Some of the highlights are:

- ***science requirements***
  Through internal meetings and contact with external scientists, we developed an extensive list of requirements, documented as science problems. These were the problems that any VO would be expected to enable scientists to tackle. From these we selected ten key science cases, which the AstroGrid project would take as its drivers.
- ***architecture***
  From the key science drivers, we derived use cases: formal definitions of the steps that a scientist would follow with the VO to tackle each problem. The use cases were then used to derive an architecture for the AstroGrid software framework. This architecture is still under development and is being created fully using the standard Unified Modelling Language (UML).
- ***external relations***
  Our team has had extensive contact with people and projects worldwide, in the VO, grid and e−Science arenas:
  - ***VO***: AstroGrid is a founding member (along with the European AVO and US−VO projects) of the International Virtual Observatory Alliance (IVOA). We work closely with our IVOA colleagues in the definition of interoperability standards. The ***VOTable*** data exchange format has been a great success. We are currently engaged in discussions about an image access protocol.
  - ***grid***: We have documented our experiences of implementing of a working data grid and have provided feedback to the developers of Globus' CAS component for OGSA based on our use of early beta software
  - ***e−Science***: Following on from our involvement in the e−Science Database Task Force, we have been chosen (along with myGrid) as *early adopters* of the OGSA−DAI technologies. We are in discussion with several other e−Science projects (eg DataGrid, GridLab) about re−using some of their components in the AstroGrid VO.

- ***pilot studies***
  Our original pilot studies were quite extensive. The optical/near−IR and x−ray ones were delayed due to late data arrival but the solar, STP and radio studies have achieved their goals, both in terms of providing input to the architecture development and demonstrating the feasibility of certain technology approaches.
- ***technology demonstrations***
  On the technical side, we have established a working grid over several institutions (currently using Globus V2 toolkit and soon to implement beta versions of OGSA) and have begun demonstration projects in:
  - ♦ grid authentication and authorization,
  - ♦ ontology usage in registry and workflow development, and
  - ♦ database access via the grid.
- ***online collaboration tools***
  We have deployed a set of ground−breaking collaboration tools which have enabled team members to share experiences, seek opinions and expert advice, and create an extensive library of documents pertaining to all aspects of the project. These tools are now being deployed on other project web sites.

(0.3) Financial Overview

The total Phase A expenditure up to the end of August 2002 has been approximately £661K. (This is the PPARC expenditure only). Staff costs have been expended through a variety of grants to the consortium institutes. For University based staff, on top of salary, these costs include standard fractions of secretarial and system management support staff, and the standard PPARC grant overhead of 46%. Some of the grants concerned had other minor costs attached. For RAL based staff, time is charged at a uniform staff rate agreed by negotiation between PPARC and CLRC.

The total funded staff effort over the year has been 8.25sy, some of which, the AVO funded staff effort, is at zero cost to PPARC. Grant claims are not all made yet so the accurate final cost is not known, but our out−turn forecast for the one year staff−related costs to PPARC, including all the above related costs and overheads, is £492K.

Non−personnel expenses were provided by PPARC through three grants, two for capital equipment and one central budget grant. These were seen however as representing a single budget controlled by the PM, who set out an overall budget plan in November 2001. Procedures were then put in place for managing these finances. Actual expenditure is very close to that budgeted.

(0.4) Phase B Proposal

Our assumption is that Phase B will commence on 1st January 2003 (ie that Phase A is extended to 31st December 2002 to enable the completion of the system architecture and the demonstration projects). The end date will be approximately December 2004, but in fact the effort profile will not be flat, and a small amount of staff effort will extend into 2005. The end goal of the project is to produce software which will enable the creation of a working, grid−enabled Virtual Observatory (VO) based around key UK astronomical data centres.

Our approach to Phase B will be based upon the Unified Software Development Process (UP, or our own variant of it, UPeSc). The UP is a software development methodology which is both iterative and incremental: each iteration contains analysis, design, code, test and deployment activities and each iteration incrementally adds to the functionality of the system components. This approach will replace the work−package approach of Phase A.

The work we anticipate breaks into a few major strands:

- Continuing research and development;
- Developing the software infrastructure that will make a VO possible;
- Developing user tools to make it possible to do science with AstroGrid (portals, visualisation tools, analysis tools, datamining algorithms, workflow editors, and so on).

The iterative/incremental approach has allowed us to specify milestones in terms of delivered functionality, though these may be switched around or altered depending on the evolving needs of the astronomers who test the early releases of the VO.

We have developed a detailed, component−level breakdown of the work required and, from that, estimates for completing each of those components. Our estimates show a total effort required of 48 staff years. Assuming a two year Phase B, this equates to 24 staff. Note that these estimates do not include management, co−ordination, and support tasks.

At the moment, the project employs 13.1 FTEs spread across 18 individuals who are actively involved in the project. (This does not include the AGLI but does include 3.0 FTEs funded by AVO at no cost to PPARC). Of these, 2.7 FTES across 4 people are primarily employed in non−development tasks − management and co−ordination, science leadership, support tasks (Project Manager, Project Scientist, Web Developer (0.5 FTE), and RAL co−ordination (0.2 FTE)). This leaves 10.4 FTES spread across 14 indviduals available for development and related research tasks.

The project therefore needs an additional 14 development staff. In addition to this, to deliver such a complex programme of software development, we need a new senior position − the ***Technical Lead*** who will directly co−ordinate developer tasks. In total then, we are asking PPARC for ***15 additional staff***.

(0.5) Conclusion

In summary, these are our ***scientific aims*** :

- to improve the quality, efficiency, ease, speed, and cost−effectiveness 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.

And these are are our top−level ***practical goals*** :

- to develop, with our IVOA partners, internationally agreed standards for data, metadata, data exchange and provenance
- to develop a software infrastructure for data services
- to establish a physical grid of resources shared by AstroGrid and key data centres
- to construct and maintain an AstroGrid Service and Resource Registry
- to implement a working Virtual Observatory system based around key UK databases and of real scientific use to astronomers
- to provide a user interface to that VO system
- to provide, either by construction or by adaptation, a set of science user tools to work with that VO system
- to establish a leading position for the UK in VO work

(1) Project Vision

(1.1) The AstroGrid Vision

The Virtual Observatory will be 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 Registry should embody some kind of "ontology" which encodes the meaning of quantities in the databases served and the relationships between them, so that user software can for example collect fluxes at various wavelengths from various databases and then plot a spectral energy distribution.

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, along with the other VO projects worldwide, is to develop the standardised framework which will allow such creative diversity.

However, the intentions of AstroGrid go beyond this framework. We will develop a ***working implementation*** of immediate use to astronomers. As a consortium of data centres and software providers, we will pool resources, including key UK databases, storage, and compute facilities. On top of this, the AstroGrid project *per se* will provide the first data services, along with a standard "point of entry" user interface, and a set of datamining tools. AstroGrid will also provide central resource on top of that provided by the participating data centres − first and foremost the construction and maintenance of an Astronomical Registry, but also one or more data warehouses, further CPU dedicated to search and analysis tools, and storage and software to create "MySpace", a kind of virtual workspace for grid−users.

Implementing such a functioning VO capability will support UK astronomy in several ways. It should make doing astronomy faster, more effective, and more economic, by standardising the data analysis process and by freeing the astronomer from many mundane tasks. It also has the potential to influence the discovery process in astronomy in a dramatic way − by encouraging new styles of data−intensive exploratory science, by removing interdisciplinary barriers, and by encouraging the pooling of resource and the formation of distributed collaborative teams. We also expect that it will be a liberating force in that the resource available to astronomers will become almost independent of their location.

(1.2) Development of the Project

AstroGrid has its origins during the Long Term Science Reviews (LTSR) undertaken by PPARC in 1999/2000, which placed IT initiatives in astronomy in general, and large database initiatives in particular, as high priorities in all the panel areas. (Similar ideas were developing across Europe and the US, and for example construction of a US "National Virtual Observatory" was recommended by the NSF decadal review). Meanwhile e−science and the Grid played a large part in Government thinking during the 2000 spending review, and an AstroGrid project concept developed by astronomers from Leicester, Cambridge Edinburgh and RAL was used by PPARC in its bid. A "white paper" on AstroGrid was reviewed by PPARC Astronomy Committee in October 2000, and debated around the community. The result was an expansion of the consortium to seven institutions, and an increased remit to cover solar and solar−terrestrial physics as well as optical, IR, X−ray and radio astronomy. A formal proposal was submitted to the PPARC e−science AO in April 2001, and a funded project finally began in September 2001. Initial funding was for a one−year Phase A study, with final project funding to be determined by a review at the end of Phase A.

During Phase A we have concentrated on the following main activities. (i) Requirements analysis, including community consultation, development of key science problems, and articulation as formal use cases. (ii) Development of a UML−based architecture. (iii) Technology assessment reports. (iv) A series of small software demos to test ideas and show them to others. (v) Development of interactive collaborative web pages − a static portal, a News site, a Forum site, and a Wiki for collaborative construction of documents and software. This document is a report on those Phase A activities, accompanied by a Phase B plan, with this section (Project Vision) being an overall summary of where we are and where we are headed. The Phase A study is being reviewed by PPARC's Grid Steering Committee in Oct 2002, following which we expect to begin our construction phase at the beginning of 2003.

(1.3) General Science Drivers

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

- to improve the quality, efficiency, ease, speed, and cost−effectiveness 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.

The first driver is then to accelerate the quality of ***on−line research***. Astronomers already do much of their research on−line through data centres. The idea is to step up the quality of service offered by those data centres, beyond simple access to archives by downloading subsets. This will mean the ability to make *complex queries* of catalogues of objects or catalogues of observations, and the ability to *analyse* the data in situ − for example to transform or pan across an image, or to draw a colour−colour−colour plot for selected objects and rotate it. Such improved service can be seen as part of a long trend in astronomy to develop communally agreed *standard tools* so that the astronomer can concentrate on doing the science rather than wiring their own instruments, or hacking their own data reduction software. Following facility−class instrumentation, then facility−class data reduction tools (Starlink, IRAF, Midas etc), then easy access to data and information (on−line archives, ADS, Vizier, etc), the next step is facility−class analysis tools. However we are also driven to this solution by the expected ***data explosion*** in astronomy. For very large datasets, such as the optical−IR sky survey which VISTA will accumulate at hundreds of TB per year, users can't afford to store their own version, or have time to download it. Data centres are therefore driven to provide analysis services as well as data access.

Along with improved query and analysis tools, the next driver is the ability to make ***multi−archive science*** easy. The study of quasars requires data at all wavelengths; finding rare objects such as brown dwarfs involves simultaneous searching in optical and IR data; study of the solar cycle involves putting together data from many different satellites over eleven years or more; and so on. There is increasing interest in combining data from different disciplines, such as linking solar observations of coronal mass ejections to changes found in monitoring of the Earth's magnetosphere. The idea is to transform this kind of science from slow and painful hand−driven work to push−button easy, so that through a single interface one can make *joint queries* such as "give me all the objects redder than so−and−so in UKIDSS that have an XMM ID but don't have an SDSS spectrum", or ask higher−level questions, such as "construct the spectral energy distribution of the object at this position". Sometimes the tasks will involve predetermined lists of data services, but often they will involve the AstroGrid system making a trawl and deciding what is relevant, using some kind of *registry of services*.

As well as offering improved data services, and multi−archive services, we wish to facilitate ***data intensive science***. Some of the most interesting science comes from manipulation of huge numbers of objects. This can mean looking for rare objects, for example those with strange colours or proper motions, or constructing a correlation function, or fitting gaussian mixtures to N−D parameter sets, and so on. At the moment such projects are the province of specialist "power users", but the vision is to make such analysis easy, as a service through data centres. This will require data centres to provide not just storage but also high−powered search and analysis engines. In addition, we need to develop *standard tools* for such kinds of analysis, and a way for users to *upload their own algorithms* to run on the data. We see all this as *empowerment of the individual* in astronomy. One doesn't need to be at Caltech or Cambridge to have the very best resources at one's finger tips. AstroGrid and other VO projects will not provide everything needed for this new kind of science − many others will invent the algorithms and write the software tools − but we need to put the framework in place to make this possible, and to provide at least some tools in our early working system.

(1.4) Specific Science Requirements

The previous section summarises the general scientific aims of AstroGrid. There is no specific scientific topic driving the project − the infrastructure should serve a whole range of present and future scientific concerns. However, in order to actually build the system we need concrete requirements, and in order to construct these we need to look at some specific scientific questions in detail. We therefore collected a series of ***Science Problems*** and analysed them in a fairly formal blow−by−blow manner. These were both contributed by Project members, and collected by community consultation. There were too many of these to use as formal requirements for the system architecture. We therefore selected the ***AstroGrid Top Ten***, chosen to represent a range of science topics, and to encapsulate the key recurrent technical issues. From these we then developed more formal *use cases* and *sequence diagrams* to feed into the architectural design. The Top Ten Science Problems used were :

- Brown Dwarf Selection
- Discovering Low Surface Brightness Galaxies
- The Galaxy Environment of Supernovae at Cosmological Distances

- Object Identification in Deep Field Surveys
- Localising Galaxy Clusters
- Discovering High Redshift Quasars
- The Solar−Stellar Flare Comparison
- Deciphering Solar Coronal Waves
- Linking Solar and STP events
- Geomagnetic Storms and their impact on the Magnetosphere

(1.5) The Virtual Observatory Concept

The science drivers described above are closely related to the popular concept of a "Virtual Observatory", especially the ideas of multi−archive science, and transparent use of archives. The idea can be summed up in one sentence. ***The aim of the Virtual Observatory is to make all archives speak the same language.***

- all archives should be searchable and analysable by the same tools
- all data sources should be accessible through a uniform interface
- all data should be held in distributed databases, but appear as one
- the archives will form the_Digital Sky_

To this now standard VO vision, AstroGrid adds the desire that more advanced *analysis and visualisation tools* should be available for studying the digital sky, and that high−powered computational resources should be available for undertaking *data intensive studies*

(1.6) The Grid Concept

The "Grid" concept originally referred to *computational grids*, i.e. distributed sets of diverse computers co−operating on a calculation. However, the idea 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 term "Grid" is an analogy with the electrical power grid. Spread over the nation there is a network of huge power stations, but the user doesn't need to know how to connect to them. One simply plugs one's hair−dryer into the socket, and electricity flows. The history of computing can seen as an 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 popular concept now 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 that 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. Beyond this, the Grid community talk of *information grids*, *knowledge grids*, and Virtual Organisations_.

The general Grid idea of transparent access to resources is then central to AstroGrid and the VO concept. At first sight our vision of a service network, where data access and computations are provided by one data centre at a time, and results are combined by the client, doesn't seem to embody the Grid version of pooled managed resources and communal collaboration. However, what we expect is that the collaboration and pooling will be by consortia of data centres, on behalf of the community, to give the best possible service to users. Therefore although we don't often expect to make diverse computers collaborate on calculations, we do expect, within our consortium, to route queries to multiple nodes, in awareness of the various hardware resources and their state at the time, and to establish dynamically updated mirrors and warehouses of our combined key databases. This will need a collaborative approach to resource and fabric management, job scheduling and job control, and so many of the key Grid concepts and software technologies will be of direct relevance. We also need ***dedicated high speed networking*** between collaborating data centres.

(1.7) The technical route forward

Standards, Standards, Standards. Our prime targets for progress are as much sociological as technological. We have to evolve agreed standard formats for data, metadata, provenance, and ontology. Astronomy has actually been in the vanguard of data standardisation, with the FITS format, bibcodes, and so on, but we now must go further, and need to produce XML−based standards to fit into the commercial computing world. Obviously this cannot be done by AstroGrid in isolation, but by international discussion. A key step forward has been the recent development of *VOTable*, an XML−based format for table data. ***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. ***Ontology*** refers to recording the *meaning* of columns in a database, and the relationships between them. 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 quantitities, but to specify their relationships in order to allow software *inference* using the data. CDS Strasbourg have made an excellent start in this area with their huge tree−structured list of Universal Column Descriptors (UCDs) but we need to improve these ideas and translate them into new XML−based ontology markup languages (DAML and OIL, and eventually the emerging W3C standard OWL).

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). 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. However, before some portal software can connect a user to web 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 will write a specialised *AstroGrid Registry*. As well as simply advertising service availability, the Registry will collate coverage information and other metadata (including ontology) 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. The next 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 is a variety of 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*. Finally, there is the issue of ***managing distributed resources***, implicit in our intention to act as a consortium of data centres − job control, resource scheduling, query routing and so on. These are the key issues in the developing world of Grid Technology, from which we will select and deploy as necessary.

To enable the kind of data intensive science we envisage, we need to take advantage of improved ***datamining algorithms and visualisation techniques***. This is an important area for AstroGrid, as it provides added value to the basic multi−archive data services framework. To a first approximation, it is the job of scientists world−wide to invent new algorithms, of a variety of sofware providers to realise these as software tools, and of participating data centres to implement them as services. However, in our effort to kick−start the VO world, AstroGrid will work on the development of example techniques. Also of course, the portal software needs to understand the kind of services available in order to provide an interface to them.

There are two simple technical issues which dictate the structure of the framework that we set up. 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 (clusters and/or multi−processor machines) 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 motto is *shift the results not the data*.

The above conclusion, together with the fact that the human expertise on any new and exciting set will usually live next to the data, dictates the ***geometry of AstroGrid***. 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 *data service users*. AstroGrid itself offers a specialist *Registry service* as well as a portal. In the future other organisations could offer competing registries. The purist model of independent data centres is in practice likely to be complicated by collaborations between those services. (The AstroGrid consortium is precisely such a collaboration.) For example, very often 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 users, will be connected by dedicated fat pipes, and an intelligent approach to cross−matching can minimise traffic. We also expect that as part of natural competition, any one data centre could choose to offer a warehouse with many catalogues, although typically this would not be the latest version of a currently growing archive such as SOHO, HST, or UKIDSS.

(1.8) World context

AstroGrid is not an isolated project. Firstly it is connected to a variety of UK e−science projects from whom it can take both lessons and actual software. The two most important examples are *GridPP* and *MyGrid*. GridPP, the UK contribution to the LHC Grid project, is by far the most advanced in terms of actually constructing a working Grid, and will be the prime source of experience and software for resource management and job control. MyGrid is a UK e−science biology project. The requirements of bio−informatics are very similar to those of astronomical e−science, with a variety of heterogeneous databases, an increasing need to search and analyse multiple databases, and a desire to manipulate large amounts of data, along with an even stronger emphasis on metadata and provenance. In addition, computer scientists involved in the MyGrid project are world experts in ontology, an area we are sure will grow in importance for astronomy.

Next, AstroGrid has good working connections with the UK e−science *core programme* centred around the National E−Science Centre (NeSC) in Edinburgh and Glasgow, the Grid Support Team at RAL, and regional centres in Belfast, Manchester, Cambridge and UCL. The most important connection is that with the OGSA−DAI project. OGSA (Open Grid Services Architecture) is a joint Globus−IBM project aimed at merging the ideas of the web services world and the Grid world. The UK programme has identified structured database access over the grid as a key problem which the UK will lead, through a GGF working group, and by forming the OGSA−DAI (Database Access and Integration) project. AstroGrid and MyGrid have been declared "early adopters" of OGSA−DAI products, and we are already working with the team.

Finally AstroGrid has close relations with other VO projects worldwide. The two most important are the US−VO project, and the EU funded Astrophysical Virtual Observatory (AVO) project. The AstroGrid consortium is formally a partner within AVO, which funds two posts in Edinburgh and one at Jodrell Bank. In return, a similar number of PPARC funded FTEs within AstroGrid are available as effort to AVO. AstroGrid has special responsibility for technology development within AVO. Our work packages are carefully aligned to maximise the joint usefulness of work done.

Specific implementations (such as AstroGrid) of working datagrids, user tools, portals, and data services do not have to be globally identical. The framework being developed should encourage creative diversity. There can even be rival registries. However we do have to evolve towards a situation where the ***underlying infrastructure*** of standards, protocols, and key software elements are universal. To this end, since late 2001, the three major funded projects (US−VO, AVO, and AstroGrid) have held both joint workshops and monthly Lead Investigator telecons. In June 2002 at the Garching conference "Towards an International Virtual Observatory" we officially formed the ***International Virtual Observatory Alliance (IVOA)*** , agreed a Roadmap, and added members from further nascent projects in Germany, Australia, Canada, and Russia. The IVOA is certain to grow in importance.

(1.9) Project Methodology

The project began with a fairly standard workpackage structure. However we soon decided to run the project along the lines of the ***Unified Process***. This means being *use−case centric*, *architecture driven*, and *iterative*. The formal architecture is being developed in the Unified Modelling Language (UML). We began constructing formal blow−by−blow use cases, but soon found that before this was possible we needed one layer of abstraction above, formulating ***Science Problems*** from which use cases can then be articulated. We collected a large number of these, and so picked the ***AstroGrid Top Ten*** science problems, selected, not as the most important, but as representing a good spread of the kinds of problems we need to solve. The formal approach to architecture is important, but on the other hand, the aim of iteration is that the project design does not freeze too early, but converges along with implementation in quarterly cycles, while code developers remain *agile*.

Working as a distributed project is not trivial. To this end, we developed several web−based ***collaboration tools***. All of them are interactive, with registered members able to make postings as well as read entries. The first is a *News* site. The second is a *Forum* aimed at discussion of technical topics. But the most interesting is the *AstroGrid Wiki*, where we jointly develop documents, record meeting minutes, collate links to other work, deposit code, and so on. Any member can directly edit any of the web pages. This has been an enormous boon to productivity, and to keeping track of developments.

(1.10) Building the VO

To create the Brave New World, several strands of work are needed − by the VO projects, by data centres, and by astronomical software providers.

(a) The VO projects will work together to ***develop agreed standards***, for data formats, tables, metadata, provenance, and ontology. This is already happening as a natural consequence of the work programmes of the various VO projects, which are

creating de facto standards. (Eventually they should achieve a more formal endorsement through the IAU.) A significant amount of AstroGrid effort will be expended in this direction.

(b) Each VO project has expended substantial effort in ***research and development***, and in assessing new technologies. For the AVO, and for Astrogrid Phase−A, this has been the main purpose of the work to date, and for AVO will continue to be so. For Astrogrid Phase−B we will be concentrating on implementation, but we still expect continuing R&D at approximately 20% of staff effort. Partly this is because of the iterative converging nature of our software development process (see below), but it is also necessary because both the commercial and the academic technologies that we are building on are changing rapidly. Also of course, we need to be in a strong position for whatever e−science work follows on from completion of AstroGrid.

(c) Next, the VO projects will be developing ***software infrastructure*** − components such as a job scheduler, data router, query optimiser, authorisation server, registry, and so on, along with choices of technology such as SOAP and WSDL, OGSA, OIL, etc. Building this infrastructure will take the largest part of the AstroGrid Phase−B staff effort. The software components used by VO projects worldwide do not have to be globally identical, but in practice as we exchange experience, they are likely to converge. However, the major VO projects, while starting at much the same time, have different timescales. US−VO is a five year project. AVO is a three year R&D project, with the intention of an ensuing Phase B build phase. AstroGrid is intending to complete a software infrastructure in three years. We expect that a large fraction but not all of the AstroGrid code will still be used in later VO work.

(d) Once the necessary standards and software infrastructure are in place, and on the assumption that at least one registry is constructed and maintained, then *Data Centres* around the world can ***publish data services***, i.e. can make available queries on, and operations with, their data holdings. This implies some work by those Data Centres to play the game, but this will be seen as being as normal and as necessary as today writing an organisation's web pages is seen to be. The Data Centres will establish and maintain their data in whatever format they like, and will build engines to deliver queries, analysis, visualisation, or whatever, but will use the VO−provided infrastructure to build a standard interface to their services.

(e) In order to actually do science with the data returned we will need some ***front end tools*** − for example some kind of portal; tools to view images, spectra, time series, etc; tools to plot spectral energy distributions from returned multi−wavelength data and to fit models; tools to make N−D plots and rotate them; and so on. Likewise the services offered by data centres need to offer not just data extracts, but ***data manipulation tools*** such as Fourier transforms, cluster analysis, outlier detection, and so on. Most such tools will not be developed by the VO projects, or even by the data centres, but by a variety of software providers all over the world, just as now, but with the addition that such tools will need to be *VO compatible*.

(f) The first three strands above are work to be done by the VO projects, whereas the latter two strands (publishing data services and developing data mining algorithms) represent work that will be done by many different organisations and individuals worldwide. However, AstroGrid, as well as being a VO development project, is a consortium of data centres and software providers, and expects to develop an ***early working system***. This is partly to act as a proof of concept and exemplar to other future users of the VO infrastructure, but also to build a tool of real daily use to scientists. This will mean constructing and maintaining a Registry, writing data services for key databases of UK interest (.e.g. UKIDSS, SOHO, XMM and so on), a user portal, and some user tools and datamining tools. Full development of such tools is too large a job for AstroGrid, so we are likely primarily to adapt existing tools, such as Gaia or Querator.

(g) A working implementation such as that described above has to run on real physical resources. The final task of the AstroGrid project is therefore to ***establish and manage a physical service grid***. Much of the resource (data, storage, search engines, analysis engines) will be provided by the data centres that are members of the AstroGrid consortium, but further resource − some storage and CPU − will be supplied by the AstroGrid project *per se*. This will be to establish one or more data warehouses, to maintain and operate the AstroGrid Registry, and to provide "MySpace", a virtual storage and workspace system for AstroGrid users. We will also be investigating how to maximise the bandwidth between the participating data centres.

(1.11) Goals of the AstroGrid Project

In summary, these are our ***SCIENTIFIC AIMS*** :

- to improve the quality, efficiency, ease, speed, and cost−effectiveness 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.

And these are are our top−level ***PRACTICAL GOALS*** :

- to develop, with our IVOA partners, internationally agreed standards for data, metadata, data exchange and provenance
- to develop a software infrastructure for data services
- to establish a physical grid of resources shared by AstroGrid and key data centres
- to construct and maintain an AstroGrid Service and Resource Registry
- to implement a working Virtual Observatory system based around key UK databases and of real scientific use to astronomers
- to provide a user interface to that VO system
- to provide, either by construction or by adaptation, a set of science user tools to work with that VO system
- to establish a leading position for the UK in VO work

(2) The Science Analysis Summary

(2.1) Summary

This report describes the derivation and formulation of the science requirements that are being used to define the scope of the Phase−B development of AstroGrid. It is noted that the AstroGrid project has determined a sample set of representative science cases, the 'AstroGrid Ten, from which the necessary AstroGrid product deliverables are derived. Each science case sets challenging demands, in the areas of resource discovery (from data through published literature sources), manipulation of large, multi−location, mult−TB data sets, application of sequences of algorithmic processing, and so forth. AstroGrid will provide tools and systems to aid the astronomical researcher as they transform experimental and theoretical model data into the information by which the physical processes under investigation can be understood.

(2.2) Introduction

The AstroGrid 'vision' is described elsewhere in Chapter 1 of the RedBook. Briefly it is one whereby AstroGrid will provide the UK astronomy community in particular, and the global astronomy community in general, access to powerful, sophisticated, distributed data and advanced processing capabilities. An emphasis will be in enabling more efficient science (e.g. by speeding processes that are currently undertaken via access to improved tools for accessing and manipulating multiple and distributed data sets). Key challenges here are in providing a system which can provide rapid access to large, distributed data sets. Likewise, AstroGrid will enable more effective science via its focus on providing improved workflow capabilities, e.g. in the development of its ontological processes which aim to provide directed workflows for common sets of tasks.

This report focuses on the key science drivers that have been taken by the AstroGrid project in order to determine the which deliverables should be produced by the project by the end of its three year development cycle. The process by which the science drivers were obtained is described, noting the importance of the initial gathering of a wide set of requirements which were subsequently narrowed to give a final, well defined, set of drivers. These were chosen to be representative of science of current and near term relevance to the UK community, with a good coverage of astronomy, solar and solar−terrestrial physics cases. The 'AstroGrid Ten' science cases are described in detail, along with the major areas of functionality that are required by these cases. Thus the 'AstroGrid Ten' provide the primary science drivers for the Phase−B development of AstroGrid.

Specific targets demanded by the science drivers which will need to be satisfied by the final project deliverables are noted. These are set in terms of volumes and types of data sets that may need to be discovered and processed during a researchers analysis on that specific science topic.

This document comments upon the fit of the science drivers for the AstroGrid project and how these fit with the development of the Astrophysical Virtual Observatory (AVO) project (of which AstroGrid is a primary member) and other virtual observatory initiatives. In closing, the future areas of astronomical science that may be focussed on in any future development of the AstroGrid or related projects, are commented upon.

(2.3) Capturing the Science Requirements

The AstroGrid project initially determined to seek out a wide and demanding set of science drivers which might shape a generic virtual observatory. These VO science cases have been written up, to varying degrees of completeness, on the AstroGrid Wiki VO science requirements area.

The following sections outline the primary mechanisms by which the project has attempted to gather it's science requirements both from activities internal to the project, and external engagement with the community.

(2.3.1) Gathering Virtual Observatory Science Drivers

The AstroGrid consortium members generated science cases reflecting the scientific interests of their research groups. Because the consortium contains representatives involved in a wide range of UK astronomy, solar and STP research activity, a broad range of science cases stressing areas such as radio astronomy, solar physics and solar/terrestrial physics resulted. This was a major and early task of the Project.

In the first instance a number of specific use cases were formulated. These were often concerned with how a part of science

problem might be approached, for instance, running a query on a database to locate and show the positions of known QSO's. Other cases are aimed at easing the process of acquiring sufficient data to address a particular problem, e.g. returning the colours of galaxies and their bulges as a function of redshift to study alternative theories of galaxy evolution. The distinction between 'science cases' and more generic 'use cases' rapidly became apparent. Emphasis was placed on capturing and formalising the science cases as listed at http://wiki.astrogrid.org/bin/view/VO/ScienceProblemList.

Further input for the science cases was sought from a number of areas. The project scientist was responsible for assessing current major scientific research strands with a view to identifying those areas likely benefit from the promise of access to the distributed data and processing capabilities to be opened up by a virtual observatory. Areas here included those science areas serviced by large scale multi−wavelength data survey's or those requiring access to large scale computational facilities. Note was made of the similar requirements survey undertaken by the Science Definition Team of the NVO.

The project scientist and other team members have been involved in discussions with their research colleagues in a variety of situations. A series of presentations have been made to representative research departments and at national meetings such as the National Astronomy Meeting and RAS AGM. Importantly the project has engaged with younger researchers in astronomy, with ad−hoc meetings arranged with PhD students and new Post Docs at the IoA, MMSL and other institutes. The focus of these discussions was which elements of AstroGrid would most likely support young researchers. The key area of concern to them was in haviving access to capabilities which would speed the process of mundane data processing and manipulation tasks, thus the concept of assisted workflows appealed.

Engagement with a number of large scale projects has also led to significant scientific input, e.g. with the UKIDSS consortium, the ING Wide Field Survey, the XMM−Survey Science Centre and its role in the XMM Serendipitous Sky Survey, etc.

(2.3.2) The Pilot Programme: Feeding Back Science Requirements

The AstroGrid Phase−A Pilot Programme, as described in the RedBook section Pilot programme report, has also provided a number of inputs into the Science Requirements of the project. In a similar vein, five science topics were fed into the pilot programme as the basis for the pilots as described in Section 7.2 of Pilot programme report.

In a joint AstroGrid/SpaceGRID initiative, the space science research (SSR) community was invited to comment on the requirements of a possible virtual observatory system providing distributed access to data and processing assets. The questionnaire issued is located in full at http://www.spacegrid.rl.ac.uk/spacegrid/SSR_online_q.htm. Summary results of this exercise are reported in the Pilot programme report and the WPA5.5 report. This report does not outline specific scientific drivers. However, it did produce an indication of the priority areas in generic capabilities of interest to the space science community, and these have been input into derivation of the AstroGrid science drivers.

(2.3.3) Presenting the AstroGrid Project: Receiving Input

The project scientist and other AstroGrid staff have been active in giving presentations and seminars about the AstroGrid project throughout Phase A of the project. The list of talks and seminars is given in the Progress against Goals section of this report. This has widened the UK communities appreciation of the possibilities of the project and led to significant input into the generation of new science cases and amendments to pre−existing ones.

A number of general articles have been presented in journals such as PPARC's Frontiers magazine, the RAS's Astronomy and Geophysics magazine etc. These have invited scientific feedback, and have led to a number of comments.

(2.4) Shaping the AstroGrid Drivers from the Virtual Observatory Science Cases

The previous section (2.3) has described the process by which the AstroGrid project captured its science drivers. The complete list is held at http://wiki.astrogrid.org/bin/view/VO/ScienceProblemList. It is anticipated that this selection of science drivers will be continually expanded upon throughout the project lifetime. These drivers will form an important resource for input into partner VO initiatives, especially the AVO. Partner projects such as EGSO and SpaceGRID are also likely to draw from them.These 'VO' science cases, together with those being highlighted by other virtual observatory projects, will be used in shaping the evolution of longer term VO initiatives.

With the collation of the VO science drivers, the project recognised that AstroGrid would not be able to produce a virtual

observatory capable of meeting the demands of all of these cases. Thus a formal and rigorous process was undertaken in order to select a well defined set of science drivers, the AstroGrid Ten, which would be used to shape the AstroGrid deliverables. The project scientist and scientific members of the team analysed the science drivers over a number of review meetings. Based on selection criteria the science problems were distilled to produce the key set of ten drivers.

These drivers were chosen to:

- Represent a cross section of currently topical Astronomy, Solar and Space Science research areas
- functionalities covering a wide spread of technical areas
- Be achievable within the AstroGrid Phase−B project span both in terms of technical complexity of solution, but also in terms of avilability of input science data sets.
- Have a well defined user base who would benefit from capabilities generated by the project
- But at the same time, the tools generated to satisfy the science project would be of use across a wide range of problem areas.

TheAstroGrid Ten drivers though will be under a more formal version control from the beginning of AstroGrid's Phase B.

(2.5) 'The AstroGrid Ten' Science Drivers

For each science driver a typical flow of events has been constructed which decomposes the tasks required to complete that process. Sequence Diagrams have thus been generated for each of the science cases, and for the generic technical use cases (these covering activities such as [[Astrogrid.NegotiateAccessToJobSD][negotiating access to jobs], logon to the system etc. The sum of these tasks represents the components of the system that are required to form the AstroGrid Phase−B product, to be developed within the framework laid down by the project architecture.

It is worth noting that AstroGrid aims to provide tools and capabilities to help the researcher in producing solutions to these science topics. However, AstroGrid will not initself provide the answers, the researcher will be presented with new capabilities to make them more efficient and effective. This will be especially so in the areas of data discovery, transformation of data into information via access to processing facilities, and management of the processing flow of events. AstroGrid will mean that the researcher will be able to devote more time to the understanding of the astrophysics revealed by the results, in other words, more time can be given to the important step of transforming information into knowledge.

The capabilities derived for the AstroGrid Ten will have a usefulness to a wider scientific audience. Any science problem with a similar workflow to one of the AstroGrid Ten will be supported. For instance, searching for AGB candidates would benefit from the system developed to support searches of high redshift Quasars − the difference being one of types of input catalogues, and constraints on the discovery space.

From Science Driver to AstroGrid Product

The AstroGrid Ten science drivers are used to define the scope of the AstroGrid deliverables. Each use case was analysed and decomposed into a work flow, with the tasks required by the science cases being identified.

- ***Generic Use Cases***
  A number of use cases were captured which have generic utility. These cases are those that would be needed in any baseline virtual observatory system, and are largely infrastructural in nature. Use cases in this category include those dealing with security (e.g. Determine Identity), monitoring, job control etc. Generic use cases were captured in a wide sense as generic 'Virtual Observatory' use cases in the VO area of the Wiki − see VO.UseCaseList.

- ***Specific Use Cases***
  Analysis of each science use case revealed that in addition to the need for the generic use cases, each would require more specialised use cases. For instance, the HiZQuasars case requires use cases such as ones to allow the determinations of redshifts of objects in the field.

- ***The AstroGrid Use Case Set***
  The analysis of the Ten science cases revealed the minimum set of use cases that would be required to enable the construction of the capabilities require to meet the demands of these science drivers.

In parallel to the scientific requirements process the architectural shape of AstroGrid was being formulated. The use cases demanded by the science drivers, together with any of those indicated by the outline system architecture thus represent the

reduced set of use cases that will need to be developed by the AstroGrid project. These are listed in the AstroGrid Use Case wiki area at http://wiki.astrogrid.org/bin/view/Astrogrid/UseCases. For a full description of the further process by which the AstroGrid project will construct its products, refer to the RedBook sections, Architecture Overview and Phase B Plan.

(2.5.1) Brown Dwarf Selection

This problem involves aiding the discovery of Brown Dwarfs from large scale survey data sets. Brown dwarfs are intrinsically faint and rare objects, so their detection is not straightforward. It can be done, however, through a combination of selection criteria using colour and proper motion information. Colour selection is the more important, because brown dwarfs populate a well−defined photospheric temperature range (although the coolest brown dwarfs have unusual spectral energy distributions, peaking around

1 micron, due to the absorption of near−infrared continuum flux by water and methane), but proper motion selection can help, too, since any detectable brown dwarfs must be nearby and, so, on average, they will have relatively high proper motions. The use of wide field optical/near−IR survey's to localise Brown Dwarfs is discussed by Basri, 2000, his Figure 7 shows the colour magnitude diagram for low mass Pleiades members.

The key areas of AstroGrid functionality required are:

- access to large area optical and near−IR data sets
- the ability to search for objects in colour−colour space, with objects referenced against model predictions for that colour−colour space
- the ability to cross match samples selected in the colour−colour search with possible multi−epoch data to determine the objects proper motion.

The resulting brown dwarf sample data sets can then be used as input into spectroscopic confirmation programmes, confirming the nature of the objects by means of tests such as the 'Lithium Test' (see Martin et al, 2000).

A comprehensive flow of events is contained in the Brown Dwarf Sequence Diagramme wiki area. For the specific example where Brown Dwarfs are discovered in Galactic Clusters from multi−colour survey data the following flow of events occurs:

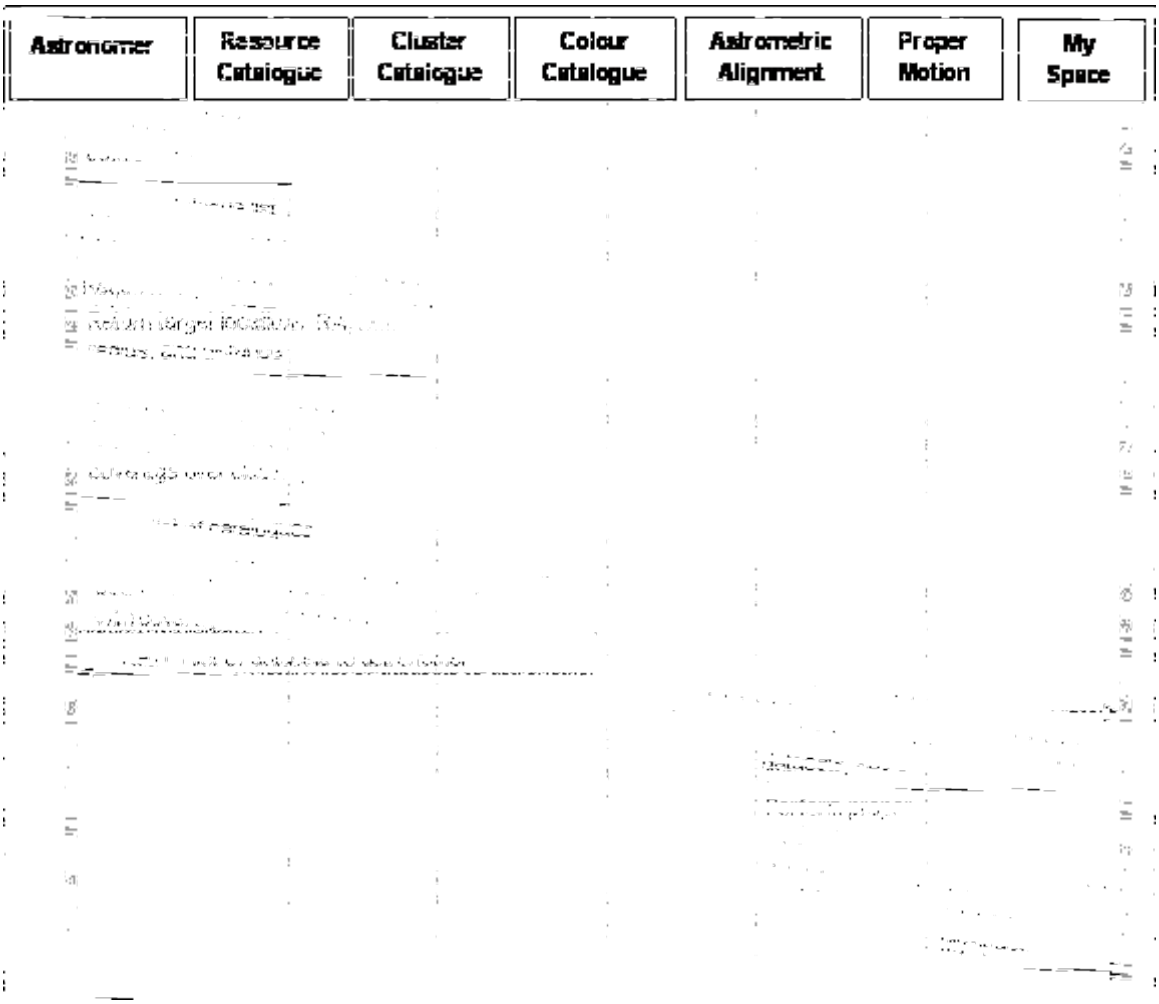
1. The astronomer searches the resource catalogue for catalogues containing Galactic clusters via PerformRegistrySearch.
2. A list of cluster catalogues is returned via MySpaceStoreResults, and the astronomer selects one or more cluster catalogues via SelectCatalogue.
3. Next, the astronomer searches the selected catalogues for cluster locations via PerformCatalogueSearch.
4. A list of locations, defined by right ascension, declination, radius, and distance, is returned via MySpaceStoreResults.
5. The astronomer then returns to the resource catalogue and executes a complex query for catalogues with coverage of I, K, or R wavelengths over each cluster location via ComplexQuery.
6. A list of catalogues with I, K, or R coverage of the cluster location is returned via MySpaceStoreResults.
7. The astronomer selects 1 or more catalogues via SelectCatalogue and searches them for 2 or more datasets covering the cluster location in the same wavelength (either all datasets with I coverage or all with K coverage) via PerformCatalogueSearch.
8. The datasets are stored to MySpace via MySpaceStoreResults.
9. Now the astronomer can prepare the data for the proper motion survey. The datasets are astrometrically aligned using a library function, a web service, or user code via DetermineProgram.
10. Next, the proper motion can be applied to the datasets by user code, a web service, or a library function via DetermineProgram.
11. The program calculates a proper motion vector−point diagram of objects in the dataset. The diagram is stored on MySpace via MySpacePublishDerivedData and returned to the astronomer.

and thus the UseCases required to deliver this science case are:

1. PerformRegistrySearch
2. PerformCatalogueSearch
3. MySpaceStoreResults
    - ♦ NegotiateAccessToResource
        - ◊ AuthenticateIdentity
        - ◊ DetermineAuthority

4. ComplexQuery
    - ♦ NegotiateAccessToJob
5. DetermineProgram
    - ♦ AccessLibraryFunction
        - ◊ NegotiateAccessToResource
            - · AuthenticateIdentity
            - · DetermineAuthority
    - ♦ AccessWebService
        - ◊ NegotiateAccessToResource
            - · AuthenticateIdentity
            - · DetermineAuthority
    - ♦ UploadUserCode
        - ◊ NegotiateAccessToResource
            - · AuthenticateIdentity
            - · DetermineAuthority
        - ◊ MySpaceStoreResults
            - · NegotiateAccessToResource
                - • AuthenticateIdentity
                - • DetermineAuthority
    - ♦ NegotiateAccessToJob
6. MySpacePublishDerivedData
7. SelectCatalogue

This sequence diagram represents a possible flow of events for this problem:



Detailed breakdowns such as these have been performed for each science case in turn, with full details accessible via the links in this document (or via the wiki pages). Further analysis of each sub case is performed to reveal the complete case set. At this

stage class diagrammes, and eventualy software construction is undertaken.

(2.5.2) Discovering Low Surface Brightness Galaxies

Low surface brightness systems are often missed from wide field survey catalogues due to selection effects acting against their discovery. However, it is important to locate and understand the properties of this population as they can contain significant mass (e.g. Impey & Bothun, 1997). A knowledge of the number and mass distribution of low surface brightness galaxies is also vital when comparing theories of galaxy formation and evolution.

The key areas of AstroGrid functionality required are:

- access to image surveys with relevant magnitude and depth
- dual pass algorithms to remove initially bright structures then localise extended low surface brightness features.
- comparison of structures across multiwavelength data sets, e.g. optical from the WFS, infrared from UKIDSS

A comprehensive flow of events is contained in the Low Surface Brightness galaxy sequence diagramme wiki area.

(2.5.3) The Galaxy Environment of Supernovae at Cosmological Distances

Supernovae searches (e.g. Perlmutter et al, 1999) typically programme observations of a set area of sky (the area imaged being dependent on the size of the SN sample desired, for SN samples at lower redshift larger areas of sky are required due to volume effects). The selection of the correct sample of Type Ia's at the imaging search stage is important because confirmation of the SN comes from spectroscopy often obtained on the largest ground based telescopes, such as the VLT, for the higher (z>0.7) redshift SN. Therefore it is important to minimise 'wasted' spectroscopic and followup time on Type II SN.

A problem with current techniques, is that for any candidates discovered there is an uncertainty as to whether or not the candidate is in fact the desired Type Ia SN. Whilst Type Ia SN are typically brighter than Type II core collapse SN, some (~10%) Type II's can contaminate the sample.

A rapid knowledge of the environment in which any SN is discovered can improve the situation. Pre−determination of the galaxy redshifts utilizing photometric means enables an estimate to be made of the candidates distance upon discovery, and thus a better estimation of which type of SN it is. Further, information on the galaxy, for instance it's morphological type, may also aid in rapid classification of the SN. Type Ia's being formed by an explosion resulting from the accretion of matter onto a degenerate star are found in all classes of galaxy. However, Type II's, which result from the catastrophic explosion of a massive star have not been found in early elliptical galaxies.

The key areas of AstroGrid functionality required are:

- search literature and published sources for possible spectroscopic redshifts of galaxies in SN survey fields.
- search archives for spectroscopic data of objects in field > determine redshifts of galaxies in fields using perhaps automated techniques such as developed for the 2dFGRS (see Colless et al, 2001)
- Locate multicolour broadband optical data for the search fields
- Determine photometric redshifts to galaxies in the fields using a variety of techniques (e.g. hyperz, more recently (2002) Z−Peg).
- identify possible galaxy clusters
- cross reference position of newly discovered SN from search. If located in a galaxy for which the redshift is known from one of the above techniques, return an assigned redshift for that SN.
- return morphological information of the galaxy in which the SN candidate is located (if applicable)

A comprehensive flow of events is contained in the Supernova Galaxy Environment sequence diagramme wiki area.

(2.5.4) Object Identification in Deep Field Surveys

The Hubble Deep Field (HDF) is a 'blank' area of sky observed with unprecedented resolution and sensitivity by the HST, revealing about 3000 faint galaxies within a 3 arcmin−square region (also including flanking fields). Fields of up to 40' centred on the HDF have since been imaged at wavelengths from radio to x−ray. In order to better understand the nature of the objects in the HDF, it is vital to be able to correctly cross identify sources seen in various wavelength regimes. This involves effort in aligning the data sets, and searching for significant correlations between sub−sets of properties. For example, it turns out that there are an excess of radio sources (including those too faint to be catalogued) within the error boxes of selected optical

sources in the HDF.

Only recently has the nature of the brightest sub−mm source (HDF850.1) in the HDF−N been unravelled, as described by Dunlop et al, 2002. The key to this discovery was the combination of new deep imagery in the infrared combined with careful astrometric alignment and association techniques to relate the various data sets. Techniques developed by AstroGrid to support further work in this area will be applicable to the data source identifications from the substantial numbers of fields for which deep multiwavelength data sets are becoming available (e.g. HDF, CDF, Subaru/XMM−Newton Deep Survey fields etc.

The key areas of AstroGrid functionality required are:

- Automatic registration and calibration
- Search of all available published data
- Tests for correlations (based on user−supplied criteria) across many catalogues
- Searches of image (or other) data for uncatalogued sources which become significant if found to co−incide with detections at other wavelengths
- Search for sources not detected in optical − thus identify objects such as dust−enshrouded starbursts

A comprehensive flow of events is contained in the Deep Field Surveys sequence diagramme wiki area.

(2.5.5) Localising Galaxy Clusters

Clusters of galaxies can be used to trace distribution of matter in the universe over large scales. Clusters are typically X−ray or optically selected. Many optically selected cluster samples have suffered from various selection effects − such as the use of only one colour data (e.g. Dalton et al, 1992).

New techniques (e.g. Gal et al, 2000) select clusters using multicolour data to localise clusters which are predicted to contain an overabundance of red, early type galaxies. Cluster identification using Optical and Near−IR data uses positional information to select clusters (e.g. Gladders & Yee, 2000)

Cluster distributions can be compared to matter distributions generated by e.g. Lambda CDM models (e.g. Nagamine et al, 2001) or Warm Dark Matter models (e.g. Bode et al, 2001). These models now have sufficient resolution to show dwarf galaxies. Morphologies of the cluster galaxies will be directly compared with predictions from models of galaxy formation (e.g. [#JumpToEke2000][Eke et al, 2000]]).

The key areas of AstroGrid functionality required are:

- select sources marked as galaxies, select only those in a particular locus of the (g−r) vs (i−r) colour space, and then create density maps
- determination of photometric and/or spectroscopic redshifts of the sample cluster galaxies
- comparison with n−body code model outputs: issues include interfacing to large model data sets, vizualisation of model vs real data − e.g. matter vs clusters at ranges of redshift, statistical correlations etc.

A comprehensive flow of events is contained in the Galaxy Clustering sequence diagramme wiki area.

(2.5.6) Discovering High Redshift Quasars

Quasars at high redshifts will provide vital clues to the processes involved in the formation of the first bound objects. Near−IR survey data from UKIRT's WFCAM (via the UKIDSS survey) and later VISTA survey programmes will enable many quasars in the redshift range $5.5 < z < 7$ to be discovered. This will enable a number of principal scientific goals to be met. A key primary rational is that quasars at the highest redshifts may enable the investigation of the epoch of reionisation of the Universe. Such an effect is already being reported for the re−ionisation of He II via studies of quasars between $3 < z < 4$ (see Theuns et al, 2002). Higher redshift HiZQuasars would probe the neutral Inter Galactic Medium at this at this epoch.

The key areas of AstroGrid functionality required are:

- Access to large scale optical and near−IR survey's, especially those in the IR to be provided by UKIDSS
- Selection of candidate samples in colour−colour space via comparison with model predictions (c.f. optical techniques as described for the SDSS survey by Richards et al, 2002.

A comprehensive flow of events is contained in the High Z Quasars sequence diagramme wiki area.

(2.5.7) The Solar−Stellar Flare Comparison

Flare stars are generally low temperature red, M−class, dwarf stars. Our Sun also experiences flares, and these are related in some poorly understood fashion to Coronal Mass Ejections. Schaefer et al, 2000 have noted that a number of nearby solar type (F−G) stars have undergone super flare events, with the energy in the flares >100 times the most energetic measured from our Sun. The census of stars with 'superflares' is incomplete due to the difficulty in collating the various data sources for nearby flaring stars. This case will aid in provide a full sample of superflare stars. Investigation of the linkage of CME's to flares could be studied by investigating eveidence for CME's in the sample of superflare stars. One technique is to discover evidence of absorption in for instance Si UV lines during a CME event for those flare stars in binary systems with a hot white dwarf (see e.g Bond et al, 2001).

The key areas of AstroGrid functionality required are:

- Localisation of flare stars from the literature
- Lightcurve generation for flare stars from published photometry − estimation of energy in the flare
- Determine availability of high res UV spectra for stars identified as having super flares.

A comprehensive flow of events is contained in the Solar Stellar Flare Comparison sequence diagramme wiki area.

(2.5.8) Deciphering Solar Coronal Waves

There is a current debate as to whether large scale coronal waves and chromospheric waves (Moreton, 1961) are related. Moreton waves were found to propagate at large distances from a solar flare site with velocities ranging from a few hundred to several thousand km/s. Due to the high speeds observed, it was assumed that the origin of the Moreton wave was in the corona and not in the chromosphere. Coronal waves were first observed by the EUV Imaging Telescope (EIT) onboard the Solar and Heliospheric Observatory (SOHO) spacecraft (Thompson et. al, 1999). They appeared in difference images as a bright front with a following dimmed or depleted region of the corona with propagation speeds of a few hundred km/s.

A key goal is to determine whether coronal waves are MHD fast mode waves occuring from a solar flare site, or if they are a global coronal mass ejection lifting off the surface of the disk. This will be achieved by searching for flares, preferably occuring on disk centre, isolating the times, and then finding the necessary datasets (EUV and Halpha spectra, EUV/SXR imaging).

The key areas of AstroGrid functionality required are:

- Localisation of coronal wave via image subtraction technique
- Discovery of supporting multi−wavelength observational data sets covering correct spatial, temporal space (e.g. Zhang et al, 2001).
- Visualistion of flare, wave datasets

A comprehensive flow of events is contained in the Solar Coronal Waves sequence diagramme wiki area.

(2.5.9) Linking Solar and STP Events

Solar models are currently used to predict STP events as impacting on the local solar−earth space environment. This information can be used to advise the telecommunications and power industries of geomagnetic disturbances (e.g. via the US's Space Environment Center'sSpace Weather page at http://www.sel.noaa.gov/today.html. Solar events such as flares, CME's, and the progression of the solar cycle can cause electromagnetic disturbances in the Earth's magnetosphere. Satellites, radio and television broadcasts, and mobile telephones all experience service interuptions during periods of high solar activity.

Several existing models take solar activity parameters (i.e., time, duration, location, and intensity of events) as input and predict the resulting STP events that will occur in the Earth's magnetosphere, e.g. Geomagnetic Storms. The solar models, solar datasets used as input, and STP datasets used to verify output predictions, are not, however, accessible from a single interface. Models include the Relativistic Electron Forecast and Wang Sheeley models from NOAA.

A key goal is to provide a selection of these models as Astrogrid web services. An individual model can be tested with several solar datasets to compare modelled predictions with actual STP datasets during different stages of solar activity . Also, one

solar dataset may be chosen as an input to several models in order to ascertain which model mostly closely predicts STP events during a given time period.

The key areas of AstroGrid functionality required are:

- Provide unified point of web service access to distributed models
- Capture relevant STP data to compare with predictions from models

A comprehensive flow of events is contained in the Solar STP Event Coincidence sequence diagramme wiki area.

(2.5.10) Geomagnetic Storms and their Impact on the Magnetosphere

Study of the morphology of the tail of the Earth's magnetosphere during the onset of geomagnetic storms is important in understanding the processes involved, and the impact of the storm on the magnetosphere. Geomagnetic storms can influence the performance of satellite systems such as the GPS (e.g. Skone & de Jong, 2000) and also in severe cases impact power transimission (see e.g.http://www.mpelectric.com/storms/). The observational data can be compared against models of the magnetosphere (e.g. Raeder at al, 2001).

The key areas of AstroGrid functionality required are:

- Determine temporal location of storm
- Retrieve list of in−situ satellites with suitable instrumentation located in the magnetosphere during the relevant time periods.
- Conversion of the position data to a defined coordinate system and the magnetic field data to specific units. The appropriate coordinate system will depend on the application.

A comprehensive flow of events is contained in the Magnetic Storm Onset sequence diagramme wiki area.

(2.6) Providing Functionality as Indicated by the AstroGrid Ten.

The ten science drivers itemised in section 2.5 above cover a number of representative science topics. Additionally they require the provision of a range functionalities to be provided by AstroGrid.

(2.6.1) The Required Capabilities of AstroGrid

As described above, the decomposition of each of the science use cases, together with a consideration of the minimum generic infrastructure required by the project architecture, leads to the required set of system use cases. Analysis of these leads to the derivation of the requirements for the software development (as discussed in the Phase−B Plan). These main functionality areas which will be provided by the AstroGrid project are described in the ArchitectureOverview section of the RedBook.

Science Case	***Science Cat***	***Functionality Area***							
		Advanced Algorithms	***Astronomical Query Language***	***Compute Intensive***	***Database Access***	***Data Mining***	***IAA***	***AstroOntology and WorkFlow***	***Resource Location***
Brown Dwarfs	A−S	y	y	y	y	Y	y	Y	y
LSB Galaxies	A−G	y	y	Y			y		y
SN Environment	A−G	y	y	Y	Y		y		Y
Deep Fields	A−C	Y	Y	y	Y		y	y	y
Galaxy Clusters	A−C	Y	y	Y	Y	Y	Y		
Hi−z QSO's	A−C	Y	y	y	y	Y	Y		
Solar/Stellar Flares	A−S/S	y	Y	y		Y	y	Y	Y
Coronal	S	y	y	y	y	y	y		Y

Waves									
STP/Solar Events	S−STP	y	y		Y	y	y	Y	Y
Magnetic Storms	STP		Y		Y	Y	y	Y	Y

Science Cat = Science category: A−S (Astronomy: Stellar), A−G (Astronomy: Galaxy), A−C (Astronomy: Cosmology), S (Solar), STP (Solar−Terrestrial Physics). A matrix element marked 'Y' indicates that this area will be important, whilst 'y' indicates a lesser degree of importance.

The analysis of the processes involved in meeting the requirements set by the science drivers shows that a number of areas are highlighted, with the relative importance of these areas varying from case to case.

For instance, the MagneticStormOnset case does not deal with large data volumes. Rather the problem is one of discovering data from a number of dispersed and hetereogeneous data sets, where data streams need to be isolated according to the spatial and temporal position of the in−situ detectors. Non standard data sets need to be addressed, as the STP data is often of the 'pen−plotter' variety, many differing variables being monitored by on−flight detectors measuring in−situ flows (e.g. the magnetic field at a point in space).

The Galaxy Clusters use case involves the manipulation of large multi−TB scale multi−wavelength data sets, and the ability to run sophisticated alogorithms on the pixel and catalogue data. The extension to this case additionally will require tools to directly compare the observational cluster data with outputs from theoretical models.

(2.6.2) Specific Targets for Data Manipulation Tasks

The AstroGrid Ten set demands upon the capabilities that must be provided in order to service these science cases. The following table gives lists example data manipulation problems which researchers would wish to complete in a certain time by using Phase−B AstroGrid deliverables.

Science Case	*Science Cat*	*Data Manipulations Problem*	*Completion Time*
Brown Dwarfs	A−S	Perform complex query on 1000 deg^2 of multi−colour (e.g. 6 band, Sloan & UKIDSS) to return 1000+ Brown Dwarf candidate sample, plus postage cutouts of pixel data (~20Mb pixel data per cutout). For each candidate return propoer motion estimate via astrometry of multi−temporal data sets.	One hour
LSB Galaxies	A−G	Select candidates from IR data sets (e.g. 1000 sq deg of UKIDSS Large Area Survey data), cross match with Sloan optical, perform background analysis, determine luminosity calculations	One hour
SN Environment	A−G	For search fields, classify galaxies, determine redshifts (literature search, spectral or photmetric methods)	30 mins
Deep Fields	A−C	Automatic registration and calibration of multicolour data sets (~10 wavebands, few x 100Mb/band), object identification, cross identification and association	One hour
Galaxy Clusters	A−C	Generate cluster maps from smoothing functions applied to red early type galaxies identified in catalogue data (e.g. VST, UKIDSS). Assign redshifts to clusters via photometric techniques, compare sanple with model outputs as function of z	Four hours
Hi−z QSO's	A−C	Perform complex query on 2000 deg^2 of multi−colour (e.g. 6 band, Sloan & UKIDSS) to return 1000+ qso candidate sample, plus postage cutouts of pixel data (~20Mb pixel data per cutout). Ability to cross reference the sample qso set against X−ray, Radio, NED, catalogue lists.	One hour
Solar/Stellar Flares	A−S/S	Recover flare photometry for all flare stars in 25pc of the solar neighbourhood, generate energy in the flare estimates, localise high resolution UV spectra for candidate sample	Two hours
Coronal Waves	S	Localise previous three months of solar imaging data, process image to determine where and when coronal waves occured. Retrieve all supporting observational data	Two hours

		with coverage for these waves	
STP/Solar Events	S−STP		
Magnetic Storms	STP	Determine period of geomagnetic storm from Dst index, located satellite data, return in geocentric coordinate space	30 mins

(2.6.3) The AstroGrid System: Core Functions and User Add−Ons

The evolving Phase−B plans detail AstroGrid's roadmap for Phase−B of the project, to deliver software which will enable the creation of a working, grid−enabled Virtual Observatory (VO) based around key UK astronomical data centres. The key areas of activity will be concentrated on the following areas:

- ***Component Services***
  This is the main activity of the build phase and concerns the building of the web and grid service−based components from which the VO will be constructed.
- ***Library Services***
  These are also service−based components but will be wrappers or interfaces to existing tools or libraries.
- ***Portal & Client Programs***
  These are stand−alone programs with which the user will interact; they will make use of the service components defined above.
- ***Demonstrations***
  These activities are technology trials required to test whether certain technologies work the way we require, or to check how areas of research can be exploited.
- ***Research***
  This activity is, as it says, research into areas which are still insufficiently understood to be incorporated into the system.
- ***Test Implementations***
  This involves the implementation of working versions of the software in one or more data centres to test the feasibility of the software being developed.

The end user will interact with the system through the ***AstroGrid Portal & Client Programs*** and be able to manipulate key data sets held both by AstroGrid, and others accessible via compatible interfaces (e.g. those made available by the AVO, NVO). Manipulations of these data sets will be possible by the application of the key AstroGrid library of services and tools (e.g. image classifiers, database manipulations tools etc.).

Running User Provided Algorithms

However, it is recognised that the work flow of many astronomical tasks indicates that the individual researcher will want to apply some forms of algorithmic processing to their data set that is not provided for in the core AstroGrid offer. AstroGrid will devote effort into making available existing software packages such as starlink, iraf, aips++, solarsoft within the system. AstroGrid's Phase−B architecture will also support a limited ability for users to integrate the use of their specific processing algorithms into the system. This capability is foreseen to be implemented in the late 2004 iteration of the project. For those cases where the individual user code can not be uploaded into the system, the user will of course be able to access their data products held by the AstroGrid system and process these data locally.

(2.6.4) Data Sources for AstroGrid

As initially stated in the original project submission for AstroGrid, the scientific focus is one with an aim to exploit key data or resources held by the UK data centres. For instance, the Brown Dwarf Selection case would mine SuperCOSMOS, UKIDSS and WFS data as primary data sources. These UK sourced data sets held by the AstroGrid consortium data centres are listed in full at http://wiki.astrogrid.org/bin/view/Astrogrid/DataCentres. They represent a hetereogeneous set of data, covering a wide range of wavelegths from Radio (e.g. Merlin at Jodrell) to X−Ray (e.g. XMM−Newton as Leicester). Limited access to model data will be provided, initial conversations are begining in this area. AstroGrid will additionally enable access to many data sets available held elsewhere. This will include data sets held by AstroGrids partners in the AVO consortium, these including data held by ESO, and catalogue data accessible through the CDS−Strasbourg.

(2.7) The AstroGrid Science Input into Partner Virtual Observatory Projects

(2.7.1) AstroGrid and the Astrophysical Virtual Observatory

AstroGrid has adopted an approach based on suitable scoping of a widely drawn set of virtual observatory science drivers. AstroGrid is also part of the European Astrophysical Virtual Observatory project. The AVO is currently under taking it's three year Phase−A study which will lead to a fully fledged proposal to develop a facility class Euro−VO.

A key area of the AVO is it's Science work area. Input to the Science WA is provided by the Science Working Group, membership drawn from the European astronomical community.

At it's second meeting a sub group of the SWG, chaired by the AstroGrid project scientist, was formed to derive the science requirements for the AVO's initial science/technology January 2003 demonstrator product. Development of the AVO demonstrator is now progressing, with AstroGrid taking a lead role in the definition and production of the web services aspects of the demo.Specifically AstroGrid is working to make available an image extractor (SExtractor, see Bertin & Arnouts, 1996) as a web service which will enable on−the fly user defined cut−outs and re−extractions of distributed GOODS data sets.This work is described at http://wiki.astrogrid.org/bin/view/Astrogrid/AVODemo.

(2.7.2) AstroGrid and the National Virtual Observatory

Because AstroGrid has consortium partners involved in Astronomy, Solar AstroPhysics and STP, its scientific remit is somewhat more extensive than that of the NVO,as outlined in the Science Definition Teams report dated April 2002. However, AstroGrid is more constrained in its shorter three year project timescale compared to the 5 year NVO project span. Thus the AstroGrid project has based its approach on supporting rather specific programmes to ensures that the project delivers a functional, although limited in scope, virtual observatory for the UK. AstroGrid is in regular contact with the NVO project, both formally through joint membership of the International Virtual Observatory Alliance, and informally through cross attandance at each others project meetings. The AstroGrid science drivers show a limited degree of overlap with the current NVO set of driver, for instance in the area of offering enhanced tools to analyse the low surface brightness universe. The possibility of the production of a diverse set of independently developed tools in this area is seen as being of benefit to the community. In general though, the AstroGrid science drivers are different to those of the NVO.

(2.7.3) AstroGrid and the International Virtual Observatory Alliance

The International Virtual Observatory Alliance (IVOA) is an alliance of virtual observatory initiatives. It's aim is to provide a forum in which the common elements, required to ensure that the systems developed by the various partners are interoperable with one another, can be identified and standards agreed upon. Most of these common elements have to do with standards for data and interfaces. Other common or shared elements may be in the form of software packages, source code libraries, and development tools. Some others have to do with issues of policy, funding and securing international support at governmental levels. The first significant milestone of the IVOA was the agreement and release of the VOTable interoperability standard.

Astrogrid is a founder/member of the IVOA, which is currently composed of representatives from the major (i.e. AstroGrid, AVO, the US National Virtual Observatory (NVO)) and all other major and currently funded virtual observatory initiatives (i.e. eAstronomy Australia, Canadian Virtual Observatory, German Virtual Observatory, Russian Virtual Observatory, Virtual Observatory India). The Chair of the IVOA is Bob Hanisch (NVO), Deputy Chair: Peter Quinn (AVO), Technical Chair: Roy Williams (NVO), and Secretary: Nic Walton (AstroGrid).

In this international context, AstroGrid is stressing the importance of science drivers in shaping the development of the global VO initives. The importance of these concepts have been agreed upon, and recognised in the concept of the IVOA supporting a roadmap of international development where demonstrations of science driven capability are featured.

(2.8) The Evolutionary Path

At the completion of its Phase−B, AstroGrid aims to provide a fully functional Virtual Observatory capability, with a specific focus on meeting the scientific demands of its UK user community. The AstroGrid product is defined in scope by the science drivers listed in this document, and by the resulting minimum architecture needed to deliver the capabilities demanded by these drivers. The implementation plan by which the AstroGrid product will be delivered is described in the Phase−B plan at http://wiki.astrogrid.org/bin/view/Astrogrid/RbPhaseBPlan.

The AstroGrid framework is being constructed in a manner that will ensure that it is both durable, but capable of further expansion to offer increased capabilities in the future. This philosophy is in line with AstroGrid Vision, where the focus is on producing an organic system that supports and facilitates scientific endeavour, rather than actually 'doing' the science.

Future UK virtual observatory initiatives would build on the AstroGrid product. It is clear that the AstroGrid system would provide the UK with a significant entry point into the planned Euro−VO initiative, whereby a facility class european virtual observatory will be created.

AstroGrid is providing capabilities in a number of areas, for examples:

- Creating the multi−space digital sky, whereby seamless access and manipulation is enabled to diverse data sets covering the sky
- Allowing seamless integration of model and observational data sets
- Providing discovery and access to neccessary compute and data storage assets

However, many exciting and challenging issues will need to be addressed in future virtual observatory initiatives, including:

- Allow the creation of 'topic−specific' workspaces, giving access to all data and tools relevant to a certain astrophysical problem
- Facilitate the creation of aided work−flows, whereby the use is able to construct their personalised data pipeline, using VO components, and in a manner where the system provides sophisticated guidance.
- Provide a means whereby the outputs of data manipulations can be automatically fed back into the operations of telescopes, both for real−time and ordinary observational programmes
- Increase the offer of powerful visualisation capabilities, especially in the domain of multi−dimensional visualisation. via the application of technical advances in immersive computing.
- Ultimately provide for a dynamic, self accreting digital sky, whereby all global astronomical observational endeavour is captured, tagged for quality and ingressed, for future manipulation and analysis by the global community.

AstroGrid, or its UK successors, will play a key role in providing solutions to these issues. Concerns as to the financial aspects of the creation of a facility class virtual operation will be discussed outside of this report.

(2.9) Closing Remarks

AstroGrid is a modest three year project which has analysed a wide range of possible science drivers for the creation of a virtual observatory. It has defined a key set of drivers − the AstroGrid Ten − and is using these as a basis upon which it determines the capabilities that it will produce in its Phase−B. It is noted that the science drivers will require the construction of a sophisticated system capable of providing access to significant hetereogeneous pera−byte scale data sets located in the UK and elsewhere. Tools to discover and manipulate these data will significantly aid the research community. In particular, AstroGrid will increase both the efficniency and effectiveness of the UK astronomer, and enable them to devote more time to the vital task of understanding the physical processes at work as revealed by the results from their data manipulations. This will undoubtedly lead to significant advances in the scientific productivity of the UK astronomical community.

This science summary closes by noting that with the completion of AstroGrid's Phase−B, the UK will be well positioned for a leadership role in future larger scale European virtual observatory initiatives.

(2.10) References

(Astronomy and Geophysics) The Royal Astronomical Societies in house journal − see A&G at Blackwell

(AstroVirtel):http://www.stecf.org/astrovirtel/ and accepted proposals at http://archive.eso.org/wdb/wdb/vo/avt_prop/query

(AVO) The Astrophysical Virtual Observatory − a three year EC funded programme charged with mapping out the structure of a facility class virtual observatory for Europe. See http://www.eso.org/avo

Basri, G, 2000 ARA&A, 38, 485, 'Observations of Brown Dwarfs'

Bertin and Arnouts, 1996, A&AS, 117, 393, 'SExtractor: Software for source extraction'

Bode et al, 2001, ApJ, 556, 93, 'Halo Formation in Warm Dark Matter Models'

Bond et al, 2001, ApJ, 560, 919, 'Detection of Coronal Mass Ejections in V471 Tauri with the Hubble Space Telescope'

Dalton et al, 1992, ApJ, 390, 1, 'Spatial correlations in a redshift survey of APM galaxy clusters'

Dunlop et al, 2002, MNRAS, in press, 'Discovery of the host galaxy of HDF850.1, the brightest sub−mm source in the Hubble Deep Field'

(EGSO) European Grid of Solar Observatories − an EU IST funded programme. See http://www.mssl.ucl.ac.uk/grid/egso/

Eke et al, 2000, MNRAS, 315, 18, 'The cosmological dependence of galactic specific angular momenta'

(Euro−VO) The current working title for the European Virtual Observatory initiative (shortly at http://www.euro−vo.org). AstroGrid is providing vital scientific and technical input into the development of this future programme.

(Frontiers) PPARC's in house magazine. A number of articles describing aspects of the AstroGrid project and Virtual Observatory initiatives have been published: http://www.pparc.ac.uk/frontiers

Gal et al, 2000, AJ, 119, 12, 'The Northern Sky Optical Cluster Survey. I. Detection of Galaxy Clusters in DPOSS'

Gladders & Yee, 2000, AJ, 120, 2148, 'A New Method For Galaxy Cluster Detection. I. The Algorithm'

The Great Observatories Origins Deep Survey (GOODS) is a public, multiwavelength survey that will cover two 150 arcmin2 fields. These fields are centered around the HDF−N (Hubble Deep Field North) and the CDF−S (Chandra Deep Field South): see http://www.eso.org/goods.

Impey & Bothun, 1997, ARA&A, 35, 267, 'Low Surface Brightness Galaxies'

The International Virtual Observatory Alliance. It's home page will shortly be found at http://www.ivoa.net. The IVOA Mission and Roadmap is located at http://wiki.astrogrid.org/bin/view/IVOA/RoadMap[http://wiki.astrogrid.org/bin/view/IVOA/RoadMap

Martin et al, 2000, 543, 299, 'Membership and Multiplicity among Very Low Mass Stars and Brown Dwarfs in the Pleiades Cluster'

Nagamine et al, 2001, ApJ, 558, 497, 'Star Formation History and Stellar Metallicity Distribution in a Cold Dark Matter Universe'

(NAM2002) National Astronomy Meeting 2002 in Bristol, see http://www.star.bris.ac.uk/nam

(NVO) The US National Virtual Observatory project: http://www.us−vo.org

Perlmutter et al, 1999, ApJ, 517, 565, 'Measurements of Omega and Lambda from 42 High−Redshift Supernovae'

Raeder et al, 2001, Solar Physics, 204, 323, 'Global Simulation of Magnetospheric Space Weather Effects of the Bastille Day Storm'

(RAS AGM) Royal Astronomical Society AGM 2002 at http://www.ras.org.uk/

"Relativistic Electron Forecast Model", USAF and NOAA Space Environment Center, http://www.sec.noaa.gov/refm/

Richards et al, 2002, AJ in press, 'Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample'

Schaefer et al, 2000, ApJ, 529, 1026, 'Superflares on Ordinary Solar−Type Stars'

(SDT) National Virtual Observatory Science Definition Team (SDT) − final report at http://nvosdt.org/sdt−final.pdf

Skone & de Jong, 2000, Earth Planets Science, 52, 1067, 'The impact of geomagnetic substorms on GPS receiver performance'

(SpaceGRID) An ESA funded programme to investigate possible uses of Grid technology in supporting the scientific and technical functions of ESA: http://spacegrid.esa.int

SuperCOSMOS Sky Survey: http://www−wfau.roe.ac.uk/sss/

Theuns et al, 2002, ApJ, 574, 111, 'Detection of He II Reionization in the Sloan Digital Sky Survey Quasar Sample'

Thompson et al, 1999, ApJ, 517, 151, 'SOHO/EIT Observations of the 1997 April 7 Coronal Transient: Possible Evidence of Coronal Moreton Waves'

(UKIDSS) UKIRT Infrared Deep Sky Survey

"Wang Sheeley Model", USAF and NOAA Space Environment Center, http://www.sec.noaa.gov/ws/

(WFS) The Isaac Newton Group's Wide Field Survey programme: http://www.ast.cam.ac.uk/~wfcsur/index.php

(XMM−SSC) XMM−Newton Survey Science Centre: http://xmmssc−www.star.le.ac.uk/

Zhang et al, 2001, ApJ, 559, 452, 'On the Temporal Relationship between Coronal Mass Ejections and Flares'

(3) Architecture Overview

(3.1) Introduction

This document will present a brief overview of the AstroGrid ***Architecture***. A system Architecture is a high level description, in formal language (we use the ***Unified Modelling Language − UML*** (1)), of the system to be built. It focusses on the ***key decisions***. These are decisions about structure, components, technology etc which are key to the success of the project. From the architecture will be derived the design documents for all components of the ***Virtual Observatory*** (VO) as well as the plans and milestones for the ***build*** phase of the project.

The Architecture consists of a number of documents stored on the AstroGrid ***wiki*** (2), many linked from the *ArchitectureDocs* (3) page (most of the references below will be hyperlinks to wiki pages). Given the interactive nature of the wiki, it will be obvious that the Architecture is no static, monolithic document. It reflects the constantly evolving nature of the project.

The architecture−related documents include discussion papers, technology assessments, reports etc. From these, we have developed more formal documents, including use cases, technology choices, analysis and design models, etc.

Although the Architecture is not complete, enough is now known about the requirements for a VO, and sufficient high level design has been carried out to enable estimates for funding and personnel to be generated. Detailed estimates are provided in another part of this RedBook. Here we show milestones in the development of AstroGrid components over the two years, 2003/4, to enable project progress to be tracked.

The rest of this document consists of:

- ***Approach***: an outline of the approach taken to develop the architecture
- ***Use Cases***: our approach is driven by *use cases* which specify the system−oriented content of the AstroGrid VO
- ***Conceptual Model***: a whole−system model which enabled development of science−based use cases
- ***Services Model***: a component−based model of the VO
- ***Technology Demonstrations***: an overview of several subprojects which tested aspects of new technology
- ***Technology Choices***: technologies which have been chosen for the project

(3.2) Approach

Early in the project, it was decided (following a recommendation by the Project Manager) that we should follow the ***Unified Process*** methodology (4). In fact, AstroGrid has contributed its own variant of the Unified Process (***UPeSc*** (5)).

The UP is both iterative and incremental. From general requirements, use cases are documented. This allows the creation of an outline model of the architecture: including subsystems, classes and components. The use cases are incorporated into the model, documented as ***Sequence Diagrams***, (7) and this leads to changes in the architecture as well as the use cases, and so the model changes and evolves.

Within the AstroGrid project, the Project Scientist, together with other members of the project and outside astronomers, contributed a large number of ***Science Problems*** (6), which it was considered would be the types of science which a VO will enable. These are covered in more detail in another part of the RedBook. This was our main addition to the Unified Process. Ten of these science problems were chosen as those which the AstroGrid VO should enable astronomers to tackle (8).

In order to visualise how this science might be conducted, a ***Conceptual Model*** (see below) of a Virtual Observatory was developed. This model documented the concepts important to the VO ***domain*** (topic or subject area) and how they were related. More importantly, it allowed the creation of sequence diagrams, for key science problems. This was critical to understanding how a VO should operate. Eventually, enough was understood about the type of VO that AstroGrid would build, and the team moved to construct a more realistic model.

We were considering the idea of adopting a web service−based approach when the Globus team announced plans for version 3 of its toolkit, Globus OGSA (9), in which the grid would be enabled using web−service based components (hence ***grid services***). We took the decision to support this move and committed the project to using this new approach. We will, however, err on the side of caution and will ensure that the services within the AstroGrid system can work independently of OGSA or with some other commercial grid implementation if such is developed before the project completes.

The next generation of model for the AstroGrid VO therefore defined components as services and component invocation as messages between those components: the ***Services Model*** (see below).

(3.3) Use Cases

The Unified Process describes itself in three key phrases (19) as:

- Use Case Driven
- Architecture−Centric
- Iterative and Incremental

The whole of this document deals with the architecture and another document in the Phase A Reports, ***Phase B Plan***, will deal with how we will implement the iterative feature of the UP. In this section we will describe our approach to ***Use Cases***.

A use case is a scenario in which a user (or an agent or other piece of software) initiates an action which leads to some benefit to the user: eg a result is returned or the software is put into a state which will enable another action to take place. In parallel with the definition of science problems, we also defined use cases which would resolve the science problems. In general, a science problem involved the execution of several use cases, and most use cases were applicable to several science problems (20).

As an example:

- AuthenticateIdentity:
  in which a ***gatekeeper*** checks the identity of a user (or agent) certificate and verifies whether it is trusted;
- MySpaceStoreResults:
  in which an astronomer is presented with the option of storing a dataset resulting from a query in the MySpace area;
- UploadUserCode:
  in which a user uploads their own code to a computer on the grid so as to run a specific analysis on data held there.

Over the next three months, and into 2003 for the later components of the architecture, we will continue to write use cases and realise them as sequence diagrams.

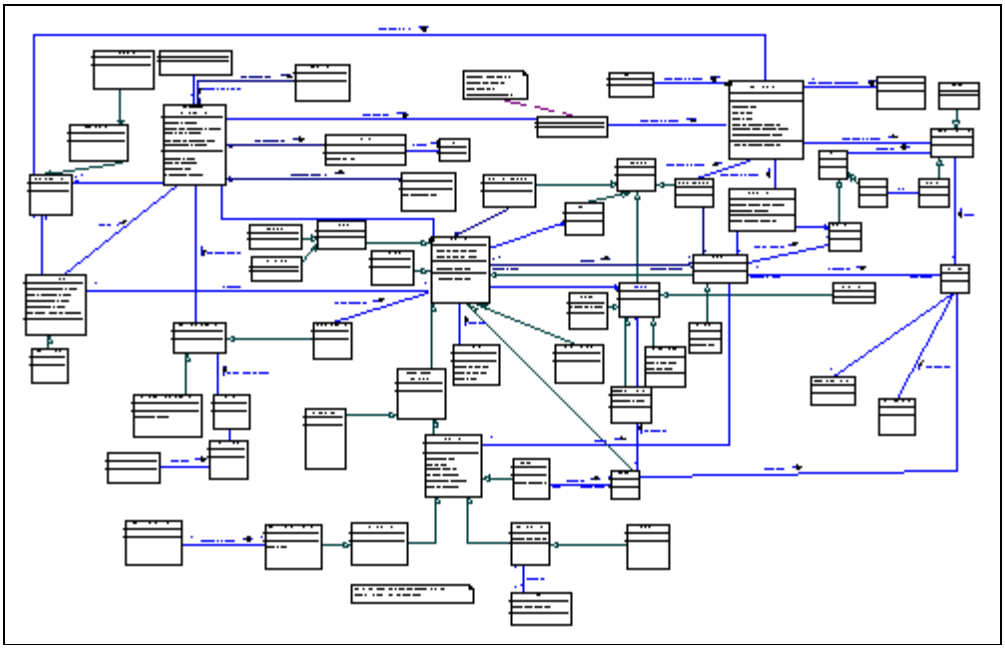
As well as assisting in defining the requirements of the system, use cases have two other important functions:

- ***iteration funcitonality***:
  Before beginning a period of building software, a number of use cases are chosen as those which will be realised during that period. Component software will then be developed or enhanced so that the use cases can be undertaken.
- ***test cases***:
  Each use case is also a test case. After building software, the team, and users, will check the software to ensure that the use cases are correctly realised.

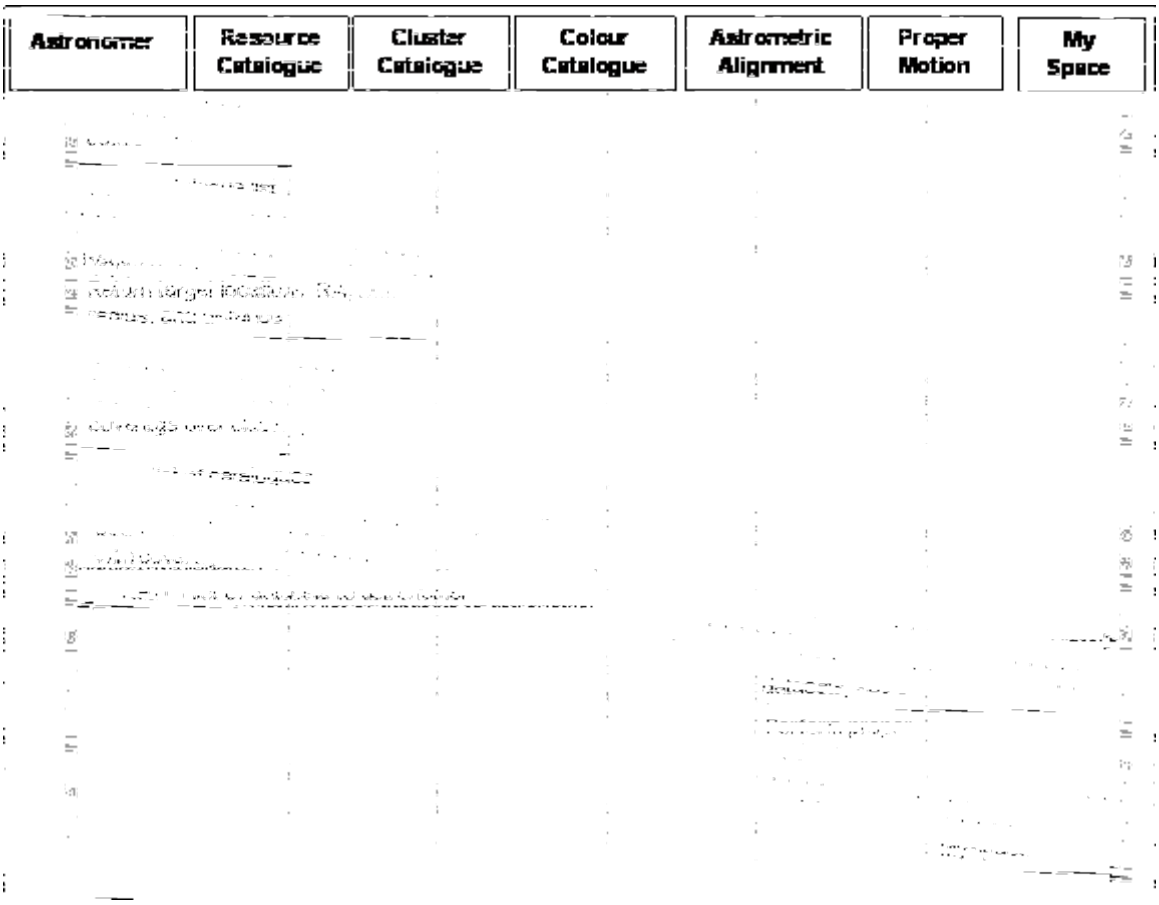
(3.4) Conceptual Model

The ***Conceptual Model*** (10) is a ***whole−system*** model, ie it looks at the VO as if it were a single system. This is useful in the early stages of a project to allow analysts to model the dynamic behaviour of a system without worrying about the separation of objects into components. It is also referred to as a ***Domain Model***. (***Note:*** one unexpected benefit of the conceptual model was that the concepts listed provided a good starting point for the Ontology Demonstration − see later.)

The domain model is too large to show here. A reduced picture will show the size and scope of this model (if you are viewing this document online, click the picture to view the full−size model):



Modelling the system concepts as classes allowed us to model the dynamic behaviour of the science problems in sequence diagrams (11). This proved a key endeavour as it allowed both the development of system use cases and the further elaboration of the domain model. One example of a sequence diagram is shown here:



The conceptual model served as the basis for developing the services model. The concepts − those that were key to the architecture − were partitioned into components (12) which would be delivered as web/grid services.

(3.5) Services Model

Key to any modelling enterprise is the creation of models which look at the system from different viewpoints. The ***Services***

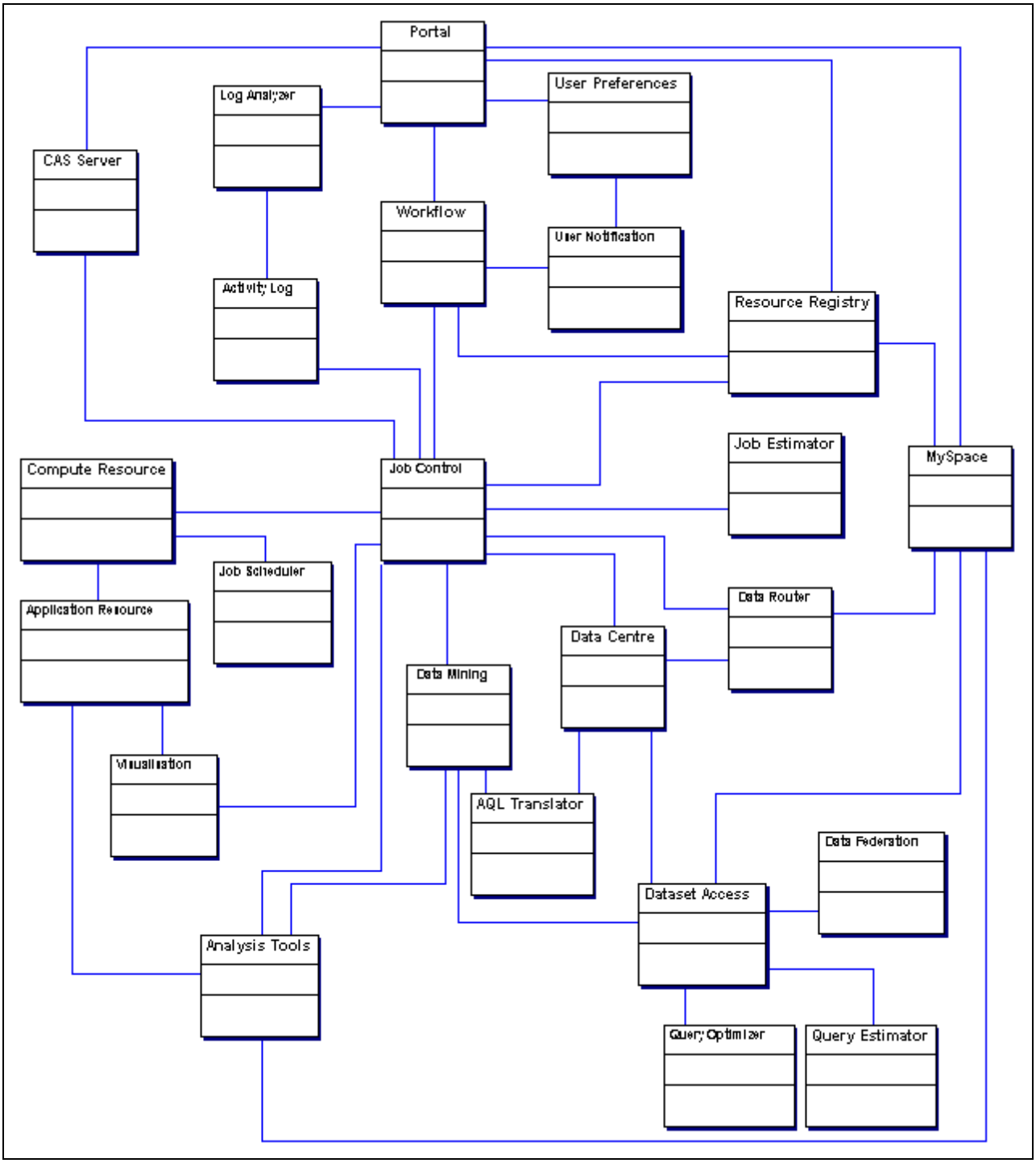
Model (13) starts from a component view of the system and then looks at the interactions required between those components in order to deliver the required functionality.

The next step of this modelling workflow is to take the sequence diagrams developed under the conceptual model and re−engineer them using the component services. This will determine the properties and methods that each service needs to implement. Detailed design for each component will not be done until the build phase of the project.

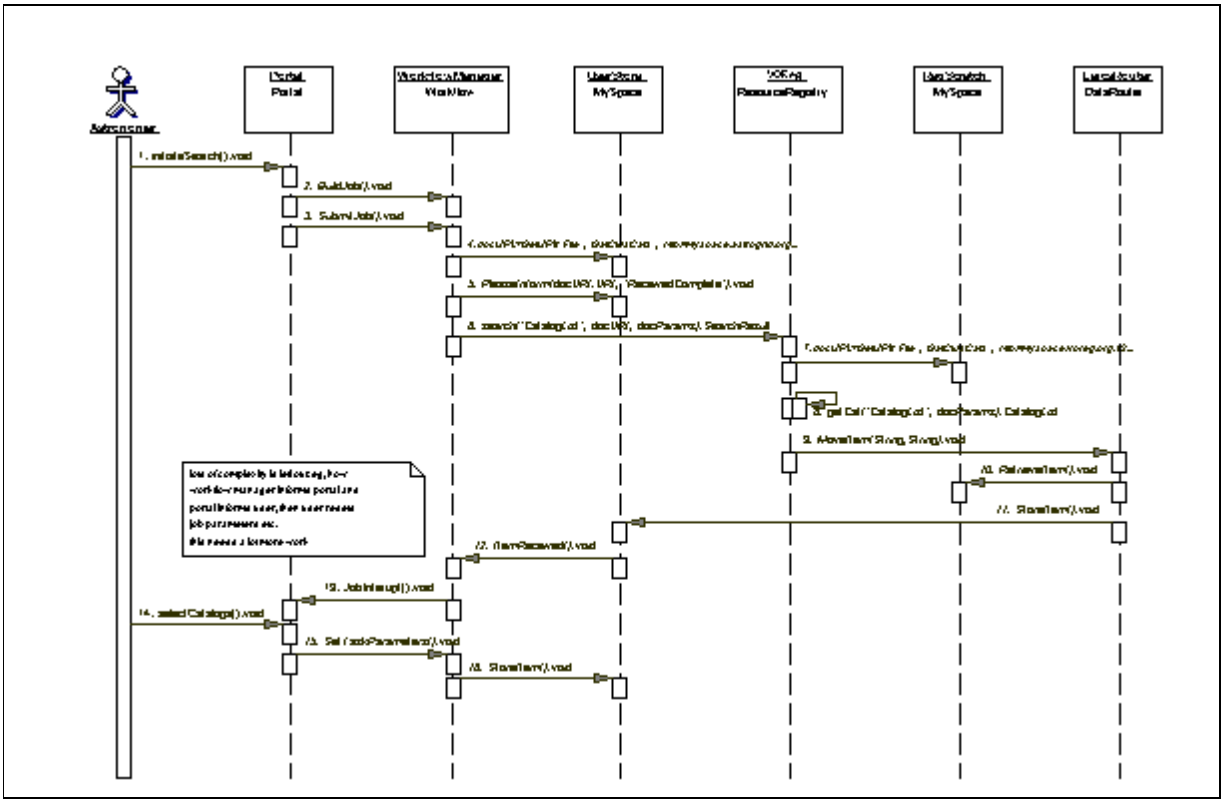
The creation of the services model has only just begun (and is expected to be complete by end of December 2002). The services in the model to date include:

- Activity Log
- Analysis Tools
- Application Resource
- AQL Translator
- Cas Server
- Compute Resource
- Data Mining
- Database Export
- Dataset Access
- Data Router
- Job Control
- Job Estimator
- Job Scheduler
- MySpace
- Query Estimator/Optimizer
- Replica Builder
- Resource Registry
- User Notification
- User Preferences
- Workflow

In addition to these services, it is envisaged that a web−based ***Portal*** and a PC−based ***Client*** program will be developed to enable the user to discover resources and construct jobs to run on the Virtual Observatory as well as a web−based ***Log Analyzer*** to provide resource analyses. For a rather simplistic view of the linkages between these services, the following is a good overview:



As an indication of the greater detail required in the services model, the following is based on the sequence diagram above but using services and only (so far) modelling the first two flows in that diagram:



Some of the core services include:

(3.5.1) Compute Resource

This is an *abstract* service (ie, one which does not actually exist but serves as a template for other services) which provides a standard set of properties and methods implemented by services which provide access to computing resources. This will enable another service to discover what facilities are available on the computer and how they can be accessed. Methods will be made available for user−written code to be uploaded and executed to make use of the facilities.

(3.5.2) Application Resource

This is also an *abstract* service. Implementations of this service will provide access to applications pre−installed at service sites. As an example, the AstroGrid team is currently working on wrapping the SExtractor tool in a web service; in the future, this web service would be expected to implement the Application Resource interface. Other likely examples are interfaces to IDL installations, visualisation tools etc.

(3.5.3) Resource Registry

This is the heart of the VO. The registry itself will contain a description and link to every resource in the VO. For AstroGrid, the decision was taken to implement a ***fine−grained registry***, so that as many queries as possible about resources available can be answered from the registry rather than having to forward those queries to the resources themselves. The registry service will provide a set of properties and methods which will enable the discovery of any resource in multiple ways. Metadata for each resource will be drawn from an astronomical ontology (see below for description of on−going technology investigation) allowing a linked inference engine to discover resources relevant to a user's enquiry.

(3.5.4) CAS Server

CAS (14) (Community Authorization Service) is a development of Globus to provide the ability to specify and certify the groups which a person may belong to on the grid. Each community will have the ability to define groups and members and assign rights to each. Data centres will then authorise access to their resources by communities, groups or individuals. AstroGrid will develop its own implementation of a CAS server as well as a user interface via client or portal software.

(3.5.5) Workflow

This service will be mainly used by the user interface programs, both the web−based and PC−based ones. It enables a user to create a programme of work, create jobs within that programme, add tasks to a job and then to submit the job and monitor its progress. The user may also create a job 'template', so that certain tasks can be rerun many times (possibly varying one or two parameters).

(3.5.6) Job Control

This is that part of the workflow service which controls the submission and monitoring of a job. It will also detect the completion of one task within a job and submit the next in the workflow.

(3.5.7) User Notification

This service will provide access to a range of notification services. A user can elect to be notified about the progress of a job: when certain tasks finish, when final or intermediate results are available etc. They can also specify how to be notified. This will typically be via email but might also be by logging into a notification web page, or even by SMS text message to a mobile phone.

(3.5.8) AQL Translator

This is less of a 'core' service but was worth including here because of the concept of AQL, ***Astronomical Query Language***. It has long been recognised in astronomy that the standard SQL language used to query databases is inadequate for many astronomical tasks. It is hoped that we might begin development of an AQL as part of this project and in conjunction with partner VO projects. What form it might take or how it might be put to use is still very much open. The AQL Translator service will parse an AQL query to determine the catalogs or data sets which need to be queried and create the relevant SQL queries.

(3.5.9) Dataset Access

This is an *abstract* service which will exist in front of any accessible dataset. This service is critical to the success of AstroGrid: many of these datasets are (or will soon be) overwhelmingly large and AstroGrid will provide the astronomer with the means of selecting a subset of data on which analysis can be performed. A dataset might be a catalog, a collection of FITS files or any other set of astronomical data. The service will provide details of how a subset of data is to be selected and retrieved (eg variant of SQL to be used, dbms type etc.) and will process queries to the dataset and return results (or pointers to where the results might be located if they are too large to move).

(3.5.10) Data Mining

This service will provide an interface to intensive data mining tools. We will, in order to address the key science problems, require more than simple data selection tools. Some methods of data selection will require, for example, intensive statistical analyses to be performed as the selection is taking place, all the time altering or tuning the selection criteria. We will develop some of these mining tools to enable the AstroGrid VO to prove the concept but, more importantly, we will create the interfaces and standards which will allow others to create similar tools.

(3.5.11) Data Router

This service will provide data movement facilities. Query results could be moved from the dataset scratch space to a user's permanent storage; data could be moved to a more powerful machine or to more complex software for detailed analysis etc. This service will need to be 'aware' of its environment, so that a movement from one part of the same machine to another is done with a simple copy while movement between separate sites is done by the most efficient method possible: eg GridFTP if both sites support the protocol, FTP if not.

(3.5.12) Server−based Analysis Tools

As with the data mining tools, we will require some analysis tools to be developed as a proof of concept. The key issue is that these tools must operate in a service−based, server environment. We will also develop the interfaces and standards that will allow other such tools to be developed.

(3.5.11) MySpace

This service is perhaps the most interesting new concept developed by the AstroGrid project. The concept allows the user to 'own' space on the VO. This space could be distributed across many computers and disks, all transparent to the user. For example, a user might run a query on a dataset and the results stored on that machine (providing security constraints are satisfied), then transferred to another machine when required as the model for another application, yet when the user looks at the MySpace directory they see only one object in their MySpace. A user will also be able to give access to any object in their MySpace to others and might make an object publicly readable, for instance when publishing the results in a paper. The service will provide methods for reserving space, adding objects to the MySpace, listing a user or group's objects etc.

(3.6) Technology Demonstrations

AstroGrid has set up a number of technology trials. These are designed to test new ideas, check the feasibility of new technologies or simply get a head start on probable components of the VO. The ***Pilot Projects*** (described in another part of the Phase A Report) were the most significant demonstrations. Smaller trials were also set up, and are still underway (15). These are:

- ***CAS Server***
  The goal was to produce a working CAS server (see above for explanation of CAS), enabling the creation of a community with groups of members having different access rights to a number of resources.
- ***Ontology trials***
  The goals are to produce: a first draft, skeletal, AstroOntology, incorporating UCD and VizieR information; a registry of (a few) UK−based astronomical catalogues, each described in ontology−based terms; an ontology−based registry access method; and an ontology−based workflow, driving a registry access web site.
- ***DBTF Technologies***
  The goal was to assess OGSA−DAI (16) technologies for access to XML and relational databases.
- ***AVO Science Demo***
  This is the AstroGrid contribution to the AVO science demonstration scheduled for January 2003. At this stage, our effort will concentrate on producing a web service which wraps the SExtractor tool and provides methods to be used by a modified version of the Aladin service accessing GOODS data.
- ***Data Centre trials***
  The goal is to demonstrate the issues involved in providing a web service front−end to an astronomical data centre and its archives. At this stage, it will focus on the Cambridge and Leicester data centres.
- ***Working Grid***
  The goals are to establish a working grid with: at least one machine at each of five sites running Globus 2 and able to GridFTP between each site; and at least one web service deployed at each site.

(3.7) Technology Choices

A number of technology choices have been made within the project (17). The first decision was to forego the existing Globus grid technologies and embrace the (as it was then) new concept of ***grid services*** within the Globus OGSA effort. We felt that web services offered significant benefits to future VOs: it was compliant with the direction of the W3C and industry; components could be packaged as discrete entities, running on servers without the problems of library conflicts that come with client−based programs; replacement components could be developed, deployed and slotted into astronomers' workflows with minimal effort.

Next choice was the development and deployment platform. The obvious choices were between the .Net platform and the Java platform. Although .Net offers technical advantages over Java, it is currently only available on Windows machines and is relatively new. For those reasons, plus the greater availability of Java developers, we decided to adopt the Java platform but expect to be able to make use of .Net deployed web services within our workflow.

The most significant technology choice still outstanding is that of database platform. We have evaluated several open source and commercial databases (18):

- MySQL
- PostgreSQL
- Oracle
- Microsoft SQLServer
- IBM DB2 (still being investigated)

No choice has yet been made but we are likely to select one or more to cover the following broad requirements:

- small internal tables
- MySpace
- data warehouse and data mining

Whatever our choices for the project, it is our firm intention that all data access will be via industry standard libraries (eg Java JDO) so that different databases can be used by those who choose to implement the AstroGrid VO.

(3.7.1) Open Source development

We are committed to developing the AstroGrid components in an ***Open Source*** way. This means that the source code will be freely available for anyone to download and make use of in any way they choose. We have not yet selected a ***license*** but will probably choose one from the LGPL, Apache, Berkeley style of licenses (21), which allows any use of the source code whether in other open source products or in commercial products.

Whether we also allow other people to participate in the project development process, by contributing changes to the code, is an issue we have not yet addressed. If people outside the project do express a wish to participate in the coding of our components, we will look at their request carefully.

(3.9) References

(1) UML Explained, *Kendall Scott*, Addison−Wesley, 2001

(2) Wiki: this is a web−based tool which allows any registered user to modify a set of pages. The AstroGrid project uses the wiki for all document storage. An explanation can be found at: http://wiki.astrogrid.org/bin/view/Main/WebHome.

(3) http://wiki.astrogrid.org/bin/view/Astrogrid/ArchitectureDocs

(4) See the brief explanation at: http://wiki.astrogrid.org/bin/view/Escience/UnifiedProcess.

(5) See the explanation at: http://wiki.astrogrid.org/bin/view/Escience/UPeSc.

(6) See the full list at: http://wiki.astrogrid.org/bin/view/VO/ScienceProblemList.

(7) A Sequence Diagram is a UML tool which shows a sequence of object interactions in time−ordered manner. In this context, it allowed the team to visualise how a VO would work as an astronomer used it on specific science problems.

(8) These ten key science problems are documented at: http://wiki.astrogrid.org/bin/view/Astrogrid/ScienceProblems.

(9) OGSA (Open Grid Services Architecture): effectively Globus Toolkit v3, this is is a proposed evolution of the current Globus Toolkit towards a Grid system architecture based on an integration of Grid and Web services concepts and technologies. See http://www.globus.org/ogsa/

(10) Conceptual Model: downloadable in document form as *AGProjectReport_d2.doc* from http://wiki.astrogrid.org/bin/view/Astrogrid/ArchitectureDocs, or as zipped Together directory as *astrogrid20020809.zip*.

(11) See: http://wiki.astrogrid.org/bin/view/Astrogrid/SequenceDiagrams

(12) This took place over a two day meeting in Leicester: the outcome of the meeting is documented at http://wiki.astrogrid.org/bin/view/Astrogrid/ArchitectureMeeting20020819 and the list of services at http://wiki.astrogrid.org/bin/view/Astrogrid/GridServiceList

(13) Services Model: downloadable as a zipped Together directory, *AGServices.zip*, from http://wiki.astrogrid.org/bin/view/Astrogrid/ArchitectureDocs.

(14) Community Authorization Service (CAS): see Globus page at: http://www.globus.org/Security/CAS/

(15) See http://wiki.astrogrid.org/bin/view/Astrogrid/DemoProjects.

(16) OGSA−DAI (http://umbriel.dcs.gla.ac.uk/NeSC/general/projects/OGSA_DAI/) is a subproject of the Globus OGSA (see above) effort. Initially started by the DBTF (Database Task Force, one of the UK e−Science teams: http://umbriel.dcs.gla.ac.uk/NeSC/general/teams/), it later became a working group of the GGF (Global Grid Forum: http://www.globalgridforum.org/6_DATA/dais.htm).

(17) See http://wiki.astrogrid.org/bin/view/Astrogrid/TechnologyDocs.

(18) See http://wiki.astrogrid.org/bin/view/Astrogrid/DbmsEvaluations.

(19) The key reference manual for the Unified Process is: The Unified Software Development Process, *Ivar Jacobson, Grady Booch, James Rumbaugh*, Addison−Wesley, 1999
For the three key−phrases and an explanation of them, see p4 onwards.

(20) The AstroGrid wiki lists two sets of use cases. In the VO web, http://wiki.astrogrid.org/bin/view/VO/UseCaseList, the use cases refer to any potential VO; in the AstroGrid web, http://wiki.astrogrid.org/bin/view/Astrogrid/UseCases, they refer to the resolution of the AstroGrid key science problems.

(21) The Open Source Initiative, http://www.opensource.org/, maintains a reference of approved open source licenses at http://www.opensource.org/licenses/index.php.

(4) Virtual Observatory Prototypes

(4.0) Introduction

The idea of the Virtual Observatory arose gradually over a period of time in which astronomical data archive sites continually improved their facilities and user−interfaces, but in a piece−meal and uncoordinated way. Many of us realised that a coherent and planned approach, with standardised interfaces, would permit a radical advance. As part of AstroGrid's initial programme, therefore, we carried out a short survey of the most advanced existing web sites and software packages to determine the facilities they provided and how they worked. Nearly all of these rely on technical solutions which pre−date Data Grids and the Web Services paradigm, so we did not expect there to be much scope for direct technology transfer, but the facilities reflect the perceived needs of the astronomical community, and we hoped to learn a lot from them about what AstroGrid needs to provide, and note the strengths and weaknesses of the current solutions.

These investigations were carried out as a joint exercise between our grid technology and database technology teams. It should be noted that they represent a snapshot of practice in the early part of 2002, and that many facilities have changed since then.

(4.1) Current Services, Sites, and Software

4.1.1 Astrobrowse

Astrobrowse is one of the web interfaces provided by the High Energy Science Archive Research Center (HEASARC) at NASA Goddard Space Flight Center (GSFC). Its most advanced facility is a distributed cone search: users can enter the coordinates of an object and a search radius into a single form and then search a large number of different on−line astronomical catalogs from around the world. Alternatively one can specify the name of an object, and NED or Simbad will be used to determine its coordinates.

There is a choice of a quick or full search, the latter form giving many more options and allowing a more selective search of archive sites. Bandpass, data type, and other keywords can be specified to refine the selection. One can also select which types of service to interrogate (e.g. optical or radio), or which individual servers, with considerable flexibility.

Astrobrowse uses AstroGLU software, a development of the GLU system written at CDS (Strasbourg). This contains a list of URLs of each data archive service and the parameters each requires, which it uses to generate a customised CGI query for each of them in turn. Only one or two sites have adopted the conventions proposed by CDS, so the GLU database is essentially a loving compilation of the idiosyncrasies of each site.

Astrobrowse then waits for the various replies to come back. This may be slow if one selects a large number of servers around the world: the results page has a side−bar reporting the status of each service, with an option to refresh this at intervals. In practice it takes only a few seconds to get results from the most servers, but it may be necessary to wait for a few minutes for all of them to respond (and at times some responses never appear). Astrobrowse has not attempted to solve the problem of integrating the results, which appear in a wide variety of formats, requiring some expertise to understand. This is obviously a difficult problem, but one which the VO needs eventually to solve (e.g. with the aid of VOTable and UCDs).

The software is available for download, and a version is also in use at Harvard/Smithsonian Astrophysical Observatory. Our tests were carried out on version 1.7.

4.1.2 Browse/W3Browse

Browse also comes from HEASARC at GSFC: it is essentially a search engine for tabular data, originally designed for data from high−energy observatories, but now broader in scope, including many radio and optical catalogues, as well as links to Vizier. The underlying DBMS is currently Sybase, but some attempt has been made to keep the software DBMS−independent. Version 6.3 was current when these comments were made. As with many astronomical archives, the primary search is by position or object name, resolved using Simbad or NED. Results can be produced in four formats: plain text, HTML tables, FITS tables, or Astrores XML.

Although the basic facilities are similar to many other data archives, Browse has two particular strengths:

- It can cross−correlate (join) two or more tables on celestial position (or date/time of observation). The results of the join can be sorted, or plotted if the browser is Java−enabled (but this often turns out to be slow).

- Browse also has links to original datasets so that, having found the required observation, users can download data products from observatory missions. These two features: joining tables, and downloading data products, are very valuable, and must form a feature of our VO design.

HEASARC's web−based browse service was for a time known as W3Browse to distinguish it from the original BROWSE service accessed by `telnet`. The software was originally written in ESOC around 1980 as the interface to the EXOSAT Observatory's data archive. The service later moved to ESTEC, then to HEASARC, and versions were subsequently installed at other sites including LEDAS (Leicester) and MPE (Garching). The original telnet service is still available at LEDAS, although usage is dwindling. A few of the useful features of the original Browse have been lost in the web version, e.g. the ability to save the results of a filtering (select) operation and then make further selections on that. These features may be beyond the scope of a simple web service, but certainly ought to be provided by a data mining service. This is something that our MySpace concept should be able to address.

4.1.3 CURSA

CURSA is a Starlink package for manipulating astronomical catalogues and tables. It is mostly concerned with accessing catalogues held as local files, but also provides some facilities for searching remote catalogues. Remote catalogue searches are available within the GUI−based catalogue browser `xcatview` and from the Unix command−line by using the application `catremote`. The only type of remote search supported by either application is the `cone search' to find objects within a specified angular separation of a specified central celestial coordinate. Optionally, the name of an astronomical object may be given instead of a central coordinate and the SIMBAD or NED name−resolver is used to replace the object name with the corresponding coordinates. For some catalogues `catremote` also allows additional selections on pre−defined columns (for example, limiting the selected objects in the specified region of the sky to also lie in a given magnitude range).

CURSA uses exactly the same mechanisms and formats as SkyCat and GAIA for submitting queries to a remote catalogue and returning the table of results, and has the same advantages and limitations. The most notable limitation is that it is only possible to search one remote catalogue at a time. `xcatview`'s GUI for searching remote catalogues has a rather different layout to SkyCat's. `catremote` is suitable for embedding in scripts as well as for interactive use from the command−line. Searching elements of the VO from within scripts which perform specialised, bespoke tasks seems a likely requirement for the VO.

It is also worth noting that CURSA is based on the FITSIO library written at GSFC, which has FTP and HTTP protocols for data access, so that all the CURSA tasks can access remote tables if they are present on FTP or HTTP repositories in FITS table format. They do this, however, by copying the entire file to local memory or scratch disc. The same facility is built in to the FTOOLS utilities for handling FITS files provided by GSFC. For small tables this is fine, but for large tables bandwidth can be a problem.

The tests were done on CURSA version 6.4. Documentation is provided in two Starlink documents, SUN/190 and SSN/76.

4.1.4 ISAIA

The ISAIA project, led by GSFC, was intended to develop a number of virtual observatory concepts, such as the integration of results from queries such as those sent out by systems like Astrobrowse. Although the project is no longer active, and those involved are now part of the NVO team. The website contains several useful documents, but these are now being overtaken by more recent developments.

4.1.5 MAST

MAST (Multi−mission Archive for Space Telescope) comes from the Space Telescope Science Institute (STScI) and is the optical/UV/near−IR component of NASA's distributed Space Science Data Services and aims to provide integrated access to data from a range of missions/projects, namely: Hubble Space Telescope (HST), Far Ultraviolet Spectroscopic Explorer (FUSE), International Ultraviolet Explorer (IUE), Extreme Ultraviolet Explorer (EUVE), Hopkins Ultraviolet Telescope (HUT), Ultraviolet Imaging Telescope (UIT), Wisconsin Ultraviolet Photo Polarimetry Experiment (WUPPE), Copernicus (OAO−3), Orbiting and Retrieval Far and Extreme Ultraviolet Spectrograph (ORFEUS), Berkeley Extreme and Far−UV Spectrometer (BEFS), Interstellar Medium Absorption Profile Spectrograph(IMAPS) (first flight), Tübingen Echelle Spectrograph (TUES), Digitized Sky Survey (DSS), Guide Star Catalog II (GSCII), Sloan Digital Sky Survey (SDSS), FIRST (VLA radio data), Roentgen Satellite (ROSAT).

MAST offers a series of cross−mission search tools, which vary from tools directed to specific science cases (e.g. find all data in MAST archives close to Abell clusters) to a Single Target quick search interface, where one can enter coordinates or source name (to be resolved by NED or SIMBAD), and get a list of data available within MAST: e.g. entering M31 yields the predictable long list of data − one nice feature is that preview versions of images (in GIF format) are provided. This interface also enables searches by data type − e.g. checking the "*X−ray spectra*" box returns a link to the top page of MAST's ROSAT site, as well as a helpful note that HEASARC provide access to a wider range of high−energy data. There is also the MAST Scrapbook, which offers users "*representative images or spectra of an astronomical object*" (specified by coordinates or resolvable name, as before) − basically another way of asking for preview data.

MAST also hosts a series of Prepared Science Products. Examples of these are the Hubble Deep Field (North and South) datasets, various UV spectral atlases and the SDSS Quasar Catalog, derived from the Early Data Release of the Sloan Digital Sky Survey. A page on Data Analysis Software lists "*some of the data analysis software packages used for the MAST archived data*", mostly comprising IRAF packages written for specific instruments. Data transfer varies between the different archives in the MAST system, but is usually either via anonymous ftp or direct downloading from a WWW browser.

Overall, MAST is a good example of a current−generation archive site: it provides access to data from a number of sources in a reasonably coherent and user−friendly manner, and with a reasonable amount of documentation (the documentation for the SDSS EDR site has improved markedly in recent months) and a Helpdesk which is staffed Mon−Fri 09.00−17:00 (EST), but it seems very interactive − there's no obvious batch mode facility, or any such means of submitting large numbers of queries in an automated fashion.

4.1.6 NED

NASA's Infra−red Processing and Analysis Centre (IPAC) at CalTech provides the National Extragalactice Database (NED). It is built around a master list of extra−galactic objects for which cross−identifications of names have been established, accurate positions and redshifts entered to the extent possible, and some basic data collected. Bibliographic references relevant to individual objects have been compiled, and abstracts of extra−galactic interest are kept on line. Detailed and referenced photometry, position, and redshift data, have been taken from large compilations and from the literature. NED also includes images for over 700,000 extra−galactic objects from 2MASS, from the literature, and from the Digitized Sky Survey. NED's data and references are being continually updated, with revised versions being put on−line every 2−3 months. In essence, therefore, NED provides facilities somewhat similar to those of CDS but specially only for identified extragalactic objects, and specially adapted to the needs of those studying them.

4.1.7 Querator

Querator is a tool for extracting images of a given region of sky from image surveys. It was developed by Francesco Pierfederici at the European Southern Observatory (ESO). In a single query it can extract images of the same region of sky from several different surveys, thus allowing a stack of images, typically in different colours or wavelength ranges, to be returned. Currently images from the HST and various ESO telescopes are available. The region of sky to be extracted can be specified in a number of ways:

- object name,
- sky box,
- external server search,
- user file upload.

The object name, sky box and user file upload options are all as would be expected. In the first two cases the user gives, respectively, an object name or the meridians of Right Ascension and parallels of Declination defining a region of sky. Additional constraints (exposure time, observation date, wavelength range, instrument etc.) can be specified to refine the search. In the user file upload option the user gives the name of a prepared file containing a list of object names or coordinates. This useful option allows images to be retrieved for a number of regions in a single operation.

The "external server search" option is more interesting and innovative. Here the user submits a query to a remote catalogue archive (such as LEDA or the NASA ADC catalogue collection), which is quite separate from the data centre holding the image surveys, in order to search one or more catalogues according to an arbitrary criterion the user has supplied. The remote catalogue archive returns a list of objects which satisfy the query. Querator takes this list and retrieves images for all the objects listed.

Access to Querator is solely through a Web interface, which is generally easy to use. However, constructing the query for the

remote archive in the "external server search" option is complicated. The query is constructed using the native syntax of the remote service and thus varies between different services. Querator seems to still be under active development. The query pages crash `netscape` running on a Compaq/Alpha but are ok on Sun/Solaris (though this could be a bug in `netscape` for all that I know). The surveys currently accessible mostly seem to consist of pointed observations. However, presumably, there is no reason why Querator could not access contiguous surveys, such as the DSS.

Querator has a number of features which seem likely to be required in the VO, including the ability to retrieve a stack of images from several surveys in one query and something analogous to the "external server search" option. However, to make the latter easy to use a unified (and simple) query syntax to specify queries on all the remote catalogues is required.

No version number was given. The tests were conducted on 13 February 2002.

4.1.8 Simbad, Vizier, and Aladin

These interwoven services are provided by the Centre de Doneés de Strasbourg (CDS), the oldest and probably largest collection of astronomical data resources in the world. There are many interconnections between the separate services which make the system easier to use, but make it harder to see which part does what. Essentially:

- **Simbad** is principally a bibliographic archive, which includes information about all papers in the primary astronomical literature about objects beyond our solar system, including the properties of the celestial objects listed therein. This means that Simbad holds many of the small surveys in the literature (e.g. fields of the 5th Cambridge catalogues of radio sources) but not the massive data−collections like SDSS. Naturally, the type of measurements are very varied. Simbad's web interface only allows queries on a few basic parameters, most of which are biased towards stellar astronomy. We note that Simbad's object classification scheme needs to be considered in our ontology efforts

- **Vizier** is billed as a "catalogue of catalogues" which underplays what it can be used for. There are two main uses: selecting catalogues (by criteria such as "contains QSOs") and listing the descriptions of those catalogues; selecting objects from the union of all the catalogues by various criteria. That is, Vizier allows both metadata and data searches. The fact that these two modes are driven from the same interface−page makes Vizier harder to use than it need be. The "union of all catalogues" seems to mean the catalogues absorbed into Simbad plus major external data−collections such as 2MASS.

- **Aladin** is an image display with advanced overlay−features. The Vizier and Simbad operations can display results by returning a web page in which Aladin runs as an applet with the data preloaded. Alternatively, Aladin can run as an application and can send queries to Simbad and Vizier in response to user actions. Aladin allows overlay plots from many catalogues to be stacked up, and provides good controls for manipulating the stack (e.g. controlling visibility of particular planes).

- **The bibliographic database** lists the papers from which data were taken for Simbad. This makes it excellent for use with Simbad and dangerous to use for any other purpose due to the specialized pre−selection. The service also holds on−line abstracts of recent papers in selected journals.

The popularity of these services is shown by the existence of several mirrors: Simbad has a mirror in the USA, while there are already half dozen mirrors of Vizier, including one in the UK (Cambridge). Simbad is also the principal name resolver (translator from celestial object name to coordinates) used by many other sites around the world: it is an obvious candidate for conversion to a *Web Service* using SOAP/WSDL.

4.1.9 Skyview and Skymorph

Skyview is another service of HEASARC at GSFC. It describes itself in these terms: *SkyView* is a Virtual Observatory on the Net generating images of any part of the sky at wavelengths in all regimes from Radio to Gamma−Ray.

The SkyView server contains copies of images of the sky taken in a wide range of wavebands from radio to gamma ray, mostly (perhaps all) stored as FITS files. The SkyView software, written at GSFC, selects and overlays these images, giving results in one's chosen resolution, and it automatically handles rotation, precession, coordinate transformations, and pixel re−sampling. The results can be seen on the screen, or a FITS image can be downloaded from an FTP area in a number of formats including FITS, TIFF, GIF, and PostScript. There are actually five different interfaces: for the non−astronomer, basic, advanced, Java, and X−windows. The latter is regarded as obsolescent, now that Java can provide the required controls. More advanced image options, such as changes to color tables, overlays on extent images, image rescaling, zooming, etc. require a

Java−enabled browser.

Advanced options include the ability to overlay data from two or three different data sources, perhaps mapping each to a different primary colour, producing a pseudo−colour result. This is even possible for those using 8−bit displays. SkyView can also implement boxcar averaging of an image, to obtain a smoothed result. There is also a batch option, with Perl scripts which can be downloaded and run on a Unix/Linux system. The software is freely available for download.

The software, written by Tom McGlynn and his team, is all available for downloading and external use. The comments here apply to Web Version 4.1, with version 3.2 of their Geometry Engine. A new interface is currently (2002−02−01) on beta−test and allows the interface to be customized; a few functions did not seem to be working correctly when tested.

Overall the SkyView facilities for image selection and display are so comprehensive, and so well covered by the documentation, that it is hard to think of features still lacking. But it must be noted that facility has been provided entirely using *local* storage, by taking copies of datasets produced elsewhere and, where necessary, reformatting them to suit SkyView. It was, apparently, decided when Skyview was designed that the Internet did not have enough bandwidth to allow the retrieval of images from remote sites.

SkyMorph specialises in searches for variable, moving or transient objects. It provides convenient access to optical images and catalogs generated by the Near Earth Asteroid Tracking (NEAT) program. These include more than 67,000 CCD images covering a large fraction of the sky. The same region is typically observed several times each night, and is revisited on monthly and yearly timescales. SkyMorph appars to be based on Skyview, and seems to have few unique features of VO importance, but it is one of the few services which supports the time dimension, which AstroGrid must not neglect.

4.1.10 Skycat, GAIA, and JSkyCat

SkyCat is an image display tool developed by Allan Brighton and colleagues as part of the ESO VLT project. GAIA is an enhancement of SkyCat by Starlink, which has added numerous astronomical analysis facilities, including: astrometric calibration, automatic object detection and aperture, optimal and surface photometry. Both SkyCat and GAIA are mostly concerned with accessing local files. However, they both contain some limited facilities for accessing remote catalogues and image surveys. GAIA's facilities in this area are identical to SkyCat's and the following notes apply to both applications.

SkyCat and GAIA can access a reasonably extensive remote collection of standard astronomical catalogues and a few image surveys, principally the HST Digitised Sky Survey (DSS). The principal purpose of remote catalogue searches in SkyCat and GAIA is to find objects which overlay an image that has already been displayed by the application (though searches can be made which are not connected with any image). Consequently, the only type of remote search supported is the "cone search" to find objects within a specified angular separation (or `radius') of a specified central celestial coordinate. Optionally, the name of an astronomical object may be given instead of a central coordinate and the SIMBAD or NED name−resolver is used to replace the object name with the corresponding coordinates. For some catalogues additional selections are also supported on pre−defined columns. For example, it may be possible to select objects which lie in the specified region of sky and which also lie within a given magnitude range.

Regions of sky can be extracted from image surveys by specifying the central coordinates and size of the field required. Again, optionally, an object name can be substituted for the central coordinates.

Skycat and GAIA have a convenient user−interface which is well−integrated with the rest of the display functions. Retrieved objects are automatically plotted on top of a displayed image if they overlay it. It is easy to highlight a given object in both a table of the selected objects and in an image overlay plot.

The list of remote catalogues and image surveys available to Skycat and GAIA is held as a text file. This arrangement is good in that the list is not hard−wired into the code and can be customised, but is bad in that the file has to be edited manually, rather than maintained automatically by a `resource register' of the sort that we have been discussing.

Queries are submitted and results returned using HTTP protocols. The query format is somewhat restricted (and is similar to, but not identical with, the ASU query standard). Tables are returned in the Tab−Separated Table (TST) format, which is somewhat deficient in catalogue metadata, though it does contain enough information to define how objects are to be plotted on overlays (ellipses etc). Images are returned as FITS files.

A VO client or portal would need to provide at least all the remote access facilities of SkyCat and GAIA. Their principal disadvantage is that they can only search one catalogue at a time.

The version tested was GAIA version 2.6, derived from SkyCat version 2.4. On−line documentation for SkyCat is available from its home page at ESO.

JSkyCat is a re−implementation of Skycat (above) in Java. It was also developed by ESO. It has similar functionality to the original Skycat, but has fewer features because it is still under development. JSkyCat is written using the JSky Java class library, elements of which are also used in the Gemini Observation Planning Tool.

The remote catalogue and image survey access facilities in JSkyCat are essentially identical to those in SkyCat: it provides the same functionality, uses the same mechanisms and formats for submitting queries and returning tables of results and has the same advantages and disadvantages.

The version tested was JSkyCat 1.2; on−line documentation is available from the ESO web pages.

4.1.11 Starcast

Starcast, also from STScI, is MAST's prototype implementation of Astrobrowse, described above. The Starcast implementation currently uses a Perl interface to the profile database, not the CDS GLU system as used in the original Astrobrowse prototype at HEASARC, but is intended to migrate to using GLU at some point: this will mean that the Starcast administrator will not have to input the profiles manually, as is currently the case.

The Starcast query form allows the user to search for data around a sky position or an object with a name that can be resolved into a sky position by NED or SIMBAD. The user then specifies the Bandpass (running from radio to gamma ray) , Data Source (with choices *Any*, *Derived*, *Observations*, *Pointed*, *Proposal*, *Survey*, *Survey Data*), the Data Type (*Any*, *Catalog*, *Image*, *Images*, *Other*, *Spectra*, *Spectrum*, *Time−series*), and Location (*Any*, or a selection from a list of about 30 international data centres) and sets the query running. The browser moves to a new page, with two frames, one of which lists the specification of the query, and the second gives links to the services which might have data satisfying the query: next to each of these links is an icon showing whether the search on that resource is running, has completed successfully or has crashed, and these may be updated by pressing a *Check status* button at the top of the frame.

The implementation of this service seems incomplete, in some sense. For example, a test query asking for EUV data from *Any Source* of *Any Type* and at *Any Location* within 10 arcmin of `10 00 00 −10 00 00` returned a number of links, one of which was to the IMPReSS interface at the NASA ADC at Goddard. Clicking on that link took me to a WWW page generated within the IMPReSS system, which listed the sky position for my search and presented me with a list of archives (not just EUV, but also X−ray and optical) with data around that position, asking me which I wanted to choose. Clearly, since I'd already specified my query on the Starcast WWW form, I should have been taken one stage further within IMPReSS. This is something of minor quibble, for what is, after all, just a work−in−progress prototype implementation, but it does highlight the difficulty of fitting a top−level query interface on top of existing data centres, each of which provide access to their archives in different ways.

4.1.12 Starview

StarView comes from the Space Telescope Science Institute (STScI) and its blurb says that "*StarView is an astronomical database browser and research analysis tool. Developed in Java, StarView provides an easy to use, highly capable user interface that runs on any Java enabled*

platform as a standalone application." Download and installation (under Windows NT) was remarkably simple, and the Java GUI is very nice. Starview can be used to search for data in MAST archives, examining the calibrations used for a particular dataset, and look at proposal information relating to past HST projects.

Downloading of data through Starview requires registration with the STScI archive, and is performed either by leaving the results file in an anonymous ftp site, or by ftping them to the user's machine: the latter requires supplying the user's password, and this is very unpopular with some system managers. One nice feature is that you can track the progress with your query on a WWW site.

Queries are defined by starting with a form (a number are provided, as templates for searching particular archives or making particular kinds of query) and then the user adds *qualifications* to narrow the search. Queries can be written out as SQL, which is nice, and there is also a *Cross−Qualifier* feature, which allows the results of one query to be used as input to constructing a second: this seems a very useful feature, but the instructions are not clear enough to enable a user to use this option at a cursory reading.

The results of a query are listed in another GUI, and datasets can be selected from that window for futher operations − e.g. previewing images or spectra, looking in ADS for references known to have resulted from that HST proposal, overlaying the instrument footprint on a DSS image − while the list itself can be exported as an ASCII file. One very nice feature is that the list of returned datasets can include proprietary ones, for which the date of public access is listed: slightly annoyingly, one has to remove those datasets from the list manually before asking to retrieve the data...surely a better default is not to include them, and then only the PI (who can access them) would have to do anything. A variety of data types can be retrieved − it is interesting that one can retrieve data quality information and/or observing logs, in addition to the data themselves.

All in all, this is a very interesting tool, displaying much of the functionality that one would require for the VO. As with the MAST WWW interface, this is still very interactive, but it does have the advantage that one can store and reload queries one has formulated interactively using the GUI. Again as for the MAST WWW interface, there is no description of the technology used, beyond Java.

4.2 Testing CDS and NED with use cases

In order to find out what these services can do, we tried to use them to do the Astrogrid use cases, but found that only a very limited subset of what we wanted to do was currently possible. The section names in the following are the Wiki−names of the use cases in the VO Wiki−web.

4.2.1 FindQSOsByPosition

Vizier and Simbad can do the main flow of this use case, using Aladin as the display tool. Ironically, Aladin itself can't make the necessary selection. None of these tools can merge the tables of results.

NED can do some of this work. It can't select on radius from the search centre, but it can select on ranges of RA and dec., which is almost as good. It understands "QSOs" and "QSO clusters" as selection criteria. The plotter ("skyplot") from NED is poor (line graphics only) and is non−interactive: there is no way to select objects on the display and get to their details from the catalogue.

4.2.2 GetLiteratureReference

This feature is available in Simbad. Results of searches carry hyperlinks to entries in CDS' bibliographic service. However, only selected references are shown (to explain where the Simbad data came from, not to refer to the science). It is possible to query the bibliographic service directly, but the number of references returned is surprisingly small (e.g. 7 for a search on NGC1068).

NED allows one to query the database of abstracts directly by object name. This finds many (all?) references (e.g. 1249 for NGC1068).

4.2.3 GetReducedSpectra

There is no obvious a way to get any actual spectral pixel−data from any of these systems except one small part of IPAC. The SWAS mission, available through IPAC, serves spectra as either web graphics or in FITS files.

4.2.4 InstrumentFootprint

None of the systems seem to allow an instrument footprint as a search area in a query.

Aladin allows a user to draw one of a set of limited footprints as an overlay on an existing plot. If one then measures the footprint in Aladin one can get a search radius that encloses the footprint, and can search on that radius in Vizier. This allows the work to be done manually.

4.2.5 ObservingProposalCheckForData

None of the systems do this use case. There are no links to observation−proposal systems.

4.2.6 PhotometrySearch

None of the systems appear to cover this case, and there are no references to software for interconverting magnitudes and fluxes except in NED, which is trying to go in the opposite direction, from photometry to coarse spectra.

4.2.7 PosteRestante

None of the systems even attempt this except for 2MASS (accessed through IPAC) which has a batch system for producing image extracts.

4.2.8 SelectAstrometricStandards, SetImageWCS

None of the systems can do these cases as written. There is no support for actually doing the astrometric fit, nor for plotting the residuals on the fit.

Vizier and Simbad can do most of SelectAstrometricStandards, but they cannot select the "best" catalogue out of the many available. Aladin doesn't help with this case, since the idea is to automate the process, not to do it interactively.

NED is not very helpful, since stars are needed, not extra−galactic objects (but QSOs may be valuable in future).

4.2.9 SyntheticSpectra

NED can do this very nicely, but only for one object at a time. The initial selection of data is not quite as general as in the use case. The plot is done as a web graphic displayed in a web browser.

Simbad, Vizier and Aladin can't do this work. It isn't even straight−forward to extract the photometry so that one can do it manually.

4.2.10 SelectionOfTrustedCatalogues

None of the systems allow this work to be done as stated.

NED allows references to be looked up, but not using bibcodes.

Vizier, Simbad and Aladin do not support bibcodes as a search term, but they do return bibcodes in the results of their results. The CDS bibliographic service does good searches by bibcodes.

None of the software helps with handling the list of data and bibcodes as suggested in the use case. Aladin could be used to display the objects in the user's catalogue and the user could then cross them off as the bibcodes were checked by drawing into the graphics overlay.

4.2.11 Use cases involving authorization and authentication

None of the systems inspected here deal with these issues.

4.3 Discussion

4.3.1 Use−cases

Some of the systems have query interfaces somewhat like those we shall want to provide in the "VO Portal", especially Vizier, the fancier bits of NED, and the facility in Astrobrowse which allow the concurrent searching of multiple web−sites. However, all the systems have the same basic philosophy: *display lots of data and metadata in a web page and give chains of hyperlinks to even more data.* They make no attempt to provide consistency in the results from disparate sources, as this would be very difficult with the existing infrastructure.

The sites, notably Vizier, Aladin, and Skyview, which make it possible to search a number of datasets in an integrated fashion have managed this by providing all the data in the right format locally. One of the principal aims of the VO projects will be to provided similar facilities but from federations of data accessed from their original locations.

Most of the use cases were not supported because they involved the technique "do a search, then do something specific with the results of the search". The VO−like archives are not set up to handle the "do something with the results" part, since they only represent the results as web−pages, not as semantically−useful data held for further processing. The exception is the making of synthetic spectra in NED, and this is a specific application − a vertical integration − that has clearly been coded in specially. It's not the kind of processing that a user can set up using a script and separate services at NED.

Some of the use−cases failed because the various archives do not have uniform criteria for selecting objects. In any given query, the selection criteria must either be on quantities that the interface designer coded into the UI, or there must be a free−form interface for specifying other criteria: a query language known to the user. The existing systems don't expose a query language, and their web interfaces only deal with a few quantities.

The use case GetReducedSpectra fails because the systems do not seem to provide reduced spectra. They only deal in images and tables.

The systems don't seem to deal in identified usage. Presumably, this means that they allow less access to data than a given user is entitled to.

In general, the systems reviewed let you look up more easily data that you could get by trawling through paper journals or by using interfaces to individual large archives. They require you either to know what you are looking for at the start (e.g. which catalogues to search) or to be prepared to spend a long time browsing. The output of the search is as for searches in paper collections: text you can read, but not machine−readable data products.

4.3.2 Conclusions

The systems studied here have a wealth of good features, many of which we need to emulate, but we were also able to identify a number of missing features and weaknesses in current systems which the VO alliance needs to address. These include:

- Searches over distributed resources are important but difficult, because of a lack of agreed standards for queries (both simple and advanced), for metadata, and for the results (both extracts from tables and from images).
- Resource discovery at present requires expert knowledge − a scalable resource discovery mechanism is needed.
- These web sites all support interactive queries, but few have any facilities for batching them up, e.g. to retrieve results from a list of interesting celestial positions.
- The ability to do cross−identifications between catalogues on different sites is important (via the fuzzy−join algorithm) but facilities for this are rare and hard to use at present, and bandwidth may limit what can be done over the network.
- It is possible to construct services, such as Simbad, Vizier, and Aladin, which are separate but so well−linked that they appear as an integrated system, but these are exceptional and they are all co−located. If services on separate sites could be as well integrated, this would be a good step towards the VO.
- We need to consider how best to support the study of time−varying and transient phenomena, somewhat neglected at present.
- These archive sites used a variety of commercial and free DBMS (Sybase, Ingres, Oracle, SQL Server, MySQL, and probably others) as well as some home−grown database systems. Web Services interfaces will be needed for almost all of them.

(5) Report on Grid Technology

(5.1) Introduction

(5.1.1) What is "grid technology"?

Grid technology can usefully be defined as technology that helps a user to exploit distributed computing without needing to know *a priori* the disposition of resources about the network, or the exact nature of those resources. By extension, technology that allows computing jobs to run remotely without continual human intervention also counts as Grid technology. This definition widens the scope of the enquiry beyond the triad of products − Globus Toolkit, Condor and Storage Resource Broker − identified by the core e−Science programme at Astrogrid's inception.

(5.1.2) AstroGrid's approach to the Grid

AstroGrid is about astronomy, not computer science. Our remit is to support working scientists and to show how a virtual observatory may be built. We use Grid technology as and when it suits our purposes and not to demonstrate the grid to other disciplines.

Our preferred approach is to define a system architecture from the science requirements and then to choose available technology to suit that architecture. We wish to build our own grid technology only where no existing products suit our purposes.

However, technology choices themselves tend to dictate architecture. We have therefore followed an iterative approach of defining architecture first according to what is natural for available technology, then altering this abstract architecture to bring it back in line with AstroGrid requirements, and finally reassessing the technological products against the revised architecture. Since there are many griddish products to choose from, our early priorities have been to detect and avoid the unsuitable ones.

As a result, most of our investigations into Grid technology to date have been thought experiments rather than executable prototypes. This is a cheaper way, in terms of staff effort, to eliminate the unhelpful products. Now that the architecture is known, we are starting to run more practical experiments to verify our *positive* choices of technology.

(5.2) The natural architectures

We looked initially at the architectures that are natural to grid products, on the off−chance that one of these might entirely meet AstroGrid's needs. There are three principal approachs, outlined below, and a Grid can combine them as necessary. However, given products tend to promote just one of the three architectures.

(5.2.1) Compute−grid

A natural compute−grid values processing power ahead of network bandwidth and richness of applications. The resources in the grid are "unfurnished" processors which are leased by users for the duration of a job and to which all data and executable software must be shipped in; results must be shipped out at the end of the job.

Compute grids are good when jobs are limited by CPU power; when software is self−containing and easy to run on arbitrary computers; and when data sets are small and simple enough to be readily portable between nodes of the grid. These grids work poorly when data sets and software are large or complex.

(5.2.2) Data−grid

A natural data grid values the software resources at any given location above processing power and network bandwidth. The resources in the grid are storage locations from which any programme can easily read and write data. The grid is effectively a distributed file−system from which data are copied to a central location for processing. Data−grids tend to work in terms of files rather than database tables.

Data grids are good when sites have large collections of complex software that needs to run locally, but need data from elsewhere. These grids do poorly when the data−sets for a job are too big or complex to move easily, or when the site with the software has too little CPU power.

(5.2.3) Service−grid

A natural service grid values data locality and the conservation of network bandwidth above CPU power and locality of application software. It makes operations on data available as network services at the sites where the relevant data are held. In the purest form of a service grid, no data ever passes from service to service and only a bare minimum between services and the user's client−software.

Service grids are excellent when data sets are too large to move about *in toto*, or when access to the data sets depends on local infrastructure such as a particular DBMS. Service grids also help in the curation of data, since they keep both the data and the specialist software for those data at the curators' sites. The grids do less well when a job requires data and software from many services to be combined.

(5.3) Astrogrid's technology needs

AstroGrid's architecture is described in section 3 of this report. AstroGrid is now seen as basically a service grid, but with a data−grid incorporated to move data between services and to store intermediate results. At present, there are no plans to include a compute grid. Although some substantial processors (e.g. Beowulf clusters) are available to AstroGrid, these will initially be the sites for pre−installed application−services, not resources to which code is uploaded by end−users. The crucial points of the architecture are as follows.

All access to bulk data−sets in archives is through data−selection services that isolate data−extracts small enough to move about the data−grid. In the rare cases where a job needs to access all of a major data−set, then the processing services for that job are constrained to run at the data−centre where the data−set lives. Hence, large−scale data−mining operations are likely to be at major data−centres, not at arbitrary places on the grid where there is spare CPU power or storage.

Data−selection services are needed for data stored in flat files (mainly pixel data) and also for structured data in DBMS.

Data−processing services are needed that wrap up existing, general software in astronomy, such as the Interactive Data Language, the Astronomical Image Processing System and the Starlink Software collection. These services should be available at many points on the Grid. Where possible, they should be available at the places where data extracts are made, such that the extracts do not have to cross the network.

Most of AstroGrid's work will be the analysis of data in archives that have already been processed in standard data−reduction "pipelines". AstroGrid's mission is, in part, to promote the use of these standard reduction in preference to reprocessing of the raw data by individual researchers. However, there will be cases where pre−reduced data are not archived (notably in solar and radio astronomy) or where re−reduction with improved techniques is desirable. Specialist data−processing services are needed that wrap up the data−reduction pipelines for the major surveys and facilities. These will naturally be co−located with the data they process.

Many jobs will require the combining of services. There must be some means of defining this combination as a workflow that states the sequence of operation and the flow of data. Some workflows for common jobs may be coded by professional developers and provided by AstroGrid; in future, many workflows may be invented and coded by end−users using simple tools, possible a graphical programming system. Ideally, each running workflow is itself a service, which is started interactively but then runs unattended. The user should be able to detach from a workflow and later reattach to it to check intermediate results and to steer it.

If data are to be exchanged between services, then they must be intelligible to the receiving services. We need standards for data and particularly for metadata that make the services interoperable.

Some data may come into the system from the end−users' desktops. We need a way of acquiring this data, decorating it with sufficient metadata to make the data intelligible to the system, and making the decorated data accessible to services.

Services may migrate around the grid as facilities change. The details of services may evolve with time. Therefore, client and workflow programmes should not hard−code the location and nature of services, but should refer to a registry of services. We need technology to set up that registry.

Results from services must remain in the data−grid for later use. A user typically wishes to feed the results back into another calculation; or to share the results with collaborators; or to publish the results directly from the Grid. We need a general system for storing and cataloguing results. "MySpace" is its working title.

Some data and services must have restricted access. This applies at least to "personal" results in MySpace and to running workflows, even in the extreme case where all archives have read−only access to public data. In fact, some of the data centres also have access policies to their data that must be respected. We need an access−control mechanism built into the service infrastructure that limits access according to the user's identity and assigned roles. By extension, we must have a way of authenticating that identity that works all across the grid.

These then are the technological needs. The codes in paratheses refer to the technology−choice matrix below.

- A basic invocation system for services. (Inv.)
- A way or representing state in service between invocations. (State)
- The ability to pass metadata to and from services (Pass−metad.)
- The ability to describe details of services in machine−readable form.
- A uniform interface to data archives (Data−sel.)
- A file−transfer capability (FT)
- A DB−transfer capability (DBT)
- A way of caching and publishing data in the grid. (MySpace)
- A facility for registering and discovering services (Reg.)
- Interoperable metadata for data extracts (Interop.)
- Access control for data and resources, allowing delegation of acccess rights by users (IAA)
- A workflow engine (Workflow)
- The ability to detach and re−attach clients to workflows (Detach)

(5.4) Technology choices

(5.4.1) Globus Toolkit (GT) (1)

The ***Globus Resource Allocation Manager*** (GRAM) is the part of GT that enables submission of compute jobs on remote machines: it is the core of Globus. GRAM is intended for pure compute grids. It provides reasonable support for long−running jobs on major processors such as supercomputers, but we find it unwieldy for finer−grained work on smaller installations. GRAM has built−in support for grid authentication of users but inadequate support for authorization. We do not intended to use GRAM; we prefer web services for remote computation.

The ***Metacomputing Data Service*** (MDS) provides a registry of services for use in a GT−based grid. MDS stores for each processing resource metadata listing capabilities, current load, etc. In principle, MDS can store any programmer−defined schema for any resource, and could therefore be the basic for AstroGrid's registry of services. In practice, MDS has not been tested as a repository for very rich schema. Users of GT are now predicting that MDS will be completely replaced in subsequent releases of GT. This means that the current MDS is not useful to AstroGrid, and the replacement will not be available or stable soon enough for AstroGrid's timescales. Therefore, we do not intend to use MDS.

GridFTP is an extension of the commonly−used File Transfer Protocol to support grid security. It also provides some high−performance features that are standard but optional and commonly left out of normal FTP implementations. GridFTP is Globus data−grid support for its compute−grids. We find GridFTP very suitable for our data−grid. It is efficient for large data−files; it has streaming semantics (c.f. HTTP); it allows reference to data on the Grid by URL. We intend to use GridFTP for moving data−extracts between services.

(5.4.2) Other compute−grid kits

Condor (2) is a well−known, mature system for running a compute grid inside one trust domain: typically, the trust domain is one university department. ***Condor−G*** is a recent version that interoperates with Globus. With Condor−G, a Globus grid can treat an entire Condor network as one computing resource (Condor handles the local allocation to queues and processors for jobs accepted from Globus). Alternately, a Condor network can "sub−contract" its excess load to a Globus compute−grid.

Grid Engine (3), from SUN Microsystems, is a proprietary product similar to Condor in scope. It may have better integration with SUN hardware (e.g. their Throughput Engine series of processor farms) and with Solaris. Grid Engine does not interoperate with Globus at the moment.

Neither Condor nor Grid Engine solve any AstroGrid problems. It might be useful for a service provider to run a private grid *behind* a service facade, but this is not really AstroGrid's responsibility. Thus, we have no plans to use Condor or Grid Engine for our own grids.

(5.4.3) SRB (4)

Storage Request Broker (SRB) from San Diego Supercomputing Centre implements a pure data−grid as a distributed file−system. Application software calls a low−level API in SRB to access files and a network of broker daemons makes the file contents available. SRB has many utilities for managing the data−grid by transparent replication and migration of files.

We initially investigated using SRB to build a data−grid that was exposed to applications. When AstroGrid was recast as a service grid, we dropped this idea. In any case, SRB as a data−grid cannot use data in a DBMS. We may now reconsider SRB for the data−grid hidden behind MySpace.

(5.4.4) Web services

Web services come in many forms. The kind that we find most apt are those that work by exchange of messages in virtual envelopes defined by the ***Simple Object Access Protocol*** (SOAP) and pass those messages using the ***Hypertext Transport Protocol*** (HTTP). This is the chosen method for invoking our services. The SOAP envelopes allow arbitrary XML fragments to be transmitted, so SOAP is good for moving our metadata between clients and services.

It is now becoming standard to describe web services using the ***Web Service Definition Language*** (WSDL). We intend to use this to describe some aspects of our services; it will be part of our service−discovery mechanism and our registry. However, WSDL will not describe all aspects of the services that AstroGrid software needs to understand. We have to provide extra metadata.

In e−Commerce, services may be registered and discovered using the *Universal Description and Discovery* (UDDI) system. We have considered using UDDI for our service registry, but we find it too tightly bound to North American business conventions. It is not at all clear that we can express our service metadata in UDDI; we may check this again when the ontological experiments have given us a clearer idea of what needs to be recorded. For now, we do not intend to use UDDI.

The ***Open Grid Services Architecture*** (OGSA) (5) extends normal web−services with facilities to express statefulness of services. In particular, OGSA covers the case where the state of a given service instance is part of a private transaction between a user and the service provider (c.f. the case where all users see the same state of the service and the same changes of state). This is important for a small but significant part of our system. Most services in AstroGrid don't need to remember and express per−user state. However, the MySpace services and the workflow service do need this feature. OGSA defines facilities for producing private instances of services that remember a user's state, and for managing that state (e.g. resources allocated by a service instance can be automatically released and recycled if the user loses contact with the service instance.) OGSA also provides a much−needed framework for using Grid security with web services.

Web−services run in a hosting environment at the server. We have considered four such environments and tested two of them, as described below. We plan to use ***Axis*** (6), from the Apache Foundation, because early support for OGSA is based on Axis; the ***SOAP−Lite*** (7) module from the Comprehensive Perl Archive Network (CPAN) where a service needs to be coded in Perl; and probably *WebSphere* (8) (from IBM) when we have a free choice of hosting environment. Microsoft's ***.NET*** system is a web−service hosting−environment for Microsoft Windows; it is not currently portable to other platforms. Since few service sites run Windows, we have not investigated .NET in detail.

(5.4.5) Data−selection services

The ***OGSA−DAI*** project (9) (OGSA Data Access and Integration) is defining standard interfaces to XML and relational databases. These interfaces will run as web services, probably following OGSA conventions. AstroGrid is an "early adopter" of this technology. We have had considerable input into the early stages of the OGSA−DAI project. Currently, two different kinds of service are being promoted within OGSA−DAI.

The main architectural thrust is to a fine−grained service which exposes to the grid the exact schema of a database and the inbuilt query language (SQL, XPath etc.) of that database. This basic service needs to be matched with translation services for the incoming query and outgoing data−extract that ensure query and results can be written in standard form. IBM (Hursley) have an alternative approach in their "Matrix" prototype. Matrix puts the translation code and the database access into one service, with an internal workflow system to manage the translation stages. Either approach could work for AstroGrid, but in each case we would need to translate to and from formats that are standard for astronomy. Until we know what those formats are, it is hard to assess the products.

Spitfire (10), from EU DataGrid, is a grid−enabled database−access product. It is already in operation for GridPP, and a version with a SOAP/HTTP interface is due out soon. We may have a use for this product, and will investigate the SOAP version when it arrives.

(5.4.6) Access Control

The ***Grid Security Infrastructure*** (11) (GSI) is a system for authenticating identity of users with a Public Key Infrastructure (PKI); i.e. it uses public−key cryptography instead of passwords. GSI is the central and common part of "grid security"; it largely defines what makes a Grid. GSI differs from other PKIs (e.g. in e−Commerce) in that it allows "delegation by impersonation". A service can use this feature to call restricted, subsidiary services in the name of the user. The delegation feature is essential to grids, but is contentious: many commentators consider GSI to be too insecure for e−Commerce and the IETF rules that certificates of identity suitable for use with GSI are technically invalid (this affects all certificates issued by the e−Science certificate authority). For now we are happy to use GSI; it is standard in e−Science, it works well enough for our low−security system, and we need the delegation feature. We expect GSI to change radically or to be discarded as grid technology evolves to support e−Commerce. For that reason, we intend that no other parts of our grid should depend on the fine details of the authentication; we must be able to switch to a new standard when it emerges.

The ***Community Authorization Service*** (12) (CAS) is a database of authorization information. It is intended to replace and expand the inadequate authorization system in GRAM. CAS allows service providers to "outsource" the sociological part of the database − the record of groupings in the community of users − to central authority, leaving the service sites to manage only the allocation of privileges to users. By extension, if users are allowed a limited ability to change the authorization database, CAS can support the ability to share data in the grid by delegating access rights. We like the idea of CAS; it supports the grid metaphor nicely; we like the idea of reducing management workload by reducing duplication; we think that the CAS principle is the way to get the delegation ability that we need. However, the current (alpha) implementation of CAS is unserviceable, and the current design depends too much on GSI. We are looking at a reimplementation of the ideas: see the section on *CoPS* below.

The ***Virtual Organization Management System*** (13) (VOMS), from EU DataGrid, is similar to CoPS but is a finished, working product. We may be able to use VOMS in our production grid.

(5.4.7) Metadata

We express as much metadata as possible in ***XML***. Some existing astronomical systems such as the ***Flexible Image Transport System*** (14) (FITS) will also be retained for compatibility with existing programmes.

Extracts of pixel data will be transported in FITS files. The file−header system built in to FITS will express some basic metadata such as the size and shape of the rasters and the world coordinate systems. It is not yet decided whether other metadata for the extracts will be in the FITS header or in accompanying XML fragements.

Data−extracts that are tabular will be expressed in ***VOTable*** (15) format. VOTable exploits ***Unified Column Descriptors*** (16) (UCDs) to make its tables machine readable, so UCDs are necessarily a part of AstroGrid. We may come to use UCDs in building our registry of services.

UCDs encode semantics of data, but we also need to record the relationships between data: e.g. between the columns of a table. This is the field of ontology; it is a research area, both for AstroGrid and for computer science generally. AstroGrid's experiments to date (17) have used the *Darpa Agent Mark−up Language* (DAML) and the Ontology Inference Layer (OIL).

One precursor of a service grid, "VizieR" (18), uses a system called *Générateur de Liens Uniformes* (19) (GLU) as an abstracting layer. VizieR is extremely successful and there is some motivation to use GLU, or its successor, in AstroGrid. A GLU dictionary could be part of our service registry. Further investigation is needed.

(5.5) Technology−choice matrix

	Inv	*State*	*Pass metad*	*Serv metad*	*File metad.*	*Data sel*	*FT*	*DBT*	*MySpace*	*Reg.*	*IAA*	*Workflow*	*Detach*
GRAM	R												
MDS				R						R			
GridFTP							C		P				
Condor	R		R										

SRB	R		R		R	R	?		?	?			
GridEngine	R												
SOAP	C		C										
HTTP	C		C						P				
WSDL				C									
UDDI										R			
OGSA		P							P		P	P	P
Axis	P												
SOAP::Lite	P												
WebSphere	P												
.NET	R												
OGSA−DAI						P							
Spitfire						?							
GSI											C		
CAS											?		
CoPS											P		
VOMS											?		
XML	C		C	C					P	P			
DAML+OIL				P	P				?	?			
GLU									?	?			
VOTable					C				P	P			
UCDs					C				P				

Key: C = confirmed choice; P = probable choice, pending validation; ? = possible choice; R = technology considered and rejected.

(5.6) Experiments

(5.6.1) Starlink's data−grid

The Starlink project contributed to AstroGrid the results of their experiment "StarGrid". This system includes a pure data−grid, built to test the data−grid technology in the Globus Toolkit.

Starlink's non−networked application software has an abstraction layer for access to files of pixel data. In StarGrid, this layer was altered to call "Globus Access to Secondary Storage" which itself calls GridFTP to access remote files at known locations on the Grid.

StarGrid is a considerable technical finesse, and its data−grid aspects were cheap: they needed little change to existing code. The existence of StarGrid validates the use of GridFTP to build a data−grid.

However, StarGrid does not address the production of data−extracts at the data archives, so does not scale to large data sets. The experiment was not extended to indexing the remote data; the calling software has to know *a priori* the location of each file. Hence, StarGrid is not a complete solution for AstroGrid's data grid.

(5.6.2) Web portal (21)

In 2001, a small data−grid was assembled from WWW technologies. It had a portal on the web by which a user could browse metadata for data−products on the Grid, and could then make informed selection about which data to download. We provided a custom user−interface on client computer in order to get the best possible presentation of the metadata and an acceptable visualization of the data. This allowed us to test ideas about metadata and presentation of results, and to try out the basic concept of a service−grid. (At the time, most grids were seen as compute−grids or data−grids.)

The prototype demonstrated:

- The basic strengths of the service−grid concept.
- The need for uniform interfaces to services.
- The validity of separate channels for data and metadata.
- The need for additional contextual metadata over and above those normally provided in astronomy.
- The need for processing power at various points in the grid (i.e. the system would not have worked had it been a pure data−grid).

(5.6.3) CoPS (22)

The Community Privilege Service (CoPS) is an authorization service: data services can call CoPS to ask if an identified individual has the necessary roles and privileges to carry out requested work.

CoPS was written by AstroGrid. It was inspired by the Community Authorization Service (CAS) from the Globus project and uses many of the same ideas. We could have used the CAS code, but it proved easier for the experiment to build our own sub−system. When CoPS has proven the ideas, it may be best to feed the results back into the CAS product.

The current CoPS experiment is working with authorization structures to show how services can share knowledge about the human and economic structure of the astronomical community. CoPS will allow end−users to modify these structures in a controlled way to share their data with collaborators and the community. To date, the CoPS prototype can respond to requests from users working from a fixed database of authorities, but the code to update the database has not been finished.

(5.6.4) Web services (23)

We are learning to build web services. There is no doubt that this technology works, since it is already widely used, but it is a very flexible technology and the best way of using it is not yet clear. We are building prototypes to determine best practices for AstroGrid.

To date, we have investigated the techniques and tooling for web services in Java, using the Axis toolkit, and in Perl, using the SOAP::Lite module. Simple services have been run sucessfully at Cambridge and accessed sucessfully from Sheffield. We picked Axis not because it is good − it is primitive, unfinished and badly documented − but because initial support for OGSA is based on Axis.

The next step is to place simple services (a trivial "echo" service as a first example) at all AstroGrid sites in order to prove that the relevant technology can run at those sites. The third stage is to publish protoype data−selection and data−transformation services. These will appear initially at Cambridge, Leicester and Edinburgh, but we hope to have useful services on all AstroGrid sites by the end of the year.

We have started an investigation of the OGSA technology preview. We are able to build web services using this equipment, and to make the basic state management work. We have not yet tested the higher functions of OGSA. The OGSA specification is currently being revised and we are waiting for it to stabalise before investing much time in tests.

(5.6.5) Trial grid using the Globus Toolkit (GT2) (24)

We are building a trial grid using components of the Globus toolkit: GRAM and GridFTP. This grid will allow members of the project to obtain and try out their identity certificates, and let us begin building our data−grid using the GT2 version of GridFTP.

So far, the Grid only exists at one site: Cambridge. It will shortly expand to link all AstroGrid project sites.

Findings so far:

- GRAM is not a comfortable or efficient technology; it compares badly with web−services for usability. Its virtue is its intrisic use of grid security.
- GridFTP works well, at least on the LAN at Cambridge.
- Interactions with the UK e−Science certificate authority have been successful so far.

(5.7) References

1. Globus Toolkit: http://www.globus.org/toolkit/default.asp

2. Condor: http://www.cs.wisc.edu/condor/

3. Grid Engine: http://wwws.sun.com/software/gridware/

4. Storage Resource Broker: http://www.npaci.edu/DICE/SRB/

5. Open Grid Services Architecture: http://www.globus.org/ogsa/

6. Apache Axis: http://xml.apache.org/axis/index.html

7. SOAP::Lite: http://theoryx5.uwinnipeg.ca/CPAN/data/SOAP−Lite/SOAP/Lite.html

8. WebSphere: http://www−3.ibm.com/software/webservers/appserv/

9. OGSA data access and integration: http://www.gridforum.org/Meetings/ggf5/pdf/dais/document3.pdf

10. Spitfire: http://hep−proj−spitfire.web.cern.ch/hep−proj−spitfire/doc/index.html

11. Grid Security Infrastructure: http://www.globus.org/security/

12. Community Authorization Service: http://www.globus.org/security/CAS/

13. Virtual Organization Management Service:

14. Flexible Image Transport Service: http://fits.gsfc.nasa.gov/

15. VOTable: http://cdsweb.u−strasbg.fr/doc/VOTable/

16. Unified Column Descriptors: http://cdsweb.u−strasbg.fr/doc/UCD.htx

17. AstroGrid ontology experiments: http://wiki.astrogrid.org/bin/view/Astrogrid/OntologyDemo

18. VizieR: http://archive.ast.cam.ac.uk/vizier/

19. Générateur de Liens Uniformes: http://simbad.u−strasbg.fr/glu/glu.htx

21. Web−portal experiment: http://www.ast.cam.ac.uk/astrogrid/publications/adass−gtr−20010930/

22. Community Privilege Service experiment: http://wiki.astrogrid.org/bin/view/Astrogrid/CASDemo

23. Web−service experiments: http://wiki.astrogrid.org/bin/view/Astrogrid/WebServiceExperiments

24. Demonstration grid: [http://wiki.astrogrid.org/bin/view/Astrogrid/GridDemo

(6) Interoperability Report

(6.1) Interoperability

In the Virtual Observatory (VO) context the word ***"Interoperability"*** is understood to cover all issues relating to the ***combination of the resources*** (databases, software, computational resources, etc) needed to construct a comprehensive VO that appears as a seamless, coherent whole to the user. Interoperability therefore covers the definition and use of standards across a range of areas, some of which are specific to astronomy (e.g. standards regarding the representation of astronomical data in different archives) and some of which (e.g. data transport protocols) are common to all users of the rapidly−developing computational infrastructure of the Internet. The key to interoperability is the collaborative definition and widespread use of agreed standards, and, as described below, AstroGrid has followed this approach in both the astronomy−specific and generic infrastructure arenas during Phase A. AstroGrid members have played an active role in the definition of VO standards (such as ***VOTable*** (1)) and the project has adopted an architecture which reflects current best practice as regards the adoption of relevant standards defined under the aegis of bodies such as the W3C(2) and the GGF(3): the former has been advanced through a series of meetings with colleagues from AVO(4), US−VO(5) and the new International Virtual Observatory Alliance (Strasbourg, Jan 2002; Garching, June 2002; Strasbourg, Harvard and Baltimore, October 2002), while the latter has been informed by participation in a number of meetings organised by the National eScience Centre (NeSC, (6)), and attendance of GGF meetings.

(6.2) Computational Infrastructure Interoperability

As detailed in the Grid Technology section, the core to ensuring interoperability with developing computational infrastructure is the adoption of a ***service−based architecture***, using web services where appropriate, and grid services, on the OGSA(7) model, where "statefulness" is necessary. AstroGrid is committed to this approach, which also seems favoured by the other VO projects (although their more relaxed timescales mean that they have not had to make firm technology choices yet), so that a picture is emerging of the VO mediated by the transfer of SOAP (Simple Object Access Protocol, (8)) messages.

(6.3) Astronomy−specific Interoperability

Given accepted standards for interoperability across computational infrastructure, it behoves the VO community to agree the standards and protocols that are required to integrate astronomical resources into a VO using that infrastructure. Such standards are required in a number of areas, the most important of which are discussed below, namely:

- Resource discovery and the resource registry
- Data archive queries
- Specifying compound VO operations via workflows
- Results returned from VO operations
- Metadata associated with datasets.
- Access to resources

Progress to date on defining them has been good, with one concrete success − VOTable, an XML standard for presenting tabular data in astronomy − and a widespread acceptance that standards must be developed collaboratively within the VO community, with the International Virtual Observatory Alliance (IVOA) as the authority endorsing such developments.

(6.3.1) Resource Discovery and the Resource Registry

VO queries may require a search of a number of data sources which may be located at many different archive sites, and the most general possible VO queries − such as "give me all information that is known about the object at position X" − implicitly assume the possibility of querying more data sources than the user knows about, so there is a clear need for some mechanism for locating all the data sources relevant to a given query. The number of such sources is unlikely to be very large, so it seems appropriate that this function be performed via interaction with an astronomical ***Resource Registry***, rather than the use of something like a WWW search engine.

The need for a astronomical Resource Registry was recognised early on by the AstroGrid Project, and is now agreed by our partners in the IVOA. There seems a good prospect for agreement on a single but replicated Resource Registry which can be used by all VO projects in the world. The times−scales of the other projects, however, appear to be more relaxed than ours, so this is an area in which AstroGrid probably needs to take a lead. There are a number of issues that have to be addressed in the

design of such a registry.

Granularity:
A user may require information about a number of facets of a given data source before knowing whether it is relevant to a particular query, for example:

- Name of service, URL, physical location, etc.
- Web interfaces supported (CGI, ASU, SOAP, WSDL, ...)
- Type of holding (source catalogues, images, spectra, photometry, observing logs, etc.)
- Waveband (radio, IR, optical, UV, X−ray, etc.)
- Sky coverage (see notes below)
- Any access restrictions (e.g. by date of observation, maximum download volume)
- Spatial resolution (of images) or typical positional error (source catalogues)
- Sensitivity (e.g. limiting magnitude)
- Data volume (table sizes, image sizes, etc.)
- Export formats supported (e.g. FITS, VOTable, CSV, PNG, GIF, PS, PDF)

In principle, the registry could contain just the top−level URLs of each data archive site, and all this information could be obtained via repeated querying of the site's web services descriptions (assuming use of WSDL) by the user's portal. This may be inefficient, so we favour a richer Registry, the entries within which may contain all the information listed above; however, there may be situations in which a multi−step interaction between the user's portal and the registry is preferred − see the discussion of Sky Coverage below

The storage of this registry information will clearly require a DBMS of some type, but the data volume will be relatively small, and since the information will all be imported and exported in XML formats, perhaps an XML−based DBMS such as Xindice(9) would be suitable. The registry clearly needs to be a ***replicated*** resource, and it also needs to take account of the fact that many popular datasets are present at a number of points on the web, since this information may be important for a VO query optimiser which can decide which of a set of replica datasets is the best to use for a given query posed by a user at a particular location. The work of keeping a detailed resource registry up−to−date is significant, but if all the individual archives use SOAP for their interfaces and WSDL for their service descriptions, then it might be possible for a robot to update the registry at regular intervals, for example every 24 hours. Search engines use similar methods to maintain their indices based on considerably more heterogeneous collections of web pages, so this ought to be feasible.

Sky Coverage:
One aspect which needs more thought is whether it is possible to store detailed information about sky coverage in the Registry. For surveys the coverage limits are usually fairly simple shapes, limited by declination or galactic latitude, but many observatories will have data arising from a large number of individual pointings, and it is not yet clear what is the most efficient way of making this coverage information available.

This may be one situation in which it is more efficient to have a multi−step interaction between the user portal or the registry and the data archive, rather than having all the relevant information stored in the registry. For example, for something like the HST archive, it may make more sense for the Registry sky coverage entry to be "whole sky (sparse)" with a procedure for querying the observing catalogue in more detail, than to have the Registry entry include a list of thousands of WFPC2 field centres, each a couple of arcminutes in size. Whatever the location of the sky coverage information, it is likely that it will have to be expressed in some ***hierarchical format***, since it is required on a wide range of scales − from, say, a whole hemisphere (in the case of an astronomer wanting to find an optical sky survey catalogue suitable to use in finding counterparts for a northern sky radio survey) to a fraction of an arcsec (for a user wanting target positions for a spectroscopy proposal on an 8m telescope).

Possible technical solutions:

- UDDI: The commercial world's solution to the problem of registering web services is UDDI (Universal Description, Discovery and Integration, (10)), an initiative led by IBM, Microsoft and SAP and advanced within the W3C framework. Whilst solving a somewhat similar problem to that motivating the astronomical resource registry, it does not look as if UDDI (at least in its current form) will be of use to the VO community. This is primarily due to its commercial roots, which mean that it contains hard−wired categories with no relevance to astronomy − for example, "yellow pages" classifications using standard business taxonomies, like the North American Industry Classification Scheme.

- AstroGLU(11): This is a software package created by CDS (Strasbourg) and subsequently used there and by NASA/GSFC. It contains a resource directory with much of the required detail, using the GLU (Générateur de Liens Uniformes, (12)) system for symbolic service names instead of hardcode URLs. In its current form, it has to be maintained by hand, and uses CGI interfaces, so an upgrade to handle XML−based standards and automated updating would be required;

it is thought that a web service upgrade to GLU is in preparation.

- RDF: The Resource Description Format(13) is a W3C recommendation for structured metadata and may form a good basis for future work, but as yet the work seems not to have progressed as far as having a database of RDF information.

These and other possibilities will be pursued further in Phase B, in collaboration with our IVOA partners.

(6.3.2) Data archive queries

Existing data archives provide many common features typically using a CGI−based query mechanism. The AstroGLU system from Strasbourg has a translator from a uniform set of CGI parameters into the terms needed by a number of major archives, but this is a work−around rather than a standard and is not very scalable. There now seems general agreement that we need to move towards a XML/SOAP/WSDL interface which can be implemented at each archive site. The US−VO has put forward an interim standard for a ***cone−search*** (14) web service (finding all sources in a cone of small angle around a given celestial position) and a number of prototypes have recently been set up. AstroGrid is in the process of setting up compatible services at Cambridge, Edinburgh and Leicester.

Less progress has been made on a standard for more advanced queries. Most if not all data archives use a DBMS which speaks SQL, but the SQL standard is both inadequate and poorly implemented in practice. Nearly all of the queries used in our DBMS evaluations (covered in the Database Technology and Data Mining report) had to be modified to suit the different DBMS under test. In addition, functions that are going to be frequently used, such as that for great−circle distance between points on the celestial sphere, need to be encapsulated in the query language, and the implementation of user−defined functions is very DBMS−specific. There is widespread agreement that some form of ***Astronomical Query Language*** (AQL) will be needed, with translators for the several SQL implementations used in archives around the world, but no standard has yet been developed. More work is needed on this in Phase B. Fundamentally, what is required is a standardised data−selection service, that provides a translation between the peculiarities of individual databases (both different DBMSs and different instances of them in different data centres, with different schemas) and some Grid−friendly format.

(6.3.3) Specifying compound VO operations via workflows

Querying a database is only one class of VO operation. If the VO maxim of "ship the results, not the data" is to be followed, then it must be possible to construct ***compound operations*** to be run within the VO; for example, a query on a set of databases, followed by running some analysis algorithm on the set of results, after they have been shipped to a common location. The construction of ***"workflows"*** like this is a common requirement within e−science, and in computing generally, so it likely that AstroGrid (and the VO) will not define the necessary interoperability standard in this case, but rather adopt an existing standard which can be made to work for the particular case of astronomical operations. That said, initial experiments are being conducted within AstroGrid to use XML fragments in SOAP messages to provide inputs for complex services − for example, supplying sets of input parameters with which to run some piece of code − to start to assess the workflow functionality required for the VO.

(6.3.4) Results returned from VO operations

The most obvious and serious weakness of existing web−based systems of querying multiple data archives has been the heterogenous nature of the results which are sent back. The VO community therefore recognised the need for the ***standardisation of results formats*** at an early stage, and this process was launched with the definition of an XML standard for tabular data, called ***VOTable***, which has now been endorsed by the IVOA. AstroGrid members were fully involved in this activity, which was undertaken through a round of email debates, followed by a meeting in Strasbourg in January 2002 to hammer out the final details required for release of the specification for VOTable version 1.0. Clearly, there is a need to repeat this procedure to produce agreed standards for the other data formats required in the VO (e.g. pixel data) and initial discussions about these are starting to commence within the VO community.

The use of XML as the basis for VOTable was motivated by several considerations. Firstly, XML is the *lingua franca* of the web services world, so its use would aid the interoperability of the VO and external computational infrastructure. Secondly, use of XML enables applications to ***validate*** a VOTable document readily, using standard rules, which is something that FITS cannot do so readily, as many people who have written pipeline reduction systems driven by FITS headers can attest. Thirdly, the existence of the XSLT (eXtensible Style Language Transformation, (15)) standard means that result sets in VOTable format can be easily transformed into other formats, as required.

Another important interoperability feature is that VOTable requires not only the tagging of all physical quantities with their ***units***, but also the use of Uniform Content Descriptors (***UCDs***, (16)), which express the nature of the quantity. The set of ~1500 UCDs were abstracted from the column names of the ~3000 tables included in the VizieR(17) system at CDS, and they provide a means of recognising synonyms, as well as a preferred taxonomy for use in astronomical databasing. The derivation of the UCDs from VizieR means that the current set of UCDs bears the same biases and selectivities as the VizieR system itself. So, one of the tasks of the radio astronomers within AstroGrid and AVO has been to develop the additional UCDs required for interferometry data. Similarly, as described in the Pilot Programme Report, the solar and STP pilots assessed the utility of VOTable in their respective areas, noting that, while additional UCDs would have to be defined before it could be used, it should be suitable for their requirements once they have been.

One criticism frequently levelled at XML in relation to its use in astronomy is that its inherent verbosity means that data files in XML format are many times larger than they would be in, say, FITS binary format. VOTable provides a means of circumventing that, however, as it allows metadata and data to be stored in separate files, but linked according to the Xlink model. This has many advantages − for example, processes can then use metadata to 'get ready' for their input data, or to organize third−party or parallel transfers of the data − but it is somewhat unsatisfactory to have to employ two methods to manipulate a VOTable document: an XML parser for the metadata and some less standard tool to handle the binary file. There are XML schemas for binary data being developed − for example, the Binx(18) proposal developed as part of the OGSA−DAI(19) project − and it is possible that one of these can be applied to extend VOTable in such a way that an application can manipulate a tabular dataset of any size, without needing to know whether it is all written in XML or has a binary file linked to it.

(6.3.5) Metadata associated with datasets.

The FITS Standard is very widely used and allows an unlimited amount of (scalar) metadata, but it is recognised that conventions for its use are inadequate and AstroGrid members are active in the discussions within the VO community concerning the definition of the metadata systems required for the VO. One area where FITS is clearly inadequate is its treatment of ***provenance*** information. The FITS standard allows for any number of "HISTORY" tokens to be placed in the FITS header to record the provenance of the contents of the file's data units, but there is no convention for writing them and so the information they contain is not readily extracted by any means other than being read by a human. This is clearly inadequate in the era of the VO, in which it is desirable to have ***machine−readable metadata***. For example, AstroGrid's MySpace concept allows for the storage of the results of VO operations, which could be complicated workflows as well as single database queries, and it would be very useful to be able to interrogate these result sets regarding their provenance, so that, for example, a complicated analysis does not have to be repeated if it has already been performed.

(6.3.6) Access to resources

Interoperability standards are also necessary to provide access to resources in the VO. The ability to access to computational resources at distant sites is the very essence of the Grid, so this is not an area where AstroGrid expects to develop the fundamental standards, but rather to adopt them. However, it is essential for the VO community to develop a consistent way of using these standard protocols (such as the Community Authorization Service described in the Grid Technology report) to implement an agreed policy for access controls in the VO. For example, most observatories impose proprietary periods on data, and, even if all agree to implement these using some Grid−standard digital certificate protocol, there must be agreement about how the existence of data access restrictions is made apparent within the VO and about exactly what credentials must be included in the digital certificate to effect access to a given resource.

(6.4) Ontology

Cutting across the whole area of Interoperability is the concept of ***"ontology"***. In this context an ontology is an explicit formal specification of the terms in a domain and the relations between them, and the reason that ontology is of relevance here is that it can provide the conceptual framework to ensure that interoperability is implemented in a meaningful way. At some level, the VO will have an ontology, whether implicit (assuming astronomers' common knowledge) or explicit, and, for example, the set of UCDs form some sort of an ontology, since they define the relations between the column names in the VizieR system.

A more all−encompassing ontology could be used within the VO, however, and AstroGrid staff are investigating this possibility, both in conjunction with colleagues from other VO projects and with ontology experts from other disciplines, notably the bioinformaticians in the myGrid(20) project (with which AstroGrid already has a collaborative relationship, by reason of these two projects being the chosen "early adopters" of OGSA−DAI deliverables).

(6.5) References

(1) VOTable: http://cdsweb.u−strasbg.fr/doc/VOTable/
(2) World Wide Web Consortium (W3C): http://www.w3.org
(3) Global Grid Forum (GGF): http://www.gridforum.org
(4) Astrophysical Virtual Observatory (AVO): http://www.eso.org/avo
(5) US Virtual Observatory project: http://www.us−vo.org
(6) National eScience Centre (NeSC): http://www.nesc.ac.uk
(7) Open Grid Services Architecture (OGSA): http://www.globus.org/ogsa
(8) Simple Object Access Protocol (SOAP): http://www.w3.org/TR/SOAP
(9) Xindice: http://xml.apache.org/xindice/
(10) Universal Description, Discovery and Integration (UDDI): http://www.uddi.org
(11) AstroGLU: http://simbad.u−strasbg.fr/glu/cgi−bin/astroglu.pl
(12) Générateur de Liens Uniformes (GLU): http://simbad.u−strasbg.fr/glu/glu.htx
(13) Resource Description Framework (RDF): http://www.w3.org/RDF
(14) US−VO cone−search: http://us−vo.org/metadata/conesearch
(15) Extensible Stylesheet Language Transformations (XSLT): http://www.w3.org/TR/xslt
(16) Unified Column Descriptors (UCDs): http://cdsweb.u−strasbg.fr/doc/UCD.htx
(17) VizieR: http://vizier.u−strasbg.fr/viz−bin/VizieR
(18) Binx: http://www.epcc.ed.ac.uk/~gridserve/WP5/Binx
(19) Open Grid Services Architecture Database Access and Integration (OGSA−DAI): http://umbriel.dcs.gla.ac.uk/NeSC/general/projects/OGSA_DAI
(20) myGrid: http://www.mygrid.org.uk

(7) Database Technology and Data Mining

(7.1) Introduction

In the commercial world almost all large data collections are stored within database management systems (DBMS), but in astronomy DBMS are used very little except in the management of data archives. Even within astronomical archives the bulk of the data are generally stored in external files (such as FITS files) with just the filenames stored in the database.

The aim of work package A4 was to investigate whether modern DBMS technology could be used more widely in large−scale astronomical data management, especially in view of the data avalanche from new instruments. In particular we wanted to investigate the value of DBMS for what is loosely called *data mining* − the attempted discovery of valuable scientific information buried in large data collections such as from sky surveys, or from sets of observations originally gathered for some other purpose.

Ideally the astronomer will make use of the virtual observatory by posing *questions* and getting *information* back, but all we can reasonably expect is to allow users to submit *queries* and get *results*. The translations between the scientific to computing domains are likely to require at least some astronomical expertise for the forseeable future. The forms of queries and results are inevitably a compromise between what the astronomer wants and what data archives can handle. In order to understand the requirements better, we have analysised a large number of potential scientific problems and use−cases.

(7.2) Use−cases

The AstroGrid project agreed at an early stage to be use−case driven, and this analysis of database requirements draws heavily on our prior experience of using and managing astronomical data archives, and specifically on:

- AstroGrid's set of (so far) 47 Science Problems set out on our Science Problem List, which were expanded to some 68 more basic use−cases;
- The US−NVO's collection of 100 Science Questions for the NVO which so far contains 29 use−cases ;
- The 35 queries described in the document Data mining the SDSS Sky Server Database by Jim Gray and colleagues.

These use−cases were further deconstructed into the basic data management operations.

It should be noted that the following classification is set out in terms of extra−solar astronomy, in which celestial position (usually specified as a pair of spherical−polar coordinates) is the most important indexing parameter, but a similar classification can be made in solar physics (in which heliocentric coordinates play a similar role), and in STP where event time is usually the primary data locator.

In database terms, there seem to be two main classes of query:

- Positional queries: where the user wants information about a small patch of sky around a given celestial position, or about a named celestial object (which can be converted to celestial coordinates using a name resolving service such as Simbad or NED).
- Non−positional queries: essentially everything else. This includes queries which require a sequential scan of all (or much of) large datasets, or which involve I/O−intensive joins between two or more tables.

(7.2.1) Positional Queries

Positional queries are especially important in astronomy since many astronomers spend long periods investigating one celestial object in detail. Indeed such queries are so common that most archives have been set up to handle them optimally or even exclusively. Positional queries are also special in database terms because they can be handled efficiently by indexing on celestial position, (although in practice two−dimensional indexing presents some interesting problems). Although in theory the user of a database should not need to know which parameters have been indexed, in practice the sequential scan of a billion−row table may take hours, while and an indexed look−up in the same table will take no more than tens of milli−seconds, so indexing cannot be ignored as a mere implementation detail.

Positional queries can be further subdivided based on the main types of information stored in astronomical data archives:

- Queries of source catalogues, for example:

    - ♦ Are there any radio sources near to HD123456?
    - ♦ Has anyone measured the B magnitude of 3C678 between 1990 and 2000?
    - ♦ Can I have a list of all known objects within 2 arc−minutes of PSR1234−567?
- Queries of image repositories, for example:
    - ♦ Can I see an IR image of M87?
    - ♦ Is there an X−ray image of the sky around (12:34:56,−45:59)?
- Queries of specific observatory archives, for example:
    - ♦ Show me the results from the IUE observation of Beta Hydri?
    - ♦ Can I download the raw data from the XMM−Newton observation of NGC9876 to re−analyse it myself?
- Bibliographical queries. No work has been done on this area, because existing services such as Simbad and NED already do an excellent job.

Source catalogues are essentially tabular datasets listing the positions, fluxes, and other properties of all the astronomical objects detected in some area of sky in some waveband. The larger ones currently have nearly a thousand million rows and a few have over 100 columns. Some arise from the systematic survey of large areas of the sky; others are produced by analysis of the sources detected serendipitously in images of the sky obtained in other programmes.

Images: Collections of images of the sky are another important astronomical resource. The total volume of data in an image archive may be large: for example sampling at 1 arc−second resolution with 4 bytes/pixel one gets 2 terabytes for the whole sky, but the individual images are usually only of modest size, and the indexing requirements are also easy to meet. Although most DBMS can store images as binary large objects (BLOBs), in practice almost all archives have chosen to store images as external files (mostly as FITS images), to simplify interworking with other data analysis software.

Observatory archives are repositories of raw data or semi−reduced data. They are the province of the more expert users, seeking to download data for further reduction and analysis on their own computers. The access is usually via the observation log, a table of modest size usually indexed by celestial position. These logs are usually fairly small (under a million rows), so indexing is simple. The main complication is that the telescope field−of−view is often a complex shape, so determining whether a given celestial position was in the field−of−view or not may require fairly complex calculations.

To summarise: positional queries are generally handled fairly well by current archives, although many of them still depend on inefficient methods of sky−indexing (more on this below). But they do present a problem as far as data mining is concerned because of the variety of conventions they use for submitting queries and the variety of data formats for the data that they return. The standards being devised (described further in the interoperability chapter) will, we hope, result in these archives being accessible as Web Services, and so easy to integrate into an overall data mining infrastructure.

(7.2.2) Non−positional queries

Non−positional queries are those typically involving the study of a class or group of astronomical objects. Astronomers carry out such a wide variety of operations on the data retrieved from archives that it is virtually impossible to provide an exhaustive classification, but the following types of operation appear to be among the most important:

- Cross−identification of sources from two (or more) source catalogues on the basis of positonal match and perhaps other criteria. (Note: a join is, of course, just a series of positional queries, but since an outer−join requires a complete scan of the first table, it really belongs in this category)
- Selection from a source catalogue by reference to the properties of the sources, for example finding all stars in a stellar catalogue with a particular spectral type and range of proper motions.
- Statistical estimations and searches: for example finding all sources with a colour index which is more than 3−sigma above the mean for that table.
- Data mining queries of more advanced types, these include using:
    - ♦ Classification and clustering algorithms to find groups of objects with similar properties;
    - ♦ Regression analysis: finding combinations of properties which are significantly correlated;
    - ♦ Sequence analysis: carrying out time−series analyses to find periodicities, bursts, or other temporal anomalies;
    - ♦ Similarity or dissimilarity searches, for example analysing sets of images or source lists from different epochs to find objects which have moved or changed in strength.
    - ♦ Measuring properties of the sky on the large scale, using for example power−spectral densities.

(7.2.3) Cross−identifications

The cross−identification of sources from different source catalogues is an important basic operation often the precursor to more advanced data mining investigations. The operation occurs in most of the "top ten" AstroGrid science cases. This requires, at its simplest, what is often termed a fuzzy−join, since the source coordinates in each catalogue will have associated errors, so an approximate match within limits of error replaces an exact match, and one needs to compute the great−circle distance between points expressed in spherical−polar coordinates (RA,DEC).

Many existing data archives just use an index on one spatial coordinate (sometimes RA, sometimes declination) which does not produce an efficient or scalable solution. What is really required is a spatial index, for example one based on the R−tree, but but few DBMS have spatial indexing built−in, and the few which have them do not cope well with the singularities of the spherical polar system.

A new solution to this problem was discovered as part of the AstroGrid Data Mining investigations, based on covering the sky with a grid of approximately equal−area pixels, and using a one−dimensional index on these pixel values. The PCODE algorithm solves the problem cases in which error−circles overlap two or more pixels by inserting additional rows in the tables, and using the SQL `DISTINCT` function to remove duplicates from the resulting table of matches. The method can be used with any suitable pixelation of the sky, for example the Hierarchical Triangular Mesh (HTM) or Hierarchical Equal Area iso−Latitude Pixelation (HEALPix) algorithms, but in our tests only the latter was used. This PCODE method has the significant merit of only requiring an equi−join on integer fields, which occurs so often in commerce it is well optimised by all DBMS, and only needs simple B−tree indices on the PCODE columns. It can also be used with an outer join to determine sources in one table unmatched with those in another, which is an operation often scientifically useful.

The use of a uniform sky pixelation, such as HTM or HEALPix, will also permit existing data achives to implement a much more efficient sky indexing method; at present many of them just utilise a B−tree on one of the two spatial coordinates, a solution which will get more inefficient as datasets get larger.

It is important to note that source identication often depends on more than just positional coincidence (see this note on association methods) but the other operations are relatively fast once the fuzzy−join has been carried out.

One of the DBMS in our evalution, Postgres, has R−tree indexing built in. We therefore used it to compare the traditional R−tree solution with the PCODE algorithm using identical hardware and software. This showed the PCODE join to be about twice as fast, and while using substantially less disc space. Index creation was speeded up even more. Since the PCODE algorithm can be used with any DBMS, not just the few which support spatial indexing, this significantly improves our freedom of choice.

(7.2.4) Selection Operations

Simple `SELECT` operations are easy to do in all relational DBMS, but the speed with which they can be carried out depends entirely on whether a suitable index exists. Since many astronomical tables contain a large number of columns, it is not always easy to arrange this. For example the source catalogues produced from the Hubble Deep Field contain 123 columns. Even worse, astronomers often want to select on some expression involving multiple columns, such as a magnitude difference, or a flux ratio. In a wide table there could be thousands of possible simple combinations, and it is impractical to maintain indices on them all.

(7.2.5) Statistical Operations

Most DBMS support simple statistical functions such as mean and variance, but SQL is not an ideal language for statistical work. Ideally the entire table needs to be scanned to measure its statistical properties, but in practice a carefully−selected sampling scheme may speed up some operations. Nevertheless statistical operations are inherently slow, being mostly unable to benefit from the presence of indices.

(7.2.6) More Advanced Operations

Many of the more advanced operations will require the use of specialised software packages. If the data are stored within a DBMS this means either exporting them in some format (e.g. as a plain ASCII file, or preferably in something XML−based such as VOTable), or accessing the DBMS through some API such as ODBC, JDBC, JDO. We hope that current projects such as OGSA−DAI and Spitfire (from CERN) will produce even more powerful grid−enabled interfaces. But whichever option is chosen, many types of data mining operation will require sequential reading of substantial parts of large tables.

(7.3) Data Exploration and Data Mining

Just as in the real world of exploiting mineral resources, before any nuggets of astronomical value can be extracted from the mass of data, the landscape has to be explored thoroughly: indeed in science the term *data mining* is usually shorthand for both the exploration phase and the phase in which bulk data are sifted.

Ideally a data exploration and mining facility would give the astronomer the ability to access any dataset on the web without needing to copy it explicitly to a local machine. It is interesting to note that software based on the well−known FITSIO library, such as the FTOOLS collection from HEASARC and the CATPAC package from Starlink have supported transparent remote access for some years, as both FTP and HTTP protocols are built in to the lower layers of the library, while DBMS packages normally require all tables to be local, and they have to be explicitly imported to a given database. A few DBMS have very limited support for access to foreign files, but usually only to tables stored in their own format, or sometimes in the databases of one of the rival vendors, and only when the relevant clients have been installed.

Many of the functions needed in data exploration and mining are those that any DBMS supports, such as selecting data, sorting it, grouping, finding means and extrema, joining with other tables, projecting new columns, etc. Other essential operations can also be done in SQL−based systems but only with rather more difficulty, for example finding the median and other quantiles, finding outliers, computing trend−lines and regressions, computing statistics, etc. Specialised statistical packages support many of these operations rather better, but all of them seem to be based on memory−resident datasets, which seriously limits their use on large astronomical datasets. Astronomers will also need graphical output and simple visualisations of their datasets, e.g. 1−d and 2−d graphs, histograms, contour maps, etc. and ways of overlaying source positions on images. These are also well beyond the facilities of any current DBMS. There are many visualisation packages which have most of the right functionality, but they all have their own data formats, so data conversion problems abound, and metadata loss is usually inevitable on format conversion. These problems have no easy solution. The problem of preservation and propagation of metadata is especially acute when using a relational DBMS, and none of them seem to preserve any attributes of a column or a table, beyond the most basic (name, and data type).

Our conclusion is that if a DBMS forms the core of a system designed to support astronomical data mining, it will be just one of a number of software packages installed, because of breadth of functionality required by astronomers is so large.

It is perhaps worth noting the international dimension: although the UK is producing many datasets of world class which will be important for astronomical research over the next few years, many other important datasets are resident overseas. Almost every one of the important science cases depends on the use of one or more foreign data collections.

It has also become obvious that many operations will require access to remote datasets which the remote systems are not prepared to support. Obstacles and limitations include:

- The necessary software will not always be installed on the remote system.
- Some operations require scans of all or substantial parts of a large table, and some algorithms are not only I/O−intensive but also cpu−intensive.
- Astronomical institutions are noted for putting a high proportion of their data on open access, but this does not necessarily mean that they also provide facilities for compute−intensive use. The majority of existing data archives restrict the amount of work that can be done, either by limiting the operations to those which are utilise the available indices, or by putting restrictions on cpu time or the volume of downloaded data. Such limits are entirely reasonable, since these facilities are currently open to the world, and unlimited access could easily be misused. It is to be hoped that the general adoption of an improved security infrastructure, based on the authentication and authorisation mechanisms currently being explored in the e−science community will allow these restrictions to be relaxed gradually. But for the foreseeable future it must be accepted that astronomers will often want to perform operations on remote datasets which the remote systems themselves cannot support.
- Some operations will require a network bandwidth which is not yet available to the remote site, and network bandwidth is growing more slowly than either available processing power or disc space.
- Some operations will be feasible only on links with low latency. At the recent "e−Science All Hands Meeting" in Sheffield the panellist from Microsoft Research, Dr Andrew Herbert, cited latency as the major obstacle to grid−based data−intensive operations, and no amount of technology will defeat the speed of light.

(7.3.1) The Astronomical Data Warehouse

In order to overcome these obstacles, we think that the UK astronomical community will want to set up a small number of data centres which we propose to call "data warehouses". The term is borrowed from the commercial world, where many large

organisations have found it necessary to set up specialised facilities for analytic data processing and data mining, containing essentially static copies of datasets from elsewhere in their organisations.

The astronomical data warehouse has many elements on common, but differs in detail. We envisage a data warehouse as containing:

- A powerful computer system, for example a Beowulf cluster;
- Good network connections;
- Local copies of the more popular astronomical datasets;
- Ample temporary disc space for additional datasets to be downloaded on demand.
- At least one powerful DBMS;
- A fairly comprehensive set of standard astronomical data processing packages.
- Grid software components at the very least for authentication and authorisation, and for efficient data transfer.

Clearly an Astronomical Data Warehouse will not be able to get copies of external datasets except with the permission and indeed cooperation of those responsible for them, but based on past experience, we expect no difficulties here. Indeed most sites seem fully prepared to make copies available to other sites who will in turn make them available to the astronomical community.

Although the term "warehouse" may be new, the concept is not much more than an obvious extrapolation of a trend which has been apparent for some years. Many of the most important astronomical data collections have at least one mirror site somewhere else in the world, and some of the most popular ones, such as Vizier, exist in more than half−a−dozen locations. Powerful data mining centres will be needed anyway for new archives such as those from e−MERLIN, XMM−Newton, WFCAM, and VISTA, and since all these datasets will be used in conjunction with those from other sites, the warehouse concept will follow fairly naturally. We propose to set up a dedicated AstroGrid data warehouse only to develop and prove the concept: we expect that the major data centres in the UK will find it worthwhile to provide data warehouse facilities, so that several will soon be set up within the UK. It will obviously than make sense to connect these separate warehouses to using data grid techniques, and in time it might become a seamless data mining facility for the UK community.

(7.3.2) MySpace

The MySpace concept is that authorised VO users should have a semi−permanent storage area allocated to them on the virtual observatory − no doubt some of this will physically reside on various astronomical data warehouses. The facilities will allow the user to *publish* fully reduced data and make it available to the rest of the world.

Neither of these concepts is yet fully defined, and their development and implementation will be undertaken during Phase B.

(7.4) DBMS Evaluations

Our DBMS evaluations were undertaken, not in the expectation that it would be possible to identify a single outstanding product, but in the hope that we could eliminate at least a few possibilities on the grounds of serious incompatibility with our general requirements.

(7.4.1) Object−oriented DBMS

The properties of the OO−DBMS look well−matched to the requirements of astronomical data centres in many ways, as they allow a schema to be exactly matched to the complex data structure, as often arises from astronomical observations. They also offer the prospect of greater efficiency, as many operations requiring a cpu−intensive join can be performed merely by following pointers. Unfortunately no OO−DBMS seems to have gained significant market share. Our experience is also that that few standards exist for interfaces and queries, and support for them is poor. We have links with two astronomical projects which have tried OO−DBMS with unsatisfactory results:

- The XMM−Newton Survey Science Centres (at Leicester, MSSL, Strasbourg) adopted O2 for both the data pipeline management and the subsequent archive; after successive take−overs of O2 by Unidata, Ardent Software, Informix, and finally IBM, the product is essentially defunct, and maintenance has become extremely expensive.
- The SLOAN Digital Sky Survey data centres adopted Objectivity/DB for their data pipelines and archive access, but after experiencing poor performance and unsatisfactory software support, they have switched to using SQL Server instead.

As a result we have only taken an interest in relational DBMS (although most of them now claim to be object−relational, a term which has no very precise definition).

(7.4.2) Relational DBMS

We chose five DBMS for examination, because of their existing usage in astronomy or their known characteristics:

- DB2 (from IBM) − heavyweight commercial product with reputation for good standards−conformance.
- MySQL − open source product with a reputation for speed; used by many existing astronomical data archive sites; now supports transactions but still has less complete coverage of SQL than the others.
- Oracle − leading heavyweight commercial product.
- Postgres − open source product with multi−dimensional indexing using R−trees.
- SQL Server (from Microsoft) − full featured and easy to use, but only available for Windows platform.

(7.4.3) Results

Detailed results from our evaluations to date may be seen on the wiki but the principal finding are summarised below.

- No complete show−stoppers were found, although a number of unexpected obstacles to easy use by astronomers were uncovered.
- None of the DBMS provided a good way of importing bulk data from, or exporting to, binary files − the use of text files wastes bandwidth, cpu time, and disc space, and risks losing precision and metadata.
- All DBMS supported indexed operations very efficiently, but were slower than expected on sequential scans of large datasets (more on this below).
- All had similar performance, although because of different platforms used we were only able to compare MySQL and Postgres directly; here MySQL was often twice as fast as Postgres, and much faster in a few special cases where results were cached.
- None of these DBMS implemented full standard SQL92, despite having now had 10 years to do so. Almost every one of our test queries had to be modified when moving from one DBMS to another. This underlines the need for a single astronomical query language (AQL) which is translated into the dialect of SQL handled by each DBMS. A recent Sourceforge project called Liberty Database Connectivity (LDBC) has much the same idea, with a translator from a standard query language to the appropriate dialect of SQL, but at present this package does not support the trigonometric functions that an AQL will need.

We have not yet had time to evaluate any of these products on parallel hardware; the necessary hardware has just been installed, and work will start very shortly.

Sequential scans of large tables or substantial parts of them are bound to occur in any data mining facility, for a number of reasons already listed, although one would obviously try to avoid them whenever possible by creating of suitable indices.

Even if sequential scans only form a small proportion of all database accesses, the fact that they tend to take thousands (if not millions) of times longer, means that they may dominate the overall time for a typical sequence of data mining operations. Hence we felt it necessary to evaluate the effective speed (or bandwidth) of these DBMS when doing simple operations involving sequential access to a column of a large table. The results of computing the mean and standard deviation about the mean for a single column in a medium−sized table were surprisingly slow, when expressed in bandwidth terms. (Note: the figures for Postgres are as installed "out of the box" − some scope remains for tuning).

Software package	**Data Bandwidth** (Mbyte/s)
Postgres DBMS	0.3
MySQL DBMS	1.9
FITSIO row−orientated	7.0
FITSIO column−orientated	21.2

The last two rows report results using the same table (A 3.5 million row sample of the USNO−A2 catalog) converted to a FITS binary table, with a simple custom application to compute the simple statistics (there is an FTOOL which does this, but it computes more detailed statistics so would not have been exactly comparable). The last line reports an experiment in which

the FITS file was set up with columns adjacent, rather than rows (as suggested by Prof Peter Buneman of Edinburgh). This was indeed even more efficient, and gets fairly close to the raw speed of the disc drive. We do not have exactly comparable figures for the commercial DBMS we tested, but they appeared to lie between the figures for Postgres and MySQL.

(7.4.4) Conclusions

A DBMS is needed to power the data warehouse, and to manage data within the MySpace domain.

Since our evaluations are not yet complete we have made no decision, but it is clear that no relational DBMS is well suited for the manipulation of scientific data, and that any of these products would, at a pinch, be usable. We may reach a different conclusion as a result of testing these DBMS on a clustered system, tests which are just about to start.

(8) Pilot Programme Report

(8.1) The AstroGrid Pilot Programme

The AstroGrid Pilot Programme was designed to complement the Phase A technology assessment workpackages. While the latter were intended to evaluate some of the technologies likely to be deployed to help meet AstroGrid's science requirements, the Pilot Programme was designed to help elucidate those requirements in more detail, by developing prototype systems to deliver small portions of AstroGrid's desired functionality. These prototype systems were intended to be constructed using technologies that staff were familiar with; the accent here was on learning from the experience of producing pilot software to meet aspects of AstroGrid's needs, without necessarily doing so in the way that AstroGrid would ultimately meet them. An important goal from the outset was that all pilots should proceed to the point of delivering software that could be used by test users to do real science, but, equally, it was stressed that no undue effort should be expended in producing pretty user interfaces, etc, as the lessons to be learnt concerned the delivery of new functionality, not its presentation to users.

(8.2) Pilot Selection

A set of five pilot topics was selected, one each for the five broad fields covered by AstroGrid, namely

- optical/near−IR astronomy: *Large object catalogues*
- X−ray astronomy: *Association methods*
- radio astronomy: *Fourier data*
- solar physics: *Data selection from summary information*
- solar/terrestrial physics (STP): *Time−series data*

The reason for this selection was two−fold. Firstly, each of these areas has specific requirements, and by selecting a pilot topic of particular importance to each of the five areas we could help ensure that the final AstroGrid system meets the needs of its full user community. Secondly, it was intended that the test users for the pilots would be selected from the likely "early adopters" of the final AstroGrid system, so, by covering all disciplines, it was hoped that the Pilot Programme would start to engage appropriate members of all communities in the work of AstroGrid.

The five pilots are described in turn below, together with an ***Example Science Case*** for each, illustrating the type of research problem each is intended to address.

(8.2.1) The Optical/Near−IR pilot − *Large Object Catalogues*

Example Science Cases:
A scientist wishes to search for halo white dwarf stars, which requires selection criteria making use of both colour and proper motion information for a large sample of stars (for these are rare objects). This can be achieved by querying the multi−epoch/multi−colour dataset produced by federating the Sloan EDR dataset with the SuperCOSMOS Sky Survey coverage of the same region.
A scientist wants to determine the optical and infrared colours of an object, e.g. an X−ray (XMM) or radio (FIRST) or far infrared (ISO) source. This can be achieved by querying the multi−colour dataset produced by federating the INT WFC five colour optical dataset with the 2 colour INT CIRSI dataset of the same region.

The optical/near−IR object catalogues to be included in the AstroGrid system greatly exceed all its other databases in size, so there was a clear need for a pilot that addressed the practical problems of federating large object catalogues. Initially, it was intended to study this using two methods. Firstly, data from the SuperCOSMOS Sky Survey (1) covering the fields of the Sloan Digital Sky Survey (2) Early Data Release (EDR) would be federated with the EDR data themselves, using a version of the SDSS science archive software (called SX, (3)), modified for use with the SSS data model. Secondly, catalogues of objects derived from INT imaging with the Wide Field Camera (optical) and CIRSI (near−infrared) would be federated by making them both accessible via the VizieR (4) system developed by the Centre de Données astronomiques de Strasbourg (CDS (5)). The functionality provided by the two approaches could then be compared by providing access to both federations to test users, who can assess how well they each match the needs of scientists using future federations of large object catalogues in the VO. In the event, the evaluation of VizieR in this regard was performed by AVO (6)Work Area 2 (Interoperability), not as part of this pilot: the results of this work will be reported elsewhere and we shall only consider the SSS−SDSS federation in what follows.

(8.2.2) The X−ray pilot − *Association Methods*

Example Science Case:
A scientist wishes to determine the variation of some X−ray hardness ratio as a function of X−ray flux and optical/near−infrared colour, to constrain models for the properties of the population of obscured AGN. To do this requires the association of objects in X−ray and optical/near−infrared catalogues, followed by the selection of subsamples of associated sources on the basis of X−ray properties.

The association of entries in different databases identified as being observations of the same astronomical object lies at the heart of the Virtual Observatory (VO) concept, but it appears to have received little attention from the VO community to date. Astronomical data has a natural indexing − spatial location in the sky − which aids the making of such identifications, but, in many cases, association by spatial proximity alone is not adequate. This is most clearly the case in situations, such as the determination of optical counterparts for infrared sources from ISO or submillimetre sources from SCUBA, where the angular resolution of one catalogue is so poor, and the surface density of objects in the other so high, that there can be many objects from one catalogue located within the positional error ellipse of each source from the other. In this case, the easy identification of the true optical counterpart is not possible, and a probabilistic method must be used, to assess which of the candidate counterparts is the most likely to be the true match. Astronomers already use several such probabilistic prescriptions, but they are frequently used in situations in which each possible association can be checked manually for plausibility. The VO provides a much greater challenge. Not only does the size of the datasets to be made available by the VO mean that associations could be sought between databases containing millions of objects each, raising concerns about the scalability of the assocation techniques commonly used, but there are also issues relating to how the method employed to make a set of associations can be recorded so that a later user can judge whether s/he can employ them with confidence, rather that recomputing them all anew. This pilot was designed to start addressing these important issues, through seeking optical counterparts for sources in the first XMM−Newton catalogue, produced by the XMM−Newton Survey Science Centre. This catalogue would be large enough (several tens of thousands of sources), and X−ray and optical data are disimilar enough, that this pilot can address several aspects of the general association problem (discussed in (29)) at once.

(8.2.3) The radio pilot − *Fourier data*

Example Science Case:
The incidence of AGN in star−forming galaxies is an important test of theories of galaxy evolution. An astronomer addresses this issue by taking X−ray (e.g. Chandra) and optical/near−IR (e.g. CFHT or Subaru) catalogues, selecting a sample of candidate AGN and generating a radio image around the position of each from archival visibility data. The radio structure on various scales (including any evidence of mergers) and the radio spectral index can then be used to reveal starburst regions and obscured AGN.

High resolution radio interferometer arrays produce data sets which are samples of the Fourier transform of the radio sky. These 'visibility data' can be processed in different ways depending on the astronomical requirements. Since sampling in the Fourier plane can be sparse, non−linear deconvolution is a necessary and critical step in the production of images which can be easily interpreted. Although the field−of−view of high resolution interferometer arrays is often large, the information content can be significantly smaller, due to the limited Fourier sampling. It is therefore more efficient and more productive to maintain access to the data in the Fourier plane and produce images or data products on demand, with options as to whether to carry out a particular deconvolution, fit models to the data in the Fourier plane, combine with data from another interferometer, or to select a particular region on the sky. This pilot, taken together with work from AVO WA3.3 (*Scalable computing and storage*), was designed to produce a test−bed system that allows users to access visibility data remotely and then launch image production on−the−fly, tailored to the requirements of the specific science goal, and implemented in a parallel computing environment to enhance speed.

(8.2.4) The solar pilot − *Data selection from summary information*

Example science case:
A scientist is studying a particular X−class solar flare. S/he wishes to identify and explore the data available which cover its site during the 24 hours leading up to the event. GOES X−ray flux data can be used to identify the timing of the event but the location is often very approximate and will be taken from the reported position of the associated active region. Catalogue data for all observations taken during the required time−period can be used to refine the flare location. This stage may also require access to original data (possibly in the form of quick−look images made on−the−fly) rather than relying on metadata. Due account will need to be taken of solar rotation during the 24 hrs. With accurate time and location parameters, the search for supporting data will be refined. The morphological and photometric history of the region will be investigated using imaging data and its plasma properties can be characterized from spectroscopic data found to match the event. Magnetogram

data will be needed in a form suitable for automatic spatial− and temporal−registration with any monochromatic images available from a variety of space and ground−based sources.

This pilot is principally a test bed for improving methods of data selection in solar physics. This is important because it is not typical in solar physics for all data from a particular experiment to be automatically pipeline−reduced all the way to science−quality products. What is more common is that summary data products are made available, together with catalogues listing observational parameters, which the user may then interrogate, to select datasets containing observations of solar features of the desired sort, and s/he can then request a copy of the processed data from the primary archive for that experiment. The main goal here, therefore, is to develop mechanisms for making it easier for the user to select which datasets are of interest, since this is the stumbling block to doing science with the data. This pilot involves much more interactive work than the others, so it does more to address on−the−fly federation of datasets than the others, which are principally implementing static federations, and much of its work is undertaken in collaboration with the European Grid of Solar Observations (EGSO,(7)).

(8.2.5) The STP pilot − *Time−series data*

Example science case:
A scientist wishes to study the propagation and effect of a coronal mass ejection. This requires use of: (i) the coronagraph on SOHO(8); (ii) upstream solar wind measurements from ACE; (iii) Cluster(9) plasma and field measurements near the magnetopause; (iv) plasma composition measurements in the mid altitude cusp; (v) ring current enhancements, in situ, remote sampling and ground−based geomagnetic indices; (vi) position and timing information. These data sets range from simple scalar time series data, to sequences of images and higher dimensionality arrays. They currently have different locations, query specifications and are returned in different formats. The data may need to be transformed into a consistent co−ordinate frame or combined to produce ancillary products. A uniform, and flexible, metadata specification is therefore crucial to ensure that manipulation of data from different archives can be done in a consistent and correct way.

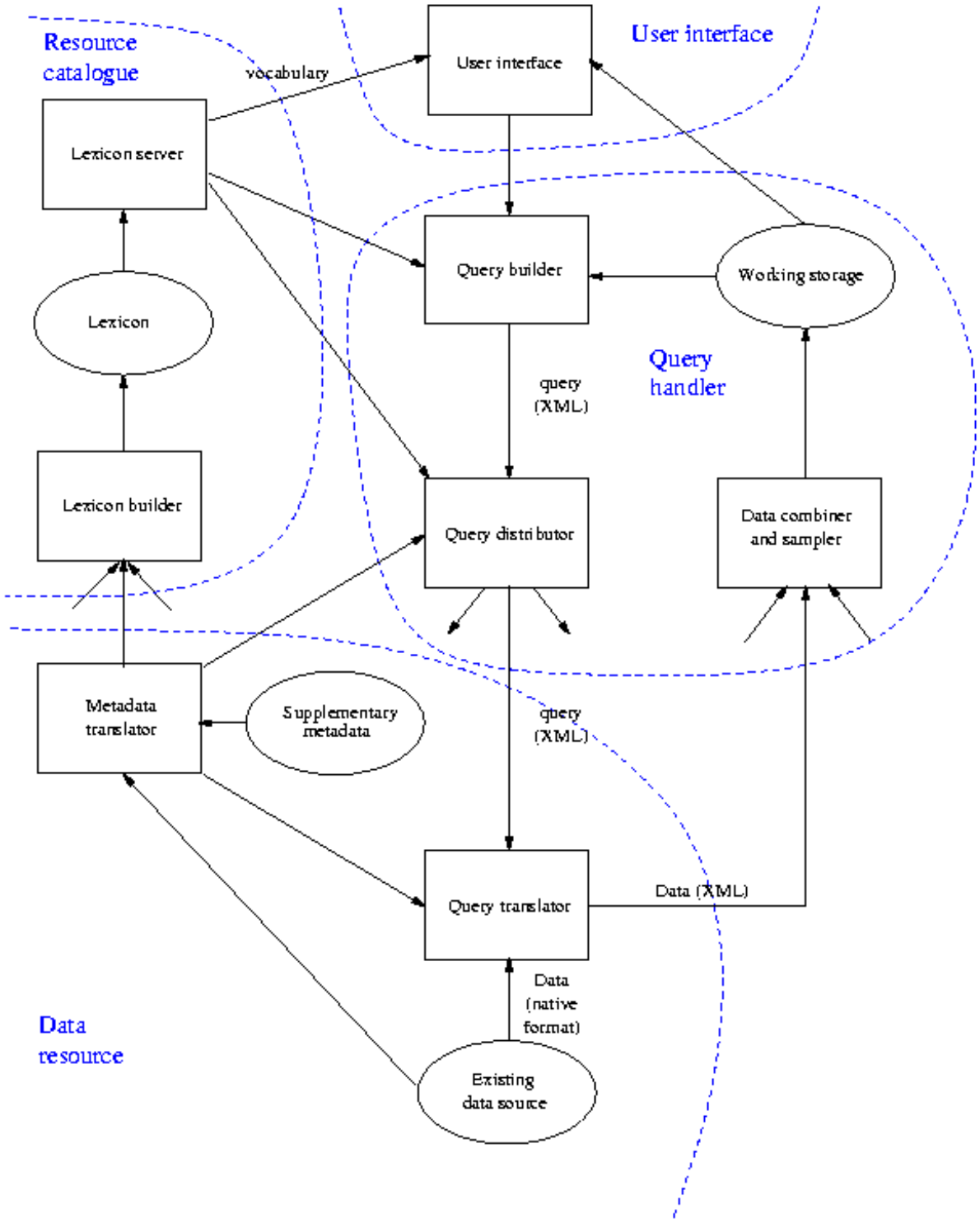
This test federation was designed to investigate the federation of heterogeneous time−series data. This is of particular relevance to Solar Terrestrial Physics (STP) data sets due to the large number of in situ, multi−point and remote sensing measurements made across a wide range of scales in both time and space. STP data sets are relatively small compared to the other AstroGrid domains. The main issues to address come from the complexity of the analysis and in particular the need to locate, search, extract, manipulate and combine multiple data sets. It is also important to consider the international perspective since many of the key datasets that will be required by UK STP scientists originate from non−UK instruments and facilities.

(8.3) Highlights from the Pilot Programme

(8.3.1) The STP pilot − *Time−series data*

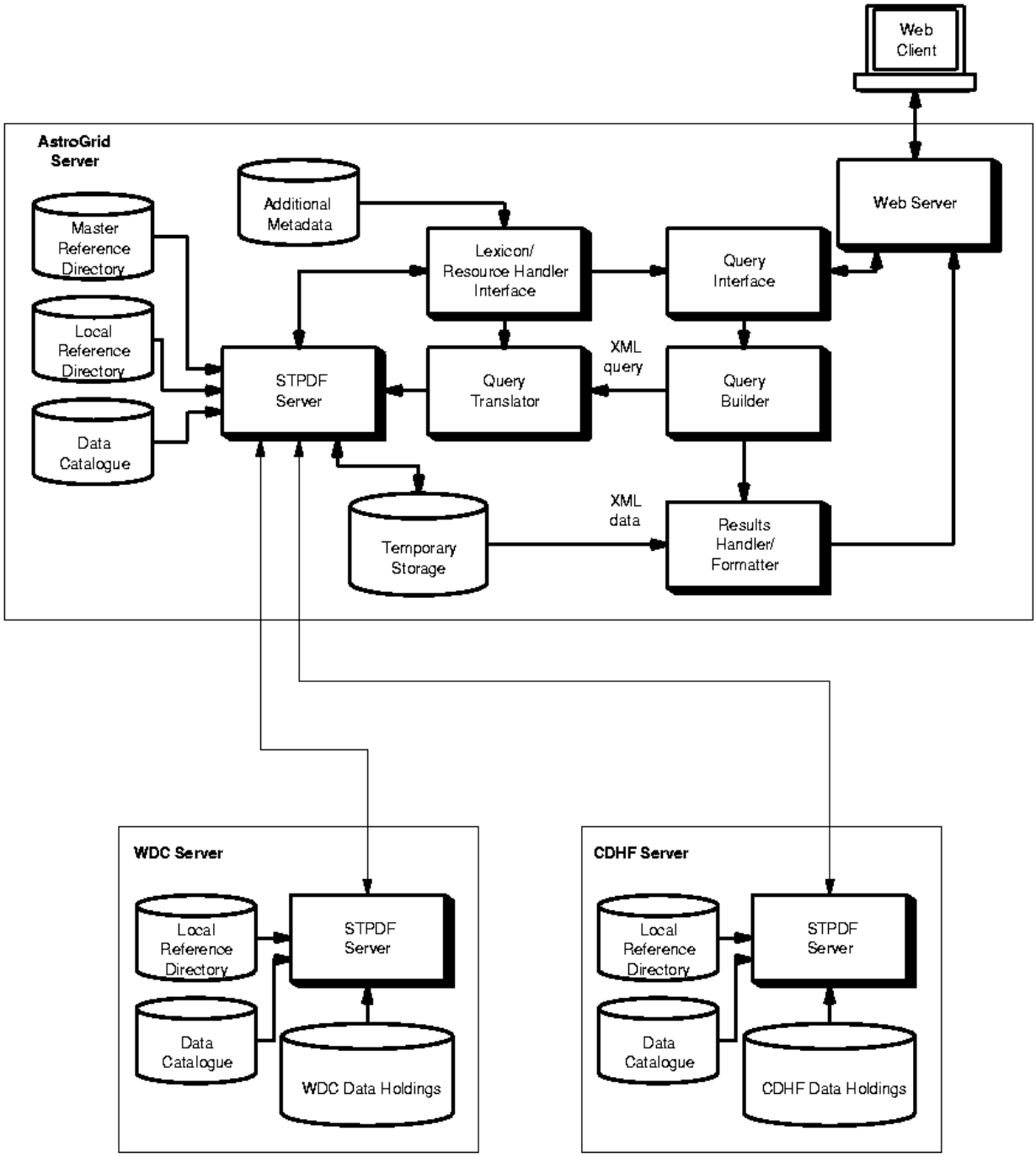
The teams running the STP and solar pilot conducted a wide−ranging questionnaire in conjunction with ESA's SpaceGRID(10) initiative, addressed to a wide cross−section of the international solar system research community and elicting more than one hundred responses. The details of these responses are beyond the scope of this document − a summary is available (11) − but a few points from it are worth noting here. This exercise produced some quite explicit performance requirements, not specified so concretely yet elsewhere within AstroGrid, for example: the system should provide feedback on an action with 30s; simple, online tasks should be completed within an average time of a minute; and complex, offline tasks should be completed within an average time of 24 hours. Interestingly, this survey also identified some Intellectual Property Rights issues not discussed much within AstroGrid: some respondents thought that there should be a possibility of keeping workflows and query results within AstroGrid's MySpace.

The results of this requirements survey influenced the design of the pilot's top−level architecture, which is sketched below:



where the rectangles are programs, the ellipses data stores, the dotted lines demarcate abstract entities (i.e. the *Resource catalogue*, an arbitrary *Data resource*, the *Query handler* and the *User interface*) and the arrows indicate major data flows. Where possible, this architecture was implemented using existing software, such as the Solar Terrestrial Physics Data Facility (STPDF (12)) system, and using data sources selected from those of the UK Cluster Data Centre (UKCDC (13)) and the World Data Centre for Solar−Terrestrial Physics (WDC (14)), already on line at RAL. These comprise about 35 million time series records in total, of several types: from UKCDC would come data from ACE, as well as GOES Key Parameters, while the WDC would provide geomagnetic indices, Dst and aa data. The UKCDC data are held as CDF files (one per day), while the WDC data are stored as ASCII or binary tables; onestrength of STPDF is that it can provide a uniform view across this heteregeneous set of data resources.

The full pilot system looks like this:

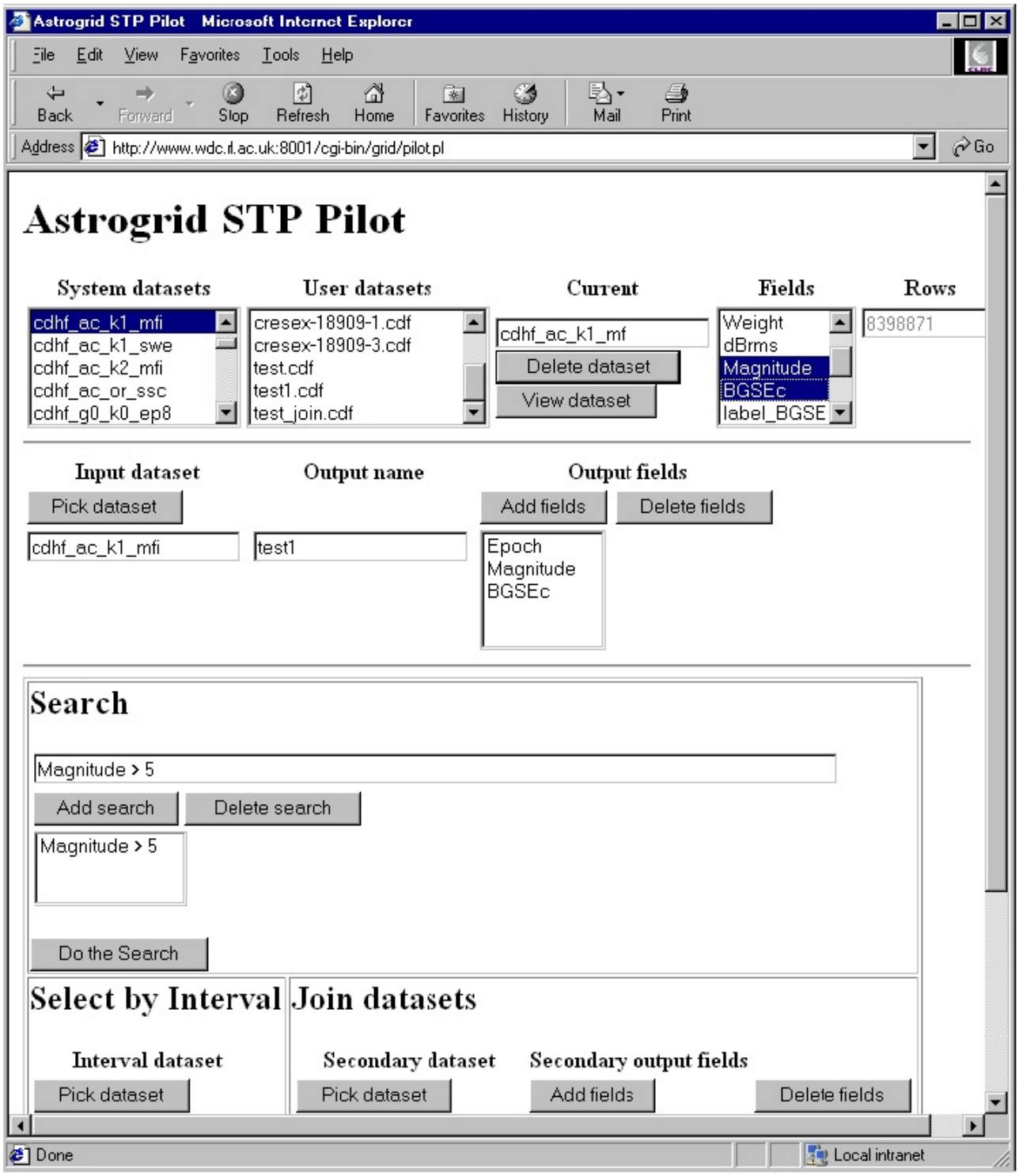


and involves three servers at RAL:

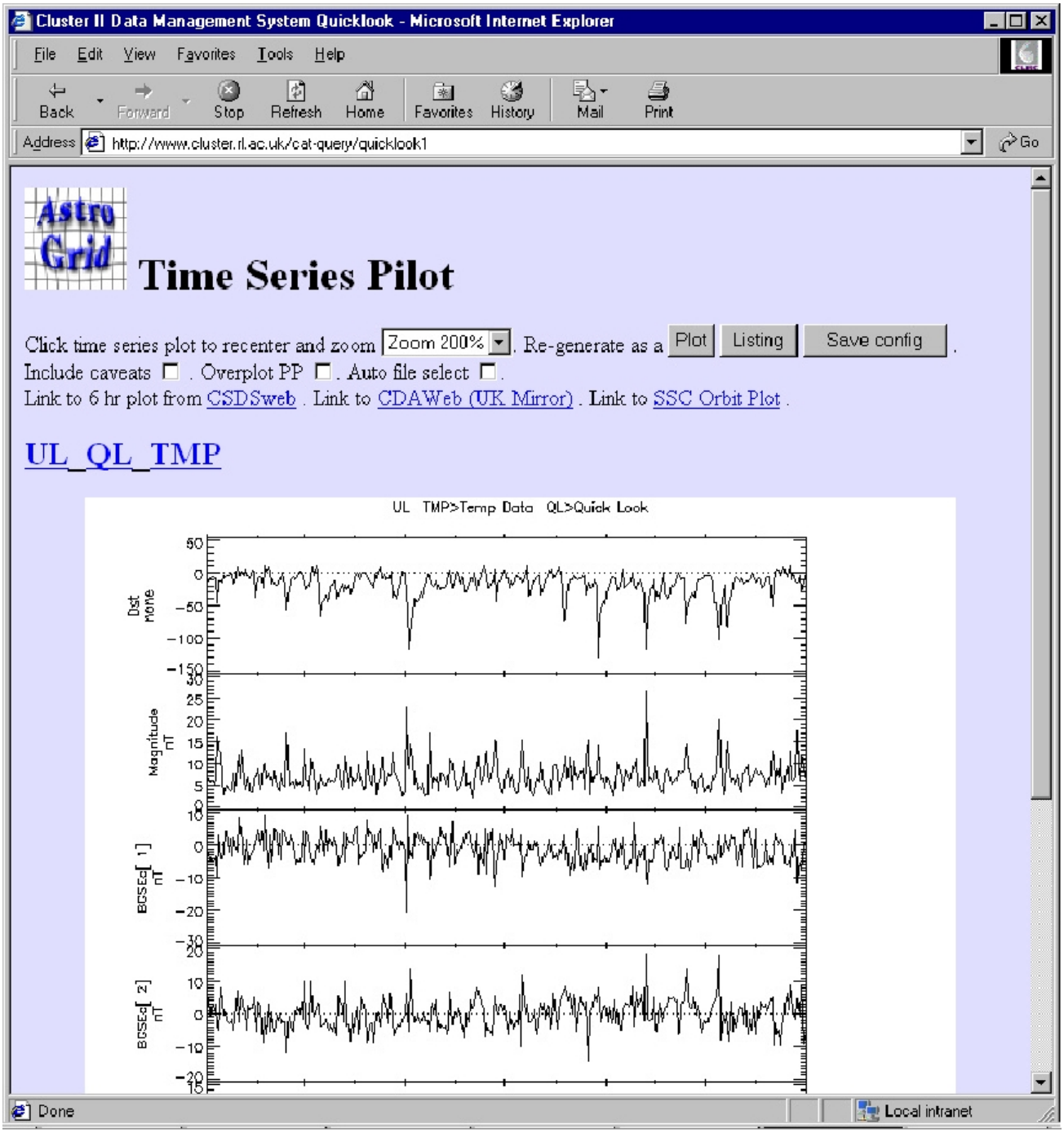
- The AstroGrid Server: is a Sun workstation, with STPDF installed on it. The local reference directory and data catalogues provide information stored on the local server. The master reference directory provides information about the location of network−accessible resources. In the case of the pilot work, this contains entries for the UKCDC and WDC servers. A web−based query is dynamically generated from information extracted from the STPDF system, plus additional metadata not available from it. The query builder generates an XML query file that is passed to the query translator, which handles the interface to the STPDF system. STPDF then handles the requests for data from each of the required archives and returns a result. That is formatted and returned to the user through the web server.
- The WDC server: provides access to its data holdings via the STPDF server. The STPDF system handles the translation from the underlying data format and the sub−setting of the data.
- The UKCDC server provides access to its holdings in the same way as those of the WDC server, described above.

The development of the pilot's metadata translation layer started with an assessment of a number of existing XML formats that might be used. A report on this (15) may be found on AstroGrid Wiki page, and it may be summarised as follows: XSIL(16) is too simple; XDF(17) is too complicated; CDFML(18) is too STP−specific; and VOTable(19) looks usable. Despite the fact that VOTable was considered usable for STP work, it was decided to use an *ad hoc* XML format for the pilot work, for several reasons. Firstly, as with the consideration of use of VOTable in the solar pilot, its use in a new area would require the definition of a new set of Unfied Column Descriptors (UCDs(20)), for which there was insufficient time in the pilot. Secondly, it would be easier to define the restricted DTD or schema in its own namespace, without having to implement the whole of VOTable. Whilst the XML format used here was *ad hoc*, in some sense, it was decided to make use of International Solar−Terrestrial Physics (ISTP(21)) guidelines for defining its terms. For the STPDF software to be able to pick up the metadata for the WDC data, it was necessary to hand−craft text files, while the UKCDC metadata could be read straight from their CDF files. (*N.B.* SpaceGRID (in collaboration with GSFC, CDPP, Southwest Research Institute and PDS) are defining a space physics query language, which will yield a data dictionary likely to become the default for use in STP: this effort is starting from low level terms (i.e. Dublin Core) and then building up domain−specific metadata within a namespace. The AstroGrid STP pilot is constructing domain−specific descriptors based on the SpaceGRID work, and it is suggested that these would have to be included in any more general AstroGrid ontology via an STP namespace.)

Queries to the pilot system are constructed using a simple UI



which allows selection of system or user datasets (i.e. the results of previous operations, thereby allowing compound queries to be built). The UI is dynamically updated to show available fields, and the number of records for each selected dataset. Four basic operations are supported: select/output fields; query data set; find/select time interval; and time series join (nearest neighbour). Finally, a Quicklook UI (based on the UKCDC UI) was produced to view the selected dataset(s). The user selects a time range subset of the dataset and choose which parameters to plot. The example plot below shows a joined dataset comprising Dst data from the WDC and ACE magnetic field data from the UKCDC.



The user can then zoom in on a portion of the plot and can also obtain an ASCII dump of the data plotted by the Quicklook UI.

(8.3.2) The solar pilot − *Data selection from summary information*

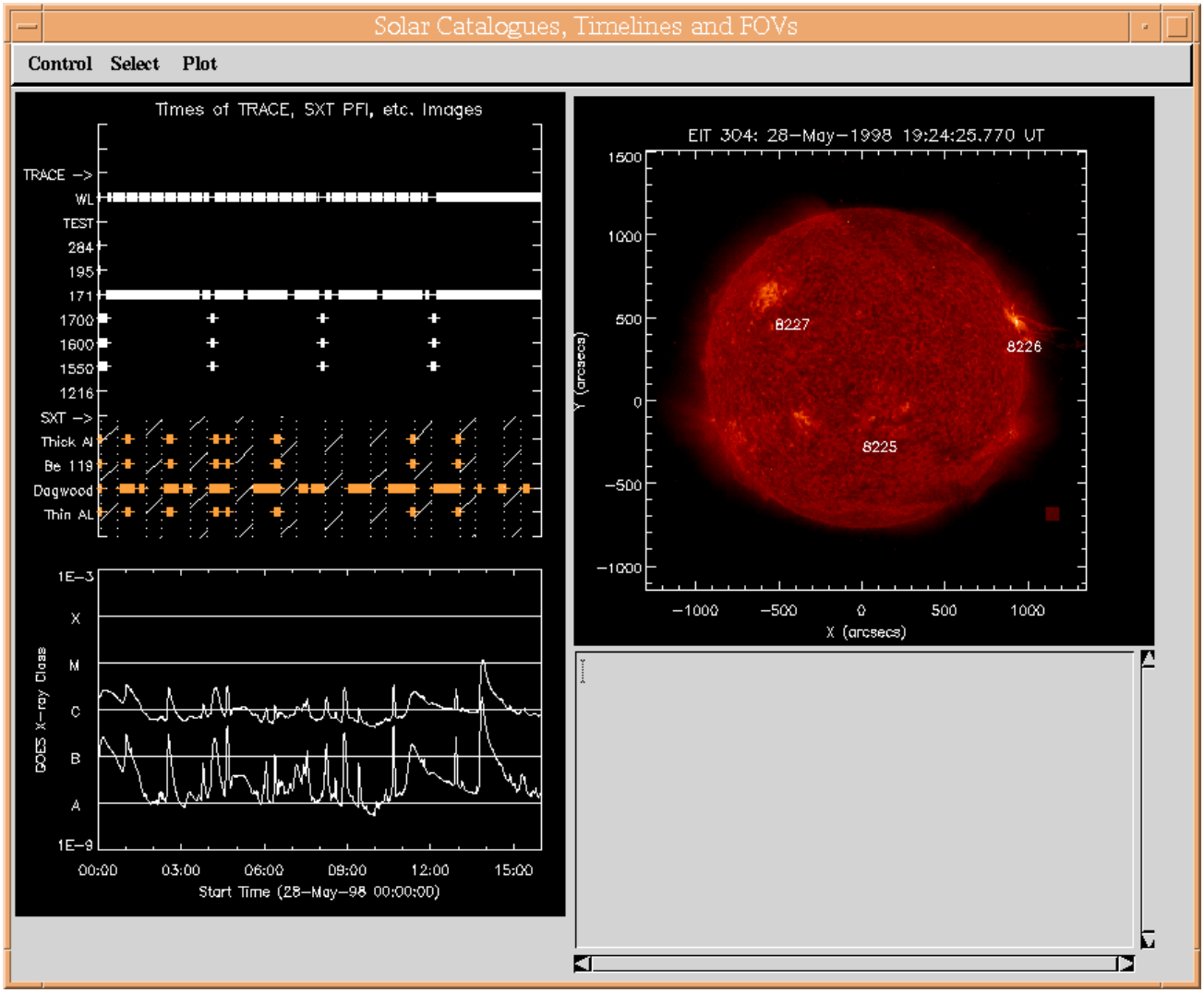
The central design issue for this pilot was the definition of the minimal set of parameters required for solar observing catalogues. This resulted in a document issued under the aegis of EGSO and entitled "EGSO Unified Observing Catalogues" (22). This was constructed from the perspective of catalogue searches, collating the set of parameters that a user might need to query in order to select a particular dataset, but bearing in mind that the set defined must work within the context of the standard SolarSoft(23) software package ubiquitous within the solar physics community. In addition to this minimal parameter set, general information about the observatory (instrument description, contact info, etc) should be available to the user: it was decided that the definition of the requirements for such ancillary data should be left to EGSO, since this information in unlikely to be interrogated via the kind of catalogue queries being prototyped in this pilot.

Another design issue was the format for storing the solar observing catalogue data. Currently, this is usually held with an IDL(24) database, for ease of manipulation using SolarSoft, but there is a desire to remove the reliance on IDL, to ease the

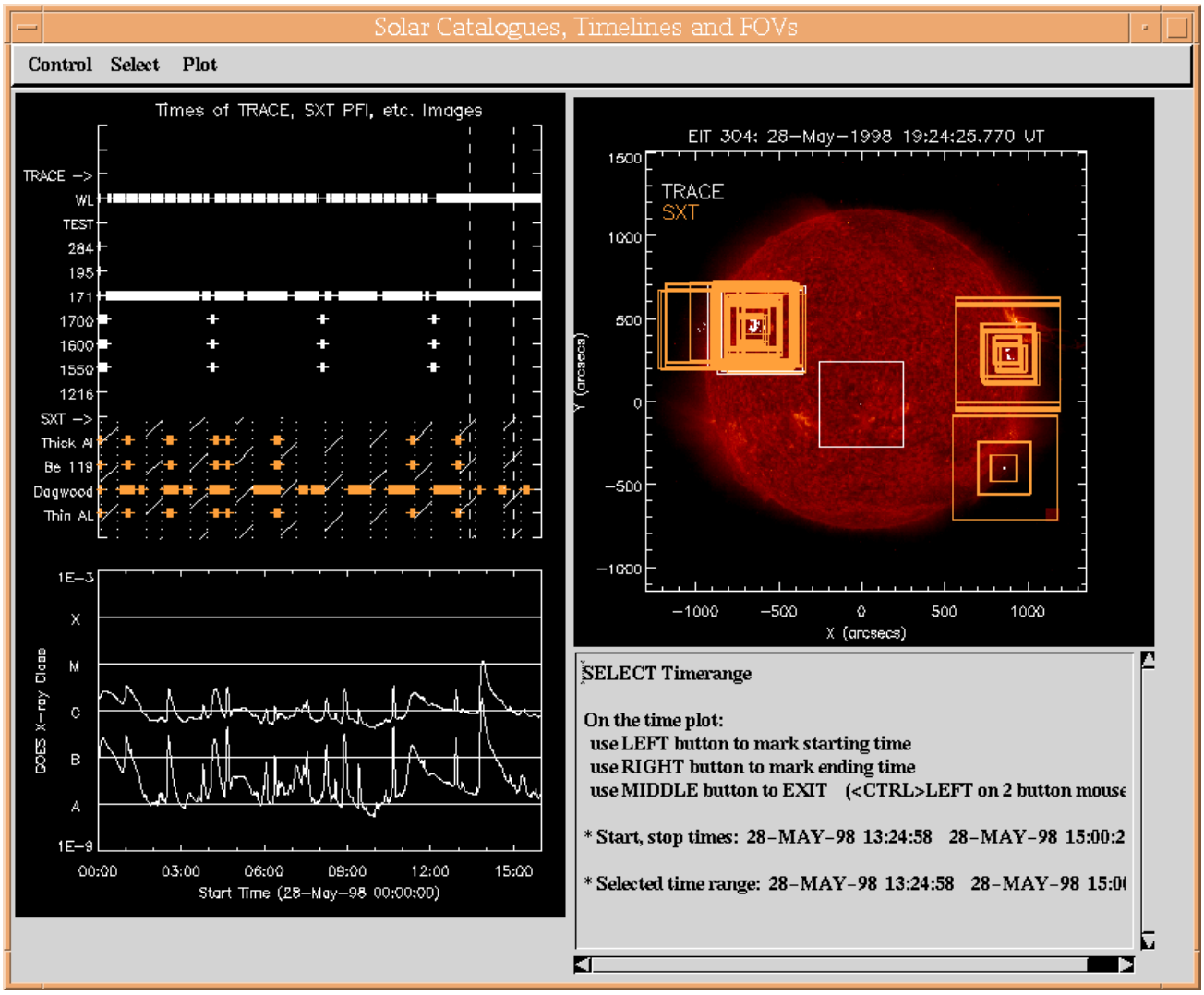
integration of solar physics software and Grid middleware. Another possibility would be to store these data in an XML repository, such as Xindice(25), possibly using VOTable documents for each entry. This would require the addition of further UCDs relevant to solar physics: this avenue is being explored within the context of EGSO, and it was decided that the time constraints on the delivery of working prototype software within this pilot would necessitate the use of an IDL−based system for this pilot.

The series of screenshots below follows the course of a query using this IDL−based system.

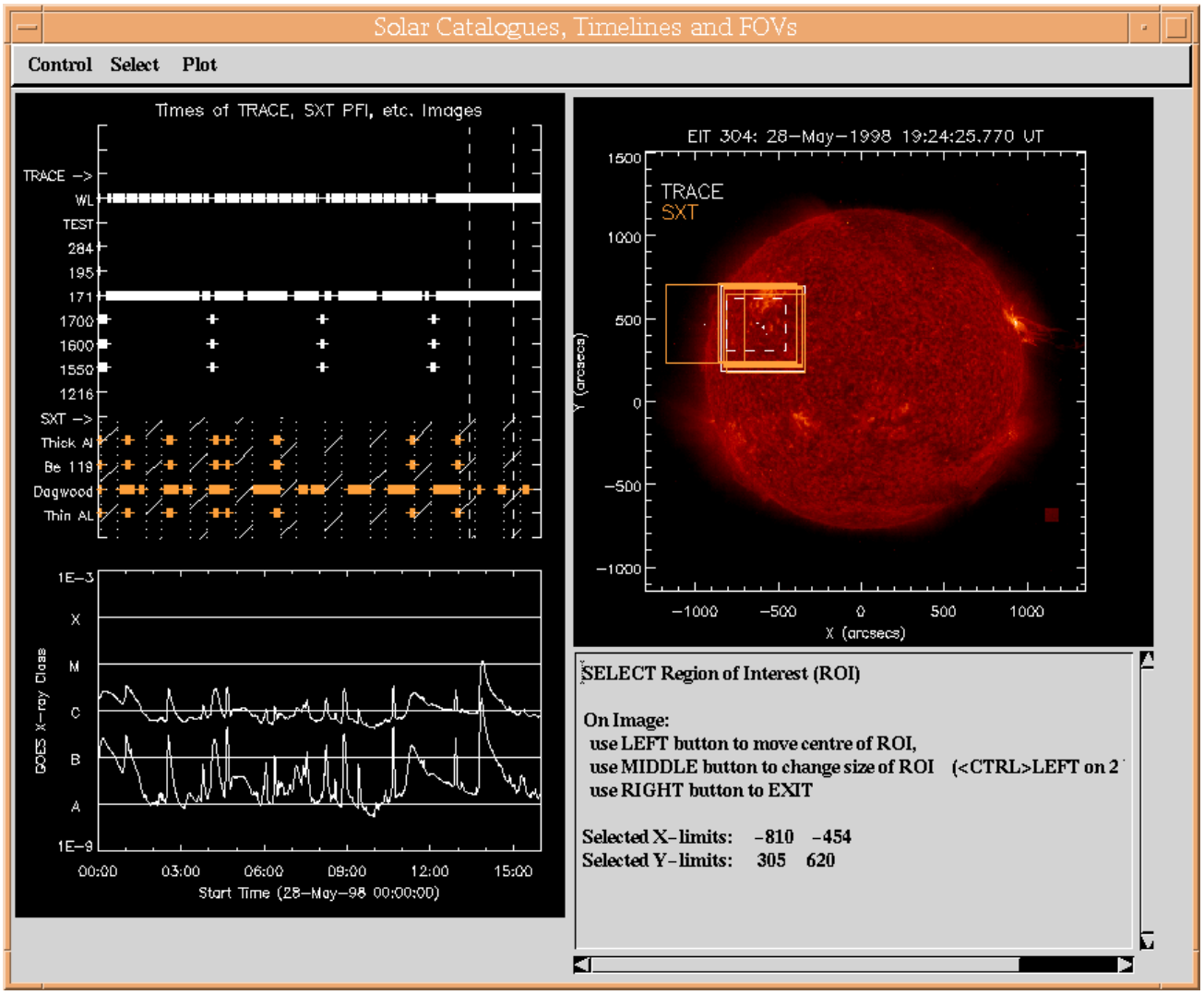
Step 1: The image below shows the start of the selection process. The right hand pane shows a SOHO−EIT EUV image, marking known features. The left hand pane displays information from three different satellites. At the bottom is the time series of GOES X−ray flux measurements in a number of bands. In the middle, orange crosses mark Yohkoh−SXT observations made in various modes during the same time period, but white crosses in the top portion mark TRACE EUV observations in various filters during that interval.



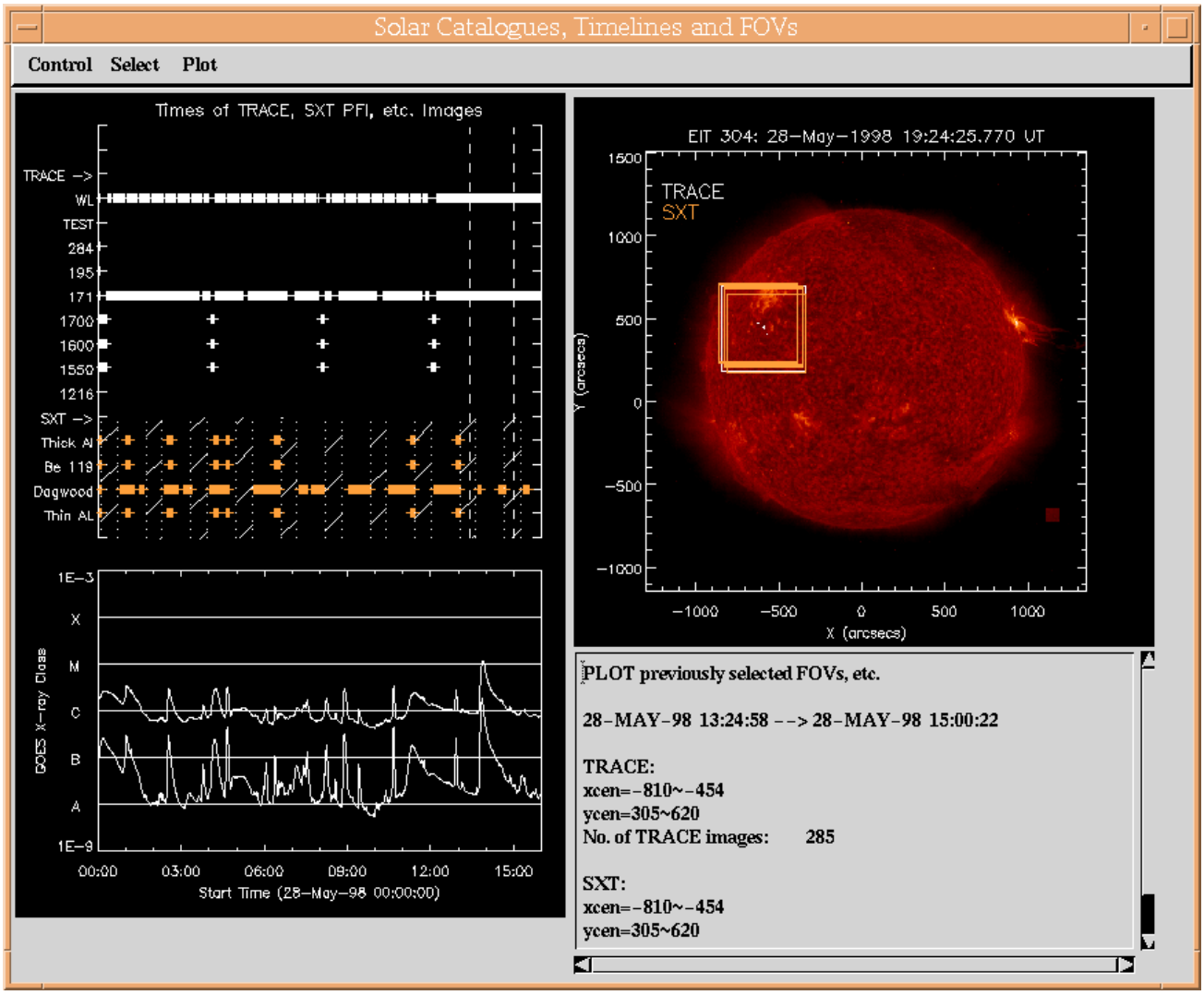
Step 2: The user notes that the GOES X−ray data are exhibiting a solar flare at about 14:00. Using a mouse the user can the define a time interval in the left hand panel (shown by the dashed lines) and then, only the image on the right hand panel are plotted the fields of view (FOVs) of all observations (TRACE in white, SXT in orange) taken during that interval.



Step 3: The user can then zoom in on the region of interest, by selecting an area using the mouse: the selected region is indicated by the dashed square on the EIT image in the right hand pane.



Step 4: Once that region has been defined, the right hand pane is replotted, showing the FOVs of only those observations whose FOVs intersect with the selected region. The lower portion of the right hand window then reports the temporal and spatial selection criteria the user defined, as well as the number of images from each instrument that these have selected.



At the moment, the user then has to go away to the normal UI for the respective databases and extract the desired data via inputting these selection criteria, but it is intended that this pilot will be extended so that the selection can be made directly from this UI. One slight complication is that this current selection procedure can return a large amount of data − 285 TRACE images in the example above − and it would be desirable to add some additional cadence criteria to the selection (e.g. only extract an image every five minutes, say).

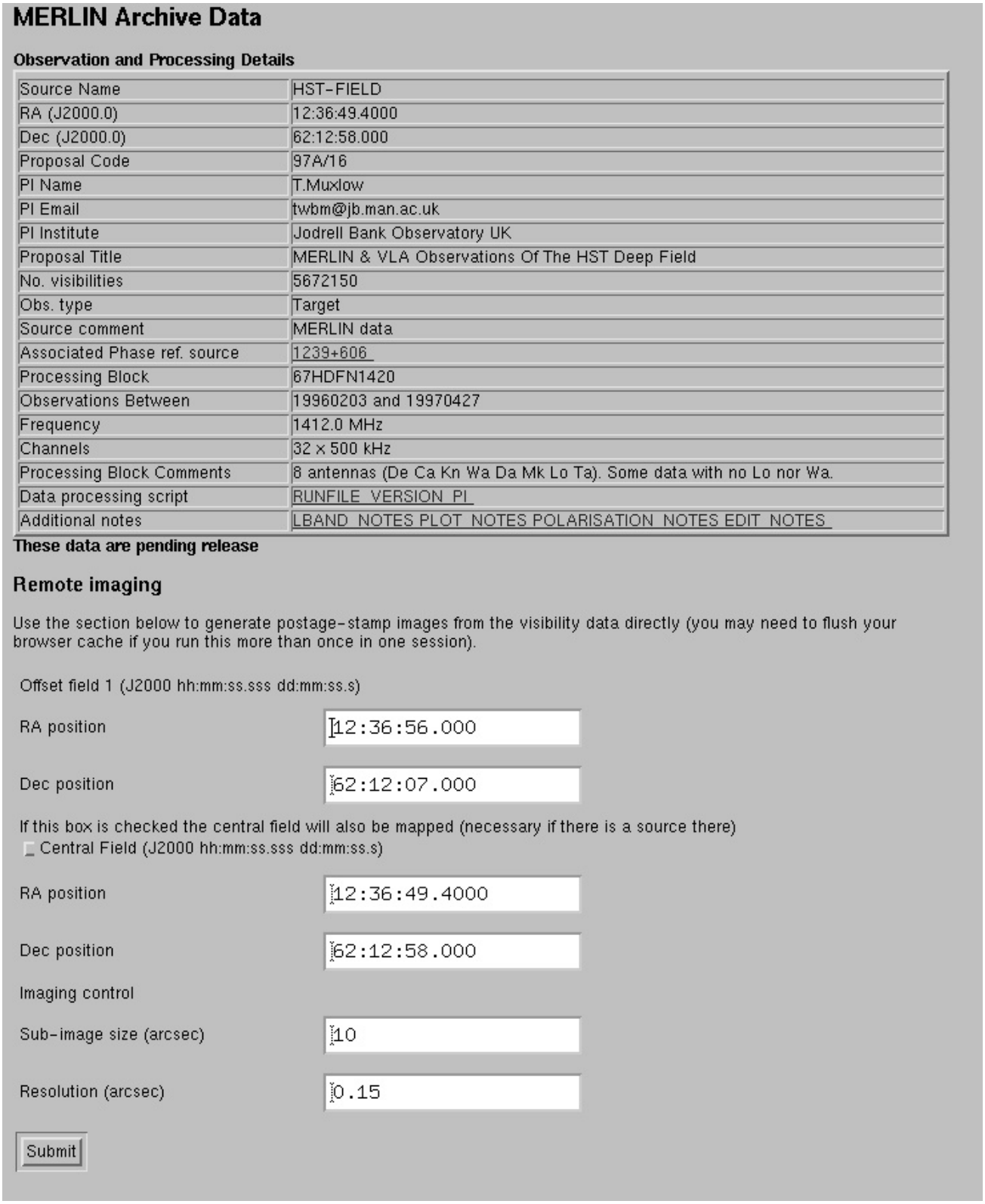
(8.3.3) The radio pilot − *Fourier data*

It was decided to undertake this pilot using data from a region of sky which has been extensively observed right across the electromagnetic spectrum, namely the Hubble Deep Field. The unprecedented sensitivity of the observations required and produced very large data sets, from which there is still significant scientific information to be extracted, making it an ideal candidate for the AstroGrid pilot.

The development of a prototype environment for access to visibility data centred on running AIPS within a cgi wrapper. By this route, the MERLIN Archive(26) now supports simple queries (returning text and ready−made plots and FITS images) either via its web page or via CDS. A prototype interface for on−the−fly imaging of visibility data was produced and has been tested locally and remotely. This enables users to extract maps from calibrated visibility data at any position within the field of view. Typically, only the central few arcsec of archive data have ever been imaged: in the case of the HDF data, only about 1/7 of the total field of view has ever been imaged, but the sensitivity means that more science can undoubtedly be done (e.g. the background radio flux or barely detectable sources at the position of Chandra sources). The calibrated visibility data are several GB, making it impractical to supply it to off−site astronomers even if they could use AIPS locally. Thus a Virtual Observatory service like the one developed here is not merely a convenience but a necessity for such data sets.

The user has only to specify the size, position and resolution required to obtain an image. In the example shown below, the astronomer has interrogated the MERLIN archive and learnt of the Muxlow et al. observations in the Hubble Deep Field. The

user then enters information about the field ("Offset field 1") for which s/he wants to generate an image from the visibility data.



MERLIN Archive Data

Observation and Processing Details

Source Name	HST-FIELD
RA (J2000.0)	12:36:49.4000
Dec (J2000.0)	62:12:58.000
Proposal Code	97A/16
PI Name	T.Muxlow
PI Email	twbm@jb.man.ac.uk
PI Institute	Jodrell Bank Observatory UK
Proposal Title	MERLIN & VLA Observations Of The HST Deep Field
No. visibilities	5672150
Obs. type	Target
Source comment	MERLIN data
Associated Phase ref. source	1239+606
Processing Block	67HDFN1420
Observations Between	19960203 and 19970427
Frequency	1412.0 MHz
Channels	32 x 500 kHz
Processing Block Comments	8 antennas (De Ca Kn Wa Da Mk Lo Ta). Some data with no Lo nor Wa.
Data processing script	RUNFILE VERSION PI
Additional notes	LBAND NOTES PLOT NOTES POLARISATION NOTES EDIT NOTES

These data are pending release

Remote imaging

Use the section below to generate postage-stamp images from the visibility data directly (you may need to flush your browser cache if you run this more than once in one session).

Offset field 1 (J2000 hh:mm:ss.sss dd:mm:ss.s)

RA position 12:36:56.000

Dec position 62:12:07.000

If this box is checked the central field will also be mapped (necessary if there is a source there)

Central Field (J2000 hh:mm:ss.sss dd:mm:ss.s)

RA position 12:36:49.4000

Dec position 62:12:58.000

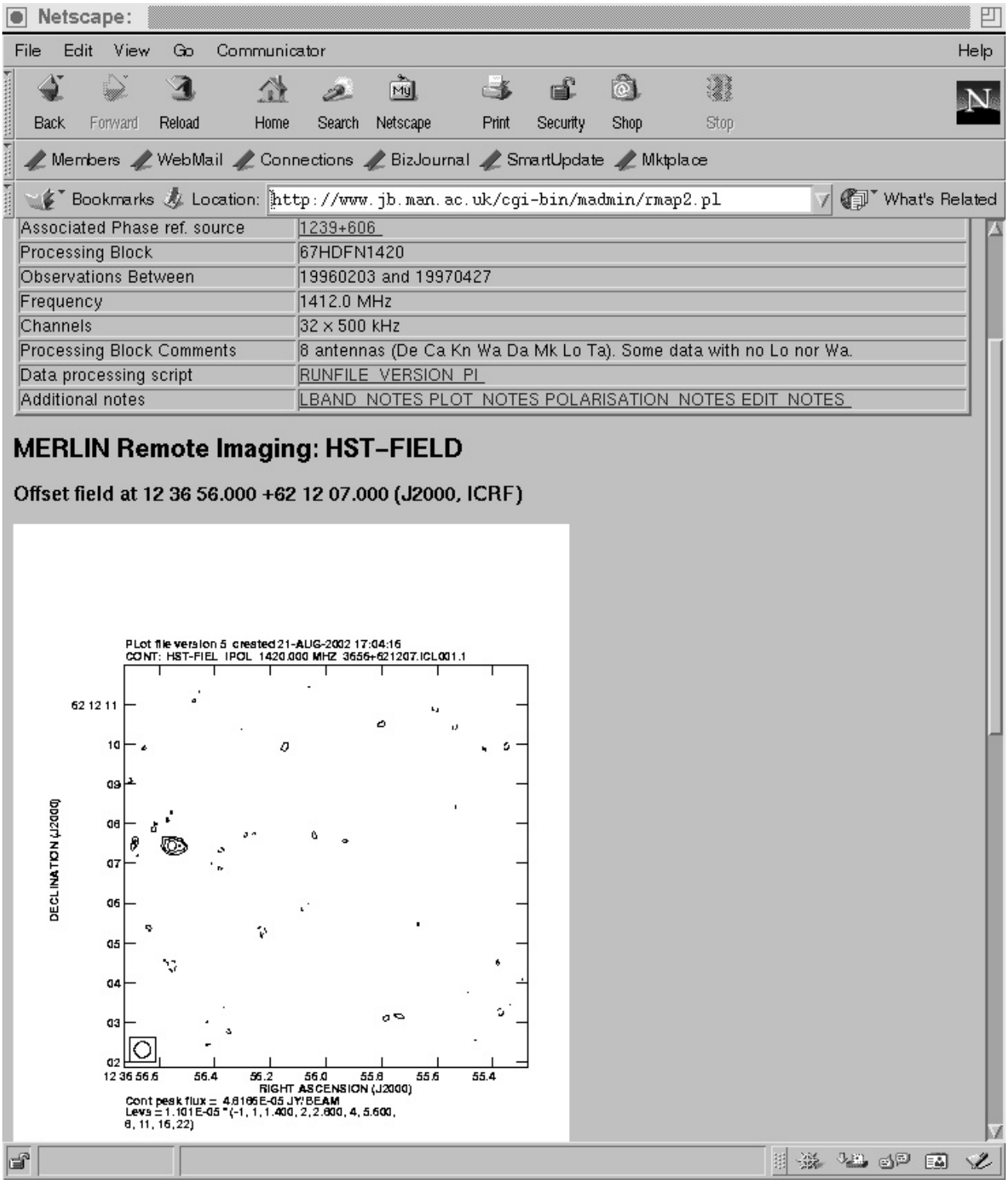
Imaging control

Sub-image size (arcsec) 10

Resolution (arcsec) 0.15

Submit

The AIPS script then generates the image specified by the user on−the−fly and displays a postage stamp showing the resulting image,



and the data may then be downloaded in FITS format. Astronomers using this facility do not need any knowledge of AIPS or of radio data, but behind the scenes the archive uses two alternative routes depending on the size and complexity of the input data sets. MERLIN−only visibility data around a chosen region can be imaged on demand. A small amount of further development is needed to provide informative feedback if the user requests images at a position or resolution which the data cannot provide, based on individual data set properties. For very large data sets taken from more than one array (e.g. the HDF MERLIN+VLA data) two−stage data processing is required and only one route is currently implimented (combining images in the map plane); other possibilities will be investigated, in due course.

User feedback on this prototype system was positive, confirming that it produces (in an automated fashion) data of the same quality as that produced by an expert reducing the same data, although it was noted that the system ran slower than might be expected. Several performance bottlenecks have been identified, and possible enhancements are currently being investigated.

The test user did note that the lack of defaults and restrictions in the current system means that the inexperienced user could be led to a poor specification of the image to be produced.

Work on deriving interferometry metadata standards has proceeded in conjunction with AVO Interoperability work. A library of terms to describe radio interferometry observations has been drawn up and these have been translated into CDS UCDs and work is ongoing to expand these where they are not at present sufficient. A fuller report on this activity can be found in (27) and consultations are underway with experts on higher frequency Fourier data to expand the scope of these metadata to ensure that it covers the requirements of interferometry in general, not just those of radio astronomy.

(8.3.4) Progress with the optical/near−IR and X−ray pilots

In common with the other three, the optical/near−IR and X−ray pilots were designed to be scientifically interesting, as well as functionally instructive, so both planned to use very topical data sets. Problems with the availability of these data sets − i.e. delays with the installation of a copy of the Sloan EDR database at Edinburgh, and with the production of the first XMM source catalogue by the Survey Science Centre, respectively − meant that neither of these pilots was able to proceed all the way to the delivery of prototype software to test users in August 2002. Indeed, work on them both was suspended midway through Phase A, as a result of delays with the test datasets, and the effort associated with them allocated to other parts of the Phase A programme.

However, both pilots did yield interesting results. For example, as part of his work for the X−ray pilot, Clive Page derived the PCODE method for performing spatial cross−matches in relational databases (28, 29). This was used to find optical associations of sources in a ROSAT catalogue, but the surface density of objects in this situation was not sufficient to exercise the algorithm thoroughly. More generally, the X−ray pilot provided the motivation for some thought of how to perform associations in the VO (29), although this remains a problem in need of a proper solution. The results of the optical/near−IR pilot are less obvious, since most of the work undertaken for it centred on gaining an understanding of the SX(3) archive system, which has since been dropped by the SDSS (2) consortium. Nevertheless, the effort expended on the optical/near−IR pilot before its suspension was not all wasted, since much of the design work for a SX−like schema for the SuperCOSMOS Sky Survey could be re−used once it was decided to produce the new SSS archive using SQL Server and not Objectivity/DB.

(8.4) Summary and Conclusions

The Pilot Programme was designed to complement the AstroGrid Science Requirements exercise and the Phase A technology evaluation workpackages, by trying to implement some small parts of AstroGrid's desired functionality using existing technologies. In this way, their limitations could be assessed, at the same time as offering the opportunity to supplement the AstroGrid Science Requirements with feedback from test users from each of AstroGrid's user communities exposed to prototype software from the pilots, which would give them an indication of what AstroGrid would eventually deliver.

The set of five pilots was chosen to cover different aspects of the general database federation problem of particular interest to the different parts of AstroGrid's user community. In each case, the pilot was intended to be performed using scientifically−interesting datasets, to motivate the use of the prototype software by test users. This led to difficulties with both the optical/near−IR and X−ray pilots, where delays with the delivery of data led to the suspension of work on the pilot mid−way through Phase A. This meant that neither of these pilots delivered prototype software to test users in August 2002, as originally intended. However, the data required for both pilots are now becoming available and it is intended that a sigificant fraction of the originally−intended work of both pilots will be completed by the end of December 2002, when the extended Phase A comes to a close.

The remaining three pilots succeeded in producing prototype software for test users to evaluate by the end of August 2002, which marks the originally−planned close of Phase A. User response to all three was positive, reassuringly confirming that initial ideas of desirable functionality were correct. One thing noted by test users of several pilots was the ease with the prototype systems could generate undesirably large volumes of data. This leads to one of the most important lessons from the Pilot Programme, which is the importance of having a resource estimation capability, both to indicate how long a job will take to run and how much data it is likely to generate.

Many useful lessons have been learnt from the AstroGrid Pilot Programme. At the most general level, test users responded positively to the additional functionality they were offered, but quickly wanted the ability to do more. This reassuringly confirms that the VO enterprise is worthwhile, but also that it will be difficult to meet expectations, and that considerable flexibility will have to be designed into VO systems to help them meet the range of user requirements. A number of more specific results emerged, too, for example that the VOTable prescription for presenting tabular data in XML may well be useful well beyond its originally intended domain, if appropriate UCDs (e.g. for interferometric, solar physics and STP data)

can be defined. It was also clear that the VO must provide the means of estimating the resource implications (e.g. result dataset volume, time taken to run, etc) of proposed operations, lest users grind the system to a halt and/or generate much more data than they can handle.

A questionnaire circulated by the teams running the STP and solar pilot, in conjunction with ESA's SpaceGRID initiative, and addressed to a wide cross−section of the international solar system research community, produced some quite explicit performance requirements, for example: the system should provide feedback on an action with 30s; simple, online tasks should be completed within an average time of a minute; and complex, offline tasks should be completed within an average time of 24 hours. Interestingly, this survey also identified some Intellectual Property Rights issues not discussed much within the VO community to date, as some respondents thought that there should be a possibility of keeping workflows and query results private within the VO.

Perhaps the most valuable lesson learnt from the AstroGrid Pilot Programme is the importance of keeping (at least some) users closely engaged in the VO development process. New technologies, and the vast wealth of astronomical data now available, mean that the possible directions that the VO can take greatly exceed what can possibly be delivered, given the finite funding for the various VO projects, and it is essential that the course of VO development is decided by what users most need to do their science and not what the technology can deliver most readily.

A fuller account of the Pilot Programme can be found in (28).

References

(1) SuperCOSMOS Sky Survey: http://www−wfau.roe.ac.uk/sss
(2) Sloan Digital Sky Survey: http://www.sdss.org
(3) SX archive system: http://www.sdss.jhu.edu/ScienceArchive/home.html
(4) VizieR: http://vizier.u−strasbg.fr/viz−bin/VizieR
(5) Centre de Données astronomiques de Strasbourg (CDS): http://cdsweb.u−strasbg.fr
(6) Astrophysical Virtual Observatory (AVO): http://www.eso.org/avo/
(7) European Grid of Solar Observations (EGSO):http://www.mssl.ucl.ac.uk/grid/egso
(8) SOHO: http://sohowww.nascom.nasa.gov/
(9) Cluster: http://www.cluster.rl.ac.uk/
(10) SpaceGRID: http://spacegrid.esa.int
(11) Summary of solar system research requirements: http://wiki.astrogrid.org/bin/view/Astrogrid/SpaceGRIDRequirements
(12) Solar Terrestrial Physics Data Facility: http://www.ssd.rl.ac.uk/stpdf/
(13) Cluster Coordinated Data Handling Facility: http://www.cluster.rl.ac.uk/cdhf.htm
(14) World Data Centre for Solar−Terrestrial Physics: http://wdcc1.bnsc.rl.ac.uk/
(15) Report on "XML for STP data": http://wiki.astrogrid.org/pub/Astrogrid/PilotDocs/agstp−0002.pdf
(16) Extensible Scientific Interchange Language(XSIL): http://www.cacr.caltech.edu/SDA/xsil/
(17) eXtensible Data Format (XDF): http://xml.gsfc.nasa.gov/XDF/XDF_home.html
(18) CDF Markup Langugage (CDFML):http://nssdc.gsfc.nasa.gov/cdf/html/cdf_xml.html
(19) VOTable: http://cdsweb.u−strasbg.fr/doc/VOTable/
(20) Unified Column Descriptors: http://cdsweb.u−strasbg.fr/doc/UCD.htx
(21) International Solar−Terrestrial Physics (ISTP): http://www−istp.gsfc.nasa.gov
(22) EGSO Unified Observing Catalogues: http://www.mssl.ucl.ac.uk/grid/egso/public/documents/EGSO_UOC_Contents.doc
(23) SolarSoft: http://surfwww.mssl.ucl.ac.uk/surf/surf_software.html
(24) Interactive Data Language (IDL): http://www.rsinc.com/idl
(25) Xindice: http://xml.apache.org/xindice/
(26) Merlin Archive: http://www.merlin.ac.uk/archive
(27) Interferometry metadata report: http://www.jb.man.ac.uk/~amsr/WP5.3/radiometadata.html
(28) Pilot Programme Final Report: http://wiki.astrogrid.org/bin/view/Astrogrid/WPA5FinalReport
(29) Association Methods report from X−ray pilot: http://wiki.astrogrid.org/bin/view/Astrogrid/WPA5AssociationMethods

(9) Financial and Management Report

(9.1) Introduction

This section of the Phase A Report outlines the way we have chosen to manage the AstroGrid project, and summarises expenditure on personnel and finances up to the end of Aug 2002.

Other parts of this report (*Architecture Overview* and *Phase B Plan*) have described the ***Unified Process*** (UP); in this report we will outline its impact on the management of the project and the implications for a work package approach to project structure. AstroGrid non−salary finances have been managed centrally as opposed to the usual up−front distribution to institutes and we explain how this was set up and controlled.

In the rest of this report, we refer to the four quarters of Phase A (plus a fifth quarter extension to Phase A which we have requested of PPARC). The actual dates of these quarters are:

- ***Q1***: Sep'01 to Nov'01
- ***Q2***: Dec'01 to Feb'02
- ***Q3***: Mar'02 to May'02
- ***Q4***: Jun'02 to Aug'02
- ***Q5***: Sep'02 to Dec'02 (*yes, **four** months long*)

Note that expenditure in Q1 included, by PPARC's permission, travel expenditure undertaken in the project planning phase before the official start of Phase A. Also note that some aspects of Q4 expenditure are not yet known, as invoices from institutes have not yet been received.

(9.2) Management

(9.2.1) Project Policy and Oversight

Overall responsibility for the project rests with the ***AstroGrid Lead Investigators (AGLI)***. In alphabetical order, they are :

- Peter Allan (RAL)
- Simon Garrington (Jodrell Bank)
- Louise Harra (MSSL)
- Andy Lawrence (Edinburgh, Project Leader)
- Richard McMahon (Cambridge)
- Fionn Murtagh (Belfast)
- Mike Watson (Leicester)

The AGLI set policy and direction, monitor the progress of the project and the performance of the Project Manager and Project Scientist, and make decisions on resource allocation. Monthly telecons are held which also normally include the Project Manager and Project Scientist. One member of the AGLI is elected as Project Leader, who acts with the Project Manager and Project Scientist as a small executive team. Nearly all day−to−day management is delegated to the Project Manager. The Project Leader takes an overall leadership role, and is responsible for liaison with PPARC. AstroGrid is overseen by two bodies − PPARC's Grid Steering Committee, which holds the budget which funds AstroGrid, and the AstroGrid Oversight Committee.

(9.2.2) Work Structure

In AstroGrid Phase A we have had both a traditional workpackage structure, and a modern flexible incremental approach using the Unified Process. Here we explain how this came about and how it works.

The majority of medium to large sized academic projects, especially those distributed across several institutes, tend to have a work package structure. A work package is focussed on one aspect of the project − an area of technology or research − and led by an expert in that subject. The work packages cooperate on standards and work together to ensure that their products integrate correctly. This was the way that AstroGrid was conceived and set up, with the following Phase A workpackages (8):

- WPA0 Project Management and Infrastructure

- WPA1 Science Requirements Analysis
- WPA2 Data Grid Technology
- WPA3 Storage/Compute Technology
- WPA4 Database Technology
- WPA5 Pilot Programme
- WPA6 WFCAM/VISTA Liaison
- WPA7 External Liaison
- WPA8 Phase B plan

The workpackages were used for formal project control (see below) but the actual work carried out differed in some respects. Most importantly, a large amount of work has formally been in *Architecture* development (see below), which was formally reported through A0. Package A6 (WFCAM/VISTA development) soon ceased to be a formal AstroGrid workpackage, although such liaison remains important. While our effort deployment was flexible, much of the work completed by the project did reflect the original structure, as can be seen by the fact that chapters of the Phase A report have similar titles to the above.

In recent years, medium to large software projects in the commercial world have adopted different methodologies (now being referred to as *Agile* methods) which emphasised an iterative, incremental approach to system building. These methods drastically increase a project's chance of being successful.

AstroGrid hired, as its project manager (PM), Tony Linde, with long−term experience in commercial applications development. After he had explained the ***Unified Software Development Process*** (UP) (1), one of the new methodologies, the project team decided to adopt the UP as its project development methodology.

As different as these two approaches seem, there was little long−term conflict during Phase A. This is because the initial stages of the UP are concerned with the development of an ***Architecture*** (see *Architecture Overview* document) which would represent a high level model of the proposed system. The responsibility for development of the architecture devolved to the PM and was subsumed into work package WP−A0 (Project Management and Infrastructure).

The only area of conflict came in the early days of development of the architecture when team members within other work packages were needed, at some length, to work on the architecture. This was resolved by putting together a small team to concentrate on the architecture with other team members' involvement confined to reviewing the models and other documents produced and taking part in periodic focus meetings where direction was discussed and set.

(9.2.3) Communications

This also touches upon another important aspect of the management approach: project communication. From the beginning, the project had a web site with several pages relating to background and project structure as well as an internal mailing list. The new PM, with the agreement of the AGLI and project members, implemented several additional measures which opened the majority of project communications to anyone who chose to view them, though the mailing list was still occasionally used for private, group−wide communications. This was somewhat akin to the software industry's Open Source initiative, which we therefore nicknamed ***Open Science*** (2).

The three initiatives (in addition to a revamped website at http://www.astrogrid.org/) were (3):

- ***News***:
  The News site (http://news.astrogrid.org/) is most often used for announcements to the whole project team (replacing the main use of the mailing list). Updates can be posted several times a day and content includes news stories, meeting, conference and other event announcements and polls. Any registered user can create these items and most allow the opportunity for feedback to be added.
- ***Forum***:
  The Forum (http://forum.astrogrid.org/) provides an area for topic−based discussions to develop as well as a place for novices to ask questions and those with problems to seek solutions. Again, any user can create a new topic or can add comments to existing topics.
- ***Wiki***:
  The Wiki (http://wiki.astrogrid.org/) is the most unusual and exciting area of the site. A wiki is a form of website that allows the registered user to change the content of the site, adding comments or new sections to existing pages or creating their own pages. The site includes help pages, a tutorial and an experimental wiki where users can try things out without fear of damaging existing content. It is divided into a number of webs: one for the AstroGrid project, general webs for VO, Grid and e−science topics, and support webs such as the tutorial and test pages.

These online facilities helped with keeping team members in touch with developments in other areas of the project but, with such a widely distributed project, it was equally important to ensure people had opportunities for direct contact.

The original plan was to have full consortium meetings, bringing together everyone involved with the project, every three months. This proved unmanageable, so it was decided that the consortium meetings would be held every six months: Sep'01, Dec'01, Jun'02 so far and the next will be in Jan'03.

At six month intervals, the consortium meetings were unsuitable as a means of keeping people up to date with progress with ongoing work. We therefore instituted ***Focus Meetings*** (4). These were subject specific meetings in which a small group of people interested in the subject could work together. The focus meetings were highly successful in developing the science problems, use cases and architecture.

(9.2.4) Control

Control of project activities was managed through the ***Work Package Managers*** (WPMs) each of whom reported on a quarterly basis (5). A meeting of all WPMs would be held just before the end of a quarter to review progress and plan tasks for the coming quarter: during the intensive architectural activity, this meeting would be subsumed by an architecture progress meeting, but the goals and ends were the same.

A quarterly ***Forecast Report*** contained:

- ***Summary***:
  A textual summary of the goals set for the coming quarter.
- ***Resources***:
  People and equipment which would be deployed on the work package activities.
- ***Planned Tasks***:
  A coded list of tasks (codes were retained from previous quarters if they overran) showing planned start and end dates and the people who would work on the task.
- ***Deliverables***:
  List of concrete deliverables that would come out of the tasks with due dates.

The quarterly ***Progress Report*** on the previous quarter's activities contained:

- ***Summary***:
  A textual summary of the achievements and issues arising from the past quarter.
- ***Resources***:
  People and equipment which was actually deployed.
- ***Tasks***:
  A list of the tasks from the forecast plus any unplanned activity. Each task showed actual start and end dates along with percentage complete.
- ***Deliverables***:
  List of actual deliverables produced with links to where they might be found.

As with everything else on the AstroGrid project, these reports were freely available on our web sites. Only personnel−related information and other sensitive material was kept offline.

(9.3) Personnel

In addition to the Lead Investigators, the AstroGrid team contained the following project funded personnel at the Sept 2001 start of Phase A :

Location	*Name*	*Position*	*FTE Aug 2002*	*sm to Aug 2002*
Leicester	Clive Page	RA/Developer	1.0	10
Cambridge	Guy Rixon	RA/Developer	1.0	12
Edinburgh	Bob Mann	RA	0.5	6
MSSL	Bob Bentley	RA/Developer	1.0	12

Edinburgh	Clive Davenhall	RA/Developer	0.5	6
RAL	David Pike	RA	0.5	6
RAL	Chris Perry	RA	0.5	6
Jodrell Bank	Anita Richards (to end March)	RA	0.0	3
Jodrell Bank	Ant Holloway (from 1−April)	RA	0.2	1
RAL	John Sherman	Co−ordination	0.2	2.4
RAL	Dave Giaretta	RA/Developer	0.2	2.4
RAL	Peter Allan	Lead Investigator	0.1	1.2
TOTAL			***5.7***	***68***

Note−1 : Ant Holloway replaced Anita Richards at Jodrell when Anita took up the AVO post below.

_Note−2 : FTE=Full Time Equivalent is the fractional effort as at August 2002. This is a rate of expenditure, so that someone rated at 0.5FTE is expending 50% of their effort on AstroGrid.

Note−3 : sm=staff months is the integrated effort expended to the end of August 2002. Likewise sy=integrated staff years, so that someone working at 0.2FTE for 3 years has expended 0.6sy.

The Lead Investigators, as academic staff, are not funded by the project, except for Peter Allan. (The RAL system requires charging for all effort used). All these staff were funded by PPARC. Funding for the above staff was routed through a variety of existing grants, except for RAL staff, who are funded through the PPARC−CLRC Service Level Agreement.

New recruits to the project were:

Location	***Funding***	***Name***	***Position***	***Start Date***	***FTE Aug 2002***	***sm to Aug 2002***
Cambridge	PPARC	Nic Walton	Project Scientist	November 1	1.0	10
Leicester	PPARC	Tony Linde	Project Manager	November 1	1.0	10
Jodrell Bank	AVO	Anita Richards	RA	April 1	1.0	5
Leicester	PPARC	Tim Goodwin	Web Developer	July 1	0.5	1
Leicester	PPARC	Patricio Ortiz	RA	July 1	1.0	2
Cambridge	PPARC	Kona Andrews	Developer	July 8	1.0	2
Edinburgh	AVO	Martin Hill	Developer	Aug 1	1.0	1
Edinburgh	AVO	Alan Maxwell	Developer	Sept 2	1.0	0
TOTAL					***7.5***	***31***

Recruitment of the later RAs was seriously delayed by issues relating to the funding of AVO posts, which had a knock−on effect of delaying the recuitment of the PPARC posts. This was finally resolved by switching one of the AVO posts from Cambridge to Edinburgh, and one of the PPARC posts in the reverse to provide the above. All posts on the project to date are of three year duration.

The size of the team (excluding the AGLI) is now 18 people, adding to 13.1 FTEs. Given the start dates, the effort expended to the end of August 2002 (including the cost of Peter Allan) is 8.25sy.

(9.4) Finances

The total Phase A expenditure up to the end of August 2002 has been approximately £661K. (This is the PPARC expenditure only).

AstroGrid finances have been split into personnel costs and non−personnel costs, so that the non−personnel costs could be centralised into a single central budget controlled by the Project Manager in Leicester. This was done in order to keep flexibility in non−personnel costs, and to make sure the PM had adequate control over expenditure.

(9.4.1) Personnel Costs

Staff costs have been expended through a variety of grants to different institutes. For University based staff, on top of salary, these costs include standard fractions of secretarial and system management support staff, and the standard PPARC grant overhead of 46%. Some of the grants concerned had other minor costs attached. For RAL based staff, time is charged at a uniform staff rate agreed by negotiation between PPARC and CLRC. A detailed breakdown has been provided to PPARC.

The total funded staff effort over the year has been 8.25sy, some of which, the AVO funded staff effort, is at zero cost to PPARC. Grant claims are not all made yet so the accurate final cost is not known, but our out−turn forecast for the one year staff−related costs to PPARC, including all the above related costs and overheads, is £492K.

(9.4.2) Non−personnel Costs : Financial Control Mechanisms

Non−personnel expenses were provided by PPARC through three grants, two for capital equipment and one central budget grant. These were seen however as representing a single budget controlled by the PM, who set out an overall budget plan in November 2001. Procedures were then put in place for managing these finances.

The procedures made each WPM responsible for forecasting expenditure during the coming quarter. These forecasts were reconciled and adjusted (if necessary) by the PM and then combined into the budget for the next quarter. Any team member needing to travel or make a purchase on behalf of the project claimed for the expenditure from his or her institution.

At the end of a quarter, each institution collated all claims made and invoiced Leicester for the total amount. The PM checked these invoices against the budget and authorised payment. It took some time to get this process established in the consortium institutes but it now seems to be working well. Invoices up to and including Q3 have been received from all institutes and two, to date, have been received for Q4.

Finances, as they are known at the time, are submitted to meetings of the AstroGrid Oversight Committee (6) for approval.

(9.4.4) Non−personnel costs : Expenditure Report

The current forecasted expenditure (based on invoices submitted to date plus forecasted expenditure for Q4) is:

Monthly Non-staff Forecast			
	Forecast		Budget
Travel	£80,000		£80,000
Equipment:			
Web Server	£4,000		£5,300
Leicester Beowulf	£30,000	!	£23,000
Cambridge Sun server	£18,000	!	£12,000
Other Servers	£4,500	!	£3,500
Laptops & Workstations	£25,200	!	£24,000
Other	£800		£1,400
Software	£3,500		£8,900
Consumables	£500		£500
Books	£2,000		£5,500
Training	£0		£5,200
Other	£800	!	£200
Total	£169,300		£169,500
Quarterly Total			
	Difference:		£200

Of these items, two stand out:

- ***Travel***:
  This figure is difficult to predict. Invoices to date only total £55K but Q3 and Q4 involved large amounts of travel by many of the team members. The reason for this is two−fold. Firstly, it is the traditional season of conferences in the academic community. Secondly, AstroGrid is now seen as a major player in the VO and Grid domains. We therefore were invited to attend and speak at many more meetings and conferences and, in order to maintain our lead in certain

areas (authorisation, data access and mining, astronomical ontology, to name a few) and to keep up in other areas, we sent team members to other meetings.
We anticipate that the final actual expenditure on travel will be much the same as that budgeted.

- ***Equipment***:
More was spent on the Leicester Beowulf cluster and Cambridge Sun server than was budgeted. This was due to expenditure being less in other areas than expected. We decided to use the underspend to increase the number of nodes in the cluster from 4 to 8, enabling a more realistic investigation of cluster−based data mining techniques, and to purchase a newer type of Sun server. Both of these machines were also upgraded to have around ***one terabyte*** of disk storage to enable the loading of complete data archives.

Phase A has been extended into a non−standard, four month Q5. This will allow the team time to complete the architecture and the demonstration projects. The budget for this extended quarter will be £15K for travel but may be further extended after the meeting of the Grid Steering Committee to consider the proposal for Phase B funding.

(9.6) References

(1) Unified Process: See the more detailed description in the Architecture Overview, elsewhere in this Phase A report.

(2) See also http://wiki.astrogrid.org/bin/view/Escience/OpenScience.

(3) Also described in PPARC Frontiers magazine, Issue 13, p26−27

(4) For a schedule of focus meetings, together with reports from them, see http://wiki.astrogrid.org/bin/view/Astrogrid/FocusMeetingNotes.

(5) All forecasts and reports are linked from the wiki page: http://wiki.astrogrid.org/bin/view/Astrogrid/WpReports

(6) AGOC: committee set up by PPARC to oversee the progress of the AstroGrid project. This committee has met twice so far; for reports from those meetings, see: http://wiki.astrogrid.org/bin/view/Astrogrid/OversightCommittee.

(7) The advertisement for this position (for which applications closed on 13th September) can be found at http://www.le.ac.uk/personnel/jobs/p9048a.html and a full description at http://www.le.ac.uk/personnel/jobs/p9048.html.

(8) Detailed descriptions of the work packages can be found by following links at http://wiki.astrogrid.org/bin/view/Astrogrid/WorkPackages.

(10) Phase B Plan

(10.1) Introduction

This document presents our plans for ***Phase B*** of the AstroGrid project. It assumes that Phase B will commence on 1st January 2003 (ie that Phase A is extended to 31st December 2002 to enable the completion of the system architecture and of several technology demonstration subprojects). The end date will be approximately December 2004, but in fact the effort profile will not be flat, and a small amount of staff effort will extend into 2005. The end goal of the project is to produce software which will enable the creation of a working, grid−enabled ***Virtual Observatory (VO)*** based around key UK astronomical data centres.

Our approach to the project is different from the usual work package−based approach of other scientific and academic projects. This approach is outlined in the next section but it is a continuation of the approach used within Phase A. The estimates of effort involved are presented along with an explanation of how they were derived. We then present the implications of these estimates on the personnel required on the project and some functionality milestones by which the project can be measured.

The overall goals, and the current state of the architecture, are described in other sections of the Phase A report. However, to set the scene, we re−iterate here our aims and goals.

These are our ***SCIENTIFIC AIMS***

- to improve the quality, efficiency, ease, speed, and cost−effectiveness 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.

And these are are our top−level ***PRACTICAL GOALS*** :

- to develop, with our IVOA partners, internationally agreed standards for data, metadata, data exchange and provenance
- to develop a software infrastructure for data services
- to establish a physical grid of resources shared by AstroGrid and key data centres
- to construct and maintain an AstroGrid Service and Resource Registry
- to implement a working Virtual Observatory system based around key UK databases and of real scientific use to astronomers
- to provide a user interface to that VO system
- to provide, either by construction or by adaptation, a set of science user tools to work with that VO system
- to establish a leading position for the UK in VO work

(10.2) General Approach

Our approach to Phase B will be based upon the ***Unified Process*** (UP, or our own variant of it, ***UPeSc***), as was Phase A (1). The UP is a software development methodology which is both ***iterative*** and ***incremental***: each iteration contains analysis, design, code, test and deployment activities and each iteration incrementally adds to the functionality of the system components.

Earlier development methods all made the fatal mistake of assuming that a system could be designed up−front and the rest of the project was simply a matter of implementing that design and then deploying the system. Statistics which show that 'Only about 10% of software projects are delivered successfully within initial budget and schedule estimates' (2) give the lie to this assumption, despite regular attempts to improve the estimation techniques of these methods.

Iterative methodologies like UP assume that an up−front architecture is all that is needed in order to estimate the resources required on a project but leave the detailed design work to each iteration (3). Each iteration has a fixed duration (we have chosen 3 month iterations at the outset). The technical management team determine which use cases will be completed during that iteration (a decision based on risk and priority) and then create teams of designers and programmers to complete these tasks. The end−point of each iteration is fixed and immutable − if a component cannot be completed in the timeframe, it is put aside for another iteration. At the end of the iteration, a working system should result.

Another shortfall of earlier methods was the assumption that software requirements do not change during the lifetime of the project build phase (and if they did, it was the users' or the analysts' fault). UP assumes that ***requirements will change*** as users begin to get to grips with the developing software. We will hold ***user workshops*** at the end of each iteration where astronomers will be invited to test the current release of software and provide feedback.

Each iteration selects the use cases to be completed in that iteration based on risks and priorities as they exist at the time. This also means that previously coded software will need to be changed. An important part of iterative development methods is ***refactoring*** (4), a technique to restructure code in a disciplined way.

At the beginning of Phase B we will establish procedures and install automated tools to enable the team to use the UP method in a more productive way.

(10.3) Management

(10.3.1) Project Policy and Oversight

Leadership of the project will be carried out in the same way as for Phase A. Overall responsibility rests with the seven ***AstroGrid Lead Investigators (AGLI)***, who will have monthly telecons. The AGLI will set policy, make decisions on overall resource allocation, and monitor progress. From amongst the AGLI, A.Lawrence acts as overall ***Project Leader (PL)***, although during the lifetime of the project this task can pass by agreement to another member of the AGLI. The PL works closely with the ***Project Manager (PM)*** and the ***Project Scientist (PS)*** to make a small executive team. We assume that AstroGrid will continue to be overseen by both the PPARC Grid Steering Committee, and the AstroGrid Oversight Committee.

(10.3.2) Work Management

The management approach to Phase B will wholeheartedly embrace the Unified Process (UP) approach. Given the potential conflict with the work package approach, and that the majority of the work will be either development of software modules or short−term (average about six months) research and demos, this phase will have no fixed work packages.

Management of the research activities will fall to either the Project Manager (PM) or Project Scientist (PS), depending on the technical or astronomical nature of the work. All software development activity will be managed by the ***Technical Lead*** (TLd). This is a new post which is currently being recruited for Leicester (7). He/she will be supported in that role by a ***Technical Support Panel*** (TSP), a small, cross−institute group of people (including PM and PS) with experience in the project, the science and software development.

At the beginning of each quarter, the TLd and the TSP will meet to determine the software to be written (according to UP precepts, by selecting the highest priority and highest risk use cases). They will assign tasks to individuals and groups of individuals. Where it is feasible, work on a given component will be spread across more than one institute, and individuals will work on several aspects of the AstroGrid over the course of Phase B. This will ensure an even spread of knowledge across institutes and people.

The TLd will implement procedures for source control and standards, daily builds and testing. He/she will be responsible for ensuring compliance with those procedures and standards. Quarterly task plans and forecasts will also be this post's responsibility.

(10.3.3) Financial Management

As for Phase A, financial planning and control will be divided into two parts : ***staff costs*** and ***central costs***.

Funding required to ***employ staff*** will be handled through the individual institutes where the staff are employed, and routed through a variety of mechanisms − existing grants, new grants, and the CLRC SLA. A separate agreement with each institute is reached concerning what items are included as staff related costs. It will normally include salary and employer on−costs, standard university overhead charges, and attributed time of support staff such as secretaries and computing support staff. It will usually not include equipment or travel, but may do in some cases. Although these funds are handled through each institution separately, the PM is still required to track these costs in order to understand the overall AstroGrid expenditure.

In addition to the staff−related costs, a ***central budget*** will be controlled by the PM, and held as one or more grants by the Leicester LI (M.Watson). For Phase B, this will include funds for travel; for major pieces of equipment; for paying for outsourced software development; for staff training; and for other general items. Some of these items will require AGLI debate

before expenditure. Others will be expended by project staff from local funds in an agreed controlled manner, and then invoiced quarterly against the Leicester budget.

In the absence of work packages, the responsibility for quarterly financial forecasting, to allow expenditure against the central budget, will be delegated to the TSP institute representatives (see above). This will resolve some of the problems we had with the work package approach in that people not in the same institute as the WPM were unsure what expense codes were set up.

(10.4) Project Work Plan

(10.4.1) General Points

As explained in the Project Vision section of the report, the work we anticipate breaks into a few major strands. (a) Continuing *research and development* will be needed, both to assess technologies and possible solutions as we go along, and to participate in the international work on *standardisation*. (b) The bulk of the AstroGrid work will be in developing the *software infrastructure* that will make a VO possible. (c) A series of *user tools* will be needed to actually make it possible to do science with AstroGrid. This can include portals, visualisation tools, analysis tools, datamining algorithms, workflow editors, and so on. This is a huge open−ended area, but only a modest part of the work of AstroGrid as most of the work here will be in adapting existing tools, and encouraging other software providers to provide new tools. In particular we expect to work closely with the Starlink team, and with the eDIKT development team in Edinburgh and Glasgow. (d) We need to construct an actual working system, using specific physical resource and database collections at the data centres that are part of the AstroGrid consortium. Much of this work will be done in collaboration with non−AstroGrid staff at those data centres, but with AstroGrid taking a central and leading role.

Given the iterative nature of the UP method and that each iteration can lead to new requirements being added and less useful ones being removed, it is obvious that we cannot provide the type of detailed estimates and week−by−week progress charts that are typical of project plans under the earlier methods (but since they were effectively useless in guaranteeing project success, this scarcely concerns us). The architecture, however, has provided a high−level view of the overall requirements of a Virtual Observatory and it is from this that we produce our estimates.

(10.4.2) Components of Work

The architecture allows us to specify the types of development required under the following headings:

- ***Component Services***
  This is the main activity of the build phase and concerns the building of the web and grid service−based components from which the VO will be constructed.
- ***Library Services***
  These are also service−based components but will be wrappers over, interfaces to or rewrites of existing tools or libraries. Some new tools will also be written specifically for AstroGrid.
- ***Portal & Client Programs***
  These are stand−alone programs with which the user will interact; they will make use of the service components defined above.
- ***Demonstrations***
  These activities are technology trials required to test whether certain technologies work the way we require, or to check how areas of research can be exploited.
- ***Research***
  This activity is, as it says, research into areas which are still insufficiently understood to be incorporated into the system. This includes the standards development programme.
- ***Test Implementations***
  This involves the implementation of working versions of the software in one or more data centres to test the feasibility of the software being developed.

At the end of this document is a detailed estimate of effort required in the above areas of activity. These estimates are only a first approximation at the detailed level but can be taken as an accurate guide to the overall effort required − i.e. where we exceed estimates in one area, we would expect to make up in another.

These are no more than guidelines based on previous experience but have allowed the development of estimates for required additional personnel and milestones based on component functionality. The estimates take into account the detailed design effort, coding, testing and refactoring of early code subsequent on new requirements.

The following subsections very briefly discuss aspects of the estimates under the above headings.

(10.4.3) Component Services

This is where most of the work will go. The largest components (in terms of effort devoted to them) are:

- application and compute resource
  These interfaces to other resources are likely to require much thought and development and are likely to be refactored each iteration as we test the components with different resources.
- AQL translator
  This component must parse the new AQL language and translate it into SQL and calls to other components, possibly constructing a mini−workflow.
- data mining and access
  Simple data access routines will be developed early on but the more complex access and datra mining routines are likely to be developed later in the project as the research and demos are winding down.
- job control & workflow
  These are typically difficult processes to get right: the job control having to interact with many other processes; and the workflow suffering from many requirements changes as people make more use of them.
- MySpace
  The hardest part of this is going to be making the 'space' look like one single resource across many distributed computers; and the issue of security complicates matters even more.
- resource registry
  How we construct and make use of this is a complex issue still undergoing research. We're likely to start with a simple lookup list of data centres or datasets and try to incorporate ontology−based lookups and metadata later.

(10.4.4) Library Services

As stated elsewhere, we'll need to build wrappers or interfaces to several existing tools or libraries; the effort on this will be dependent on what is required for the early demonstrations of the VO. Some research will be needed into this − we do not want to rewrite different code for every library. Some library tools might be better recoded. While AstroGrid will do some work in this area, we intend to engage with tool developers (especially Starlink and eDIKT) to help them port their applications to AstroGrid.

For AstroGrid to prove immediately useful to astronomers, we will commit to writing some analysis tools. In the first instance these will relate to our key science drivers (8)). Other tools may come from users of the early releases of AstroGrid.

Visualisation is a *hot topic* in grid circles. It is likely too complex for us to develop from scratch or contribute substantive research to but we'll need to cover the research underway, conduct our own to see which efforts are likely to benefit AstroGrid then take what is available and adapt to our circumstances.

(10.4.5) Portal & Client Programs

The portal is likely to be under development for the whole of the the build phase as new components are constructed and need to be integrated into it. Creating workflow tools that make the job easier for novices but don't hamper the work of experts is key to the portal's success. We will want to create a VO front−end which also allows easy addition of new components, along the lines of the Microsoft Digital Dashboard (now incorporated into SharePoint Portal Server (5)).

Much the same problems will apply to the Client so we're likely to delay development of this until later in the build phase.

(10.4.6) Demonstrations

In general, the demo projects are expected to take no longer than two iterations (6 months). Their purpose is to take the outcome of some aspect of research and test how the technology can be implemented within the AstroGrid project.

(10.4.7) Research

Research also should take no longer than two iterations. The goal, as always, is to find a way of incorporating the feature required into AstroGrid. Each of the areas under research has features which are either unknown at this stage or which have not been implemented in a grid (or even simple distributed) environment. Where possible, we will join our efforts to those of

other e−Science projects here, in Europe and world−wide. This is especially true for the crucial *standards development programme*, which is being carried out in collaboration with other members of the International Virtual Observatory Alliance (IVOA) and needs agreement at each stage. As each of these standards becomes agreed, their actual implementation in code, construction of parsers and so on, moves out of the *research* component and into various other components depending on the nature of the standard concerned.

(10.4.8) Test Implementations

The best way to test whether we have created tools which will enable a VO when installed in data centres is to install those tools in real data centres. This effort will require intensive on−site hand−holding. We expect to implement tools in more than one site though it is doubtful that we could cover all seven of the consortium institutes. We will also be looking for in−kind effort from the data centres themselves.

(10.5) Personnel Plan

(10.5.1) Estimate of Required Volume

The final section of this document tabulates our current estimate of staff effort needed in the components listed above, including a plan of how to dispose the effort versus time. These detailed estimates show a total effort required of 48 staff years. Assuming a two year Phase B, this equates to 24 staff. Note that these estimates do not include management, co−ordination, and support tasks.

At the moment, the project employs 13.1 FTEs spread across 18 individuals who are actively involved in the project. (This does not include the AGLI but does include 3.0 FTEs funded by AVO at no cost to PPARC). Of these, 2.7 FTES across 4 people are primarily employed in non−development tasks − management and co−ordination, science leadership, support tasks (Project Manager, Project Scientist, Web Developer (0.5 FTE), and RAL co−ordination (0.2 FTE)). This leaves 10.4 FTES spread across 14 indviduals available for development and related research tasks.

The project therefore needs an ***additional 14 development staff***. In addition to this, to deliver such a complex programme of software development, we need a new senior position − the ***Technical Lead*** who will directly co−ordinate developer tasks. In total then, we are asking PPARC for 15 additional staff.

(10.5.2) Staff skills and experience.

Many of the existing staff employed on AstroGrid are primarily researchers. All of these have an astronomical background, although many have been primarily active throughout their careers in astronomical software. The key requirement of the additional staff will be that they are experienced and active professional software developers. It is likely that we will need to recruit the majority of these staff from industry and will target advertisements with this in mind.

In recruiting these software developers, we are aware that they are unlikely to be offered extensions beyond the two−year period. We actually see this as a broad benefit of AstroGrid, as these individuals will return to industry taking grid and web service skills back with them.

In order to recruit such professional developers, we need to offer pay at a reasonably competitive level, although this will never be the main attraction of a temporary job in an astronomical project. We will mostly aim at young developers, keen to learn new skills in a forefront project. However, in order for such a distributed project to work, we need to recruit at least some reasonably *senior developers*, preferably one or more at each site.

(10.5.3) Staff deployment plan.

We have agreed internally a plan for deployment of the new staff across the AstroGrid institutes, with the following requirements in mind : (1) Create the main teams at three key centres (Leicester, Cambridge, Edinburgh); (2) Keep all the consortium institutes fully involved by at least one and preferably two appointments; and (3) Ensure at least one senior developer at each institute (including some already in place). This results in the following teams, with new positions labelled PPn for developer or PPn(S) for senior developer.

Leicester : 7.5 FTEs

Watson(LI), Linde(PM), TLd, Page(sci/dev), Ortiz(sci/dev), Goodwin(0.5,tech), PP1(S), PP2, PP3.

Cambridge : 5 FTEs

McMahon(LI), Walton(PS), Rixon(S,sci/dev), Andrews(dev), PP4, PP5

Edinburgh : 5 FTEs

Lawrence(PL), Mann(0.5,sci), Davenhall(0.5,sci/dev), Hill(S,dev,AVO), Maxwell(dev,AVO), PP6, PP7

RAL : 3.5 FTEs

Allan(0.1,LI), Sherman(0.2,coordn), Giaretta(S,0.2,sci/dev), Pike(0.5,sci), Perry(0.5,sci), PP8, PP9

Jodrell Bank : 3.2 FTEs

Garrington(LI), Richards(sci,AVO), Holloway(0.2,sci/dev), PP10(S), PP11.

MSSL : 2.0 FTEs

Harra(LI), Bentley(sci/dev), PP12.

Belfast : 2.0 FTEs .

Murtagh (LI), PP13(S), PP14.

Note that the FTEs of effort above include the AVO funded effort. We also expect that both RAL and MSSL will have further positions funded through the EGSO and SpaceGrid projects. We expect to collaborate with the relevant team members, but don't have the same formal relationship that we do with AVO, so do not formally estimate attributable staff effort. Rather, we simply expect that AstroGrid, EGSO, and SpaceGrid will simply share work to each other's mutual benefit.

(10.6) Financial Plan

(10.6.1) General Points

We are still refining our budget model, but our current estimate of the total requirement for the whole lifetime of the AstroGrid project is £4.5M, with £3.7M for staff−related costs, and £0.8M for hardware, travel, outsourcing, and other costs. We do not detail the overall financial request here, partly because of sensitive salary information, but partly also to avoid unnecessary detail in this report. Full information is being provided to PPARC's Grid Steering Committee, who will make a final budget allocation to the AstroGrid project. However here we summarise the main components of our budget.

(10.6.2) Staff cost elements

Effort expended to date on the one−year funded Phase A has been 8.25sy, 0.5 of which has been funded by AVO. The Phase A extension to the end of 2002 will be running at a rate of 13.2 FTEs (3.0 paid by AVO) for 4 months, coming to 4.4sy (1.0sy paid by AVO). The 15 additional staff for Phase B are planned to run for 2 years, coming to 30sy. Of the existing staff, not all will run to the end of the two year Phase B, and some will extend past the official end−date. This is partly inevitable given various contract constraints, but anyway desirable in order to have a smooth wind−down of the project. Our current estimate is that in Phase B these existing staff will expend a further 25.25sy, of which 19.75 requires funding from PPARC.

The total expended staff effort integrated over the lifetime of the AstroGrid project is therefore curently estimated as 68sy, and the required funding from PPARC is 61sy. We are still refining our staff cost model, but this is likely to require a total of approximately £3.7M funding.

(10.6.3) Non−staff cost elements

Our total intended 2 year non−staff budget is £475K. The non−staff costs for Phase A (including the extension) are forecast as approximately £200K.

The new 2 year budget is made up as follows :

Travel : budget £200K. The multiple e−science and international VO connections has required a large amount of travel. This has been absolutely crucial to the success of AstroGrid, and will continue to be necessary. So far we have expended on travel at a rate of approximately 6K/hd/year. We expect the travel for new developers to be much less demanding, but they will still need to attend frequent team meetings. With roughly 12 core members at 6k/hd/yr, and 12 others at 3K/hd/yr, this would give 216K.

Personal Equipment : budget £100K. This is intended to keep staff supplied with up−to−date workstations and laptops. Costed at the standard PPARC rate of 2K/hd/year for 24 active staff gives 48K/year.

Special Equipment : budget £50K. Most of the physical resource needed to construct the AstroGrid system will be donated by the participating data centres, and either already exists or is in their respective budgets. (For example, the WFCAM archive plans to purchase a 50 node Beowulf cluster, and also to use a share of the new NeSC 32 processor SMP machine). However, AstroGrid needs to provide at least some storage and search−CPU for MySpace, and to establish at least one data warehouse. In Phase A, AstroGrid purchased a small cluster in Leicester and a high−end server in Cambridge, with around a TB of storage. To complement this, we are likely to need around 20 TB more with attached CPU, costing around 50K. The plans in this area are still very uncertain, so that this figure represents a reasonable amount to set aside.

General : budget £25K. We anticipate continued expenditure on books, training, software, and consumables. To date, this has run at around 500/hd/year, so a similar level for 24 staff gives 12K/yr.

Outsourcing : budget £100K. We anticipate that most of what we need to construct will be either be done by our own staff, or written by other projects such as Starlink or eDIKT or the other VO projects, or other e−Science projects such as MyGrid, and available to us at no cost. However we envisage two kinds of situation where we might want to effectively buy software from others. Firstly, small very specialist chunks may well be best done by commercial companies. Second, other UK groups who can write relevant software (such as user tools) may in practice only be able to do so with extra grant resource. It is hard to judge how much of either of these kinds of outsourcing may be necessary, but we feel strongly it is wise to keep some budget aside for this purpose.

(10.7) Milestones

From the outset, AstroGrid has strongly believed in setting goals and reporting progress towards them. Every three months, for each work package, we have published a report on the previous three months' work and a forecast of the next three months' goals (6).

For Phase B however, as can be seen from the above explanation of the ***iterative and incremental*** approach, setting milestones several months before the project has started for periods some two and a half years ahead is not only problematic but counter to the spirit of the UP methodology. We do recognise, however, that it is important that the community and those providing our funds can see and measure progress on the project.

To that end, we have devised the following milestones for Phase B of the project. They are phrased in terms of system capability as we fully intend to have a working VO−like system available from at least the second iteration. We will report to the AstroGrid Oversight Committee (AGOC) (7) on progress against the milestones. Where it is decided to change the use cases to be realised in an iteration such that it impacts one of the milestones, we will likewise report that to the committee.

NOTE: These milestones are preliminary. As we complete the architecture by December 2002, so we will publish the final list of milestones on our web site and to the AGOC. These are likely to be in use case format.

The milestones are set every six months (two iterations) and refer to the major features of the system:

- ***Iteration 2 (June 2003)***
  - ♦ ***Portal***:
    - ◊ astronomer will be able to log into portal, recall preferences and past activities, set preferences, and search for registered resources (probably only catalogs)
  - ♦ ***Dataset Access***:
    - ◊ from the portal, astronomer will be able to select two catalogs and run simple JOIN across the two
  - ♦ ***MySpace***:
    - ◊ astronomer will reserve space at specific data centre and have results of query returned there

- ***CAS Server (demo)***:
  - community administrator will be able to create community on demonstration CAS server, register and de−register members, create and modify groups of members and set data access rights for both groups and members
  - data centre administrator will be able to create permissions for access to a resource for community, group and member

- ***Iteration 4 (December 2003)***
  - ***Analysis tools***:
    - Ability to run server−based analysis tools against datasets
  - ***CAS Server***:
    - CAS server integrated into system with above functionality
    - access to data resources is governed by community, group and member rights combined with data permissions
  - ***Job Control***:
    - user can manually construct set of sequential tasks
    - job control will monitor progress of tasks
    - user can query job control for task progress and completion
  - ***MySpace***:
    - user can request space on number of distributed servers
    - tasks will make use of scratch space on servers, moving results into MySpace area
    - user can make seamless queries on distributed MySpace
  - ***Data Mining (demo)***:
    - for complex queries, data can be loaded in optimal form
    - custom algorithms can be run against loaded data
  - ***Registry/Workflow (demo)***:
    - drag−and−drop workflow editor can be used to construct job
    - workflow editor can query registry for details of resources

- ***Iteration 6 (June 2004)***
  - ***Data Mining***:
    - user can move data resources to warehouse and run complex queries
  - ***Workflow***:
    - workflow editor integrated into portal
    - jobs can be stored and rerun with different parameters
    - workflow editor will detect if joined tasks are incompatible, suggesting data converter if possible
  - ***Visualisation***:
    - results of job can be visualised via server
  - ***OGSA Integration (demo)***:
    - system is implemented on grid servers
    - all authorisation and permissioning via grid certificates

- ***Iteration 8 (December 2004)***
  - ***Data Mining***:
    - user can upload algorithm code to mining centre
    - user can request run of complex algorithm on data in warehouse
  - ***Query Estimator/Optimizer***:
    - workflow can estimate job/task length
    - queries are optimised before being run
  - ***Resource Registry***:
    - registry has ontology−based metadata and inference engine to assist with queries
  - ***Visualisation***:
    - visualisation integrated into workflow and portal
  - ***OGSA Integration***:
    - system fully OGSA compliant

(10.8) References

(1) See the more detailed description in the Architecture Overview, elsewhere in this Phase A report.

(2) from: Software Project Management: A Unified Framework, Walker Royce, Addison−Wesley, 1998, p5: http://www.awprofessional.com/catalog/product.asp?product_id={53E5E335−C3FD−4D92−BDB6−679A63DBFC2F}

(3) Some variants of the eXtreme Programming (XP) methodology eschew even the architecture, preferring to simply collect a number of use cases (user stories in the parlance) and then begin coding. Look at http://www.extremeprogramming.org/ for more information on XP.

(4) See: http://www.refactoring.com/

(5) See: http://www.microsoft.com/sharepoint/evaluation/features/default.asp

(6) See: http://wiki.astrogrid.org/bin/view/Astrogrid/WpReports

(7) AGOC: committee set up by PPARC to oversee the progress of the AstroGrid project. This committee has met twice so far; for reports from those meetings, see: http://wiki.astrogrid.org/bin/view/Astrogrid/OversightCommittee.

(8) For an explanation of the AstroGrid key science drivers, see 'The Science Analysis Summary', another document in these Phase A Reports.

(10.9) Detailed Effort Estimates

The following chart provides, for each component considered, the following estimates:

- ***s.mths***: staff months effort required (summed on the right as ***Total s.mths / s.yrs***)
- ***s.itns***: effort in 'staff iterations' = s.mths / 3 (assuming three month iterations)
- ***itns***: expected number of iterations required to complete this component
- ***staff***: computed staff therefore required per iteration to work on that component

	Effort				Total Effort	
	s.mths	s.itns	itns	staff	s.mths	s.yrs
Components					285	23.75
Activity Log	6	2	2	1.00		
Analysis tools	18	4	4	1.00		
Application Resource	15	5	3	1.67		
AQL Translator	24	8	3	2.67		
CAS Server	12	4	3	1.33		
Compute Resource	18	6	3	2.00		
Database Export	9	3	2	1.50		
Data Mining	18	6	2	3.00		
Dataset Access	18	6	3	2.00		
Data Router	6	2	2	1.00		
Job Control	24	8	4	2.00		
Job Estimator	9	3	3	1.00		
Job Scheduler	9	3	2	1.50		
MySpace	24	8	4	2.00		
Query Estimator/Optimizer	12	4	2	2.00		
Replica Builder	9	3	1	3.00		
Resource Registry	24	8	4	2.00		
User Notification	6	2	2	1.00		
User Preferences	6	2	2	1.00		
Workflow	18	6	2	3.00		
Library Services					36	3.00
Wrappers	18	6	6	1.00		
Visualisation	18	6	3	2.00		
Programs					54	4.50
Portal	24	8	4	2.00		
Client	18	6	2	3.00		
Log Analyzer	12	4	2	2.00		
Demonstrations					72	6.00
Data Centre	9	3	2	1.50		
Data Federation	6	2	2	1.00		
Data Mining	6	2	2	1.00		
CAS / Permissioning	12	4	2	2.00		
File Transport	9	3	2	1.50		
OGSA Integration	9	3	2	1.50		
Ontology/Registry/Workflow	6	2	2	1.00		
User Notification	6	2	2	1.00		
Visualisation	9	3	2	1.50		
Research					78	6.50
AQL	12	4	2	2.00		
Data Federation	9	3	1	3.00		
Data Warehouse/Mining	12	4	2	2.00		
Job/Query Estimation	12	4	2	2.00		
Ontology/Registry/Workflow	9	3	2	1.50		
Server-based analysis tools	6	2	2	1.00		
User Notification	6	2	2	1.00		
Visualisation	12	4	2	2.00		
Implementation					48	4.00
Data Centre	12	4	2	2.00		
Portal / Client	6	2	2	1.00		
CAS	12	4	2	2.00		
Data Mining Centre	18	6	2	3.00		
TOTAL					573	47.75

One possible deployment scenario could be:

	itns	staff	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 6	Iteration 7	Iteration 8
Components										
Activity Log	2	1.00			1.00	1.00				
Analysis tools	4	1.50			1.50	1.50			1.50	1.50
Application Resource	3	1.67		1.67	1.67				1.67	
AQL Translator	3	2.67					2.67		2.67	2.67
CAS Server	3	1.33			1.33	1.33	1.33			
Compute Resource	3	2.00			2.00	2.00				2.00
Database Export	2	1.50	1.50	1.50						
Data Mining	2	3.00						3.00	3.00	
Dataset Access	3	2.00	2.00	2.00	2.00					
Data Router	2	1.00	1.00	1.00						
Job Control	4	2.00			2.00	2.00			2.00	2.00
Job Estimator	3	1.00					1.00	1.00		1.00
Job Scheduler	2	1.50					1.50	1.50		1.50
MySpace	4	2.00	2.00		2.00	2.00	2.00			
Query Estimator/Optimizer	2	2.00							2.00	2.00
Replica Builder	1	3.00			3.00					
Resource Registry	4	2.00		2.00	2.00				2.00	2.00
User Notification	2	1.00					1.00		1.00	
User Preferences	2	1.00			1.00	1.00				
Workflow	2	3.00		3.00				3.00		
Library Services										
Wrappers	6	1.00	1.00	1.00	1.00	1.00	1.00	1.00		
Visualisation	3	2.00						2.00	2.00	2.00
Programs										
Portal	4	2.00	2.00	2.00		2.00	2.00			
Client	2	3.00						3.00	3.00	
Log Analyzer	2	2.00					2.00	2.00		
Demonstrations										
Data Centre	2	1.50	1.50			1.50				
Data Federation	2	1.00		1.00						
Data Mining	2	1.00				1.00	1.00			
CAS / Permissioning	2	2.00	2.00	2.00						
File Transport	2	1.50		1.50				1.50		
OGSA Integration	2	1.50					1.50	1.50		
Ontology/Registry/Workflow	2	1.00	1.00			1.00				
User Notification	2	1.00						1.00	1.00	
Visualisation	2	1.50				1.50	1.50			
Research										
AQL	2	2.00				2.00	2.00			
Data Federation	1	3.00	3.00							
Data Warehouse/Mining	2	2.00		2.00	2.00					
Job/Query Estimation	2	2.00	2.00			2.00				
Ontology/Registry/Workflow	2	1.50	1.50	1.50						
Server-based analysis tools	2	1.00		1.00			1.00			
User Notification	2	1.00				1.00	1.00			
Visualisation	2	2.00		2.00	2.00					
Implementation										
Data Centre	2	2.00						2.00	2.00	
Portal / Client	2	1.00				1.00				1.00
CAS	2	2.00						2.00		2.00
Data Mining Centre	2	3.00							3.00	3.00
Total Staff			20.50	25.17	24.50	24.83	22.50	24.50	26.84	22.67
(diff to 24)			3.50	-1.17	-0.50	-0.83	1.50	-0.50	-2.84	1.33