AstroGrid

Submission to PPARC e-science AO
April 2nd 2001

Peter Allan, Bob Bentley, Clive Davenhall, Simon Garrington, David Giaretta, Louise Harra, Mike Irwin, Andy Lawrence, Mike Lockwood, Bob Mann, Richard McMahon, Fionn Murtagh, Julian Osborne, Clive Page, Chris Perry, Dave Pike, Anita Richards, Guy Rixon, John Sherman, Richard Stamper, Mike Watson.

School of Computer Science, Queens University of Belfast
Institute of Astronomy, University of Cambridge
Institute for Astronomy, University of Edinburgh
Dept of Physics and Astronomy, University of Leicester
Mullard Space Science Laboratory, University of London.
Jodrell Bank Observatory, University of Manchester
Space Data Division, Rutherford Appleton Laboratory

www.astrogrid.ac.uk

SUMMARY

A tidal wave of data is approaching Astronomy, requiring radical new approaches to database construction, management, and utilisation. The next two explosion points are in 2003 and 2005 when UKIRT WFCAM and VISTA (both UK facilities) come on-line and start producing hundreds of Gbytes of data *every night*. The UK also has a lead role in other key databases accumulating now in X-ray, solar, radio, and space plasma astrophysics. These datbases present rich scientific opportunities, but also serious worries; current data manipulation methods will not scale to these volumes. The UK is in a position of strength internationally because of the importance of these datasets, but will remain in a strong position only if it solves the data-handling problems. AstroGrid proposes to deliver the tools needed to meet that challenge.

At the same time, a new style of science is emerging, which requires the searching, filtering, manipulation, or browsing of *entire* vast datasets. Sometimes this is because the scientific problem requires massive analysis (e.g. the spatial clustering power spectrum of a billion galaxies), sometimes because rare events or objects need to be found (e.g. locating the one in ten million faint stellar images that *might* be one of the first generation of pre-galactic stars) or sometimes because one wants to examine, explore, and get ideas. These systematic and "data-mining" styles of working are already a strength of a minority of astronomical specialists, but require a long difficult slog, and will become impossible as the data explosion hits unless we transform our technology. In fact our intention is to develop hardware, architecture and data management solutions based round a small number of expert data centres, along with query, analysis and exploration tools that will make it *easier* and *quicker* to take this approach from any desktop, thus democratising it. Everybody can be a power-user. We therefore expect an acceleration of real science.

The real fruits however will come from the ability to take this approach to multiple databases simultaneously. For example, one wants to search solar and STP databases to find a coronal mass ejection followed by its effect on the Earth a little later; to browse optical, radio and X-ray images of the same piece of sky taken by different research teams, to click on an interesting object and to find to ones surprise that somebody just got an IR spectrum of it at UKIRT last month ... and of course to retrieve the calibrated spectrum immediately. This kind of science is already done, but slowly and painfully by hand, as the databases are spread all over the globe, with a variety of formats, access levels and policies, and it is hard to keep up to date with what's available. A key aim of astronomers world-wide is to stitch together essentially all such databases in a "Virtual Observatory" together with intelligent discovery agents so that the archive becomes like a second sky. This is a very ambitious aim, with many technical and organisational challenges.

The aim of the UK AstroGrid project is to focus on short term deliverables, both relevant application tools and the federation, by the data centres that manage them, of key sky survey datasets, namely: (a) SuperCOSMOS, Sloan, INT-WFC, UKIRT WFCAM, XMM-Newton, Chandra, MERLIN and related VLA datasets; (b) SOHO and Yohkoh; and (c) Cluster and

EISCAT. The differences between the data types involved in these federations means that each brings distinct challenges, whose solutions will shed light on generic problems to be faced by the developing global "Virtual Observatory". We set out below a three year programme, starting straight away and estimated to cost £4M (on the assumption of existing funding to establish the archives and provide on-line storage) which will add value to existing UK astronomy resources in the short term, as well as positioning the community strongly with respect to wider Grid initiatives, in astronomy and beyond.

The goals of the AstroGrid project are :

►A working datagrid for key UK databases
►High throughput datamining facilities for interrogating those databases
►A uniform archive query and data-mining software interface
►The ability to browse simultaneously multiple datasets
►A set of tools for integrated on-line analysis of extracted data
►A set of tools for on-line database analysis and exploration
►A facility for users to upload code to run their own algorithms on the datamining machines
►An exploration of techniques for open-ended resource discovery

Many of these goals are common to other nations and other disciplines. We will work in collaboration with related projects worldwide to deliver these goals.

CONTENTS

8.3 Budget estimates.

(1) PREAMBLE

The AstroGrid project began to emerge during 1999 and 2000 as several groups concerned with upcoming large databases made submissions to PPARC's long term science reviews. The review process placed IT infrastructure in general, and large database initiatives in particular, in the "Priority 1" shopping list. The opportunity to do something about it occurred as it became clear that the OST wished to fund "e-science" initiatives. The AstroGrid consortium formed during summer 2000, and felt convinced that a major and coherent project was needed for PPARC to put forward in the CSR process, if Astronomy was to stand any chance of getting a slice of this cake. In October 2000 we presented our plans as a formal proposal to Astronomy Committee, followed by a progress report to Astronomy Committee on March 2nd. The project also began a process of community consultation, circulating the proposal as a position paper, issuing a Call For Comments, and holding a workshop in Belfast. At the same time, contacts began with related international projects, notably the US NVO project, and the European AVO proposal, in which AstroGrid is a partner.

In the meantime, PPARC issued an AO for proposals in the e-science area. The current document is the AstroGrid submission to that AO. It repeats the general science case from the October 2000 proposal (with some updates, notably in the area of radio astronomy), but goes on to provide a management plan, budget plan, more detail on the activity to be undertaken, and a detailed Phase A plan with Gantt chart.

(2) GENERAL SCIENCE CASE

(2.1) COLLECTIVISATION AND EMPOWERMENT OF THE INDIVIDUAL

Over three decades astronomy has moved inexorably towards a more communally organised approach, but paradoxically this has increased the power of individuals to undertake research. First, it became normal for observatories and satellites to be equipped with "common-user" or "facility class" instruments rather than private equipment, built to a high standard of robustness, simplicity of use, and documentation. Most astronomers did not need to build equipment, only to learn to drive it, and so could get straight to the science. More than anything else, this change led to the flowering of astronomy in the UK. The next step was the communal development of data reduction tools (Starlink, IRAF, Midas) which again shortened the route to scientific results. Then came the provision of calibrated electronic archives, with some simple tools, but mostly intended for file download and open-ended use. Over the last few years, another big change in our working lives has come from the development of on-line information services (ADS, NED, Simbad, astro-ph, LEDAS, HEASARC). The latest phase, still ongoing, is the collectivisation of data collection - i.e. large pre-planned projects producing coherent databases undertaken by organised consortia (MACHO, 2dF-z, SDSS, VISTA). The classic astronomer's individualist approach will re-assert itself as the databases are opened up and trawled.

This paper concerns itself with what we see as the next steps in this direction. The first step is for key databases, and eventually all databases, to become *interoperable*, i.e. to become seamlessly jointly browsable and queryable in a simple way. The second step is the communal development of database analysis and exploration tools, to enable easy *data mining*. The third step is the development of *information discovery tools*. The overarching idea is that data, computing power, and services are distributed but that the astronomer uses a single simple interface - she plugs into the information grid. We shall explain these ideas in a little more detail in later Sections. These developments will make it easier for astronomers to do science, and will enable exciting new kinds of science. Furthermore, as we shall see, the explosion of data volume that we are facing means that developments of this kind are needed anyway if we are not to grind to a halt.

(2.2) GROWTH OF DATA AND ARCHIVES

Today a substantial fraction of astronomical research papers are based wholly or in part on archive data. The archive is a second sky. A recent ESA paper reports that 10% of the ISO archive is downloaded every month, and that the volume equivalent of the whole archive has been downloaded twice. For the HST archive the volume of archive downloads is growing faster than the accumulated data, and is now at around 20 GB/day. The HST archive is growing at 1-2 TB/yr and currently holds ~ 7 TB. Likewise the SOHO archive is growing at 1TB/yr. This is also fairly typical of major ground-based observatories around the world. (Recall that Giga=10^9, Tera=10^{12}, Peta=10^{15} ; also note that B=bytes and b=bits). This data growth is largely a consequence of the growth of detector size, which has been exponential for three decades. Quite soon now VISTA will be using a Gpixel array.

A benchmark data volume is that needed to store a 16 bit image of the whole sky with 0.1" pixels - 100 TB. The raw data may be substantially larger depending on the observing technique, but will rarely be needed by users. On the other hand, the normal "science database", with which nearly all users will normally interact, may be an order of magnitude smaller. Astronomical object catalogues derived from current surveys are typically ten times smaller than the pixel data. The pixel image itself can also be compressed by about an order of magnitude without losing much information using a variable blocking technique such as the H-compress algorithm developed at STScI. So a simple survey may require PB-scale shelved-tape storage (raw data for occasional re-reduction), 100TB near-line storage (full pixel data), and 10TB on-line storage (compressed pixels and objects). For multi-passband surveys, large-scale monitoring programmes, and spectroscopic surveys, or sampled in situ data (as for STP satellites) the volume mutiplies correspondingly.

(2.3) THE NEW ERA OF BIG SURVEYS

Sky surveys have been of great historical importance, and a particular UK strength (3C, Ariel-V, IRAS, UK Schmidt/COSMOS/APM). Several US surveys are currently attracting attention as representing the new age of digital surveys with on-line archives - SDSS, 2MASS, FIRST, and NVSS, each of which are of multi-TB sizes. The real data explosion will come however with three UK projects which will be the premier survey facilities on the world scene during the coming decade - the UKIRT Wide Field Camera (WFCAM), VISTA, and e-MERLIN, each of which will accumulate PB size archives. Finally the most ambitious (as yet unfunded) project is the US Large Synoptic Survey Telescope (LSST) which aims to map most of the sky every night! Here we provide a few details on some current UK projects, followed by plans for WFCAM, VISTA, and e-MERLIN.

The SuperCOSMOS and APM Programmes

The SuperCOSMOS Sky Surveys programme (see www-wfau.roe.ac.uk/sss/) is providing on-line access to digitised scans of photographic plates covering the whole southern hemisphere, in three bands (BRI) and with two epochs at R. Currently 5000 sq. deg. of the Southern Galactic Cap are available, from which users can extract pixel data for regions up to 15x15 arcmin and/or user-designed object catalogues covering up to 100 sq. deg., while the eventual southern sky survey database will be ~2 TB in size. Multi-colour object catalogues from the APM high-latitude scans with 500 million objects are also available. The vast areal coverage and, in particular, the temporal baseline, of the survey data makes this a very valuable resource even in the era of CCD-based photometric surveys, but to make full use of this important legacy dataset within the context of the developing "Virtual Observatory" will require the database federation and data-mining tools that AstroGrid would provide for the community.

The INT Wide Field Survey

The INT Wide Field Camera survey programme uses 20% of UK and NL time on the 2.5m INT to carry out publicly accessible multi-colour and multi-epoch surveys. The survey was carefully designed by a consortium of astronomers to be appropriate for many different projects, and to be a long term resource. The raw and pipeline processed images have no proprietary period in the UK and NL. Broadly the goal is to cover 100 sq.deg of high galactic latitude sky in the u,g,r,i,z wavebands with multi-epoch data in two bands over 50 sq.deg. The survey is a factor of 5 deeper than the SDSS over the whole region .i.e. m(r)=24.5 (5sigma). Smaller regions of around 10 sq.deg. are being surveyed 3-4 magnitudes deeper. Infrared coverage of the same area of sky with CIRSI has been proposed. The survey is described, and initial data products are available at www.ast.cam.ac.uk/~wfcsur/.

The XMM Survey Science Centre

XMM-Newton, by virtue of its sensitivity and wide field of view, is producing the largest and deepest serendipitous X-ray source survey to date with a projected estimated content of 500,000 objects for an assumed 10 year mission lifetime. The XMM-Newton serendipitous source catalogue will clearly become a major resource for a wide range of astrophysical projects. The XMM-Newton Survey Science Centre (SSC), an international collaboration led by Leicester University, has the responsibility within the XMM-Newton project of producing and maintaining the XMM-Newton serendipitous source catalogue, as well as coordinating follow-up identification programmes aimed at delineating the nature of the faint X-ray source populations. The data volume, excluding optical follow-up data, will be around 4TB.
See xmmssc-www.star.le.ac.uk/.

The SOHO data archive

The ESA/NASA Solar and Heliospheric Observatory (SOHO) spacecraft is the principal space-based research tool for the solar community world-wide. Official ESA data-archives have been set up at the NASA Goddard Space Flight Center near Washington, at the Institut d'Astrophysique Spatiale near Paris, and at the Rutherford Appleton Laboratory. The RAL archive contains all SOHO data, which can be accessed via a dedicated Web site. SOHO has been in operation for almost 5 years and operations are anticipated to continue to at least 2003. The mission produces about 1 terabyte per year. The RAL archive also includes all data from the associated NASA TRACE (Transition Region and Coronal Explorer). UK investment in SOHO has been very significant, with one UK-led instrument, another instrument with a significant UK hardware contribution, and scientific involvement in most of the 12-instrument payload. Many UK research groups regularly use SOHO and this has maintained a very strong position in world-wide solar physics for the UK. Data from the mission are regularly used by a dozen UK research groups spread around the UK. Along with SOHO, TRACE, and Yohkoh, the next few years will see more high-profile solar missions with strong UK involvements (Solar B, Stereo, Solar Orbiter) with a wide variety of instruments and data types - there is an urgent need to be able to explore all these data coherently.

The Cluster data archive

The Cluster Joint Science Operations Centre (JSOC) and Co-ordinated Data Handling Facility (CDHF) for Solar-Terrestrial Physics data, including Cluster, both located at RAL, provide a natural location for a full Cluster archive. As yet, ESA has not selected official archive sites but, as with SOHO, we would anticipate an RAL facility. This would serve the strong UK Cluster community which includes a number of university groups, such as Imperial College, MSSL, QMW and Sheffield as well as RAL. This would not be a large archive, compared to SOHO, for example, producing about 1 terabyte for the entire mission. However, exciting prospects come from the prospect of being able to search and examine space-based in situ particle and wave data that are contemporaneous with ground based radar and lidar data, or shortly after coronal mass ejection events studied by solar observatories with imaging and spectroscopy, and in the future even heliospheric in situ data. This involves many different datasets with a rich variety of data formats. This kind of work is attempted now but is extremely difficult in practice, and so presents a classic opportunity for AstroGrid.

UKIRT WFCAM

By early 2003, UKIRT will have a new wide-field camera (WFCAM: see www.roe.ac.uk/atc/projects/wfcam/), which will be the most capable JHK survey instrument in the world. Through a large allocation of UKIRT time over several years WFCAM will undertake both private PATT programmes and a large coherent public survey programme undertaken by the UKIDSS consortium (see www-wfau.roe.ac.uk/ukidss). Together these will provide the UK community with an archive of near-infrared photometric data unrivalled in the northern sky. For example there is a plan to image ~4000 sq. deg. of the Sloan Digital Sky Survey region to matching depths in JHK, yielding an 8-band UV/optical/near-infrared photometric database unique in the north, and only surpassed in the south by another UK project, VISTA. Federation of the SDSS and WFCAM archives is a key science goal for AstroGrid, and a concrete deliverable on the path to development of a global "Virtual Observatory".

VISTA

One of the most exciting prospects for ground-based astronomy in the UK in the coming decade is VISTA (see www.vista.ac.uk), a 4m optical/near-IR survey telescope funded by £24.8M from JIF and to be located in Chile. VISTA will dominate post-Sloan survey astronomy, thanks not only to its technical capabilities, but also to its sophisticated observational strategy: rather than produce a monolithic sky atlas to uniform depth, as traditional in wide-field astronomy, it will undertake a set of surveys of varying depth and areal coverage, as required to meet the needs of a series of key programmes, expressing the scientific goals of the 18 UK universities in the VISTA consortium. VISTA will produce ~1TB of raw data each night, yielding a vast (~300TB) science archive after ten years of operation: the VISTA database represents the sternest challenge for UK wide field astronomy, but, with the tools deliverable by AstroGrid, it can provide the UK community with a world-beating resource, that will drive research for several decades.

MERLIN, e-MERLIN, and EVLA

The wide-field capabilities of high-resolution radio interferometers are often overlooked. MERLIN can produce 20k pixel square images and EVN can produce 200k pixel square images. The CPU/ storage resources (and determination!) to tackle such projects have only just become available, but as the HDF, XMM and Trapezium cluster projects have shown, the scientific return from these high sensitivity wide-field multi-wavelength studies is worth the effort. The real challenge however is the ability to allow on-the-fly image construction from visibility data, especially combining visibility data from different telescopes. The proposed e-MERLIN and EVLA facilities will transform our radio capabilities, providing an order of magntitude sensitivity increase and default wide-field capability, but at the cost of a 100-1000 fold increase in the raw data rate - hundreds of Gbytes per day, as challenging as WFCAM or VISTA. A single e-MERLIN observation will produce a wide-field image much more sensitive than even the longest current radio Deep Field, for free. The ability to search these background fields, which would contain each thousands of radio sources, in conjunction with the new generations of optical/IR and X-ray surveys is an important aim for AstroGrid.

(2.4) NEXT STEPS IN USE OF ARCHIVES

Current use of archives can be broken into three types. First, and most common, is *support of other observations*, i.e. examination of small images ("finding charts") and other datasets, for use as illustration, as exploration, and as assistance in other research programmes. Second is *directed research*, the download of specific observation datasets for off-line analysis by the user. Third is *discovery based programmes* - the "power-user", running statistical analyses of huge amounts of data, re-reducing the data in novel ways, searching for rare objects and so on. The aims of AstroGrid are to enhance all three classes of use, but we have the power user particularly in mind, as we believe an increasing fraction of astronomers will want to become power users. To date, this sort of work has been the restricted domain of specialists, requiring privileged access to the datasets, detailed technical knowledge, and months of work. The advent of the new generation of databases will motivate many more astronomers to become power users, as new science becomes possible, through trawling the multidimensional data spaces made available by database federation. The aim is to make this kind of science much faster and easier and as standard as reducing your spectrum with FIGARO or IRAF.

The idea of *interoperability* is that one can browse and query multiple databases from a single point of contact. For example given a region of sky, one could log-on to AstroGrid and view the Sloan, WFCAM, and XMM-Newton images simultaneously. Then one could click on an interesting object and download the X-ray spectrum. Or one could issue joint queries, such as "give me a flux-ordered list of all the XMM-Newton sources in this RA and Dec range with optical colours redder than X that have no radio counterpart". Such things can be done now with the small current archives, and with considerable effort. The aim is to log on to AstroGrid and undertake such explorations automatically without having to know where the databases are, or having to learn six different packages, and for all this to work fast enough to be interactive. More ambitiously, one would also want a more sophisticated level of *visualisation and exploration* - very large image plotting and scrolling (possibly on power-walls), interactive subset selection and plotting, projections of multi-dimensional datasets, 3D examination with CAVEs or VR headsets and so on. Next we want large data-set *manipulation tools* - Fourier transforms, kd-trees, Gaussian mixture models, fuzzy joins, finding of outliers and so on. We will never invent enough analysis tools for the most imaginative astronomers, and so we also need a method for *uploading user code* to run on the databases. Finally and most ambitiously we need *information discovery tools*. After finding that interesting object in the federated XMM/Sloan/WFCAM databases, one would want to send off a search agent which returns with a report that this is probably the same object observed at Kitt Peak last week. Or with a suggestion that a recent paper on astro-ph seems to be about objects with very similar properties. The route to such information discovery could either be through intelligent search agents of some kind - a challenging technical problem - or through a global "Virtual Observatory" club - a challenging sociological problem !

(2.5) LARGE DATABASE SCIENCE

At the very least, the kind of advances we have sketched above will greatly improve the general infrastructure of research, making science easier and more efficient (and indeed, possible at all, in the face of the huge expected data volumes). The most striking difference, however, will be that the ambitious data-mining projects become the norm, rather than the province of the occasional stubborn specialists. One can already feel the pressure as astronomers realise the value of this style of working. It is difficult to guess what may be thought of in the future - and, indeed, one of our key arguments is that these advances open up unexpected avenues of research - but some inferences may be made on the basis of exciting examples from current or planned work:

- *Rare object searches* involve trawling through large databases, looking for the (quite literally) one in a million object. Cool white dwarfs have been found in the SuperCOSMOS database, constraining the nature of Galactic dark matter (Fig. 3); z=4 quasars have been found in the APM data, and a z=5.8 quasar by a combined SDSS(optical)/2MASS(IR) search (see Fig 4). Adding UKIRT WFCAM data, or turning to VISTA, will produce quasars at z=6 or 7. Brown dwarfs have been found from joint analysis of SDSS/2MASS data (Fig.5). VISTA will probe further down the luminosity function. Astronomers will also look for the missing dwarf galaxies, Trans-Neptunian objects, and Near Earth asteroids.

- Huge *statistical manipulations* are required to deduce the structure and history of the Galaxy from stellar motions, or to determine cosmological parameters from the anisotropy of the microwave sky. The use of techniques from computational geometry to structure databases in sophisticated ways can make such computations much easier - improvements by orders of magnitude are reported in the time taken to compute the N-point correlation functions of galaxy catalogues when they are indexed using k-d trees, for example.

• *Population modelling* could yield galaxy cluster histories from ugrizJHK diagrams, with the same descriptive power for galaxy evolution as H-R diagrams of stellar clusters had for the evolution of stars, or detect ancient galaxy mergers from groupings in stellar phase space (tidal streamers produced by the Galaxy swallowing the Sagittarius dwarf have been found by Irwin et al using APM).

• *Massive photometry*, monitoring thousands or millions of objects simultaneously can yield exciting advances in: micro-lensing and dark matter; high-redshift supernovae and the cosmological constant; quasar variability and the growth of black holes; parallaxes and the solar neighbourhood problem; planet searches; and stellar seismology and the age problem; the nature of gamma-ray bursts.

• Finally, there is the lure of the *unknown* - what astrophysical phenomena lurk in hitherto unexplored corners of parameter space that can only be reached through the federation of large and disparate datasets?

Finally, we note that part of our Phase A plan is to construct a series of "use-cases", encouraging UK astronomers to develop blow-by-blow examples of imaginary projects that could be carried out with the AstroGrid facility.

(2.6) STORAGE and DATA MANAGEMENT REQUIREMENTS

As described earlier, data volumes are growing alarmingly, especially for the major surveys. Multi-TB databases are becoming normal now, and PB databases (WFCAM and VISTA) will be with us in a few years. Technology is keeping pace, but requires moderately specialised equipment. For example the SuperCOSMOS object catalogue and compressed pixels (2TB) are stored on-line on a highly reliable RAID array costing about £60K. (The raw data (15TB) are accumulating on tape but will probably migrate to DVD jukebox soon). We have made an estimate of the UK science data volumes (from SuperCOSMOS, AAO, UKIRT, ING, XMM, Gemini, SOHO, and Cluster) accumulating in the next few financial years and find +20TB in 01/02, +30TB in 02/03, and +35TB in 03/04. The ramp up comes from UKIRT WFCAM starting in early 2003. So by April 2004 we need to deal with 85TB. VISTA will start in 2005 and by the time it has been running for three years in 2008 it may have accumulated a full database of PB size and a science database of around 100TB. Over the last few years the world storage capacity (and by implication a characteristic system size) has doubled about every 11 months, so that by 2004 we can seriously anticipate on-line storage systems at the 50-100TB level, and hopefully some kind of near-line storage (tape robots, or FMDs) will make the full data available within minutes. (Our future technology forecasts come from a paper given by Jim Gray of Microsoft Research at the recent Caltech NVO conference : see http://research.Microsoft.com/~Gray)

There is then no gross technical problem, but we need to make sure we have the financial provision to keep pace with our data. As today, mass storage equipment will be moderately specialised, and so available only at a few specialist centres. More importantly, although the storage technology is cheap, the associated staff effort is not. The management of such mass storage systems is far from trivial, and maintenance of the databases will require specialised staff, including for example migrating the data to new storage technologies every few years. Because the storage management is intimately linked with the problems of understanding, calibrating and documenting the data, it is however not appropriate to store the data at a national "computer centre" which provides such services in abstract - the data should stay near the experts. This is one of several factors that leads us towards the idea of building a grid around *expert data centres*.

(2.7) DATA INTENSIVE COMPUTING

If one requests an image or catalogue of a random small piece of sky from the SuperCOSMOS database it will be available in seconds. If one searches through all the objects over 5000 square degrees looking for those that pass a particular parameter filter, this will take about 2 hours. Such a search is limited by the I/O bandwidth of a SCSI disk, currently about 10MB/s. In recent history, although CPU power and mass storage have increased at the rate of 100x/decade, device bandwidths have increased at only 10x/decade, and current technology forecasts expect similar trends. Some kinds of problems are limited by the "seek time" from a disk, and this has improved by only a factor of two over the last decade (from about 10msec to 5msec). Extrapolating to a 100TB science database, searching this even at 100MB/s would take 12 days - unacceptable as a general service. Considerable improvement for many tasks can be made by intelligent structuring and indexing of the database, and caching of recent search results. Such techniques are being actively pursued by the SDSS team now. However, there will always be some queries that require streaming all the data. The only way around this is *parallelism* - to partition the data and set many CPUs/discs off searching simultaneously. In other words, we need supercomputer database engines.

Fortunately, many simple queries are quite strictly parallelisable, so we don't need proper Cray-type supercomputers, but rather massive "PC farms" with commodity components and relatively simple message-passing software. Currently the most popular version is the "Beowulf" cluster, costing around £2K/node. To achieve reasonable turn-round on a 100TB database we will need such machines with >100 nodes however. As with mass storage, we expect the technology to be in place to solve our problems, but the equipment will be expensive and specialised. We do not expect that everybody has a Beowulf on their desk, but rather that everybody has access to one provided by a specialist centre. (To achieve the required speeds, the disks and CPUs will need to be co-located, i.e. at the expert data centres). Once again, the real issue is the staff effort involved in developing and maintaining such specialised services. Finally, we should note that although PC farms look very appealing for simple queries, other kinds of analysis (Gaussian mixture models, Fourier analysis, model fitting) require much more message passing and so may run optimally on a different (more expensive !) type of machine, such as a shared memory architecture, or a vector machine. This will probably require collaboration and timesharing with a supercomputer centre. One of the most urgent tasks for AstroGrid will be to predict and define the types of problems users will want to solve in order to design the best architecture for database engines.

(2.8) REMOTE ANALYSIS SERVICES

A key aim of AstroGrid is to design new data analysis and data mining techniques, and to implement them in applications packages that will be developed to the degree of completeness, robustness, ease of use, and documentation that is currently considered the norm in instrumentation, and in data reduction software. This is where we expect the greatest scientific benefit to flow, where a large part of the cost will fall, and where the greatest degree of community involvement is likely to be. However this will not just be a matter of constructing and delivering applications packages which the user installs on their desktop PC. Rather, it will require a commitment to services provided by the expert data centres. This is because of the problem of network bandwidth.

Prediction of future network capacity is considerably harder than for CPU or storage capacity. There is no technical problem. In principle a single fibre can carry up to 25Tb/s. Network speeds in practice are determined by economic factors, and Wide Area Network (WAN) prices have changed little over the last decade. Local Area Networks (LANs) have improved, and most departments have somewhere in the range 10 to 100 Mb/s. However the characteristic real end-to-end speed if one is for example downloading a file from a web page is nothing like this - one will be very lucky to get 1Mb/s. This is just good enough for downloading single observation files but copying large databases is impossible, even if one had the space to store them. The future price/performance ratio for networks is anybody's guess, but an optimistic view is that when the public start to demand downloadable movies on demand then the prices will at last come tumbling down. With luck then in a few years we might hope for end-to-end performance of 1 Gb/s. (This is the number being assumed by the CERN Grid plans). This would still make it impractical to download large databases.

The logic is clear - that large database exploration and data analysis calculations need to be performed on remote computers attached to the databases (in parallel). The motto is *shift the results not the data*. The necessity of offering a remote analysis service forces us into well designed and robust software tools, an aim which is also consistent with our belief that scientific gold lies in systematic large database analysis, and that we can liberate many more scientists to undertake this sort of work. As well as providing click-and-play packages, we will need to provide a toolkit for users to develop their own programs, as well as the facility to upload and run such code, including the monitoring and allocation of CPU time. Along with the constraints of storage, I/O speed, and database management, this is yet another factor leading us to the idea of expert data centres. Such centres will need to ingest, calibrate, and manage the databases, and provide services in data subset access, database queries, on-line analysis tools, and remote visualisation. Users will also want to correlate the databases managed by the various centres, which may require multiple copies of the largest databases at key centres.

(2.9) THE GRID CONCEPT

We have argued that services need to be *remote*, but they also need to be *distributed* because the expertise is distributed around our community. In the last few years the idea of *Grids* has emerged in the computer science world, where many computers collaborate transparently on a problem. The original idea was one of a computational grid; jobs or portions of a job may be allocated to machines over the network, choosing the optimum architecture for particular calculations, filling spare capacity in CPU cycles, or simply aggregating the CPU power of many machines at times of peak demand. The term "grid" is used by analogy with the electrical power grid - a user can simply plug into the grid and obtain computing power, without needing to know where the power station is or how it works. All the user needs is an appliance which uses the power from the grid. Prototype grids are in operation now, but the technique is developing and growing in importance. Networked users will have supercomputer power at their fingertips.

Such computational grids will be relevant to theoretical astronomy, but the data intensive problems we are concerned with here present different problems, being I/O bound rather than CPU bound, and being driven by average load rather than peak load. However the concept leads naturally on to the idea of a *service grid*, and ultimately to a *knowledge grid*. One can imagine an AstroGrid gateway of some kind (not necessarily a web page) where, once a user is logged on, a variety of databases will be browsable, searchable, and manipulable, actually running on one of several database engines and analysis machines, but without the user needing to know where or how. Whether a typical job is simply remote or actually distributed in a computational-grid-like fashion will depend on the kind of job and on future technological and economical developments, but the surface concept remains the same.

(2.10) GRID TECHNOLOGY

Our themes are common ones in modern science (and commerce), requiring mass storage management, remote use of high-throughput database engines, distributed systems, and globalisation of data and meta-data standards. Some of the work we need to undertake is quite specific to astronomy. Some of it is fairly typical of several science areas, but is high level applications work that we would expect to undertake ourselves, with the possibility of our solutions migrating outwards. Much of the work required however is fundamental computing technology. We will not describe this work in detail here, but simply append a list of some of the key issues to give a flavour of some of the problems. Most of these technical issues will not be solved by AstroGrid, but neither will they soon be commercial off-the-shelf solutions. Rather, they will be addressed by computer science research, by the largest scientific-grid projects (especially the CERN/US LHC grid project), and by commercial concerns with similar data-mining problems. (The UPS database is already 17TB). Our task will be to interact with such larger grid projects and adopt appropriate solutions, but also potentially to drive them as an interesting application area.

Some of the key technical issues are as follows :

- data format standards
- metadata and annotation standards
- information exchange protocols
- presentation service standards
- security and access management
- user registration and verification
- request translation middleware
- workload scheduling, resource allocation
- mass storage management
- computing fabric management
- differentiated service network technology
- distributed data management - caching, file replication, file migration
- visualisation technology and algorithms
- data discovery methods
- search agents and AI
- database structure and query methods
- data mining algorithms
- s/w libraries and tools for upload requests
- data quality assurance (levels of club membership ?)

(3) PROJECT PLAN

(3.1) LONG TERM PERSPECTIVE

The issues, desires, and ambitions we are concerned with are common themes in Astronomy worldwide, and across scientific disciplines. The vision of a "Virtual Observatory" only makes eventual sense as a global entity (GVO ?), with the problems being almost as much sociological as technical. Agreeing standards will of course will be a key issue, but so will the attitude to the myriad of private and semi-private databases as opposed to the key large public datasets. It is our impression that such aspirations will only become reality on a timescale of perhaps 6 years or so, but that the components and toolkits will start to develop much sooner. Likewise, the underlying computer science will on the one hand evolve as we try to apply it, and on the other hand will be looking for challenging applications to influence its evolution.

In this perspective we believe that a major but short-term (three year) project is the correct first step, as after this we must take stock and re-group. The UK cannot build the GVO by itself, but we should not wait either. Our overall aims are therefore (i) to provide a working system for UK astronomers to exploit key databases, (ii) to develop tools and methods that the international community will see as key contributions to an eventual GVO, and (iii) to learn technical lessons as fast as possible that will influence the design of an eventual GVO.

(3.2) OVERALL PROJECT PLAN

As argued above, our intention is a focused three year project, with any follow-on being seen as a separate project requiring peer review in the new circumstances. Within the three year project, we divide into two phases. Phase A is essentially an intensive one-year R&D study, which we are developing in as concrete a manner as possible. Phase B delivers the agreed AstroGrid facility, and has preliminary goals, but we believe we should be cautious about the precise scope until completion of the Phase A study. We now take a little time to justify this approach. The October proposal deliberately put forward an ambitious vision. We make no apology for this, as it was necessary to make it clear how important and exciting the subject is, and to get a feeling for what can be achieved. However as the project moves towards reality, we need now to be very careful about the goals, and about the realistic achievables within three years. There are several reasons for this caution :

(a) We need a far more detailed analysis of the scientific requirements before we construct the functionality to meet them. As well as long careful thought, this requires detailed interaction with the community. We have begun this process, but the response in itself makes it clear that extensive consultation is needed. Our intention is to develop a series of blow-by-blow "Use Cases", and to analyse these to produce a Science Requirements Document. We expect this process to take six months.

(b) Pressure from the community is inevitably towards including everything possible in our programme. However it would be a pipe-dream to imagine the UK building an all-encompassing "Virtual Observatory". In practice this concept will emerge into reality on a global scale on a timescale of perhaps 6 years. A better idea is for the UK (i) to deliver some specific facilities on a short timescale that match its prime opportunities and needs, and (ii) to establish itself as a key player within the Virtual Observatory game, with real product to offer. This strongly argues for selecting key datasets and limited goals.

(c) This programme will be complex to design and manage. Unless we keep it reasonably focused, it will fail.

(d) Politically, PPARC having been given new money for this work, it is very important to visibly succeed.

We have therefore structured our programme into two distinct Phases. Phase A is a one-year intensive R&D study, with two main components - requirements analysis, and early experimentation with hardware and software, to bring us rapidly up the learning curve. There are two main outputs - the Science Requirements Document, and a Phase B plan. In Section (5) we describe our current understanding of the scope of our whole programme, divided into "Activity Areas". We have deliberately avoided casting these as "Work Packages" in order to indicate that the precise work and deliverables remain undecided. For Phase A however, we have developed outlines of concrete Work Packages.

We note also that a two-phase plan sits well with a cautious approach to committing the money, and make a specific proposal along these lines in Section (8)

(3.3) THREE YEAR GOALS

The final deliverables of the project will not be agreed until completion of the Phase B plan. However, our goals are as follows :

►A working datagrid for key databases, requiring :

- associated distributed storage systems and high-throughput datamining machines
- implementation of middleware to enable distributed processing and data management
- ingestion of key target databases : WFCAM, SOHO, Yohkoh, Cluster, EISCAT, SuperCOSMOS, INT-WFC, Sloan (by permission), XMM, Chandra (by permission), Merlin, FIRST, 2dF, 6dF, plus more by agreement.

►A uniform archive query and data-mining interface for the above

- simple public query interface by standard web page
- advanced query and datamining interface using supplied software environment
- intelligent on-line help system (AstroGrid Assistant).

► Ability to browse simultaneously multiple datasets
  - agreed data, metadata, and annotation standards for subsets of above databases
  - agreed world co-ordinate system
  - visualisation tool for examining :
    observation catalogues, images, source catalogues, spectra, and time series

► Tools for integrated on-line analysis of data (images and spectra)
  - on-the-fly imaging from UV data
  - measuring fluxes, astrometry, fitting curves.

► Tools for on-line database analysis
  - high speed whole database queries via indexing, caching, and parallelism
  - database subset exploration tools - parameter plotting, rotating, statistics etc
  - whole database manipulation tools - eg FFTs, gaussian mixtures, etc

► Ability for user to upload code to run own algorithms.
  - new algorithms
  - user-supplied modelling techniques

► Tools for database ingestion
  - linking private datasets
  - procedures for future public database ingestion

► Tool for open-ended resource discovery
  - proposal for protocols/interfaces

(3.4) ONE YEAR GOALS

The Phase A plan is decsribed in the next Section. The key deliverables are as follows :

► Phase B plan
► Science Requirements Document
  - including set of example Use Cases)
► Functional Requirements Document
► Functionality Market Survey Report
► Agreed division of labour with international partners
► Working data grid
  - including demonstration of multi-site browsing and database searching
► Federation of SOHO and Yohkoh databases
► Federation of SuperCOSMOS, SDSS, and INT-WFC databases
► Federation of XMM and Chandra data products
► Federation of CLUSTER and EISCAT databases

► Federation of selected MERLIN and VLA database subsets
► Preliminary visualisation tool
► Preliminary database management system

(4) PHASE A PLAN

(4.1) GENERAL APPROACH

We all realise that we are working in an area that is extremely important and exciting, but that is in danger of being driven by fashion and half-thought-through ideas. This drives us towards speed and caution simultaneously. On the one hand, we should start concrete work as soon as possible, so that we know what we are talking about, learn lessons, and get into a position of strength with respect to international partners. A top-down debate on standards and so on will be going on internationally, to which we should contribute, but for the completion of which we cannot wait. On the other hand, we need to step back and carefully analyse both the scientific requirements, and current national, international, and commercial capabilities, before freezing the goals, techniques, and project structure. Our Phase A plan therefore contains an intensive requirements analysis, but in parallel, an experimental programme of benchmark tests, and several pilot database federations. These are intended primarily to learn lessons, but will also be real working scientific facilities.

(4.2) WORK PACKAGE DESCRIPTIONS

WP-A0 ESTABLISHMENT OF PROJECT INFRASTRUCTURE

TASKS Set up project web pages, FTP areas, and software repositories, and automate information sharing. Decide document classes and formats, set up templates. Decide software standards and set up libraries and toolkits. Set up document and email repositories, e-mail exploders, diaries, and schedules.

DELIVERABLES Handbook of project procedures. Public and consortium web pages. Working project information system

RESOURCE 3 staff months

WP-A1 REQUIREMENTS ANALYSIS

TASKS Consult community; circulars, public meetings, key committees, and private discussion visits. Commission and develop Use Cases and deduce requirements. Organise public meetings and invited brainstorming sessions.

DELIVERABLES Science Requirements Document, including set of Use Cases.
Functional Requirements Document. Public meetings. Reports on meetings.

RESOURCE 12 staff months

WP-A2 FUNCTIONALITY MARKET SURVEY

TASKS Investigate options for commercial software and assess capabilities (applications, DBMS, etc). Investigate options for academic software and assess capabilities (IRAF, Starlink, IDL; rival middleware toolkits). Likewise for options in data/metadata standard, interfaces etc - FITS, XML, etc. Investigate capability for new s/w construction within UK astronomy community. Investigate progress and working assumptions being made in other disciplines and other countries

DELIVERABLES Technical Reports on applications/products etc.

Functionality Market Survey Report. Decisions on route forward

RESOURCE 9 staff months

WP-A3 EXPERIMENTAL PROGRAMME.

TASKS Procure and deploy 16-node Beowulf clusters at selected sites. Make arrangements for borrowed use of other machines, eg. EPCC SMP machine, Jodrell 128-node pulsar search machine. Install and test middleware packages (Globus, Legion, CORBA). Deploy two or more clusters as test data-grid. Design benchmark problems for data access, database searching, and data analysis problems. Quantify performance of various machines, configurations, scale-sizes, and packages on the benchmark problems, using version-0 tools.

DELIVERABLES Working data grid (for experimental purpose only). Grid-enabled versions of selected applications packages. Demonstration of multi-site browsing and database searching. Documented performance tests of various machines/packages. Technical reports on lessons learned from experiments. Decisions on route forward.

RESOURCE 24 staff months

WP-A4 DEMONSTRATION FEDERATION-1 : SOLAR

TASKS Federate SOHO database (RAL) with Yohkoh database (MSSL). Agree criteria for federation success. Define agreed data, metadata and database standards as necessary for the pilot federation (i.e. not necessary to use "final" standards. Construct simple user interface for interrogation of databases simultaneously, using version-0 tools.

DELIVERABLES Successful pilot federation. Technical report on lessons learned. Documentation of sufficient standard to allow user-testing.

RESOURCE 7 staff months

WP-A5 DEMONSTRATION FEDERATION-2 : OPTICAL-IR

TASKS Federate SuperCOSMOS and early-release SDSS databases (US via Edinburgh) with INT-WFC and INT-CIRSI databases (Cambridge). Agree criteria for federation success. Define agreed data, metadata and database standards as necessary for the pilot federation (i.e. not necessary to use "final" standards. Construct simple user interface for interrogation of databases simultaneously, using version-0 tools.

DELIVERABLES Successful pilot federation. Technical report on lessons learned. Documentation of sufficient standard to allow user-testing.

RESOURCE 7 staff months

WP-A6 DEMONSTRATION FEDERATION-3 : X-RAY

TASKS Federate XMM database (Leicester) with Chandra database (US via Leicester). Agree criteria for federation success. Define agreed data, metadata and database standards as necessary for the pilot federation (i.e. not necessary to use "final" standards. Construct simple user interface for interrogation of databases simultaneously, using version-0 tools.

DELIVERABLES Successful pilot federation. Technical report on lessons learned. Documentation of sufficient standard to allow user-testing.

RESOURCE 7 staff months

WP-A7 DEMONSTRATION FEDERATION-4 : STP

TASKS Federate CLUSTER and EISCAT databases. Agree criteria for federation success. Define agreed data, metadata and database standards as necessary for the pilot federation (i.e. not necessary to use "final" standards. Construct simple user interface for interrogation of databases simultaneously, using version-0 tools.

DELIVERABLES Successful pilot federation. Technical report on lessons learned. Documentation of sufficient standard to allow user-testing.

RESOURCE 7 staff months

WP-A8 DEMONSTRATION FEDERATION-5 : RADIO

TASKS Federate VLA and MERLIN visibility data and image archives. Consultation with community. Define agreed data. Define/Develop access methods and metadata for continuum visibility data. Prototype techniques for on-the-fly image formation and deconvolution. Agree criteria for federation success. Define agreed data, metadata and database standards as necessary for the pilot federation (i.e. not necessary to use "final" standards. Construct simple user interface for interrogation of databases simultaneously, using version-0 tools.

DELIVERABLES Successful pilot federation. Demonstration of on-the-fly imaging. Technical report on lessons learned. Documentation of sufficient standard to allow user-testing.

RESOURCE 7 staff months

WP-A9 VERSION ZERO VISUALISATION TOOL.

TASKS Construct simple image viewer and/or source catalogue viewer for use in the demonstration federations. Not necessarily intended to be final visualisation tool - just sufficient functionality to make the federation test possible. Probably choose, upgrade and deploy existing package (e.g. GAIA) .

DELIVERABLES Working visualisation package. Report on lessons learned. User documentation.

RESOURCE 7 staff months

WP-A10 VERSION ZERO DATABASE SYSTEM.

TASKS Construct simple database management system, and simple data access tools, observation catalogue browsing tools, and exploration tools for use in the demonstration federation. Not necessarily intended to be development towards final DBMS, or data-mining tools, just working code with sufficient functionality to make the federation test possible. Probably choose, upgrade and deploy existing packages (e.g. SYBASE, SolarSurf, Objectivity, CURSA) .

DELIVERABLES Working database management system capable of addressing two physically separate databases. Simple data exploration tools. Report on lessons learned. User documentation.

RESOURCE 7 staff months

WP-A11 CONSTRUCT PHASE B PLAN.

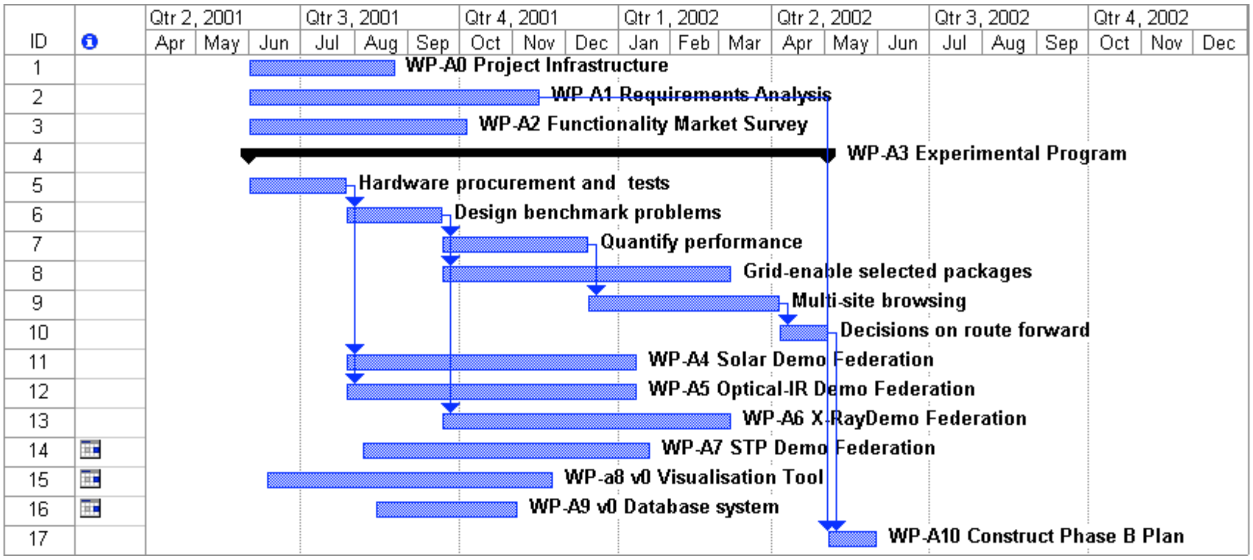
TASKS Monitor progress of Grid work across all disciplines. Monitor progress of international projects, and begin contributions to AVO project. Digest results from all other work-packages. Assess political scene and financial prospects and make technology forecast. In

	the light of all the above, set realistic goals for two year programme and design workpackages to achieve them.
DELIVERABLES	Phase A completion report. Phase B Plan (project goals; whole life cost estimates; detailed workpackage definitions; management plan; operational plan.) Agreed division of activities with international partners. Agreed stance with cross-disciplinary partners.
RESOURCE	6 staff months

(4.3) PHASE A SCHEDULE

This is still under development and is expected to change rapidly over the next few weeks. The attached Gantt chart expresses our plan to date. The science requirements analysis and functionality survey need to start straight away and complete in half a year in order to influence the pilot federations. The experimental programme spreads across the whole year, and the pilot federations are designed to deliver towards the end of the year. The year finishes with almost all staff working together in an intensive debate aimed at delivering the Phase B plan, digesting the lessons learned in the Phase A Work Packages.

Crucial to this plan is of course putting in place a Project Manager and Project Scientist as soon as possible, and also making some other early recruitments in addition to our re-deployed staff. We are currently working on modelling the effort profile over the year, and the matrix of effort across individuals and institutions.



(5) PHASE A/B ACTIVITY AREAS

The twin goals of AstroGrid are the provision of the infrastructure and tools for the federation and exploitation of large astronomical, solar and space plasma datasets, and the delivery of federations of current datasets for its user communities to exploit using those tools.

The work required varies between these communities. For example, the solar community has a well established software system (SOLARSOFTWARE, based in IDL) which is used world--wide, and to which any user can submit a software module for general use. What it lacks is the transparent access to multiple, remotely--held datasets from different instruments, as required to make full use of these common analysis tools in understanding solar phenomena seen at different wavelengths, and over different scales in time and space. By studying the federation of different types of data, from different communities, with different requirements and current expertise, AstroGrid will shed light on generic problems from a number of directions, aiding the development of solutions with applicability well beyond the realms of astronomy and solar--terrestrial physics.

The AstroGrid programme is divided into a set of eleven Activity Areas, whose union delineates the scope of the project. The list below is not ordered by importance, nor is it sequential: while some of these Activity Areas are restricted to Phase A or primarily to Phase B, many will continue for the life of the project. It is only upon the completion of Phase A that all these broad and interconnected Activity Areas can be converted into well--defined Work Packages with definite resource allocations, deliverables and schedules. In particular, for those Activity Areas operating in Phase B, the following descriptions can only indicate what we currently envisage they will comprise.

AA1 Project Management

The coordination of a project of this size and complexity will require an experienced Project Manager, supported by a Project Assistant, and the importance of this position argues strongly for flexibility in the location of this Project Office, to help secure the best person for the job. The Project Manager will liaise closely with the AstroGrid Lead Investigators and the Project Scientist who will provide the scientific direction for the project on the strategic and day--to--day level, respectively, and with local coordinators who will be responsible for the deliverables from particular Work Packages and the work of particular teams. The Project Manager will be responsible for the progress of the AstroGrid programme as a whole, from the preparation of detailed descriptions of Work Packages to the allocation of the resources to undertake them. This will include the decision as to whether a particular activity should be performed within a Consortium institution or be outsourced, to an external academic or commercial body; this decision being based on a balance between concerns of cost and schedule, and the desirability of developing and maintaining relevent and re--usable expertise within UK universities and research establishments.

AA2 Project Infrastructure

This Activity Area covers the implementation, in software and computer--based procedures, of the policy decisions of the project management. This involves the setting up and maintenance of:

● WWW pages, both public and internal to the project, static and dynamic (including perhaps CGI scripts to reflect actual project progress in near real-time)

● FTP areas and software repositories with mirror scripts to ensure full and regular information flow around all project sites, perhaps with automatic and regular software builds of key components.

● Standards for documentation (what document classes are to be used, which formats, standard templates for some document classes, etc.)

● Standards for software (which languages/compilers, which libraries can be assumed to be present, which platforms to support, etc.)

● Configuration and version control of software packages.

● Document and e-mail repositories, e-mail exploders, diaries and schedules.

AA3 Requirements Analysis

The core of AstroGrid's Phase A will be its requirements analysis process. This will survey what the user communities can currently do with existing datasets, what they will require to do the same science on the next generation of much larger datasets, and what additional functionality they will need to do the additional science that these new datasets will make possible, both individually and, in particular, when federated. We intend to make this consultation exercise as wide--ranging as possible within our user communities, and propose that its results be distilled into a set of Use Cases that together comprise the communities' requirements. A subset of these will then be selected as defining the requirements of the programme that AstroGrid can deliver (given its time and resource limitations, plus the complementary activities being undertaken by analogous projects internationally) and this subset will provide the evaluation criteria to be used in AA5.

AA4 Dataset Federation

This Activity Area divides into three components, covering the development of expertise in federating datasets and the delivery of dataset federations to AstroGrid's user communities.

Pilot Federations

The federation of datasets in several pilot projects within Phase A is desired, for several reasons:

(i) the experience gained will help decide which hardware and software solutions should be developed within Phase B; (ii) the early availability of pilot federations (together with beta versions of some restricted set of tools) will enable AstroGrid's various user communities to help refine the detailed science requirements of Phase B tasks; (iii) UK astronomers can start doing the new science made possible by these pilot federations as soon as possible.

The proposed sets of pilot federations are :

(a) Imaging data from SOHO and Yohkoh, to be followed (in Phase A if time permits, or else in Phase B), by the further inclusion of spectral data;

(b) SuperCOSMOS + INT--WFS + INT--CIRSI + Sloan initial data release;

(c) XMM--Newton + Chandra data products, to be followed (in Phase A, if possible, but otherwise extending into Phase B) by combination with the optical/near--IR federation of (b).

(d) CLUSTER and EISCAT databases

(e) Selected subsets of MERLIN and VLA imaging and visibility datasets, plus the ability to create images from fringes on-the-fly, followed by combination with (b) and (c) to make a complete multi-wavelength database.

These five pilots have been chosen because they focus on different aspects of the general database federation problem. In the first, the data products are similar, but they are located in different places -- SOHO data at RAL and Yohkoh data at MSSL -- so this pilot project studies the federation of geographically--distributed datasets. The optical and near--IR survey catalogues federated in the second pilot are very similar, and relatively simple, but they are very large, so the focus of this pilot project is learning how to federate very large data volumes. The salient feature of the third pilot project is the multi--faceted nature of X--ray data: from a photon event list one can derive image, spectroscopic and timing information, so, by federating data products from XMM--Newton and Chandra, we can learn about how to combine consistent data of different types so that the resultant datasets are both rich and meaningful. The fourth federation involves databases that are modest in size, but extremely complex and heterogeneous, presenting the greatest challenge in visualisation and in keeping requirements focused. The final federation involves datasets that are fundamentally similar, but greatly different in spatial resolution, and with the greatest benefits to come from real-time re-processing of raw data, which has not been previously attempted.

Generic Federation Techniques

This component consolidates the experiences from the pilot federations and elsewhere, to produce a generic system for the inclusion of datasets within the AstroGrid system. This must include facilities for restricting access to portions of the federated archive, both so that users can upload private datasets to share only amongst their collaborators, and so that telescope and survey archives can be included in the federation in such a way that the proprietary periods on particular datasets are handled automatically. The deliverables from this component will include a minimal set of conditions that must be met for a dataset to be ingested into the AstroGrid federation: it must obey certain data format and metadata conventions, and may also have to satisfy some basic data quality criteria. These conditions will have to be developed in coordination with other, international bodies, to ensure consistency within the developing global Virtual Observatory and its analogues in solar and space plasma physics.

Federation of Existing and Near--Future Datasets

Once the generic federation techniques have been developed, they will be deployed to provide the federations of a wide range of current and near--future datasets for the three user communities of AstroGrid, as follows:

- Astronomy: SuperCOSMOS, INT--WFS, INT--CIRSI, 2MASS, FIRST, NVSS, Merlin, 2dF, ISO, Chandra, XMM, SDSS, UKIDSS, etc;

- Solar Physics: SOHO, Yohkoh, TRACE, HESSI, SMM, plus ground--based datasets;

- Space Plasma Physics: Cluster, Ulysses, Wind, Polar, ACE, IMAGE, plus ground-based data (in particular EISCAT/ESR, SuperDARN/CUTLASS, SAMNET, IRIS and SPEAR) and global activity indices held in the World Data Centres;

as well as for the federation of cross-disciplinary datasets. For example, astronomical and solar physics datasets can be used in conjunction, to compare and contrast solar and stellar flare events, while the combination of solar and space plasma data can be used to investigate the impact of solar phenomena on the terrestrial environment.

The datasets listed above include a number that will be released incrementally by their respective consortia (with separate proprietary periods for successive data releases from each survey). This component must, therefore, cover the preparation and implementation of the software required to ingest data from these surveys as they appear and become public, in such a manner that the federated dataset is kept as up to date and well integrated into the rest of AstroGrid (and wider virtual observatories) as possible. This will require the periodic assignment of new associations between objects in these and already--federated datasets, as well as, possibly, the dynamic re--indexing/re--structuring of the federated database in the light of those new associations. Federating these datasets will yield additional experience, so it is likely that the detailed specification of AstroGrid's generic federation procedures may evolve during Phase B.

AA5 Design and Evaluation of AstroGrid Architecture

This Activity Area will define our basic architecture and develop sets of fundamental tools. Some of these we can borrow from other disciplines, while for others we expect to take the lead and make software generally available. The interoperability of AstroGrid with other e-science developments will be an important architectural consideration. Work will include the development of procedures for:

- the easy location of, and access to, remote data (and its associated metadata) in a secure manner;

- the consolidation of prototype software into reliable, reusable products, by means of proper testing, generalisation and documentation;

- the maintenance of delivered software, through formal releases upgrading the code, together with the issuing of temporary patches as and when required between releases;

- the registration and authentication of users;

- the monitoring and control of the usage of AstroGrid, possibly including resource allocation for major archival projects, via a PATT--like system.

It is important that the performance of the AstroGrid system defined by this basic architecture be evaluated against objective criteria, defined by a set of Use Cases describing realistic research programmes undertaken on it. This evaluation will take place at several levels, from the assessment of individual components (and, in particular, the decision as to whether an existing software package can be used for a given task, or whether bespoke software must be produced) to the operation of the system as a whole, as implemented on a range of different hardware/ software configurations.

AA6 Database Systems

AstroGrid's database federation software is likely to sit upon one or more commercial database management systems (DBMSs), and an important task will be to assess the range of such products to determine which are best suited to AstroGrid's needs, nature and budget. This will require study at several levels, from the general (comparing AstroGrid requirements with the generic advantages and disadvantages of relational, object-relational and object-oriented DBMSs, and assessing the relevance of new XML--based systems to AstroGrid) through to the specific (the evaluation of particular products, implemented on particular hardware in particular configurations) and will include the gathering of relevant experience from similar users of DBMSs, as well as the conducting of realistic benchmarking tests within the project. This is, of course, part of the evaluation procedure outlined in AA5 but it is singled out as a separate Activity Area, both because of its importance, and because it constitutes a large body of work that can be done in isolation from other tasks. The result of this process will be the purchase and installation of the chosen DBMS(s) and, once again, this choice must not be taken in isolation, but must be informed by parallel international developments within AstroGrid's user communities.

The second component of this Activity Area is the development of specialised software to sit upon the chosen commercial DBMS(s), such as indexing schemes and query systems, including procedures for performing the basic operations (mathematical, statistical and astronomical) required to deliver a sophisticated query system. Once again, this will require original research by AstroGrid, plus the evaluation and assimilation of existing work from other disciplines -- for example, there is a well--developed literature on database indexing schemes and tools for the statistical description of multi--dimensional datasets in the field of computational geometry.

AA7 GRID-enabling existing pacakages

A key principle for AstroGrid will be that existing packages will be used wherever possible, in preference to the development of entirely new ones. This will be particularly important for the community assessment and exploitation of the pilot federations produced in Phase A, since these will be produced on such a short timescale that the only option is for the user interaction with them to be via existing tools, extended as required for distributed operations.

So, an initial task within this Activity Area will be to decide what minimal set of tools are necessary for users to work with the pilot federations, and to amend them as required for that task. More general work in this Area will cover the assessment of which existing packages (in Starlink, IRAF, IDL, etc) can be amended to meet the requirements set for AstroGrid within its schedule and resource allocations, to be followed by the implementation of this GRID--enabling work and the testing of its performance, as part of AstroGrid's system evaluation procedure. A natural part of GRID-enabling the applications will be the enhancement of the underlying data model, which will be promoted to other areas of science.

AA8 Visualisation

Visualisation tools are an essential component of any database system, both for directing the course of queries and for displaying their results. Perhaps the most important visualisation tool needed by AstroGrid is a good browser, capable of combining image, spectral and catalogue data in a useful way. Current browsers can do this when the data are stored locally, but it will be necessary to expand this functionality to include the seeking out of relevant data over the network (including some level of deciding what is ``relevant'' through metadata manipulation), to experiment with ways of presenting in a meaningful graphical way what may be a large quantity of data, and of linking other data search and manipulation modules to the browser, so that the browser becomes the basis for the user's directed querying of the archive.

Equally essential will be the provision of software which enables the user to plot different combinations of object parameters, as well as overlaying theoretical model predictions and empirically--derived statistical correlations: visualisation therefore represents an important interface between the theoretical modeller and the observer. Several commercial and public domain products exist that cover some of this functionality, so the work in this Activity Area will include their evaluation against the requirements set by AA5, as well as, possibly, the development of new software, where required.

It is envisaged that the tasks within this Activity Area will be divided between Phases A and B: a relatively simple browser and a few other basic visualisation tools will need to be provided for the study of the pilot federations made in Phase A, and then more sophisticated tools will be produced in Phase B, with a functionality informed by the experience of users working with the pilot federations. There is an increasing diversity of image and information display devices to be considered for possible Phase B development, and user profiling will be used to validate new applications of visualisation for AstroGrid's purposes: for example, methods for the navigation of virtual or augmented reality information spaces are available now at low cost, but the important issue for AstroGrid is when and where they are of relevance to its user communities.

AA9 Data Mining and Information Discovery.

The major science driver for AstroGrid is the realisation that the federation of new and existing datasets will make possible new science, and the task of AA9 is to provide the tools to facilitate that: we distinguish here between Data Mining, which is taken to denote unsupervised analysis, while Information Discovery is understood to mean supervised analysis.

Data Mining

Data Mining includes looking for patterns within datasets, whether images, time series, spectra, catalogues, or bibliographic resources, and its corollary, which is the identification of unusual events, which might either be rare phenomena or flaws in the data. AstroGrid must, therefore, provide algorithms for running on its database federations regression analysis, deviation detection, sequence analysis, classification through clustering and similarity searches, amongst other things.

The output from the Data Mining component will be an assessment of the requirements and possibilities for unsupervised analysis of the datasets to be federated by AstroGrid, and the adaptation or development of the software to do it. Since success in unsupervised analysis relies on the structure of the database, this component will be coupled strongly to AA6, to ensure that the database architecture adopted facilitates data mining activities: this coupling may include the imposition of some semantic structure (XML, leading to close linkage with Information Discovery), incorporation of aspects of resolution scale to aid in data interpretation, and ``smart" data processing such as signal and noise modelling. It will also be important for the data mining component to interact with the definition of generic federation techniques, so that the criteria for allowing data ingest into the AstroGrid system include the specification of sufficiently rich metadata that data mining algorithms can work effectively on them.

Information Discovery

To complement the unsupervised analysis of the Data Mining theme, this component covers the provision of tools to enable the user to derive new knowledge from directed searching of the federated datasets (e.g. using a browser as the basis for a search, as outlined in AA8). An important concept here is the interaction of the user with summary data (possibly produced automatically, through information characterisation methods developed in the Data Mining component) so that s/he can *see the wood for the trees*, facilitating directed searching of the data. Producing such summations -- of text, of images in conjunction with catalogues and the published literature, etc. -- is a difficult task from the point of view of cognitive science. The wealth of increasingly federated online data and information, including image archives, ADS and the literature, the contents of data centres, and educational and other presentation materials, points to the central task of information fusion.

Tools to be produced in this component will reinforce links between the varying types of information and data referred to in the previous paragraph, and will be of help in the data ingest and quality control work of data centres and data archive sites. It is expected that many of the tools such as software agents and ontology capture will be jointly developed with GRID activities in other disciplines.

AA10 Data Processing, Curation, and Quality Assurance Paradigms

A prerequisite for extracting good science from a database federation is having data worth combining. So, while AstroGrid will not provide the core funding for setting up or maintaining data centres, it must cooperate with them in the definition of data products and quality assurance and curation procedures that ensure that the data centres feed into AstroGrid's federation infrastructure well--documented data of appropriate quality, and continue to do so despite the inevitable migration of data over storage media as the years pass.

It is highly preferable that the pipeline processing of a particular dataset not be considered in isolation from the creation and curation of its archive. It makes sense for the data models of the pipeline and archive to match as closely as possible, for ease of ingestion of data (including metadata describing, for example, the provenance of the data -- including the particular version of the pipeline that produced it) into the archive. Also, few archives are completely static, and the interface between pipeline and archive should ease the iteration of reprocessing and recalibration of the data, as the understanding of the instrument that created it improves. Finally, the archive, and tools written to work with it, can help the quality assurance process, both through the use of some of the visualisation tools of AA8 in undertaking testing of the data quality, and through the flagging of "oddities'" in the data by autonomous data mining processes, as mentioned in AA9. There are discussions underway in the optical/near--IR sky survey community to develop common standards for data products and, perhaps, data processing systems, and it is the goal of this Activity Area to interact with these (and similar developments in AstroGrid's other user communities) to develop integrated quality assurance programmes for future projects, like VISTA, to encompass both pipeline processing and database curation activities.

AA11 External Relations

This broad Activity Area covers all relationships with those outwith the AstroGrid project. This includes analogous projects (both internationally and in other subject areas) addressing similar problems, and those who will use AstroGrid's ouput, both the scientific user communities exploiting those federations in their research, and the educators and general public, wishing to use AstroGrid as a way of learning about astronomy and the solar--terrestrial environment.

Collaborative Programme with International Partners

It is vital that AstroGrid's data and software products are consistent with those produced by similar initiatives being developed internationally, and by starting in earnest earlier than most of them, AstroGrid has a great opportunity to influence the standards that are adopted by these projects, which are discussed in Section (7).

Collaborative Programme with Other Disciplines.

As discussed in several of the Activity Area descriptions above, the expertise required for a number of the developments that AstroGrid must make already exists in other disciplines. We must make use of this expertise wherever possible, developing a collaborative programme, which will cover:

(i) generic Grid issues (e.g. extension of the Globus toolkit) which will be common to all e-science initiatives;

(ii) bilateral interactions between AstroGrid and similar groups in other application areas (e.g. Earth observation) which address similar problems, and may already have found adequate solutions to them;

(iii) learning from experts in informatics and computer science researching relevant subjects, such as database indexing schemes and statistical descriptions of multi--dimensional data spaces.

At the core of this activity will be regular consultation with both PPARC's E--science Director and the OST's "Grid Czar", to facilitate the two--way transfer of ideas between AstroGrid and other initiatives being funded by the Government's e--science initiative. These issues are discussed in more detail in Section (7).

Community Training.

AstroGrid must also develop a community training programme, to provide its user communities with the information and skills necessary to exploit AstroGrid's database federations to the full. This is likely to combine public meetings, the production of introductory documentation (e.g. online tutorials) and more hands--on training, which might either entail AstroGrid staff giving demonstrations to research groups and/or having researchers attend more formal training courses at an AstroGrid site.

Public education

AstroGrid will also present an excellent opportunity for advancing the Public Understanding of Science, and an outreach programe should be established, with a two--pronged approach:

(i) to help educators make the most of the vast array of information collated and linked by AstroGrid as a teaching resource;

(ii) to provide tailored access to AstroGrid for the general public, so that they can learn something of the combined power of the astronomical facilities they fund, as evinced by the datasets brought together by AstroGrid itself and as illustrated using the visualisation tools produced under AA8.

(6) MANAGEMENT PLAN

(6.1) GENERAL APPROACH

The management philosophy of the AstroGrid project is that the overall aspirations should emerge from the desires of the community, but that the project in practice should have finite specific goals, and should be run as a tightly defined project by a relatively small consortium, equivalent to building an instrument. There are two reasons for this approach. First, we believe that if the project is allowed to become too ambitious, or has too diffuse a structure, it will fail. Second, the grandest aspirations of a "Virtual Observatory" like environment will certainly not be achieved by the UK in isolation, but may be achieved on a global scale over five years or more. Seen in this context, the UK AstroGrid project needs to make clear concrete achievements seen as contributions to the world-wide "VO" agenda, which will place us centrally on that stage.

Although a small number of institutions retain management responsibility for AstroGrid, a much larger number of institutions are likely to be involved, as (a) a large fraction of the work will be outsourced, either as grants or commercial contracts, and (b) many individuals will participate in the advisory structure, and (c) the community as a whole will be extensively consulted.

(6.2) MANAGEMENT and ADVISORY STRUCTURE

PPARC e-Science Steering Committee (PESC)

AstroGrid will be almost entirely a PPARC funded project. (Some HEFCE/SHEFC/DENI funded effort is of course included, and EU funds are hoped for). AstroGrid activities will therefore be subject to the usual grant conditions and management constraints imposed by PPARC. In particular, PPARC has announced the creation of an e-science director and e-science steering committee, covering the whole e-science programme and not just AstroGrid. The steering committee will monitor the progress of AstroGrid, control its budget allocation, and oversee its goals and match to PPARC requirements. For convenience we refer below to the combination of e-science Director and Steering Committee as PESC.

AstroGrid Lead Investigators.

Responsibility for the goals, design, and implementation of the project rests with the AstroGrid Lead Investigators (AGLI). The individuals concerned are A.Lawrence, R.McMahon, M.Watson, F.Murtagh, L.Harra, P.Allan, M.Lockwood, and S.Garrington, although this list may change by agreement of the members. The AGLI also direct the project and set policy, subject to the constraints and oversight of the PESC. There is no formal Project Director - policy and direction are achieved by mutual agreement. However at any one time there is an agreed Project Spokesperson and figurehead. Currently this is A.Lawrence. Normally the AGLI meet within the context of AGWC meetings (see below) but may arrange extra meetings and teleconferences as necessary.

AstroGrid Working Consortium (AGWC)

A considerable number of people will be employed towards the ends of the AstroGrid project, either within the consortium organisations, or in other organisations, or through commercial contracts and secondments. However a number of core staff have already been extremely active in both technical work and in the design and goals of AstroGrid, and it will remain useful to identify these "key staff" as having a special role. The current list is M.Irwin, J.Sherman, R.Mann, D.Pike, C.Page, C.Davenhall, G.Rixon, D.Giaretta, R.Bentley, R.Stamper, C.Perry, and A.Richards. The list may change by agreement. On the other hand it would not be appropriate to place top-level responsibility on such a large group of people, especially as many of them will have their own salary dependent on AstroGrid. This is why we have separated the smaller AGLI group as taking responsibility for direction. The concept is that the AGWC is the body through which debate concerning AstroGrid policy and implementation and technicalities takes place, but that formal responsibility rests with the AGLI. The AGWC maintains a continous email discussion including complete email archive. It should meet approximately quarterly, and has indeed done more than this so far. It will be expected that further staff will normally be welcome at AGWC meetings as requirements suggest, but that the formal membership of the AGWC (in effect the default circulation list) will only evolve by agreement.

Project Manager and Project Scientist.

The project will appoint a full time Project Scientist (PS) and a full time Project Manager (PM). The PM will be an open recruitment. The Project Scientist may also be an open recruitment, but may be selected from within consortium institutions. Draft job descriptions for these two posts were included in the March Astronomy Committee paper and are available on request. Both postholders report to the AGLI, normally through documents and reports presented to AGWC meetings. The PM and the current AGLI spokesperson have the additional reponisbility of being the principal liaison points with the PESC. The PM will expect to make regular reports to the PESC. The PS has prime responsibility for seeing that AstroGrid meets its science goals. The PM has prime responsibility for the budget and schedule of the project. It is intended that the AstroGrid Work Packages will be fairly clearly devolved to particular individuals and organisations. The PM will allocate the work, but in close interaction with the AGLI and the PESC. It is not yet clear whether there will be a distinct "Project Office" (as with e.g. Gemini or VISTA) or simply a distributed programme (as e.g. with XMM) . We wish to leave PM candidates the freedom to indicate their preferred structure. Our intention is to recruit the PM as soon as possible. In the meanwhile, we will not appoint an interim PM, but will appoint one or more of the re-deployed staff to undertake the administrative side of the PM responsibilities.

AstroGrid Advisory Board (AGAB)

We are undertaking a programme of community consultation in a variety of ways, for example by meetings and by an open "call for comment". This is crucial in developing the Science

Requirements Document. However we do not see this as simply a once-for-all process. Furthermore, there are considerable skills and experience on both scientific and technical matters in the wider community that we wish to benefit from. Finally, we wish to strike a balance between creating a wide sense of ownership on the one hand, and keeping a manageable project structure on the other hand. Our intention to strike this balance is to create an Advisory Board with invited members. Around ten members may be large enough to be representative but still small enough to be useful. We intend that the AGAB will include a mixture of ordinary astronomers, those with special skills and interest in astronomical software or data mining research, those with key future project interests (e.g. VISTA, GAIA, ALMA, WASP, the Liverpool Telescope), and finally computer scientists and particle physicists. The Advisory Board will normally meet immediately before AGWC meetings, and formally provides advice to the AGLI.

(6.3) RESOURCE ALLOCATION - the BUSINESS PLAN.

Our suggestion is that the PESC agrees an overall financial envelope for AstroGrid, but does not announce a single large grant. Rather, the AstroGrid PM develops a Business Plan, iterating this between the AGLI on the one hand, and the PESC on the other hand. At intervals as appropriate the PM asks PPARC to announce a grant to a particular institution to carry out work to agreed workscopes. This system has the advantage of administrative simplicity for the participating Universities, whilst still giving the PM considerable control and flexibility. It also means that grants to Universities outside the consortium work just the same way. These grants could be to allow recruitment of PDRAs, or to purchase particular pieces of equipment, and could also carry fractional quantities of administrative assistance etc as recommended by the PM. In addition however, the Institution at which the PM resides could hold a grant acting as a central fund, for example for travel, or for rapid response on procurement.

(7) EXTERNAL RELATIONS

(7.1) UK user community.

We see the project as rather like building an instrument. The general scientific opportunity is clear, and expert groups have proposed a concept for the instrument. Before proceeding to actual design and construction however, there must be a careful science requirements analysis, which must then become the construction bible. This is even more crucial than usual in this case, and needs to involve the user community as widely as possible, for several reasons. (i) AstroGrid is not concerned with a specific scientific problem, but with a whole range of possible problems. (ii) Much of the user community has key skills as well as interest, and so has much to contribute. (iii) We all know there is fog as well as gold in these hills... it is important to keep the science goals clearly in sight, and to do something that astronomers actually want.

What we wish to work towards is both an open dialogue with the user community, and an actual list of example "use-cases". We have started this process by publishing our October 2000 proposal as a position paper, along with a Call For Comments. This was very successful, with around 15 quite lengthy and detailed responses from which we have learned a lot, and leading to the expansion of the consortium and its aims. The second step was the holding of the first AstroGrid workshop in Belfast in Jan 2001. This was not an open meeting, but rather an invited brainstorming, including international as well as UK participants. The next steps include (i) a talk to the UK National Astronomy Meeting in Cambridge; (ii) the construction of our AstroGrid Advisory Board (AGAB), (iii) further public meetings and position papers, and (iv) visits to individual groups to start to solicit use-cases.

(7.2) Other scientific disciplines

The work of the AstroGrid consortium will not take place in isolation and will need to be well connected with similar programmes being undertaken in other scientific disciplines. The obvious connection is with developments in Particle Physics as this area has been at the forefront of the building of the GRID as a whole and has a common funding source through PPARC, as well as a common oversight mechanism through the e-science Director and the e-science steering committee. Astronomy will benefit from the development of toolkits that are being driven by Particle Physics. P.Allport and A.Lawrence have been named as the respective link persons between the PP DataGrid and AstroGrid projects respectively. P.Allport will attend our Advisory Board meetings.   We also intend to organise specific collaborative meetings between AstroGrid partners and Particle Physics groups at the Universities and RAL. In addition, there will be a specific focus on working with the UK Particle Physics Tier 1 GRID node.

We will also need to have strong ties with other areas of science as they have particular strengths and requirements that will complement those of astronomy. In the area of climate studies, a consortium is planning to develop the necessary infrastructure for building coupled models that can communicate over the Internet. This mode of operation is directly applicable to problems that exist in the area of Solar Terrestrial Physics, where there is a desire to link models that cover different physical regimes. (Strictly speaking this area of coupled models is outside the AstroGrid remit but well inside the astro e-science remit, so we mention it here for completeness). In the area of bio-sciences, the requirements for handling the burgeoning amounts of data that are coming out of the human genome project, which will be dwarfed by the human brain project, have strong parallels with the requirements from astronomical data and metadata. The heterogeneity of biological data is most unlike that of Particle Physics, and even worse than that of Astronomy. When searching for data on the Internet, the Earth Observation community already has sophisticated facilities available to them that could be used more widely. Additionally, with the growing need to access telescopes at remote sites with great ease via the network, astronomy will benefit from development of control systems for large-scale instruments such as synchrotron light sources and neutron sources that include the processing chain as part of the overall system. Astronomy will benefit from many external developments, but we expect it be a net contributor of tools for the storage, access and processing of heterogeneous data. We expect that there will be collaborative generic projects in the area of coupled models.

In addition to these interactions with researchers in other sciences which are facing similar problems to those of AstroGrid, we shall also seek collaborative links with computer scientists. Several such links already exist. The group at QUB is a full partner in AstroGrid. We also have close working connections with both the RAL e-science department, and the Edinburgh Parallel Computer Centre. The EPCC is leading a proposal (GenGrid) to undertake generic and linking work; A.Lawrence is a co-I of that proposal. Regardless of the success of that proposal, we will expect to approach other computer science groups, such as the database group at Manchester, in order to make use of their specific expertise. Much of the expertise that AstroGrid will require in a number of key areas already exists in academic computer science, where researchers are keen to see their work put into practice on challenging applications, as provided by the large databases to be federated by AstroGrid. For example, there exists a mature and active field of computational geometry concerned with the description of multi-dimensional datasets, which informs both the choice of indexing schemes to use in databases for efficient querying and design of methods for analysing the data they contain. Computer scientists developing generic Grid tools are also seeking realistic examples of distributed processing upon which to test their protocols, and are eager to contribute effort to the development of such applications.

(7.3) The Global Virtual Observatory movement.

The idea of a "Virtual Observatory" has been building worldwide over the last year or so, with several conferences and workshops devoted to this idea. The US has the strongest track record of developments in this area in recent years (eg NED, the HEASARC website, NASA SkyView, and the Astrophysical Data Service). However the CDS in Strasbourg, responsible for SIMBAD and ALADIN, also occupies a central place. Two major projects have arisen.

The first is the US "National Virtual Observatory (NVO)" project. This was given a boost by being highlighted in the recent report of the Decadal Survey Committee as the top priority medium sized project, with an anticipated budget of around $60M over ten years. It is not yet clear whether this initiative will develop primarily through NSF funding, or as a NASA project. In November a "pre-proposal" was submitted to the NSF for a five year programme, headed by Messina and Szalay. The final proposal is in preparation. A paragraph in the NSF pre-proposal describes the AstroGrid project, and A.Lawrence and F.Murtagh are listed as international collaborators (along with other AVO principals - Quinn, Genova, and Benvenuti).

The second major project is the "Astrophysical Virtual Observatory (AVO)". This has developed out of the OPTICON working group on archive interoperability. A proposal to the EU Framework V RTD programme was submitted on February 15th. The partners are ESO, ESA, CDS, Terapix, AstroGrid, and Jodrell Bank. Each partner commits 2 staff years per year for three years to the AVO programme, and requests a further 2 staff years per year from the EU. The goals of the AVO programme are similar to that of AstroGrid, but without solar physics and space plasma physics. The key archives highlighted are those of ST-ECF and the VLT. The initial three year programme is seen as a development project , with an expected second three year programme to follow establishing a Europe-wide working system and possibly a physical user support centre. Like AstroGrid, the AVO plan has a strong emphasis on requirements analysis, but this is seen as extending over the whole three years. Within AVO, AstroGrid has agreed a lead responsibility in the area of grid computing developments and their application.

A possible third major project is EGSO, the European Grid of Solar Observations. Organisations across Europe active in solar physics archives are starting to come together with the intention of an EU bid analagous to the AVO project.

NVO, AVO, AstroGrid, and related developments in Canada led by CADC, represent a kind of fluid jigsaw. It is not even quite clear what the pieces are, let alone how they fit together. There has been no suggestion so far to develop a coherent "world project", but rather to keep the various projects in dialogue. Some duplication of work is inevitable and indeed desirable, as we explore alternative solutions. We obviously must work towards common data, metadata and interface standards, but these may evolve by both competition and co-operation. Our relationship with the US project is likely to remain informal. With AVO however we are committed to a much closer relationship, with part of AstroGrid inside AVO and part outside. We must therefore work towards more closely defined complementary tasks. For now, such close definition is not practicable, but one of AstroGrid's key deliverables from Phase A is a Phase B plan, and this should include agreed division of tasks with both AVO and NVO.

We cannot expect to *dominate* the construction of a world-wide Virtual Observatory. However by making clear and simple well recognised contributions, we can become a *leader* in this area. This points very strongly to AstroGrid (a) getting off the ground fast, and (b) having limited rather than ambitious goals.

(7.4) Work in external groups and commercial companies.

Only a limited number of groups are actually proposing to take responsibility for design and construction of AstroGrid, but this doesn't necessarily mean that all the work is done in-house. In part this means that we anticipate commissioning commercial software development, but we also expect to work with university astronomy groups outside the consortium. Sometimes this may be more or less a contract for the development of a particular software tool or component, but we also expect other examples to be of a more open-ended and collaborative nature. This seems particularly important as the whole idea of e-science has attracted considerable interest and attention across the astronomy community. It is obviously hard to predict events, but our working assumption has been that after top-level appointments (PM, PS, and the AGWC staff) that roughly half the staff effort will be external.

Collaborative relationships with the telecommunications and information technologies industries (both hardware and software) will also be pursued. The industry will be aware of the likely demands from commercial users and thus the potential market. The largest commercial databases are growing at a rate (TB/yr) quite similar to that of astronomy. Their desire for database access distributed across sites, correlation hunting among object fields, and resource discovery also has striking similarity. Thus, although Particle Physics will undoubtedly be leading the way in development of GRID toolkits, astronomy is probably closer to the commercial problem in nature and scale. Software vendors can be willing to work closely with demanding users of their products, as this can help them further develop the functionality of their products, and both software and hardware companies will provide expertise or products at a discount, in order to use their association with prestigious research projects for PR purposes. Initial conversations with some companies suggests however that their horizons are short and they will not invest in their own R&D programmes. Rather than simply waiting for us to deliver the R&D, a suggestion that has arisen is to second commercial programmers into academic groups, working to our tasks, but returning to their companies with new knowledge and skills. The development of these relationships may be fruitfully sought through the bodies coordinating e-science at the OST and PPARC level, as well as by AstroGrid itself.

(8) PROJECT BUDGET

(8.1) GENERAL ISSUES

Abbreviations. Below we use the term "HEFC" to mean HEFCE, SHEFC, or DENI as appropriate, and Y1,Y2,Y3 to refer to the assumed three financial years of the project, i.e. Y1=April 2001 - April 2002.

Total versus additional cost. The total project cost, the cost to PPARC, and the additional cost to PPARC, may all be different. The total cost may include the HEFC funded staff cost, and non-PPARC external funds such as EU funding, or commercial subsidy. The cost to PPARC includes all activity/equipment required to achieve the AstroGrid goals. It is recognised that some of these activities may already be underway, funded under existing lines. In this case the *additional* cost to PPARC may be smaller. Our aim here is to estimate the *total cost to PPARC*, regardless of what is considered new or not. HEFC staff effort is listed but at zero cost.

Limits of staff costs. Our intention is that AstroGrid starts with databases in place, adding value. Our cost estimates therefore do *not* include staff effort to develop or run pipelines, or to curate data. For AstroGrid to succeed however, that effort *must be present elsewhere in the PPARC programme*.

Limits of hardware costs. The data-mining machines that we plan to procure are an integral part of the AstroGrid project. Whether the storage hardware is inside or outside AstroGrid is harder to be clear on. This might seem clearly part of the parent projects (SOHO, XMM SSC, WFCAM, etc) but we are planning a greater degree of on-line availability, and more robust and higher quality solutions than might have otherwise been assumed. In addition, the correct storage solution (for example RAID vs Beowulf hang-ons) will not be clear until we have completed our experimental programme. We wish to make sure that the possible implied costs are clearly visible. We therefore include hardware storage costs in our table below, but clearly separated as an implied cost elsewhere rather than an integral AstroGrid cost.

Downstream implications. AstroGrid is a three year project. We do not intend to become a standing item. Activities in this area will obviously develop and continue, but we believe the community should re-group and re-propose in the light of the UK and international situation at that time. As well as general activity of this kind, there are obviously major projects which will dominate future datagrids - most obviously VISTA, ALMA, GAIA, and LISA. Most of these are some way downstream, but VISTA is only just out of the window. The tools we develop should be ready for VISTA, but PPARC should be aware that the storage and data-mining compute facilities for VISTA are not included here but hit us almost straight away after the current spending review cycle.

Staff effort volume. Our original estimate of required staff effort (in the October 2000 proposal) was based on the division in that paper into Work Packages that seemed reasonable for

1-2 dedicated staff each, arriving at 59 staff years in total. We maintain roughly the same approximate total, which obviously must be taken with a pinch of salt until the delivery of a Phase B plan. The Phase A plan is however much firmer and consumes 8 staff years of effort.

Internal and External spend. We have always assumed that a considerable fraction of staff effort would be placed outside the consortium, either as grants to other universities, or as commercial contracts. Of the astronomical proposals to the PPARC AO, it is quite possible that some of them may be effectively funded within the AstroGrid umbrella, but we cannot be sure until Phase A is well underway.

Commitment versus hold-back. A clear implication of our Phase A / Phase B structure, as well as the uncertainty in costs estimates, and the ambitions of other groups and projects, is that while a nominal budget envelope may be planned for AstroGrid, the majority of the money should be held back for a while. On the other hand, we need some degree of commitment for planning, for negotiation with international partners, and for recruitment. In our budget model, we make a specific proposal for what should be committed and what held back.

(8.2) BUDGET MODEL

Hardware requirements. The hardware we require will depend sensitively on the result of Phase A studies. For simplicity we have modelled this in two parts - mass storage, assumed to be RAID-like systems, and datamining machines, assumed to be PC clusters. These asssumptions may turn out to be incorrect - for instance, we may decide that SMP machines are better suited to the datamining research that users actually want to undertake, or we may decide that hanging storage off the back of cluster-nodes makes more sense than monolithic RAID systems. Or of course events may get overtaken by technological developments, such as the promised FMDs.

Mass storage needs. In the October 2000 proposal we estimated UK on-line data growth across all areas (OIR, X, solar, STP) to be +20TB, +30TB, +35TB in Y1,2,3. (Note that the full raw data is considerably larger). We have taken on-line storage costs to be £30K/TB, £20K/TB, £15K/TB in Y1,2,3. This represents fairly robust upper RAID-level SCSI storage, and is meant to be an effective price including maintenance, DBMS support, associated servers, and tape robot back-up systems. It is quite possible that cheaper mass storage options may be available but it is not yet clear what the best choice will be. We may not be able to commit very much early money, so we have chosen to slip the first two years requirements, buying nothing in Y1, 20TB in Y2, and 65TB in Y3. Funds needed are therefore 0K, 400K, 975K. As explained above, we list these costs separately, as an implied cost required elsewhere rather than an integral AstroGrid cost.

Datamining machines. Our immediate priority is for prototype data-mining machines for optical/IR, for XMM, for solar work, and for grid-experimentation - we assume four 16-node clusters. In Y3 the WFCAM project will need a much faster machine, assumed to be 200 nodes. (Other later projects may also need high-throughput machines, but are outside our window). We assume the use of PC clusters, and use a cost of 2K/node, 1K/node, 1K/node in Y1,2,3. This includes a 50GB disk hanging off the back of each node. We need four 16-node machines in Y1, one 200-node machine in Y3. Funds needed £128K, 0K, £200K.

Commercial licences. We are likely to require public access use of commercial software, for example DBMS systems. An example is the Objectivity license costing of the order £20K/site if being used for web access. We will need at least a couple of copies in Phase A, plus maybe similar items for evaluation. The total need is hard to estimate. We have allowed £50K/year.

Standard staff rate. For modelling purposes, we are assuming £60K/staff year including a travel allowance and personal computing allowance. This corresponds to a senior PDRA salary, say 30K + 22% NIS + 46% overhead = 54K, plus 3K per year travel and 3K per year towards workstation/laptop provision. This is also reasonably similar to the RAL dsy rate. The actual spend will vary wildly, but this is a reasonable average. This staff rate is assumed to apply whether staff are employed within an AstroGrid consortium institution, or as a grant to another university. We also expect however to use commercial effort. For purchased external programmer effort, a budget of more like £90K/yr would be expected, and sometimes we will certainly pay rates like these. However a promising alternative is that software companies may second staff to University groups for finite periods, charging actual salary cost. (The University would receive a grant-like overhead). For this method, £60K/yr is again a reasonable cost.

Project Management. We wish to hire a Project Manager (PM), Project Scientist (PS), and Project Assistant (PA) as soon as possible. Our working assumption is that this means staff in post in October 2001. The PM will be an open recruitment. The PS may also be an open recruitment, but could be internal selection. Project Assistance will probably in reality be several bodies - for example one whole office assistant co-located with the PM, and distributed clerical assistance across our institutions. For simplicity we model this as 1.0sy at the standard staff rate. The total staff cost is therefore 3sy/yr , but only for half the first year.

Lead Investigator Effort. We assume that the eight academic/senior establishment staff listed as Lead Investigators will each contribute 0.1sy/yr. In the table below all this staff effort is listed, but the academic staff are considered HEFC-funded, and so zero-cost. The Lead Investigators from RAL (Allan, Lockwood) are costed at our standard staff rate.

Re-deployment of existing staff. We hope to begin some recruitment immediately, but want to start real work immediately. A number of staff have already been active, and are committed to begin dedicating a large fraction of their effort to AstroGrid. These staff are all currently funded through other programmes, so are re-deployments. The level of deployment is an interim working arrangement but may be revisited as we begin our recruitment programme. The individuals are :

70%	Page, Rixon
50%	Mann, Osborne, Davenhall, Bentley, Stamper, Pike

30%	Richards	TOTAL 5.4 sy/yr
20%	Sherman, Giaretta, Perry	
10%	Irwin	

PDRA recruitments. To provide the total staff effort needed to complete the Phase A workplan, we need to recruit two further PDRAs/programmers in Y1. We assume we will achieve staff in post by October 2001, thus costing 1.0sy total in Y1, continuing the same staff on three year contracts, and so committing 2sy in each of Y2 and Y3. Beyond this we do not have a concrete workplan, but do have a preliminary model. We assume the total staff effort as in the October 2000 proposal, and assume that roughly half of this is in-house. To achieve this, we then make a preliminary assumption of three further PDRAs in April 02 (two-year contracts), and two more in October 02 (eighteen month contracts).

External staff effort. We are unlikely to make external contracts until Phase B, so the estimate is very preliminary. We model this as 11sy/yr at the standard staff rate, starting in Y2. (If substantial commercial contracts are used, the number of staff years of effort returned for the same money will be less.

(8.3) BUDGET MODEL

Below is our current budget model using the above assumptions.

Proposal on committment of funds. Our proposal is that we be given permission to begin recruitment immediately for PM, PS, PA, and two PDRAs. In addition, we need some immediate equipment and license procurement for Phase A. The rest of the estimated funds required could be held back. In other words the proposed committment = Y1+Y2 requisitions, staff obligations to 5.6sy from April 01, and 5.0sy from October 01).

ASTROGRID BUDGET TABLE

	Y1	Y2	Y3	total	
REQUISITIONS (k£)					
datamining	**128**	0	200	328	
licenses	**50**	**50**	50	150	
TOTAL REQUISITIONS	*178*	*50*	*250*	*478*	
STAFF EFFORT (sy)					
HEFC Leads	0.6	0.6	0.6	1.8	(zero cost)
CLRC Leads	**0.2**	**0.2**	**0.2**	**0.6**	(at 60K/sy)
re-deployment	**5.4**	**5.4**	**5.4**	**16.2**	
PM/PS/PA	**1.5**	**3.0**	**3.0**	**7.5**	
P1,P2 (Oct 01)	**1.0**	**2.0**	**2.0**	**5.0**	
P3,P4,P5 (April 02)	0.0	3.0	3.0	6.0	
P5,P6 (October 02)	0.0	1.0	2.0	3.0	
External effort	0.0	11.0	11.0	22.0	

TOTAL STAFF	*8.7*	*26.2*	*27.2*	*62.1*
STAFF COSTS (k£)	486	1536	1596	3618
REQUISITIONS (k£)	178	50	250	478
TOTAL COSTS (k£)	*664*	*1586*	*1846*	*4096*
Proposed commitment	*664*	*686*	*636*	*1986*
Proposed hold-back	*0*	*900*	*1210*	*2110*
ADDITIONAL IMPLIED COST OUTSIDE ASTROGRID				
mass storage	0	400	975	1375
creation of databases	TBD	TBD	TBD	TBD

Note : bold items are proposed committments.